Speech and Language Impairment Detection by Means of AI-Driven Audio-Based Techniques

Luca Corvitto[1], Lorenzo Faiella[1], Christian Napoli[1], Adriano Puglisi[1] and Samuele Russo[2]

[1]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, Roma, 00185, Italy
[2]Department of Psychology, Sapienza University of Rome, Via dei Marsi 78, Roma, 00185, Italy

Abstract

Speech and Language Impairments (SLI) affect a large and heterogeneous group of people. With our work, we propose a novel, easy, and immediate detection tool to help diagnose people who suffer from SLI using speech audio signals, along with a new dataset containing English speakers affected by SLI. In this work, we experiment with feature extraction methods such as Mel Spectrogram and wav2vec 2.0, as well as classification methods such as SVM, CNN, and linear neural networks. We also work on data audio augmentation trying to overcome the very common limitations imposed by data scarcity in the medical field. The overall results indicate that the wav2vec 2.0 feature extractor, paired with a linear classifier, provides the best performance with a reasonably high accuracy of over 96%.

Keywords

SLI, AI, audio, healthcare, speech, learning disease, feature extraction, data augmentation

1. Introduction

The rapid development of the use of Artificial Intelligence (AI) techniques in a broad range of scientific fields has helped solve real-life problems, in particular, the new advancements revolutionized a wide variety of areas such as Natural Language Processing (NLP) [1], computer vision [2, 3], robotics and many more. Due to the huge volume of medical data being generated worldwide, there is a clear need for efficient use of this information to benefit health sectors around the world [4, 5]. The medical community has taken strong notice of the potential of these new technologies in AI. Machine learning (ML) thrives in areas where there are lots of data, therefore ML is one of the essential and most effective tools in analyzing highly complex medical data [6]. For example, analyzing medical data originating from disease diagnosis with the aid and benefits given by these tools could be a lot more financially efficient. In healthcare, it is also vital that diseases are detected early on during diagnosis and prognosis. The success of these AI methods has also spread across other domains, including speech recognition and the music recommendation task [7]. Due to the relevance of such systems in our day-to-day lives, there is an increasing need for effective and efficient audio classification systems. Automatic classification technologies are widely applied in voice assistants [8], chatbots [9], smart safety devices [10, 11], and in different real-world environments [12, 13, 14].

Our project aims to conciliate these two worlds and design a Deep Learning (DL) model that can detect, from a given audio input, if the speaker could be affected by a speech and language impairment. Individuals with a Speech and Language Impairment (SLI), generally, despite normal hearing, normal nonverbal intelligence, adequate social functioning, and no obvious signs of brain injury represent a heterogeneous group of people with significant difficulty in learning languages [15]. One of the defining characteristics of SLI is speech disfluency, more specifically impaired acquisition of pattern-based components in language, such as morphology, syntax, and some aspects of phonology such as stuttering. This commonly used definition leads to early hypotheses regarding the etiology of SLI that an impaired language-specific learning mechanism underlies language development and disorders [16, 17, 18]. This disorder is deemed "primary" or "specific" when there is no clear explanation for these lags in language skills, a defining characteristic of primary language disorder is that its cause is unknown [19]. Language disorders are also linked to a heightened risk for psychiatric concerns, attentional difficulties, social-behavioral problems, and learning disabilities [20, 21]. Many current trends in audio signal processing rely on data-driven machine learning approaches to achieve state-of-the-art results [22, 23, 24]. However, the quantity and quality of available data influences heavily the achieved performance for a task. Depending on the specific task, as for our case study, such data can often be hard to obtain and costly to label particularly in

ICYRIME 2024: 9th International Conference of Yearly Reports on Informatics, Mathematics, and Engineering. Catania, July 29-August 1, 2024

corvitto.1835668@studenti.uniroma1.it (L. Corvitto); faiella.1835950@studenti.uniroma1.it (L. Faiella); cnapoli@diag.uniroma1.it (C. Napoli); puglisi@diag.uniroma1.it (A. Puglisi); samuele.russo@uniroma1.it (S. Russo)
0000-0002-3336-5853 (C. Napoli); 0009-0007-6307-7194 (A. Puglisi); 0000-0002-1846-9996 (S. Russo)

the audio domain. As a consequence, researchers often have to deal with datasets of insufficient size or quality. Usually, diagnosis of this type of problem is carried out with human experts, with special in-loco tests [25, 26] or with the aid of tools such as electroencephalogram (EEG) [27]. We want to design an easy and accessible model that can detect if a person could be affected by an SLI without having to go through complex and time-consuming procedures. In this manner, such a model could also be implemented in robots from a human-robot interaction (HRI) perspective, allowing the machine to detect people with SLI and change its behavior and form of interaction accordingly.

This study proposes an analysis of a novel, yet simple, approach of using exclusively audio recordings for SLI detection. Specifically, in Section 2 we start by exploring the current literature, and then we will talk about the problems faced in collecting our data and how we handled them in section 3. After that, in Section 4, we will go through an analysis of the techniques and models used to perform the detection, the trials and results we obtained from them in Section 5, and then we will discuss the limits of our approach in Section 6. We will finally draw our conclusions in Section 7.

2. Related Works

In the ever-evolving landscape of computer science and artificial intelligence, the domains of audio data augmentations and feature extraction are undergoing very rapid changes and revolutions thanks to groundbreaking research and advancements. In the following sections, we will delve into the story and explore the state of the art of these fields.

2.1. Audio Data Augmentation

One of the most important challenges in developing an efficient and effective audio classification system is accessing a large and well-annotated dataset. One of the main obstacles in developing sound classifications is a lack of a sufficient quantity of labeled data. This is due to the following main reasons: class imbalance, data privacy issues, time constraints involved in data collection, high dependency on expertise for effective annotation, etc. [28, 29, 30] Data Augmentation (DA) is defined as the creation of new data by adding deformations to increase the variety of the data so that these deformations do not change their semantic value. It is well known that DA can improve the algorithm's performance, tackle the issue of overfitting [31, 32], and improve the generalization ability of Deep Neural Networks (DNN); this happens because DA averages over the orbits of the group that keeps the data distribution invariant, which leads to variance reduction [33]. DA is key when dealing with problems regarding audio signals because the Convolutional Neural Network (CNN) is the most widely used model in audio applications and when faced with small datasets, CNN's capacity for information retention becomes a flaw; the models memorize the training data and lose performance on new data [34, 35]. In addition to increasing generalization capabilities, the augmentation of data also allows the designed system to improve data significance, regardless of the available data samples [36, 37]. These strategies include methods on raw audio signals, as well as applying other techniques on samples converted into spectrograms or even more complex approaches such as interpolation and nonlinear mixing on the spectrum. We will now list and briefly explain the most used audio augmentation techniques.

