Tuning In : Analysis of Audio Classifier Performance in Clinical Settings with Limited Data

Hamza Mahdi[*†1,2,3,4,5] hmahdi2026@meds.uwo.ca
Eptehal Nashnoush[*†1,2,3,4] e.nashnoush@mail.utoronto.ca
Rami Saab[1,2,3,4] rami.saab@utoronto.ca
Arjun Balachandar[1,2,3,4] arjun.balachandar@mail.utoronto.ca
Rishit Dagli[†1,2,6] rishit@cs.toronto.edu
Lucas X. Perri[1,2] lucas.perri@outlook.com
Houman Khosravani[†1,2,3,4] h.khosravani@utoronto.ca

[1] *Temerty Centre for Artificial Intelligence Research and Education in Medicine, University of Toronto, Canada*
[2] *Hurvitz Brain Sciences Program, Toronto, Canada*
[3] *Division of Neurology, Department of Medicine, University of Toronto, Canada*
[4] *Sunnybrook Research Institute, Toronto, Canada*
[5] *Western University, Canada*
[6] *Department of Computer Science, University of Toronto, Canada*

Abstract

This study assesses deep learning models for audio classification in a clinical setting with the constraint of small datasets reflecting the prospective collection of real-world data. We analyze CNNs, including DenseNet and ConvNeXt, alongside transformer models like ViT, and SWIN, and compare them against pretrained audio models such as AST, YAMNet and VGGish. Our method highlights the benefits of pretraining on large datasets before fine-tuning on specific clinical data. We prospectively collected two first-of-its-kind patient audio datasets from stroke patients. We investigated various preprocessing techniques, finding that RGB and grayscale spectrogram transformations affect model performance differently based on the priors they learn from pretraining. Our findings indicate CNNs can match or exceed transformer models in small dataset contexts, with DenseNet-Contrastive and AST models showing notable performance. This study highlights the significance of incremental marginal gains through model selection, pretraining, and preprocessing in sound classification; this offers valuable insights for clinical diagnostics that rely on audio classification.

Data and Code Availability In this research, in addition to patient data, we used publicly available datasets such as ImageNet (Deng et al., 2009), AudioSet (Gemmeke et al., 2017), US8K (Salamon et al., 2014) and ESC50 (Piczak), along with a proprietary clinical dataset. Due to confidentiality, patient privacy regulations, and local research ethics board (REB) constraints, the clinical portion of the dataset cannot be shared at this time. Our project code is available on Github [1]. We are working with our local data sharing hub at our university institute to make the clinical data available in the near future through an REB amendment and the framework the institute has in place for such purposes.

Institutional Review Board (IRB) This study involved human subjects and received approval from the Research Ethics Board of our local hospital and institution. Throughout, we ensured compliance with local institution guidelines. All individuals or SDMs who participated in the study provided their consent.

1. Introduction

Auditory biomarkers have been widely incorporated as the first line of evaluation in medical applications; non-speech and non-semantic sounds in particular have been used for decades to detect respi-

[*] These authors share first-authorship
[†] These authors have contributed equally to describing this work.

1. https://github.com/UofTNeurology

ratory problems (Pahar et al., 2021). Modern tools for collecting and analyzing audio data have revolutionized the diagnosis of common symptoms such as coughing, making voice analysis a critical first step in the diagnostic process (Larson et al., 2012; Tracey et al., 2011). Furthermore, clinical problems ranging from continuous health monitoring (Alqudaihi et al., 2021), stroke (Saab et al., 2023), psychiatric conditions (Moedomo et al., 2012; Fagherazzi et al., 2021), neurodegenerative diseases (Fagherazzi et al., 2021; Vizza et al., 2019) to cardiac applications (Dwivedi et al., 2018; Emmanuel, 2012) and lung pathology detection (Gavriely et al., 1994; Sello et al., 2008; Aykanat et al., 2017; García-Ordás et al., 2020) among others have adopted non-speech health acoustic data as an important biomarker. However, understanding the clinical effect of different preprocessing and modeling techniques for the audio domain is a long-standing challenge.

The domain of modeling auditory data borrows heavily from the advances made in sequential and vision models. Such auditory models have made great progress through the development of large-scale unsupervised pretraining for audio encoders (Baevski et al., 2020; Valk and Alumäe, 2021). This transposition is usually facilitated by the conversion of audio signals into log-mel spectrograms or superlets, which are then analyzed using algorithms originally designed for vision or sequential data (Radford et al., 2023; Choi et al., 2018). Such approaches for modeling auditory data leverage the significant progress made in the vision and sequence modalities.

Among audio transformation techniques, log-mel spectrograms and superlets (Moca et al., 2021) represent two leading methodologies. Log-mel spectrograms are widely recognized for their ability to approximate the human auditory system's response to sound. This method converts audio signals into a spectrogram using the Mel scale. This results in a compact, yet effective representation of sound, which highlights the elements most relevant to human auditory perception and have been widely used for acoustic problems (Böck and Schedl, 2011; Radford et al., 2023; Choi et al., 2018). Superlets (Moca et al., 2021), include a set of wavelets that are iteratively applied across different cycles with a specific central frequency, potentially capturing more nuanced information within complex audio signals. We were particularly interested in the comparative efficacy of these methods for downstream clinical applications in neurology and beyond, which remains a subject of ongoing research.

Audio data is increasingly used as a biomarker in clinical settings for disease classification and assessment, and our aim was to expand the range of possibilities for both analysis and characterization of changes using different modeling techniques and at different stages of processing. Our analysis focuses on clinical data, in a neurologic setting but with broader applications, and on comparing different model approaches. Using first-of-its-kind prospectively collected data sets, *Dataset NIHSS* and *Dataset Vowel* aim to expand the existing area of research in the context of disease state classification when starting with limited real-world data. This opens the way for use in other clinical settings (neurologic and beyond), where audio data can be used as a disease biomarker, and in settings where limited data are available. This also includes clinical settings and datasets where data is limited to not only collection limitations but also rare diseases where data scarcity is an intrinsic factor.

