

# IMPERIAL

## AUDIO EXPLAINABLE ARTIFICIAL INTELLIGENCE

Alican Akman

Submitted in part fulfilment of the requirements for the degree of  
PhD in Computing and the Diploma of Imperial College London.

Department of Computing  
Imperial College London  
2025

# Abstract

This thesis explores the potential of audio Explainable Artificial Intelligence (XAI) to improve the interpretability of deep audio processing models. While most existing XAI methods focus on visual and textual explanations, audio explanations offer a more intuitive approach for audio-based tasks. They provide a unique level of expressiveness, particularly where visual explanations require specialised knowledge. By aligning explanations with the audio domain, this research aims to bridge the interpretability gap and enhance understanding of complex audio models.

As a case study, this thesis examines COVID-19 detection from cough and speech audio, considering both classifier performance and explainability. To enhance interpretability, CoughLIME, a modified version of LIME tailored for cough data, is introduced. CoughLIME generates faithful and listenable explanations, addressing a key challenge in trust for audio-based COVID-19 classifiers.

With transformer models excelling in audio processing, the need for interpretability of their complex decision-making has grown. This thesis proposes a technique to explain audio-processing transformers by integrating their attention mechanisms with non-negative matrix factorisation (NMF). NMF decomposes audio into spectral patterns, while attention weights identify the most relevant time activations. By reconstructing key audio components, the method generates high-fidelity, listenable explanations, validated through audio classification tasks.

Additionally, a novel explanation method is introduced, leveraging the meaningful representation space and generative capacity of audio foundation models. By integrating feature attribution techniques, significant features in their embedding space are identified, enabling the generation of meaningful audio explanations. Extensive evaluations explaining audio classification models confirm the effectiveness of this approach.

Finally, we propose a novel framework, extending beyond traditional feature attribution which emphasise only the most relevant features, overlooking the broader representational space, including less important features. Rather than removing features, the framework uses generative audio language models to replace removed features with contextually appropriate alternatives, offering a more comprehensive understanding of model behaviour.

# Declaration of Originality

I declare that this thesis is my own original work and that all sources have been properly acknowledged. Any collaborative contributions have been clearly indicated. This thesis has not been submitted for any previous degree or qualification.

# Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC).

Under this license, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the license terms clear to others by naming the license and linking to the license text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this license or permitted under UK Copyright Law.

# Acknowledgments

First and foremost, I would like to express my deepest gratitude to my wife, Hande, for her unwavering support, love, and for being an inseparable part of my life. Her presence has been my greatest source of strength throughout this journey. I also extend my heartfelt love to our baby girl, Alina, who joined our lives just a month before my viva. Both in the anticipation of her arrival during my final year and in her earliest days with us, she has brought immense joy and motivation to the final stages of my PhD.

I am immensely grateful to my family and Hande's family for their endless support, not only during my PhD but throughout my entire life. Their selflessness, prioritising our well-being and making sacrifices along the way, has given me the confidence to take bold steps without hesitation. Knowing that their support is infinite has been a cornerstone of my resilience.

I would like to sincerely thank my PhD supervisor, Björn, for his invaluable intellectual guidance, insightful ideas, and mentorship. His positive attitude has shaped both my research and career, not just from an academic perspective but also as a mentor who has motivated me and pushed me to expand my boundaries.

To my dear friends at Imperial, thank you for creating such a collaborative and welcoming environment. A special thanks to George, whose warm welcome in my early years in the group made a significant impact, and who has since become a great friend. I am also grateful to Harry for our early research collaborations, which helped me adapt to the research environment, and for his friendship. A big thanks to Qiyang for the time we spent together, both working and having fun; our moments together made the final days of my PhD truly enjoyable. I also want to thank Yasin for our coffee and lunch meetings, our friendship, and our many productive discussions about our goals, which have been both inspiring and motivating. I am also grateful my lifelong friends, Ahmet Serdar Mutluer, Ali Ari, Furkan Eris, Mustafa Emin Oktay, and Yusuf Aydin. Our friendship of more than fifteen years has been a constant source of support, and this is not their first appearance in my acknowledgments; I hope it will not be the last.

I would also like to extend my sincere gratitude to the Turkish Ministry of Education for their financial support throughout my PhD. Their sponsorship has played a crucial role in making this

journey possible, allowing me to focus on my research and academic development without financial concerns.

Lastly, I extend my appreciation to the excellent staff at Imperial for their continuous support throughout this journey. A special thanks to Amani, who has always looked after us PhD students, patiently answering our questions and solving our problems, making this process as smooth as possible.

This journey has been a collective effort, and I am forever grateful to everyone who has been a part of it. Like every achievement in my life, what makes completing my PhD truly meaningful is sharing it with the people I love.

*To my beloved wife and partner, Hande, and to our little princess, Alina, whose arrival during my doctoral journey became the greatest gift of our lives.*

# Contents

|                                                                           |            |
|---------------------------------------------------------------------------|------------|
| <b>Abstract</b>                                                           | <b>i</b>   |
| <b>Declaration of Originality</b>                                         | <b>ii</b>  |
| <b>Copyright Declaration</b>                                              | <b>iii</b> |
| <b>Acknowledgments</b>                                                    | <b>iv</b>  |
| <b>List of Publications</b>                                               | <b>x</b>   |
| <b>List of Acronyms</b>                                                   | <b>xii</b> |
| <b>List of Figures</b>                                                    | <b>xiv</b> |
| <b>List of Tables</b>                                                     | <b>xv</b>  |
| <br>                                                                      |            |
| <b>1 Introduction</b>                                                     | <b>1</b>   |
| 1.1 Research Questions . . . . .                                          | 3          |
| 1.2 Contributions . . . . .                                               | 5          |
| 1.3 Thesis Outline . . . . .                                              | 6          |
| <br>                                                                      |            |
| <b>2 Background</b>                                                       | <b>8</b>   |
| 2.1 Deep Learning . . . . .                                               | 8          |
| 2.1.1 Recent Advancements . . . . .                                       | 9          |
| 2.1.2 Applications . . . . .                                              | 12         |
| 2.1.3 Deep Models for Audio Processing . . . . .                          | 13         |
| 2.2 Explainable Artificial Intelligence . . . . .                         | 14         |
| 2.2.1 XAI Foundations . . . . .                                           | 15         |
| 2.2.2 XAI Methods . . . . .                                               | 16         |
| 2.2.3 XAI Evaluation . . . . .                                            | 24         |
| 2.3 Audio Explainability for Deep Models . . . . .                        | 28         |
| 2.3.1 Explaining Audio Models using Domain-Agnostic XAI Methods . . . . . | 29         |
| 2.3.2 Audio-specific XAI Methods . . . . .                                | 34         |

---

|          |                                                                             |           |
|----------|-----------------------------------------------------------------------------|-----------|
| <b>3</b> | <b>Interpretable Audio Classification with Data Decomposition</b>           | <b>39</b> |
| 3.1      | Introduction . . . . .                                                      | 40        |
| 3.2      | Datasets . . . . .                                                          | 42        |
| 3.2.1    | ComParE . . . . .                                                           | 42        |
| 3.2.2    | DiCOVA . . . . .                                                            | 43        |
| 3.3      | Methodology . . . . .                                                       | 43        |
| 3.3.1    | Classifying COVID-19 . . . . .                                              | 43        |
| 3.3.2    | Explaining COVID-19 Classifiers using CoughLIME . . . . .                   | 45        |
| 3.4      | Experiments . . . . .                                                       | 48        |
| 3.4.1    | COVID-19 Classification Results . . . . .                                   | 48        |
| 3.4.2    | Explanation Results . . . . .                                               | 50        |
| <b>4</b> | <b>Understanding Audio Transformers through Attention and Factorisation</b> | <b>54</b> |
| 4.1      | Introduction . . . . .                                                      | 55        |
| 4.2      | Methodology . . . . .                                                       | 56        |
| 4.2.1    | NMF Fundamentals . . . . .                                                  | 56        |
| 4.2.2    | AttHear Design . . . . .                                                    | 57        |
| 4.2.3    | Explanation Generation . . . . .                                            | 58        |
| 4.3      | Experiments . . . . .                                                       | 59        |
| 4.3.1    | Implementation Details . . . . .                                            | 59        |
| 4.3.2    | Quantitative Evaluation . . . . .                                           | 59        |
| 4.3.3    | Qualitative Evaluation . . . . .                                            | 62        |
| <b>5</b> | <b>Audio Explanation Synthesis with Generative Foundation Models</b>        | <b>63</b> |
| 5.1      | Introduction . . . . .                                                      | 64        |
| 5.2      | Methodology . . . . .                                                       | 65        |
| 5.2.1    | Audio Foundation Models . . . . .                                           | 66        |
| 5.2.2    | System Design . . . . .                                                     | 67        |
| 5.2.3    | Explanation Generation . . . . .                                            | 67        |
| 5.3      | Experiments . . . . .                                                       | 68        |
| 5.3.1    | Implementation Details . . . . .                                            | 68        |
| 5.3.2    | Quantitative Evaluation . . . . .                                           | 70        |
| 5.3.3    | Qualitative Evaluation . . . . .                                            | 72        |

---

|          |                                                                 |            |
|----------|-----------------------------------------------------------------|------------|
| <b>6</b> | <b>Improving Audio Explanations using Audio Language Models</b> | <b>74</b>  |
| 6.1      | Introduction . . . . .                                          | 75         |
| 6.2      | Methodology . . . . .                                           | 77         |
| 6.2.1    | Audio Language Models . . . . .                                 | 77         |
| 6.2.2    | Explainer Design . . . . .                                      | 78         |
| 6.3      | Experiments . . . . .                                           | 79         |
| 6.3.1    | Implementation Details . . . . .                                | 79         |
| 6.3.2    | Quantitative Evaluation . . . . .                               | 82         |
| 6.3.3    | Qualitative Evaluation . . . . .                                | 84         |
| <b>7</b> | <b>Conclusions</b>                                              | <b>86</b>  |
| 7.1      | Contributions Summary . . . . .                                 | 86         |
| 7.2      | Ethics . . . . .                                                | 89         |
| 7.3      | Future Work . . . . .                                           | 93         |
| <b>8</b> | <b>Appendix</b>                                                 | <b>121</b> |
| 8.1      | Codes . . . . .                                                 | 121        |
| 8.1.1    | CoughLIME (Used in Chapter 3) . . . . .                         | 121        |
| 8.1.2    | AttHear (Used in Chapter 4) . . . . .                           | 122        |
| 8.1.3    | AudioXgen (Used in Chapter 5) . . . . .                         | 122        |
| 8.1.4    | AudioXLM (Used in Chapter 6) . . . . .                          | 123        |
| 8.2      | Datasets . . . . .                                              | 124        |
| 8.2.1    | DiCOVA Challenge Dataset . . . . .                              | 125        |
| 8.2.2    | Speech Commands . . . . .                                       | 126        |
| 8.2.3    | Environmental Sound Classification (ESC-50) . . . . .           | 127        |
| 8.2.4    | Toronto Emotional Speech Set (TESS) . . . . .                   | 128        |

# List of Publications

In the following, the asterisk symbol (\*) denotes equal co-authorship credit.

## A. Core Publications Associated with Thesis

- Anne Wullenweber\*, **Alican Akman**\*, & Björn W. Schuller. (2022). Coughlime: Sonified explanations for the predictions of covid-19 cough classifiers. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1342–1345. <https://doi.org/10.1109/EMBC48229.2022.9871291>
- **Alican Akman**\*, Harry Coppock\*, Alexander Gaskell, Panagiotis Tzirakis, Lyn Jones, & Björn W. Schuller. (2022). Evaluating the covid-19 identification resnet (cider) on the interspeech covid-19 from audio challenges. *Frontiers in Digital Health*, 4. <https://doi.org/10.3389/fdgth.2022.789980>
- **Alican Akman** & Björn W. Schuller. (2024b). Audio explainable artificial intelligence: A review. *Intelligent Computing*, 3, 0074. <https://doi.org/10.34133/icomputing.0074>
- **Alican Akman** & Björn W. Schuller. (2024a). Atthear: Explaining audio transformers using attention-aware nmf. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7015–7019. <https://api.semanticscholar.org/CorpusID:268596433>
- **Alican Akman**, Qiyang Sun, & Björn W. Schuller. (2025). Audio explanation synthesis with generative foundation models. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://arxiv.org/abs/2410.07530>
- **Alican Akman**, Qiyang Sun, & Björn Schuller. (2025). Improving audio explanations using audio language models. *IEEE Signal Processing Letters*, PP, 1–5. <https://doi.org/10.1109/LSP.2025.3532218>

## B. Additional Publications

- Dominika Woszczyk, Anna Hedlikova, **Alican Akman**, Soteris Demetriou, & Björn Schuller. (2022). Data augmentation for dementia detection in spoken language. *Interspeech 2022*, 2858–2862. <https://doi.org/10.21437/Interspeech.2022-10210>
- Harry Coppock, **Alican Akman**, Christian Bergler, Maurice Gerczuk, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, Jing Han, Shahin Amiriparian, Alice Baird, Lukas Stappen, Sandra Ottl, Panagiotis Tzirakis, Anton Batliner, Cecilia Mascolo, & Björn W. Schuller. (2023). A summary of the compare covid-19 challenges. *Frontiers in Digital Health*, 5. <https://doi.org/10.3389/fdgth.2023.1058163>
- Björn W. Schuller, **Alican Akman**, Yi Chang, Harry Coppock, Alexander Gebhard, Alexander Kathan, Esther Rituerto-González, Andreas Triantafyllopoulos, & Florian B. Pokorny. (2024). Ecology & computer audition: Applications of audio technology to monitor organisms and environment. *Helicon*, 10(1), e23142. <https://doi.org/https://doi.org/10.1016/j.helicon.2023.e23142>

- Chia-Hsin Lin, Charles Jones, Björn W. Schuller, Harry Crippock, & **Alican Akman**. (2024). Synthia’s melody: A benchmark framework for unsupervised domain adaptation in audio. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7450–7454. <https://doi.org/10.1109/ICASSP48485.2024.10446369>
- Qiyang Sun, **Alican Akman**, Xin Jing, Manuel Milling, & Björn W. Schuller. (2025). Audio-based kinship verification using age domain conversion. *IEEE Signal Processing Letters*, 32, 301–305. <https://doi.org/10.1109/LSP.2024.3515811>
- Qiyang Sun, **Alican Akman**, & Björn W. Schuller. (2025). Explainable artificial intelligence for medical applications: A review. *ACM Trans. Comput. Healthcare*, 6(2). <https://doi.org/10.1145/3709367>

# List of Acronyms

**AI** Artificial Intelligence

**XAI** Explainable Artificial Intelligence

**MLP** Multilayer Perceptron

**CNN** Convolutional Neural Network

**RNN** Recurrent Neural Network

**NMF** Non-negative Matrix Factorisation

**ALM** Audio Language Models

**NLP** Natural Language Processing

**ViT** Vision Transformers

**LSTM** Long Short-Term Memory

**ASR** Automatic Speech Recognition

**MFCC** Mel-Frequency Cepstral Coefficients

**VAE** Variational Autoencoders

**GAN** Generative Adversarial Networks

**DTD** Deep Taylor Decomposition

**RISE** Randomised Input Sampling to Provide Explanations

**LORE** Local Rule-based Explanation

**IG** Integrated Gradients

**CAM** Class Activation Maps

**Grad-CAM** Gradient-weighted CAM

**LRP** Layer-wise Relevance Propagation

**DeepLIFT** Deep Learning Important FeaTures

**SHAP** Shapley Additive Explanation

**TCAV** Quantitative Testing with Concept Activation Vectors

**SSL** self-Supervised Learning

**COVID-19** Coronavirus Disease 2019

**LFT** Lateral Flow Test

**ComParE** the INTERSPEECH 2021 Computational Paralinguists Challenge

**DiCOVA** Diagnosing COVID-19 using acoustics

**LR** Logistic Regression

**RF** Random Forrest

**UAR** Unweighted Average Recall

**STFT** short-time Fourier transform

**GRU** Gated Recurrent Unit

# List of Figures

|     |                                                                                                                                                                      |     |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 2.1 | Illustration of the architectures of (a) VAE, (b) GAN where Enc, Gen, Dec, and Disc denote the encoder, generator, decoder, and discriminator, respectively. . . . . | 10  |
| 2.2 | Illustration of the transformer architecture adapted from [Vaswani et al., 2017]. . .                                                                                | 11  |
| 2.3 | An overview of the taxonomy of the XAI methods. . . . .                                                                                                              | 18  |
| 2.4 | Examples of feature-attribution-based XAI methods on audio models. . . . .                                                                                           | 32  |
| 2.5 | Example explanations from CoughLIME [Wullenweber et al., 2022] using loudness decomposition. . . . .                                                                 | 37  |
| 2.6 | Examples of interpretation of audios generated from [Parekh et al., 2022]. . . . .                                                                                   | 38  |
| 3.1 | A schematic of the <b>COVID-19 Identification ResNet, (CIdR)</b> . . . . .                                                                                           | 44  |
| 3.2 | Example loudness decomposition of a cough audio file with threshold 75 dB. . . . .                                                                                   | 46  |
| 3.3 | Example NMF decomposition of a cough audio file into 6 components consisting of spectral profiles (a) and temporal activations (b). . . . .                          | 47  |
| 3.4 | Pixel flipping for audio for the loudness decomposition with $\gamma = 75dB$ shows CoughLIME is able to correctly identify the most important components. . . . .    | 51  |
| 3.5 | Example explanations for COVID-19 positive (a), (b) and COVID-19 negative (c), (d) coughs generated with the loudness decomposition. . . . .                         | 53  |
| 4.1 | An overview of AttHear. . . . .                                                                                                                                      | 57  |
| 4.2 | Sample spectrogram visualisations for the mixed audio experiments. . . . .                                                                                           | 61  |
| 5.1 | An overview of our method: The top row depicts the role of a foundation model with autoencoder architecture. . . . .                                                 | 66  |
| 5.2 | An overview of the Encodec model: an encoder-decoder architecture trained with both reconstruction and adversarial losses. . . . .                                   | 69  |
| 5.3 | Sample spectrogram visualisations for the qualitative audio experiments. . . . .                                                                                     | 73  |
| 5.4 | Confusion matrix for the classifier for TESS dataset after explanation removal by a ratio of $\beta = 0.1$ . . . . .                                                 | 73  |
| 6.1 | An overview of our method. . . . .                                                                                                                                   | 77  |
| 6.2 | An overview of the AudioGen framework adapted from [Kreuk et al., 2023]. . . . .                                                                                     | 80  |
| 6.3 | Sample audio waveform visualisations from the qualitative audio experiments. . . . .                                                                                 | 85  |
| 8.1 | Distribution of subjects in the DiCOVA Challenge dataset based on symptoms, age, and gender. . . . .                                                                 | 125 |

# List of Tables

|     |                                                                                                                                                               |     |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 2.1 | Overview of generic XAI methods applicable to audio processing models. . . . .                                                                                | 30  |
| 2.2 | Overview of XAI methods specialised on audio processing models. . . . .                                                                                       | 36  |
| 3.1 | ComParE sub-challenge dataset splits. . . . .                                                                                                                 | 42  |
| 3.2 | DiCOVA sub-challenge dataset splits. . . . .                                                                                                                  | 43  |
| 3.3 | Results for CIDEr and a range of baseline models for 4 sub-challenges across the DiCOVA and ComParE challenges. . . . .                                       | 49  |
| 3.4 | The results for cross dataset experiments. . . . .                                                                                                            | 50  |
| 3.5 | Comparison of the $\Delta$ AUC achieved with different decompositions. . . . .                                                                                | 52  |
| 4.1 | Fidelity and faithfulness results on the Speech Commands dataset, with mean and standard deviation over five runs. . . . .                                    | 60  |
| 4.2 | Fidelity and faithfulness results on the ESC-50 dataset, with mean and standard deviation over five runs. . . . .                                             | 60  |
| 5.1 | Fidelity results by measuring explanation classification agreement on the Speech Commands dataset, with mean and standard deviation over five runs. . . . .   | 70  |
| 5.2 | Fidelity results by measuring explanation classification agreement on the TESS dataset, with mean and standard deviation over five runs. . . . .              | 71  |
| 5.3 | Fidelity results by measuring accuracy drop over explanation removal on the Speech Commands dataset, with mean and standard deviation over five runs. . . . . | 71  |
| 5.4 | Fidelity results by measuring accuracy drop over explanation removal on the TESS dataset, with mean and standard deviation over five runs. . . . .            | 72  |
| 6.1 | Explanation Generation Time for Different Methods . . . . .                                                                                                   | 82  |
| 6.2 | Fidelity results by measuring explanation classification agreement on the Speech Commands dataset, with mean and standard deviation over five runs. . . . .   | 83  |
| 6.3 | Fidelity results by measuring explanation classification agreement on the TESS dataset, with mean and standard deviation over five runs. . . . .              | 83  |
| 6.4 | Automatic speech recognition results by measuring WER on the TESS dataset, with mean and standard deviation over five runs. . . . .                           | 84  |
| 8.1 | Word list and corresponding number of utterances. . . . .                                                                                                     | 127 |
| 8.2 | Categorisation of sound classes in the ESC-50 dataset, organised into five major groups . . . . .                                                             | 128 |

# 1

## Introduction

### Contents

---

|                                        |                   |
|----------------------------------------|-------------------|
| <a href="#">1.1 Research Questions</a> | <a href="#">3</a> |
| <a href="#">1.2 Contributions</a>      | <a href="#">5</a> |
| <a href="#">1.3 Thesis Outline</a>     | <a href="#">6</a> |

---

Recent advancements in Artificial Intelligence (AI) technology have revolutionised numerous domains and have been widely used in various industries such as healthcare, finance, manufacturing, transportation, and education. Popular AI methods include a range of techniques, from standard machine learning models such as decision trees and linear regression to more complex deep learning models such as Multilayer Perceptron (MLP) [Rumelhart et al., 1986], Convolutional Neural Network (CNN) [Krizhevsky et al., 2012], Recurrent Neural Network (RNN) [Sherstinsky, 2020], and transformers [Vaswani et al., 2017]. Nowadays, the transformer-based models have demonstrated their versatility across various domains, ranging from natural language processing tasks [Devlin et al., 2018, Peters et al., 2018, Lan et al., 2019] such as machine translation, sentiment analysis, and question answering, to computer vision tasks [Touvron et al., 2021, Radford et al., 2021, Khan et al., 2022] such as image generation, object detection, and image segmentation. With advancements in both model architectures and hardware technologies, state-of-the-art deep models, such as GPT-3 [Brown et al., 2020b] and ViT [Dosovitskiy et al., 2020] evolved to large sizes and complexities, pushing the boundaries with hundreds of billions of parameters. Although an increase in complexity has improved the accuracy of these models, the challenge of interpretability

has intensified, making it difficult to understand the inner workings and decision-making processes of these black-box models.

Explainable Artificial Intelligence (XAI) methods play a crucial role in the decision-making processes of AI systems by providing transparency and trust. These XAI properties justify the AI model prediction, which builds trust among human users and debugs the model when necessary. Exploiting XAI methods is inevitable, particularly in sectors such as healthcare [Amann et al., 2020, Reddy, 2022], law [Deeks, 2019], transportation [Atakishiyev et al., 2021] and finance [Weber et al., 2023] where the decisions made have critical consequences. For example, XAI helps doctors and medical professionals understand and trust the recommendations made by AI models, thereby facilitating better patient care and treatment planning. In finance, XAI explains credit scoring decisions, aiding banks in understanding factors that influence the credibility of individuals or companies and mitigating potential bad debts. XAI aids in interpreting legal documents and precedents, assisting lawyers in building stronger cases, and ensuring fair judgments. Furthermore, XAI empowers autonomous vehicles by explaining their decision-making processes and ensuring safety and reliability while building public acceptance of this cutting-edge technology.

The modality of explanations generated by an XAI model plays an important role in increasing the level of interpretability and helping human users better understand the model, depending on use-case scenarios. Numerous XAI methods use numerical, rule-based, textual, and visual explanations, all presented visually, to enhance communication with human users. Examples of numerical explanations include providing feature importance scores, class probabilities, or weights such as coefficients in linear models or importance values in tree-based models to quantify the impact of different variables on the model’s inference. Rule-based explanations include presenting logical rules or decision paths that the model follows to arrive at a decision, making the logic behind the model’s output clear, and making its decision-making process transparent. Visual explanations leverage the communication power of images, heat maps, or saliency maps to highlight important features or parts of an input data point, enabling users to understand the factors influencing the model’s decision-making process. Textual explanations involve the clarification of natural language statements. Recent surveys [Barredo Arrieta et al., 2020, Vilone and Longo, 2020] summarise the field of XAI regarding methods that provide explanations using these visually appealing or human-readable formats.

In addition to these formats, audio explanations offer a unique way of presenting explanations that appeal to auditory perception. This includes providing speech, a sound signal, or parts of the input audio that are derived from the original input of audio tasks, or sonification of non-

audio formats for other modalities. In this thesis, my research goal is to explain audio models by providing listenable explanations, leveraging the expressiveness of audio to enhance interpretability. To achieve this, I approach the problem from both the audio data and deep model perspectives, aiming to answer three key questions: which parts of the audio are most important for a model’s decision, how this information is processed within the embedding space of neural networks, and what is the best way to present these as audio explanations to end-users to ensure transparency.

In this thesis, we aim to improve audio-based [XAI](#) methods by developing an approach that targets different aspects of audio interpretation. A key aspect of producing meaningful audio explanations is decomposition, which allows for the investigation of relevant components contributing to the final decisions of AI models. We explore both simple audio decomposition techniques as well as more sophisticated methods, such as Non-negative Matrix Factorisation ([NMF](#)). Another major focus of this work is the focus on explaining state-of-the-art models by exploring their internal structure. Specifically, we apply our approach to transformer-based models by leveraging their attention mechanisms in combination with NMF.

Deep models inherently encode high-level features, providing a natural basis for audio decomposition. To leverage this, we also utilise audio foundation models, employing their embedding space to extract semantically meaningful features that contribute to more interpretable explanations. This allows us to capture essential aspects of audio beyond raw waveforms or spectrograms. Building on this, we further exploit audio-language models to generate complete and coherent audio explanations, aligning audio representations with textual descriptions. This integration not only improves the interpretability of audio-based AI models but also provided a more structured and human-understandable way of explaining complex auditory patterns.

## 1.1 Research Questions

The primary research questions explored in this thesis can be summarised as follows:

**Research Question 1 (RQ-1):** Given that explainability refers to the details that make an AI model’s decision process understandable, what are the best practices for explaining audio-processing models? Furthermore, how does the explainability of audio-processing models differ from that of AI models in other major domains, such as computer vision and natural language processing, considering the unique dynamics of audio processing?

**Research Question 2 (RQ-2):** Beyond the unique characteristics of audio processing and the specialised properties of AI models designed for audio modality, the effectiveness of the output format in audio XAI methods is a critical consideration. While visual representations, such as frequency interval visualisations, may aid expert users in understanding model decision processes, they are often not intuitive for end-users from diverse backgrounds. What is the significance of providing listenable explanations in enhancing the interpretability of audio models for end-users?

**Research Question 3 (RQ-3):** XAI methods primarily focus on analysing the importance of the "smallest meaningful units" in different modalities, such as superpixels in images or words in natural language processing, in the decision-making process of AI models. These methods typically operate by first identifying meaningful features and then assigning importance to them. In the context of audio signals, what constitutes the smallest meaningful units, and how can these features be effectively extracted using audio decomposition techniques to facilitate the application of standard feature attribution methods?

**Research Question 4 (RQ-4):** Unlike image matrices and discrete language tokens, audio data can be represented through various methods, such as spectrograms and raw waveform signals. What role do these different representations of audio signals play in their decomposition, and how do they contribute to the extraction of explainable features?

**Research Question 5 (RQ-5):** Instead of relying on deterministic audio decomposition techniques, can more advanced linear algebra-based methods, such as NMF, be leveraged? Do these methods inherently produce meaningful features that are important for audio tasks, and can these features be reconstructed into listenable explanations?

**Research Question 6 (RQ-6):** As AI algorithms advance and hardware technologies continue to improve, deep learning models are becoming increasingly complex. For instance, transformers are widely employed to achieve state-of-the-art performance across various tasks. Can the inherent architectural features of transformers, such as attention weights, be leveraged to enhance feature attribution methods for audio explainability?

**Research Question 7 (RQ-7):** Audio foundation models have gained popularity due to their ability to learn meaningful representation spaces and capture distinctive audio features by training on large-scale audio datasets. Can the high-level features embedded in the rich representational space of audio foundation models serve as an alternative to manual data decomposition methods for extracting meaningful explanations?

**Research Question 8 (RQ-8):** Audio language models have demonstrated significant performance in generating audio content. However, existing feature attribution methods typically focus on highlighting the most relevant features while disregarding less important ones, thereby narrowing the representational space. By shifting this paradigm, can the generative capabilities of audio language models be harnessed to replace removed features with contextually appropriate alternatives, offering a more comprehensive view of the model’s decision-making process?

## 1.2 Contributions

Building on the research objectives outlined above, the key contributions of this work can be summarised as follows.

We advance the explainability of audio-based machine learning models in healthcare, focusing specifically on COVID-19 detection. The chapter first provides empirical evidence of the challenges in developing universally reliable COVID-19 classifiers, highlighting issues such as dataset bias and limited generalisability. We then introduce CoughLIME, a novel interpretability method that translates model explanations into the audio domain, allowing researchers and clinicians to better understand the acoustic features driving predictions. Together, these efforts address the broader question of how trustworthy and interpretable audio health diagnostics can be designed by providing listenable explanations, while also underscoring the current limitations that constrain the reliability of such systems.

We then propose AttHear, a novel audio explanation method designed specifically for audio transformers. AttHear leverages the attention module in transformers and integrates it with NMF, eliminating the need to fine-tune the original prediction model or train a surrogate model. To achieve this, we introduce a new formulation for NMF reconstruction that incorporates attention weights into NMF time activations. This approach enables AttHear to generate audio explanations in a listenable format, making them interpretable and accessible to end-users. We evaluate AttHear on keyword spotting and environmental sound classification tasks, demonstrating that it produces faithful explanations with high fidelity scores while effectively capturing meaningful high-level audio components relevant to these tasks.

Next, by incorporating common feature attribution methods into the latent space of foundation models, we introduce an innovative audio explanation approach. This method leverages the meaningful latent space to avoid the challenges of interpreting feature relevance directly in the

input space, such as audio frequencies, where individual features can be difficult to understand. Instead, it utilises the generative capabilities of foundation models to create audio explanations in a listenable format, making them both meaningful and interpretable for end-users. We evaluate this approach on keyword spotting and speech emotion recognition tasks, demonstrating that it generates high-fidelity explanations while capturing significant high-level audio components. Additionally, we propose a strong baseline using NMF with random attention to quantitatively validate the performance of our method.

Finally, we combine Audio Language Models (ALM) with foundation model-based audio classifiers to provide a deeper and more comprehensive understanding of important audio features. By leveraging the high-level latent space of foundation models, our approach identifies meaningful audio features that influence model decisions. The generative capabilities of ALMs are then used to refine and complete these features within the latent space, producing listenable, interpretable audio explanations. This offers a compelling alternative to traditional feature selection methods in the XAI literature, enabling richer and more coherent explanations. We evaluate our framework on keyword spotting and speech emotion recognition tasks, showing that it not only delivers high-fidelity explanations but also enhances the overall understandability of these explanations. Notably, emotion-focused explanations lead to improved speech recognition performance, demonstrating the effectiveness of our method in clarifying the model’s decision-making process.

## 1.3 Thesis Outline

The thesis is structured in seven chapters as follows:

**Chapter 1 – Introduction:** Presents the motivation behind the thesis, followed by an overview of the technical research, outlining its progression chapter by chapter. It concludes with a summary of the key contributions of the work.

**Chapter 2 – Background:** Explores the theoretical foundations and key literature on AI and audio processing. It then reviews relevant research in XAI that explains audio processing models, including evaluation methods used to assess explanation effectiveness.

**Chapter 3 – Interpretable Audio Classification with Data Decomposition:** Explores the explainability of audio classification models using a healthcare case study on COVID-19 prediction. It introduces CoughLIME, a novel method for generating listenable explanations of machine learning model predictions.

---

**Chapter 4 – Understanding Audio Transformers through Attention and Factorisation:** Introduces AttHear, a novel explanation method for audio transformers by leveraging the attention mechanism and integrating it with NMF.

**Chapter 5 – Audio Explanation Synthesis with Generative Foundation Models:** Presents an audio explanation approach using generative foundation models. By integrating feature attribution methods into the latent space, it avoids the challenges of interpreting feature relevance in raw audio and instead generates intuitive, listenable explanations.

**Chapter 6 – Improving Audio Explanations using Audio Language Models:** Combines ALMs with foundation model-based audio classifiers to enhance the understanding of important audio features. The proposed method refines and synthesises meaningful features, producing listenable and interpretable audio explanations by leveraging the latent space and generative capabilities of ALMs.

**Chapter 7 – Conclusions:** Highlights the key contributions of this thesis, discusses ethical considerations of the presented research in audio model interpretability, and explores future research directions to enhance audio explainability and broader applications.

**Chapter 8 – Appendix:** Provides detailed descriptions of the codebases corresponding to each component of the work, including implementation code and listenable sample audio explanations generated using the proposed methods. It also includes comprehensive information about the datasets utilised in this thesis.

# 2

## Background

### Contents

---

|                                                                 |           |
|-----------------------------------------------------------------|-----------|
| <b>2.1 Deep Learning</b>                                        | <b>8</b>  |
| 2.1.1 Recent Advancements                                       | 9         |
| 2.1.2 Applications                                              | 12        |
| 2.1.3 Deep Models for Audio Processing                          | 13        |
| <b>2.2 Explainable Artificial Intelligence</b>                  | <b>14</b> |
| 2.2.1 XAI Foundations                                           | 15        |
| 2.2.2 XAI Methods                                               | 16        |
| 2.2.3 XAI Evaluation                                            | 24        |
| <b>2.3 Audio Explainability for Deep Models</b>                 | <b>28</b> |
| 2.3.1 Explaining Audio Models using Domain-Agnostic XAI Methods | 29        |
| 2.3.2 Audio-specific XAI Methods                                | 34        |

---

## 2.1 Deep Learning

The era of deep learning was catalysed by significant advancements in algorithms and hardware technologies, which enabled the development of more powerful models and systems. Initially,

deep learning models focused on relatively simple tasks, such as predicting handwritten digits in the MNIST dataset [Deng, 2012]. However, over time, these models have evolved to tackle far more complex challenges, including generative models and sophisticated multimodal tasks. In this section, we will explore the recent breakthroughs in deep learning that have propelled artificial intelligence into real-world applications, highlighting its impact across various domains. We will focus on popular use cases in computer vision and natural language processing, where AI is increasingly used for detection and generation. Additionally, we will examine the unique characteristics and applications of deep learning in the audio domain, highlighting its domain-specific properties and innovations.