Pitch Shifting. The tone of each audio signal in the dataset is lowered or raised by a factor preserving its duration.

Time Stretching. The audio sample is slowed down or sped up by a ratio without altering the pitch drastically.

Time Shifting. Time is shifted to the left or to the right by a random factor or by a predetermined amount.

Volume Adjustment. The volume of the audio file is altered, there is a change in loudness, or sometimes a dynamic range compression is applied.

Noise addition. Noise is introduced into the samples, other than a simple random Gaussian noise there are many types of noises such as white noise [38], babble noise, static noise [39], factory noise, etc.

SpeedUp. The signal is resampled at a preset sampling rate and later returned at the original sampling rate, resulting in a speed change.

Filtering. Several kinds of filters are applied to the input audio. Most of the common filters are band-pass, band-stop, high-pass, high-shelf, low-pass, low-shelf, and peaking filters.

This topic is so important that researchers also developed and designed methods that generate entirely new samples, for example with the aid of a Generative Adversarial Network (GAN) in [40] people created new variants of the audio samples that already existed in their dataset and then utilized an evolutionary algorithm to search the input domain to select the best-generated samples, in this way they were able to generate audio in a controlled manner that contributed to an improvement in classification performance of the original task. One very recent DA method proposed by Google is SpecAugment [41], in this method, the two-dimensional spectrum diagram is treated as an image with time on the horizontal axis and frequency on the vertical axis. Encoder-decoder networks are becoming very popular in fields different from NLP, this is because they can convert a high-dimensional input into a lower-dimensional vector in latent space, researchers in [42] have experimented with a Long Short

Term Memory (LSTM) based auto-encoder to produce artificial data.

2.2. Audio Feature Extraction and Models

It should be noted that data augmentation is not the only way to reduce overfitting and improve the generalization ability of DL models. Model structure optimization, transfer learning, and One-shot and Zero-shot learning are also known strategies that deal with overfitting from different aspects. We will now focus on the most common processing flow of audio classification: preprocessing the original audio data, feature extraction, and feeding the features into the DL model. Audio signals have very high dimensionality, so thousands of floating point values are required to represent a short audio signal, raising the need for exploring dimensionality reduction and feature extraction methods. The degree of how great or poor a model performs is also determined by the choice of features used feature representation is crucial to improve the performance of learning algorithms in the sound classification task. One of the first features that comes to mind when thinking of an audio signal is the spectrogram, its characteristics have been widely used by previous researchers in different domains of sound classification, such as heartbeat sounds to detect heart diseases [43]. Another method used to extract features implements the Mel-Frequency Cepstrum (MFC), which is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency, where the Mel-Frequency Cepstral Coefficients (MFCC) were successful in representing sounds for the detection of respiratory diseases [44]. Some methods that also use the MFC are the long-mel [45], mel filter bank energy [46], inverted MFCC [47], and many more. Although mel spectrogram and MFCC are commonly used, people also implement bag of audio words [48], Discrete Gabor Transform (DGT) audio image representation [49], ZCR, entropy of energy, spectral centroid, spectral spread, spectral entropy [50], and so on.

Classification is a common task in ML and pattern recognition. DL methods applied in these tasks, such as CNN models, often do not perform as well as more traditional ML methods such as random forest, Adaboost, etc., especially in small data [51]. On the other hand, typical ML algorithms, such as ensemble classifiers have been shown to learn features better and adapt more with improved generalization abilities even in the case of small and imbalanced datasets. Over the past years, different ML algorithms have been used for detecting sound events and medical sounds, and the achieved results were of great significance. Classifiers, such as Support Vector Machine (SVM), have shown to be very effective in sound classification tasks [52], also MultiLayer Perceptrons (MLP) were very useful in person identification using speech and breath sounds [53], Hidden Markov Models (HMM) [54], logistic regression and linear discriminant analysis [55] and others. Some studies exploited the effectiveness of multiple simpler methods with ensemble methods such as random forests [56, 51], XgBoost [57], and so on. Unfortunately, considering the complexity of sound and the need to sometimes train an extremely sensitive classifier that can identify different representations of sound features, traditional ML still suffers in these kinds of tasks from having less complex models. In this case, the choice of DL methods has been proven to be more efficient. DL methods differ from traditional ones because they can extract meaningful features from data through the application of a hierarchical structure [58] CNNs were able to achieve significant and more accurate training results [59]. People tried to combine the best of these two worlds by implementing hybrid methods, for example, researchers merged an SVM and a GRU-RNN in [60].

3. Dataset

In the medical field, in particular, regarding specific problems such as the one presented in this paper, data is not always freely available or available at all. This is mostly due to privacy concerns [61, 62]. Another important reason, which is also related in some ways to privacy [62], lies in the overall low level of digitization of healthcare information [63]; in fact, according to Gopal G. et al. [64], healthcare has the lowest level of digital innovation compared to other industries, such as media, finance, insurance, and retail, contributing to limited growth of labor productivity. In addition to this, it is also worth noting that not every dataset containing the desired medical information is also in the desired format, in which case the only remaining option is to create an entirely new dataset from scratch, that is what we did.

3.1. Data Collection

The process of collecting audio data is a pivotal phase in this research. For our dataset, we aimed to collect a sufficient amount of pure, non-multimodal, audio data in a waveform representation. Audio data can be stored in various formats, each with its characteristics, trade-offs, and use cases. Common audio formats include Waveform Audio File Format (WAV), MPEG-1 Audio Layer 3 (MP3), Free Lossless Audio Codec (FLAC), and more. These formats differ in terms of compression, quality, and compatibility. For this study, we opt for the WAV format [65], which is an uncompressed audio file format, developed by IBM and Microsoft, that efficiently stores audio data in a waveform representation without any loss of in-

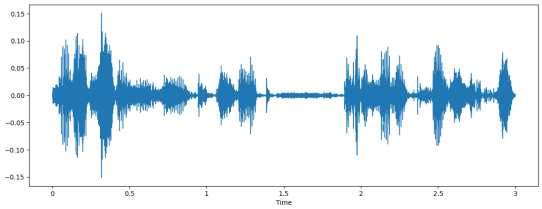

Figure 1: Audio sample from our dataset

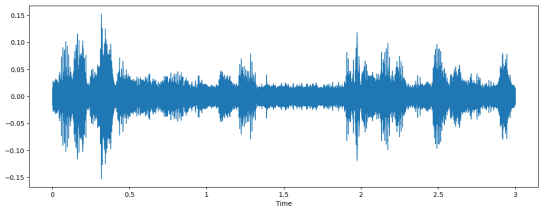

Figure 2: Audio sample with noise from our dataset

formation. Thanks to its characteristics, which guarantee the highest amount of information for an audio signal, WAV is the audio format used as input by wav2vec 2.0 [66], a state-of-the-art speech model developed by the Facebook AI Research group (FAIR) that is one of the models used in this work.

