Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU34282

Robust Voice-Based Laryngeal Cancer Diagnosis via Machine Learning

Yichen Jiao

Browning School, New York, USA
lucas.jiao.ny@gmail.com

Abstract. Laryngeal cancer is traditionally diagnosed using advanced clinical methods, which are both uncomfortable and inaccessible to patients with limited healthcare access. Recently, using Machine Learning (ML) based methods to classify voice data has been considered as a friendly alternative for laryngeal cancer diagnosis. However, existing models often suffer from scarce training data and thin decision boundaries, limiting the achievable diagnosis accuracy. Additionally, they used closed-source data and code, limiting experimental reproducibility. To mitigate these issues, we present a open-source ML and voice-based framework for laryngeal cancer detection. The framework includes a weighted data sampler, a fine-tuned audio feature extractor, and a metric learning loss function to discriminate distinct audio samples. It's trained and tested on the Far Eastern Memorial Hospital (FEMH) and Saarbrücken Voice Datasets (SVD). Furthermore, we used model ensemble to boost the diagnose accuracy. Our results show that fine-tuning the feature extractor and class weighting can substantially improve the overall performance: our best model achieved 85.98% for the balanced accuracy compared to 61.28% from the SVM-based baseline model. These results suggest that carefully designed feature learning and data preprocessing substantially improve the feasibility of voice-based laryngeal cancer screening.

Keywords: laryngeal cancer, voice analysis, machine learning, feature extraction, model ensemble

1. Introduction

Laryngeal cancer is a type of cancer that affects the voice box (larynx). The National Institute of Health (NIH) reports that it is among the most prevalent head and neck cancers, with alarming mortality rates, especially in regions with limited access to surgical care, [1]. Over 188,000 people were diagnosed with laryngeal cancer in 2022, and cases are expected to rise by 66% in 2050 [2]. Early diagnosis improves survival rates, treatment options, and mitigates long-term impairments, making early detection a crucial task. Currently, diagnosis involves techniques such as Laryngoscopy or Computerized Tomography (CT) scans, performed by trained specialists using professional equipment [3]. Although clinically effective, these methods directly interfere with the larynx and need to be performed in special settings, making the treatment uncomfortable for patients and inaccessible for people with limited healthcare access.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU34282

Recently, voice-based diagnosis for laryngeal cancer has been considered as an alternative to the above-mentioned traditional methods. Because Laryngeal cancer alters vocal vibration and airflow, a patient's voice carries useful signals for diagnosis. Hence, by leveraging Machine Learning (ML) and pathological voice datasets, we can train ML models to detect laryngeal cancer [4]. This approach reduces the need for complex medical tests, presenting a sustainable, non-invasive alternative for diagnosing laryngeal cancer.

Nonetheless, voice-based diagnosis has several drawbacks that limit its actual applications in clinical settings. First, experimental reproducibility and external validation are limited: most existing research papers use private datasets and do not open-source their models [5]. Even among the few publicly available voice datasets that contain patients with laryngeal cancer, their data distributions are long-tailed [4]; an overwhelming majority of samples are classified as benign, while only a few are malignant in comparison. This imbalanced class distribution presents great challenges for both training and evaluating ML models. Additionally, it is also difficult to diagnose laryngeal cancer from voice data alone. Laryngeal cancer alters an individual's vocal cords, making their voice sound very hoarse. However, other conditions like Reinke's Edema [6], classified as benign, do the same. Due to their acoustic similarities, voice signals alone are often insufficient for ML models to differentiate between cancerous cases and benign voice conditions [7].

To address these challenges, we propose a robust and open-source[1] framework for non-invasive laryngeal cancer diagnosis. Our models are evaluated on publicly available voice datasets and we report our findings using balanced accuracy, precision, and recall score. Our contributions to this field include (1) a reproducible, open-source pipeline for laryngeal cancer diagnosis, (2) a custom sampling strategy to adapt to imbalanced pathological voice data, and (3) fine-tuned feature extractor representations rather than frozen features, and (4) ensemble methods for robust predictions. Through detailed ablation studies, we further show that proper data handling and fine-tuning contribute more to the model's performance than metric learning approaches.

2. Framework

Our framework consists of three primary parts, as shown in Fig.1: we first pre-process and augment our data for a robust training process. These signals are then processed through a fine-tuned variant of Wav2Vec 2.0 [8]. Finally, the output embeddings are used for a custom loss function to discriminate different classes.