### 2.1.1 Recent Advancements

The early methods in deep learning revolved around fundamental end-to-end architectures such as MLPs [Rumelhart et al., 1986], CNNs [Krizhevsky et al., 2012], and RNNs [Sherstinsky, 2020]. MLPs, the simplest of these, provided the basis for many early deep learning models by stacking layers of neurons to learn complex patterns. CNNs, originally designed for image recognition tasks, became widely known for their ability to automatically learn spatial hierarchies in data, revolutionising fields like computer vision. RNNs, on the other hand, were introduced to process sequential data, excelling in tasks like language modeling and speech recognition. Although CNNs and RNNs are now considered relatively older models, they remain integral to modern architectures, often serving as foundational building blocks in more complex, hybrid models like attention-based transformers [Vaswani et al., 2017]. Their continued relevance underscores their robustness and adaptability in solving a variety of tasks across different domains [Brown et al., 2020b, Raffel et al., 2019, Grechishnikova, 2019, Baevski et al., 2019, Lin et al., 2023, Hayes et al., 2025].

As deep learning has evolved, more sophisticated architectures have emerged to enhance representation learning, allowing models to better capture and understand complex relationships within high-dimensional spaces. Autoencoders [Vincent et al., 2010], for instance, have been pivotal in learning efficient, compact representations of data by encoding inputs into a lower-dimensional latent space and then decoding them back to the original form. Variational Autoencoders (VAE) [Kingma and Welling, 2014] extend this concept by introducing probabilistic modeling in the latent space, enabling them to generate new, meaningful variations of the input data while ensuring smoother and more continuous latent representations. This ability to learn abstract features has been crucial in tasks like anomaly detection and data compression. Generative Adversarial Net-

works (GAN) [Goodfellow et al., 2014], another groundbreaking architecture, introduced a framework consisting of two neural network, a generator and a discriminator, that compete adversarially to generate and evaluate realistic data. This innovative approach has made GANs highly effective in image generation and various other creative applications. Figure 2.1 details a schematic of the VAE and GAN architectures. Similarly, Siamese networks [Koch et al., 2015], designed to learn similarity between pairs of inputs, have been widely applied in tasks like face verification and metric learning. These architectures contribute to more powerful models capable of representing complex, high-level abstractions, pushing the boundaries of what deep learning can achieve in fields ranging from computer vision to natural language processing.

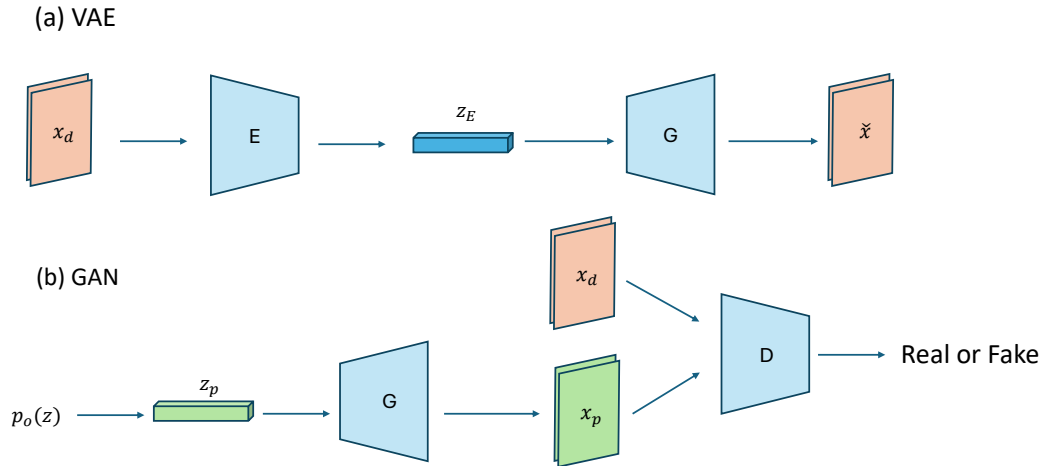


Figure 2.1: Illustration of the architectures of (a) VAE, (b) GAN where Enc, Gen, Dec, and Disc denote the encoder, generator, decoder, and discriminator, respectively.

More recent advancements in deep learning have introduced architectures that further refine model efficiency and performance, with a particular focus on eliminating the limitations of sequential processing. The transformer architecture [Vaswani et al., 2017] shown in Figure 2.2, introduced in 2017, revolutionised the field by replacing recurrent processing with self-attention mechanisms, enabling models to process entire sequences of data simultaneously rather than one step at a time. This innovation drastically improved training efficiency and model scalability, making it the backbone of state-of-the-art applications in Natural Language Processing (NLP) and beyond. Transformers have become the foundation of large language models like GPT [OpenAI et al., 2024], BERT [Devlin et al., 2018], and their variants, which have demonstrated remarkable performance in tasks such as machine translation, question answering, and text generation. Similarly, the transformer architecture has been extended to vision tasks, leading to the development of im-

age foundation models, such as Vision Transformers (ViT) [Dosovitskiy et al., 2020], that process images as sequences of patches rather than pixels. These foundation models, capable of learning generalisable features across vast datasets, have reshaped the landscape of AI, offering a unified approach to tackling a variety of tasks across different modalities, from text to images and even multimodal applications.

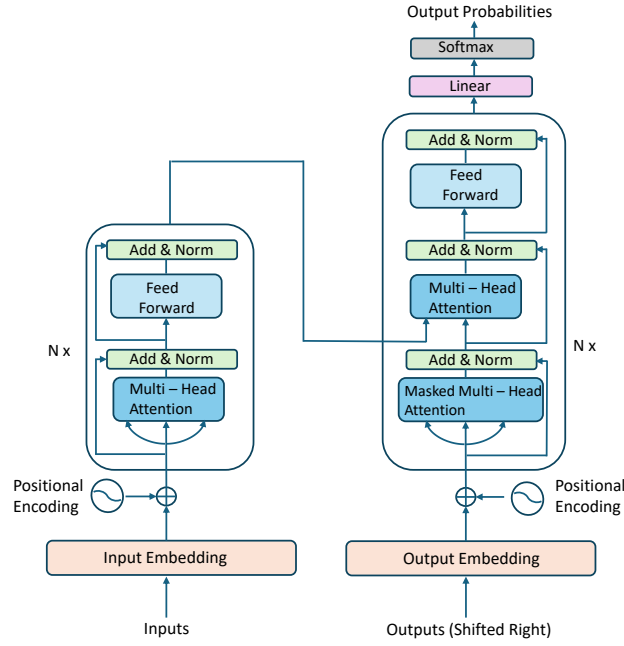


Figure 2.2: Illustration of the transformer architecture adapted from [Vaswani et al., 2017].

As deep learning models have grown larger and more complex, traditional supervised learning strategies, which rely heavily on labeled data, have faced significant challenges. Feeding vast amounts of labeled data into models is often not feasible due to the immense cost and time required for labeling. To address this, self-supervised learning has emerged as a powerful alternative. This approach allows models to learn from unlabeled data by creating pretext tasks, where the model generates its own supervision from the data itself. For example, in natural language processing, models can predict the next word in a sentence, fill in missing parts of text by masking [Devlin et al., 2018, Yang et al., 2019, Brown et al., 2020b], replace some tokens with plausible alternatives sampled from a small generator network [Clark et al., 2020] or exploit contrastive learning strategies [Gao et al., 2021]. This strategy not only reduces the dependency on labeled data but also enhances scalability, as vast amounts of unlabeled data, which are widely available across the internet, can be utilised. Self-supervised learning has become a key component in training large models, especially in the context of foundation models, where it allows them to learn rich representations without the need for expensive human labeling efforts. This shift has significantly accelerated the development

of AI systems, enabling models to generalise more effectively across a wide range of tasks and domains.

### 2.1.2 Applications

Deep learning excels in handling a wide variety of tasks, broadly categorised as predictive, generative, and transformative. Predictive tasks involve analysing input data to produce specific outputs, such as classifying images [Simonyan and Zisserman, 2015], detecting objects in scenes [Ren et al., 2015], or performing regression to predict numerical values [Goodfellow et al., 2016]. Generative tasks, on the other hand, focus on creating new content conditioned on input parameters or learned patterns, exemplified by applications like image generation [Goodfellow et al., 2014], text generation [Radford et al., 2019], and audio synthesis [van den Oord et al., 2016]. Transformative tasks involve altering the content or structure of input data to produce a modified output, such as applying style transfer to images [Gatys et al., 2016], translating text between languages [Bahdanau et al., 2015, Vaswani et al., 2017], or performing Image-to-Image Translation like converting photographs into sketches [Isola et al., 2017]. The versatility of deep learning across various tasks highlights its adaptability to a wide range of applications, establishing it as a fundamental tool in modern artificial intelligence.

Deep learning's applications span a variety of domains and even integrate multiple modalities, demonstrating its broad utility and adaptability. Domain-specific tasks include visual applications like medical image analysis in healthcare [Ronneberger et al., 2015, Litjens et al., 2017], object detection for autonomous driving [Redmon and Farhadi, 2018, He et al., 2017], and facial recognition in security [Taigman et al., 2014, Schroff et al., 2015]. In the textual domain, tasks such as legal document analysis [Chalkidis et al., 2020], sentiment analysis [Socher et al., 2013], and natural language processing facilitate efficient data handling in industries like law and customer service. For the audio domain, deep learning powers speech recognition [Hinton et al., 2012, Amodei et al., 2015] in virtual assistants and music generation tools [Dhariwal et al., 2020, Civit et al., 2022] in the entertainment industry. Beyond individual domains, multimodal tasks integrate multiple data modalities to address complex objectives. For instance, image captioning generates descriptive text from visual inputs [Vinyals et al., 2015], speech-to-text transcription enhances accessibility and facilitates automation [Prabhavalkar et al., 2023], and text-to-speech synthesis plays a crucial role in conversational AI [Shen et al., 2017]. These applications, spanning diverse domains and industries, underscore the pioneering potential of deep learning in addressing real-world challenges.

### 2.1.3 Deep Models for Audio Processing

Audio processing models are at the core of this thesis, representing a domain with unique characteristics that demand specialised approaches. Unlike visual or textual data, audio data has distinct dynamics, including its continuous nature, temporal dependencies, and diverse representations such as waveforms, spectrograms, and feature embeddings [Purwins et al., 2019]. These complexities require tailored models and architectures designed to effectively capture and interpret the underlying patterns in audio signals. Unlike tasks in natural language processing or computer vision, which often involve structured text or spatially arranged pixels, audio processing tasks must deal with the intricacies of time-frequency analysis and the variability of acoustic environments.

Audio processing encompasses a wide range of tasks with diverse applications across various fields. Generative tasks, such as music generation and audio synthesis [van den Oord et al., 2016, Yang et al., 2023], involve creating new auditory content, enabling innovations in creative industries. Recognition tasks, particularly Automatic Speech Recognition (ASR), transform spoken language into text and are foundational to voice-controlled systems and accessibility tools [Yu and Deng, 2016, Kheddar et al., 2024]. In healthcare, audio processing is applied to diagnostic tasks using auditory signals like heartbeats, coughs, or breathing sounds, helping identify conditions linked to specific body parts, such as cardiac anomalies [Zhang et al., 2020] or respiratory diseases [Akman et al., 2022, Coppock et al., 2021b]. Another emerging area is affective computing, where audio cues like tone, pitch, and rhythm are analysed to detect emotions or mental states, enhancing human-computer interaction [Schuller et al., 2018]. These tasks leverage the distinctive characteristics of audio data to address critical challenges and open new possibilities in both academic research and industrial applications.

Audio models utilise various input data types, each suited for specific tasks and capturing different aspects of the audio signal. Raw waveforms represent the sound in its most basic form, with models processing the time-domain signal directly to identify patterns, often using deep learning techniques that bypass traditional feature extraction [Zeghidour et al., 2021b, Mehri et al., 2016]. Spectrograms, on the other hand, transform the waveform into a time-frequency representation, making it easier to analyse frequency content over time and widely used in tasks like speech recognition and music classification [Hershey et al., 2016]. Mel-Frequency Cepstral Coefficients (MFCC) are another popular feature, especially in speech and speaker recognition, as they capture the essential characteristics of sound by modeling the human ear’s sensitivity to different frequencies [Davis and Mermelstein, 1980, Gourisaria et al., 2024]. Deep features, which are high-level rep-

representations learned by deep neural networks, can provide more abstract insights into the audio signal, making them useful for complex tasks like emotion recognition or sound event detection [Baeovski et al., 2020, Hsu et al., 2021, Baeovski et al., 2022]. Additionally, engineered features like OpenSMILE [Eyben et al., 2010] focus on extracting specific attributes such as pitch, rhythm, and energy, which are particularly effective for affective computing and audio analysis in contexts like healthcare or human-computer interaction. These diverse input data types empower audio models to tackle a broad spectrum of applications by representing sound in formats that are optimally aligned with the specific requirements of each task.

Model architectures specifically designed for audio processing have evolved significantly, beginning with sequential models such as Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] networks and other RNNs. These architectures effectively handle the temporal nature of audio signals, making them suitable for tasks such as speech recognition and audio classification. As the field progressed, transformers emerged as a game-changer due to their ability to model long-range dependencies and parallelise computation, leading to improved performance in various audio tasks. Noteworthy advancements like audio language models, exemplified by Wav2Vec [Schneider et al., 2019] and other audio self-supervised learning methods [Liu et al., 2022], highlight the effectiveness of self-supervised learning in extracting meaningful representations from raw audio data without extensive labeled datasets. This evolution signifies a transition towards more robust and adaptable models, capable of effectively addressing the inherent complexities of audio data at scale.

## 2.2 Explainable Artificial Intelligence

XAI aims to address the black-box nature of AI models by making their operations and decision-making processes transparent and interpretable. In this context, explainability encompasses terms like transparency, interpretability, and understandability, all of which strive to clarify how AI models arrive at their outputs. Just as humans rely on reasoning and can articulate the basis for their decisions or creative processes beyond just presenting the outcome, explainability seeks to enable AI to provide similar clarity. This involves not only showing the results but also making the logic, data patterns, or mechanisms behind those results comprehensible to users. By aligning AI models more closely with human-like thinking processes, often considered the ultimate goal, these efforts ensure the quality and authenticity of the “intelligence” underpinning the system. A more recent and widely accepted definition of explainability is proposed by [Barredo Arrieta et al., 2020],

stating that “Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand.” This definition explicitly highlights the dependence of a model’s explainability on its intended audience.

### 2.2.1 XAI Foundations

#### Need for XAI

The key motivations for explainability can be categorised into four distinct yet interrelated aspects [Adadi and Berrada, 2018]. First, explanations serve as justifications, particularly in cases where AI systems produce biased or unexpected results. Providing clear reasoning behind decisions not only enhances trust but also ensures compliance with regulations like the GDPR’s “right to explanation.” Second, explainability enhances control by improving system transparency, allowing for the identification and correction of errors, and preventing unintended consequences. Third, it facilitates continuous improvement as understanding the reasons behind a model’s outputs enables users to refine and optimise its performance. Finally, explainability facilitates discovery by enabling AI systems to share new insights and knowledge, as seen in breakthroughs like AlphaGo Zero’s strategic advancements. In the future, XAI could play a crucial role in uncovering hidden principles in various scientific fields, driving both innovation and understanding.

#### XAI Concepts

In XAI, several terms are commonly used, often with overlapping meanings [Barredo Arrieta et al., 2020, Lipton, 2016, Miller, 2019, Vilone and Longo, 2021b]. Understandability refers to a model’s ability to be grasped by humans without needing an explanation of its internal workings. Comprehensibility, on the other hand, is specific to machine learning models and relates to how well the learned knowledge can be represented in a human-understandable way, often tied to model complexity. Interpretability focuses on making a model’s decisions meaningful to humans, while explainability extends this by serving as an interface between the model and users, providing an understandable and accurate proxy of the decision-making process. Furthermore, transparency refers to the inherent understandability of a model, which can be classified into three categories: simulability (where humans can mentally simulate decisions), decomposability (where individual components are interpretable), and algorithmic transparency (where the underlying logic is clear).

The XAI pipeline includes several key components essential for generating meaningful explanations. The AI model to be explained ranges from inherently interpretable models, such as decision trees, to complex black-box models, including deep neural networks [Lipton, 2016, Guidotti et al., 2018b]. The choice of XAI method depends on the model and context, with intrinsic methods designed for interpretability and post-hoc methods providing explanations for pre-trained models [Barredo Arrieta et al., 2020]. The purpose of explanation varies across stakeholders, from model debugging and regulatory compliance to improving user trust [Miller, 2019]. To evaluate XAI methods, key metrics include fidelity, ensuring explanations accurately reflect model behaviour, completeness, capturing essential decision factors, and human interpretability, determining usability for end-users [Vilone and Longo, 2021b]. Balancing these properties is crucial for developing reliable and effective explainability frameworks.

### 2.2.2 XAI Methods

Existing XAI methods cover different types of techniques, including rule-based explanations, feature importance analysis, local surrogate models, prototype/example-based explanations, counterfactuals, model distillation, and attention mechanisms. XAI methods are typically classified according to their stage: ante hoc vs. post hoc; scope: local vs. global; input data: numerical/categorical, pictorial, textual, and time series; and output format: numerical, rules, textual, visual, audio, and mixed [Vilone and Longo, 2020]. The stage of providing explanations refers to the stage at which explanations are generated, where ante-hoc methods make models naturally explainable by considering interpretability before and during the training process, and post-hoc methods exploit external explainer models to ensure explainability without changing the original model after training. While global explanations target the entire model, local explanations explicitly explain the inference of a specific input. In addition to methodological differences, the input data representation and output format used by XAI models differ with respect to the domain of the AI task and the demographics and expertise level of human users.

A number of comprehensive survey papers review XAI methods from diverse perspectives, contributing significantly to the field. Some of these surveys focus on specific modalities, such as explainability techniques tailored for deep neural networks in audio processing [Akman and Schuller, 2024b], whereas the majority of XAI research has traditionally centered on image and text-based models. Others examine XAI methods from a domain-specific standpoint, including applications in medical AI [Sun et al., 2025]. A broader categorisation of XAI methods by application

domains, such as forecasting, healthcare, transportation, finance, and social media, is provided in [A. and R., 2023].

Additionally, several surveys aim to systematically classify existing XAI approaches, offering structured taxonomies. Some categorise methods based on their output type, conceptual foundation, or functionality [Speith, 2022, Vilone and Longo, 2021a, Barredo Arrieta et al., 2020, Adadi and Berrada, 2018, Ali et al., 2023], providing a well-organised overview of the field. A conceptual and theoretical perspective is taken in [Vilone and Longo, 2021b], which not only reviews XAI research but also discusses evaluation strategies using formal notions and metrics. Meanwhile, [Linardatos et al., 2021] contributes a taxonomy of XAI techniques alongside links to their programming implementations, serving as a valuable reference for both researchers and practitioners.

Beyond their technical focus, many of these survey papers also offer insights into the broader role of XAI in AI research, its significance, associated challenges, evaluation methodologies, and future research directions. Notably, [Ali et al., 2023] presents an in-depth analysis of existing survey papers, detailing their scope, including the number of studies reviewed, the types of XAI methods covered, the availability of evaluation frameworks, and the core concepts addressed. Together, these surveys provide a well-rounded understanding of the current landscape of XAI, guiding further advancements in the field. Following the categorisation in [Barredo Arrieta et al., 2020, Ali et al., 2023], although there are XAI methods specifically targeting data explainability, we briefly review inherently interpretable models and post-hoc XAI methods due to their relevance to our research in this thesis. An overview of the taxonomy of the XAI methods is shared in Figure 2.3.

### Self-interpretable Machine Learning Models

**Decision trees** are hierarchical models [Quinlan, 1986, Blockeel et al., 2023] used for both regression and classification that inherently satisfy transparency constraints. In their simplest forms, they are small enough for a human to simulate and understand easily; however, as their size and complexity grow, they shift from being easily interpretable to only partially decomposable or fully algorithmically transparent. Their long-standing use in decision-making contexts highlights the importance of balancing model complexity with interpretability, a focus that is evident in the increasing number of studies on decision tree simplification and generation. Although decision trees can meet all transparency requirements, expanding their size or feature interactions can make them less accessible to non-experts. Moreover, despite their immediate clarity, decision trees tend to have lower generalisation performance compared to other models, which has led to the adoption of tree

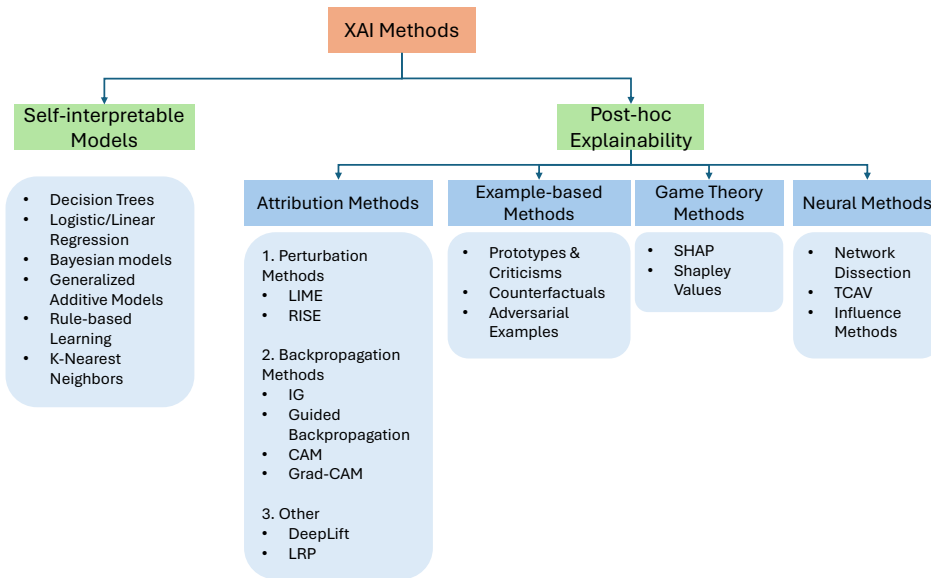


Figure 2.3: An overview of the taxonomy of the XAI methods.

ensembles to boost predictive accuracy. Unfortunately, combining multiple trees in an ensemble sacrifices the original transparency of individual trees, thereby necessitating the use of post-hoc explainability techniques.

**Logistic regression** is typically used for binary classification, while **linear regression** handles continuous outcomes; both models assume a linear relationship between predictors and the response variable [Tripepi et al., 2008, Chao-Ying Joanne Peng and Ingersoll, 2002]. This assumption, which restricts their flexibility, actually contributes to their transparency by making them easier to understand and simulate when kept simple. These models inherently offer algorithmic transparency, decomposability, and simulatability, but when explaining them to non-experts, post-hoc visualisation techniques are often needed. Widely used in social sciences, the regression models are evaluated through overall model performance, statistical tests like the Wald chi-square, goodness-of-fit measures such as the Hosmer–Lemeshow statistic, and the validation of predicted probabilities. However, caution is advised when interpreting log odds ratios, as they can be misleading due to unobserved heterogeneity, and comparisons across different models or groups may be problematic. Ultimately, maintaining transparency in regression models depends on keeping the model size manageable and ensuring that the variables used are easily understandable.

Building on the theme of transparent machine learning models, several additional approaches such as **k-nearest neighbors** [Taunk et al., 2019, Halder et al., 2024], **rule-based learning** [Urbanowicz and Moore, 2009, Weiss and Indurkha, 1995], **generalised additive models** [Hastie

and Tibshirani, 1986, Wood, 2017], and **bayesian models** [Derks and de Waal, 2021] stand out for their interpretability. K-Nearest Neighbors is a transparent method that classifies or predicts outcomes based on the similarity between samples, much like human decision-making using past experiences. Its clarity depends on the chosen number of neighbors and the distance function; too many neighbors or overly complex features can reduce its interpretability. Similarly, rule-based learning produces interpretable if-then or fuzzy rules that express data relationships in natural language, although an excessive number of rules or overly detailed conditions can compromise understandability. Generalised Additive Models enhance transparency by modeling the output as a sum of smooth functions of predictors, allowing users to easily visualise and assess each variable’s impact. Likewise, Bayesian models, typically structured as probabilistic networks that map conditional dependencies among variables, offer clear insights into feature interactions, but their interpretability may diminish if the model becomes overly complex.

### Post-hoc Explainability Methods

While intrinsic methods rely on building inherently transparent models, they often compromise on accuracy due to their simplicity. For situations that demand both high performance and interpretability using more complex models, post-hoc methods offer a compelling alternative by providing explanations after the model is built. These techniques work by constructing a secondary, often surrogate model that replicates the behaviour of the complex original, thereby illuminating its decision-making process without needing to fully understand its inner workings. Post-hoc explainability methods can be categorised into six key types: attribution methods, visualisation methods, example-based explanation methods, game theory methods, knowledge extraction methods, and neural methods [Ali et al., 2023].

#### A. Attribution Methods

Attribution methods in explainable AI quantify how each input feature contributes to a model’s final decision. Often described in terms of relevance, contribution, or importance, these methods help to elucidate which aspects of the input, such as individual pixels or image patches in computer vision, or tokens in natural language processing, are most influential in driving the output. By assigning a numerical value to each feature, attribution techniques provide a granular view of model behaviour, offering insights into how decisions are made. Attribution methods can be broadly categorised into several groups, including Perturbation Methods, which evaluate changes in output when inputs are modified; Backpropagation Methods, which trace gradients back through

the network to assess feature impact; Deep Taylor Decomposition ([DTD](#)), which decomposes the decision into feature contributions; and Deep Learning Important FeaTures ([DeepLIFT](#)), which compares activations to a reference baseline.

**A1. Perturbation Methods** analyse how changes to input features affect a model’s output. By systematically modifying or occluding parts of the input, these methods reveal which elements are most critical to the model’s decision-making process.

**LIME** is an effective technique in explainable AI that creates local surrogate models to elucidate individual predictions of complex black-box systems. By perturbing the input, such as toggling super-pixels in images, modifying words in text, or independently altering features in tabular data, LIME generates a new dataset paired with the corresponding outputs from the original model. It then fits an interpretable model (e.g., LASSO, logistic regression, or decision trees) that is weighted by the proximity of the perturbed samples to the instance of interest, ensuring that the explanation faithfully reflects the model’s local behaviour. This approach, known as local fidelity, focuses on accurately approximating the decision boundary around a specific prediction rather than capturing the global dynamics of the model. The general formulation of LIME is presented in Equation 2.1.

$$\xi(x) = \arg \min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_\phi) + \Omega(g) \quad (2.1)$$

Here,  $f$  denotes the black-box model,  $g$  represents the interpretable surrogate model chosen from the class  $\mathcal{G}$ ,  $\mathcal{L}(f, g, \pi_\phi)$  is a loss function that measures the fidelity of  $g$  in approximating  $f$  within the locality defined by the proximity measure  $\pi_\phi$ , and  $\Omega(g)$  is a complexity penalty to ensure interpretability.

**Randomised Input Sampling to Provide Explanations ([RISE](#))** is an innovative method in explainable AI that quantifies the significance of image regions in deep neural network predictions. In this approach, an input image is multiplied elementwise with a randomly generated binary mask that sets the brightness of selected pixels to zero, thereby simulating their removal. The resulting masked images are then fed into the deep neural network to compute confidence scores for the target class. RISE produces a heatmap by linearly combining these binary masks, with each mask weighted by its corresponding confidence score, which reveals the influence of individual pixels or regions on the model’s decision. This technique provides clear visual explanations for predictions made by object detectors and other vision models while offering an intuitive, model-agnostic way to understand the inner workings of complex AI systems.

Extensions of LIME have been proposed to address its limitations and enhance the interpretability of complex models by offering more detailed and diverse explanations. For example, Local Rule-based Explanation (LORE) [Guidotti et al., 2018a] employs an ad-hoc genetic algorithm to generate a balanced set of neighbor instances for a given input and extracts a decision tree classifier, yielding clear, rule-based insights. Similarly, Cluster Representatives with LIME [El Bekri et al., 2020] extends the LIME framework by clustering the data first and then assigning an unknown instance to the closest cluster, subsequently visualising each cluster’s feature validity ranges to highlight the most significant aspects influencing the target class. Submodular Pick LIME [Ribeiro et al., 2016] further improves model understanding by selecting a diverse subset of key predictions, ensuring that users examine representative events that capture the model’s overall behaviour. Additionally, the Anchors method [Ribeiro et al., 2018] departs from LIME’s linear surrogate models by employing perturbations to generate intuitive IF-THEN rules, known as anchors, that clearly delineate decision boundaries with high local fidelity. Together, these extension methods provide a richer and more nuanced toolkit for explaining black-box predictions, thereby improving both local and global trust in AI systems.

## A2. Backpropagation Methods

Backpropagation methods in explainable AI efficiently compute feature attributions by propagating gradients through a deep neural network. Unlike perturbation-based approaches, they require only a single forward and backward pass, making them computationally efficient. However, their attributions may not always directly correspond to output variations, which can limit interpretability.

**Integrated Gradients (IG)** [Sundararajan et al., 2017] quantifies the contribution of each input feature to a model’s prediction by computing the integral of the gradients along a straight-line path from a baseline to the actual input. The baseline is typically a zero vector or a neutral reference. This approach ensures that attributions satisfy key properties like sensitivity, which ensures important features receive non-zero attributions, and implementation invariance, which guarantees consistency across functionally equivalent models. Unlike simple gradient-based methods, IG addresses issues like gradient saturation by accumulating gradients over multiple scaled versions of the input. The Integrated Gradients attribution for an input feature  $x_i$  is defined in Equation 2.2.

$$IG_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (2.2)$$

where,  $x$  is the input instance,  $x'$  is the baseline input,  $F(x)$  is the model function mapping inputs to outputs,  $\frac{\partial F}{\partial x_i}$  represents the gradient of  $F$  with respect to  $x_i$ , and  $\alpha$  scales the input from the baseline to the actual input.

**Guided Backpropagation** computes the gradient of the target output with respect to the input but modifies the standard backpropagation algorithm by selectively allowing only positive gradients to flow backward through the network. This is achieved by overriding the ReLU function's behaviour so that negative gradients are suppressed during the backward pass, ensuring that only features that positively influence the activation of higher layers are considered. Specifically, the ReLU function is applied to the input gradients in Guided Backpropagation, whereas in deconvolution, it is applied to the output gradients before direct backpropagation. As a result, Guided Backpropagation produces sharp and detailed saliency maps that emphasise relevant structures in the input data while reducing noise from less important features.

**Class Activation Maps (CAM), Gradient-weighted CAM (Grad-CAM), and Guided Grad-CAM** are methods in XAI that help interpret CNNs. CAM utilises global average pooling to reduce model complexity while retaining performance by computing spatial averages of feature maps and using them to highlight discriminative image areas. However, CAM requires architectural modifications. Grad-CAM improves on this by being model-agnostic and does not require retraining. It uses gradients from the last convolutional layer to assign importance scores to neurons for a given target class, producing a heatmap that highlights relevant regions in the input image. Guided Grad-CAM combines Grad-CAM with Guided Backpropagation to enhance pixel-level detail. By backpropagating only positive gradients and suppressing negative ones, Guided Backpropagation emphasises input features that activate neurons. By combining this with Grad-CAM through element-wise multiplication, Guided Grad-CAM produces high-resolution, class-discriminative visualisations that offer both global interpretability and fine-grained insights into model behaviour.

**Layer-wise Relevance Propagation (LRP)** [Bach et al., 2015] and **DeepLIFT** [Shrikumar et al., 2017] are two prominent attribution methods in XAI that help interpret deep neural networks by tracing model predictions back to input features. LRP is based on Taylor Decomposition, which decomposes the model's output by summing relevance scores to explain its behaviour. Inspired by sensitivity analysis, LRP redistributes the function output backward through the network while preserving the total relevance at each layer. During classification, the input is processed as a feature vector, and the network assigns it to a specific class. LRP then decomposes the classification output into feature-wise relevance scores, ensuring that important features receive higher attributions. DeepLIFT, on the other hand, follows a similar decomposition approach but frames significance in

terms of deviations from a user-defined reference input. Unlike traditional gradient-based methods, DeepLIFT can propagate importance signals even when the gradient is zero, avoiding issues such as saturation and discontinuities. Additionally, by separately considering the effects of positive and negative contributions at nonlinearities, DeepLIFT can capture dependencies missed by other approaches. Since DeepLIFT scores are computed using a backpropagation-like algorithm, they can be obtained efficiently in a single backward pass after a prediction is made.

## B. Example-based Explanation Methods

Example-based explanation methods interpret model decisions by identifying representative instances, offering intuitive and interpretable insights. Prototypes are instances that summarise the dataset by capturing central patterns, while criticisms highlight deviations, providing complementary perspectives. Traditional prototype selection methods like K-medoids [Park and Jun, 2009] focus solely on representative instances, but the Maximum Mean Discrepancy method [Kim et al., 2016] improves on this by integrating both prototypes and criticisms. It identifies prototypes by analysing data densities and criticisms by measuring distribution differences, ensuring a more comprehensive understanding of model behaviour. In contrast to prototypes, counterfactual explanations focus on “what-if” scenarios by generating modified instances that demonstrate how a slight change in input could alter the model’s decision. Unlike prototypes, counterfactuals do not need to be actual training instances; they can be synthetically generated to explore decision boundaries. Counterfactual explanations are formulated in [Wachter et al., 2017] by defining a loss function that optimises the difference between an instance of interest and its modified counterpart while achieving a desired outcome. Additionally, adversarial examples, while primarily known for exposing vulnerabilities in deep neural networks, can also serve as analogical and contrastive explanations. Analogical explanations use reasoning by similarity [Gentner, 1983], comparing familiar and unfamiliar examples, while contrastive explanations highlight differences by presenting competing or opposing evidence [Stepin et al., 2021].

## C. Game Theory Methods

**Shapley Values** and **Shapley Additive Explanation (SHAP)** [Lundberg and Lee, 2017] are powerful methods in XAI that quantify the contribution of each feature to a model’s prediction using concepts from cooperative game theory. Shapley Values determine how much each feature influences a particular prediction by distributing the model’s output among input features in a fair and mathematically rigorous way. SHAP builds on this by providing a unified framework for explaining any machine learning model using the optimal Shapley values. In this context, a

player represents a feature or a group of features, such as superpixels in an image, that collectively contribute to the model’s decision. SHAP offers both local interpretability, where values are assigned to each instance to explain why a specific prediction was made, and global interpretability, where aggregated SHAP values reveal the overall importance and directionality of features in influencing the target variable. This makes SHAP particularly useful for interpreting complex models, as it provides insights similar to variable importance plots while also showing whether each feature’s contribution is positive or negative.