The data collection process began with the identification of audio samples containing English speakers affected by Speech and Language Impairment (SLI) originating from different conditions. This diverse dataset was intentionally curated to optimize the performance of SLI detection. By including speakers with a range of impairments, the model is exposed to a broad spectrum of speech patterns and anomalies, thereby enhancing its ability to accurately detect SLI in real-world applications. To source such data, we turned to YouTube, a vast and user-friendly repository of video and audio content. The videos found were then converted into audio files in WAV format using an online converter.

We finally paired the collected data with a subset of the LibriSpeech dataset [67] containing healthy English speakers only.

3.2. Data Preprocessing

To feed the waveform signals to the model, we needed to ensure that they were appropriately prepared and processed. Effective data preprocessing is fundamental to enhancing the model's performance, as it directly impacts the model's ability to extract meaningful patterns and insights from raw input data. This was performed in different steps. **Firstly** we identified different time windows from each audio file to cut out unnecessary information from them, keeping just human speech sounds (with or without background noise). After that, we analyzed the time windows by dividing each one of them into smaller ones containing the speech of one single person each. Even if there exist different tools available to detect human speech, considering the scarcity of data we suffer, we decided to perform this step manually to be sure that the quality of our dataset is not affected.

Table 1
Dataset samples

	Train		Test	
	SLI	Healthy	SLI	Healthy
Non-augmented	1010	1010	124	125
Time-shifted	893	1010	104	125
Time-stretched	1010	1010	124	125
Pitch-shifted	2020	2020	248	250
Noise-addition	1010	1010	124	125
Total	5943	6060	724	750
Dataset	**12003**		**1474**	

Secondly we split the time windows that we obtained in 3-second clips. We chose this length as a trade-off between a sufficient length, to capture fluency information and a brief duration. Our decision was also based on the standard approach used in the state-of-the-art working with wav2vec 2.0 in these kinds of tasks [68, 69]. Then these clips were saved in two different subsets, creating the Train and the Test set, ensuring that the same speakers do not overlap in both datasets.

Finally the acquired data was augmented to increase its dimension. We applied the following audio augmentations techniques: *Time shifting*, *Time stretching*, *Pitch shifting*, and *Noise addition*, using Gaussian noise. To do so, we used the python library *audiomentations* [70]. For *Time shifting* we resampled the time windows shifting the starting time further by 1.5 seconds; For *Time stretching* we slowed down the speed of the audios by a ratio of 0.8; For the *Pitch shifting* we both lowered and raised the pitch tone by a value of 3, obtaining for each clip two additional ones; Finally for the *Noise addition*, we added a $0.01m$ amplitude Gaussian noise. Audio waveforms before and after noise addition are shown in Fig. 1 and Fig. 2. All the augmentation techniques were applied on the original audio; *Time shifting* was directly applied on the time windows, while the other ones on the initial 3 seconds clips.

The number of samples in the created dataset is shown in Table 1, while in Table 2 we collect the audio data augmentation techniques used and their respective parameters.

Table 2
Pamateres used for augmentation methods

Augmentations	Parameters
Time-shifting	shift = +1.5 seconds
Time-stretching	ratio = 0.8
Pitch-shifting	shift = ± 3 tones
Noise-addition	amplitude = 0.01 meters

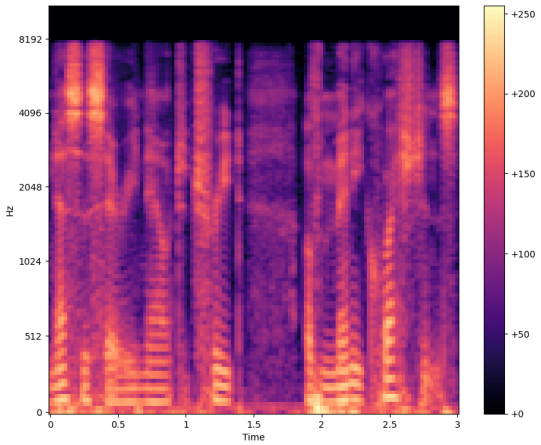

Figure 3: Log Mel Spectrogram of a sample from our dataset

3.3. Data Management

The dataset contains audio files in the WAV format, its data is affected not only by its advantages but also by its drawbacks. The complete dataset, which comprehends both original and augmented data, was too large to be loaded in an online manner using the original files. To overcome this problem we loaded the data in batches and concatenated them in subsets that were saved in the *.arrow* format [71], a columnar memory format for flat and hierarchical data, organized for efficient analytic operations. In this way, large data can be saved, loaded, and processed avoiding memory usage problems.

4. Models and Techniques Used

The best way to approach a problem is to know deeply every factor that influences it and how the key components work, after that, one can tackle it and try to capture its essence with the maximum capabilities. In the following subsections, we present a brief description of the techniques we used and the models we implemented.

4.1. Log Mel Spectrogram

The way humans hear frequencies in sound is known as *pitch*, it is a subjective impression of the frequency. They do not perceive frequencies linearly, on the contrary, humans are more sensitive to differences between lower frequencies than higher ones. For example, the difference between audios of frequency $100Hz$ and $200Hz$ is way bigger than $1000Hz$ and $1100Hz$, even though the absolute difference is the same amount. Humans perceive sounds on a logarithmic scale rather than a linear scale. The Mel Scale [72] was developed to take this into account by conducting experiments with a large number of listeners. It is a scale of pitches, such that each unit is judged by listeners to be equal in pitch distance from the next. The human perception of the amplitude of a sound is called *loudness*, similarly to frequency, also loudness is heard logarithmically rather than linearly. The Decibel scale is used to measure the loudness of a sound, for example, a sound with an amplitude of $20Db$ is 10 times louder than one with an amplitude of $10Db$. We can see that, to deal with sound realistically, we need to use a logarithmic scale via the Mel Scale and the Decibel Scale when dealing with Frequencies and Amplitudes in our data. Spectrograms are generated from sound signals using Fourier Transforms. A Fourier Transform (FT) [73] is a mathematical formula that allows us to decompose the signal into its constituent frequencies and displays the amplitude of each frequency present in the signal. Spectrograms are generated from sound signals using FTs. In other words, an FT converts the signal from the time domain into the frequency domain, and the result is called a spectrum. A *spectrogram* consists in dividing the sound signal into smaller time segments, then applying the FT to each segment, and finally, the combination of these segments in a single plot is called *spectrogram*. A Mel Spectrogram makes two important changes relative to a regular spectrogram that plots frequency vs time: it uses the Mel scale instead of frequency on the y-axis and uses the Decibel scale instead of amplitude to indicate color. In Fig. 3 we can see a normalized version of the Mel spectrogram of one of the audios present in the dataset.