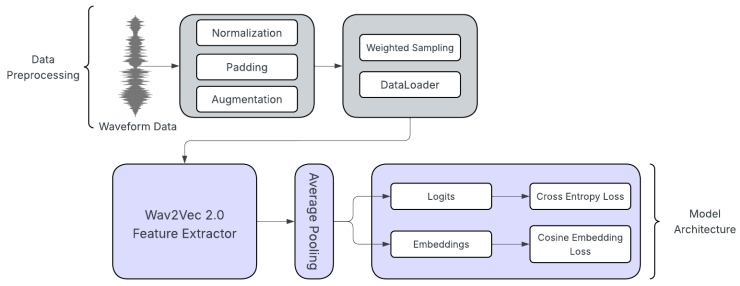

Figure 1. The Flowchart of the proposed framework

[1] Our code is available here: https://github.com/lLukii/Non-Invasive-Laryngeal-Cancer/tree/main/src

DOI: 10.54254/2755-2721/2026.GU34282

2.1. Wav2Vec 2.0

Wav2Vec 2.0 is a self-supervised speech representation model that creates contextualized audio representations. Although originally designed for Automatic Speech Recognition (ASR), several studies show that the model works well on classification tasks, including voice-disorder detection, obtaining results that outperforms previous feature extraction methods and models [9-12]. An illustration of the Wav2Vec 2.0 architecture is presented in Fig. 2

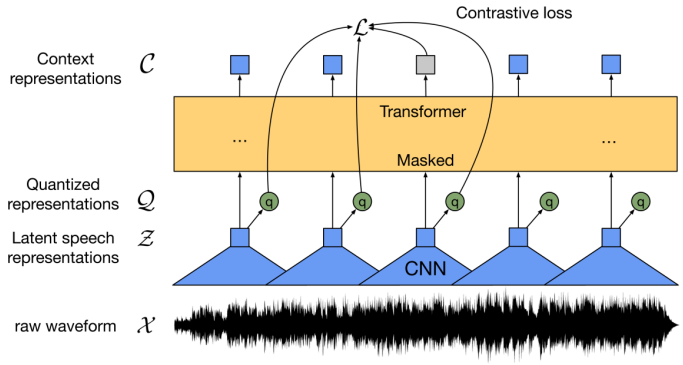

Figure 2. Illustration of the Wav2Vec 2.0 pre-training process

We propose to build a network that fine-tunes from the Wav2Vec 2.0 architecture, further tweaking its parameters to adapt to the specific patterns of pathological voices rather than being an all-purpose model. Our pre-trained model is Facebook's wav2vec2-xlsr-53 variant, which we choose for its ability to learn cross-lingual speech representations. From the model we extract the final contextual embeddings, aggregate them with average pooling, and pass the resulting vector through a linear head to predict class probabilities.

2.2. Data augmentation

Recordings under clinical conditions are likely to be affected by background noise, which could affect the model's predictions. To address these issues, we randomly augment our training data using three tools so that the model can adapt to background anomalies. Specifically, we shift the volume of our recordings by a scale between 0.8 and 1.2, add pitch shift between -3 and 3 steps, and perturb the speed of the recording by a factor of 0.9 to 1.1. Each augmentation is applied to a sample, with specific values chosen at random from our desired ranges.

Since Wav2Vec takes in raw waveform, we refrain from applying any frequency-domain augmentations such as frequency masking, as doing so would require us to convert our signal into a spectrogram. Though the Griffin-Lim algorithm can revert spectrograms back into waveform data, it only provides a loose estimation. We fear that doing so would harmfully distort our audio data beyond realistic scenarios, motivating us to stick with time-domain data augmentations.

2.3. Weighted sampling

To mitigate the class imbalance described in Section 1, we reshape our training data using a customized sampler rather than relying on pure random sampling. With naive random sampling, many mini-batches would contain no positive cases, while others could contain several, causing training to be very noisy.

DOI: 10.54254/2755-2721/2026.GU34282

To ensure consistent exposure to positive samples during training, our sampler ensures that exactly one positive case is present in each batch: for each batch, we draw an index from the pool of positive data, and fill the rest of the batch with negative samples. Adding one case preserves the imbalanced nature of our data while guaranteeing that our model receives a positive signal every batch. We found that a single positive example works best for us, as adding too many cases increases the likelihood of overfitting and complicates our loss function design.