#### D. Methods for Neural Networks

XAI methods for neural networks include example-based and representation-based approaches. Example-based methods identify influential training samples or inputs that strongly activate neurons, using influence functions or optimisation techniques to highlight key model dependencies [Koh and Liang, 2017, Pruthi et al., 2020, Olah et al., 2017]. Representation-based methods analyse internal feature representations, with Network Dissection [Bau et al., 2017] mapping neurons to human-labeled concepts like objects or phonemes, while Quantitative Testing with Concept Activation Vectors (TCAV) [Kim et al., 2018] measures a model’s sensitivity to user-defined concepts without requiring detailed annotations.

### 2.2.3 XAI Evaluation

Evaluating XAI methods is essential to assess their usefulness, reliability, and overall quality in different contexts. High-quality explanations should satisfy key properties such as generalisability, ensuring applicability across different models and datasets; consistency, meaning similar inputs yield similar explanations; fidelity, where explanations accurately reflect the true model behaviour; and privacy and fairness, which ensure explanations do not expose sensitive information or introduce bias. However, these are only a few of the many important properties that define a good explanation. Various studies classify XAI evaluation methods into application-grounded, human-grounded, and functionally-grounded evaluations, each focusing on different aspects of explainability. In [Vilone and Longo, 2021b], the authors identify two primary evaluation approaches: objective evaluations, which use quantitative metrics to assess explanation accuracy and robustness, and human-centered evaluations, which measure interpretability, trust, and usability from a user perspective. In this thesis, we adopt this categorisation, evaluating XAI methods through both objective and human-centered approaches to ensure a comprehensive assessment of their effectiveness.

### Quantitative Metrics

Fidelity, also referred to as faithfulness, is a widely used metric for quantitatively evaluating the performance of explanation methods in XAI. It primarily measures how well the explanation aligns with the model’s true decision-making process by assessing the prediction agreement between the identified important features and the original input. One common approach to measuring fidelity is the removal-based strategy, where the most important features identified by an explanation method are removed, and the model’s performance drop is analysed. A higher accuracy drop indicates higher fidelity, as it suggests that the removed features were indeed crucial for the prediction. In computer vision, for instance, an explanation method may highlight key image regions, which are then extracted while the remaining pixels are set to zero, followed by re-evaluating the model’s classification performance. However, applying fidelity evaluation in audio is more challenging, especially when raw waveforms are transformed into the frequency domain. Feature removal in this domain and reconstructing back to the original waveform introduce complexities, making fidelity assessment in audio less straightforward than in images.

Evaluating XAI methods by comparing their explanations against other methods and human-annotated references is a widely used approach to assess their reliability and effectiveness. This evaluation can be conducted by analysing how well explainability techniques align with human-labeled ground truth or by using perturbation-based testing to measure the impact of altering key features. For instance, in image classification, filter interpretability metrics assess the overlap between CNN activation maps and semantic categories such as objects, textures, or colours [Zhang and Zhu, 2018], while location instability measures whether relevant object parts are consistently identified across different images. Additionally, saliency maps generated by various explainability methods have been compared against those from randomly initialised untrained networks or datasets with randomly permuted labels [Adebayo et al., 2018a, Adebayo et al., 2018b]. The degree of correlation between these saliency maps was quantified using Spearman Rank Correlation coefficients, providing insights into the consistency and reliability of different XAI approaches.

Quantitative evaluation of trustworthiness in XAI methods involves assessing key properties such as fairness, robustness, and reliability in a measurable way [Deck et al., 2024, Stevens et al., 2020]. One method for evaluating fairness is to analyse whether explanations reveal biased features that influence predictions, thereby enabling bias detection and mitigation through well-defined fairness metrics. Fairness assessment is particularly critical in high-stakes applications such as mortgage lending, hiring, and prison sentencing, where biased AI-driven decisions can have serious

societal consequences. Tools like AI Fairness 360 [Bellamy et al., 2019] provide a systematic framework for fairness evaluation by offering diverse fairness metrics, explanation techniques, and bias mitigation strategies. This open-source toolkit supports the transition of fairness research into real-world applications and enables benchmarking to ensure models adhere to fairness standards, ultimately contributing to the trustworthiness and accountability of AI systems.

Robustness in XAI methods refers to the stability of explanations when small, non-impactful changes are made to the input. Ideally, if a minor input modification does not alter the model’s prediction, the explanation should remain consistent [Alvarez-Melis and Jaakkola, 2018]. Studies have tested this by applying perturbations such as Gaussian noise [Alvarez-Melis and Jaakkola, 2018] or constant shifts [Kindermans et al., 2019, Ghorbani et al., 2019] and then measuring the changes in saliency maps using metrics like Local Lipschitz Continuity, Spearman Rank Correlation, and Top-k Intersection. However, findings indicate that many explainability methods are sensitive to even minor perturbations, violating the input invariance requirement [Kindermans et al., 2019]. In time-series classification, perturbing relevant time segments has been used to evaluate whether explanations truly highlight discriminative features [Nguyen et al., 2020]. Additionally, the completeness of explanations has been analysed by comparing various gradient-based, weight-based, and perturbation-based methods [Gevrey et al., 2003]. These evaluations highlight that many XAI techniques struggle with robustness, affecting their reliability in practical applications.

Computational complexity is a critical consideration in AI, particularly given the resource-intensive nature of many applications, and this concern extends to XAI methods as well. While the primary goal of XAI is to provide insightful explanations, efficiency is equally important, especially for real-time or large-scale applications. Therefore, it is essential to report the computational cost of explanation methods to assess their practicality. For instance, backpropagation-based approaches such as Grad-CAM and LRP require an additional backward pass through the network, increasing their runtime compared to simple forward evaluations. In contrast, game theory based SHAP method computes attributions by considering all possible feature subsets, making them considerably more computationally demanding. Additionally, the computational burden of XAI techniques often depends on hyperparameters, such as the number of iterations in perturbation-based approaches, which can dramatically affect both runtime and explanation quality. Balancing explanation accuracy and computational efficiency is crucial for deploying XAI methods in real-world scenarios.

### Human-centered Evaluation

As the primary goal of XAI is to enhance human interpretability, human-centered evaluation of these methods is crucial. Such evaluations primarily focus on assessing the understandability of the explanations provided, ensuring they effectively clarify the model's decision-making process. This is particularly important in scenarios where auxiliary machine learning models assist human decision-makers, such as medical professionals diagnosing patients. Evaluating whether these explanations genuinely improve decision-making is essential for validating their usefulness. To achieve this, experiments involving human participants were conducted in two ways [Vilone and Longo, 2021b]. First, individuals with no prior technical or domain knowledge were randomly selected to interact with various explanatory tools and provide feedback through questionnaires. Second, domain experts were engaged to critically assess the explanations, ensuring their consistency with established domain knowledge and verifying their reliability. This dual approach provides both general and expert perspectives, strengthening the overall evaluation of explainability in AI models.

Several review papers have examined the role of human-centered evaluation in XAI, emphasising the need to assess AI explanations from a user perspective. A systematic review by [Kim et al., 2024] analysed 73 studies, highlighting the necessity of user evaluations to align AI explanations with human cognitive processes. Similarly, [Rong et al., 2024] reviewed 97 core papers, categorising them based on measured characteristics such as trust, understanding, usability, and human-AI collaboration performance. Their study also proposed practical guidelines for designing user studies in XAI research. To standardise evaluation procedures, [Donoso-Guzmán et al., 2023] introduced a comprehensive human-centered evaluation framework that integrates explanation aspects and summarises properties to assess the effectiveness of XAI methods. Additionally, [Ma et al., 2024a] developed OpenHEXAI, an open-source framework facilitating human-centered evaluation of XAI methods. OpenHEXAI provides tools for designing user studies and enhancing reproducibility in XAI research. Complementing these efforts, [Dawoud et al., 2023] conducted a user-centric study comparing three leading explanation methods, Prototypical Part Network, Occlusion, and Layer-wise Relevance Propagation, finding that despite varying highlighted regions, all methods offered users a comparable depth of understanding, thereby reinforcing their value in enhancing AI transparency.

Several studies have explored different approaches to evaluating explanation quality and its impact on human-AI interaction. One study investigates the role of virtual agents in XAI, analysing

how different levels of human-likeness (text, voice, and visual presence) influence users' trust in AI systems [Weitz et al., 2020]. The findings suggest that more human-like virtual agents can improve the perceived trustworthiness and helpfulness of AI explanations, particularly in complex tasks like speech recognition. Another study examines the impact of explanations on reducing bias in human-agent decision-making [Malhi et al., 2020]. By comparing AI recommendations with and without explanations, researchers assessed how two post-hoc explanation algorithms, SHAP and LIME, influenced user awareness of bias in a loan application dataset. Although the study did not yield significant findings, it highlighted differences in how explanations affect human reliance on AI-driven decisions. A third study introduces the System Causability Scale, a framework for measuring the quality of AI explanations [Holzinger et al., 2019]. This scale evaluates user perception of an explanation's utility by incorporating concepts from usability research, emphasising transparency, traceability, and interactive learning.

[Kim et al., 2024] conducted a systematic review analysing 73 studies to assess XAI methods from a user-centered perspective, addressing the lack of consensus on what makes AI-generated explanations meaningful and how they should be evaluated. The authors identify 30 key components of meaningful explanations and categorise them into three dimensions: in-context quality, which includes attributes such as understandability, sufficiency, trustworthiness, and usefulness; contribution to human-AI interaction, which examines the impact on trust, transparency, cognitive load, and user engagement; and contribution to human-AI performance, which assesses whether explanations improve user decision-making and task performance. The review highlights inconsistencies in evaluation methodologies, noting that only 19 of the 73 papers applied a framework used in at least one other study, making cross-study comparisons difficult. Building on taxonomies by [Doshi-Velez and Kim, 2018], [Mohseni et al., 2021], and [Lopes et al., 2022], the study extends previous work by offering a more detailed categorisation of evaluation measures. The authors advocate for a more unified approach to human-centered XAI evaluation, emphasising the need for standardised assessment frameworks to improve the reliability and applicability of user studies in XAI research.

## 2.3 Audio Explainability for Deep Models

In this thesis, we explore XAI from the perspective of the audio domain, rather than the more commonly studied field of image processing. This section specifically examines the potential of XAI methods to enhance the interpretability of audio-processing models. The significance of XAI in the audio domain lies not only in its ability to highlight human-understandable audio features

but also in its unique capacity to generate explanations in the form of directly listenable audio. By doing so, these methods offer more intuitive and accessible insights into model behaviour. This section incorporates content from our published review paper on audio XAI [Akman and Schuller, 2024b].

Audio models can be made interpretable and transparent either by applying XAI methods from other domains to these models or using audio-specific XAI methods. The main idea behind developing XAI audio methods is to provide more meaningful and comprehensible explanations by leveraging the expressive power of audio, which serves as an intuitive means of explaining audio tasks [Schuller et al., 2021b]. Although most XAI methods have been proposed to provide visual explanations or tested only on visual tasks and datasets, certain methods have the potential to explain audio models with simple modifications. When using these methods in audio tasks, similar properties between the data representations used in audio processing tasks and the original data representation of the XAI method (e.g., audio spectrograms and standard images) are considered to provide meaningful explanations. Conversely, there are also a substantial number of works that propose XAI methods tailored towards explaining audio models. Whereas some methods are built on existing XAI methods with substantial improvements, others propose completely new methods built from scratch to make audio models interpretable. In this study, we provide an overview of the existing XAI techniques, either generic or audio-specific, that are used to explain audio models. We also aim to highlight the potential of developing XAI methods that generate audio explanations, leading to promising research in this field.

The remainder of this section is organised as follows. First, we provide an overview of existing XAI methods applicable to audio models, categorising them according to the model’s input data representation. Next, we focus on current audio-specific XAI methods, highlighting their potential for specialised audio explanations. Finally, we discuss several open problems and outline potential research directions, including testing XAI methods on audio data, generating listenable explanations, and providing sonified explanations from non-audio modalities, before drawing a conclusion.

### 2.3.1 Explaining Audio Models using Domain-Agnostic XAI Methods

Audio representations, such as waveforms, spectrograms, and MFCCs, provide different ways of transforming and analysing audio data, enabling tasks such as automatic speech recognition, audio classification, speaker recognition, and signal processing. Depending on the selection of audio

representations and the characteristics of the task, various deep model architectures are utilised, including MLPs, CNNs, RNNs, and transformers. In addition to these design choices for AI systems, the selection of an appropriate XAI method is vital to enabling users to understand and trust the decisions made by these systems. In this section, we provide an overview of the generic XAI methods applicable to audio processing models, considering the input data representation of these models. We also present a summary of the generic XAI methods applicable to audio processing models in Table 2.1.

Table 2.1: Overview of generic XAI methods applicable to audio processing models.

| XAI Method                                         | Category             | Input Representation  | Output Format                   | Scope  | Model Agnostic |
|----------------------------------------------------|----------------------|-----------------------|---------------------------------|--------|----------------|
| Guided Backpropagation [Springenberg et al., 2015] | Feature attributions | Spectrogram, waveform | Feature attributions            | Local  | No             |
| LRP [Bach et al., 2015]                            | Feature attributions | Spectrogram, waveform | Feature attributions            | Local  | No             |
| Integrated Gradients [Sundararajan et al., 2017]   | Feature attributions | Spectrogram, waveform | Feature attributions            | Local  | No             |
| Grad-CAM [Selvaraju et al., 2017]                  | Feature attributions | Spectrogram           | Feature attributions            | Local  | No             |
| LIME [Ribeiro et al., 2016]                        | Feature attributions | Spectrogram, waveform | Feature attributions            | Local  | Yes            |
| SHAP [Lundberg and Lee, 2017]                      | Feature attributions | Spectrogram, waveform | Feature attributions            | Local  | Yes            |
| [Koh and Liang, 2017]                              | Prototype-based      | Spectrogram, waveform | Syntetic/natural input examples | Local  | No             |
| [Pruthi et al., 2020]                              | Prototype-based      | Spectrogram, waveform | Natural input examples          | Local  | No             |
| [Olah et al., 2017]                                | Prototype-based      | Spectrogram, waveform | Syntetic/natural input examples | Local  | No             |
| Network Dissection [Bau et al., 2017]              | Concept-based        | Spectrogram, waveform | Concepts                        | Global | No             |
| TCAV [Kim et al., 2018]                            | Concept-based        | Spectrogram, waveform | Concepts                        | Global | No             |
| Attention mapping [Caron et al., 2021]             | Attention-based      | Spectrogram           | Feature attributions            | Local  | No             |

### Traditional Feature Extractors

Common feature extraction toolkits, such as openSMILE [Eyben et al., 2013], openXBOW [Schmitt and Schuller, 2017], DeepSpectrum [Amiriparian et al., 2018], and auDeep [Freitag et al., 2018] exploit various audio processing methods to extract diverse acoustic features from audio data. The numerical tabular structure of these extracted features enables users to train self-explainable AI models, such as linear regression or decision trees, directly. Additionally, meaningful descriptors

such as pitch, loudness, and the number of segments produced in openSMILE provide textual explanations. For example, in an emotion recognition task, an angry audio can be explained as “the emotion in the audio is angry because the loudness is high.”

### Spectrogram Representation

Spectrograms and their derivatives, such as MFCCs, are commonly used to represent audio. They provide a visual and quantitative depiction of the frequency content and temporal variations of an audio signal, enabling a detailed analysis of audio signal characteristics and identifying patterns.

Given their image-like nature, CNNs are primarily used to process spectrograms for a wide range of audio tasks. A popular XAI method applicable to these models uses guided backpropagation [Springenberg et al., 2015] which creates feature relevance by modifying the backpropagation by computing the imputed version of a gradient. It augments standard backpropagation with an additional guidance signal from the higher layers; that is, it backpropagates only positive gradients during gradient computation, which increases the activation of the higher-layer unit to be observed. By backpropagating the model prediction in a neural network, LRP [Bach et al., 2015], an alternative technique, awards relevance scores to specific pixels or neurons. It utilises specially crafted local backpropagation rules subject to the conservation property, which requires that an identical quantity of what a neuron receives be disseminated to the bottom layer again. Integrated Gradients [Sundararajan et al., 2017] are a potent axiomatic attribution technique that can be easily implemented with only a few calls to the regular gradient operator and require no changes to the original network. It calculates the integral of the gradients along the direct line connecting the input and baseline images, offering a thorough assessment of feature attribution. The authors of [Selvaraju et al., 2017] proposed Grad-CAM, which is a feature attribution method designed for CNNs and is typically applied to the last convolutional layer. It creates a coarse localisation map that highlights key areas in the image for any predicted class using the gradients of the relevant classification score with respect to the features in the final convolutional layer. Examples of these methods are presented in Fig. 2.4.

Model-agnostic XAI methods such as LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017] offer a unique method for explaining audio models, regardless of the model architecture. LIME approximates the original complex model using an interpretable surrogate model around the local region of a given prediction to provide faithful explanations for each prediction. It locally perturbs sample points around a particular input example to identify feature importance

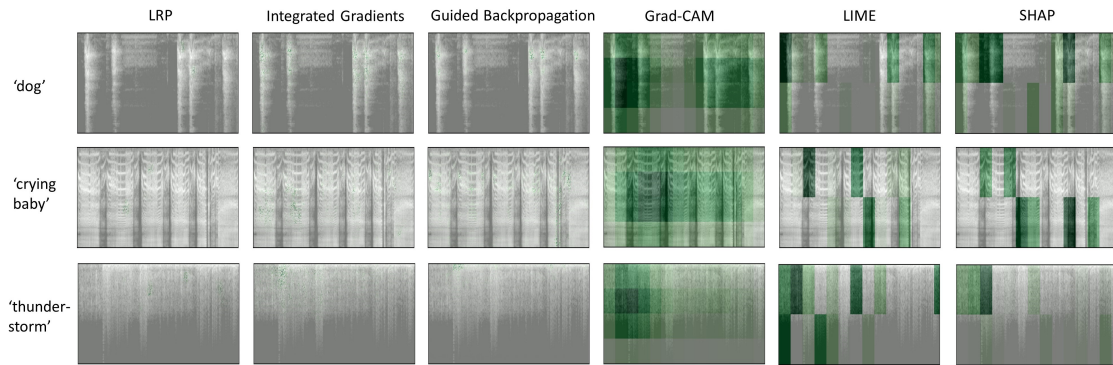


Figure 2.4: Examples of feature-attribution-based XAI methods on audio models. The positively relevant features towards the predicted class are marked in green.

based on changes in predictions. SHAP, which incorporates cooperative game theory, presents a unified framework for computing the additive feature importance for each specific prediction. This framework has several desirable properties (local accuracy, missingness, and consistency) that are novel to other additive feature-attribution approaches. Both methods also offer a method to directly assign feature importance to predefined larger data parts, such as simple spectrogram patches or sources of audio, rather than combining the additive scores of individual time-frequency points. The extraction of these important, higher-level components makes it possible to provide more meaningful and comprehensible audio explanations. The sample feature importance assigned to the predefined spectrogram patches using LIME and SHAP is shown in Fig. 2.4.

Transformer models are a strong class of neural network architectures that have recently revolutionised various fields of deep learning. Although they were initially introduced for NLP tasks, their impact has extended beyond NLP to computer vision and audio tasks. Examples of transformer models in vision include the ViT [Dosovitskiy et al., 2020] and data-efficient image transformers, which have demonstrated remarkable performance in image classification, object detection, and image segmentation by leveraging self-attention mechanisms and hierarchical representations. Recently, there has been increasing interest in whether the attention maps offered by these modules are useful in explaining the reasoning for a model’s decision [Wiegrefe and Pinter, 2019] proposed possibilities for explaining attention. Therefore, in transformer-vision models, attention mechanisms can be used to provide explanations or visual interpretations. The authors of [Caron et al., 2021] demonstrated that attention maps from different heads could attend to different semantic regions of an image. Thus, providing attention maps can be highly useful, as it allows users to visualise the model’s focus on specific components of the audio spectrogram. To interpret transformers beyond attention visualisation, [Chefer et al., 2020] proposed a new method for computing

relevancy. This method assigns local relevance based on the Deep Taylor Decomposition Principle [Montavon et al., 2017] and then propagates the relevancy scores through the layers. Propagation involves attention layers and skip connections, which pose challenges to the existing feature attribution methods.

In addition to feature importance methods, prototype/example-based explanations offer an effective method for explaining audio models with synthetic or natural input “examples” by addressing questions such as “what kind of input is the model most likely to misclassify?” “which training samples are possibly mislabeled?” or “which input activates an intermediate neuron most?” While [Koh and Liang, 2017, Pruthi et al., 2020] exploits influence functions to the training data points having the most “influence” on the test loss, [Olah et al., 2017] searches for natural examples within a specified set or synthesises examples using gradient descent that maximally activate a neuron of interest for both. The latter also uses an optimisation approach to isolate the causes of model behaviour from mere correlations, aiming to understand what a model is really looking for. Providing natural examples using these methods has a considerable advantage for audio models because the original audio corresponding to an example spectrogram can also be provided to increase model interpretability by humans. Although the original waveform audio is already included in the dataset, it is reconstructed for synthetic examples when phase information is available.

Representation-based explanations serve as a global explanation method for deriving model understanding by analysing the intermediate representations of a deep model. The Network Dissection method in [Bau et al., 2017] was the first to identify a broad set of human-labeled visual semantic concepts, including low-level concepts such as colours and higher-level concepts such as objects. It then gathers the responses of the hidden variables from the CNNs to these known concepts to quantify the alignment of the hidden variable-concept pairs. Although this has proven successful in computer vision, new concepts that target the audio domain, such as frequencies, pitch, phonemes, and other audio parts or concepts, have been identified. TCAV [Kim et al., 2018] measures the sensitivity of a model’s prediction to a user-provided concept by exploiting its internal representation. Whereas network dissection requires concept labels at the level of segmented and annotated samples, TCAV only requires a set of samples where the concept is present and a set where it is not present. This also provides an advantage for concepts that are not localisable, such as frequency-related features in the waveform of a sample. Therefore, the explanations provided by TCAV are not restricted to pre-existing features that provide flexibility for non-expert users to explore the audio concepts they define.

### Waveform Representation

Although MFCCs and spectrograms provide spectral representations of audio, they lose some temporal information when the phase information is discarded. The waveform preserves all temporal characteristics, allowing the model to capture subtle variations and intricate patterns in the audio signal. With advancements in deep model architectures, such as transformers and self-Supervised Learning (SSL) strategies, waveform representation has become favorable for deep audio processing to capture fine-grained temporal details without the need for any feature extraction method [Liu et al., 2022]. In the context of XAI, waveform representation has certain advantages over spectral representation. This offers a more understandable explanation without the need for additional reconstruction. Spectral representations often experience information loss during audio reconstruction to the original domain, primarily because of the discarded phase information.

Wav2Vec 2.0 [Baevski et al., 2020], VQ-Wav2Vec [Baevski et al., 2019], and HuBERT [Hsu et al., 2021] are well-known transformer-based models that use SSL to learn audio representations of large-scale data. They all use raw waveforms as their input representation and are commonly used in various audio and speech processing tasks, as analysed in [Liu et al., 2022]. Providing attention maps for audio transformers trained on raw waveform data is an effective method for enhancing the interpretability of these large models [Diwan et al., 2023].

The raw waveform representation is typically higher in dimension than the spectrogram representation because of the discarded content, such as phase information. Sharing time-point importance in high-dimensional waveform-represented audio using feature attribution methods may be difficult for users to understand. Thus, SHAP and LIME have an advantage in explaining waveform-fed models by directly assigning importance to decomposed audio patches rather than to single time points. Similarly, this could be achieved using additive attribution methods, such as LRP, by aggregating the additive scores of individual time-frequency points. Thus, users are provided with more meaningful audio segments, which enables them to listen to and enhance their understanding.

#### 2.3.2 Audio-specific XAI Methods

Audio processing models have distinctive characteristics compared with models in other domains, such as computer vision and NLP. These characteristics are sourced from both the properties of audio data and the deep model architectures specifically designed to process them. Thus, the de-

velopment of XAI methods specialised in audio models is critical for enhancing the interpretability of these models beyond a superficial understanding. Audio-specific XAI methods aim to explain complex audio signals, leveraging humans’ adeptness at interpreting harmonies, rhythms, and other high-level concepts through listening. Table 2.2 presents a summary of audio-specific XAI methods.

Exploiting existing feature-attribution methods to understand the predictions of audio models is a naive approach. [Becker et al., 2019] explores the interpretability of deep models in the audio domain by using LRP. They aimed to explain multiple deep models trained with raw waveforms and spectrograms for the classification tasks of spoken digits and speaker gender. For both tasks, the authors investigated the connections between feature relevance scores and concepts, such as phonemes or certain frequency ranges. In [Frommholz et al., 2023], the authors used DFT-LRP [Vielhaben et al., 2024], which is a recently modified version of LRP, to explain audio event detection models with different architectures. They scored the importance of each time-frequency component for the predicted class. The insights are then used to determine the most suitable input representation for audio event detection models and to understand the model’s reasoning if it aligns with human requirements.

Table 2.2: Overview of XAI methods specialised on audio processing models.

| XAI Method                                           | Built on                                         | Usage                                                                                                                                                       | Tested on                                                                                   | Category             | Input Representation  | Output Format   | Scope |
|------------------------------------------------------|--------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|----------------------|-----------------------|-----------------|-------|
| <a href="#">[Becker et al., 2019]</a>                | LRP                                              | Investigating feature relevance scores' connections to concepts such as phonemes and certain frequency ranges                                               | UrbanSoundSk <a href="#">[Salamon et al., 2014]</a>                                         | Feature attributions | Spectrogram, waveform | Attribution map | Local |
| <a href="#">[Frommholz et al., 2023]</a>             | DFT-LRP                                          | Determining the most suitable input representation for audio event detection models, understanding the model reasoning if it aligns with human requirements | AudioMNIST <a href="#">[Becker et al., 2019]</a>                                            | Feature attributions | Spectrogram, waveform | Attribution map | Local |
| CoughLIME <a href="#">[Wullenweber et al., 2022]</a> | LIME, loudness/NMF decomposition                 | Generating faithful audio explanations for COVID-19 detection specifically tailored towards cough data                                                      | DiCOVA challenge <a href="#">[Muguli et al., 2021]</a>                                      | Feature attributions | Spectrogram, waveform | Attribution map | Local |
| audioLIME <a href="#">[Haunschnid et al., 2020a]</a> | LIME, Spleeter                                   | Explaining music tagging models with respect to the sources generated by a source separation system                                                         | MillionSongDataset <a href="#">[Bertin-Mahieux et al., 2011]</a>                            | Feature attributions | Spectrogram, waveform | Attribution map | Local |
| <a href="#">[Parekh et al., 2022]</a>                | Surrogate self-exainable model NMF decomposition | Explaining audio processing models with meaningful audio objects, keeping them listenable to the users                                                      | ESC-50 <a href="#">[Piczak, 2015]</a> , SONYC-UST <a href="#">[Cartwright et al., 2020]</a> | Feature attributions | Spectrogram, waveform | Attribution map | Local |
| X-ASR <a href="#">[Wu et al., 2023]</a>              | LIME, SFL, Causal                                | Providing a minimal and sufficient cause for an ASR transcription as a subset of audio frames                                                               | Commonvoice <a href="#">[Ardila et al., 2019]</a>                                           | Feature attributions | Spectrogram, waveform | Attribution map | Local |

Decomposing audio input into meaningful components offers an effective method for disclosing the audio parts to be considered by XAI methods. CoughLIME [Wullenweber et al., 2022] extended the LIME method to explain audio processing models specifically tailored to cough data. The crucial part of CoughLIME, which distinguishes it from directly using standard LIME on audio spectrograms, decomposes the input audio into components that can be interpreted by humans. Two decomposition methods are emphasised: loudness and NMF. While the former targets the extraction of individual cough sounds by computing audio power and thresholding the raw waveform, the latter leverages NMF to decompose the spectrogram matrix into a spectral pattern and a time activation matrix. Following audio decomposition, CoughLIME assigns importance to each decomposed part. Given that the NMF decomposition is applied at the spectrogram level, an inverse short-time Fourier transformation is used to generate listenable explanations from the NMF components that are considered important. The authors showed that CoughLIME can generate faithful audio explanations for COVID-19 detection as illustrated in Figure 2.5. Using a parallel approach, audioLIME [Haunschmid et al., 2020a] extends LIME by creating a perturbation process based on the components extracted through source separation. It aimed to explain music-tagging models with respect to the sources generated by Spleeter [Hennequin et al., 2020] as a source separation system. For example, audioLIME explains a predicted tag “female vocalist” with separated vocals with a female singer or “rock” with components including a driving drum set and a distorted guitar.

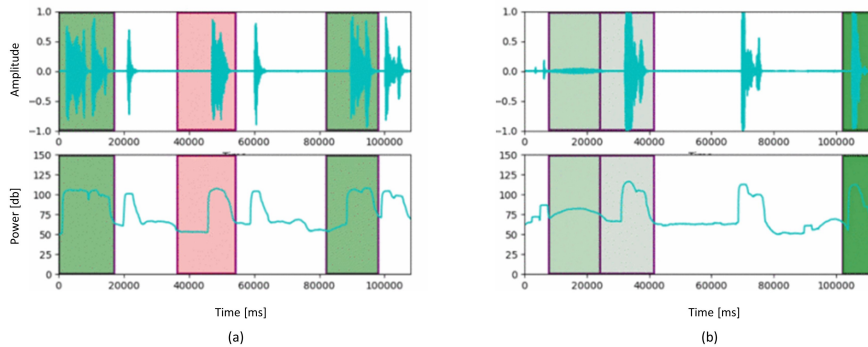


Figure 2.5: Example explanations from CoughLIME [Wullenweber et al., 2022] using loudness decomposition. While both (a) and (b) are COVID-19 positive samples, components highlighted in red (green) account for a COVID-19 negative (positive) prediction.

In addition to utilising the existing XAI methods to create audio-specific methods, the authors of [Parekh et al., 2022] proposed an interpreter network from scratch. They aim to explain audio-processing models with meaningful audio objects and keep them listenable to users. Their interpreter network incorporated NMF as the audio decomposition methodology. They designed 2

surrogate models that form the interpreter module: (1) a regularised interpreter model that takes hidden layer representations of the targeted network as input and produces time activations of pre-learned NMF components as intermediate outputs; and (2) a simple model that pools the time activations of the NMF components and uses a linear layer to mimic the output of the original classifier. It also learns a separate NMF decoder to ensure that the intermediate representation corresponds to the time activations of a pre-learned spectral pattern dictionary. For audio interpretation generation, the important components are selected by estimating their relevance using the pooled time activations and weights of the linear layer in (2). A listenable waveform explanation was obtained by performing NMF inversion using only the relevant components. The authors validated their method qualitatively by mixing audio with a sample of different classes. They also illustrated that the interpretations emphasised the audio object of interest. Fig. 2.6 presents the examples.

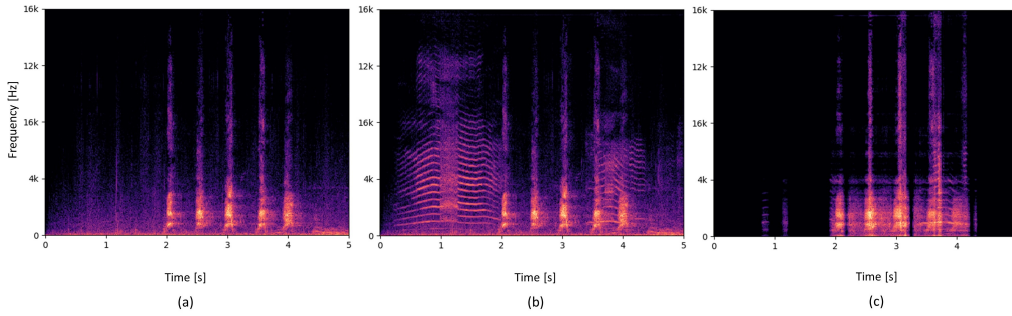


Figure 2.6: Examples of interpretation of audios generated from [Parekh et al., 2022]. (a) original signal from the class “dog,” (b) mixed signal with a signal from the class “crying-baby,” and (c) interpretation audio for the predicted class “dog.”

ASR is a primary task in audio processing, which outputs transcriptions for a given audio input. The length of the transcriptions varies depending on the length of the audio input, which makes explaining them more challenging than using simple classification labels. X-ASR [Wu et al., 2023] provides an explanation for an ASR transcription as a subset of audio frames that is both a minimal and sufficient cause of the transcription. It modifies the existing XAI methods, statistical fault localisation [Sun et al., 2019], causality [Chockler et al., 2021], and LIME to build the framework X-ASR. To generate explanations using these adapted models, it also exploits a classification step, in which correct or incorrect labels are attached based on a similarity metric.

# 3

## Interpretable Audio Classification with Data Decomposition

### Contents

---

|                                                       |           |
|-------------------------------------------------------|-----------|
| <b>3.1 Introduction</b>                               | <b>40</b> |
| <b>3.2 Datasets</b>                                   | <b>42</b> |
| 3.2.1 ComParE                                         | 42        |
| 3.2.2 DiCOVA                                          | 43        |
| <b>3.3 Methodology</b>                                | <b>43</b> |
| 3.3.1 Classifying COVID-19                            | 43        |
| 3.3.2 Explaining COVID-19 Classifiers using CoughLIME | 45        |
| <b>3.4 Experiments</b>                                | <b>48</b> |
| 3.4.1 COVID-19 Classification Results                 | 48        |
| 3.4.2 Explanation Results                             | 50        |

---

In this chapter, we explore explainability in audio-based COVID-19 classification. We begin by providing an overview of COVID-19, relevant biomarkers, and the process of detecting the disease from audio data. The second part focuses on explaining these models using various data decomposition methods. The work presented in this section is a joint effort between Harry Coppock and me, for which we received joint first authorship in the corresponding publication [[Akman et al.](#),

2022]. Additionally, the XAI component is based on a collaborative study with Anne Wullenweber, culminating in our research paper [Wullenweber et al., 2022], where we contributed equally. This chapter includes sections of the both published works.

## 3.1 Introduction

Trust is a fundamental prerequisite for using AI models in sensitive domains, particularly in healthcare, where decisions can directly affect patient outcomes. In this part, we address the trustworthiness of AI models in healthcare, with a specific emphasis on the COVID-19 detection task. Focusing on audio-based COVID-19 classifiers, we aim to explain their decision-making processes and evaluate their performance with respect to both trustability and generalisability.

The Coronavirus Disease 2019 (COVID-19) pandemic, caused by the severe-acute-respiratory-syndrome-coronavirus 2 (SARS-CoV-2), has infected confirmed 777 million people and resulted in 7,089,979 deaths <sup>1</sup>. Mass testing schemes offer the option to monitor and implement a selective isolation policy to control the pandemic without the need for regional or national lockdown [Peto et al., 2020]. However, physical mass testing methods, such as the Lateral Flow Test (LFT) have come under criticism since the tests divert limited resources from more critical services [Wise, 2020, Holt, 2021] and due to suboptimal diagnostic accuracy. Sensitivities of 58% have been reported for self-administered LFTs [Mahase, 2020], unacceptably low when used to detect active virus, a context where high sensitivity is essential to prevent the reintegration into society of falsely reassured infected test recipients [Moyse, 2021]. In addition to mass testing, radar remote life sensing technology offers non-contact applications to combat COVID-19 including heart rate tracking, identity authentication, indoor monitoring and gesture recognition [Islam et al., 2021].

