Fairness-Aware Deep Learning Model for COVID-19 Detection from Cough Audio Recordings

Dimitra Kostavasili
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
dimitrakostavasili@gmail.com

Theofanis Ganitidis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
orcid.org/0009-0006-7794-9793

Maria Athanasiou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
orcid.org/0000-0003-1575-9100

Konstantina S. Nikita
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
orcid.org/0000-0001-8255-4354

2025 IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3315-5899-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBE66822.2025.00045

Abstract— The COVID-19 pandemic intensified the demand for rapid, accessible diagnostic methods. Machine learning models using cough audio recordings have shown potential for remote COVID-19 detection but often exhibit performance disparities across demographic and clinical subgroups. This study investigates fairness-aware machine learning models for COVID-19 diagnosis using crowdsourced data from the COVID-19 Sounds dataset. A Random Forest and a VGGish-based deep neural network classifier were developed. To mitigate bias, four different methods were applied and comparatively evaluated, namely correlation remover as a pre-processing approach, exponentiated gradient and adversarial debiasing as in-processing approaches, and threshold optimizer as a post-processing approach. Evaluation across sensitive attributes including gender, age, recording device's operating system, and the intersection of age and gender revealed that fairness could be substantially improved with minimal loss in predictive accuracy. The best equalized odds ratio values were 0.956, 0.998, and 0.88 with respect to gender, age, and recording device's operating system, respectively. Similarly, for the combination of gender and age the corresponding metric was 0.804. These findings support the feasibility of fair and reliable audio-based diagnostic systems, emphasizing the importance of integrating fairness into clinical machine learning pipelines.

Keywords—COVID-19, machine learning, audio processing, fairness, bias mitigation, deep learning, VGGish

I. Introduction

The Coronavirus Disease 2019 (COVID-19) emerged from infection by the novel SARS-CoV-2 virus and primarily affects the respiratory system. In symptomatic individuals, it resembles severe pneumonia and can involve a broad range of symptoms [1]. In critical cases, COVID-19 can lead to death, particularly among high-risk groups including the elderly and individuals with pre-existing health conditions [2]. As of January 2025, COVID-19 has claimed over 7.1 million lives globally [2]. Despite progress in containment, it remains present, having exposed critical vulnerabilities in healthcare systems and spurring the need for early, accessible diagnostic approaches.

Artificial Intelligence (AI), and more specifically Machine Learning (ML), has emerged as a transformative tool in the medical domain, offering significant benefits in both disease diagnosis and treatment optimization. ML algorithms, fueled by diverse digital medical data such as electronic health records, biomarkers, medical imaging, audio, and video signals, can enhance the speed and accuracy of clinical decision-making processes [3], [4], [5]. Furthermore, ML facilitates the development of self-diagnosis tools that promote preventative care with increased accessibility and reduced cost [6], [7], [8].

In the context of COVID-19, the need for scalable diagnostic approaches became especially urgent. As SARS-CoV-2 primarily targets the respiratory system, it can alter the acoustic signals produced by patients, making audio signal processing a promising direction for ML-based diagnostic models [9]. The continuous monitoring of COVID-19 progression has also led to the collection of vast quantities of newly generated medical data, enabling the training of ML models based on both images and sounds associated with COVID-19 symptoms. Consequently, several mobile and web-based applications were developed to capture audio samples via device microphones, resulting in the creation of large open-access audio datasets, which continue to support diagnostic model training [10], [11], [12], [13]. A wide range of ML models trained on these datasets have demonstrated high levels of diagnostic accuracy, enabling fast, cost-effective, and non-invasive COVID-19 detection [6], [14].

The ongoing integration of ML into healthcare has significantly redefined disease diagnosis, prognosis, and treatment. Thus, in addition to model performance, *fairness* is considered a key evaluation criterion. Fairness in ML refers to a model's ability to make predictions without discriminatory behavior across sensitive subgroups [15]. Especially in high-stakes domains like healthcare, unfair treatment based on sensitive attributes such as gender, age, ethnicity, spoken language, socioeconomic status, or medical history, can lead to severe consequences [3], [16], [17].