2.4. The loss function

In this study, we propose applying a combination of weighted cross-entropy (CE) loss and cosine embedding loss to train our downstream task. The former is widely used for classification, and will be used on our baseline Wav2Vec model and ablation studies. It penalizes the model when its prediction for the correct class is not confident. We apply a weighted version of CE loss to address the imbalanced nature of our data, as standard CE loss can be reduced in this case by simply predicting negative for each example. The equation for our weighted CE loss is described in [ce_loss].

In Section 1, we note that malignant voice conditions are hard to distinguish from the benign ones. We address this problem by further applying the cosine embedding loss function. By projecting the Wav2Vec outputs into an embedding space, this loss aims to separate the positive case away from each negative sample in every training batch. Doing so implicitly updates the model's parameters to create more distinctive predictions for positive and negative samples, thus easing the training process. The equation for cosine embedding loss is described in [cosine_emb].

$$L_{ce} = -\frac{1}{N}\sum_{i=0}^{N} w_1 y_i \log\left(p_i\right) + w_2\left(1-y_i\right)\log\left(1-p_i\right) \tag{1}$$

$$L_{cos} = \frac{1}{N-1}\sum_{i=2}^{N} \max\left(\cos\left(x, x_i\right) - m, 0\right) \tag{2}$$

In the two equations above, N denotes the size of our batch, y_i represents the label of sample i, and p_i denotes the probability that our model predicts the correct label. w_1 and w_2 are weights denoting the emphasis Cross-Entropy loss should place on the positive and negative samples, respectively. Assume that the label for x is the one positive sample, and the labels for every x_i represent a negative sample in a training batch. $\cos\left(a, b\right) = \frac{a \cdot b}{||a||||b||}$ indicates the cosine similarity of embeddings a and b, and m is a tunable hyperparameter denoting the margin to shift each similarity by. Our final training objective is simply a weighted combination of the two losses, where λ represents a tunable hyperparameter.

$$L = L_{ce} + \lambda L_{cos} \tag{3}$$

2.5. Model ensemble

To further improve the accuracy of our model, we propose an ensemble approach, which aggregates the results of multiple independently trained models of the same architecture described above. At inference time, we combine results using a majority voting algorithm: each model returns a class label, and the label with the most votes becomes the ensemble prediction.

The motivation behind this approach is to mitigate the bias in individual models. Since the downstream parameters for each model are randomly initialized, the training process will likely lead

DOI: 10.54254/2755-2721/2026.GU34282

them to different local minima, even if they are trained on the exact same data. As a result, each model develops slightly different decision boundaries. This adds more perspectives to how our models classify each sample, improving the robustness of our overall prediction. To prevent ties in our predictions, we use an odd number of models.

3. Experiments

3.1. Datasets

We train and test our models on the Far Eastern Memorial Hospital (FEMH) and Saarbrücken Voice Datasets (SVD) [13, 14], two pathological voice datasets containing many disorders from roughly 4000 patients. In each recording, the patient sustains the vowel "a" for a few seconds. The FEMH dataset was used in a challenge hosted by IEEE back in 2018, where various research groups competed to build the best-performing voice pathology model. The challenge offered a rigorous and concrete benchmark in which many teams were assessed, allowing our results to be more compatible with existing works.

We also incorporate the Saarbrücken Voice Dataset (SVD), one of the largest pathological voice datasets in the world. It contains over 2700 samples recorded in German, and has been widely used in previous studies [10]. The recordings are stored in the KayPENTAX NSP Audio format, which we converted to waveform signals using a dedicated Python library. We trained and tested our models by combining both datasets, doing so to ensure a larger and more robust collection of patient samples. We further normalized each audio signal between -1 and 1 so that our data fell under a fixed range. Doing so would stabilize our training process, leading to significantly better results.

In the datasets, each sample was categorized as either "malignant" or "benign" (i.e., positive or negative) depending on its voice disorder. The specific criteria was determined by a specialist in the field [4], as indicated by Figures 3 and 4. We randomly partition the combined datasets into training and validation sets with an 80% to 20% split.