Key Contributions. Acoustic-based clinical diagnosis (or prognosis) has gained popularity in medical applications, leading the way to consider audio data as a biomarker in disease classification, risk prediction, and monitoring. However, the impact of modeling decisions on these medical tasks remains largely unexplored, with one variable being limited datasets. There are also implications for rare diseases that intrinsically have this limitation. In this work, we focused on stroke as a neurovascular disease process and utilized speech as a biomarker and surrogate for swallowing difficulty (dysphagia). We evaluated training health acoustic models with different preprocessing techniques—mel RGB, log-mel mono, and superlet—and clinical data representations and assessed classification based on a defined clinical outcome state of dysphagia.

- Introduced *Dataset NIHSS*: A novel data set that captures continuous speech, sentences, and words based on the National Institutes of Health Stroke Scale (NIHSS) (Kwah and Diong, 2014), an internationally established neurologic assessment scale for stroke emergencies.
- Introduced *Dataset Vowel*: A unique data set of sustained vowel sounds from patients, which further aids in the analysis of swallowing disorders.
- Analyzed Model Training Impact: Evaluated how training health acoustic models with dif-

ferent preprocessing techniques, mel RGB, log-mel mono, and superlet, affects the representation and classification of clinical data based on a defined clinical outcome state of dysphagia.

2. Related Work

Categorizing acoustic data is a problem that has been well explored throughout the years, and many deep learning-based methods have recently performed very well in classifying acoustic data, which has led to exploration of application in clinical settings. Often directly analyzing raw audio leads to improper learned representations, and acoustic data needs to be preprocessed first. We provide a brief overview of acoustic classification and acoustic event detection approaches. We are particularly interested in analyzing the downstream clinical effects of such approaches. We also provide a brief overview of the analysis done previously in these areas.

2.1. Audio Classification

Audio event detection and classification historically relied on simple representations of the underlying audio to transform the audio based on dynamic time warping (DTW), which allowed algorithms trained on these representations to measure spectral variability Sakoe and Chiba (1978); Salvador and Chan (2007). Following this, Hidden Markov Models (HMM) gained popularity for discrete speech and soon became the dominant technique for all audio-based applications, outperforming early neural approaches (Juang, 1984; Raphael, 1999). Following this, the audio classification was mainly performed using features such as Mel-frequency cepstrum coefficients (MFCC) and classifiers based on Gaussian Mixture Models (GMM) (Nilsson et al., 2002), Hidden Markov Models (HMM) (Juang, 1984; Raphael, 1999), Nonnegative matrix factorization (NMF) (Holzapfel and Stylianou, 2008; Ozerov and Févotte, 2009) or support vector machine (SVM) (Dhanalakshmi et al., 2009). Soon these models transitioned to a discriminative training strategy (Hermansky et al., 2000), which led to weighted finite-state transducers (WFSTs) becoming increasingly common and Restricted Boltzmann Machines (RBMs) becoming the first popular neural component of acoustic models. Following this, Seide et al. (2011) led to neural architectures becoming the dominant approach for modeling audio after demonstrating 30% RER on the Switchboard benchmark.

Modern neural-network-based models for acoustic tasks have demonstrated significant performance increases over previous approaches with ConvNets (Hershey et al., 2017; Schmid et al., 2023; Gong et al., 2021b), RNNs (Phan et al., 2017; Gimeno et al., 2020), Transformers (Koutini et al., 2021; Jaegle et al., 2021; Chen et al., 2022) often trained with self-supervised learning (Gong et al., 2022; Georgescu et al., 2023; Huang et al., 2022), and for some tasks diffusion (Kong et al., 2020; Lee and Han, 2021) models.

The gains offered by early deep neural networks (DNN), HMM, and hybrid models could be attributed mainly to the wider frame windows used as inputs. Although these features are valuable, they are highly correlated, and neural networks become prominent for this task by building specialized acoustic architectures as opposed to building language models over acoustic frames like most of the early neural approaches.

An important milestone in the development of acoustic event detection and classification has also been some form of feature extraction, which transforms raw waveforms into a sequence of feature vectors that can be used as inputs to deep models (Radford et al., 2023; Gong et al., 2021a; Chen et al., 2022; Georgescu et al., 2023). Although MFCC spectrograms were demonstrated to work very well for shallow models, modern deep models mainly utilize mel-spectrograms and very recently superlets (Moca et al., 2021).

2.2. Downstream Effects of Transforms

There has been a significant shift from traditional, hand-crafted audio features such as MFCCs to the use of raw audio waveforms and spectrogram representations as inputs for neural networks. Wyse (2017) showed the advantages of spectrogram representations for deep neural networks, particularly their ability to capture both time and frequency information, which is crucial for effectively modeling and generating complex audio signals. Furthermore, multiple works have employed representations based on spectrograms, coupled with convolutional neural networks, and have shown that these work particularly well together (Hershey et al., 2017; Schmid et al., 2023; Gong et al., 2021b).

A class of models is also based on directly processing raw audio (Verma and Berger, 2021). These often involve segmenting the audio input with some window length before converting it into an embedding compatible with the models, rather than producing spectrograms. However, this class of models has seemed to work well primarily for generative applications (Gardner et al., 2021). We did not focus on this class of models for our analysis.

Although there have been studies exploring the classification power of these transforms (Wyse, 2017; Ji et al., 2020; Moysis et al., 2023), none of these works demonstrate the downstream clinical effects of the preprocessing techniques that we focus on in this work. Furthermore, we also analyze additional transforms such as superlets.

3. Method

The prospective audio data of the human patient was collected at a comprehensive stroke center, and the study was approved by the local Research Ethics Board. Subjects were anonymized and the computation occurred locally according to the local institution guidelines.