Investigating the potential for digital mass testing methods is an alternative approach, based on findings that suggest a biological basis for identifiable vocal biomarkers caused by SARS-CoV-2's effects on the lower respiratory track [Quatieri et al., 2020]. This has recently been backed up by empirical evidence [Bartl-Pokorny et al., 2020]. Efforts have been made to collect and classify a range of different modality audio recordings of COVID-19-positive and COVID-19-negative individuals and several datasets have been released that use applications to collect the breath and cough of volunteer individuals. Examples include the 'COUGHVID' [Orlandic et al., 2021], 'Breath for Science'<sup>2</sup>, 'Coswara' [Sharma et al., 2020], COVID-19 sounds<sup>3</sup>, and 'CoughAgainst-

<sup>1</sup>as of 16th February 2025 <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>

<sup>2</sup><https://www.breathforscience.com>

<sup>3</sup><https://www.covid-19-sounds.org/en/>

Covid’ [Bagad et al., 2020]. In addition, to focus the attention of the audio processing community onto the task of binary classification of COVID-19 from audio, two INTERSPEECH competitions: the INTERSPEECH 2021 Computational Paralinguists Challenge (ComParE)<sup>4</sup> [Schuller et al., 2021a] with its COVID-19 Cough and Speech Sub-Challenges, and Diagnosing COVID-19 using acoustics (DiCOVA)<sup>5</sup> [Muguli et al., 2021] have been organised with this focus as their challenge.

Several studies have been published that propose machine learning-based COVID-19 classifiers exploiting distinctive sound properties between positive and negative cases to classify these datasets. [Brown et al., 2020a] and [Ritwik et al., 2020] demonstrate that simple machine learning models perform well in these relatively small datasets. In addition, deep neural networks are exploited in [Laguarta et al., 2020, Pinkas et al., 2020, Imran et al., 2020, Nessiem et al., 2021] with proven performance at the COVID-19 detection task. Although there are works that try to combine different modalities computing the representations separately, CIdER [Coppock et al., 2021a] proposes an approach computing joint representation of a number of modalities. The adaptability of this approach to different types of datasets has not to our knowledge been explored or reported.

Some proposed applications detecting COVID-19 from speech or cough data delivers good results [Qian et al., 2021]; however, in order for them to be widely accepted, trust in the systems is crucial. To establish trust, users must understand how the model determines its predictions. ML models are traditionally black boxes – their predictions can, if at all, hardly be explained in human understandable ways [Samek et al., 2021]. Explainable artificial intelligence (XAI) aims at providing explanations for the output of an AI application that are understandable to humans. This is especially crucial in sensitive domains such as healthcare or finance [Burkart and Huber, 2020]. Besides establishing human trust and acceptance of a model’s predictions, Burkart et al. list various reasons for the importance of XAI, among them improved understanding of the model and fair and ethical decision making [Burkart and Huber, 2020]. While many promising applications for detecting COVID-19 from audio data have been proposed, none of them focus on providing explanations for their models’ decisions. We argue that in order for a COVID-19 detection application to be deployed, trust, hence explaining its predictions, is crucial. Additionally, methods to explain COVID-19 prediction models can help medical experts and users understand and accept the models’ decisions. Intuitively, explanations for predictions on audio data should also be provided in form of audio. However, existing XAI techniques almost exclusively focus on visual and textual explanations.

---

<sup>4</sup><http://www.compare.openaudio.eu/>

<sup>5</sup><https://dicova2021.github.io/>

To this end, we first share a modified version of CIdER, a recently developed end-to-end deep learning neural network optimised for binary COVID-19 diagnosis from cough and breath audio [Coppock et al., 2021a], which is applicable to common datasets with further modalities. We present the competitive results of CIdER to the two COVID-19 cough and speech Challenges of INTERSPEECH 2021, ComParE and DiCOVA. We also investigate the behaviour of a strong COVID-19 classifier across different datasets by running cross dataset experiments with CIdER. We describe the limitations of current COVID-19 classifiers with these experiments regarding the ultimate goal of proposing a universal COVID-19 classifier.

In the second part, we introduce CoughLIME to obtain audio explanations for the predictions of ML applications. It extends Ribeiro et al.’s approach providing LIME [Ribeiro et al., 2016] for use with audio data, specifically tailored towards cough data. To demonstrate its functionality, we apply CoughLIME to explain the predictions of the CIdER model.

## 3.2 Datasets

### 3.2.1 ComParE

ComParE hosted two COVID-19 related sub-challenges, the COVID-19 Cough Sub-Challenge (CCS) and the COVID-19 Speech Sub-Challenge (CSS). Both CCS and CSS are subsets of the crowd sourced Cambridge COVID-19 sound database [Brown et al., 2020a, Han et al., 2021]. CCS consists of 926 cough recordings from 397 participants. Participants provided 1-3 forced coughs resulting in a total of 1.63 hours of recording. CSS is made up of 893 recordings from 366 participants totalling 3.24 hours of recording. Participants were asked to recite the phrase *“I hope my data can help manage the virus pandemic”* in their native language 1-3 times. The train-test splits for both sub-challenges are detailed in Table 3.1.

Table 3.1: ComParE sub-challenge dataset splits. Values specify the number of audio recordings, not the number of participants.

| #                 | CCS   |     |      | CSS   |     |      |
|-------------------|-------|-----|------|-------|-----|------|
|                   | train | val | test | train | val | test |
| COVID-19-postive  | 71    | 48  | 39   | 72    | 142 | 94   |
| COVID-19-negative | 215   | 183 | 169  | 243   | 153 | 189  |
| Total             | 286   | 231 | 208  | 315   | 295 | 283  |

### 3.2.2 DiCOVA

Once again, DiCOVA hosted two COVID-19 audio diagnostic sub-challenges [Muguli et al., 2021]. Both sub-challenge datasets were subsets of the crowd sourced Coswara dataset [Sharma et al., 2020]. The first sub-challenge, named Track-1, comprised of a set of 1,274 forced cough audio recordings from 1,274 individuals totalling 1.66 hours. The second, Track-2, was a multi-modality challenge, where 1,199 individuals provided three separate audio recordings; deep breathing, sustained vowel phonation, and counting from 1-20. This dataset represented a total of 14.9 hours of recording. The train-test splits are detailed in Table 3.2

Table 3.2: DiCOVA sub-challenge dataset splits. The test set labels were withheld by the DiCOVA team, contestants had to submit predictions for each test case, on which a final AUC was returned.

| #                 | Track-1     |       | Track-2     |      |
|-------------------|-------------|-------|-------------|------|
|                   | train + val | test  | train + val | test |
| COVID-19-postive  | 75          | blind | 60          | 21   |
| COVID-19-negative | 965         | blind | 930         | 188  |
| Total             | 1,040       | 234   | 990         | 209  |

## 3.3 Methodology

### 3.3.1 Classifying COVID-19

#### Classifier Model

CIdeR is a 9 layer convolutional residual network. A schematic detailing of the model can be seen in Figure 3.1. Each layer or block consists of a stack of convolutional layers with Rectified Linear Units (ReLUs). Batch normalisation [Ioffe and Szegedy, 2015] also features in the residual units, acting as a source of regularisation and supporting training stability. A fully connected layer with sigmoid activation terminates the model yielding a single logit output which can be interpreted as an estimation of the probability of COVID-19. As detailed in Figure 3.1 the network is modified to be compatible with a varying number of modalities, for example, if a participant has provided cough, deep breathing, and sustained vowel phonation audio recordings, they can be stacked in a depth wise manner and passed through the network as a single instance. We use PyTorch library in python to implement CIdeR and baseline models.

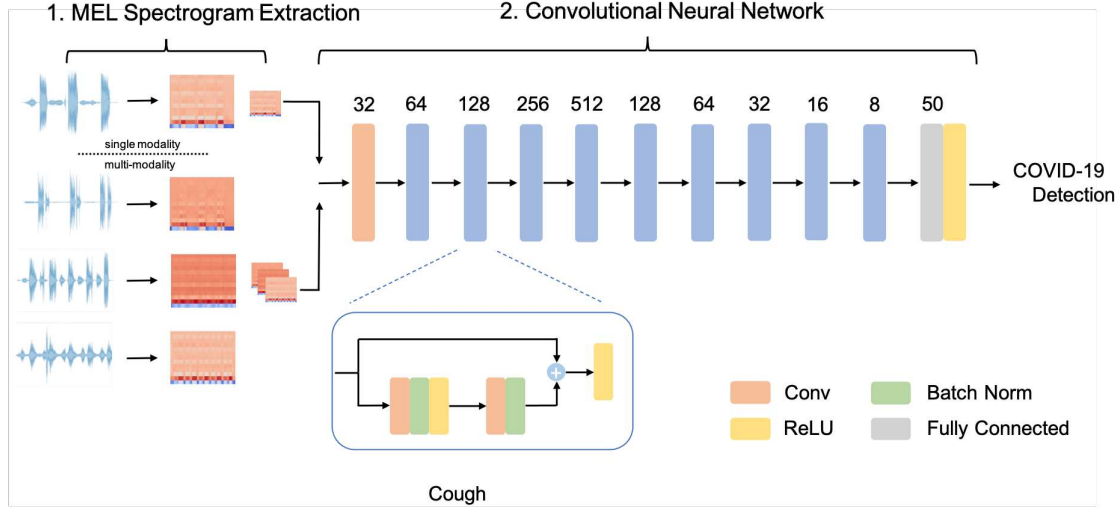


Figure 3.1: A schematic of the **COVID-19 Identification ResNet, (CIdER)**. The figure shows a blow-up of a residual block, consisting of convolutional, batch normalisation, and Rectified Linear Unit (ReLU) layers.

### Pre-processing

At training time, a window of  $s$ -seconds, which was fixed at 6 seconds for these challenges, is sampled from the audio recording randomly. If the audio recording is less than  $s$ -seconds long, the sample is padded with repeated versions of itself. The sampled audio is then converted into MFCCs resulting in an image of width  $s * \text{the sample rate}$  and height equal to the number of MFCCs. Three data augmentation steps are then applied to the sample. First, the pitch of the recording is randomly shifted, secondly, bands of the Mel spectrogram are masked in the time and Mel coefficient axes and finally, Gaussian noise is added. At test time, the sampled audio recording is chunked into a set of  $s$ -second clips and processed in parallel. The mean of the set of logits is then returned as the final prediction.

### Baselines

The DiCOVA team ran baseline experiments for the track 1 (coughing) sub-challenge; only the best performing (MLP) model's score was reported. For the track 2 (deep breathing/vowel phonation/counting) sub-challenge, however, baseline results were not provided. Baseline results were provided for the ComParE challenge but only Unweighted Average Recall (UAR) was reported rather than Area Under Curve of the Receiver Operating Characteristics curve (ROC-(AUC)). To allow comparison across challenges, we created new baseline results for the ComParE sub-

challenges and the DiCOVA Track 2 sub-challenge, using the same baseline methods described for the DiCOVA Track 1 sub-challenge. The three baseline models applied to all four sub-challenge datasets were Logistic Regression (**LR**), MLP, and Random Forrest (**RF**), where the same hyperparameter configurations that were specified in the DiCOVA baseline algorithm was used [Muguli et al., 2021].

To provide a baseline comparison for the CIdER track 2 results, we build a multimodal baseline model. We follow a similar strategy with the provided DiCOVA baseline algorithm, while extracting the features for each modality. Rather than individual training for different models, we develop an algorithm that concatenates input features from separate modalities. Then, this combined feature set is fed to the baseline models: LR, MLP, and RF.

We use 39 dimensional MFCCs as our feature type to represent the input sounds. For LR, we use Least Square Error (L2) as a penalty term. For MLP, we use a single hidden layer of size 25 with a Tanh activation layer and L2 regularisation. The Adam optimiser and a learning rate of 0.0001 is used. For RF, we build the model with 50 trees and split based on the gini impurity criterion.

### 3.3.2 Explaining COVID-19 Classifiers using CoughLIME

CoughLIME extends the basic LIME functionality developed by [Ribeiro et al., 2016]. It generates explanations by determining and highlighting the parts of the input data responsible for a model’s decision. Calculating the weights  $w_i \in \mathbb{R}$ , it determines both the components that tend the model towards a positive prediction as well as the ones that account for a negative prediction.

#### Audio Decomposition

A crucial part of generating explanations with CoughLIME is decomposing the input audio into components which can be interpreted by humans. We additionally emphasise that, in order to produce listenable explanations, it must be possible to reconstruct an audio file from these components. To create the interpretable components, we experiment with different approaches for spectral and temporal masking. Employing these rather simple decompositions is motivated by the fact that they have led to excellent results for audio data augmentation [Park et al., 2019]. To compare the obtained results with a more complex approach, we further implement a decomposition based on NMF. Due to their greater complexity, we only further illustrate the loudness

decomposition and the NMF decomposition. Results for CoughLIME with the temporal, spectral and loudness-spectral decomposition as well as the code for all experiments is publicly available<sup>6</sup>.

The proposed novel loudness decomposition aims at extracting individual cough sounds. For the purpose of this, we compute the power  $\mathbf{p}$  in  $dB$  of the cough sounds. We round  $\mathbf{p}$  to the nearest 10 to avoid an excess of components and calculate local minima using Equation 3.1:

$$\mathbf{z} = \underset{x}{\operatorname{argmin}} \mathbf{p}_{rounded}(x) \quad (3.1)$$

To avoid creating components that do not entail an entire cough, we introduce a threshold  $\gamma$  as a hyperparameter and only generate a new component if  $\mathbf{p}_{z(i)} < \gamma$ . Component  $i$  is then created by taking the subset of the audio array between indices  $z_{i-1}$  and  $z_i$  along the temporal axis. Figure 3.2 illustrates this approach. We point out that the number of generated components  $n$  varies depending on the audio file and on  $\gamma$ .

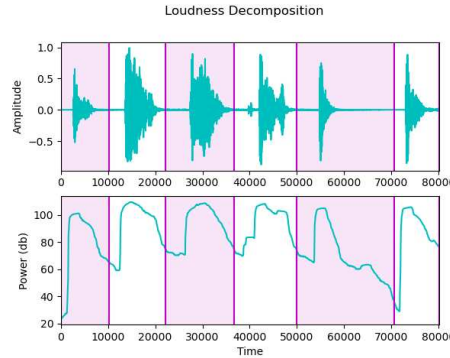


Figure 3.2: Example loudness decomposition of a cough audio file with threshold 75 dB.

NMF decomposes a matrix  $\mathbf{V} \in \mathbb{R}_{\geq 0}^{n \times m}$  into two matrices  $\mathbf{W} \in \mathbb{R}_{\geq 0}^{n \times k}$  and  $\mathbf{H} \in \mathbb{R}_{\geq 0}^{k \times m}$  such that  $\mathbf{V} = \mathbf{WH}$ . As audio signals are not typically non-negative, a cough signal is first transformed to its spectrogram using the discrete Fourier transform. Taking the magnitude of the spectrogram then results in a non-negative matrix which can be decomposed into  $k$  components using NMF. The hyperparameter  $k$  is set by the user. We ensure the decomposed signal can be transformed back to audio to generate listenable explanations by retaining the phase information when calculating the spectrogram’s magnitude. The described decomposition is implemented using the Python packages *librosa* and *sklearn.decomposition*. Figure 3.3 shows an example for the NMF decomposition of a cough audio file into six components.

<sup>6</sup><https://github.com/glam-imperial/CoughLIME>

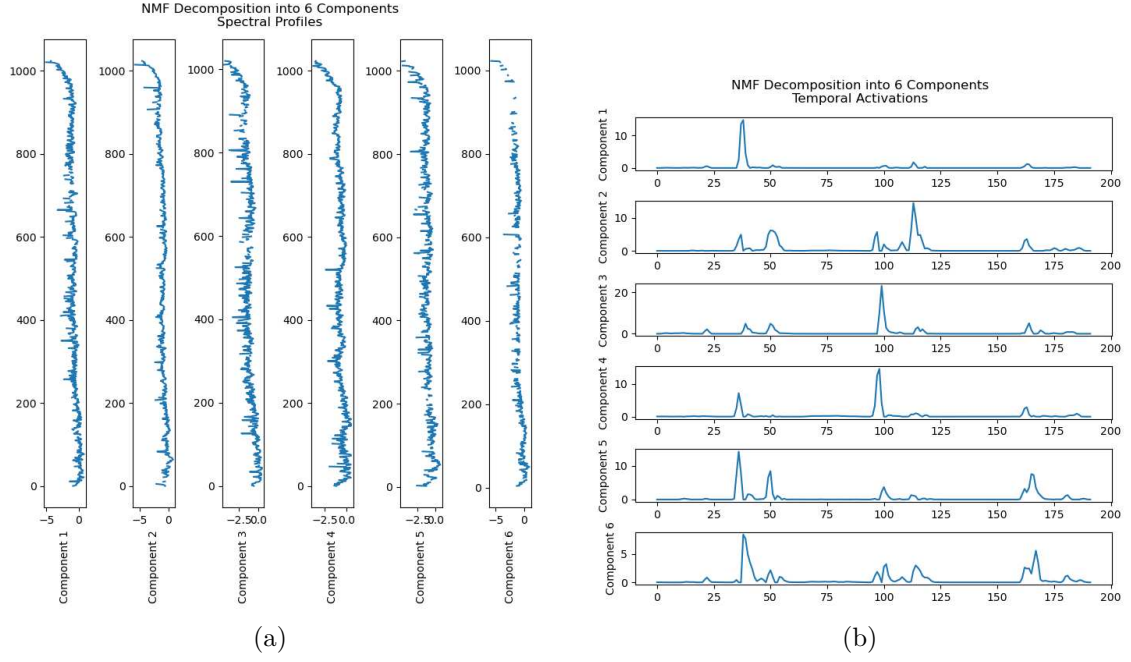


Figure 3.3: Example NMF decomposition of a cough audio file into 6 components consisting of spectral profiles (a) and temporal activations (b).

### Explanation Generation

Once the input cough audio is decomposed into interpretable components, explanations are generated according to the procedure from [Ribeiro et al., 2016]. The decomposed input is given by a binary vector  $\mathbf{x}' \in \{0, 1\}^{d'}$  indicating the presence or absence of the individual components. Hence, to obtain the representation of the original audio data, all entries in  $\mathbf{x}'$  would be set to 1. To generate the training data for the sparse linear surrogate model, the input audio is perturbed by randomly setting entries in  $\mathbf{x}'$  to 0 resulting in  $n$  training samples  $\mathbf{z}' \in \{0, 1\}^{d'}$ . The loss function of the linear surrogate model is given in Equation 3.2:

$$\mathcal{L}(f, g, \pi_x) = \sum_{\mathbf{z}, \mathbf{z}' \in \mathcal{Z}} \pi_x(\mathbf{z})(f(\mathbf{z}) - g(\mathbf{z}'))^2 \quad (3.2)$$

with  $f$  the black-box model and  $g$  the surrogate model [Ribeiro et al., 2016].  $\pi_x(\mathbf{z})$  is a proximity measure which takes the distance of the perturbed training samples  $\mathbf{z}$  to the original data instance  $\mathbf{x}$  into account. As proposed in [Ribeiro et al., 2016], we use an exponential kernel with cosine distance measure. The overall generated explanation is given by a dictionary associating the weights  $w_i \in \mathbb{R}$  of the different components with the component indices, sorted in descending order according to the absolute value of the weights. An explanation is always generated for a certain

class label. Negative weights correspond to components that tend the model towards predicting that the component does not belong to the class whereas positive weights are associated with components that tend the model towards predicting that the sample is part of the class. Sonified explanations are obtained by factoring the  $k$  most important components together in the original representation. Listenable explanations were the main focus of this work, however, complementing them with visualised explanations can arguably be beneficial. It is therefore also possible to obtain a visualised explanation highlighting the most important components of the input representation and their respective weights  $w_i \in \mathbb{R}$ . Hereby, focus was set on a modular implementation: CoughLIME which extends the LIME base class can be used with different decompositions to explain the predictions of any model taking audio as input data. We emphasise that the compatibility of the CoughLIME with any COVID-19 detection algorithm from audio data enables the comparison of different classifiers.

## 3.4 Experiments

### 3.4.1 COVID-19 Classification Results

The results from the array of experiments with CIdeR and the 3 baseline models are detailed in Table 3.3. CIdeR performed strongly across all four sub-challenges, achieving AUCs of 0.799 and 0.787 in the DiCOVA Track 1 and 2 sub-challenges, respectively, and 0.732 and 0.787 in the ComParE CCS and CSS sub-challenges. In the DiCOVA cough sub-challenge, CIdeR significantly outperformed all three baseline models based on 95% confidence intervals calculated following [Hanley and McNeil, 1982], and in the DiCOVA breathing and speech sub-challenge it achieved a higher AUC although the improvement over the baselines was not significant. Conversely, while CIdeR performed significantly better than all three baseline models in the ComParE speech sub-challenge based on 95% confidence intervals calculated following [Hanley and McNeil, 1982], it performed no better than baseline in the ComParE cough sub-challenge. One can speculate that this may have resulted from the small dataset sizes favouring the more classical machine learning approaches which do not need as much training data.

Table 3.3: Results for CIDEr and a range of baseline models for 4 sub-challenges across the DiCOVA and ComParE challenges. Testing is performed on the held-out test fold once final model decisions have been made on the validation sets. The Area Under Curve of the Receiver Operating Characteristics curve (AUC(-ROC)) is displayed. A 95 % confidence interval is also shown following [Hanley and McNeil, 1982]. CIDEr scores which are statistically higher than the best baseline results with a 95 % confidence are in bold. The three baseline models are Logistic Regression (LR), Multi-layer Perceptron (MLP), and Random Forrest (RF). All baseline models were trained on MFCC features.

|         | sub-challenge* | CIDEr                  | Baseline        |                 |                 |
|---------|----------------|------------------------|-----------------|-----------------|-----------------|
|         |                |                        | LR              | MLP             | RF              |
| DiCOVA  | Track 1**      | <b>.799</b> $\pm$ .058 | -               | .699 $\pm$ .068 | -               |
|         | Track 2        | .786 $\pm$ .057        | .647 $\pm$ .014 | .684 $\pm$ .072 | .776 $\pm$ .063 |
| ComParE | CCS            | .732 $\pm$ .068        | .722 $\pm$ .069 | .765 $\pm$ .065 | .753 $\pm$ .066 |
|         | CSS            | <b>.787</b> $\pm$ .060 | .583 $\pm$ .072 | .656 $\pm$ .070 | .628 $\pm$ .070 |

\*Track 1: coughing, Track 2: deep breathing + vowel phonation + counting, CCS: coughing, CSS: speech – ‘*hope my data can help managethe virus pandemic*’

\*\* As the demographics were not provided for the Track 1 test set, when calculating the AUC confidence intervals, it was assumed that there was an equal number of COVID-19-positive and COVID-19-negative recordings.

### Limitations

A key limitation with both the ComParE and DiCOVA COVID-19 challenges is the size of the datasets. Both datasets contain very few COVID-19-positive participants. Therefore, the certainty in results is limited and this is reflected in the large 95 % confidence intervals detailed in Table 3.3. This issue is compounded by the demographics of the datasets. As detailed in [Brown et al., 2020a] and in [Muguli et al., 2021] for the ComParE datasets and the DiCOVA datasets, respectively, not all demographics from society are represented evenly – most notably, there is poor coverage of age and ethnicity and both datasets are skewed towards the male gender.

In addition, the crowd-sourced nature of the datasets introduces some confounding variables. Audio is a tricky sense to control. It contains a lot of information about the surrounding environment. As both datasets were crowd-sourced, there could have been correlations between ambient sounds and COVID-19 status, for example, sounds characteristic of hospitals or intensive care units being more often present for COVID-19-positive recordings compared to COVID-19-negative recordings. As the ground truth labels for both datasets were self-reported, presumably the participants knew at the time of recording whether they had COVID-19 or not. One could postulate that the individuals who knew they were COVID-19-positive might have been more fearful than COVID-19-negative participants at the time of recording, an audio characteristic known to be identifiable by machine learning models [Trigeorgis et al., 2016]. Therefore, the audio features

which have been identified by the model may not be specific audio biomarkers for the disease.

We note that both the DiCOVA Track 1 and ComParE CCS sub-challenges were cough recordings. Therefore, there was an opportunity to utilise both training sets. Despite having access to both the DiCOVA and ComParE datasets, training on the two datasets together did not yield a better performance on either of the challenges’ test sets. Additionally, a model which performed well on one of the challenges test sets would see a marked drop in performance on the other challenge’s test set. We run cross dataset experiments to analyse this effect further. For these experiments, we also included the COUGHVID dataset[Orlandic et al., 2021] in which COVID-19 labels were assigned by experts and not as a results of clinically validated test. The results in Table 3.4 show that the trained models for each dataset do not generalise well and perform poorly on excluded datasets. This is a worrying find, as it suggests that audio markers which are useful in COVID-19 classification in one dataset are not useful or present in the other dataset. This agrees with the concerns presented in [Coppock et al., 2021c] that current COVID-19 audio datasets are plagued with bias, allowing for machine learning models to infer COVID-19 status, not by audio biomarkers uniquely produced by COVID-19, but by other correlations in the dataset such as nationality, comorbidity and background noise.

Table 3.4: The results for cross dataset experiments.

| Train Set | Test Set |         |          |
|-----------|----------|---------|----------|
|           | DiCOVA   | ComParE | COUGHVID |
| DiCOVA    | .799     | .554    | .464     |
| ComParE   | .512     | .732    | .552     |
| COUGHVID  | .395     | .518    | .566     |
| All       | .673     | .717    | .531     |

### 3.4.2 Explanation Results

To assess CoughLIME’s ability to faithfully explain COVID-19 prediction models, we generate explanations for the CIdER model trained on the publicly available data set from the DiCOVA challenge. The data set consists of 1040 samples which can be split for a 5-fold cross-validation with instance-per-fold lists provided by the challenge. CIdER is a deep ResNet model specifically developed for COVID-19 detection from audio data. It achieved an AUC of 79.9% on the blind test set. We point out that CoughLIME is a model-agnostic approach – it is hence of minor importance which COVID-19 classification model we explain for the purpose of this demonstration.

As explained above, we focus on sonified explanations as they are the most intuitive way to explain the behaviour of a model acting on audio data. Listenable examples can be found in our repository<sup>7</sup>.

To quantify the faithfulness of the generated explanations, we adopt pixel flipping to audio. This evaluation method was originally proposed for image data and argues that if an XAI method correctly highlights the important parts of the input representation, ‘flipping’ these components to 0 should lead to a rapid decrease in performance [Samek et al., 2017]. To this end, we generate explanations using CoughLIME with the loudness decomposition for all 218 files contained in fold 1 of the DiCOVA data set. We then calculate the percentage of the predictions having flipped the most significant components to 0 that lead to the same class prediction as the entire file. We compare this percentage to the percentage of same class predictions having flipped the same number of randomly chosen components. As the loudness decomposition generates a variable number of components, we flip a percentage of all generated components  $n$  to establish a basis of comparison. The results and the standard deviation over two runs are reported in Figure 3.4. All results were generated with a threshold  $\gamma = 75dB$ . For all  $k$  and all decompositions tested, the predictions on the components found important by CoughLIME lead to superior results than the predictions on random components.

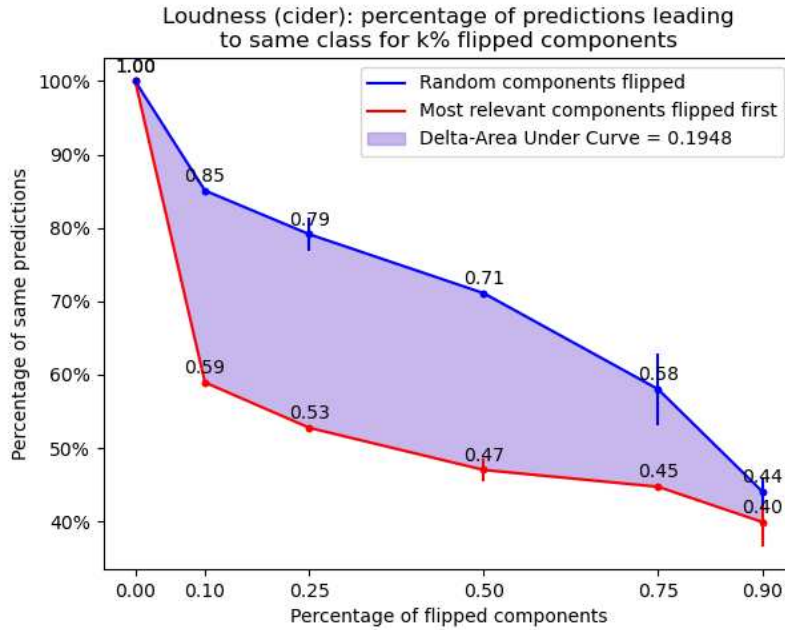


Figure 3.4: Pixel flipping for audio for the loudness decomposition with  $\gamma = 75dB$  shows CoughLIME is able to correctly identify the most important components. Results are averaged over 2 runs.

To establish a base of comparison for the pixel flipping results, we introduce a new metric, the

<sup>7</sup><https://github.com/glam-imperial/CoughLIME>

Delta-Area Under the Curve ( $\Delta AUC$ ), which is defined in Equation 3.3:

$$\Delta AUC = \frac{\int_{c_1}^{c_n} y_2(x) dx - \int_{c_1}^{c_n} y_1(x) dx}{\int_{c_1}^{c_n} 1 dx} \quad (3.3)$$

with  $\mathbf{c}$  being the percentages of components flipped,  $y_1$  the curve generated by flipping the most relevant components first and  $y_2$  the curve generated by flipping random components. The  $\Delta AUC$  is given by the percentage of the entire graph area which is covered by the area between the curves generated for the most relevant and the random components. The higher the  $\Delta AUC$ , the better the explanations. Table 3.5 compares the  $\Delta AUC$  obtained for the proposed decompositions.

Table 3.5: Comparison of the  $\Delta AUC$  achieved with different decompositions.

| Decomposition | Temporal | Spectral | Loudness | Loudness-Spectral | NMF   |
|---------------|----------|----------|----------|-------------------|-------|
| $\Delta AUC$  | 13.26%   | 9.38%    | 19.48%   | 5.68%             | 4.13% |

We argue that it can be beneficial to complement listenable explanations with visual or textual ones to obtain a better overall understanding of the system. Therefore, we implement methods to highlight the components found most important by the classification model together with their relative weights (see Figure 3.5).

In summary, the exploration of explainability in the context of audio-based COVID-19 classification presents a compelling research direction, particularly given its role in enabling trustworthy decision-making in healthcare and ensuring transparent application by medical experts in this domain. However, as discussed in the limitations section, the classification model used in this task raises concerns regarding generalisability, largely due to the limited and imbalanced datasets as well as potential biases introduced through their crowd-sourced and self-reported nature. Since explainability methods are only as reliable as the models they aim to interpret, building trustworthy explanations requires an underlying classifier that demonstrates consistent and robust performance. Without this foundation, explainability results may risk being misleading, regardless of their methodological soundness. While our findings provide valuable insights into how audio decomposition techniques can enhance transparency in healthcare-oriented AI systems, the subsequent chapters extend this investigation to more generalisable tasks, specifically audio classification and speech emotion recognition, where model performance is more robustly established. These domains provide a more stable ground for validating and refining our explainability framework, as presented in the following chapters.

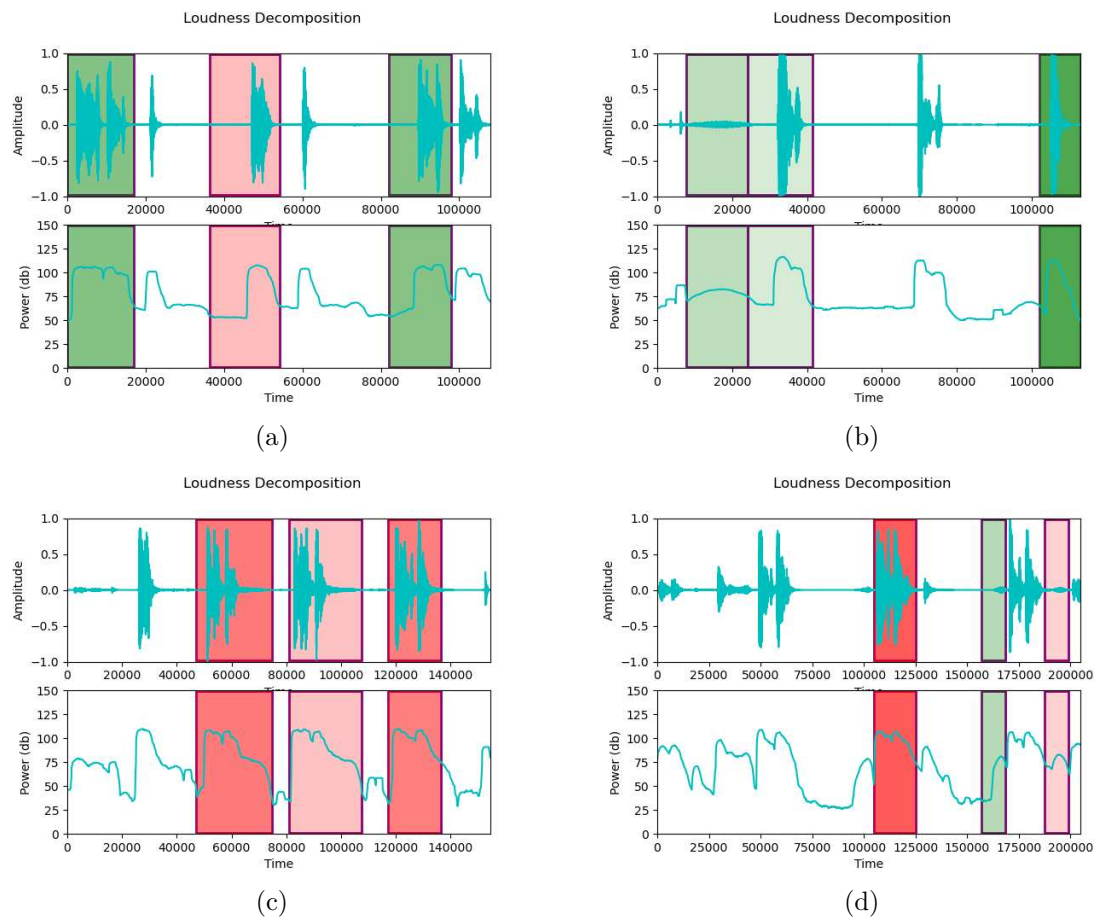


Figure 3.5: Example explanations for COVID-19 positive (a), (b) and COVID-19 negative (c), (d) coughs generated with the loudness decomposition. Components highlighted in red (green) account for a COVID-19 negative (positive) prediction.