4.2. Wav2vec 2.0

Wav2vec 2.0 [66] is an exceptional tool that learns powerful representations from speech mimicking the human learning experience. People start, in fact, since the early stages of their lives comprehending language without labeled data, i.e. kids learn from listening to adults around them. It is also able to outperform state-of-the-art models while using 100 times less labeled data, thus demonstrating the feasibility of training without huge amounts of labeled data which is very hard to achieve in a field dealing with a complex medium such as audio.

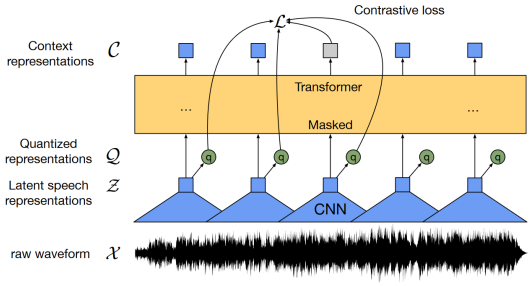

Figure 4: Wav2vec 2.0 pipeline

The model can be visualized in Fig. 4 and next, we will describe its components.

Multi-layer convolutional feature encoder. It consists of several blocks containing a temporal convolution followed by layer normalization and a GELU activation function.

Context network. It follows the Transformer architecture, differently from a normal Transformer that uses fixed positional embeddings, a convolutional layer is used instead, and it acts as a relative positional embedding. The output of the convolution followed by a GELU is added to the inputs and then a layer normalization is applied.

Quantization module. It discretizes the output of the feature encoder to a finite set of speech representations via product quantization. Product quantization amounts to choosing quantized representations from multiple codebooks and concatenating them. The Gumbel softmax enables choosing discrete codebook entries in a fully differentiable way.

The feature encoder $f : X \rightarrow Z$ takes as input the raw waveform X and outputs the latent speech representations $z_1, ..., Z_t$ for T time steps, then they are fed to the transformer $g : Z \rightarrow C$ that captures information from the entire sequence and outputs context representations. The output of the feature encoder is also discretized to q_t with a quantization module to represent the targets in the self-supervised objective. During the model's pre-training a part of the latent speech representations that are generated from the feature encoder are masked, and then the model learns the representations of speech audio by solving a contrastive task, which requires identifying the true quantized latent speech representation for a masked time step within a set of distractors. After pre-training on unlabeled speech, the model is fine-tuned on labeled data with a Connectionist Temporal Classification (CTC) loss.

4.3. Classification Methods

Classification is the part that stands out the most in an entire model because it outputs the labels that are used to compute the evaluation metrics, even though it is the most noticeable part of a model, in our case they are just the final piece of the puzzle since most of the work is done in the previous steps of the pipeline; still, we want to pay some attention to the type of classifiers we used in our work.

Support Vector Machine (SVM) [74] is one of the first algorithms learned by every ML expert, it is simple yet it can achieve excellent results, especially with small amounts of data where other ML algorithms tend to have some difficulties. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space ($N-$ the number of features) that distinctly classifies the data points. To separate the two classes of data points, many possible hyperplanes could be chosen. SVM finds a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. The biggest difficulty encountered when testing the SVM is that even with low amounts of data the model had memory issues, since audio features are extremely large and with multiple classes, while SVM excels with data that has fewer classes, thus making it hard to fully exploit SVM's strengths.

One of the best and most efficient methods to generate labels from an ML model is adding a linear layer at the end of the pipeline, that is what we did with our *wav2vec 2.0* feature extractor, we have included a linear classifier $f(x_i, W, b) = W \cdot x_i + b$ and we trained its weights to output two types of labels, one for people affected by a SLI and one for the others.

Resnet34 is a very famous residual neural network that was pre-trained on ImageNet-1k and was released by Microsoft [75], thanks to residual learning and skip connections this type of model can be much deeper than normal convolutional neural networks. We decided to fine-tune this model with the features extracted with the log mel spectrogram from our dataset.

5. Results

In this section, we will describe the different architectures that we tested in detail and then we will comment on the obtained results.

5.1. Architectures

Our first approach was to use the **wav2vec 2.0** model, in particular the pre-trained *wav2vec2-base* model from HuggingFace [76], to perform Feature Extraction on the

pre-processed non-augmented dataset and then use a **SVM**, the Support Vector Classifier (SVC) model from scikit learn [77], to perform the classification process taking the extracted features in input. As it was explained in the previous section 4, wav2vec2.0 takes a raw waveform signal as input, 3 seconds clips in WAV format in our case, then extracts audio features from them following what it had learned in its previous training. The extracted features were then standardized using the *StandardScaler* from scikit learn, removing the mean and scaling them to unit variance. The standardization of a dataset is a common requirement for many ML estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance). Finally, we fitted the SVM using a linear kernel.

Using the SVM model as a classifier was our first attempt to cope with the limited number of samples at our disposal. Once the dataset was augmented we ceased to use the SVM due to its intrinsic limitations at working with large datasets; so we opted for a complete DL approach.

For our second architecture, we substituted the classifier head with a simple Fully Connected (**FC**), or linear, layer, keeping the **wav2vec 2.0** model to perform the Feature Extraction, this time, on the augmented dataset. We trained the model for 5 epochs through the *Trainer* class by HuggingFace on a batch of 32 samples each, setting the learning rate to $2e-5$ after a warm-up period at a ratio of 0.1 and decreasing its value linearly till the end of the training.

The last architecture tested was a **CNN**, more precisely *resnet34*, that received as input the **log mel spectrogram** of the audios and generated as output the labels of the given audio. All the procedures to extract the spectrogram were carried on with the *librosa* library, firstly the sample was resampled with a new rate of 22050, then the mel spectrogram extracted was normalized and finally scaled. Regarding the CNN, only the last layer was modified, it was replaced with a linear layer that had two output channels and the whole model was fine-tuned without freezing the previous layers. Training was carried out for 50 epochs, the learning rate started at $2e-4$ and decayed by a factor of 10 every 10 epochs; the loss function used was the *CrossEntropyLoss*. All parameters used to compute the spectrum are shown in Table 3.