3.1. Participants

We enrolled 70 individuals from a comprehensive stroke center, affiliated with the University. These participants were selected during two periods: from June 13, 2022, to January 19, 2023 (epoch 1) and from January 24 to March 4, 2023 (epoch 2), to form training and testing datasets, respectively. Technical problems with the audio recordings resulted in the exclusion of two participants during Epoch 1. Therefore, a total of 68 participant audio samples (with 94% inter-rater agreement on audio quality by AB and HM) were incorporated into our study. The Toronto Bedside Swallowing Screening Test (TOR-BSST©) was administered to all participants as part of standard care, assessing voice changes, repetitive swallows, and dysphonia. Based on this assessment, 27 participants were marked as "fail" and 41 as "pass". TOR-BSST is a dysphagia screening tool that can be used by operators trained in courses of varying backgrounds Martino et al. (2009). The study split these participants into two groups: 40 (58.9%) for training and 28 (41.1%) for testing. The enrollment was ongoing and based on a randomized approach, targeting admissions to the stroke unit within 72 hours of admission. Each participant gave their informed consent and the research was sanctioned by the stroke center's REB. Inclusion criteria were recent stroke patients proficient in English, able to follow instructions, and without severe aphasia. Exclusion criteria included non-English speakers, people with significant speech impairments, or medically unstable individuals.

3.2. Data Collection

Speech data was divided into two types: a) National Institutes of Health Stroke Scale (NIHSS) speech segments and b) sustained vowel pronunciations. The NIHSS was chosen to avoid bias in test selection as it is commonly used in stroke assessments. NIHSS language tests included continuous speech, sentences, and words. The second dataset comprised vowel sounds (/a/, /e/, /i/, /o/, and /u/), with participants pronouncing each vowel for 3 seconds, three times. This choice was based on evidence showing the uniqueness of vowel sounds in detecting swallowing problems. Data collection was done using an encrypted iPhone 12 and the Voice Recorder app in a real hospital setting. The investigators in charge of data collection, segmentation, and model testing were deliberately kept unaware of each other's activities.

3.3. Data Analysis

The initial step involved a quality assessment. A three-stage data processing method was adopted, which included segmentation, transformation, and the use of machine learning. Audacity software was used for data segmentation, and a custom Python program transformed segmented audio into Mel-spectrogram image representations (see Figure 1). Given the varied audio durations, a windowing approach ensured consistent Mel-spectrogram image scaling. The Mel spectrograms, renowned for their accuracy in replicating human auditory perception, were critical for our model's success, particularly when compared against evaluations by speech-language professionals. Two distinct Mel-spectrogram images were created for separate machine learning classifier training: RGB and three-channel Mel spectrograms. The latter blends monochrome versions with different FFT lengths. Additionally, Superlet Transform (SLT) was used to create spectrogram images, to assess their performance against mel spectrograms.

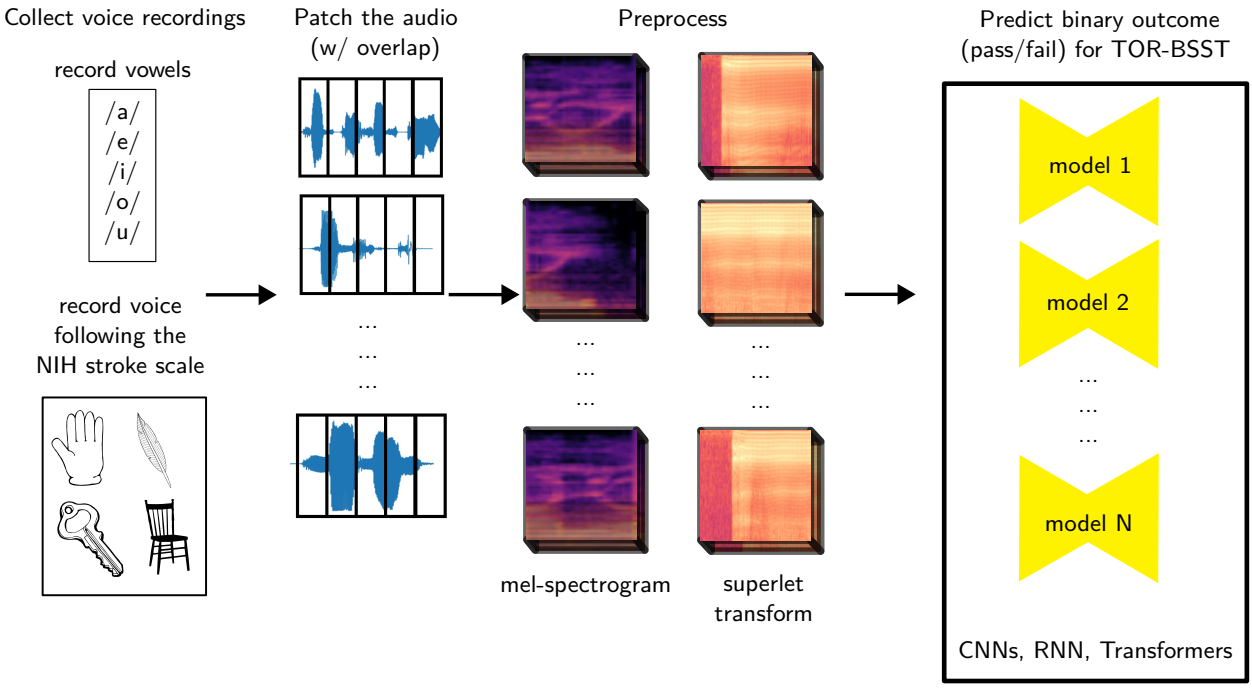

Figure 1: Schematic Overview of the Audio Classification Workflow for Stroke Assessment. The process begins with the collection of voice recordings, including both sustained vowel sounds (/a/, /e/, /i/, /o/, /u/) and speech following the NIHSS, represented by icons for hand, feather, key, chair, etc. These recordings are then segmented using an overlapping patch method to prepare for preprocessing. Subsequently, audio segments are transformed into two types of visual representations: mel-spectrograms and superlet transforms. The final stage involves inputting the processed spectrograms into an array of machine learning models—CNNs, transformers, and RNNs—to predict the outcome of the TOR-BSST© as either "pass" or "fail".