# 4

## Understanding Audio Transformers through Attention and Factorisation

### Contents

---

|                               |           |
|-------------------------------|-----------|
| <b>4.1 Introduction</b>       | <b>55</b> |
| <b>4.2 Methodology</b>        | <b>56</b> |
| 4.2.1 NMF Fundamentals        | 56        |
| 4.2.2 AttHear Design          | 57        |
| 4.2.3 Explanation Generation  | 58        |
| <b>4.3 Experiments</b>        | <b>59</b> |
| 4.3.1 Implementation Details  | 59        |
| 4.3.2 Quantitative Evaluation | 59        |
| 4.3.3 Qualitative Evaluation  | 62        |

---

In this chapter, we present our proposed method for interpreting audio-processing transformer models. Our approach leverages the attention mechanisms inherent in these models and combines them with NMF to compute the relevance of audio inputs. By integrating these techniques, we

aim to enhance the explainability of transformer-based audio models, providing deeper insights into their decision-making processes. This chapter incorporates sections from our published work [Akman and Schuller, 2024a].

## 4.1 Introduction

Generating audio explanations for audio processing AI models has been gaining importance due to its intuitiveness on audio-based tasks and higher expressiveness over other modalities in specific scenarios, such as where understanding visual explanations needs expertise [Schuller et al., 2021b]. Recent studies have adapted domain-agnostic XAI methods to specialise in providing audio explanations [Becker et al., 2019, Frommholz et al., 2023, Vielhaben et al., 2024, Wu et al., 2023, Haunschmid et al., 2020b]; however, they only provide feature importance maps as explanations which are difficult to be interpreted by humans. Aiming to generate listenable and interpretable audio explanations, [Wullenweber et al., 2022, Parekh et al., 2022] exploit NMF [Lee and Seung, 2000] to decompose audio into meaningful components. However, both methods require training a separate surrogate model around the original model.

Transformer models are extensively used in audio processing to obtain state-of-the-art performance on various tasks such as automatic speech recognition, keyword spotting, and speaker recognition [Gong et al., 2021, Hsu et al., 2021, Chen et al., 2021, Baevski et al., 2019, Chi et al., 2021]. Certain studies target to explain transformer models by leveraging their attention mechanism and presenting attention weights as explanations [Vig, 2019, Yeh et al., 2024, Abnar and Zuidema, 2020]. Although attention visualisation works well with tasks in a visual or textual domain, there is still a need to explain audio transformer models more intuitively.

To address these issues, we propose AttHear, which combines the transformer’s attention mechanism with NMF to compute relevancy for audio inputs. We first apply NMF on input audio to extract its spectral pattern dictionary with corresponding time activations. We then analyse the attention behaviour of the transformer model for spectral component groups, which are created depending on their similarity. These attention weights modify the original time activations of the spectral pattern dictionary. In the final step, AttHear reconstructs the relevant audio part to the model decision using the spectral pattern dictionary and the modified time activations. We verify that AttHear can generate faithful explanations with high fidelity scores through experiments that simulate removing relevant features and assess the original model’s performance on audio parts

containing these essential features. The main contributions of this study are as follows:

- We propose a novel audio explanation method, AttHear, that targets audio transformers. Our method combines the attention module in transformers and NMF without needing to finetune the original prediction model or train a surrogate model.
- We introduce a new formulation for NMF reconstruction that incorporates attention weights in NMF time activations. Using this formulation, AttHear generates audio explanations in a listenable format that are interpretable to the end-user.
- We evaluate our model on keyword spotting and environmental sound classification tasks. We show that while AttHear provides faithful explanations with high fidelity scores, it captures meaningful high-level audio components for the investigated tasks.

## 4.2 Methodology

This section elaborates on the AttHear system’s design and the procedures we employ to generate meaningful audio explanations. We begin by introducing NMF’s fundamentals and rationale for incorporating it into our framework. Following this, we provide a detailed description of our explainer system, which involves the integration of NMF with an attention mechanism in transformer audio models. Lastly, we outline the steps for producing meaningful audio explanations using our system. We present an overview of AttHear in Figure 4.1.

### 4.2.1 NMF Fundamentals

NMF’s ability to disentangle overlapping audio components while preserving the inherent non-negativity of audio data allows it to effectively identify underlying sources such as instruments, vocals, and background noise. The formulation for NMF is defined in Equation 4.1.

$$X \approx WH \tag{4.1}$$

In our audio decomposition case,  $X \in \mathbb{R}_+^{F \times T}$  represents the input audio spectrogram, where  $F$  and  $T$  denote frequency bins and time points respectively,  $W \in \mathbb{R}_+^{F \times K}$  is a spectral pattern dictionary with  $K$  components and  $H \in \mathbb{R}_+^{K \times T}$  is a matrix for time activations. While each

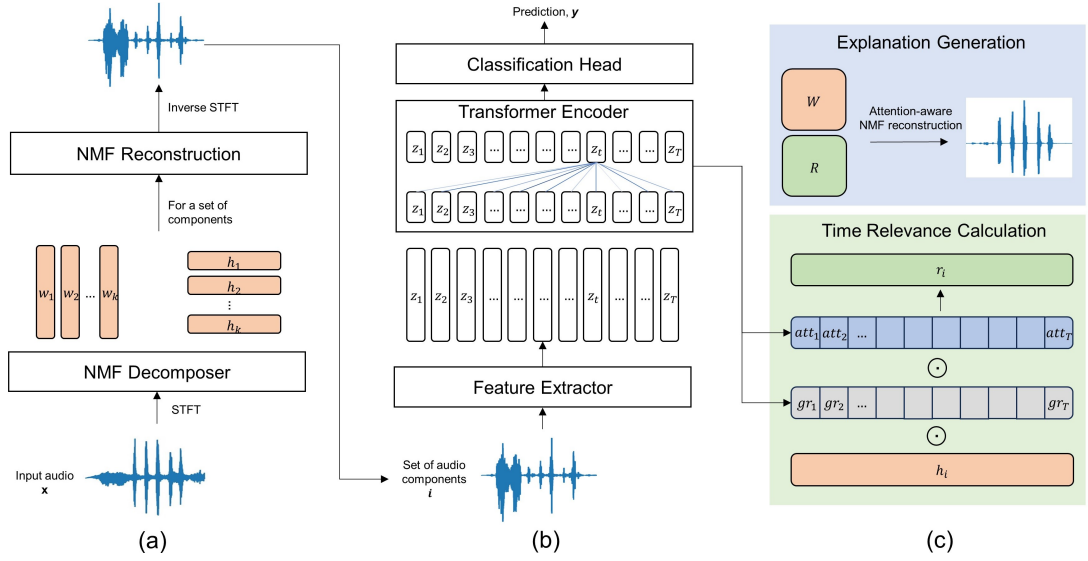


Figure 4.1: An overview of AttHear: (a) Depicts NMF procedure for an input audio. (b) Indicates the transformer model inference for an audio segment to compute corresponding attention weights  $att$  and their gradients  $gr$ . (c) Represents the calculation of temporal relevance  $r$  where  $\odot$  denotes element-wise multiplication and the explanation generation.

column  $w_k$  in  $W$  embodies a spectral pattern extracted through NMF, each row  $h_k$  in  $H$  captures time activation information for each spectral pattern. Also, NMF enables an audio spectrogram reconstruction from a selected group of spectral patterns. To reconstruct an audio spectrogram  $X_k$  using a single spectral pattern is formulated in Equation 4.2, where both  $/$  and  $\odot$  denote element-wise operations.

$$X_k = \frac{w_k h_k}{WH} \odot X \quad (4.2)$$

In AttHear, we employ NMF to decompose the input audio into meaningful components. Furthermore, our intuition for utilising NMF lies in integrating attention weights extracted from transformer models into the time activations during reconstruction in order to extract relevant audio components.

#### 4.2.2 AttHear Design

AttHear begins with decomposing the input audio into high-level components using NMF. Since common audio-based transformer models accept input in standard waveform representation, we convert the input audio into a spectrogram using the short-time Fourier transform (STFT). Subsequently, we extract the spectral pattern dictionary with its corresponding time activation matrix

as shown in Equation 4.1. To interpret NMF components in waveform representation, we initially reconstruct the corresponding audio spectrogram for a set of similar spectral patterns using Equation 4.2. Then, we transform it back into the original waveform space through an inverse STFT operation.

AttHear exploits the original transformer model in the second stage to extract its attention behaviour on audio components from the NMF stage. Each audio component is fed into the model to obtain its corresponding attention weights. Notably, the attention weights  $att_k$  and the time activations  $h_k$  for a spectral component are aligned to contain information about temporal significance. We combine them to create a unified vector with this knowledge. Furthermore, if class-specific explanations are required based on the audio-processing application, we incorporate attention gradients  $grad_k^c$  for each output class. We formulate the final temporal relevance  $r_k$  for a spectral component in Equation 4.3.

$$\begin{aligned} r_k &= h_k \odot att_k \odot \{grad_k^c\} \\ s.t. att_k &\in [0, 1], grad_k^c \in [-1, 1] \end{aligned} \quad (4.3)$$

In Equation 4.3, we use the normalised version of  $att_k$  to keep time activations in the same scale. Additionally, we apply interpolation to align the sizes of  $att_k$  and  $grad_k^c$  with that of  $h_k$  depending on the transformer model parameters.

### 4.2.3 Explanation Generation

To produce listenable explanations for an input audio, AttHear modifies the NMF reconstruction procedure in Equation 4.2, taking into consideration the temporal relevance  $r_k$  as defined in Equation 4.3. This attention-aware NMF reconstruction is defined in Equation 4.4. Using proposed novel reconstruction method, AttHear generates audio explanations that include relevant spectral components and their corresponding time activations, harnessing information from NMF time activations and attention weights from the original transformer model. Furthermore, it utilises the inverse STFT operation to convert these explanations into a raw waveform to make these explanations audible to the end-user.

$$X_k = \frac{w_k r_k}{WH} \odot X \quad (4.4)$$

## 4.3 Experiments

We evaluated our method on two datasets, Speech Commands V1-12 [Warden, 2018], and ESC-50 [Piczak, 2015], to assess its performance across different audio recognition tasks. This version of Speech Commands dataset consists of over 60 000 utterances of 10 different English words and two additional unique labels for “Unknown Word” and “Silence”. It spans multiple speakers and variations in pronunciation, offering a diverse set of challenges for audio classification algorithms. On the other hand, ESC-50 comprises 2,000 environmental sound recordings, each belonging to one of 50 distinct classes, such as dog barking, ocean waves, and car horns. In this section, we provide implementation details of our method, followed by a comprehensive quantitative and qualitative evaluation. The implementation code and sample audio explanations is available on our project page <sup>1</sup>.

### 4.3.1 Implementation Details

We choose wav2vec 2.0 [Baeovski et al., 2020] as our classifier model due to its popularity and proven performance on various audio tasks. Wav2vec 2.0 comprises two key components: a feature extractor based on a CNN, and a transformer module that models long-range dependencies. While AttHear leverages the transformer module to extract temporal relevance, the CNN architecture allows it to linearly map this relevance to the input space.

We utilise 20 and 40 components,  $K$ , for implementing NMF on data points from Speech Commands and ESC-50 datasets, respectively, as the latter contains more extended and complex audios. We select the final layer to obtain attention weights from the transformer layer, as higher layers tend to capture more global and abstract information [Lin et al., 2022]. To ensure compatibility with the dimensions of the input audio, we compute the average attention scores across multiple heads and cross-embeddings. Our classifiers achieve an accuracy of 98.0% on the test set for Speech Commands and 76.4% on the final fold of the validation set for ESC-50.

### 4.3.2 Quantitative Evaluation

We conduct fidelity experiments to measure how well the prediction of the underlying model and the generated explanation agree. Since AttHear investigates feature importance beyond the space

<sup>1</sup><https://github.com/glam-imperial/AttHear>

of the original input features by integrating NMF decomposition, it is not possible to select important features in this space. Thus, we design a novel strategy that involves selecting the highest temporal relevances  $r_k$  by a ratio of  $\alpha$  and setting remaining relevances to 0. Following this, we reconstruct the corresponding audio segments using Equation 4.4. We also propose a baseline that uses the same reconstruction method but from randomly selected temporal relevances. Although there is a random selection, note that our proposed baseline is “stronger” than a standard baseline as it utilises the learnt spectral patterns in NMF. We compute the fidelity score as the fraction of samples where the predicted class for generated explanations aligns with the original model’s prediction. The results in Table 4.1 and 4.2 show that AttHear can generate high-fidelity explanations compared to our proposed baseline for both datasets. The classifier model demonstrates considerably improved performance when evaluated on the selected features, even with a small selection ratio of  $\alpha = 0.1$ , using AttHear, compared to the baseline. Moreover, selecting important features with higher ratios enables the model to converge to the original classification performance more rapidly. For instance, when using half of the features ( $\alpha = 0.5$ ), the classification accuracy becomes comparable to that of the original classifier, indicating the effectiveness of the feature selection process in preserving model performance.

Table 4.1: Fidelity and faithfulness results on the Speech Commands dataset, with mean and standard deviation over five runs.

| Ratio, $\alpha$ | Fidelity (%)   |                | Faithfulness (% drop) |                |
|-----------------|----------------|----------------|-----------------------|----------------|
|                 | AttHear        | Baseline       | AttHear               | Baseline       |
| $\alpha = 0.1$  | $30.1 \pm 2.8$ | $2.8 \pm 1.7$  | $23.2 \pm 3.5$        | $5.4 \pm 1.3$  |
| $\alpha = 0.3$  | $65.9 \pm 2.6$ | $19.0 \pm 2.5$ | $50.2 \pm 1.8$        | $13.2 \pm 2.5$ |
| $\alpha = 0.5$  | $85.7 \pm 1.2$ | $51.7 \pm 3.5$ | $68.4 \pm 2.9$        | $28.4 \pm 4.5$ |

Table 4.2: Fidelity and faithfulness results on the ESC-50 dataset, with mean and standard deviation over five runs.

| Ratio, $\alpha$ | Fidelity       |                | Faithfulness   |                |
|-----------------|----------------|----------------|----------------|----------------|
|                 | AttHear        | Baseline       | AttHear        | Baseline       |
| $\alpha = 0.1$  | $54.2 \pm 2.6$ | $17.5 \pm 2.0$ | $14.6 \pm 3.1$ | $8.2 \pm 3.4$  |
| $\alpha = 0.3$  | $70.5 \pm 2.2$ | $26.4 \pm 2.4$ | $48.2 \pm 2.2$ | $22.1 \pm 3.9$ |
| $\alpha = 0.5$  | $78.1 \pm 3.7$ | $33.2 \pm 2.9$ | $61.7 \pm 1.6$ | $27.4 \pm 3.0$ |

In addition, we evaluate the faithfulness of AttHear to demonstrate that the generated explanations are relevant to the model’s final decision. Since NMF reconstruction is additive, we simulate faithfulness experiments by obtaining the remaining audio from subtracting a generated

explanation from the original audio. We followed the same steps with fidelity experiments to produce explanations. Faithfulness is evaluated by measuring the accuracy drop for different  $\alpha$  values compared to our proposed baseline, which is shown in Table 4.1 and Table 4.2. In this setup,  $\alpha$  represents the ratio of removed important features, where a higher  $\alpha$  is expected to result in a greater accuracy drop. The classifier’s performance declines considerably when evaluated on the remaining features, even with a small removal ratio of  $\alpha = 0.1$ , using AttHear, in comparison to the baseline. Furthermore, removing important features at higher ratios causes the model to approach random classification more rapidly. For instance, when half of the features are removed ( $\alpha = 0.5$ ), the classification accuracy drops to a level comparable to that of a random classifier, demonstrating the effectiveness of the feature selection process. Since the faithfulness metric proposed in [Parekh et al., 2022] does not align with our metric and is also dependent on the original model’s performance, which is not shared, we cannot present a fair comparison.

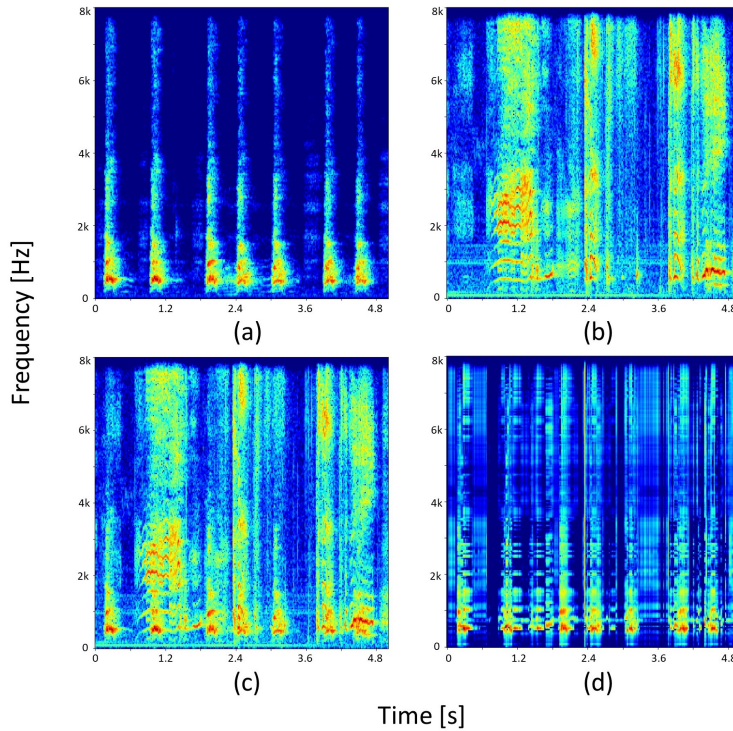


Figure 4.2: Sample spectrogram visualisations for the mixed audio experiments: (a) Original signal (“Dog”), (b) Mixing signal (“Crying-Baby”), (c) Mixed signal to be explained (d) Explanation audio generated for the target class (“Dog”). We notice that when generating the explanation, the majority of the content comes from the target class, with only slight corruption from the other class.

### 4.3.3 Qualitative Evaluation

To evaluate the quality of generated explanations, we listen to and observe our model’s behaviour on mixed audios from separate classes on the ESC-50 dataset. Ideally, we expect AttHear to capture the separate classes in these mixed audios and generate explanations based on the segments of the target class that we aim to explain. In Figure 4.2, we present an example of mixed audio, including the classes of “Crying-Baby” and “Dog.” We observe that the generation on target class “Dog” primarily consists of parts from the target class with minor corruption from the other class.

# 5

## Audio Explanation Synthesis with Generative Foundation Models

### Contents

---

|                               |           |
|-------------------------------|-----------|
| <b>5.1 Introduction</b>       | <b>64</b> |
| <b>5.2 Methodology</b>        | <b>65</b> |
| 5.2.1 Audio Foundation Models | 66        |
| 5.2.2 System Design           | 67        |
| 5.2.3 Explanation Generation  | 67        |
| <b>5.3 Experiments</b>        | <b>68</b> |
| 5.3.1 Implementation Details  | 68        |
| 5.3.2 Quantitative Evaluation | 70        |
| 5.3.3 Qualitative Evaluation  | 72        |

---

In this chapter, we introduce a novel method for explaining audio foundation models by leveraging their generative capabilities. Our approach harnesses the intrinsic representational power of the embedding space within these models, integrating established feature attribution techniques to identify key features that contribute to their decisions. By doing so, we provide deeper insights into how these models process and interpret audio data. The content of this chapter is based on our published work [Akman et al., 2025b].

## 5.1 Introduction

Generating explanations for large AI models has been gaining importance as they are used in various domains such as audio processing and computer vision. Most existing XAI methods try to extract important features in the input space towards the model’s final decision, that can be categorised into perturbation-based [Ribeiro et al., 2016, Petsiuk et al., 2018, Zeiler and Fergus, 2014, Fong et al., 2019] and backpropagation-based [Bach et al., 2015, Sundararajan et al., 2017, Shrikumar et al., 2017, Selvaraju et al., 2017, Montavon et al., 2017] techniques [Ancona et al., 2017]. These methods aim to identify relevant input features, such as pixels for computer vision tasks and tokens for natural language processing tasks. On the other hand, providing audio explanations is a useful method due to its intuitiveness on audio-based tasks and higher expressiveness over other modalities in specific scenarios, such as where understanding visual explanations needs expertise [Schuller et al., 2021b]. Aiming to generate listenable and interpretable audio explanations, [Wullenweber et al., 2022, Parekh et al., 2022] exploit NMF [Lee and Seung, 2000] to decompose audio into meaningful components.

Foundation models are extensively used in audio processing to achieve state-of-the-art performance on various tasks such as automatic speech recognition, keyword spotting, and speaker recognition [Gong et al., 2021, Hsu et al., 2021, Chen et al., 2021, Baevski et al., 2019, Chi et al., 2021, Ravanelli et al., 2020, Kharitonov et al., 2021]. In addition to that, these models offer a generalised and meaningful embedding space due to their broad range of training data; some foundation models such as EnCodec [Défossez et al., 2022] enable generation from this space. Although certain studies target to explain transformer-based foundation models by leveraging their attention mechanism and presenting attention weights as explanations [Vig, 2019, Yeh et al., 2024, Abnar and Zuidema, 2020], they do not consider computing feature importance in the meaningful embedding space to understand model behaviour. TCAV [Kim et al., 2018] and Network Dissection [Bau et al., 2017] focus on explaining a model’s behaviour with provided concepts by exploiting the model internal representation. However, these methods require user-defined concepts without considering unleashing the learnt concepts which are already embedded in the latent space of a foundation model.

To address these issues, we propose a method which combines prominent feature attribution methods with foundation models to explain model behaviour in audio processing tasks. We first exploit an audio foundation model as an encoder, and train an additional model on this backbone depending on the type of the downstream task. To understand the behaviour of the final model

on the task, we analyse important features for a decision in the latent space using a feature attribution method. In the final step, we use the generative part of the foundation model to construct the relevant audio in the input space. We verify that our method can generate high-fidelity explanations through experiments that simulate removing relevant features and assess the original model’s performance on these essential features. The main contributions of this study are as follows:

- We propose a novel audio explanation method that integrates common feature attribution methods into the latent space of foundation models. Our approach takes advantage of this meaningful space for creating understandable explanations without mapping feature relevance to the input space where individual features are difficult to interpret like audio frequencies.
- Our method leverages the generative capacity of foundation models from latent space to produce meaningful audio explanations in the input space. In this way, our method generates audio explanations in a listenable format that are interpretable to the end-user.
- We evaluate our model on keyword spotting and speech emotion recognition tasks. We show that while our method provides high-fidelity explanations, it captures meaningful high-level audio components for the investigated tasks. We also propose a strong baseline exploiting NMF with random attention to validate the performance of AttHear quantitatively.

## 5.2 Methodology

This section elaborates on the design of our method to generate meaningful audio explanations. We begin by introducing the general structure of audio foundation models and their usage for downstream tasks. Following this, we provide a detailed description of our explainer system, which involves assigning importance in the embedding space of a foundation model. Lastly, we outline the steps for producing meaningful audio explanations using our approach. We present an overview of our system in Figure 5.1.

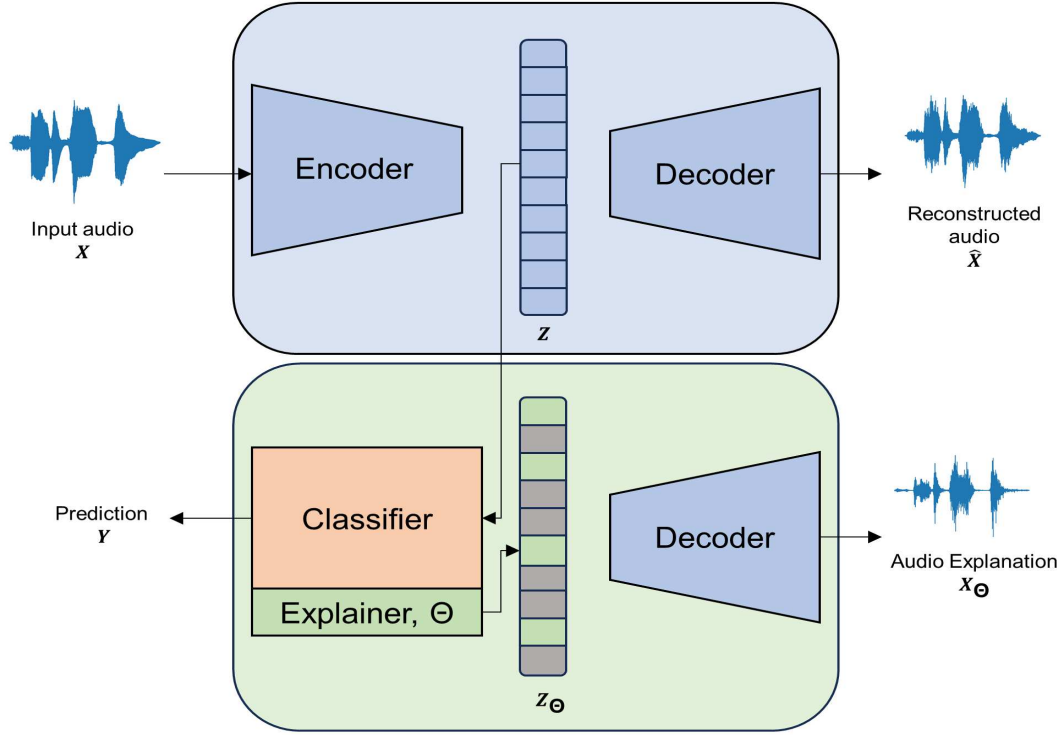


Figure 5.1: An overview of our method: The top row depicts the role of a foundation model with autoencoder architecture. The bottom row shows the process of explaining a task-specific classifier model including finding important features in the latent space and generating audio explanations based on these features.

### 5.2.1 Audio Foundation Models

Audio foundation models are typically pre-trained on large datasets of audio samples to learn patterns from the audio signals, which can then be fine-tuned on specific tasks. They exploit self-supervised learning strategies to discover general representations from large-scale data without requiring expensive labels. To learn these representations which reflect meaningful patterns in audio signals, they build a high-level embedding space using deep learning frameworks such as autoencoders and Siamese Network. While the former framework uses an encoder-decoder pair to project the audio input into the embedding space and then reconstruct it, the latter uses two encoders projecting into the same space where one latent representation is seen as a pseudo-label for the other. In this paper, we focus on audio foundation models using autoencoder architecture to be able to generate listenable explanations using its decoder component. The standard autoencoder architecture is defined in Equation 5.1.

$$Z = \text{Encoder}(X) \in \mathbb{R}^{T \times L} \quad ; \quad \hat{X} = \text{Decoder}(Z), \quad (5.1)$$

where  $X \in [-1, 1]^{D \times F_s}$  represents an audio signal input of duration  $D$  with sample rate  $F_s$ ,  $Z$  represents the latent vector with  $T$  denoting the number of audio frames after down-sampling in the encoder,  $L$  the feature dimension of the encoder, and  $\hat{X}$  represents the reconstructed audio signal.

### 5.2.2 System Design

We aim to understand the important audio features for a model decision by leveraging the high-level embedding space of foundation models. For this purpose, we target explaining foundation model-based audio models trained on specific audio tasks such as audio classification. Thus, our system starts with finetuning a foundation model with an autoencoder style framework on a desired task. To maintain the learnt representation space during finetuning, we freeze the weights of the encoder part and only update the additional task-specific model part such as a classification head. Then, our framework uses feature attribution methods to determine the most relevant features in the latent space for a model decision. Without backpropagating feature attribution computation to the audio input space, our method learns the important high-level components in the latent space, which is not restricted with the dimensions of the input space. In Equation 5.2, we formulate our feature attribution method in the latent space.

$$\text{att} = \Theta(\text{Classifier}(Z)) \in \mathbb{R}^{T \times L}, \quad (5.2)$$

where  $\Theta$  represents our explainer module that computes feature attribution,  $\text{att}$ , in the latent space. Note that the classifier module is designed to process the latent representation of an audio input extracted by the fixed encoder module.

### 5.2.3 Explanation Generation

To generate listenable audio explanations, our method first extracts the relevant latent vector based on the computed feature attributions in Equation 5.2. While keeping the latent dimensions with high importance by a ratio of  $\alpha \in [0, 1]$ , it replaces less important dimensions with a base latent vector which is obtained by encoding a noise audio with appropriate length. It then uses

the decoder part of the foundation model of interest to transform the relevant latent vector into audio explanations. The explanation generation is defined in Equation 5.3.

$$X_{\Theta} = \text{Decoder}(Z_{\Theta}), \quad (5.3)$$

where  $Z_{\Theta}$  represents the relevant latent vector for a specific prediction and  $X_{\Theta}$  represents the audio explanation in the input space. Although our method generates an audio explanation in the input space, it goes beyond only selecting features in this space by the integration of meaningful latent space.

## 5.3 Experiments

We evaluated our method on two datasets, namely, Speech Commands [Warden, 2018], and the Toronto Emotional Speech Set (TESS) [Pichora-Fuller and Dupuis, 2020], to assess its performance across keyword spotting and speech emotion recognition tasks. The Speech Commands dataset consists of over 100 000 utterances of 35 different English words by 2 618 speakers recorded. It spans multiple speakers and variations in pronunciation, offering a diverse set of challenges for audio classification algorithms. On the other hand, TESS comprises 2 800 stimuli in total. In TESS, a set of 200 target words (\*) were spoken in the carrier phrase “Say the word \*” by two actresses, and recordings were made of the set portraying each of seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. In this section, we provide implementation details of our method, followed by a comprehensive quantitative and qualitative evaluation. The implementation code and sample audio explanations is available on our project page<sup>1</sup>.

### 5.3.1 Implementation Details

We choose the EnCodec neural audio codec [Défossez et al., 2022] as our foundation model to learn the audio representations from the raw waveform, which is trained across diverse domains including general audio, speech, and music. EnCodec comprises two key components: an encoder that extracts features based on a CNN, and a decoder module that reconstructs the same audio. While our method uses EnCodec’s encoder module to extract meaningful audio representations and assign importance based on the classification model, its decoder module enables mapping these

<sup>1</sup><https://github.com/glam-imperial/AudioXgen>

features back to the input space. In our experiments, we use the EnCodec version for 24 kHz audio at 1.5 kbps bandwidth.

The EnCodec model is trained end-to-end to minimise reconstruction loss across both the time and frequency domains while incorporating a perceptual loss through multi-resolution discriminators. The original EnCodec model utilises Residual Vector Quantisation (RVQ), as introduced by [Zeghidour et al., 2021a], to map encoder outputs onto the closest entries in a codebook. However, to improve classification accuracy with higher-dimensional representations, we remove the quantisation step. This modification preserves richer feature representations, leading to better performance in classification tasks. A visual representation of the EnCodec model is shown in Figure 5.2.

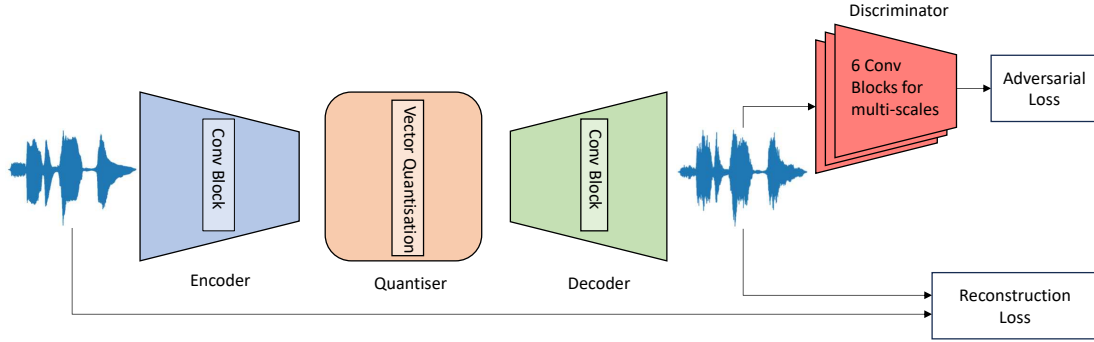


Figure 5.2: An overview of the Encodec model: an encoder-decoder architecture trained with both reconstruction and adversarial losses.

As our classification model, we train a transformer-based classifier on top of the embeddings extracted by EnCodec. Note that we only train a base model without extra tuning. In the classifier architecture for keyword spotting on Speech Commands, we use 3 layers of transformer with an 8 head multi-head attention module. We employ a dropout probability of 0.1 and the dimension of the feed-forward network model is 512 for each transformer layer. For speech emotion recognition on TESS, we use a Gated Recurrent Unit (GRU) based recurrent neural network with 2 layers and 128 hidden dimensions and employ a dropout of 0.2. Our classifiers achieve an accuracy of 85.4% on the test set for Speech Commands and 96.4% on the arranged test set for TESS. To arrange the TESS test set, we select random emotions from each spoken word using 0.2 split ratio and share the test indices on our project page for reproducibility. To compute the feature attribution in the latent space, we use IG [Sundararajan et al., 2017] which calculate the integral of the gradients of the model’s output along the straight line path from the baseline to the input.

### 5.3.2 Quantitative Evaluation

We conduct fidelity experiments to measure how well the prediction of the underlying model and the generated explanation agree. Since our method investigates feature importance beyond the space of the original input features by integrating Encodec latent space, it is not possible to select important features in this space. Thus, our strategy involves selecting the latent dimensions with the highest relevance with respect to the IG algorithm by a ratio of  $\alpha$ . We set the remaining dimensions to a base value using a base latent vector which is obtained by encoding a noise audio with appropriate length. We compute the fidelity score as the fraction of samples where the predicted class for provided explanations in the latent space aligns with the classifier’s prediction. We compare our methodology with three approaches: (1) We propose a baseline that randomly selects latent dimensions to obtain explanation embedding; (2) We select the most important features in the input audio space by a ratio of  $\alpha$  using the IG method and generate the explanation embedding with the Encodec encoder – here, eliminating the quantisation part of Encodec enables us to backpropagate the IG gradient calculation through the encoder safely; (3) We also implement the random feature selection strategy in the input audio space and follow the same explanation embedding generation steps with (2). The results in Table 5.1 and 5.2 show that our method can generate explanations with higher fidelity compared to the two baselines (1) and (3) for both datasets. It also outperforms the standard IG method (2) which validates our approach on providing explanations leveraging a high-level embedding space.