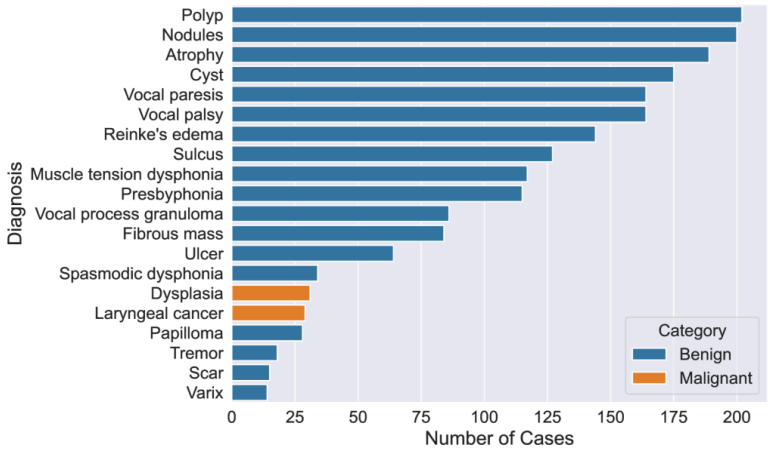

Figure 3. FEMH dataset criteria. Note that pre-malignant conditions are also classified as positive

DOI: 10.54254/2755-2721/2026.GU34282

Condition	Condition Type	Male	Female	**Total**
Vocal cord cancer	Malignant	21	1	**22**
Hypopharyngeal tumor	Malignant	6	0	**6**
Larynx tumor	Malignant	4	1	**5**
Epiglottic cancer	Malignant	0	1	**1**
Nesopharyngeal tumor	Malignant	1	0	**1**
Carcinoma in situ	Malignant	1	0	**1**
Dysplastic dysphonia	Pre-malignant	1	0	**1**
Dysplastic larynx	Pre-malignant	1	0	**1**

Figure 4. SVD dataset criteria

3.2. Training setup

Each audio recording was resampled to a frequency of 16000 Hz and normalized between -1 and 1 . Our ensemble contains five identical models, each trained using the PyTorch framework and the Adam [15] optimizer with an initial learning rate 10^{-4} . The models were trained on a batch size of 32 and 7 epochs; 32 was the maximum batch size that our hardware could support, and 7 epochs is when our training loss usually began to converge. All experiments were conducted on a Jupyter Notebook running an NVIDIA A100 GPU with 40 GB of VRAM. Our cross-entropy loss weights w_1 and w_2 were 31 and 1 , respectively, to reflect on the class distribution in each training batch. We grid searched from 0 to 1 for λ and 0 to 0.5 for m and found $\lambda = 0.6, m = 0.5$ to perform the best.

We also compare our framework with several baseline models to assess the performance gain of Wav2Vec 2.0, the loss function, and the model ensemble. We compare these results with a Support Vector Machine (SVM) using Mel-Frequency Cepstral Coefficients (MFCCs), a common design choice in the past [4, 9, 10].

3.3. Evaluation metrics

To ensure a reliable assessment of our model's performance, we discuss the reasoning and formulas behind our evaluation metric. We apply the commonly used precision and recall scores, which determine the accuracy of the model's class prediction and the percentage of the positive samples correctly identified by the model, respectively. Our imbalanced datasets makes plain accuracy an unreliable metric, since one can achieve very high accuracy by simply outputting the majority class. To combat this issue, we instead use balanced accuracy for an overall measure of the model's performance.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{5}$$

$$\text{Balanced Accuracy} = \frac{\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}}{2} \tag{6}$$

Where TP, FP, TN, FN stand for true positive, false positive, true negative, and false negative, respectively.

DOI: 10.54254/2755-2721/2026.GU34282

4. Results

In Table 1, we list the average weighted accuracy, precision, and recall for each model. Bold markers correspond to the model with the best performance on each metric. W2V-Base is trained only on cross-entropy loss, while W2V-Contrastive uses our combined loss function. The W2V-Ensemble aggregates all five instances of W2V-Contrastive described in 2.5

Table 1. Averaged balanced accuracy, precision, and recall for each model

Model	Balanced Accuracy	Precision (Benign, Malignant)	Recall (Benign, Malignant)
MFCC + SVM	61.28%	98.2%, 5.9%	87.6%, 34.9%
W2V-Base	81.19%	100%, 10.4%	70%, 92.2%
W2V-Contrastive	81.46%	100%, 10.4%	69.8%, 93%
W2V-Ensemble	84.85%	100%, 7%	70%, 100%