3.4. Pretraining on Public Datasets

This study delves into the improvement of the performance of various networks in a downstream data set through various pretraining scenarios.

Pretraining on public datasets is crucial for developing machine learning applications tailored to clinical needs, especially when working with small datasets Rasmy et al. (2021). This strategy addresses the significant challenge clinicians face in collecting large datasets.

Imagenet is a widely recognized dataset crucial for training and evaluating computer vision models.

AudioSet is a comprehensive database featuring 632 categories of audio events in 2,084,320 human-labeled 10-second clips from YouTube. It presents a hierarchical categorization of various sounds that include those of humans, animals, musical instruments, genres, and everyday life.

US8K is a dataset of 8732 labeled sound excerpts that are up to 4s in length and are taken from the UrbanSound dataset (Salamon et al., 2014). The sounds are composed of urban sound clips with labels such as air conditioner, car horn, children playing etc.

ESC50 consists of 2000 audio clips of 5-second-long recordings categorized into 50 classes. These classes fall into the following major categories: animals, natural soundscapes, water sounds, human, non-speech sounds, domestic sounds, urban noises.

3.5. Exploring Network Architectures for Audio Analysis

In this section, we explore various neural network architectures, including ConvNeXt and DenseNet for CNN-based models, ConvLSTM2D for temporal data analysis, and Vision Transformer (ViT) and SWIN Transformer for transformer-based models. Addition-

Method	Mel RGB	Mel mono	Superlet
YAMNet		✓	
VGGish		✓	
Trill		✓	
BEATS		✓	
ConvNeXT	✓	✓	✓
DenseNet	✓	✓	✓
DenseNet-Contrastive		✓	
DenseNet-Contrastive US8K		✓	
DenseNet-Contrastive ESC50		✓	
ConvLSTM2D	✓		
ViT	✓	✓	
SWIN	✓	✓	✓
AST		✓	

Table 1: List of models evaluated for performance comparison alongside the different pre-processing options used for each.

ally, we introduce pretrained audio feature extractors such as YAMNet, VGGish, and Trill. To address classification tasks, we employ different loss functions and optimizers. For CNN-based models, particularly DenseNet, we implement a hybrid loss function that combines Cross-Entropy and Contrastive Loss. We also incorporate class weights to handle dataset imbalances. Transformer-based models are trained using Cross-Entropy Loss with the inclusion of class weights. The Adam optimizer is chosen for its adaptive learning rate capabilities. Our preprocessing methods involve the use of grayscale audio spectrograms and the conversion of spectrograms into RGB images for select models. Additionally, we explore the use of Superlet transforms in preprocessing. Finally, we evaluate our classifiers using per-participant prediction aggregation (Majority Voting).

3.5.1. Convolutional Neural Networks (CNNs)

ConvNeXt is a recent adaptation of the CNN architecture that has shown impressive results in image classification tasks. Although primarily designed for typical image classification applications, it can be adapted for audio spectrogram analysis, offering potentially more effective feature extraction in sound-based medical diagnostics.

DenseNet is known for its densely connected convolutional networks, where each layer is connected to every other layer in a feed-forward fashion (Huang et al., 2018). The strong performance of the basic DenseNet architecture, as detailed in Table 3, encouraged us to explore its various adaptations, namely:

- DenseNet with binary cross-entropy loss, pretrained on ImageNet (referred to as **DenseNet**).
- DenseNet with a hybrid loss (contrastive loss and binary cross entropy), pretrained on ImageNet (referred to as **DenseNet Contrastive**).
- DenseNet pretrained on ImageNet, then on US8K, and applied to our dataset using the hybrid loss (referred to as **DenseNet Contrastive US8K**).
- DenseNet pretrained on ImageNet and ESC50, later trained on our dataset with the hybrid loss (referred to as **DenseNet Contrastive ESC50**).

Pretrained CNN-based Audio Feature Extractors:

YAMNet is a CNN that uses the popular MobileNetV1 architecture for the detection of audio events (Howard et al., 2017). It is pretrained on AudioSet to classify various sounds. This pretraining approach streamlines the creation of spectrogram-based filters without requiring extensive proprietary data.

VGGish (Hershey et al., 2017) is a variant of the VGG model (Simonyan and Zisserman, 2014), adapted for audio processing. Initially developed for image classification, the VGG architecture's adaptation to audio allows it to extract meaningful features from sound waves. Like YamNet, it is pretrained on AudioSet.

Trill "TRIpLet Loss network" is designed for sound event detection and is particularly effective in distinguishing fine-grained acoustic differences. Trill is based on the ResNet50 architecture (He et al., 2015) and has shown clear improvements over popular sound classification models (Shor et al., 2020).

3.5.2. Recurrent Neural Networks (RNNs)

In this paper, we utilized ConvLSTM2D, an amalgamation of CNNs and Long Short-Term Memory (LSTM) networks, designed to capture spatial and temporal relationships in data (Shi et al., 2015). This is particularly effective for our use case, which involves generating a series of consecutive spectrograms by splitting a single audio file into multiple fixed-length segments.

3.5.3. Transformers

Vision Transformer (ViT) is tailored for adaptation from large-scale pretraining to fine-tuning on smaller datasets, a process that involves replacing the original MLP head with a new linear layer tailored to the specific class size of the task at hand (2 in our case). This adjustment allows ViTs to be efficiently customized for new tasks, without complete retraining (Dosovitskiy et al., 2020). ViTs utilize self-attention to capture long-range dependencies, a feature that, while powerful, requires extensive training data to achieve the innate perceptual capabilities of CNNs (Zhang et al., 2023).

SWIN Transformer is a variant of ViT that features a unique "shifted window" self-attention mechanism. Unlike ViT, which applies self-attention to the entire sequence of tokens, Swin Transformer performs attention over square-shaped blocks of patches, each block being analogous to a receptive field in convolutional layers. This method enhances hierarchical feature processing with less computational demand. The SWIN Transformer architecture integrates "merging layers" for efficient token downsampling and incorporates advanced features such as layer normalization and scaled cosine attention, significantly improving performance and adaptability in transfer learning scenarios (Liu et al., 2021).