Several COVID-19 diagnostic models based on acoustic data have recently been evaluated not only in terms of predictive performance but also with respect to fairness [6], [18], [19]. The findings indicate that certain models exhibited systematic biases in their predictions, particularly across sensitive attributes such as gender, age, and smoking frequency, raising concerns about the ethical implications and clinical reliability of such systems. In response to these

2471-7819/25/$31.00 ©2025 IEEE
DOI 10.1109/BIBE66822.2025.00045

challenges, recent work has demonstrated that it is possible to mitigate such demographic disparities through fairness-aware interventions. As a result, enhanced versions of the original models have shown reduced bias and more equitable performance across demographic groups, marking a significant step toward the responsible deployment of AI-based diagnostic tools [18], [19]. However, fairness in ML for healthcare remains a key challenge, as the balance between equity and predictive accuracy is not yet well defined. While some studies report a trade-off between fairness and performance [16], [20], others show that preprocessing can reduce bias without major accuracy loss [21], [22]. Fairness outcomes depend heavily on data properties, sensitive attribute selection, and training strategies, with ML systems often at risk of reinforcing existing societal inequalities.

In this study, fairness in ML models for COVID-19 detection based on cough audio signals is investigated. A systematic assessment of variations in predictive performance across demographic groups is performed and the effectiveness of established bias mitigation techniques in reducing such disparities is explored. To this end, a binary Random Forest (RF) classifier and a VGGish-based deep neural network were trained on audio data and evaluated with respect to sensitive attributes including gender, age, and recording device's operating system, as well as the intersection of age and gender. To address the observed disparities, four representative bias mitigation methods were implemented and independently assessed. These included correlation remover as a pre-processing approach, exponentiated gradient and adversarial debiasing as in-processing approaches, and threshold optimizer as a post-processing approach. Through this analysis, it was demonstrated that model fairness can be significantly improved in several cases without major degradation in accuracy. Beyond empirical findings, this work contributes a structured evaluation of fairness-aware ML strategies within the context of an audio-based biomedical application, highlighting their effectiveness and limitations. Furthermore, it emphasizes the importance of examining fairness not only across individual sensitive attributes but also at the intersection of demographic factors, offering a more nuanced understanding of algorithmic bias in healthcare-oriented AI systems.

II. Methods

A. Proposed fairness-aware framework

The architecture illustrated in Fig. 1 outlines the proposed fairness-aware pipeline for COVID-19 detection using cough audio. It integrates two classification models, an RF classifier, selected as a lightweight and interpretable baseline, and a Convolutional Neural Network (CNN), based on VGGish [23] pretrained model, followed by a fully connected classification head. The two models are deployed alongside pre-processing, in-processing, and post-processing bias mitigation techniques applied to sensitive attributes including gender, age, recording device's operating system, and the intersection of gender and age. Evaluation combines predictive accuracy with fairness metrics to ensure both reliable and equitable model performance.

B. Data

The dataset utilized was the COVID-19 Sounds dataset, released by the University of Cambridge [10], containing cough, breath, and speech recordings from individuals with confirmed COVID-19, as well as healthy controls. Each

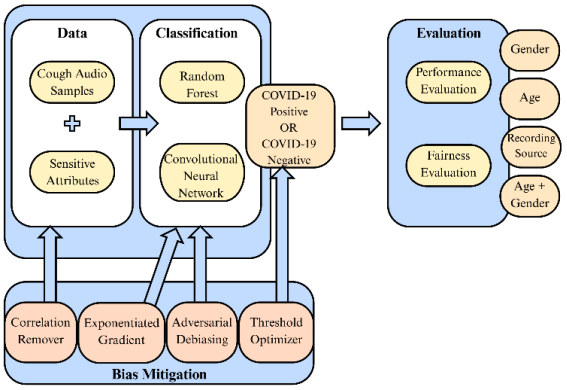

Fig. 1. Schematic overview of the classification, bias mitigation, and fairness evaluation workflow.

sample was accompanied by metadata including age, gender, and recording device's operating system. For the purpose of this study, a specific subset of the original dataset was used, tailored to the COVID-19 diagnosis task, comprising a total of 1,482 audio samples, with a balanced class distribution of 732 COVID-19 positive samples (49.4%) and 750 negative samples (50.6%).

The distributions with respect to sensitive attributes are shown in TABLE I. The dataset was relatively balanced in terms of gender. The age groups, while unbalanced, presented a relatively balanced distribution between the two classes. Regarding the device's operating system, the dataset's distribution was unbalanced and presented significant differences across classes. These differences suggested a potential association between device type and class label, able to introduce bias in model training, and emphasized the need to consider recording device's operating system as a sensitive attribute in fairness analysis.