Table 3
Parameters used to compute the Spectrogram

Log Mel Spectrum Parameters	
Sample rate	22050
Windows length	2048
Hop length	512
N mels	128

Table 4
Architectures Accuracy

Models	Accuracy
LASSO (Full Model) [78]	0.84
1NN CHI Strategy [79]	0.8832
LMT BL Strategy [79]	0.9269
MLP BL Strategy [79]	0.9013
NB BL Strategy [79]	0.9269
CNN [80]	0.8421
Our Models	**Accuracy**
Wav2vec2.0 + SVM	0.6627
Wav2vec2.0 + FC	**0.9661**
Log Mel Spectrogram + CNN	0.9362

5.2. Evaluations

In Table 4 we show the accuracy of our architectures, compared with others architectures [78] As we can see, the first model is the one with the lowest score. This means that, despite the ability of the SVM to avoid overfitting on the poor quantity of data provided, it cannot accurately detect the speakers affected by SLI. This is probably due to the magnitude of the feature space extracted by the wav2vec 2.0 model.

Using, instead, an augmented dataset together within a DL approach we manage to reach a very high value of accuracy, the highest of our models. The wav2vec 2.0 feature extractor, having enough data to work with, managed to extract the key features and information needed to correctly identify which voice belongs to a healthy speaker or an impaired one.

The CNN model that was fine-tuned with Log Mel Spectrum features achieved great accuracy in labeling samples, unfortunately, through a more accurate analysis of the confusion matrices shown in 5, 6, and 7, we discovered that the number of false negatives is extremely high compared to the false positives. In the medical field, especially for tools helping with diagnosis, it is crucial to have the smallest number of false negatives, since an undetected disease is much worse than a false positive, medical operators could be missing a lot of vital anomalies and in time they will lose trust in the system. In our case *recall* is way more important than the *precision* score, from Table 5 we can see that the CNN model reaches only a recall score of 0.85, on the other hand wav2vec 2.0 achieves a better recall and F1 Score.

6. Limitations and Future Works

It is of critical importance to examine our achievements and acknowledge the constraints that affect our work.

Table 5
Architectures overall performances

Model	F1 Score	Precision	Recall
Wav2vec2.0 + SVM	0.6316	0.6923	0.5806
Wav2vec2.0 + FC	**0.9655**	0.9641	**0.9668**
Spectrogram + CNN	0.9187	**0.9983**	0.8508

While our research has given promising results, the following section delves into the limitations that shape our results and sets a base for future possible improvements.

6.1. Limitations

The lack of data quantity and quality is one of our major constraints. The problem of data scarcity has already been addressed in section 3 so we will now talk about *quality*.

In the realm of ML and DL, it has been well documented that the issue of low-quality data and disparities in data collection methodologies exacerbate the inherent biases within the data when utilized for training algorithms, a clear example is given by the societal or political biases reflected in word embeddings or large language models [81, 82]. This concern arises when the data collected for training purposes exhibits significant variations in quality and collection techniques, resulting in a heightened vulnerability to intrinsic biases within the data. Such biases can subsequently propagate through the training process, influencing the performance and fairness of ML and DL algorithms leading to further disparities and discrimination in the real world, due to the accessibility to such tools [83, 84]. Particularly, in our work, the collection of English speakers affected by SLI presents the limitation of containing mostly speakers with American accents. In real-world applications this can have negative effects on the model performance, for example, the algorithm could achieve higher and better results with American people rather than with Mexican ones, or other English-speaking minority ethnic groups of people whose accent differs from the standard American one [84].

Another limitation of our dataset is that it does not contain children speakers. This is because finding such materials on the web is often difficult, and it is more difficult to create them from scratch due to the small number of certified children affected by SLI and, since they are minors, due to more strict privacy concerns. The most used dataset in this field [85] consists of one second clips of Czech speaking children, both healthy or affected by SLI. Although this dataset could be useful for the detection of SLI, it is limited to the Czech language and children speakers. This kind of limitation is common in the healthcare field, especially in SLI detection.

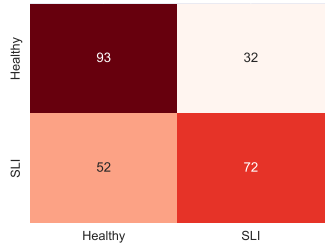

Figure 5: Wav2vec 2.0 + SVM confusion matrix

6.2. Future Works

Future works should focus on the creation of a new dataset comprising people speaking different languages, since it is not yet known, to our knowledge, whether fluency problems can be generalized in all languages and a wide age range, knowing that the features and the overall characteristics of the voice between children and adults change in general, due to their anatomical differences [86].

Given the technological advancement in the field of generative audio with astonishing tools such as the audio manipulation software produced by ElevenLabs [87], which can clone voices, generate new ones, translate them into other languages, and make them read texts, new kinds of audio enhancement can be experimented with, and although they cannot be used now, because they cannot replicate stuttering or other kinds of fluency features that characterize people affected with SLI yet, they are promising tools to take into consideration for the near future.

7. Conclusions

This work proposes a novel approach to Speech and Language Impairment (SLI) detection, based solely on audio and AI audio-based techniques, together within an entirely new dataset composed of English speakers affected by SLI. The results show that, even with some limitations related to the scarcity of data available, Deep Learning methods can achieve accurate estimations on healthy or impaired speakers. In particular, wav2vec 2.0, with a Fully Connected layer as the classification head, reaches an accuracy of over 96% on our test set. Our findings also confirm that data audio augmentation techniques are fundamental to training Deep Learning models adequately.

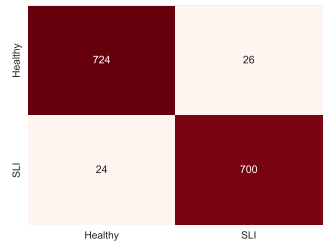

Figure 6: Wav2vec 2.0 + FC confusion matrix

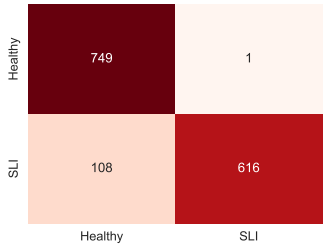

Figure 7: Log Mel Spectrogram + CNN confusion matrix

References

[1] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, C. Lemey, Machine learning and natural language processing in mental health: Systematic review, J Med Internet Res 23 (2021) e15708.

[2] D. I. Patrício, R. Rieder, Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review, Computers and Electronics in Agriculture 153 (2018) 69–81. URL: https://www.sciencedirect.com/science/article/pii/S0168169918305829. doi:https://doi.org/10.1016/j.compag.2018.08.001.