Pretrained Transformer-based Audio Feature Extractors:

AST is a convolution-free, purely attention-based model designed for audio classification (Gong et al., 2021a). We used the ast-finetuned-audioset-10-10-0.4593 version, pretrained on the AudioSet dataset. Exploring this model represents a shift from our previous approach, which was more image classification-centric, to one that is specifically designed for audio classification.

BEATs model, designed for the extraction of audio features, incorporates 12 transformer encoder layers, 768 hidden states, and 8 attention heads. Pretrained on AudioSet, BEATs has been evaluated across various audio (AS-2M, AS-20K and ESC-50) and speech (KS1, KS2 and ER) classification tasks, demonstrating its versatility and effectiveness in processing and understanding complex audio data (Chen et al., 2022).

3.6. Loss functions and optimizers

For our CNN-based models, particularly DenseNet, we implemented a hybrid loss function combining cross-entropy and contrast loss, incorporating class weights to address the imbalance in our data set, notably the underrepresentation of the "fail" class (27 patients) compared to the "pass" class (41 patients). Similarly, for Transformer-based models, we applied Cross-Entropy Loss, ensuring class weights were also used. Adam optimizer was chosen for its adaptive learning rate (Kingma and Ba, 2014), optimizing efficiency between models with varied learning rates. This approach was uniformly applied to all trainable parameters to ensure balanced learning dynamics.

3.7. Preprocessing

Given the large search space of the many options we wanted to explore in this paper, we chose to approach the problem strategically starting with the most popular approach in the literature: Grayscale (referred to as Mel mono in Table 1) audio spectrograms. In order to take advantage of ImageNet pretraining, we ran spectrograms in three different settings, and concatenated the three spectrograms into a single three-channel image. This was then used as input and compared across all models except ConvLSTM2D. The "Mel mono" method allowed us to establish a baseline to compare across all feasible models, which we then used to refine our training approach and model selection strategy.

For select models, we converted a single-channel spectrogram into an RGB three-channel ("Mel RGB") image using color-maps and tried this approach on ConvNeXt, Densenet, ViT, Swin Transformer, and ConvLSTM2D network architectures (see Table 1). The strategy here was to make use of the RGB feature extraction abilities of models trained on ImageNet, an RGB-image dataset. Additionally, Palanisamy et al. (2020) discussed a similar approach to convert the spectrogram to an RGB

Method	Mel RGB			Mel mono			Superlet		
	AUC	ST	SP	AUC	ST	SP	AUC	ST	SP
YAMNet	-	-	-	0.69	0.71	0.79	-	-	-
VGGish	-	-	-	0.82	0.71	0.93	-	-	-
Trill	-	-	-	0.71	0.57	0.86	-	-	-
BEATS	-	-	-	0.44	0.57	0.55	-	-	-
ConvNeXt	0.91	0.78	0.89	0.86	0.78	0.79	0.74	0.68	0.66
DenseNet	0.89	0.89	0.79	0.88	0.78	0.74	0.74	0.67	0.67
DenseNet Contrastive	-	-	-	0.82	0.86	0.78	-	-	-
DenseNet Contrastive US8K	-	-	-	0.89	0.78	1.00	-	-	-
DenseNet Constrastive ESC50	-	-	-	0.75	0.71	0.78	-	-	-
ConvLSTM2D	0.52	0.78	0.11	-	-	-	-	-	-
ViT	0.79	0.67	0.89	0.84	0.40	0.94	-	-	-
SWIN Transformer	0.80	0.78	0.68	0.83	0.60	0.83	0.78	0.67	0.79
AST	-	-	-	0.83	0.89	0.60	-	-	-

Table 2: Overview of model performance on participant level classification task (where AUC represents Area Under the ROC Curve, ST represents the Sensitivity, and SP represents the Specificity). For each operator characteristic we highlight the best performance values.

image or choosing different window sizes and hop lengths to create three distinct channels that are concatenated into a single image. They found that on the basis of the baseline model experiments, using mel spectrograms with different window sizes and hop lengths in each channel yielded better performance.

Another preprocessing approach used in this study is Superlet transforms, which is a relatively recent method to transform time-series data into a spectrogram that preserves both time and frequency resolution. Similarly, the grayscale Superlet spectrogram was converted into an RGB image. This approach was only applied to the best performing models, namely ConvNeXt, Densenet, and Swin Transformer.

3.8. Evaluation of classifiers

We evaluated each classifier using the aggregation of prediction by participant (majority voting), treating each participant as a single data point, regardless of the number of associated clips. The majority vote across a participant's clips determines their overall prediction. Various performance metrics including F1 Score, Precision, Recall (Sensitivity), and Specificity, both at the validation and test stages were calculated. The Receiver Operating Characteristic (ROC) curve and the confusion matrices were also plotted to visually assess performance of models.

4. Results

Evaluation of various models in a participant-level classification task reveals performance differences measured by AUC, Sensitivity (ST), and Specificity (SP).

The ConvNeXt model showcased a robust performance across all three metrics in the "Mel RGB" category, achieving an AUC of 0.91, ST of 0.78, and SP of 0.89. It maintained this strong performance in the "Mel mono" category, with an AUC of 0.86, ST of 0.78, and SP of 0.79, and in the "Superlet" category, with slightly lower scores of 0.74 for AUC, 0.68 for ST, and 0.66 for SP.

DenseNet models also performed well, with the standard DenseNet achieving an AUC of 0.89, the highest ST of 0.89, and SP of 0.79 in "Mel RGB". In "Mel mono", it scored an AUC of 0.88, ST of 0.78, and SP of 0.74, with consistent performance in "Superlet" (AUC 0.74, ST 0.67, SP 0.67). Interestingly, the DenseNet Contrastive model excelled in "Mel mono" with an AUC of 0.82, ST of 0.86, and SP of 0.78, suggesting its effectiveness in monochrome settings.

In contrast, the ConvLSTM2D model underperformed in the "mel RGB" category, with an AUC of only 0.52, although it had a satisfactory ST of 0.78. However, its SP of 0.11 was notably low, indicating a high rate of false positives.