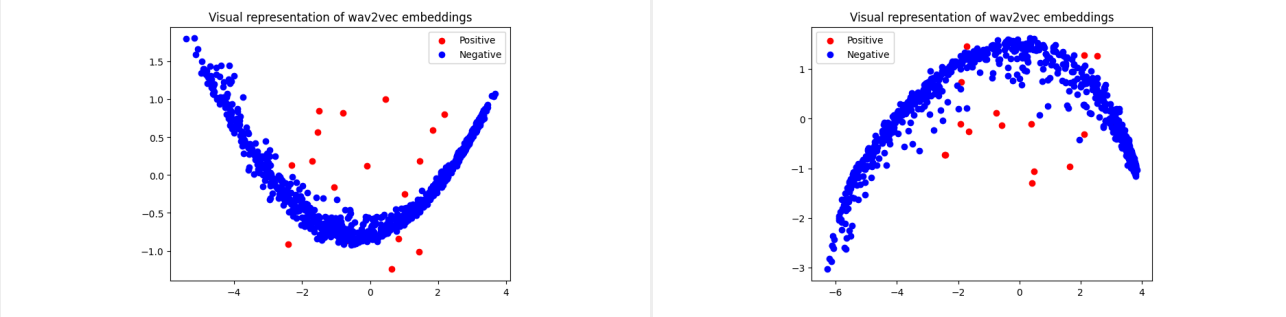

Figure 5. PCA Visualizations of embeddings in W2V-Base (left) and W2V-Contrastive (right)

4.1. Discussion

The results in Table 1 show that each of our models trained on Wav2Vec 2.0 significantly outperforms the baseline SVM, both in terms of balanced accuracy and recall score for the malignant class. These scores indicate that using Wav2Vec 2.0 as a feature extractor over traditional MFCCs significantly improves the model's overall performance, and works well when diagnosing actual cancer cases. The ensemble also provided further performance boost, indicating that it was able to address the bias present in each model.

However, there are still glaring issues with our results. For starters, W2V-Contrastive gives nearly no improvements over W2V-Base. This observation is further reflected in Figure 5, as W2V-Contrastive did not significantly increase the spread between positive and negative embeddings. Additionally, the precision score for malignant classes, although better than the baseline model, is extremely low in all of our models, with our best model achieving 10.4% accuracy. This means that all models are highly prone to falsely diagnosing non-cancerous patients with cancer, making the model undesirable in clinical settings. The benign recall scores are also lower than their malignant counterparts, suggesting that there is room for Wav2Vec 2.0 to improve detection accuracy on negative samples.

DOI: 10.54254/2755-2721/2026.GU34282

4.2. Ablation study

To further understand the effects of our design choices, we conduct a thorough ablation study with results listed in Table 2. We test another ensemble that uses our baseline model to further understand the impact of our ensemble. Moreover, we consider the performance when Wav2Vec 2.0 is frozen during training, and test both loss functions without data augmentations to better understand their behavior (codenamed No Augmentations and No Augmentations contrastive).

Table 2. Ablation study results

Model	Balanced Accuracy	Precision (Benign, Malignant)	Recall (Benign, Malignant)
Wav2Vec 2.0 Frozen	52.14%	98%, 2.2%	27%, 78%
Baseline Ensemble	85.98%	100%, 7%	72%, 100%
No Augmentations	82.28%	100%, 6.5%	70%, 95%
No Augmentations Contrastive	78.21%	99%, 16.5%	90%, 66.5%

Our ablation study results show that freezing the Wav2Vec feature extractor does indeed hinder performance, as the balanced accuracy dropped all the way down to 52.14%. Although the positive recall score is still strong, the model seems to lean more towards malignant cases in general, thus lowering our accuracy. Once again, our baseline ensemble improves the performance of the individual models, raising the accuracy from 81.19% to almost 86%.

Interestingly, though, removing data augmentations from our training slightly increased the performance of our baseline model, but further reduced that of the contrastive loss. We hypothesize that our baseline model exhibits this behavior because removing augmentations reduces the noisiness of the training data, thereby making it perform better on our datasets. This finding challenges the common belief that applying data augmentations would, in fact, better generalize to data, as our lower validation score suggests otherwise. On the other hand, the contrastive loss may have unintentionally distorted a clearer decision boundary present without data augmentations, thereby lowering the overall accuracy.