Considering the intersection of gender and age, noticeable distributional differences were observed. In younger (16-39 years) and older (above 60 years) age groups, male participants were more represented than females, with the gender imbalance being more pronounced in the above 60 years category. In the middle-aged group, females were relatively more prominent. Analyzing class labels within these intersectional subgroups revealed further asymmetries, with some age-gender combinations showing a tendency toward either COVID-19 positive or negative labels. These patterns are summarized in TABLE II. Such imbalances across intersected subgroups can potentially influence model behavior, underscoring the importance of incorporating intersectional fairness assessments in machine learning models.

Only cough recordings were used in this study. Audio files were downsampled to 16kHz and converted into mono-

TABLE I. Dataset distribution of sensitive attributes

		Healthy	Covid-19 positive	Total
Gender	**Male**	49.3%	50.7%	51.2%
	Female	51.9%	48.1%	48.8%
Age	**16-39**	48.6%	51.4%	51.1%
	40-59	50.9%	49.1%	40%
	>60	60.8%	39.2%	8.9%
Device	**iOS**	56.5%	43.5%	65.3%
	Android	40.9%	59.1%	33.5%
	Web	0%	100%	1.2%

TABLE II. DATASET DISTRIBUTION FOR AGE-GENDER INTERSECTION

		Healthy	Covid-19 positive	Total
16-39	**Male**	48.4%	51.6%	51.3%
	Female	48.6%	51.4%	48.7%
40-59	**Male**	48.4%	51.6%	47.4%
	Female	53.1%	46.9%	52.6%
>60	**Male**	56.5%	43.5%	65.9%
	Female	70.5%	29.5%	34.1%

channel WAV format. Samples were amplitude-normalized, in accordance with the requirements of the pre-trained VGGish model. A row-wise filtering for missing values was applied to exclude rows with null feature vectors from failed audio processing, ensuring better quality control. Participants with incomplete or ambiguous metadata were excluded. For each recording, two types of feature representations were extracted. First, digital signal processing features were extracted for use in traditional classifiers. More specifically, 68 values of statistical audio features were extracted from each cough signal using digital signal processing methods, including Zero-Crossing Rate, MFCCs, Spectral Centroid, and Root Mean Square Power. The extracted features were fed as input to the RF classifier. Second, mel-spectrograms served as input to the CNN model. The VGGish model, which is a CNN pre-trained on a large-scale YouTube audio dataset [23], was used for feature extraction. Input audio segments were transformed into 96×64 Mel-spectrogram patches with 10ms hop size. The extracted 128-dimensional embeddings from VGGish were used as input features for classification.

C. Bias mitigation

To address bias across sensitive attributes (gender, age, and recording device's operating system), bias mitigation methods were implemented at the pre-processing, in-processing, and post-processing stages of the ML pipeline:

- **Pre-processing methods** intervene before model training by transforming the data to reduce inherent correlations between input features and sensitive variables, without sacrificing predictive information. In this study, the correlation remover algorithm was employed, which performs linear transformations to decorrelate features from sensitive attributes while preserving information content.
- **In-processing methods** aim to mitigate biases by modifying the model architecture or incorporating fairness objective functions and constraints. Two state-of-the-art in-processing methods were implemented. The first is exponentiated gradient, which is an iterative optimization technique that minimizes classification loss while enforcing fairness constraints (e.g., equalized odds). The other method is adversarial debiasing. This method trains the main predictor alongside an adversarial network. While the predictor aims to maximize accuracy, the adversary attempts to infer sensitive attributes. The competition between the two guides the model toward representations that are informative for prediction yet agnostic to sensitive attributes.
- **Post-processing methods** adjust the outputs of a pre-trained model to reduce fairness violations. In this work, the threshold optimizer was applied, which customizes classification thresholds for each demographic subgroup independently, with the goal of equalizing fairness constraints across groups to directly address disparities in decision outcomes.

Among the above methods, adversarial debiasing was applied in the case of the CNN classifier, due to its suitability for joint optimization of fairness and predictive objectives in deep models. The remaining mitigation methods were implemented on the RF classifier, selected for its robustness and compatibility with non-deep bias mitigation techniques.