Method	AUC 95% CI	ST (Recall)	SP	Precision	F1 score
ConvNeXT(RGB)	0.77 - 1.0	0.78	0.89	0.78	0.84
DenseNet(RGB)	0.73 - 1.0	0.89	0.79	0.67	0.81
SWIN Transformer(mono)	0.5-0.9	0.6	0.83	0.55	0.57
AST(mono)	0.68 - 0.98	0.89	0.60	0.80	0.84
DenseNet-Constrastive US8K(mono)	0.76-1.0	0.78	1.0	1.0	0.88

Table 3: Detailed comparison of model performance on participant level classification task (AUC=Area Under the ROC Curve. ST = Sensitivity. SP = Specificity).

The SWIN Transformer model demonstrated versatility with competitive scores across all categories. It achieved an AUC of 0.801, ST of 0.78, and SP of 0.68 in "Mel RGB", and showed improvement in "Mel mono" with an AUC of 0.83, ST of 0.6, and SP of 0.83. In the "Superlet" category, it scored an AUC of 0.78, ST of 0.67, and SP of 0.79.

The performance of the AST model in "Mel mono" was notably effective, with an AUC of 0.83 and ST of 0.89, but its SP of 0.60 suggests a significant trade-off, with a higher tendency for false positives. According to Gong et al. (2021a), the AST model does not require as many epochs to train as the CNN-attention hybrid models, which need significantly more epochs. It is worth noting that the AST model required only 6 epochs of training on our dataset to achieve these metrics, which is fewer compared to the CNN-attention hybrid models and other Transformer models explored in this study that needed significantly more epochs to train.

Other models such as YAMNet, VGGish, Trill, and BEATS were assessed only in the "Mel mono" category. VGGish showed promising results with an AUC of 0.82, ST of 0.71, and the highest SP of 0.93 among the CNN-based feature extractors, benefiting from its pretraining on AudioSet.

Finally, a detailed comparison of the best performing models (Table 3) in terms of their statistical performance metrics provides a deeper understanding of their predictive capabilities. The ConvNeXt (Mel RGB) model's AUC confidence interval ranged from 0.77 to 1.0, indicating a high degree of certainty in its classification performance, with a commendable F1 score of 0.84, balancing precision and sensitivity. The precision and recall rates of this model, both at 0.78, suggest a harmonious balance between the positive predictive value and the true positive rate.

On the contrary, DenseNet ("Mel RGB") showed a wider confidence interval in the AUC of 0.73 to 1.0, reflecting more variability in its performance. Despite this, it had the highest sensitivity of 0.89, which shows its strength in identifying true positives. However, the trade-off is evident in its precision of 0.67, which is lower compared to ConvNeXt, leading to an F1 score of 0.81 that, while high, indicates room for improvement in precision.

The SWIN Transformer ("Mel mono") had a narrower confidence interval for AUC, ranging from 0.5 to 0.9. This range suggests more uncertainty in the model's performance, which is also reflected in the lowest F1 score of 0.57 among the evaluated models. A sensitivity of 0.6 and a specificity of 0.83 show an imbalance, with the model favoring the correct identification of negatives over positives, as also implied by a lower precision rate of 0.55.

AST ("Mel mono") showed a strong performance with an AUC confidence interval between 0.68 and 0.98 and an F1 score equal to ConvNeXt at 0.84. The model's high sensitivity at 0.89 is on par with DenseNet ("Mel RGB"), but a lower specificity of 0.60 points to a higher false positive rate.

The DenseNet-contrastive US8K model ("Mel mono") stood out with an AUC confidence interval of 0.76 to 1.0 and perfect scores for both specificity and precision, both at 1.0. This exceptional performance resulted in the highest F1 score of 0.88, indicating a very strong predictive power where the model excelled both in recognizing true positives and in avoiding false positives.

5. Discussion

Our study examines the associations between spectrogram preprocessing techniques and the ensuing performance of audio classification models, underscoring an important consideration for clinical applications: the nuanced efficacy of preprocessing approaches has a significant bearing on leveraging trans-

fer learning. Our work suggests that while RGB preprocessing exhibits superior performance in conjunction with ImageNet pretraining, the "Mel mono" approach, when pretrained on expansive public audio datasets, surpasses RGB's effectiveness. This insight is crucial, suggesting that in clinical settings, where data limitations and intrinsic differences are prevalent, adopting a more standardized and contextually tailored approach to preprocessing could significantly enhance the performance of deep learning models. Moreover, the observed variances in model architecture performance, particularly the robustness of transformer-based models versus traditional CNNs in handling limited training epochs, offer a promising avenue for refining audio classification frameworks. This suggests that through strategic selection of preprocessing techniques and models there may be more optimal audio classification strategies that can improve diagnostics with heightened accuracy and efficiency in clinical environments. This has implications for voice as a biomarker in stroke and other neurologic conditions, in addition to other states of disease where data limitations may be intrinsic to the health condition, including rare diseases.

Further to this point, the complexities of dealing with diverse patient populations, especially in a small data set necessitate careful consideration of confounding factors. For example, in the context of stroke, variables such as age, gender, stroke severity and type, medical comorbidities, medications, cognitive function, psychological factors, rehabilitation history, time since stroke, environmental factors, and nutritional status significantly impact patient responses to audio classification-based diagnostics. Properly accounting for these confounders through stratified analysis could unveil more nuanced insights into how preprocessing techniques perform across varied patient demographics, ultimately refining the clinical utility of audio classifiers.

The DenseNet model, particularly Contrastive US8K variant, excelled by leveraging a hybrid loss combining cross-entropy with a supervised contrastive loss, significantly enhancing specificity and yielding the highest F1 score of 0.88 among the evaluated models. The pretraining and fine-tuning process, progressing from general large-scale audio datasets to specialized clinical data, proved crucial in developing effective feature extractors for audio spectrogram classification.