Table 5.1: Fidelity results by measuring explanation classification agreement on the Speech Commands dataset, with mean and standard deviation over five runs.

| Ratio, $\alpha$ | Latent Space   |                | Input Space    |                |
|-----------------|----------------|----------------|----------------|----------------|
|                 | Our Method     | Random         | IG             | Random         |
| 0.1             | 79.1 $\pm$ 0.0 | 6.6 $\pm$ 1.1  | 18.6 $\pm$ 0.0 | 7.9 $\pm$ 0.5  |
| 0.2             | 87.6 $\pm$ 0.0 | 16.5 $\pm$ 1.8 | 25.4 $\pm$ 0.0 | 10.2 $\pm$ 0.7 |
| 0.4             | 91.6 $\pm$ 0.0 | 46.0 $\pm$ 1.9 | 27.3 $\pm$ 0.0 | 15.9 $\pm$ 0.8 |
| 0.6             | 93.9 $\pm$ 0.0 | 69.6 $\pm$ 1.6 | 28.4 $\pm$ 0.0 | 23.0 $\pm$ 1.1 |
| 0.8             | 97.1 $\pm$ 0.0 | 84.8 $\pm$ 1.3 | 31.2 $\pm$ 0.0 | 39.9 $\pm$ 1.9 |

Table 5.2: Fidelity results by measuring explanation classification agreement on the TESS dataset, with mean and standard deviation over five runs.

| Ratio, $\alpha$ | Latent Space   |                | Input Space    |                |
|-----------------|----------------|----------------|----------------|----------------|
|                 | Our Method     | Random         | IG             | Random         |
| 0.1             | $21.9 \pm 0.0$ | $15.5 \pm 0.0$ | $57.7 \pm 0.0$ | $52.8 \pm 0.9$ |
| 0.2             | $55.7 \pm 0.0$ | $16.0 \pm 0.9$ | $59.8 \pm 0.0$ | $54.2 \pm 0.4$ |
| 0.4             | $62.5 \pm 0.0$ | $28.3 \pm 1.5$ | $62.3 \pm 0.0$ | $55.0 \pm 0.3$ |
| 0.6             | $73.6 \pm 0.0$ | $38.3 \pm 0.7$ | $63.2 \pm 0.0$ | $55.4 \pm 0.8$ |
| 0.8             | $90.4 \pm 0.0$ | $74.1 \pm 1.4$ | $62.6 \pm 0.0$ | $60.6 \pm 0.7$ |

In addition, we evaluate the fidelity of our method by analysing accuracy drop upon important feature removal to demonstrate that the generated explanations are relevant to the model’s final decision. We measure the accuracy drop for different  $\beta$  values which represent the ratio of most important features to be removed. Then, we compare our method with the other methods by following the same steps to produce the explanation embeddings. As shown in Table 5.3 and 5.4, feature removal based on our method leads to highest accuracy drop outperforming the other methods.

Table 5.3: Fidelity results by measuring accuracy drop over explanation removal on the Speech Commands dataset, with mean and standard deviation over five runs.

| Ratio, $\beta$ | Latent Space   |                | Input Space    |                |
|----------------|----------------|----------------|----------------|----------------|
|                | Our Method     | Random         | IG             | Random         |
| 0.01           | $20.4 \pm 0.0$ | $85.0 \pm 0.4$ | $51.4 \pm 0.0$ | $76.3 \pm 1.1$ |
| 0.1            | $1.0 \pm 0.0$  | $83.2 \pm 1.1$ | $22.6 \pm 0.0$ | $56.2 \pm 0.7$ |
| 0.2            | $0.7 \pm 0.0$  | $78.8 \pm 2.3$ | $10.4 \pm 0.0$ | $37.7 \pm 0.6$ |
| 0.4            | $0.6 \pm 0.0$  | $66.4 \pm 1.1$ | $8.0 \pm 0.0$  | $20.3 \pm 1.1$ |
| 0.6            | $0.4 \pm 0.0$  | $43.8 \pm 1.8$ | $7.8 \pm 0.0$  | $12.5 \pm 0.4$ |
| 0.8            | $0.4 \pm 0.0$  | $18.9 \pm 2.3$ | $7.4 \pm 0.0$  | $7.6 \pm 0.8$  |

Table 5.4: Fidelity results by measuring accuracy drop over explanation removal on the TESS dataset, with mean and standard deviation over five runs.

| Ratio, $\beta$ | Latent Space   |                | Input Space    |                |
|----------------|----------------|----------------|----------------|----------------|
|                | Our Method     | Random         | IG             | Random         |
| 0.01           | 64.6 $\pm$ 0.0 | 96.3 $\pm$ 0.2 | 85.4 $\pm$ 0.0 | 94.9 $\pm$ 0.4 |
| 0.1            | 22.1 $\pm$ 0.0 | 90.6 $\pm$ 0.7 | 55.2 $\pm$ 0.0 | 70.3 $\pm$ 1.4 |
| 0.2            | 17.3 $\pm$ 0.0 | 74.7 $\pm$ 1.8 | 43.6 $\pm$ 0.0 | 59.6 $\pm$ 0.8 |
| 0.4            | 16.9 $\pm$ 0.0 | 38.9 $\pm$ 1.2 | 39.8 $\pm$ 0.0 | 54.5 $\pm$ 0.7 |
| 0.6            | 16.3 $\pm$ 0.0 | 28.2 $\pm$ 1.1 | 39.6 $\pm$ 0.0 | 53.4 $\pm$ 0.4 |
| 0.8            | 17.1 $\pm$ 0.0 | 15.4 $\pm$ 1.2 | 40.1 $\pm$ 0.0 | 55.2 $\pm$ 0.7 |

### 5.3.3 Qualitative Evaluation

To evaluate the quality of generated explanations, we observe our model’s behaviour on audios from the same spoken words and separate emotion classes on the TESS dataset. By following similar procedures with quantitative evaluation experiments, we select an audio sample and generate the audio explanation using our method by setting  $\alpha = 0.2$  in the first experiment. In the second experiment setting, we removed the explanation from the original audio sample in the embedding space with the same ratio. We then generate the irrelevant audio part by using our framework. Ideally, we expect that while the generated audio in the first experiment can represent the emotion in the original audio sample, the irrelevant audio in the second experiment represents a neutral state of the same word spoken. In Figure 5.3, we present an example of audio from the class “Happy” to conduct these experiments. We also present the audio of the same word from the class “Neutral” to enable comparison for explanation-removed audio. We observe that while the generated explanation can represent the emotion in the original audio, explanation-removed audio looks more similar to the neutral version of the spoken word. We use spectrogram representation to increase the interpretability of the visuals.

We also investigate the classifier model behaviour for the TESS dataset after explanation removal by a ratio of  $\beta = 0.1$  using our method. In Figure 5.4, we present the confusion matrix of the classifier for each class in TESS dataset. The results show that majority of the audios are classified as “Neutral” when explanation is removed by a small ratio which shows our explanation generation method focuses on emotions.

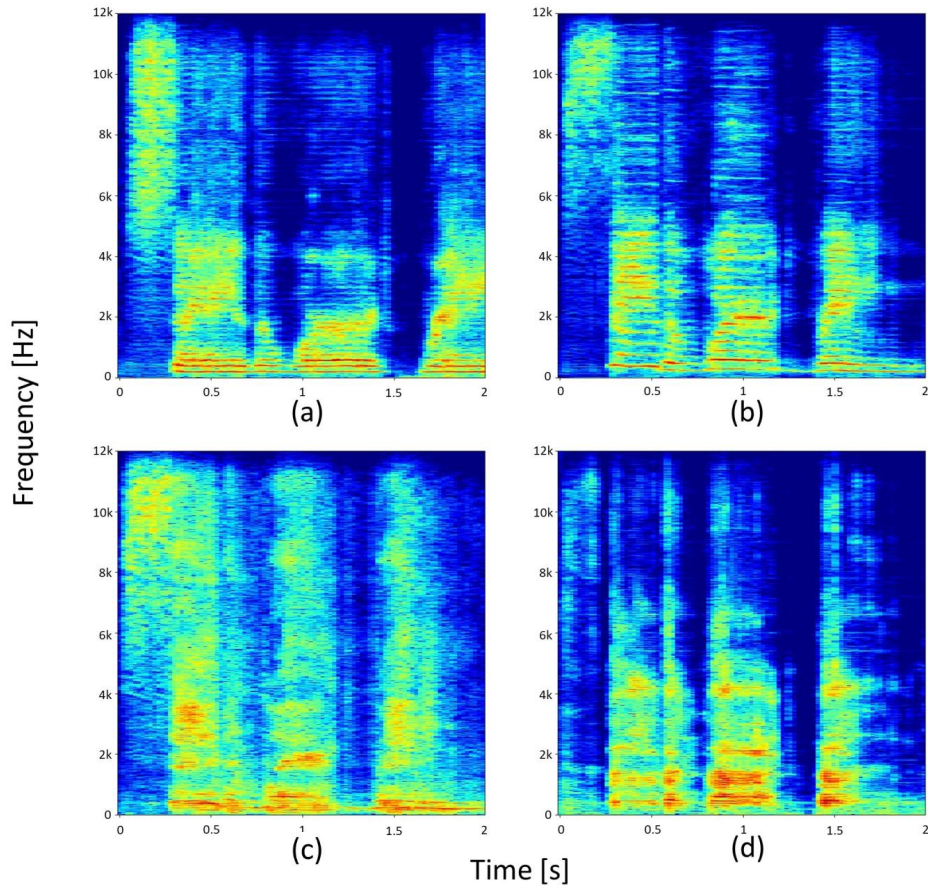


Figure 5.3: Sample spectrogram visualisations for the qualitative audio experiments: (a) Neutral audio of the word (“Rain”), (b) Happy audio of the word (“Rain”), (c) Explanation-removed audio from (b), (d) Explanation audio generated from (b).

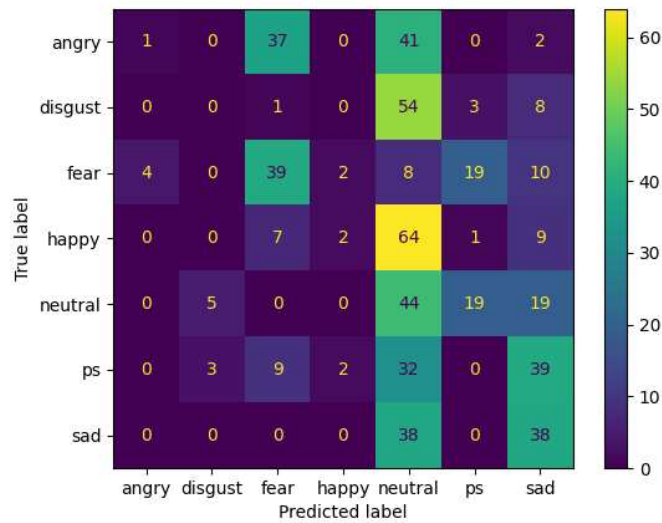


Figure 5.4: Confusion matrix for the classifier for TESS dataset after explanation removal by a ratio of  $\beta = 0.1$ .

# 6

## Improving Audio Explanations using Audio Language Models

### Contents

---

|                               |           |
|-------------------------------|-----------|
| <b>6.1 Introduction</b>       | <b>75</b> |
| <b>6.2 Methodology</b>        | <b>77</b> |
| 6.2.1 Audio Language Models   | 77        |
| 6.2.2 Explainer Design        | 78        |
| <b>6.3 Experiments</b>        | <b>79</b> |
| 6.3.1 Implementation Details  | 79        |
| 6.3.2 Quantitative Evaluation | 82        |
| 6.3.3 Qualitative Evaluation  | 84        |

---

In this chapter, we propose a novel framework for explanation generation that offers an alternative to traditional feature removal methods, enabling a comprehensive understanding of model behaviour. Our approach leverages the generative capabilities of audio language models to replace removed features with contextually appropriate alternatives, preserving the integrity of the input while providing deeper insights into the model’s decision-making process. This framework enhances interpretability by ensuring explanations reflect how the model processes information in a more holistic manner. The content of this chapter is based on our published work [Akman et al., 2025a].

## 6.1 Introduction

Foundation models has become prerequisite for designing deep models for various applications. By leveraging the principles of self-learning and large scale training data, they have established a new state-of-the-art across various tasks. While they show outstanding performance on automatic speech recognition [Baeviski et al., 2020, Radford et al., 2023, Gong et al., 2021, Hsu et al., 2021, Chen et al., 2021, Baeviski et al., 2019] and emotion recognition [Ma et al., 2024b, Wang et al., 2021, Chen et al., 2024] in audio domain, they are also primarily used for language tasks such as text classification [Touvron et al., 2023, OpenAI et al., 2024]. In addition to their predictive capabilities, foundation models, such as language models, also exhibit strong generative abilities, producing both text and sound under various conditions, including text-to-sound generation [Kreuk et al., 2023, Yang et al., 2023].

The growing popularity of foundation models, along with their complex and large-scale architectures, has made interpretability an essential concern. Most existing XAI methods focus on identifying important features in the input space that contribute to a model’s final decision. These feature attribution methods can be categorised into perturbation-based [Ribeiro et al., 2016, Pet-siuk et al., 2018] and backpropagation-based [Bach et al., 2015, Sundararajan et al., 2017, Selvaraju et al., 2017] techniques [Ancona et al., 2017]. They aim to identify relevant input features, such as tokens for natural language processing tasks time-frequency points in spectrograms for audio processing tasks. However, extracting high-level insights from these features is often challenging, particularly in tasks like audio processing, where interpreting spectrogram features typically requires domain expertise. In addition, focusing on low-level input features when generating explanations limits the ability to capture and convey a model’s high-level behaviour. In [Wullenweber et al., 2022, Parekh et al., 2022], the authors exploit NMF [Lee and Seung, 2000] to decompose audio into meaningful components, aiming to generate listenable and interpretable audio explanations. However, these methods primarily focus on computing relevance in the input space without delving into the intermediate representations within a deep model.

Certain methods target to improve interpretability by leveraging the intermediate representations of deep models. TCAV [Kim et al., 2018] and Network Dissection [Bau et al., 2017] aim to explain a model’s behaviour by utilising its internal representations with user-provided concepts. However, these approaches rely on predefined concepts and do not account for discovering the learned concepts already embedded within the latent space of a foundation model. In [Akman et al., 2025b], generative foundation models are used to produce explanations from relevant em-

bedding vectors computed by applying a feature attribution method in the latent space. However, due to the simplistic feature selection strategy, which replaces less relevant parts of an embedding vector with a base value, this approach has not explored the impact of less relevant dimensions of the latent space on explanation generation.

To address these challenges, we propose a framework that leverages the generative capabilities of ALMs, combined with feature attribution in the latent space of foundation models, to explain model behaviour in audio processing tasks. Our approach first utilises an audio foundation model as an encoder, and then trains an additional classification model on this backbone, tailored to the specific downstream task. To interpret the behaviour of the final model, we identify important features in the latent space using a prominent feature attribution method from the literature. Once the key features are selected, ALMs are employed to fill in the less relevant feature dimensions, generating a complete explanation that reflects the model’s decision-making process. In the final step, the decoder of the foundation model is used to reconstruct the relevant audio in the input space. We demonstrate that our method can generate high-fidelity explanations through experiments that assess the model’s performance on relevant features. Additionally, we evaluate the interpretability of the generated explanations by analysing speech recognition performance in relation to the speech emotion recognition task. The main contributions of this study are as follows:

- We propose a novel audio explanation framework that integrates ALMs with foundation model-based audio classifiers to provide a comprehensive understanding of important audio features. Our method leverages the high-level latent space of foundation models to identify meaningful audio features relevant to the model’s decisions.
- Our approach utilises the generative capabilities of ALMs to refine and complete the attributed features within the latent space of foundation models, generating listenable and interpretable audio explanations. This offers an alternative to traditional feature selection strategies in the XAI literature, enabling richer and more coherent explanations.
- We evaluate our framework on keyword spotting and speech emotion recognition tasks, demonstrating that, in addition to delivering high-fidelity explanations, our method enhances the understandability of these explanations. This is evidenced by improved speech recognition performance on emotion-focused explanations, showing their effectiveness in conveying the model’s decision-making process.

## 6.2 Methodology

This section details the design of our framework, which harnesses ALMs to generate comprehensive and meaningful audio explanations. We first outline the overall architecture of ALMs and their application in generative tasks. Next, we elaborate on our explainer system, which assigns importance to features within the embedding space of a foundation model-based classifier, while augmenting these features with ALMs. Finally, we describe the process for producing full audio explanations using our method. An overview of our system is illustrated in Figure 6.1.

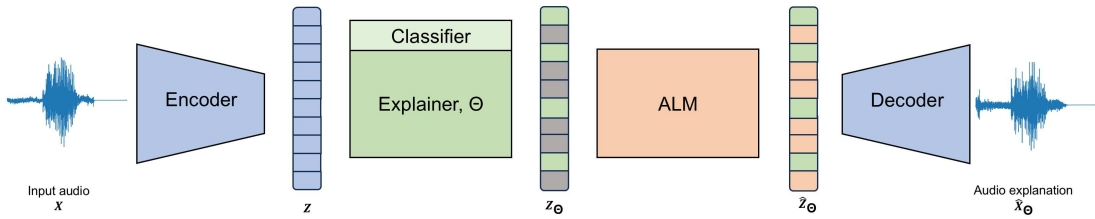


Figure 6.1: An overview of our method: The process starts by computing feature attribution for an input audio sample in the embedding space of a foundation model-based audio classifier. In the second stage, Audio Language Models (ALMs) are employed to augment the important features, generating a complete and coherent audio explanation.

### 6.2.1 Audio Language Models

Similar to large language models which are built on large textual datasets, ALMs try to utilise the expressive power of audio modality including emotions and speaker timbre [Wu et al., 2024]. They exploit neural audio codecs [Défossez et al., 2022, Zeghidour et al., 2021a, Borsos et al., 2023] as suitable tokenisers for converting continuous audio representation into discrete codes. Using the audio modality as their primary input type, certain studies also incorporate textual conditioning as a secondary input type for their ALMs [Kreuk et al., 2023, Yang et al., 2023, Du et al., 2024].

A typical ALM consists of three components in order to obtain the generative capacity. First, the tokeniser module, *Enc*, transforms a single channel audio signal into discrete audio codecs, as described in Equation 6.1. Next, the decoder-only transformer module predicts the subsequent audio token by training on a likelihood objective function, outlined in Equation 6.2. Finally, the detokeniser module acts as a lookup table, mapping the generated token sequence back into the audio waveform, as shown in Equation 6.3. The overall process of ALMs can be formulated as follows:

$$Z_u = Enc(X) \in \mathbb{R}^T \quad (6.1)$$

$$Objective = \prod_{t=1}^T p(Z_{u_t} | Z_{u_{<t}}, [Z_{v_{<t}}]) \quad (6.2)$$

$$\hat{X} = Dec(\hat{Z}_u), \quad (6.3)$$

where  $X \in [-1, 1]^{D \times F_s}$  represents an audio signal input of duration  $D$  with sample rate  $F_s$ ,  $Z_u$  represents the audio token sequence with  $T$  denoting the number of audio frames after down-sampling in the tokeniser, and  $Z_v$  represents the text tokens that is optionally used for conditional generation from text input.

### 6.2.2 Explainer Design

We aim to provide complete and listenable audio explanations using ALMs based on the extracted important features by standard feature attribution methods. To achieve this, we focus on explaining audio models that utilise the same tokeniser backbone as ALMs and are trained on specific audio tasks, such as audio classification. Our system begins by training an audio classifier on top of the EnCodec neural audio codec [Défossez et al., 2022], which serves as our tokeniser. To preserve the learned representation space during fine-tuning, we freeze the weights of the tokeniser and update only the additional task-specific components, such as the classification head. Subsequently, our framework employs feature attribution methods to identify the most relevant features in the latent space represented by audio codecs that contribute to model decisions. We backpropagate the feature attribution computations only to this latent space, allowing our method to learn the important high-level components. We formulate the process of feature attribution in the latent space in Equation 6.4, where  $\Theta$  represents a standard feature attribution method which produces the relevant subpart of the latent vector,  $Z_\theta$ , by selecting most important features by a ratio of  $\alpha \in [0, 1]$ . Note that a higher value of  $\alpha$  corresponds to selecting a larger proportion of latent features from the input for the explanation.

$$Z_\theta = \Theta(Classifier(Z), \alpha) \in \mathbb{R}^T, \quad (6.4)$$

By operating in the embedding space, our method aims to enhance important latent features with complementary features using ALMs, which can generate audio from this tokenised representation. To achieve this, we designate the remaining dimensions of the resulting latent vector,  $Z_\theta$ , as *unknown* tokens and feed it into the ALM to fill in the complementary features:  $\hat{Z}_\theta = ALM(Z_\theta)$ . Importantly, we do not utilise any text conditions in this process, even though the ALM model supports this option. This approach allows our method to prioritise high-level important features for the final prediction while providing a comprehensive understanding of the model’s behaviour through the use of an ALM with the same tokeniser backbone. Finally, we leverage the decoder component of the EnCodec model to transform the relevant latent vector into audio explanations. Our explanation generation framework to produce audio explanations in the input space,  $\hat{X}_\theta$ , is defined in Equation 6.5.

$$\hat{X}_\theta = Dec(\hat{Z}_\theta), \quad (6.5)$$

## 6.3 Experiments

We evaluated our method using two datasets: Speech Commands [Warden, 2018] and the Toronto Emotional Speech Set (TESS) [Pichora-Fuller and Dupuis, 2020], to measure its performance in keyword spotting and speech emotion recognition tasks. In this section, we detail the implementation of our approach and present a thorough quantitative and qualitative evaluation. The implementation code and sample audio explanations can be found on our project page <sup>1</sup>.

### 6.3.1 Implementation Details

We choose AudioGen [Kreuk et al., 2023] as our ALM to generate complementary audio features in the embedding space. AudioGen is trained on a diverse set of datasets that include various types of audio and text annotations. For our experiments, we use the second version of AudioGen<sup>2</sup>, which is trained on the same data but with longer audio samples, up to 10 seconds in duration, using a retrained EnCodec model on environmental sound data, following the standard EnCodec setup [Défossez et al., 2022]. This reimplementation of AudioGen is based on the ALM architecture from MusicGen [Copet et al., 2023] and employs a single-stage, auto-regressive Transformer. It uses a

<sup>1</sup><https://github.com/glam-imperial/AudioXLM>

<sup>2</sup><https://github.com/facebookresearch/audiocraft>

16 kHz EnCodec tokeniser with four codebooks sampled at 50 Hz, having an embedding space of dimension 128. This version maintains comparable audio quality to the original AudioGen while achieving faster generation speeds due to its lower frame rate. A visual representation of the AudioGen model is shown in Figure 6.2.

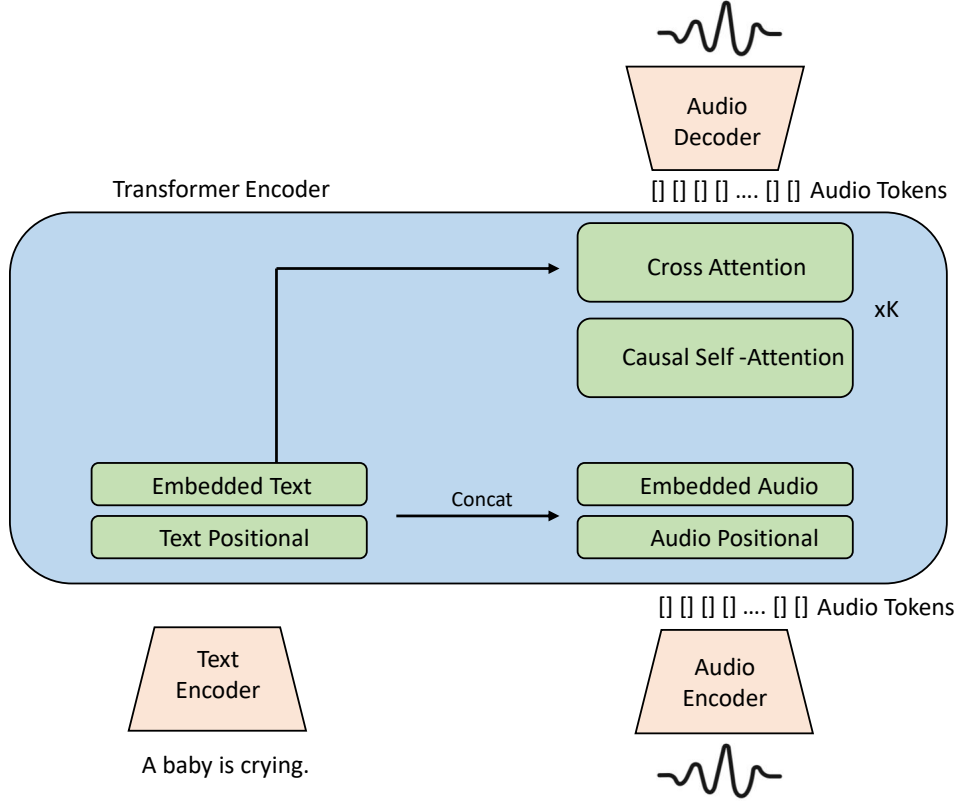


Figure 6.2: An overview of the AudioGen framework adapted from [Kreuk et al., 2023]: Both text and audio embeddings are concatenated along the time dimension and processed through K causal self-attention and cross-attention blocks, incorporating the embedded text.

Originally, AudioGen was designed for a variety of tasks, including text-to-sound generation, text-to-music generation, audio continuation, and multi-stream modeling, enabling it to generate audio content from textual input and model audio sequences in a multi-modal context. However, in our research, we adapted the AudioGen model for a novel task – audio completion. In this modified approach, we preserve the critical audio features within the latent space, which are extracted using our explainer pipeline, and complete the remaining latent dimensions that have not yet been predicted. Notably, we do not use any text prompts, as our input consists solely of explanation audio that needs to be completed. To implement this method, we introduced the use of an *unknown\_token* = -1, a default value embedded within the model’s architecture to represent positions in the latent space where audio content is incomplete. This token acts as a placeholder, signaling to the model where it needs to generate the missing information. By using

this approach, AudioGen is able to identify and generate complementary features for these missing latent dimensions, ultimately producing more comprehensive, coherent, and contextually relevant audio completions.

We train our classification models using the embeddings extracted from EnCodec, without additional hyperparameter tuning, and eliminate the quantisation step in EnCodec to improve classification accuracy by leveraging higher-dimensional representations. For each task, we train two models with different architectures to capture both average and strong classification performance. For the average classifier, used in keyword spotting on the Speech Commands dataset and speech emotion recognition on TESS, we employ a CNN with 4 convolutional layers and 2 linear layers. For the strong classifier on Speech Commands, we implement a transformer model with 3 layers, each containing an 8-head multi-head attention mechanism, a 0.1 dropout rate, and a 512-dimensional feed-forward network. On TESS, the strong classifier is a recurrent neural network based on GRU, with 2 layers, 128 hidden units, and a 0.2 dropout rate. While our models achieve test accuracy rates of 85.7%-89.7% on Speech Commands and 87.7%-99.6% on TESS, these results may not reach state-of-the-art performance because our method is constrained to use the same encoder as the ALM. For TESS, we create a test set by randomly selecting emotions for each spoken word, with a 0.2 split ratio, and we provide the test indices for reproducibility. To compute feature attribution in the latent space, we apply IG [Sundararajan et al., 2017], chosen for its theoretical rigor and suitability for neural networks, though our framework is compatible with any feature attribution method.

We conducted our timing experiments on a single NVIDIA TITAN Xp GPU, measuring the explanation generation time for three methods: (1) a random baseline, (2) the approach proposed by [Akman et al., 2025b] using IG, and (3) our ALM method. For CNN-based classifiers on the SC-TESS dataset, explanation generation times were approximately 1.5–1.5 ms, 20–24 ms, and 3.6–7.5 s, respectively, exhibiting similar trends across both classifiers. For the transformer model on the SC dataset, the measured times were 1.4 ms, 29 ms, and 3.7 s, while for the GRU model on the TESS dataset, they were 1.4 ms, 34 ms, and 7.5 s. These results, summarised in Table 6.1, highlight the computational trade-offs associated with employing large language models for explanation generation.

Table 6.1: Explanation Generation Time for Different Methods

| Dataset | Model       | Random Baseline (ms) | [Akman et al., 2025b] (ms) | Our Method (s) |
|---------|-------------|----------------------|----------------------------|----------------|
| SC      | CNN         | 1.5                  | 20                         | 3.6            |
|         | Transformer | 1.4                  | 29                         | 3.7            |
| TESS    | CNN         | 1.5                  | 24                         | 7.5            |
|         | GRU         | 1.4                  | 34                         | 7.5            |

### 6.3.2 Quantitative Evaluation

We conduct fidelity experiments to evaluate how closely the predictions of the underlying model align with the generated explanations. Similar to [Akman et al., 2025b], our framework begins by assigning feature importance in the EnCodec latent space using the IG method. We then select the most relevant features from the latent dimensions based on a specified ratio,  $\alpha$ . The key contribution of our approach lies in the final step, where we aim to provide a more comprehensive explanation by generating complementary features for the remaining latent dimensions. To achieve this, we use AudioGen to generate these complementary features, thereby enhancing the interpretability of the initially selected important features related to classification. We measure fidelity as the fraction of instances where the predicted class for the generated explanations matches the classifier’s original prediction in the latent space. We validate our approach by comparing it to two approaches: (1) a random selection baseline, which randomly selects latent dimensions based on the  $\alpha$  ratio to create explanation embeddings, and (2) the approach from [Akman et al., 2025b], which selects the most important features using IG and zeros out the remaining dimensions. As shown in Tables 6.2 and 6.3, our method consistently outperforms both approaches, producing higher-fidelity explanations across both datasets. However, at low  $\alpha$  values, the performance improvement of our method is less pronounced, particularly for the simpler convolutional model. This may be expected, as ALMs tend to generate irrelevant tokens when given more flexibility under weaker constraints. Note that the fidelity scores indicate alignment with the original model, not accuracy, explaining cases like the stronger transformer model showing lower alignment with a random selection strategy.

Table 6.2: Fidelity results by measuring explanation classification agreement on the Speech Commands dataset, with mean and standard deviation over five runs.

|             | Ratio, $\alpha$ | Random         |                | Feature Attribution   |                |
|-------------|-----------------|----------------|----------------|-----------------------|----------------|
|             |                 | Random         | Random + ALM   | [Akman et al., 2025b] | Our Method     |
| ConvNet     | 0.1             | 24.6 $\pm$ 2.5 | 11.9 $\pm$ 1.4 | 65.8 $\pm$ 0.0        | 57.4 $\pm$ 1.4 |
|             | 0.2             | 46.5 $\pm$ 1.0 | 24.8 $\pm$ 0.8 | 68.2 $\pm$ 0.0        | 69.4 $\pm$ 0.8 |
|             | 0.4             | 72.3 $\pm$ 1.6 | 62.4 $\pm$ 2.0 | 68.8 $\pm$ 0.0        | 84.0 $\pm$ 0.6 |
|             | 0.6             | 83.4 $\pm$ 1.1 | 83.9 $\pm$ 0.5 | 69.0 $\pm$ 0.0        | 91.2 $\pm$ 0.7 |
|             | 0.8             | 89.8 $\pm$ 1.6 | 93.0 $\pm$ 1.1 | 73.8 $\pm$ 0.0        | 96.4 $\pm$ 0.3 |
|             | 0.9             | 93.9 $\pm$ 1.2 | 96.1 $\pm$ 0.6 | 82.2 $\pm$ 0.0        | 98.0 $\pm$ 0.1 |
| Transformer | 0.1             | 6.0 $\pm$ 1.3  | 12.8 $\pm$ 0.8 | 95.6 $\pm$ 0.0        | 85.8 $\pm$ 1.0 |
|             | 0.2             | 19.3 $\pm$ 2.1 | 27.3 $\pm$ 2.3 | 95.7 $\pm$ 0.0        | 91.8 $\pm$ 1.7 |
|             | 0.4             | 52.5 $\pm$ 1.1 | 61.6 $\pm$ 2.5 | 96.3 $\pm$ 0.0        | 96.6 $\pm$ 0.7 |
|             | 0.6             | 72.9 $\pm$ 1.2 | 82.5 $\pm$ 1.1 | 97.0 $\pm$ 0.0        | 98.4 $\pm$ 0.3 |
|             | 0.8             | 85.1 $\pm$ 1.2 | 91.0 $\pm$ 0.6 | 97.4 $\pm$ 0.0        | 99.7 $\pm$ 0.1 |
|             | 0.9             | 91.5 $\pm$ 0.8 | 94.5 $\pm$ 0.6 | 98.4 $\pm$ 0.0        | 99.8 $\pm$ 0.0 |

Table 6.3: Fidelity results by measuring explanation classification agreement on the TESS dataset, with mean and standard deviation over five runs.

|         | Ratio, $\alpha$ | Random         |                | Feature Attribution   |                 |
|---------|-----------------|----------------|----------------|-----------------------|-----------------|
|         |                 | Random         | Random + ALM   | [Akman et al., 2025b] | Our Method      |
| ConvNet | 0.1             | 61.1 $\pm$ 7.8 | 47.5 $\pm$ 1.6 | 87.3 $\pm$ 0.4        | 93.6 $\pm$ 0.5  |
|         | 0.2             | 84.4 $\pm$ 2.1 | 67.4 $\pm$ 1.7 | 90.7 $\pm$ 0.3        | 97.0 $\pm$ 0.6  |
|         | 0.4             | 93.6 $\pm$ 1.1 | 91.4 $\pm$ 1.1 | 91.0 $\pm$ 0.3        | 98.4 $\pm$ 0.3  |
|         | 0.6             | 96.7 $\pm$ 1.2 | 96.0 $\pm$ 0.4 | 91.3 $\pm$ 0.3        | 99.6 $\pm$ 0.1  |
|         | 0.8             | 98.3 $\pm$ 0.3 | 98.5 $\pm$ 0.4 | 94.4 $\pm$ 0.3        | 100.0 $\pm$ 0.0 |
|         | 0.9             | 98.8 $\pm$ 0.4 | 99.3 $\pm$ 0.3 | 97.4 $\pm$ 0.1        | 100.0 $\pm$ 0.0 |
| GRU     | 0.1             | 21.1 $\pm$ 1.6 | 35.0 $\pm$ 1.7 | 86.5 $\pm$ 0.2        | 85.1 $\pm$ 1.7  |
|         | 0.2             | 32.7 $\pm$ 2.9 | 51.6 $\pm$ 2.1 | 89.0 $\pm$ 0.2        | 96.1 $\pm$ 0.5  |
|         | 0.4             | 66.4 $\pm$ 3.9 | 88.9 $\pm$ 0.9 | 91.3 $\pm$ 0.4        | 98.9 $\pm$ 0.4  |
|         | 0.6             | 97.1 $\pm$ 0.8 | 98.4 $\pm$ 0.2 | 91.8 $\pm$ 0.2        | 99.9 $\pm$ 0.1  |
|         | 0.8             | 99.3 $\pm$ 0.1 | 99.5 $\pm$ 0.4 | 94.7 $\pm$ 0.3        | 100.0 $\pm$ 0.0 |
|         | 0.9             | 99.6 $\pm$ 0.2 | 99.6 $\pm$ 0.2 | 97.5 $\pm$ 0.2        | 100.0 $\pm$ 0.0 |

Additionally, we evaluate the audio quality of explanations generated by our framework using quantitative metrics from an automatic speech recognition model. Since the keyword spotting task is closely related to speech recognition, we focus our analysis on the speech emotion recognition model trained on the TESS dataset to avoid redundant experiments. Our hypothesis is that incorporating language context into emotion-related features improves interpretability compared to generating explanation audios solely from abstract emotional concepts. Thus, we measure the speech recognition performance of the explanations generated by our method and the two baseline methods (1) and (2), using word error rate (WER). The results in Table 6.4 show that our method produces more understandable explanations, with lower WER scores, demonstrating improved clarity and interpretability for end-users.