D. Evaluation framework

The dataset was divided using a fixed stratified split into training (70%), validation (10%), and test (20%) subsets to maintain class balance between positive and negative COVID-19 cases. To prevent data leakage, user-level grouping was enforced, ensuring that all recordings from a given participant were allocated exclusively to a single subset. Additionally, to facilitate a focused fairness analysis, participants with unknown sensitive features or those from underrepresented groups, such as 'other' in gender, were excluded.

Classification performance was assessed using standard evaluation metrics including accuracy, precision, recall, and F1-score. To assess model fairness across sensitive attributes a comprehensive set of group fairness metrics was used. These included the equal opportunity, equalized odds, demographic parity, and accuracy parity. Each of these metrics was calculated across subgroups to capture disparities in model performance. To gain further insight into subgroup-specific behavior, additional class-specific metrics including true positive rate (TPR), false positive rate (FPR), false negative rate (FNR), and true negative rate (TNR) were computed independently for each subgroup.

Fairness disparities were quantified using both absolute differences and proportional ratios between the best-performing and worst-performing subgroups per metric. While absolute differences reflect the magnitude of disparity, ratio-based comparisons provide a relative measure of fairness that facilitates comparisons across attributes with uneven distributions. Among these fairness measures, particular emphasis was placed on the equalized odds ratio (EOR), which accounts for both TPR and FPR, thereby offering a balanced perspective on subgroup treatment. This is especially relevant in healthcare settings, where both false positive predictions and false negative predictions can lead to significant clinical consequences. EOR is calculated as:

$$EOR = \min\left(\left|\frac{\min(TPR_{group\ i})}{\max(TPR_{group\ j})}\right|, \left|\frac{\min(FPR_{group\ i})}{\max(FPR_{group\ j})}\right|\right) \quad (1)$$

with an ideal value of 1.0, and was considered alongside predictive performance metrics to ensure robust and equitable model assessment.

III. RESULTS

The obtained results of the classification performance and fairness evaluation are presented for both the traditional Random Forest (RF) classifier and the deep learning model (VGGish-based CNN) under various bias mitigation scenarios. Evaluation was conducted across the sensitive attributes of gender, age, and recording device's operating system, as well as the intersectional category combining

gender and age. Fig. 2 and Fig. 3 summarize the fairness and classification performance before and after applying the selected bias mitigation strategies.

Fairness evaluation revealed notable disparities across demographic groups, with variability in how each model responded to different debiasing techniques. For the gender attribute, the RF model exhibited relatively balanced performance at baseline, achieving an EOR of 0.832. This high initial score may be attributed to the relatively balanced gender distribution in the dataset. Applying the correlation remover significantly improved fairness, increasing the EOR to 0.956. This suggests that pre-processing techniques can be highly effective even when initial disparities are moderate. Conversely, the CNN started with a slightly lower EOR of 0.785, but achieved a notable improvement to 0.874 via adversarial debiasing.

For the age attribute, fairness differences were more pronounced, particularly in the RF model, which recorded a low baseline EOR of 0.600. This indicates potential disadvantages for underrepresented age groups, especially the one with individuals aged over 60. Application of the correlation remover led to a substantial improvement in fairness, with an EOR of 0.998, approaching perfect parity. The CNN showed more favorable baseline fairness with EOR of 0.858 and also benefitted from adversarial debiasing, which increased the EOR to 0.900, though the magnitude of improvement was smaller compared to RF.

In terms of recording device's operating system (e.g., iOS, Android, web browser), the RF exhibited significant bias, with a baseline EOR of just 0.338. This can be attributed to the data

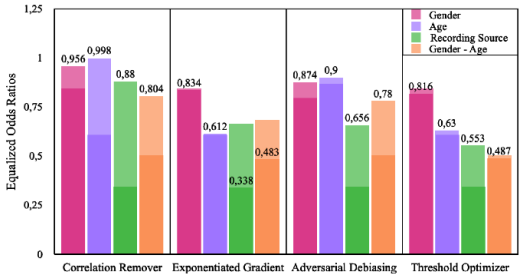

Fig. 2. Equalized odds ratio values for each sensitive attribute before and after the application of bias mitigation methods. The plot illustrates the effectiveness of the debiasing strategies applied to the RF (correlation remover, exponentiated gradient, threshold optimizer) and CNN (adversarial debiasing) classifiers in improving group fairness across gender, age, recording device's operating system, and the intersection of gender and age.