[3] I. E. Tibermacine, A. Tibermacine, W. Guettala, C. Napoli, S. Russo, Enhancing sentiment analysis on seed-iv dataset with vision transformers: A comparative study, in: ACM International Conference Proceeding Series, 2023, p. 238 – 246. doi:10.1145/3638985.3639024.

[4] S. B. Scruggs, K. Watson, A. I. Su, H. Hermjakob, J. R. Yates, M. L. Lindsey, P. Ping, Harnessing the heart of big data, Circulation Research 116 (2015) 1115 – 1119. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84930702552&doi=10.1161%2fCIRCRESAHA.115.306013&partnerID=40&md5=4244ed52d2f51fa5c08f02ca67e4103e. doi:10.1161/CIRCRESAHA.115.306013, cited by: 44; All Open Access, Bronze Open Access, Green Open Access.

[5] E. Iacobelli, V. Ponzi, S. Russo, C. Napoli, Eye-tracking system with low-end hardware: Development and evaluation, Information (Switzerland) 14 (2023). doi:10.3390/info14120644.

[6] M. Woźniak, D. Połap, R. K. Nowicki, C. Napoli, G. Pappalardo, E. Tramontana, Novel approach toward medical signals classifier, in: Proceedings of the International Joint Conference on Neural Networks, volume September 2015, 2015. doi:10.1109/IJCNN.2015.7280556.

[7] P. Zinemanas, M. Rocamora, M. Miron, F. Font, X. Serra, An interpretable deep learning model for automatic sound classification, Electronics 10 (2021). URL: https://www.mdpi.com/2079-9292/10/7/850. doi:10.3390/electronics10070850.

[8] M. Azimi, U. Roedig, Room Identification with Personal Voice Assistants (Extended Abstract), 2022, pp. 317–327. doi:10.1007/978-3-030-95484-0_19.

[9] J. Kapočiūtė-Dzikienė, A domain-specific generative chatbot trained from little data, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/7/2221. doi:10.3390/app10072221.

[10] S. Shah, Z. Tariq, Y. Lee, Audio iot analytics for home automation safety, 2018, pp. 5181–5186. doi:10.1109/BigData.2018.8622587.

[11] F. Fiani, S. Russo, C. Napoli, An advanced solution based on machine learning for remote emdr therapy, Technologies 11 (2023). doi:10.3390/technologies11060172.

[12] S. Gholizadeh, Z. Leman, B. T. Baharudin, A review of the application of acoustic emission technique in engineering, Structural Engineering and Mechanics 54 (2015) 1075–1095. doi:10.12989/sem.2015.54.6.1075.

[13] H. Lozano, I. Hernáez, A. Picón, J. Camarena, E. Navas, Audio classification techniques in home environments for elderly/dependant people, in: K. Miesenberger, J. Klaus, W. Zagler, A. Karshmer (Eds.), Computers Helping People with Special Needs, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 320–323.

[14] D. T. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, A. N. G. Kirschel, Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus, Journal of Applied Ecology 48 (2011) 758–767. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.

1111/j.1365-2664.2011.01993.x. doi:https://doi.org/10.1111/j.1365-2664.2011.01993.x.

[15] D. V. M. Bishop, Uncommon understanding: Development and disorders of language comprehension in children, Psychology Press/Erlbaum (UK) Taylor and Francis, 1997.

[16] H. CLAHSEN, The grammatical characterization of developmental dysphasia 27 (1989) 897–920. URL: https://doi.org/10.1515/ling.1989.27.5.897. doi:doi:10.1515/ling.1989.27.5.897.

[17] M. L. Rice, K. Wexler, P. L. Cleave, Specific language impairment as a period of extended optional infinitive, Journal of Speech, Language, and Hearing Research 38 (1995) 850–863. URL: https://pubs.asha.org/doi/abs/10.1044/jshr.3804.850. doi:10.1044/jshr.3804.850.

[18] H. K. van der Lely, Domain-specific cognitive systems: insight from grammatical-sli, Trends in Cognitive Sciences 9 (2005) 53–59. URL: https://doi.org/10.1016/j.tics.2004.12.002. doi:10.1016/j.tics.2004.12.002.

[19] D. V. M. Bishop, Ten questions about terminology for children with unexplained language problems, Int. J. Lang. Commun. Disord. 49 (2014) 381–415.

[20] J. H. Beitchman, B. Wilson, E. B. Brownlie, H. Walters, A. Inglis, W. Lancee, Long-term consistency in speech/language profiles: II. behavioral, emotional, and social outcomes, J. Am. Acad. Child Adolesc. Psychiatry 35 (1996) 815–825.

[21] T. L. Stanton-Chapman, L. M. Justice, L. E. Skibbe, S. L. Grant, Social and behavioral characteristics of preschoolers with specific language impairment, Topics in Early Childhood Special Education 27 (2007) 98–109. URL: https://doi.org/10.1177/02711214070270020501. doi:10.1177/02711214070270020501.

[22] J. Wagner, D. Schiller, A. Seiderer, E. André, Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?, in: Interspeech, 2018. URL: https://api.semanticscholar.org/CorpusID:52192644.

[23] Y. Tokozume, T. Harada, Learning environmental sounds with end-to-end convolutional neural network, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2721–2725. doi:10.1109/ICASSP.2017.7952651.

[24] J. Lee, J. Park, K. L. Kim, J. Nam, Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification, Applied Sciences 8 (2018). URL: https://www.mdpi.com/2076-3417/8/1/150. doi:10.3390/app8010150.

[25] L. M. Justice, W.-Y. Ahn, J. A. R. Logan, Identifying children with clinical language disorder: An application of machine-learning classification, J. Learn. Disabil. 52 (2019) 351–365.

[26] J. C. Lee, J. B. Tomblin, Reinforcement learning in young adults with developmental language impairment, Brain Lang. 123 (2012) 154–163.

[27] R. A. Ahire, Nitin, A. Wagh, Eeg based identification of learning disabilities using machine learning algorithms, J Neurol Disord (2022).

[28] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, S. Misra, Data augmentation and deep learning methods in sound classification: A systematic review, Electronics 11 (2022). URL: https://www.mdpi.com/2079-9292/11/22/3795. doi:10.3390/electronics11223795.

[29] C. Napoli, G. Pappalardo, E. Tramontana, Using modularity metrics to assist move method refactoring of large systems, in: Proceedings - 2013 7th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS 2013, 2013, p. 529 – 534. doi:10.1109/CISIS.2013.96.

[30] D. Połap, M. Woźniak, C. Napoli, E. Tramontana, Real-time cloud-based game management system via cuckoo search algorithm, International Journal of Electronics and Telecommunications 61 (2015) 333 – 338. doi:10.1515/eletel-2015-0043.