Table 6.4: Automatic speech recognition results by measuring WER on the TESS dataset, with mean and standard deviation over five runs.

|         | Ratio, $\alpha$ | Random          |                 | Feature Attribution   |                 |
|---------|-----------------|-----------------|-----------------|-----------------------|-----------------|
|         |                 | Random          | Random + ALM    | [Akman et al., 2025b] | Our Method      |
| ConvNet | 0.1             | 70.0 $\pm$ 0.8  | 41.0 $\pm$ 3.1  | 77.2 $\pm$ 1.0        | 42.9 $\pm$ 1.3  |
|         | 0.2             | 89.8 $\pm$ 1.3  | 60.7 $\pm$ 3.3  | 95.3 $\pm$ 0.8        | 64.0 $\pm$ 1.0  |
|         | 0.4             | 98.8 $\pm$ 0.3  | 97.1 $\pm$ 1.2  | 99.1 $\pm$ 0.4        | 99.1 $\pm$ 0.9  |
|         | 0.6             | 100.0 $\pm$ 0.0 | 116.9 $\pm$ 0.8 | 100.0 $\pm$ 0.0       | 116.7 $\pm$ 0.9 |
| GRU     | 0.1             | 70.0 $\pm$ 1.6  | 42.0 $\pm$ 2.3  | 68.1 $\pm$ 0.5        | 40.9 $\pm$ 0.5  |
|         | 0.2             | 89.8 $\pm$ 2.9  | 61.2 $\pm$ 2.6  | 93.2 $\pm$ 0.6        | 62.7 $\pm$ 1.0  |
|         | 0.4             | 98.8 $\pm$ 3.9  | 96.9 $\pm$ 0.4  | 99.5 $\pm$ 0.3        | 99.7 $\pm$ 1.0  |
|         | 0.6             | 100.0 $\pm$ 0.0 | 116.3 $\pm$ 0.8 | 100.0 $\pm$ 0.0       | 117.7 $\pm$ 1.2 |

### 6.3.3 Qualitative Evaluation

We evaluate the quality of the explanations generated by our method for the speech emotion recognition task using the TESS dataset. In this experiment, we set  $\alpha = 0.4$  to generate audio explanations. To evaluate the quality of the generated explanations, we compare them against both the original audio and the explanations generated using the method from [Akman et al., 2025b]. In addition, we generate audio explanations with our method using input text conditions to examine their effect on the emotion recognition task. Initially, we tested using the full script of the original audio sentence, e.g., “Say the word time” but observed no considerable changes in the output. Given the known limitations of the AudioGen model in generating realistic vocals, as

noted by its authors, we instead used a simpler text condition “Human speech” for the generation process. As shown in Figure 6.3, while our method produces clearer audio explanations compared to the noisier outputs from [Akman et al., 2025b], adding text conditioning helps improve the quality in certain cases.

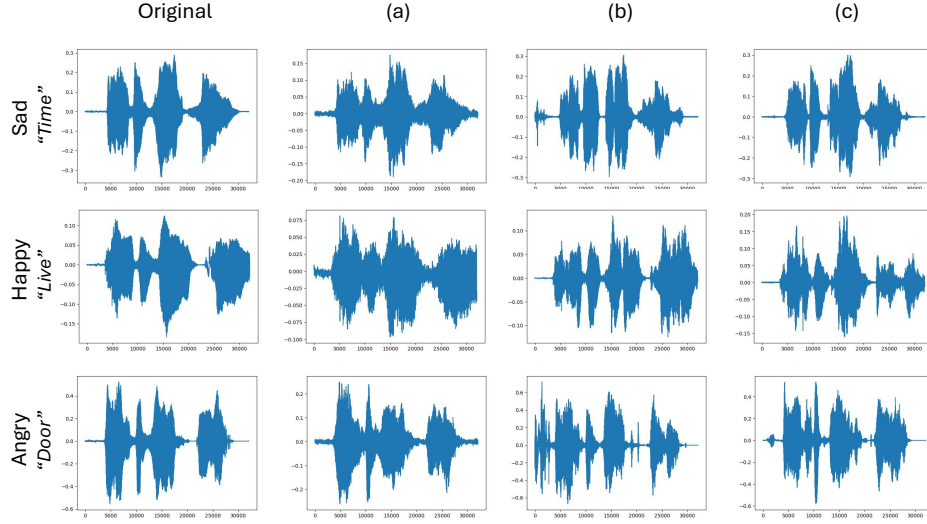


Figure 6.3: Sample audio waveform visualisations from the qualitative audio experiments. Each row represents different spoken words and corresponding emotions, while the columns compare various methods against the original audio input: (a) Method from [Akman et al., 2025b], (b) Our method, and (c) Our method conditioned with the audio script.

# 7

## Conclusions

### Contents

---

|                                  |           |
|----------------------------------|-----------|
| <b>7.1 Contributions Summary</b> | <b>86</b> |
| <b>7.2 Ethics</b>                | <b>89</b> |
| <b>7.3 Future Work</b>           | <b>93</b> |

---

### 7.1 Contributions Summary

Recent advancements in AI have transformed many fields, with audio processing models playing a key role in applications like speech recognition, music recommendations, and medical diagnostics. Unlike visual or textual data, audio has unique challenges, such as its continuous nature, temporal dependencies, and varied representations like waveforms and spectrograms. These complexities require specialised models to effectively analyse and interpret audio patterns, driving innovation across diverse applications.

Explainable AI methods are essential for enhancing the decision-making processes of AI systems, offering insights into how models reach their predictions. Utilising XAI is crucial not only for understanding a model's behaviour but also for debugging and improving its performance. In audio processing, explainability carries unique challenges and opportunities. Interpreting frequency-based features, such as those in spectrograms, often requires domain expertise, while time points

in raw waveforms may not directly correspond to meaningful events. Therefore, audio explanations should enhance understandability by leveraging auditory perception and presenting insights through significant and meaningful segments of the input audio. Additionally, these methods can include sonification, where non-audio data from other modalities is transformed into sound, enriching the interpretability experience.

In this thesis, we aimed to produce listenable and meaningful explanations for deep audio processing models by approaching the challenge from multiple perspectives. First, we explored data decomposition methods to create interpretable audio components, enabling the measurement of their importance to the model’s predictions. Next, we utilised built-in attention values in transformer models combined with NMF to identify and extract significant audio components. Additionally, we leveraged state-of-the-art audio foundation model embedding spaces to obtain high-level, meaningful features that enhance the interpretability of audio data. Finally, we incorporated audio-language models within these embedding spaces to further improve the quality and coherence of the generated audio explanations, offering a more complete understanding of model behaviour.

In **Chapter 3**, we explored the adaptability and interpretability of deep learning models for COVID-19 detection from audio data. Cross-running CIdER on the two 2021 Interspeech COVID-19 diagnosis challenges highlighted the model’s ability to operate across multiple modalities with minimal adjustments, achieving competitive results and underscoring the potential of end-to-end deep learning for audio processing. However, cross-dataset experiments also exposed the technical limitations of the datasets for joint usage, which hinder the creation of a unified COVID-19 classifier. To address interpretability, we introduced CoughLIME, a model-agnostic method for generating sonified and visual explanations for predictions. By decomposing input audio into interpretable components and fitting a surrogate model, CoughLIME faithfully identifies the most critical components influencing CIdER’s predictions. Quantitative evaluation using the novel  $\Delta$ AUC measure and pixel flipping for audio further verified its effectiveness, offering a valuable tool for enhancing the transparency of COVID-19 detection models.

In **Chapter 4**, we presented AttHear, which explains audio-processing transformer models. Unlike existing methods, AttHear integrates the attention module from transformers and NMF without finetuning the original prediction model or training a surrogate model. We introduced a novel formulation for NMF reconstruction, incorporating attention weights into NMF time activations. AttHear produces audio explanations in a user-friendly and audible format, enhancing end-user interpretability. The experiments demonstrate that AttHear delivers faithful explana-

tions with impressive fidelity scores, effectively capturing meaningful high-level audio components pertinent to the specific tasks.

In **Chapter 5**, we presented a novel audio explanation method which targets audio-processing foundation models. Unlike existing feature attribution methods which assign importance in the input space, our method integrates the latent space of a foundation model to generate meaningful explanations. In addition, we showed that our method leverages the generative capacity of foundation models to provide listenable audio explanations to the end-user. The experiments demonstrate that our method delivers high-fidelity explanations, effectively capturing meaningful audio components pertinent to the specific task.

In **Chapter 6**, we introduced a novel XAI method that enhances the quality of audio explanations using ALMs. Our approach not only generates complete explanations but also redefines existing feature attribution methods by modifying their feature selection strategy, incorporating generated features via embedding space supervision from ALMs. The experimental results demonstrated that our method produces high-fidelity explanations, accurately identifying important audio components and meaningfully completing them for audio classification tasks. Furthermore, we showed that our method leverages the generative capabilities of ALMs to create listenable and more interpretable audio explanations for end-users.

To revisit the primary research questions explored in this thesis, we first examined RQ-1, which asked about the best practices for explaining audio-processing models and how their explainability differs from AI models in other major domains such as computer vision and natural language processing. Through detailed analysis of both audio-processing models and the intrinsic properties of audio data, we identified several unique characteristics that distinguish audio from other modalities. These differences often limit the applicability of general XAI methods without adaptation, highlighting the need for tailored approaches in the audio domain.

Building on this, RQ-2 explored the role of listenable explanations in improving interpretability for end-users, as visual representations such as frequency interval visualisations, while informative for experts, are often unintuitive for users from diverse backgrounds. We demonstrated that listenable explanations leverage the inherent strengths of audio as a communication medium, providing a more intuitive and expressive way to convey model reasoning.

RQ-3 and RQ-4 addressed the identification and extraction of the “smallest meaningful units” in audio signals. Across multiple chapters, we investigated a range of decomposition techniques, from simple time–frequency masking to advanced matrix factorisation methods, to break down

audio into meaningful components. In doing so, we considered both spectrogram-based and raw waveform representations, assessing how these influence decomposition outcomes and the extraction of explainable features.

RQ-5 and RQ-6 focused on the potential of advanced linear algebra methods, such as NMF, and the use of transformer-based architectures for feature attribution. Although NMF can yield useful components for audio source separation, their interpretability is not always guaranteed. However, by combining NMF’s temporal components with transformer attention weights, we achieved high fidelity explanations, as validated by experimental results.

In addressing RQ-7, we evaluated whether the rich representation spaces learned by audio foundation models could serve as an alternative to manual decomposition methods. Our findings suggest that high-level embeddings from such models not only facilitate the extraction of meaningful features but can also be transformed into listenable explanations containing semantically rich audio concepts. Finally, in RQ-8, we investigated the use of audio language models to enhance explanations. Departing from conventional feature removal approaches, we utilised these models to replace removed features with contextually appropriate alternatives. This yielded more comprehensive and understandable explanations, improving both fidelity and end-user interpretability.

The findings across these research questions demonstrate that advancing explainability in audio-processing models requires a multifaceted approach that integrates tailored decomposition techniques, leverages the architectural strengths of modern AI models, and embraces the unique communicative potential of audio-based explanations.

## 7.2 Ethics

Ethical considerations in AI are essential to ensuring fairness, accountability, and public trust in its deployment. Key concerns include the responsible use of sensitive data, robust privacy protection, transparency in model decision-making, and bias mitigation to prevent discriminatory outcomes. AI systems must also be designed to avoid generating harmful, misleading, or unethical outputs, particularly in high-stakes domains such as healthcare, finance, and law enforcement. Furthermore, the potential misuse of AI, such as deepfakes spreading misinformation or biased algorithms reinforcing societal inequalities, poses significant risks. To address these challenges, AI research must adhere to rigorous ethical guidelines, prioritising fairness, safety, and compliance with legal and societal norms to ensure responsible and beneficial AI development.

In this thesis, we focus on the explainability of AI methods, positioning our work within the broader effort to make AI systems more transparent and accountable. XAI research aims to enhance various desirable properties, including transparency, trustworthiness, fairness, and understandability, as discussed in the background section. Our proposed XAI methods specifically target the explainability of audio processing neural networks by providing intuitive and listenable audio-based explanations. By making these models more interpretable, our approach not only improves user trust and comprehension but also helps identify potential model failures. These explanations can serve as valuable diagnostic tools, enabling researchers and practitioners to refine and improve models based on insights gained from the explanations themselves.

Ensuring fairness is a fundamental ethical consideration in XAI, particularly in applications like speech recognition, speaker identification, and medical audio analysis, where biased models can lead to unfair or discriminatory outcomes. Bias may arise from imbalanced training data or from the way explanations highlight model decision-making. To address this, our research prioritises the use of diverse and representative datasets to minimise inherent biases in both model predictions and explanations. Additionally, there is potential for the use of quantitative fairness metrics in future research to systematically assess and mitigate bias, ensuring that explanations do not reinforce or obscure unfair patterns in the data.

When used as assistive tools, XAI methods aim to support humans in making final decisions. However, their role in decision-making processes can become significant and, in some cases, irreversible, particularly in critical domains such as medical diagnosis and legal applications. In areas like medical audio analysis, human oversight remains essential to ensure that AI-generated explanations are correctly interpreted and appropriately acted upon. Therefore, it is crucial to provide clear and accurate explanations to guide human decision-making effectively in these critical decision-making tasks. Additionally, the way explanations are presented, including their modality, is an important factor in ensuring they convey the most relevant information. In this thesis, we consider audio-based explanations as a means of enhancing interpretability and aiding decision-makers in these critical applications.

While explainability methods aim to enhance trust in AI systems, they can also be exploited for adversarial purposes, such as model inversion attacks or identifying sensitive speaker characteristics. This research acknowledges these risks and promotes the responsible use of XAI techniques. Future work should explore safeguards against such vulnerabilities while maintaining interpretability.

In this research, we focus on audio datasets, where sensitive features such as gender, age, and racial information are often embedded in human audio signals. These features can unintentionally influence AI model predictions, potentially leading to biased outcomes. By employing XAI methods, we aim to illuminate on how these sensitive features are being utilised by the model, offering a deeper understanding of the model’s decision-making process. For example, explainability techniques can reveal whether certain features, like voice pitch or tone, are disproportionately contributing to the model’s predictions, which may correlate with gender or age. This insight can help identify and address biases within the model, ensuring that decisions are made based on relevant and fair criteria, rather than inadvertently reinforcing societal stereotypes or discriminatory practices.

For instance, in our exploration of COVID-19 prediction based on audio, we utilise healthcare data that includes audio recordings from participants. Given the sensitive nature of healthcare information, all datasets are anonymised to protect patient privacy and ensure confidentiality. The data collection process involves obtaining informed consent from participants, who are asked to record four audio clips after agreeing to a privacy statement that outlines the ethical considerations and data usage protocols. This process ensures that individuals are aware of how their data will be used and reassures them that their personal information is safeguarded. By anonymising the audio data and securing participants’ consent, we adhere to ethical standards that prioritise privacy while still leveraging the valuable insights that can be derived from audio-based models in healthcare applications.

In addition to the healthcare audio datasets, this thesis also makes use of several open-source audio datasets that are publicly available for research purposes. These datasets, which are commonly used in the AI community, contribute to the transparency and reproducibility of the research. All datasets incorporated in this work were obtained in compliance with legal and ethical standards. By using publicly accessible datasets or obtaining appropriate permissions, this work ensures that it adheres to established best practices for data governance while also promoting the open and transparent sharing of research resources within the AI community.

To promote transparency and reproducibility, we open-source the code for the experiments presented in this thesis. Making our code publicly available ensures that other researchers can validate our findings, conduct fair comparisons, and build upon our work. This openness not only strengthens the integrity of our research but also encourages collaboration within the AI community. By sharing our methods, we contribute to a more inclusive research environment where individuals from diverse backgrounds, regardless of institutional affiliation or geographic

location, can access, learn from, and contribute to the advancement of explainable AI.

Environmental sustainability is an increasingly important concern in AI, as the growing complexity of models leads to higher energy consumption. Large language models, in particular, have driven major companies to develop and continuously improve their computational hardware. While significant computing power is necessary for training and deploying these AI models, both the research and industry communities must prioritise energy efficiency in AI development. Excessive energy consumption contributes to environmental challenges such as climate change, and these resources could be allocated more effectively to areas of greater need. Given the finite nature of global energy resources, AI data and computation centres should operate within sustainable limits, and researchers should focus on optimising algorithms to reduce hardware demands. To support future considerations of sustainability, we report the type of GPU hardware used and the computation time required for our explanation method involving large language models.

The deployment of AI methods across various fields has a direct and significant impact on human life, making ethical considerations crucial. In high-stakes domains such as healthcare and autonomous driving, false negative cases can lead to severe consequences, potentially endangering lives. A key ethical concern is determining responsibility for incorrect or harmful decisions made by or influenced by AI. We argue that XAI should play a central role in addressing these challenges. Firstly, XAI methods can help clarify the technical sources of AI failures by identifying specific training data points or input features that led to inaccurate or unjustifiable outcomes. Secondly, XAI can assist in establishing appropriate trust in AI systems, ensuring that human decision-makers remain actively involved in critical judgments. By maintaining human oversight and accountability, XAI supports the ethical use of AI, safeguarding human values and well-being.

In addition to ethical considerations in this research, there are significant ongoing efforts to regulate AI due to its potential risks. Institutions like the UK, the European Union, and the US are working on frameworks to ensure AI is developed responsibly. For instance, the UK's AI Safety Institute (AISi) focuses on AI safety and accountability, while the European Commission has introduced the AI Act to regulate high-risk AI systems. The US is also developing AI standards through the National Institute of Standards and Technology (NIST). These initiatives aim to address key issues like bias, privacy, and transparency, helping to build public trust and safety. It is important for researchers to stay informed and contribute to these efforts, ensuring AI technologies benefit society while minimising harm.

## 7.3 Future Work

Our work enlightens the way to promising research in interpreting state-of-the-art audio models as well as encompassing their application for model debugging and justification. In this section, we highlight some major points for potential future work, exploring key areas where advancements could be made, while also acknowledging that numerous other opportunities for innovation remain underexplored.

In Chapter 3, we examined the interpretability of audio processing models within the medical domain, focusing on an application for predicting COVID-19 from audio data. Given the critical role these models play in healthcare scenarios, building trust among patients and supporting doctors in their diagnostic processes are essential objectives. Achieving this requires clear and actionable explanations of model behaviour that align with medical reasoning. To enhance trust and efficacy, it is crucial to gather feedback from medical experts, particularly regarding the specific audio features they consider important when making clinical decisions. Furthermore, to assess the practical utility of the generated explanations, conducting extensive qualitative usability evaluations is necessary. These evaluations should involve both medical professionals, who can provide insights into the clinical relevance of the explanations, and non-expert users, who can offer perspectives on the accessibility and interpretability of the information presented. This comprehensive approach ensures that the explanations meet technical standards while also effectively serving their intended purpose in real-world medical applications.

Human experimentation poses significant challenges in the evaluation of model interpretability, as it requires considerable time, effort, and resources to gather meaningful feedback from users. To address these limitations, automatic evaluation methods can serve as a valuable alternative by leveraging other models to assess interpretability across related tasks. For instance, in the context of speech emotion recognition, the quality and utility of explanations can be indirectly evaluated by examining their impact on the performance of ASR or speaker identification systems. By analysing how explanations derived from one task influence the outcomes of these auxiliary tasks, we can gain insights into their broader applicability and effectiveness. This automated approach not only provides a scalable solution for evaluation but also helps identify connections between interpretability and performance across various domains, thereby promoting a deeper understanding of how explanations can generalise to support multiple objectives.

In our work, we explored data decomposition techniques from the literature, such as NMF, which leverage the inherent mathematical properties of audio data to uncover meaningful structures. An exciting avenue for future research involves extending this approach to use audio explanations for building more reliable audio-processing models that align closely with human expertise. Achieving this requires defining high-level audio concepts that correspond to NMF-decomposed segments, providing task-specific explanations that are both interpretable and actionable. From a complementary perspective, incorporating supervision into the NMF decomposition process, where the components are guided to represent meaningful, predefined audio concepts, could further enhance the effectiveness of the explanations. By bridging data-driven decomposition with human-centric conceptualisation, this approach has the potential to create audio models that not only perform robustly but also offer insights that are accessible and valuable to human users across various applications.

Model-dependent XAI methods have inherent limitations, particularly when explanations rely on specific underlying architectures or foundation models. For example, in our work, we fine-tune specific audio foundation models to develop classifiers tailored to the tasks under investigation. While this approach is advantageous for aligning embedding spaces to uncover meaningful features relevant to the audio task, it restricts the generalisability of the explanations, as they remain tied to the specific foundation model. This dependency challenges the adoption of a model-agnostic framework that can operate across diverse architectures and classifiers. A promising direction for future research would involve modifying our method to provide explanations for any classifier, regardless of whether it is built on a specific foundation model. This shift would broaden the applicability of our approach, enabling insights into a wider range of audio models and enhancing its versatility across varied audio-processing tasks.

Our work opens exciting avenues for exploring the use of language models to interpret state-of-the-art audio models. By integrating audio-focused language models, we can enhance the quality of audio explanations, particularly by capturing and completing critical high-level features that define the audio task. This approach also enables the creation of counterfactual audio examples, offering a novel way to explore global explanations by showing how modifications to the audio input influence model decisions. Furthermore, our vision extends beyond merely identifying parts of the input audio as explanations. We proposed generating explanations as coherent, interpretable concepts, supported by creating entirely new audio examples that embody these concepts from scratch. This approach could provide more intuitive and accessible explanations, aligning technical outputs with human understanding and enabling deeper insights into model behaviour for future research.

Future research in model compression should consider the role of XAI in enhancing the efficiency and interpretability of compressed audio models. With the rapid advancements in deep learning architectures, particularly transformers, large-scale AI models trained on vast datasets have demonstrated state-of-the-art performance across diverse tasks with minimal fine-tuning. However, the deployment of such models remains challenging due to hardware constraints on edge devices, latency requirements in real-time applications, and the significant energy consumption associated with their operation, raising environmental concerns. Consequently, model compression techniques, such as the teacher-student transfer learning paradigm, have gained increasing attention, as they enable the development of smaller, more efficient models that, in some cases, even outperform their larger counterparts. While traditional model compression primarily focuses on aligning the performance of the student model with that of the teacher using accuracy-based metrics, an emerging research direction involves leveraging XAI to ensure that compressed models provide explanations similar to those of their larger counterparts. By aligning explainability across teacher and student models, future work can improve the trustworthiness and transparency of compressed models, which is particularly relevant for critical applications where model decisions require human interpretability.

Drawing from the definition of explainability, it is evident that the concept entails ensuring that the decisions of AI models are understandable to their intended audiences. The potential audience for XAI methods is highly diverse, ranging from children engaging with audio-based applications to medical experts working in healthcare. Consequently, XAI approaches should be tailored to the specific needs, cognitive abilities, and levels of expertise of the end-users they are designed to support. Developing specialised methods that align with the characteristics of different user groups represents a promising research direction, as it would enable more effective communication of AI decision-making processes and promote meaningful adoption. Future research could, therefore, focus on systematically identifying the requirements of distinct audiences and determining the forms of explanation that best facilitate their ability to utilise AI effectively.

The level of detail in explanations is another critical consideration in the development of XAI methods. Explanations frequently serve as complementary tools that enable users to take more informed actions in addition to interpreting the model's output. Both the degree of detail and the mode of presentation are central to their influence on the end-user's decision-making. Although an effective XAI method should provide accurate and transparent information regarding the model's reasoning, excessive detail may paradoxically reduce performance, for example by prompting experts to overtrust or unquestioningly accept all aspects of the model's output. Consequently,

---

maintaining human oversight and determining the most effective ways of integrating explanations into decision-making processes constitute an important avenue for future research.

A promising direction for future research is the sonification of model explanations across different modalities, extending beyond the audio processing models examined in this thesis. While visual and textual explanations are widely used to interpret deep learning models, audio-based explanations could serve as a complementary or even more effective alternative in certain contexts. For instance, in driver-assist systems, sonified alerts about potential obstacles could provide critical information without diverting the driver's visual attention from the road. More generally, the unique expressive capacity of audio signals offers an alternative means of conveying complex model insights, potentially improving interpretability and accessibility for diverse user groups. Future research could focus on designing effective sonified explanations for computer vision and natural language models, ensuring that they enhance understanding while minimising cognitive load.

# Bibliography

- [A. and R., 2023] A., S. and R., S. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7:100230.
- [Abnar and Zuidema, 2020] Abnar, S. and Zuidema, W. (2020). Quantifying attention flow in transformers. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.
- [Adadi and Berrada, 2018] Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.
- [Adebayo et al., 2018a] Adebayo, J., Gilmer, J., Goodfellow, I. J., and Kim, B. (2018a). Local explanation methods for deep neural networks lack sensitivity to parameter values. *CoRR*, abs/1810.03307.
- [Adebayo et al., 2018b] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018b). Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, page 9525–9536, Red Hook, NY, USA. Curran Associates Inc.
- [Akman et al., 2022] Akman, A., Coppock, H., Gaskell, A., Tzirakis, P., Jones, L., and Schuller, B. W. (2022). Evaluating the covid-19 identification resnet (cider) on the interspeech covid-19 from audio challenges. *Frontiers in Digital Health*, 4.
- [Akman and Schuller, 2024a] Akman, A. and Schuller, B. W. (2024a). Atthear: Explaining audio transformers using attention-aware nmf. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7015–7019.
- [Akman and Schuller, 2024b] Akman, A. and Schuller, B. W. (2024b). Audio explainable artificial intelligence: A review. *Intelligent Computing*, 3:0074.
- [Akman et al., 2025a] Akman, A., Sun, Q., and Schuller, B. (2025a). Improving audio explanations using audio language models. *IEEE Signal Processing Letters*, PP:1–5.

- 
- [Akman et al., 2025b] Akman, A., Sun, Q., and Schuller, B. W. (2025b). Audio explanation synthesis with generative foundation models. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [Ali et al., 2023] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805.
- [Alvarez-Melis and Jaakkola, 2018] Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *CoRR*, abs/1806.08049.
- [Amann et al., 2020] Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20.
- [Amiriparian et al., 2018] Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Pugachevskiy, S., and Schuller, B. (2018). Bag-of-deep-features: Noise-robust deep feature representations for audio analysis. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.
- [Amodei et al., 2015] Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J. H., Fan, L., Fougner, C., Han, T., Hannun, A. Y., Jun, B., LeGresley, P., Lin, L., Narang, S., Ng, A. Y., Ozair, S., Prenger, R., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, Y., Wang, Z., Wang, C., Xiao, B., Yogatama, D., Zhan, J., and Zhu, Z. (2015). Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR*, abs/1512.02595.
- [Ancona et al., 2017] Ancona, M., Ceolini, E., Öztireli, A. C., and Gross, M. H. (2017). A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, abs/1711.06104.
- [Ardila et al., 2019] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670.
- [Atakishiyev et al., 2021] Atakishiyev, S., Salameh, M., Yao, H., and Goebel, R. (2021). Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *CoRR*, abs/2112.11561.

- 
- [Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.
- [Baevski et al., 2022] Baevski, A., Hsu, W., Xu, Q., Babu, A., Gu, J., and Auli, M. (2022). data2vec: A general framework for self-supervised learning in speech, vision and language. *CoRR*, abs/2202.03555.
- [Baevski et al., 2019] Baevski, A., Schneider, S., and Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. *CoRR*, abs/1910.05453.
- [Baevski et al., 2020] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.
- [Bagad et al., 2020] Bagad, P., Dalmia, A., Doshi, J., Nagrani, A., Bhamare, P., Mahale, A., Rane, S., Agarwal, N., and Panicker, R. (2020). Cough Against COVID: Evidence of COVID-19 Signature in Cough Sounds. *arXiv*, (2009.08790). 12 pages.
- [Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.
- [Barredo Arrieta et al., 2020] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.
- [Bartl-Pokorny et al., 2020] Bartl-Pokorny, K. D., Pokorny, F. B., Batliner, A., Amiriparian, S., Semertzidou, A., Eyben, F., Kramer, E., Schmidt, F., Schönweiler, R., Wehler, M., and Schuller, B. W. (2020). The voice of COVID-19: Acoustic correlates of infection. 8 pages.
- [Bau et al., 2017] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations.
- [Becker et al., 2019] Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., and Samek, W. (2019). Interpreting and explaining deep neural networks for classification of audio signals.
- [Bellamy et al., 2019] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2019). Ai fairness 360:

- An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15.
- [Bertin-Mahieux et al., 2011] Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.
- [Blockeel et al., 2023] Blockeel, H., Devos, L., Frénay, B., Nanfack, G., and Nijssen, S. (2023). Decision trees: from efficient prediction to responsible ai. *Frontiers in Artificial Intelligence*, 6.
- [Borsos et al., 2023] Borsos, Z., Sharifi, M., Vincent, D., Kharitonov, E., Zeghidour, N., and Tagliasacchi, M. (2023). Soundstorm: Efficient parallel audio generation.
- [Brown et al., 2020a] Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., and Mascolo, C. (2020a). Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [Brown et al., 2020b] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020b). Language models are few-shot learners. *CoRR*, abs/2005.14165.
- [Burkart and Huber, 2020] Burkart, N. and Huber, M. F. (2020). A survey on the explainability of supervised machine learning. *CoRR*, abs/2011.07876.
- [Caron et al., 2021] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294.
- [Cartwright et al., 2020] Cartwright, M., Cramer, J., Méndez, A. E. M., Wang, Y., Wu, H., Lostanlen, V., Fuentes, M., Dove, G., Mydlarz, C., Salamon, J., Nov, O., and Bello, J. P. (2020). SONYC-UST-V2: an urban sound tagging dataset with spatiotemporal context. *CoRR*, abs/2009.05188.
- [Chalkidis et al., 2020] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In Cohn, T., He,

- Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.
- [Chao-Ying Joanne Peng and Ingersoll, 2002] Chao-Ying Joanne Peng, K. L. L. and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1):3–14.
- [Chefer et al., 2020] Chefer, H., Gur, S., and Wolf, L. (2020). Transformer interpretability beyond attention visualization. *CoRR*, abs/2012.09838.
- [Chen et al., 2021] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., and Wei, F. (2021). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *CoRR*, abs/2110.13900.
- [Chen et al., 2024] Chen, W., Xing, X., Chen, P., and Xu, X. (2024). Vesper: A compact and effective pretrained model for speech emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):1711–1724.
- [Chi et al., 2021] Chi, P.-H., Chung, P.-H., Wu, T.-H., Hsieh, C.-C., Chen, Y.-H., Li, S.-W., and Lee, H.-y. (2021). Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350.
- [Chockler et al., 2021] Chockler, H., Kroening, D., and Sun, Y. (2021). Compositional explanations for image classifiers. *CoRR*, abs/2103.03622.
- [Civit et al., 2022] Civit, M., Civit-Masot, J., Cuadrado, F., and Escalona, M. J. (2022). A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications*, 209:118190.
- [Clark et al., 2020] Clark, K., Luong, M., Le, Q. V., and Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.
- [Copet et al., 2023] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. (2023). Simple and controllable music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- [Coppock et al., 2021a] Coppock, H., Gaskell, A., Tzirakis, P., Baird, A., Jones, L., and Schuller, B. (2021a). End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: A pilot study. *BMJ Innovations*, 7(2).

- 
- [Coppock et al., 2021b] Coppock, H., Gaskell, A., Tzirakis, P., Baird, A., Jones, L., and Schuller, B. W. (2021b). End-2-end covid-19 detection from breath & cough audio. *BMJ Innovations*.
- [Coppock et al., 2021c] Coppock, H., Jones, L., Kiskin, I., and Schuller, B. W. (2021c). Covid-19 detection from audio: Seven grains of salt. *The Lancet Digital Health*.
- [Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.
- [Dawoud et al., 2023] Dawoud, K., Samek, W., Eisert, P., Lapuschkin, S., and Bosse, S. (2023). Human-centered evaluation of xai methods. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, page 912–921. IEEE.
- [Deck et al., 2024] Deck, L., Schoeffer, J., De-Arteaga, M., and Kühl, N. (2024). A critical survey on fairness benefits of explainable ai. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, page 1579–1595, New York, NY, USA. Association for Computing Machinery.
- [Deeks, 2019] Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7):1829–1850.
- [Deng, 2012] 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.
- [Derks and de Waal, 2021] Derks, I. P. and de Waal, A. (2021). A taxonomy of explainable bayesian networks. *CoRR*, abs/2101.11844.
- [Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- [Dhariwal et al., 2020] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). Jukebox: A generative model for music. *ArXiv*, abs/2005.00341.
- [Diwan et al., 2023] Diwan, A., Choi, E., and Harwath, D. (2023). When to use efficient self attention? profiling text, speech and image transformer variants. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1639–1650, Toronto, Canada. Association for Computational Linguistics.

- 
- [Donoso-Guzmán et al., 2023] Donoso-Guzmán, I., Ooge, J., Parra, D., and Verbert, K. (2023). Towards a comprehensive human-centred evaluation framework for explainable ai. In Longo, L., editor, *Explainable Artificial Intelligence*, pages 183–204, Cham. Springer Nature Switzerland.
- [Doshi-Velez and Kim, 2018] Doshi-Velez, F. and Kim, B. (2018). *Considerations for Evaluation and Generalization in Interpretable Machine Learning*, pages 3–17. Springer International Publishing, Cham.
- [Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.
- [Du et al., 2024] Du, Z., Wang, J., Chen, Q., Chu, Y., Gao, Z., Li, Z., Hu, K., Zhou, X., Xu, J., Ma, Z., Wang, W., Zheng, S., Zhou, C., Yan, Z., and Zhang, S. (2024). Lauragpt: Listen, attend, understand, and regenerate audio with gpt.
- [Défossez et al., 2022] Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. (2022). High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.
- [El Bekri et al., 2020] El Bekri, N., Kling, J., and Huber, M. F. (2020). A study on trust in black box models and post-hoc explanations. In Martínez Álvarez, F., Troncoso Lora, A., Sáez Muñoz, J. A., Quintián, H., and Corchado, E., editors, *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, pages 35–46, Cham. Springer International Publishing.
- [Eyben et al., 2013] Eyben, F., Wengner, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 835–838, New York, NY, USA. Association for Computing Machinery.
- [Eyben et al., 2010] Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.
- [Fong et al., 2019] Fong, R., Patrick, M., and Vedaldi, A. (2019). Understanding deep networks via extremal perturbations and smooth masks. *CoRR*, abs/1910.08485.