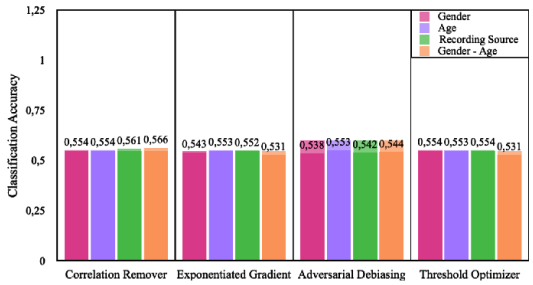

Fig. 3. Classification accuracy values for each sensitive attribute before and after the application of bias mitigation methods. The figure compares the predictive performance of the RF and CNN classifiers under different fairness interventions (correlation remover, exponentiated gradient, threshold optimizer for the RF classifier and adversarial debiasing for the CNN classifier).

imbalance across platforms as well as hardware-related discrepancies in microphone characteristics. Both the correlation remover and threshold optimizer improved fairness, with the former achieving the best outcome with an EOR of 0.880. By contrast, the CNN presented higher baseline fairness, reaching a 0.657 EOR value, which remained stable across all bias mitigation methods, suggesting more robustness, but also resistance to fairness improvement regarding this attribute.

For the intersection of gender and age, the RF model again showed poor baseline fairness, with an EOR of 0.459. This can be explained by the fragmentation of data into smaller subgroups, which affects label balance and learning capacity. Among the tested methods, only the correlation remover led to a meaningful improvement of EOR to 0.804. The CNN outperformed RF in this category at baseline with an EOR value of 0.676, which reached 0.78 following adversarial training. This result emphasizes the need for fairness evaluation beyond individual attributes, particularly in intersectional contexts where disparities are often amplified.

Regarding classification performance, both models achieved relatively low accuracy scores prior to the application of fairness interventions. The RF's baseline accuracy was 0.548, and although bias mitigation slightly influenced its performance, no clear improvement pattern was observed. Minor fluctuations across methods suggest that the RF's predictive capacity remained relatively stable, regardless of the fairness enhancement technique applied. In the CNN model, baseline accuracy was modest at 0.600. However, the application of adversarial debiasing led to a small decline in performance. This degradation occurred despite the improved fairness metrics, indicating a possible trade-off between equity and accuracy, particularly in high-capacity models, where fairness constraints may compete with loss minimization during training.

Overall, the results highlight that fairness improvements are attribute- and model-dependent, and that no single mitigation method is universally optimal. The RF classifier, though simpler, showed high responsiveness to pre-processing, while the CNN benefitted more from in-processing debiasing. Importantly, some mitigation strategies improved fairness with minimal performance cost, supporting their adoption in real-world systems where equitable decision-making is critical.

IV. Discussion

The results show that applying debiasing methods significantly improves fairness metrics, especially the EOR, across individual and intersected sensitive attributes. The adversarial debiasing and correlation remover methods yielded the most consistent fairness gains, whereas the threshold optimizer was less effective in post-hoc balancing subgroup outcomes.

Notably, the initially high fairness scores observed for the gender attribute may be partly attributed to the relatively balanced distribution of male and female participants in the dataset. In contrast, larger fairness disparities were detected for age, likely due to significant underrepresentation of individuals over 60 years old, which impacts subgroup balance during training and evaluation. The sensitive attribute of recording device's operating system further revealed that the device used for audio capture can influence prediction outcomes. This effect may stem from hardware-specific

characteristics such as microphone quality and frequency response differences across platforms, particularly relevant in audio-based classification tasks. Lastly, the intersectional category of gender combined with age exhibited greater unfairness compared to individual attributes. This is possibly due to the division of the dataset into smaller subgroups, which disrupts label distribution balance and magnifies disparities in predictive performance across intersectional categories.

In the case of the CNN model, a performance drop in the range of approximately 5% to 10% was observed following the application of adversarial debiasing, which was accompanied by a substantial enhancement in fairness, with the EOR improving by up to 28.66%, demonstrating the capacity of adversarial debiasing to effectively correct learned biases within deeper architectures. This compromise could be justified in healthcare applications where equity is a primary concern. The closest value to the ideal EOR (1.0) was achieved for the gender attribute, reaching 0.903. Although a deterioration in fairness was observed for certain attributes, the relative decreases in EOR were generally limited, indicating that the method did not introduce substantial fairness degradation.