[31] Z. Mushtaq, S.-F. Su, Q.-V. Tran, Spectral images based environmental sound classification using cnn with meaningful data augmentation, Applied Acoustics 172 (2021) 107581. URL: https://www.sciencedirect.com/science/article/pii/S0003682X2030685X. doi:https://doi.org/10.1016/j.apacoust.2020.107581.

[32] M. Woźniak, D. Połap, C. Napoli, E. Tramontana, Graphic object feature extraction system based on cuckoo search algorithm, Expert Systems with Applications 66 (2016) 20 – 31. doi:10.1016/j.eswa.2016.08.068.

[33] S. Chen, E. Dobriban, J. Lee, Invariance reduces variance: Understanding data augmentation in deep learning and beyond, ArXiv abs/1907.10905 (2019). URL: https://api.semanticscholar.org/CorpusID:198895147.

[34] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (2019) 60. URL: https://doi.org/10.1186/s40537-019-0197-0. doi:10.1186/s40537-019-0197-0.

[35] C. Napoli, G. Pappalardo, E. Tramontana, Z. Marszalek, D. Polap, M. Wozniak, Simplified firefly algorithm for 2d image key-points search, in: IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CIHLI 2014: 2014 IEEE Symposium on Computational Intelligence for Human-Like Intelligence, Proceedings, 2014. doi:10.1109/CIHLI.2014.7013395.

[36] A. Greco, N. Petkov, A. Saggese, M. Vento, Aren:

A deep learning approach for sound event recognition using a brain inspired representation, IEEE Transactions on Information Forensics and Security 15 (2020) 3610–3624. doi:10.1109/TIFS.2020.2994740.

[37] D. Połap, M. Woźniak, C. Napoli, E. Tramontana, R. Damaševičius, Is the colony of ants able to recognize graphic objects?, Communications in Computer and Information Science 538 (2015) 376 – 387. doi:10.1007/978-3-319-24770-0_33.

[38] Z. Mushtaq, S.-F. Su, Environmental sound classification using a regularized deep convolutional neural network with data augmentation, Applied Acoustics 167 (2020) 107389. URL: https://www.sciencedirect.com/science/article/pii/S0003682X2030493X. doi:https://doi.org/10.1016/j.apacoust.2020.107389.

[39] O. Novotny, O. Plchot, O. Glembek, J. H. Cernocky, L. Burget, Analysis of dnn speech signal enhancement for robust speaker recognition, Computer Speech and Language 58 (2019) 403–421. URL: https://www.sciencedirect.com/science/article/pii/S0885230818303607. doi:https://doi.org/10.1016/j.csl.2019.06.004.

[40] S. Mertes, A. Baird, D. Schiller, B. W. Schuller, E. André, An evolutionary-based generative approach for audio data augmentation, in: 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), 2020, pp. 1–6. doi:10.1109/MMSP48831.2020.9287156.

[41] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, in: Interspeech 2019, ISCA, 2019. URL: https://doi.org/10.21437%2Finterspeech.2019-2680. doi:10.21437/interspeech.2019-2680.

[42] E. K. Wang, J. Yu, C.-M. Chen, S. Kumari, J. J. P. C. Rodrigues, Data augmentation for internet of things dialog system, Mobile Networks and Applications 27 (2022) 158–171. URL: https://doi.org/10.1007/s11036-020-01638-9. doi:10.1007/s11036-020-01638-9.

[43] T. Koike, K. Qian, B. W. Schuller, Y. Yamamoto, Transferring cross-corpus knowledge: An investigation on data augmentation for heart sound classification, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2021.

[44] V. Basu, S. Rana, Respiratory diseases recognition through respiratory sound with the help of deep neural network, in: 2020 4th International Conference on Computational Intelligence and Networks (CINE), 2020, pp. 1–6. doi:10.1109/CINE48825.2020.234388.

[45] Y. Leng, W. Zhao, C. Lin, C. Sun, R. Wang, Q. Yuan, D. Li, Lda-based data augmentation algorithm for acoustic scene classification, Knowledge-Based Systems 195 (2020) 105600. doi:10.1016/j.knosys.2020.105600.

[46] V. M. Praseetha, P. P. Joby, Speech emotion recognition using data augmentation, International Journal of Speech Technology 25 (2022) 783–792. URL: https://doi.org/10.1007/s10772-021-09883-3. doi:10.1007/s10772-021-09883-3.

[47] S. Lalitha, D. Gupta, M. Zakariah, Y. A. Alotaibi, Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation, Applied Acoustics 170 (2020) 107519. URL: https://www.sciencedirect.com/science/article/pii/S0003682X2030623X. doi:https://doi.org/10.1016/j.apacoust.2020.107519.

[48] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, B. Schuller, A bag-of-audio-words approach for snore sounds' excitation localisation, in: Speech Communication; 12. ITG Symposium, 2016, pp. 1–5.

[49] H. Lachambre, B. Ricaud, G. Stempfel, B. Torrésani, C. Wiesmeyr, D. Onchis-Moaca, Optimal window and lattice in gabor transform. application to audio analysis, in: 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2015, pp. 109–112. doi:10.1109/SYNASC.2015.25.

[50] E. Garcia-Ceja, M. Riegler, A. K. Kvernberg, J. Torresen, User-adaptive models for activity and emotion recognition using deep transfer learning and data augmentation, User Modeling and User-Adapted Interaction 30 (2020) 365–393. URL: https://doi.org/10.1007/s11257-019-09248-1. doi:10.1007/s11257-019-09248-1.

[51] H. Ykhlef, F. Ykhlef, S. Chiboub, Experimental design and analysis of sound event detection systems: Case studies, in: 2019 6th International Conference on Image and Signal Processing and their Applications (ISPA), 2019, pp. 1–6. doi:10.1109/ISPA48434.2019.8966798.

[52] S. Lalitha, D. Gupta, M. Zakariah, Y. A. Alotaibi, Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation, Applied Acoustics 170 (2020) 107519. URL: https://www.sciencedirect.com/science/article/pii/S0003682X2030623X. doi:https://doi.org/10.1016/j.apacoust.2020.107519.

[53] V.-T. Tran, W.-H. Tsai, Stethoscope-sensed speech and breath-sounds for person identification with sparse training data, IEEE Sensors Journal 20 (2020) 848–859. doi:10.1109/JSEN.2019.2945364.

[54] T. A. M. Celin, T. Nagarajan, P. Vijayalakshmi, Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition, IEEE Journal of Selected Topics in Signal Processing 14 (2020) 346–354. doi:10.1109/JSTSP.2020.2972161.