The transformer-based approach, relatively new in the field of audio analysis, demonstrates the potential to adapt architectures originally designed for other domains, such as natural language processing, to audio classification. The AST model's sensitivity score suggests a strong grasp of relevant audio spectrogram features. It showed training efficiency, achieving optimal results with just 6 epochs, contrasting with CNN-based models requiring more epochs for similar performance. This highlights the potential of transformer pretrained models in audio classification, even with limited training epochs.

We observed variations in model performance, partially explained by by intrinsic differences in model architecture's as well as limited data. Surprisingly, the RGB preprocessing approach outperformed the grayscale triple channel approach when using ImageNet pretraining. Theoretically, concatenating three grayscale spectrograms constructed from different Mel transform settings should outperform a single mel spectrogram transformed into an RGB image through color mapping. However, our results suggest that the convolutional layers of models pretrained on ImageNet might be better attuned to the features present in RGB images. This had not been widely discussed in the literature before and we noticed some variations where some papers use grayscale images (Chen et al., 2022; Gong et al., 2021a; Howard et al., 2017; Mu et al., 2021), and others often use RGB representations of the spectrograms (Aykanat et al., 2017; Zaman et al., 2023). This surprising performance could be attributed to the ImageNet pretrained models' well-tuned and generalized filters that extract features from structure as well as color. Such generalized filters trained on massive datasets are robust to overfitting and could explain the improved performance on the RGB input.

6. Conclusion

In summary, our study underscores the effectiveness of modern CNN architectures, such as DenseNet and ConvNeXt, in the field of clinical audio classification. These architectures demonstrate robustness, often rivaling or even surpassing the capabilities of transformer models, particularly in scenarios involving small datasets. A key factor in this success is the strategic use of open-source pretrained weights, which not only accelerates the development process but also significantly enhances model accuracy.

A cornerstone of our research involved the prospective collection of two first-of-their-kind patient audio datasets from stroke patients. We introduce the

use of an NIHSS-based audio dataset, a novel collection that captures continuous speech, sentences, and words based on NIHSS. This well-established test for the assessment of stroke in emergency departments provides invaluable data to develop audio classification models tailored to clinical needs. In addition, we presented the Vowel dataset, a unique compilation of sustained vowel sounds from patients, offering new insights into the analysis of swallowing disorders.

Leveraging open datasets for pretraining enables generalized feature learning, essential for subsequent fine-tuning on specific datasets. Our study highlights the effectiveness of a multistage training and fine-tuning process for gradual model adaptation and improved performance. The influence of preprocessing techniques, such as Mel RGB, Mel mono, and Superlet, on model performance is significant and requires careful selection. Temporal segregation between training and testing data sets is crucial to prevent data leakage and improve model generalization. Furthermore, our findings underscore the potential of audio as a robust predictor of clinically relevant information, exemplified by our successful prediction of swallowing status based solely on audio data, suggesting promising applications in clinical settings.

Through nuanced consideration in data handling and model training, our research contributes to the advancement of clinical audio classification, with promising implications for its application in various neurologic conditions and beyond.

Acknowledgments

We acknowledge local institutional funding from the Sunnybrook AFP Innovation fund, in addition to Summer Research Studentships from TCAIREM (Temerty Centre for Artificial Intelligence Research and Education in Medicine). We thank anonymous reviewers of the CHIL Conference for their insightful suggestions which we incorporate in this work.

References

Kawther S. Alqudaihi, Nida Aslam, Irfan Ullah Khan, Abdullah M. Almuhaideb, Shikah J. Alsunaidi, Nehad M. Abdel Rahman Ibrahim, Fahd A. Alhaidari, Fatema S. Shaikh, Yasmine M. Alsenbel, Dima M. Alalharith, Hajar M. Alharthi, Wejdan M. Alghamdi, and Mohammed S. Alshahrani. Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities. *IEEE Access*, 9:102327–102344, 2021. doi: 10.1109/ACCESS.2021.3097559.

Murat Aykanat, Özkan Kılıç, Bahar Kurt, and Sevgi Saryal. Classification of lung sounds using convolutional neural networks. *EURASIP Journal on Image and Video Processing*, 2017(1):1–9, 2017.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Sebastian Böck and Markus Schedl. Enhanced beat tracking with context-aware neural networks. In *Proc. Int. Conf. Digital Audio Effects*, pages 135–139, 2011.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. A comparison of audio signal preprocessing methods for deep neural networks on music tagging. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1870–1874, 2018. doi: 10.23919/EUSIPCO.2018.8553106.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

P Dhanalakshmi, S Palanivel, and Vennila Ramalingam. Classification of audio signals using svm and rbfnn. *Expert systems with applications*, 36 (3):6069–6075, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Amit Krishna Dwivedi, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. Algorithms for automatic analysis and classification of heart sounds–a systematic review. *IEEE Access*, 7:8316–8345, 2018.

Babatunde S Emmanuel. A review of signal processing techniques for heart sound analysis in clinical diagnosis. *Journal of medical engineering & technology*, 36(6):303–307, 2012.

Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digital biomarkers*, 5(1):78–88, 2021.

María Teresa García-Ordás, José Alberto Benítez-Andrades, Isaías García-Rodríguez, Carmen Benavides, and Héctor Alaiz-Moretón. Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data. *Sensors*, 20(4), 2020. ISSN 1424-8220. doi: 10.3390/s20041214. URL https://www.mdpi.com/1424-8220/20/4/1214.

Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel. Mt3: Multi-task multitrack music transcription. *arXiv preprint arXiv:2111.03017*, 2021.

N Gavriely, M Nissan, DW Cugell, and AHE Rubin. Respiratory health screening using pulmonary function tests and lung sound analysis. *European Respiratory Journal*, 7(1):35–42, 1994.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023.

Pablo Gimeno, Ignacio Viñals, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020:1–19, 2020.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021a.