- 
- [Font et al., 2013] Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 411–412, New York, NY, USA. Association for Computing Machinery.
- [Freitag et al., 2018] Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., and Schuller, B. (2018). audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *Journal of Machine Learning Research*.
- [Frommholz et al., 2023] Frommholz, A., Seipel, F., Lapuschkin, S., Samek, W., and Vielhaben, J. (2023). Xai-based comparison of audio event classifiers with different input representations. In *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, CBMI '23, page 126–132, New York, NY, USA. Association for Computing Machinery.
- [Gao et al., 2021] Gao, T., Yao, X., and Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.
- [Gatys et al., 2016] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423.
- [Gentner, 1983] Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.
- [Gevrey et al., 2003] Gevrey, M., Dimopoulos, I., and Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, 160(3):249–264.
- [Ghorbani et al., 2019] Ghorbani, A., Abid, A., and Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688.
- [Gong et al., 2021] Gong, Y., Chung, Y., and Glass, J. R. (2021). AST: audio spectrogram transformer. *CoRR*, abs/2104.01778.
- [Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Book in preparation for MIT Press.
- [Goodfellow et al., 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

- 
- [Gourisaria et al., 2024] Gourisaria, M., Agrawal, R., Sahni, M., and Singh, P. (2024). Comparative analysis of audio classification with mfcc and stft features using machine learning techniques. *Discover Internet of Things*, 4.
- [Grechishnikova, 2019] Grechishnikova, D. (2019). Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific Reports*, 11.
- [Guidotti et al., 2018a] Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. (2018a). Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820.
- [Guidotti et al., 2018b] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018b). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5).
- [Halder et al., 2024] Halder, R., Uddin, M., Uddin, M. A., Aryal, S., and Khraisat, A. (2024). Enhancing k-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11.
- [Han et al., 2021] Han, J., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., and Mascolo, C. (2021). Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8328–8332.
- [Hanley and McNeil, 1982] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.
- [Hastie and Tibshirani, 1986] Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3):297 – 310.
- [Haunschmid et al., 2020a] Haunschmid, V., Manilow, E., and Widmer, G. (2020a). audiolime: Listenable explanations using source separation. *CoRR*, abs/2008.00582.
- [Haunschmid et al., 2020b] Haunschmid, V., Manilow, E., and Widmer, G. (2020b). audioLIME: Listenable Explanations Using Source Separation. 13th International Workshop on Machine Learning and Music.
- [Hayes et al., 2025] Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D.,

- Sercu, T., Candido, S., and Rives, A. (2025). Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858.
- [He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969.
- [Hennequin et al., 2020] Hennequin, R., Khelif, A., Voituret, F., and Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154.
- [Hershey et al., 2016] Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. W. (2016). CNN architectures for large-scale audio classification. *CoRR*, abs/1609.09430.
- [Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- [Holt, 2021] Holt, E. (2021). Newsdesk covid-19 testing in slovakia. *The Lancet Infectious Diseases*.
- [Holzinger et al., 2019] Holzinger, A., Carrington, A. M., and Müller, H. (2019). Measuring the quality of explanations: The system causability scale (SCS). comparing human and machine explanations. *CoRR*, abs/1912.09024.
- [Hsu et al., 2021] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhota, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- [Imran et al., 2020] Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, S., Ali, K., John, C. N., and Nabeel, M. (2020). AI4COVID-19: AI Enabled Preliminary Diagnosis for COVID-19 from Cough Samples via an App. *arXiv*, (2004.01275). 27 pages.
- [Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning*, volume 37, pages 448–456, Lille, France.

- 
- [Islam et al., 2021] Islam, S. M. M., Fioranelli, F., and Lubecke, V. M. (2021). Can radar remote life sensing technology help combat covid-19? *Frontiers in Communications and Networks*, 2.
- [Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.
- [Khan et al., 2022] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s).
- [Kharitonov et al., 2021] Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.-E., Douze, M., and Dupoux, E. (2021). Data augmenting contrastive learning of speech representations in the time domain. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 215–222.
- [Kheddar et al., 2024] Kheddar, H., Hemis, M., and Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, 109:102422.
- [Kim et al., 2016] Kim, B., Khanna, R., and Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, page 2288–2296, Red Hook, NY, USA. Curran Associates Inc.
- [Kim et al., 2018] Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Dy, J. G. and Krause, A., editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR.
- [Kim et al., 2024] Kim, J., Maathuis, H., and Sent, D. (2024). Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7.
- [Kindermans et al., 2019] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2019). *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham.
- [Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.
- [Koch et al., 2015] Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*.

- 
- [Koh and Liang, 2017] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 1885–1894. JMLR.org.
- [Kreuk et al., 2023] Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. (2023). Audiogen: Textually guided audio generation.
- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- [Laguarta et al., 2020] Laguarta, J., Hueto, F., and Subirana, B. (2020). COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, pages 1–1.
- [Lan et al., 2019] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.
- [Lee and Seung, 2000] Lee, D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.
- [Lin et al., 2022] Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open*, 3:111–132.
- [Lin et al., 2023] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.
- [Linardatos et al., 2021] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1).
- [Lipton, 2016] Lipton, Z. C. (2016). The mythos of model interpretability. *CoRR*, abs/1606.03490.
- [Litjens et al., 2017] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

- 
- [Liu et al., 2022] Liu, S., Mallol-Ragolta, A., Parada-Cabaleiro, E., Qian, K., Jing, X., Kathan, A., Hu, B., and Schuller, B. W. (2022). Audio self-supervised learning: A survey. *Patterns*, 3(12):100616.
- [Lopes et al., 2022] Lopes, P., Silva, E., Braga, C., Oliveira, T., and Rosado, L. (2022). Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19).
- [Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.
- [Ma et al., 2024a] Ma, J., Lai, V., Zhang, Y., Chen, C., Hamilton, P., Ljubenkov, D., Lakkaraju, H., and Tan, C. (2024a). Openhexai: An open-source framework for human-centered evaluation of explainable machine learning.
- [Ma et al., 2024b] Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., and Chen, X. (2024b). emotion2vec: Self-supervised pre-training for speech emotion representation. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.
- [Mahase, 2020] Mahase, E. (2020). Covid-19: Innova lateral flow test is not fit for test and release strategy, say experts. *BMJ*, 371.
- [Malhi et al., 2020] Malhi, A., Knapic, S., and Främling, K. (2020). Explainable agents for less bias in human-agent decision making. In Calvaresi, D., Najjar, A., Winikoff, M., and Främling, K., editors, *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 129–146, Cham. Springer International Publishing.
- [Mehri et al., 2016] Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A. C., and Bengio, Y. (2016). Samplernn: An unconditional end-to-end neural audio generation model. *CoRR*, abs/1612.07837.
- [Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.
- [Mohseni et al., 2021] Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4).
- [Montavon et al., 2017] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65(C):211–222.

- [Moyse, 2021] Moyse, K. A. (2021). Lateral flow tests need low false positives for antibodies and low false negatives for virus. *BMJ*, 372.
- [Muguli et al., 2021] Muguli, A., Pinto, L., R., N., Sharma, N., Krishnan, P., Ghosh, P. K., Kumar, R., Bhat, S., Chetupalli, S. R., Ganapathy, S., Ramoji, S., and Nanda, V. (2021). Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics.
- [Nessiem et al., 2021] Nessiem, M. A., Mohamed, M. M., Coppock, H., Gaskell, A., and Schuller, B. W. (2021). Detecting COVID-19 from Breathing and Coughing Sounds using Deep Neural Networks. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 183–188.
- [Nguyen et al., 2020] Nguyen, T. T., Le Nguyen, T., and Ifrim, G. (2020). A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In Lemaire, V., Malinowski, S., Bagnall, A., Guyet, T., Tavenard, R., and Ifrim, G., editors, *Advanced Analytics and Learning on Temporal Data*, pages 77–94, Cham. Springer International Publishing.
- [Olah et al., 2017] Olah, C., Schubert, L., and Mordvintsev, A. (2017). Feature visualization. *Distill*.
- [OpenAI et al., 2024] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich,

- A., Konstantinidis, A., Kotic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.
- [Orlandic et al., 2021] Orlandic, L., Teijeiro, T., and Atienza, D. (2021). The coughvid crowd-sourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1).
- [Parekh et al., 2022] Parekh, J., Parekh, S., Mozharovskiy, P., d’Alché Buc, F., and Richard, G. (2022). Listen to interpret: Post-hoc interpretability for audio networks with nmf.
- [Park et al., 2019] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2019-September.
- [Park and Jun, 2009] Park, H.-S. and Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Syst. Appl.*, 36(2):3336–3341.

- [Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.
- [Peto et al., 2020] Peto, J., Carpenter, J., Smith, G. D., Duffy, S., Houlston, R., Hunter, D. J., McPherson, K., Pearce, N., Romer, P., Sasieni, P., and Turnbull, C. (2020). Weekly covid-19 testing with household quarantine and contact tracing is feasible and would probably end the epidemic. *Royal Society Open Science*, 7.
- [Petsiuk et al., 2018] Petsiuk, V., Das, A., and Saenko, K. (2018). RISE: randomized input sampling for explanation of black-box models. *CoRR*, abs/1806.07421.
- [Pichora-Fuller and Dupuis, 2020] Pichora-Fuller, M. K. and Dupuis, K. (2020). Toronto emotional speech set (TESS).
- [Piczak, 2015] Piczak, K. J. (2015). Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA. Association for Computing Machinery.
- [Pinkas et al., 2020] Pinkas, G., Karny, Y., Malachi, A., Barkai, G., Bachar, G., and Aharonson, V. (2020). SARS-CoV-2 Detection From Voice. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:268–274.
- [Prabhavalkar et al., 2023] Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., and Watanabe, S. (2023). End-to-end speech recognition: A survey.
- [Pruthi et al., 2020] Pruthi, G., Liu, F., Sundararajan, M., and Kale, S. (2020). Estimating training data influence by tracking gradient descent. *CoRR*, abs/2002.08484.
- [Purwins et al., 2019] Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S., and Sainath, T. N. (2019). Deep learning for audio signal processing. *CoRR*, abs/1905.00078.
- [Qian et al., 2021] Qian, K., Schuller, B. W., and Yamamoto, Y. (2021). Recent advances in computer audition for diagnosing Covid-19: An overview. In *LifeTech 2021 - 2021 IEEE 3rd Global Conference on Life Sciences and Technologies*.
- [Quatieri et al., 2020] Quatieri, T., Talkar, T., and Palmer, J. (2020). A Framework for Biomarkers of COVID-19 Based on Coordination of Speech-Production Subsystems. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:203–206.
- [Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

- 
- [Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.
- [Radford et al., 2023] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- [Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.
- [Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.
- [Ravanelli et al., 2020] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993.
- [Reddy, 2022] Reddy, S. (2022). Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, 4(4):e214–e215.
- [Redmon and Farhadi, 2018] Redmon, J. and Farhadi, A. (2018). Yolo3: An incremental improvement. *CoRR*, abs/1804.02767.
- [Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- [Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*, page 1135–1144, New York, NY, USA. Association for Computing Machinery.
- [Ribeiro et al., 2018] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence*

- Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.
- [Ritwik et al., 2020] Ritwik, K. V. S., Kalluri, S. B., and Vijayasenan, D. (2020). COVID-19 Patient Detection from Telephone Quality Speech Data. *arXiv*, (2011.04299). 6 pages.
- [Rong et al., 2024] Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., and Kasneci, E. (2024). Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2104–2122.
- [Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241.
- [Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.
- [Salamon et al., 2014] Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 1041–1044, New York, NY, USA. Association for Computing Machinery.
- [Samek et al., 2017] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K. R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11).
- [Samek et al., 2021] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K. R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 109(3).
- [Schmitt and Schuller, 2017] Schmitt, M. and Schuller, B. (2017). openxbow – introducing the passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research*.
- [Schneider et al., 2019] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862.
- [Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

- [Schuller et al., 2018] Schuller, B., Steidl, S., Batliner, A., Marschik, P. B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F. B., Rathner, E.-M., Bartl-Pokorny, K. D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., and Zafeiriou, S. (2018). The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In *Interspeech*.
- [Schuller et al., 2021a] Schuller, B. W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., Rothkrantz, L. J. M., Zwerts, J., Treep, J., and Kaandorp, C. (2021a). The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates.
- [Schuller et al., 2021b] Schuller, B. W., Virtanen, T., Riveiro, M., Rizos, G., Han, J., Mesaros, A., and Drossos, K. (2021b). Towards sonification in multimodal and user-friendly explainable artificial intelligence. In *Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI '21*, page 788–792, New York, NY, USA. Association for Computing Machinery.
- [Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.
- [Sharma et al., 2020] Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., R., N., Ghosh, P. K., and Ganapathy, S. (2020). Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In *Proc. Interspeech*, pages 4811–4815, Shanghai, China.
- [Shen et al., 2017] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2017). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884.
- [Sherstinsky, 2020] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.
- [Shrikumar et al., 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, page 3145–3153. JMLR.org.
- [Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd Interna-*

- tional Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*
- [Socher et al., 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- [Speith, 2022] Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2239–2250, New York, NY, USA. Association for Computing Machinery.
- [Springenberg et al., 2015] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for simplicity: The all convolutional net.
- [Stepin et al., 2021] Stepin, I., Alonso, J. M., Catala, A., and Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001.
- [Stevens et al., 2020] Stevens, A., Deruyck, P., Veldhoven, Z. V., and Vanthienen, J. (2020). Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1241–1248.
- [Sun et al., 2025] Sun, Q., Akman, A., and Schuller, B. W. (2025). Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare*, 6(2).
- [Sun et al., 2019] Sun, Y., Chockler, H., Huang, X., and Kroening, D. (2019). Explaining deep neural networks using spectrum-based fault localization. *CoRR*, abs/1908.02374.
- [Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.
- [Taigman et al., 2014] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708.

- 
- [Taunk et al., 2019] Taunk, K., De, S., Verma, S., and Swetapadma, A. (2019). A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1255–1260.
- [Touvron et al., 2021] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). Training data-efficient image transformers & distillation through attention. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.
- [Touvron et al., 2023] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardaş, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.
- [Trigeorgis et al., 2016] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204.
- [Tripepi et al., 2008] Tripepi, G., Jager, K., Dekker, F., and Zoccali, C. (2008). Linear and logistic regression analysis. *Kidney International*, 73(7):806–810.
- [Urbanowicz and Moore, 2009] Urbanowicz, R. J. and Moore, J. H. (2009). Learning classifier systems: A complete introduction, review, and roadmap. *Journal of Artificial Evolution and Applications*, 2009(1):736398.
- [van den Oord et al., 2016] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V.,

- Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- [Vielhaben et al., 2024] Vielhaben, J., Lapuschkin, S., Montavon, G., and Samek, W. (2024). Explainable ai for time series via virtual inspection layers. *Pattern Recognition*, 150(C).
- [Vig, 2019] Vig, J. (2019). A multiscale visualization of attention in the transformer model. In Costa-jussà, M. R. and Alfonseca, E., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.
- [Vilone and Longo, 2020] Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. *CoRR*, abs/2006.00093.
- [Vilone and Longo, 2021a] Vilone, G. and Longo, L. (2021a). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3):615–661.
- [Vilone and Longo, 2021b] Vilone, G. and Longo, L. (2021b). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.
- [Vincent et al., 2010] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408.
- [Vinyals et al., 2015] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Wachter et al., 2017] Wachter, S., Mittelstadt, B. D., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399.
- [Wang et al., 2021] Wang, Y., Boumadane, A., and Heba, A. (2021). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *CoRR*, abs/2111.02735.
- [Warden, 2018] Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209.

- 
- [Weber et al., 2023] Weber, P., Carl, K. V., and Hinz, O. (2023). Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, pages 1–41.
- [Weiss and Indurkha, 1995] Weiss, S. M. and Indurkha, N. (1995). Rule-based machine learning methods for functional prediction. *J. Artif. Int. Res.*, 3(1):383–403.
- [Weitz et al., 2020] Weitz, K., Schiller, D., Schlagowski, R., Huber, T., and Andre, E. (2020). “let me explain!”: exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces*, 15.
- [Wiegrefe and Pinter, 2019] Wiegrefe, S. and Pinter, Y. (2019). Attention is not not explanation. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- [Wise, 2020] Wise, J. (2020). Covid-19: Concerns persist about purpose, ethics, and effect of rapid testing in liverpool. *BMJ*, 371.
- [Wood, 2017] Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. CRC Press, United States, 2 edition.
- [Wu et al., 2024] Wu, H., Chen, X., Lin, Y.-C., wei Chang, K., Chung, H.-L., Liu, A. H., and yi Lee, H. (2024). Towards audio language modeling – an overview.
- [Wu et al., 2023] Wu, X., Bell, P., and Rajan, A. (2023). Explanations for automatic speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- [Wullenweber et al., 2022] Wullenweber, A., Akman, A., and Schuller, B. W. (2022). Coughlime: Sonified explanations for the predictions of covid-19 cough classifiers. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1342–1345.
- [Yang et al., 2023] Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Wu, X., Zhao, Z., Watanabe, S., and Meng, H. (2023). Uniaudio: An audio foundation model toward universal audio generation.

- 
- [Yang et al., 2019] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA.
- [Yeh et al., 2024] Yeh, C., Chen, Y., Wu, A., Chen, C., Viégas, F., and Wattenberg, M. (2024). Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):262–272.
- [Yu and Deng, 2016] Yu, D. and Deng, L. (2016). *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 1st edition.
- [Zeghidour et al., 2021a] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2021a). Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.
- [Zeghidour et al., 2021b] Zeghidour, N., Teboul, O., de Chaumont Quitry, F., and Tagliasacchi, M. (2021b). LEAF: A learnable frontend for audio classification. *CoRR*, abs/2101.08596.
- [Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.
- [Zhang and Zhu, 2018] Zhang, Q. and Zhu, S. (2018). Visual interpretability for deep learning: a survey. *CoRR*, abs/1802.00614.
- [Zhang et al., 2020] Zhang, X., Gu, K., Miao, S., Zhang, X., Yin, Y., Wan, C., mei Yu, Y., Hu, J., Wang, Z., Shan, T., Jing, S., Wang, W., Ge, Y., Chen, Y., Guo, J., and Liu, Y. (2020). Automated detection of cardiovascular disease by electrocardiogram signal analysis: a deep learning system. *Cardiovascular diagnosis and therapy*, 10 2:227–235.

# 8

## Appendix

### 8.1 Codes

The code implementations for most of the methods proposed in this thesis are open-source and available on their respective GitHub project pages. In this section, I provide descriptions of the codebases for each part, originally prepared by me and, where applicable, my co-authors. These project pages not only include implementation code to ensure reproducibility but also feature listenable sample audio explanations generated using these methods.

#### 8.1.1 CoughLIME (Used in Chapter 3)

Project webpage: <https://github.com/glam-imperial/CoughLIME> Extends LIME: <https://github.com/marcotcr/lime>

CoughLIME generates explanations in listenable in visual form for the predictions of any COVID-19 classifier from audio data and is specifically tailored to cough data. CoughLIME decomposes the input into different interpretable components and generates explanations by determining the weights of each component towards the classifier’s prediction.

To generate the components, five decompositions are provided:

- Run `main.py` to generate sample explanations.
- **“temporal”**: Splits the audio array into equal segments along the temporal axis.

- **“Spectral”**: Splits the audio array into equal segments along the spectral axis.
- **“loudness”**: Splits the audio along the temporal axis based on minima in the power array.
- **“ls”**: Combines loudness and spectral decompositions.
- **“nmf”**: Uses non-negative matrix factorisation to split the audio array.

Explanations can be generated for the predictions of the DiCOVA baseline [Muguli et al., 2021] and CideR [Coppock et al., 2021a] models. The explanations can be evaluated using pixel flipping for audio and the Delta-AUC metric.

### 8.1.2 AttHear (Used in Chapter 4)

Project webpage: <https://github.com/glam-imperial/AttHear>

#### Contents:

- **audio\_explanation\_generation**: Inference script for generating audio explanations using AudioXgen.
- **fidelity\_performance**: Script to measure the fidelity score of explanations on the Speech Commands and TESS datasets.
- **models**: Folder containing classification models that predict on encoded datasets using EnCodec.
- **sample\_explanations**: Sample audio explanations generated by AudioXgen.

### 8.1.3 AudioXgen (Used in Chapter 5)

Project webpage: <https://github.com/glam-imperial/AudioXgen>

#### Contents:

- **audio\_explanation\_generation**: Inference script for generating audio explanations using AudioXgen.
- **fidelity\_performance**: Script to measure the fidelity score of explanations on the Speech Commands and TESS datasets.

- **models:** Folder containing classification models that predict on encoded datasets using EnCodec.
- **sample\_explanations:** Sample audio explanations generated by AudioXgen.

**EnCodec Installation:** Encodec has now been added to Transformers. For more information, please refer to [Transformers' EnCodec documentation](#).

You can find both the [24KHz](#) and [48KHz](#) checkpoints on the Hugging Face hub.

#### 8.1.4 AudioXLM (Used in Chapter 6)

Repository for Improving Audio Explanations using Audio Language Models

##### Contents

- **audio\_explanation\_generation:** Inference script for generating audio explanations using AudioXLM.
- **fidelity\_performance:** Script to measure the fidelity score of explanations on the Speech Commands and TESS datasets.
- **ASR\_WER\_performance\_TESS:** Script for evaluating automatic speech recognition performance of AudioXLM explanations in the speech emotion recognition task.
- **encode\_dataset\_AudioGen\_SC:** Script for encoding datasets into the embedding space of AudioGen for classifier models and AudioXLM, ensuring representation consistency.
- **AudioGen\_update:** Scripts for modifying the original AudioGen library. After installing AudioGen, copy these scripts into the appropriate [path](#) in the AudioGen library.
- **models:** Folder containing classification models that predict on encoded datasets.
- **sample\_explanations:** Sample audio explanations generated by AudioXLM.

##### AudioGen Installation Installation:

Follow the AudioGen installation instructions from the [AudioCraft repository](#). AudioCraft requires Python 3.9, PyTorch 2.1.0. To install AudioCraft, you can run the following:

```
# Best to make sure you have torch installed first, in particular before
installing xformers.

# Don't run this if you already have PyTorch installed.
python -m pip install 'torch==2.1.0'

# You might need the following before trying to install the packages
python -m pip install setuptools wheel

# Then proceed to one of the
# stable release
python -m pip install -U audiocraft

# bleeding edge
python -m pip install -U
git+https://git@github.com/facebookresearch/audiocraft#egg=audiocraft
# or if you cloned the repo locally (mandatory if you want to train).
python -m pip install -e .
# if you want to train a watermarking model
python -m pip install -e '.[wm]' # if you want to train a watermarking model
```

It is also recommended having `ffmpeg` installed, either through your system or Anaconda:

```
sudo apt-get install ffmpeg

# Or if you are using Anaconda or Miniconda
conda install "ffmpeg<5" -c conda-forge
```

## 8.2 Datasets

In this section, I provide details about the datasets used in this thesis to validate the proposed methodologies. The information is presented in a structure similar to that of the original papers, with appropriate citations, to accurately reflect the technical properties of the datasets. These details include anonymity, copyright, data distribution, demographic information, collection methodology, and data preprocessing.

### 8.2.1 DiCOVA Challenge Dataset

The DiCOVA Challenge datasets [Muguli et al., 2021] were developed to analyse respiratory sound recordings for COVID-19 detection. These datasets were derived from the Coswara [Sharma et al., 2020] dataset, a global crowd-sourced collection of audio samples gathered through an online platform. Launched in April 2020, this platform enabled volunteers to submit recordings of their respiratory sounds along with self-reported health and demographic details.

Each participant contributed a total of nine recordings, covering different types of respiratory and vocal sounds. These included coughs (both shallow and heavy), breathing sounds (both shallow and deep), sustained vowel phonations ( $\text{æ}$  as in bat,  $\text{[i]}$  as in beet, and  $\text{[u]}$  as in boot), and number counting (both at a normal and fast pace). Alongside the audio samples, participants provided information on their COVID-19 status, symptoms, and any pre-existing respiratory conditions.

For the challenge, two distinct datasets were created: Track-1 and Track-2. This thesis focuses on Track-1, which consists exclusively of cough sound recordings. Participants were categorised into two groups based on their self-reported health status. The COVID group included individuals who confirmed a positive COVID-19 diagnosis, regardless of whether they exhibited symptoms. The non-COVID group comprised individuals who either had no symptoms or reported symptoms associated with other respiratory conditions, such as asthma, pneumonia, or chronic lung disease, but were confirmed as COVID-19 negative. The Track-1 dataset consists of 1,040 recordings, including 965 from non-COVID subjects. A demographic breakdown of subjects by symptoms, age group, and gender is presented in Figure 8.1.

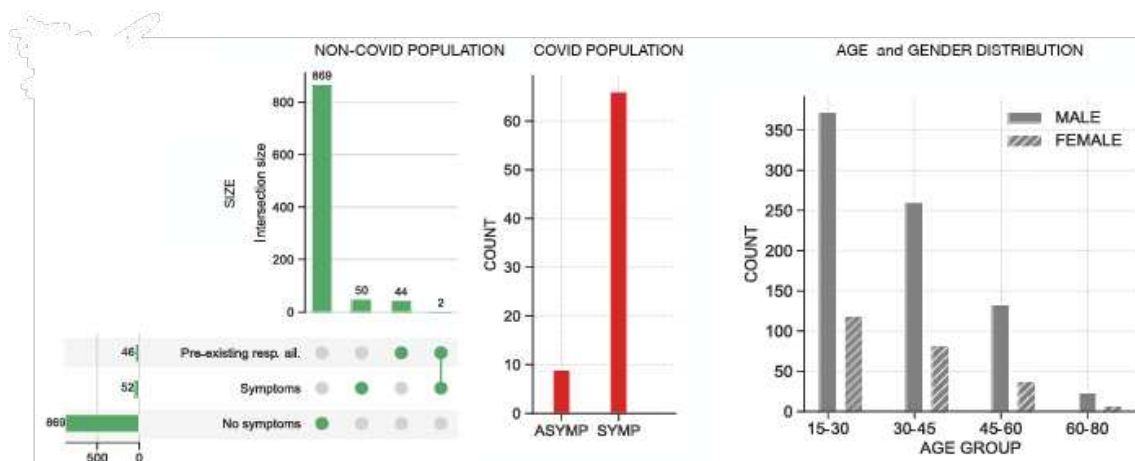


Figure 8.1: Distribution of subjects in the DiCOVA Challenge dataset based on symptoms, age, and gender. Adapted from [Muguli et al., 2021].

Since the Coswara dataset was collected via crowd-sourcing, the audio recordings exhibit vary-

ing quality levels, reflecting real-world data variability. While most recordings are relatively clean, informal listening confirms occasional background noise. Over 90% of the files were originally recorded at a sampling rate of 48 kHz and stored in WAV format. To standardise the challenge dataset, all recordings were resampled to 44.1 kHz and converted to FLAC format. The Track-1 dataset consists of 1,040 cough recordings, each from a unique subject, with an average duration of 472 seconds ( $\pm 7$  seconds standard error). This dataset serves as a valuable resource for analysing respiratory sounds in the context of COVID-19 detection.

### 8.2.2 Speech Commands

The Speech Commands dataset [Warden, 2018] is a standardised resource for training and evaluating keyword spotting models, enabling lightweight speech recognition with minimal false positives. Designed for detecting a small set of predefined words, it is particularly useful for voice-activated systems. Released under the Creative Commons BY 4.0 license [5], it is freely accessible for research and commercial use without registration or special permissions.

The dataset was designed with a limited yet diverse vocabulary to ensure a lightweight collection process while maintaining real-world applicability. Initially, twenty common words were selected, including digits (e.g., zero to nine) and command words relevant to IoT and robotics (e.g., “Yes,” “No,” “Up,” “Down”). In the second version, additional command words like “Backward,” and “Learn” were added. To improve keyword recognition, the dataset also includes non-command words as distractors, such as “Tree,” chosen for phonetic similarity to command words, and others like “Dog,” “Happy,” and “House” to ensure phonetic diversity. Overall, the dataset contains 105,829 utterances covering 35 distinct words, categorised and distributed as shown in Table 8.1.

To meet the project’s requirements, an open-source, web-based application was developed to record speech utterances. Users are prompted to grant microphone access, with the process of granting permissions being more acceptable than the alternative of installing native mobile apps, which users often resist due to privacy concerns. The application involves users consenting to participate in the study, after which they can record random words displayed for a short duration. Audio clips are stored locally on the user’s device before being uploaded to a server where they are saved in cloud storage. To ensure diverse speaker representation, a cookie prevents users from repeating the process after they’ve completed a recording session. The collected audio underwent quality control, with clips under 5 KB discarded as they were too short or silent. An “Extract Loudest Section” tool identified and removed quiet recordings, aligning the spoken word at the loudest

| Word     | Number of Utterances |
|----------|----------------------|
| Backward | 1,664                |
| Bed      | 2,014                |
| Bird     | 2,064                |
| Cat      | 2,031                |
| Dog      | 2,128                |
| Down     | 3,917                |
| Eight    | 3,787                |
| Five     | 4,052                |
| Follow   | 1,579                |
| Forward  | 1,557                |
| Four     | 3,728                |
| Go       | 3,880                |
| Happy    | 2,054                |
| House    | 2,113                |
| Learn    | 1,575                |
| Left     | 3,801                |
| Marvin   | 2,100                |
| Nine     | 3,934                |
| No       | 3,941                |
| Off      | 3,745                |
| On       | 3,845                |
| One      | 3,890                |
| Right    | 3,778                |
| Seven    | 3,998                |
| Sheila   | 2,022                |
| Six      | 3,860                |
| Stop     | 3,872                |
| Three    | 3,727                |
| Tree     | 1,759                |
| Two      | 3,880                |
| Up       | 3,723                |
| Visual   | 1,592                |
| Wow      | 2,123                |
| Yes      | 4,044                |
| Zero     | 4,052                |

Table 8.1: Word list and corresponding number of utterances.

part of the clip. Any remaining unclear or incorrect utterances were filtered through crowdsourcing, where workers transcribed the clips, and any mismatches were excluded. To anonymise the data, speaker IDs were hashed, removing any personally identifiable information. Additionally, background noise recordings were included, some captured from real environments and others synthetically generated, to aid in distinguishing speech from noise for keyword spotting.

### 8.2.3 Environmental Sound Classification (ESC-50)

The ESC-50 dataset [Piczak, 2015] is a collection of 2,000 labelled environmental sound recordings, equally balanced across 50 classes with 40 clips per class, organised into five major categories as

shown in Table 8.2: animal sounds, natural soundscapes and water sounds, human (non-speech) sounds, interior/domestic sounds, and exterior/urban noises. The recordings, sourced from publicly available contributions on the Freesound project [Font et al., 2013], were carefully selected to maintain balance across sound event types while considering the limitations of available source recordings. Sound samples were extracted by annotating relevant fragments from field recordings, and these samples were then converted to a standardised format of 44.1 kHz, single-channel, Ogg Vorbis compression at 192 kbit/s. The dataset is organised into five cross-validation folds, with each fold containing clips from a single original source file, and it is available under a Creative Commons non-commercial licence, along with an IPython notebook for further analysis and study replication.

While the dataset aims to provide varied exposure to both common and distinct environmental sounds, such as laughter, dog barking, and glass breaking, it has some limitations, including a relatively small number of clips per class due to the high cost of manual annotation and extraction. Despite these constraints, the dataset offers a valuable resource for environmental sound classification research and could be expanded upon in the future if there is sufficient interest from the academic community.

Table 8.2: Categorisation of sound classes in the ESC-50 dataset, organised into five major groups: animal sounds, natural soundscapes and water sounds, human (non-speech) sounds, interior/domestic sounds, and exterior/urban noises. Each category contains ten representative environmental sounds used for classification tasks.

| Animal sounds    | Natural soundscapes,<br>water sounds | Human (non-speech)<br>sounds | Interior/domestic<br>sounds | Exterior/urban<br>noises |
|------------------|--------------------------------------|------------------------------|-----------------------------|--------------------------|
| Dog              | Rain                                 | Crying baby                  | Door knock                  | Helicopter               |
| Rooster          | Sea waves                            | Sneezing                     | Mouse click                 | Chainsaw                 |
| Pig              | Crackling fire                       | Clapping                     | Keyboard typing             | Siren                    |
| Cow              | Crickets                             | Breathing                    | Door, wood creaks           | Car horn                 |
| Frog             | Chirping birds                       | Coughing                     | Can opening                 | Engine                   |
| Cat              | Water drops                          | Footsteps                    | Washing machine             | Train                    |
| Hen              | Wind                                 | Laughing                     | Vacuum cleaner              | Church bells             |
| Insects (flying) | Pouring water                        | Brushing teeth               | Clock alarm                 | Airplane                 |
| Sheep            | Toilet flush                         | Snoring                      | Clock tick                  | Fireworks                |
| Crow             | Thunderstorm                         | Drinking, sipping            | Glass breaking              | Hand saw                 |

#### 8.2.4 Toronto Emotional Speech Set (TESS)

The Toronto Emotional Speech Set (TESS) dataset [Pichora-Fuller and Dupuis, 2020] consists of 2800 stimuli, modeled on the Northwestern University Auditory Test No. 6. It includes 200 target words spoken in the carrier phrase “Say the word \*” by two actresses, aged 26 and 64 years, portraying seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The two actresses, both native English speakers from the Toronto area, are university-

---

educated and have musical training. Audiometric testing confirmed that their hearing thresholds are within the normal range. This dataset is available under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license, published by the University of Toronto, Psychology Department, in 2010.

The TESS dataset includes the following words: back, bar, base, bath, bean, beg, bite, boat, bone, book, bought, burn, cab, calm, came, cause, chain, chair, chalk, chat, check, cheek, chief, choice, cool, dab, date, dead, death, deep, dime, dip, ditch, dodge, dog, doll, door, fail, fall, far, fat, fit, five, food, gap, gas, gaze, germ, get, gin, goal, good, goose, gun, half, hall, hash, hate, have, haze, hire, hit, hole, home, hurl, hush, jail, jar, join, judge, jug, juice, keen, keep, keg, kick, kill, king, kite, knock, late, laud, lean, learn, lease, lid, life, limb, live, loaf, long, lore, lose, lot, love, luck, make, match, merge, mess, met, mill, mob, mode, mood, moon, mop, mouse, nag, name, near, neat, nice, note, numb, pad, page, pain, pass, pearl, peg, perch, phone, pick, pike, pole, pool, puff, rag, raid, rain, raise, rat, reach, read, red, ring, ripe, road, room, rose, rot, rough, rush, said, sail, search, seize, sell, shack, shall, shawl, sheep, shirt, should, shout, size, soap, soup, sour, south, sub, such, sure, take, talk, tape, team, tell, thin, third, thought, thumb, time, tip, tire, ton, tool, tough, turn, vine, voice, void, vote, wag, walk, wash, week, wheat, when, which, whip, white, wife, wire, witch, yearn, yes, young, youth.