[55] J. Ye, T. Kobayashi, M. Murakawa, Urban sound event classification based on local and global features aggregation, Applied Acoustics 117 (2017) 246–256. URL: https://www.sciencedirect.com/science/article/pii/S0003682X16302274. doi:https://doi.org/10.1016/j.apacoust.2016.08.002, acoustics in Smart Cities.

[56] V. Ramesh, K. Vatanparvar, E. Nemati, V. Nathan, M. M. Rahman, J. Kuang, CoughGAN: Generating synthetic coughs that improve respiratory disease classification(), Annu Int Conf IEEE Eng Med Biol Soc 2020 (2020) 5682–5688.

[57] N. Yella, B. Rajan, Data augmentation using gan for sound based covid 19 diagnosis, in: 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), volume 2, 2021, pp. 606–609. doi:10.1109/IDAACS53288.2021.9660990.

[58] H. Lee, J. Lee, Neural network prediction of sound quality via domain knowledge-based data augmentation and bayesian approach with small data sets, Mechanical Systems and Signal Processing 157 (2021) 107713. URL: https://www.sciencedirect.com/science/article/pii/S0888327021001084. doi:https://doi.org/10.1016/j.ymssp.2021.107713.

[59] D. Koszewski, B. Kostek, Musical instrument tagging using data augmentation and effective noisy data processing, Journal of the Audio Engineering Society 68 (2020) 57–65. doi:10.17743/jaes.2019.0050.

[60] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, B. Schuller, Snore-GANs: Improving automatic snore sound classification with synthesized data, IEEE J Biomed Health Inform 24 (2019) 300–310.

[61] K. P. Seastedt, P. Schwab, Z. O'Brien, E. Wakida, K. Herrera, P. G. F. Marcelo, L. Agha-Mir-Salim, X. B. Frigola, E. B. Ndulue, A. Marcelo, L. A. Celi, Global healthcare fairness: We should be sharing more, not less, data, PLOS Digital Health 1 (2022) 1–13. URL: https://doi.org/10.1371/journal.pdig.0000102. doi:10.1371/journal.pdig.0000102.

[62] M. Paul, L. Maglaras, M. A. Ferrag, I. Almomani, Digitization of healthcare sector: A study on privacy and security concerns, ICT Express 9 (2023) 571–588. URL: https://www.sciencedirect.com/science/article/pii/S2405959523000243. doi:https://doi.org/10.1016/j.icte.2023.02.007.

[63] T. M. Stoumpos AI, Kitsios F, Digital transformation in healthcare: Technology acceptance and its applications, Int J Environ Res Public Health (2023). doi:10.3390/ijerph20043407.

[64] G. Gopal, C. Suter-Crazzolara, L. Toldo, W. Eberhardt, Digital transformation in healthcare – architectures of present and future information technologies, Clinical Chemistry and Laboratory Medicine (CCLM) 57 (2019) 328–335. URL: https://doi.org/10.1515/cclm-2018-0658. doi:doi:10.1515/cclm-2018-0658.

[65] Wave file format specification, ???? URL: https://www.mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/WAVE.html.

[66] A. Baevski, H. Zhou, A. Mohamed, M. Auli, Wav2vec 2.0: A framework for self-supervised learning of speech representations, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[67] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210. doi:10.1109/ICASSP.2015.7178964.

[68] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, M. Kurimo, Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering, in: Proceedings of the 30th ACM International Conference on Multimedia, MM '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 7026–7029. URL: https://doi.org/10.1145/3503161.3551572. doi:10.1145/3503161.3551572.

[69] J. Liu, A. Wumaier, D. Wei, S. Guo, Automatic speech disfluency detection using wav2vec2. 0 for different languages with variable lengths, Applied Sciences 13 (2023) 7579.

[70] I. Jordal, A. Tamazian, E. T. Chourdakis, Angonin, askskro, N. Karpov, T. Dhyani, O. Sarioglu, kvilouras, E. Berk, F. Mirus, J.-Y. Lee, K. Choi, MarvinLvn, SolomidHero, T. Alum, iver56/audiomentations: v0.33.0 (????). doi:10.5281/zenodo.7010042.

[71] N. Richardson, I. Cook, N. Crane, D. Dunnington, R. François, J. Keane, D. Moldovan-Grünfeld, J. Ooms, Apache Arrow, arrow: Integration to 'Apache' 'Arrow', 2023. Https://github.com/apache/arrow/, https://arrow.apache.org/docs/r/.

[72] B. Truax, Handbook for acoustic ecology, Leonardo 13 (1980) 83.

[73] R. N. Bracewell, R. N. Bracewell, The Fourier transform and its applications, volume 31999, McGraw-Hill New York, 1986.

[74] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.
[75] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.
[76] T. W. Clément Delangue, Julien Chaumond, Wav2vec2, 2022. URL: https://huggingface.co/docs/transformers/model_doc/wav2vec2.
[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[78] L. M. Justice, W.-Y. Ahn, J. A. Logan, Identifying children with clinical language disorder: an application of machine-learning classification, Journal of learning disabilities 52 (2019) 351–365.
[79] J. Gaspers, K. Thiele, P. Cimiano, A. Foltz, P. Stenneken, M. Tscherepanow, An evaluation of measures to dissociate language and communication disorders from healthy controls using machine learning techniques, in: Proceedings of the 2nd acm sighit international health informatics symposium, 2012, pp. 209–218.
[80] C. Kanimozhiselvi, S. Santhiya, Communication disorder identification from recorded speech using machine learning assisted mobile application, in: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, 2021, pp. 789–793.
[81] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.
[82] D. Rozado, The political biases of ChatGPT, Soc. Sci. (Basel) 12 (2023) 148.
[83] M. A. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk, Potential biases in machine learning algorithms using electronic health record data, JAMA Intern. Med. 178 (2018) 1544–1547.
[84] H. Ibrahim, X. Liu, N. Zariffa, A. D. Morris, A. K. Denniston, Health data poverty: an assailable barrier to equitable digital health care, The Lancet Digital Health 3 (2021) e260–e265. URL: https://www.sciencedirect.com/science/article/pii/S2589750020303174. doi:https://doi.org/10.1016/S2589-7500(20)30317-4.
[85] P. Grill, J. Tučková, Speech databases of typical children and children with SLI, PLoS One 11 (2016) e0150365.
[86] A. McAllister, P. Sjölander, Children's voice and voice disorders, in: Seminars in speech and language, volume 34, Thieme Medical Publishers, 2013, pp. 071–079.
[87] M. S. Piotr Dabkowski, Eleven labs, 2023. URL: https://elevenlabs.io/voice-lab.