Yuan Gong, Yu-An Chung, and James Glass. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021b.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma. Tandem connectionist feature extraction for conventional hmm systems. In *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1635–1638. IEEE, 2000.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

Andre Holzapfel and Yannis Stylianou. Musical genre classification using nonnegative matrix factorization-based features. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2): 424–434, 2008.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Maaten Laurens van der, and Kilian Q. Weinberger. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2018. URL https://arxiv.org/pdf/1608.06993.pdf.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *arXiv preprint arXiv:2207.06405*, 2022.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

Shulei Ji, Jing Luo, and Xinyu Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*, 2020.

B.-H. Juang. On the hidden markov model and dynamic time warping for speech recognition — a unified view. *AT&T Bell Laboratories Technical Journal*, 63(7):1213–1243, 1984. doi: 10.1002/j.1538-7305.1984.tb00034.x.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/abs/1412.6980.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

Li Khim Kwah and Joanna Diong. National institutes of health stroke scale (nihss). *Journal of physiotherapy*, 2014.

Sandra Larson, Germán Comina, Robert H Gilman, Brian H Tracey, Marjory Bravard, and José W López. Validation of an automated cough detection algorithm for tracking recovery of pulmonary tuberculosis patients. 2012.

Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321*, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Rosemary Martino, Frank Silver, Robert Teasell, Mark Bayley, Gordon Nicholson, David L Streiner, and Nicholas E Diamant. The toronto bedside swallowing screening test (tor-bsst) development and validation of a dysphagia screening tool for patients with stroke. *Stroke*, 40(2):555–561, 2009.

Vasile V Moca, Harald Bârzan, Adriana Nagy-Dăbâcan, and Raul C Mureșan. Time-frequency super-resolution with superlets. *Nature communications*, 12(1):337, 2021.

Ria Lestari Moedomo, M Sukrisno Mardiyanto, Munawar Ahmad, Bachti Alisjahbana, and Tjahjono Djatmiko. The breath sound analysis for diseases diagnosis and stress measurement. In *2012 International Conference on System Engineering and Technology (ICSET)*, pages 1–6. IEEE, 2012.

Lazaros Moysis, Lazaros Alexios Iliadis, Sotirios P Sotiroudis, Achilles D Boursianis, Maria S Papadopoulou, Konstantinos-Iraklis D Kokkinidis, Christos Volos, Panagiotis Sarigiannidis, Spiridon Nikolaidis, and Sotirios K Goudos. Music deep learning: Deep learning methods for music signal processing-a review of the state-of-the-art. *IEEE Access*, 2023.

Wenjie Mu, Bo Yin, Xianqing Huang, Jiali Xu, and Zehua Du. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11(1):21552, 2021.

Mattias Nilsson, Harald Gustaftson, Søren Vang Andersen, and W Bastiaan Kleijn. Gaussian mixture model based mutual information estimation between frequency bands in speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–525. IEEE, 2002.

Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE transactions on audio, speech, and language processing*, 18(3):550–563, 2009.

Madhurananda Pahar, Marisa Klopper, Byron Reeve, Rob Warren, Grant Theron, and Thomas Niesler.

Automatic cough classification for tuberculosis screening in a real-world environment. *Physiological Measurement*, 42(10):105014, nov 2021. doi: 10.1088/1361-6579/ac2fb8. URL https://dx.doi.org/10.1088/1361-6579/ac2fb8.

Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*, 2020.

Huy Phan, Philipp Koch, Fabrice Katzberg, Marco Maass, Radoslaw Mazur, and Alfred Mertins. Audio scene classification with deep recurrent neural networks. *arXiv preprint arXiv:1703.04770*, 2017.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL http://dl.acm.org/citation.cfm?doid=2733373.2806390.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Christopher Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE transactions on pattern analysis and machine intelligence*, 21(4):360–370, 1999.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Rami Saab, Arjun Balachandar, Hamza Mahdi, and Eptehal Nashnoush. Machine-learning assisted swallowing assessment: a deep learning-based quality improvement tool to screen for post-stroke dysphagia. *Frontiers in Neuroscience*, 17, 2023. doi: 10.3389/fnins.2023.1302132. URL https://www.frontiersin.org/articles/10.3389/fnins.2023.1302132/full.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014.

Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

Florian Schmid, Khaled Koutini, and Gerhard Widmer. Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Frank Seide, Gang Li, Xie Chen, and Dong Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 24–29, 2011. doi: 10.1109/ASRU.2011.6163899.

Stefano Sello, Soo-kyung Strambi, Gennaro De Michele, and Nicolino Ambrosino. Respiratory sound analysis in healthy and pathological subjects: A wavelet approach. *Biomedical Signal Processing and Control*, 3(3):181–191, 2008.

Xingjian Shi, Zhourong Chen, Hao Wang, DY Yeung, W-K Wong, and WC Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. arxiv 2015. *arXiv preprint arXiv:1506.04214*, 2015.

Joel Shor, Aren Jansen, Ronnie Zvi Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, and Dotan Emanuel. Towards learning a universal non-semantic representation of speech. 2020. URL https://www.isca-speech.org/archive/Interspeech_2020/abstracts/1242.html.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Brian H. Tracey, Germán Comina, Sandra Larson, Marjory Bravard, José W. López, and Robert H. Gilman. Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6017–6020, 2011. doi: 10.1109/IEMBS.2011.6091487.

Jörgen Valk and Tanel Alumäe. Voxlingua107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658, 2021. doi: 10.1109/SLT48900.2021.9383459.

Prateek Verma and Jonathan Berger. Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions. 2021.

Patrizia Vizza, Giuseppe Tradigo, Domenico Mirarchi, Roberto Bruno Bossio, Nicola Lombardo, Gennarina Arabia, Aldo Quattrone, and Pierangelo Veltri. Methodologies of speech analysis for neurodegenerative diseases evaluation. *International Journal of Medical Informatics*, 122:45–54, 2019. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2018.11.008. URL https://www.sciencedirect.com/science/article/pii/S1386505618307639.

Lonce Wyse. Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint arXiv:1706.09559*, 2017.

Khalid Zaman, Melike Sah, Cem Direkoglu, and Masashi Unoki. A survey of audio classification using deep learning. *IEEE Access*, 2023.

Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, pages 1–22, 2023.