The RF model exhibited overall lower predictive accuracy compared to the CNN. Nevertheless, it showed only slight fluctuations in performance before and after the application of bias mitigation methods. In terms of fairness, the correlation remover yielded the most favorable results, with EOR values being closer to the ideal. The highest proximity to the ideal EOR value (1.0) was observed for the age attribute, reaching 0.998 with the correlation remover. Additionally, the largest relative improvement in EOR was achieved for the recording device's operating system attribute, with a 59.54% increase following the application of the correlation remover.

Despite small trade-offs in classification accuracy, the fairness improvements suggest a positive fairness-accuracy trade-off, which is acceptable in medical applications prioritizing ethical AI deployment. Moreover, the methods used are modular and generalizable, making them suitable for other medical sound-based detection tasks (e.g., asthma, COPD). The approach also aligns with ethical AI principles, promoting equity in ML applications where clinical decision-making may be influenced by biased predictions. In the context of future pandemics or telemedicine frameworks, such audio-based tools can support remote screening in underserved populations, provided they are trained and validated responsibly.

The observed partial improvements in fairness metrics following the application of individual bias mitigation strategies suggest promising directions for further research. In particular, future work could explore the combined or sequential application of multiple mitigation techniques across pre-processing, in-processing, and post-processing levels, to assess whether synergistic effects may lead to greater gains in both fairness and predictive performance. Additionally, expanding the dataset to include a broader and more representative sample with increased coverage and numerical balance across sensitive attributes may significantly improve model fairness, especially within intersectional or minority subgroups.

V. Conclusion

In this study, the integration of fairness-aware ML methods into audio-based COVID-19 diagnostic pipelines was investigated. Using a dataset of cough recordings, both an RF classifier and a CNN model were trained and evaluated while representative bias mitigation techniques across pre-processing, in-processing, and post-processing stages were deployed. Our analysis revealed that methods such as correlation remover and adversarial debiasing can significantly enhance group fairness without incurring substantial losses in predictive accuracy. The importance of addressing intersectional subgroups and technical factors such as recording hardware was also highlighted, as they can subtly influence model behavior. Although classification accuracy remained modest, the achieved fairness gains demonstrated the feasibility of ethical and equitable audio-based diagnostics. These findings underscore the need for fairness to be treated as a core evaluation criterion in clinical ML systems.

References

[1] A. Çalıca Utku, G. Budak, O. Karabay, E. Güçlü, H. D. Okan, and A. Vatan, "Main symptoms in patients presenting in the COVID-19 period," *Scott. Med. J.*, vol. 65, no. 4, pp. 127–132, Nov. 2020, doi: 10.1177/0036933020949253.

[2] "COVID-19 deaths | WHO COVID-19 dashboard," datadot. Accessed: Jan. 21, 2025. [Online]. Available: https://data.who.int/dashboards/covid19/cases

[3] S. Ding *et al.*, "Fairly Predicting Graft Failure in Liver Transplant for Organ Assigning," *AMIA. Annu. Symp. Proc.*, vol. 2022, p. 415, Apr. 2023.

[4] K. Zarkogianni, M. Athanasiou, A. C. Thanopoulou, and K. S. Nikita, "Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1637–1647, Sep. 2018, doi: 10.1109/JBHI.2017.2765639.

[5] M. E. Vlontzou, M. Athanasiou, K. V. Dalakleidi, I. Skampardoni, C. Davatzikos, and K. Nikita, "A comprehensive interpretable machine learning framework for mild cognitive impairment and Alzheimer's disease diagnosis," *Sci. Rep.*, vol. 15, no. 1, p. 8410, Mar. 2025, doi: 10.1038/s41598-025-92577-6.

[6] J. Han *et al.*, "Sounds of COVID-19: exploring realistic performance of audio-based digital testing," *Npj Digit. Med.*, vol. 5, no. 1, pp. 1–9, Jan. 2022, doi: 10.1038/s41746-021-00553-x.

[7] A. Kondaka, "Evaluating Gender Bias and Fairness in Skin Lesion Diagnoses using Convolutional Neural Networks," 2024.