5. Conclusions

In this study, we design and evaluate an open-source framework for voice-based laryngeal cancer diagnosis which addresses existing challenges with data. To do so, we employ an model-ensemble approach, fine-tune the Wav2Vec 2.0 model, and examine the effects of a metric-learning loss function. Our results show that fine-tuning Wav2Vec 2.0, combined with the ensemble and our preprocessing strategy, significantly outperforms the traditional Machine Learning methods and improves our results much more than loss function design. In summary, our study shows that fine-tuning, feature learning, and careful data sampling outweigh loss function design and data augmentation, substantially improving the feasibility of voice-based laryngeal cancer diagnosis. These results also challenge previous methods for voice-related classification, where the latter two approaches were key performance contributors.

In the future, we also encourage researchers to incorporate demographic and symptom data with our proposed approach to further boost the performance, particularly for precision scores. This

DOI: 10.54254/2755-2721/2026.GU34282

method supports our goal for the detection of laryngeal cancer, as demographic and symptom data can be acquired from patients without the need for advanced clinical equipment or interference with a patient's organs. Both SVD and FEMH contain the necessary labels, and previous works have shown that including these features significantly improves model performance, thus adding another possible layer of improvement for non-invasive laryngeal cancer detection.

References

[1] J. Huang et al., "Updated disease distributions, risk factors, and trends of laryngeal cancer: A global analysis of cancer registries, " Int. J. Surg., vol. 110, no. 2, pp. 810–819, Feb. 2024.
[2] F. Bray et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, " CA Cancer J. Clin., vol. 74, no. 3, pp. 229–263, May 2024.
[3] "Cancer.org." https: //www.cancer.org/cancer/types/laryngeal-and-hypopharyngeal-cancer/detection-diagnosis-staging/detection.html.
[4] M. Paterson, J. Moor, and L. Cutillo, "A classification benchmark for artificial intelligence detection of laryngeal cancer from patient voice." 2025. Available: https: //arxiv.org/abs/2412.16267
[5] M. Paterson, J. Moor, and L. Cutillo, "Detecting throat cancer from speech signals using machine learning: A scoping literature review, " IEEE Access, vol. 13, pp. 58465–58480, 2025, doi: 10.1109/access.2025.3555767.
[6] S. A. S. Nolan N. Jones, Reinke's edema. National Institute of Health, 2023.
[7] C. Wang, T.-M. Chen, N. Lee, and S. Fang, "AI detection of glottic neoplasm using voice signals, demographics, and structured medical records, " The Laryngoscope, vol. 134, pp. 4585–4592, Jun. 2024, doi: 10.1002/lary.31563.
[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations." 2020. Available: https: //arxiv.org/abs/2006.11477
[9] R. Gupta, C. Madill, D. R. Gunjawate, D. D. Nguyen, and C. T. Jin, "Addressing data scarcity in voice disorder detection with self-supervised models, " in ICASSP 2024 - 2024 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2024, pp. 11866–11870. doi: 10.1109/ICASSP48485.2024.10446075.
[10] D. Ribas, M. A. Pastor, A. Miguel, D. Martínez, A. Ortega, and E. Lleida, "Automatic voice disorder detection using self-supervised representations, " IEEE Access, vol. 11, pp. 14915–14927, 2023, doi: 10.1109/ACCESS.2023.3243986.
[11] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning Wav2Vec2 for speaker recognition, " in ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2022, pp. 7967–7971. doi: 10.1109/ICASSP43922.2022.9746952.
[12] A. Koudounas et al., "Voice disorder analysis: A transformer-based approach, " in Interspeech 2024, in Interspeech_2024. ISCA, Sep. 2024, pp. 3040–3044. doi: 10.21437/interspeech.2024-1122.
[13] S.-S. Wang, C.-T. Wang, C.-C. Lai, Y. Tsao, and S.-H. Fang, "Continuous speech for improved learning pathological voice disorders." Feb. 2022. doi: 10.31219/osf.io/xpcez.
[14] M. Pützer and J. Koreman, "A german database of patterns of pathological vocal fold vibration." Zenodo, 1997. doi: 10.5281/zenodo.16258835.
[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." 2017. Available: https: //arxiv.org/abs/1412.6980