[8] H. M. Thompson *et al.*, "Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 28, no. 11, pp. 2393–2403, Oct. 2021, doi: 10.1093/jamia/ocab148.

[9] E. S. Adamidi, K. Mitsis, and K. S. Nikita, "Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 2833–2850, Jan. 2021, doi: 10.1016/j.csbj.2021.05.010.

[10] T. Xia *et al.*, "COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening".

[11] D. Bhattacharya *et al.*, "Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection," *Sci. Data*, vol. 10, no. 1, p. 397, Jun. 2023, doi: 10.1038/s41597-023-02266-0.

[12] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Sci. Data*, vol. 8, no. 1, p. 156, Jun. 2021, doi: 10.1038/s41597-021-00937-4.

[13] K. Zarkogianni *et al.*, "The smarty4covid dataset and knowledge base as a framework for interpretable physiological audio data analysis," *Sci. Data*, vol. 10, no. 1, p. 770, Nov. 2023, doi: 10.1038/s41597-023-02646-6.

[14] T. Ganitidis, M. Athanasiou, K. Mitsis, K. Zarkogianni, and K. S. Nikita, "A Comprehensive Drift-Adaptive Framework for Sustaining Model Performance in COVID-19 Detection From Dynamic Cough

Audio Data: Model Development and Validation," *J. Med. Internet Res.*, vol. 27, no. 1, p. e66919, Jun. 2025, doi: 10.2196/66919.

[15] Z. Xu, J. Li, Q. Yao, H. Li, M. Zhao, and S. K. Zhou, "Addressing fairness issues in deep learning-based medical image analysis: a systematic review," *Npj Digit. Med.*, vol. 7, no. 1, pp. 1–16, Oct. 2024, doi: 10.1038/s41746-024-01276-5.

[16] M. E. Vlontzou, M. Athanasiou, C. Davatzikos, and K. S. Nikita, "Comparative assessment of fairness definitions and bias mitigation strategies in machine learning-based diagnosis of Alzheimer's disease from MR images," May 29, 2025, *arXiv*: arXiv:2505.23528. doi: 10.48550/arXiv.2505.23528.

[17] P. Gajane, S. Newman, M. Pechenizkiy, and J. D. Piette, "Investigating Gender Fairness in Machine Learning-driven Personalized Care for Chronic Pain," Jun. 14, 2024, *arXiv*: arXiv:2402.19226. doi: 10.48550/arXiv.2402.19226.

[18] T. Saeed, A. Ijaz, I. Sadiq, H. N. Qureshi, A. Rizwan, and A. Imran, "An AI-enabled Bias-Free Respiratory Disease Diagnosis Model using Cough Audio: A Case Study for COVID-19," Jan. 04, 2024, *arXiv*: arXiv:2401.02996. doi: 10.48550/arXiv.2401.02996.

[19] R. Pfeifer, S. Vhaduri, and J. E. Dietz, "Mitigating Sex Bias in Audio Data-driven COPD and COVID-19 Breathing Pattern Detection Models," Sep. 16, 2024, *arXiv*: arXiv:2409.10677. doi: 10.48550/arXiv.2409.10677.

[20] Y. Zong, Y. Yang, and T. Hospedales, "MEDFAIR: Benchmarking Fairness for Medical Imaging," Feb. 17, 2023, *arXiv*: arXiv:2210.01725. doi: 10.48550/arXiv.2210.01725.

[21] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney, "Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing," in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Nov. 2020, pp. 2803–2813. Accessed: Jul. 10, 2025. [Online]. Available: https://proceedings.mlr.press/v119/dutta20a.html

[22] Y.-Y. Huang, V. Chiuwanara, C.-H. Lin, and P.-C. Kuo, "Mitigating Bias in MRI-Based Alzheimer's Disease Classifiers Through Pruning of Deep Neural Networks," in *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging: 12th International Workshop, CLIP 2023 1st International Workshop, FAIMI 2023 and 2nd International Workshop, EPIMI 2023 Vancouver, BC, Canada, October 8 and October 12, 2023 Proceedings*, Berlin, Heidelberg: Springer-Verlag, Jul. 2023, pp. 163–171. doi: 10.1007/978-3-031-45249-9_16.

[23] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135. doi: 10.1109/ICASSP.2017.7952132.