Explainable and Interpretable AI for Voice and Speech Analysis in Clinical Care: A Systematic Review

Mohamed Ebraheem, Jamie Toghranegar, Bridge2AI-Voice Consortium, Yael Bensoussan, John Michael Templeton

Submitted to: Journal of Medical Internet Research
on: September 09, 2025

Table of Contents

Explainable and Interpretable AI for Voice and Speech Analysis in Clinical Care: A Systematic Review

Mohamed Ebraheem[1] MSCS; Jamie Toghranegar[2] SLPD; Bridge2AI-Voice Consortium[3]; Yael Bensoussan[2*] MD; John Michael Templeton[1, 4*] PhD

[1] Bellini College of Artificial Intelligence, Cybersecurity and Computing University of South Florida Tampa US
[2]USF Health Voice Center Department of Otolaryngology Head & Neck Surgery University of South Florida Tampa US
[3] University of South Florida Tampa US
[4]Medical Engineering College of Engineering University of South Florida Tampa US
[*]these authors contributed equally

Corresponding Author:
Mohamed Ebraheem MSCS

Bellini College of Artificial Intelligence, Cybersecurity and Computing
University of South Florida
4202 E Fowler Ave.
ENB 247
Tampa
US

Abstract

Background: Driven by recent advances in artificial intelligence, particularly in medicine, audio-based voice and speech biomarkers are increasingly investigated for various medical applications as a complementary or even alternative modality to traditional medical devices. The adoption of deep learning techniques in recent literature is motivated by their superior performance compared to classical machine learning (ML) methods. However, ethical and regulatory concerns regarding the black-box nature of these models have limited their integration into clinical workflows. Consequently, Explainable AI (XAI) has recently been employed to address this issue by generating explanations for opaque model output. Ideally, medical XAI systems aim to provide human-understandable, clinically grounded explanations essential for enhanced AI trustworthiness and, thereby, facilitated adoption into real-world clinical settings.

Objective: We conduct a systematic literature review of XAI methods applied for explaining deep learning techniques in audio-based voice and speech clinical applications. We present a taxonomy of XAI methods in the literature and discuss the limitations of these methods, particularly for their application to clinical audio, evaluation of XAI outputs, and stakeholder relevance of generated explanation. Then, we identify opportunities and recommendations for future clinical audio XAI design.

Methods: This review follows the Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Six databases (IEEEXplore, ACM, Scopus, PubMed, Web of Science, and Nature) were searched for articles between January 2015 and February 2025. Included studies applied explainability and/or interpretability methods to deep learning techniques for clinical voice and speech audio.

Results: A taxonomy of XAI methods is presented for 30 eligible studies. These methods are grouped into four categories: visualization-based techniques, feature-importance and attribution methods, attention-based explanations, and concept detectors and model intrinsic approaches. We find that current XAI methods and implementations lack rigorous evaluation and validation, are not suitable for the unique nature of clinical audio, and do not align with stakeholder expectations and needs.

Conclusions: This survey presents a categorization of XAI techniques employed for voice and speech AI. We discuss several gaps and considerations and identify several opportunities for future clinical audio XAI design.

(JMIR Preprints 09/09/2025:83790)
DOI: https://doi.org/10.2196/preprints.83790

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Explainable and Interpretable AI for Voice and Speech Analysis in Clinical Care: A Systematic Review

Mohamed Ebraheem[1], MSc; Jamie Torghranegar[2], SLPD; Bridge2AI-Voice Consortium; Yael Bensoussan[2*], M.D. John Michael Templeton[1*], PhD;

[1]Bellini College of Artificial Intelligence, Cybersecurity and Computing, University of South Florida, Tampa, FL, United States

[2]Department of Otolaryngology Head & Neck Surgery, Morsani College of Medicine, University of South Florida, Tampa, FL United States

[*]These authors have equally contributed to this work.

Corresponding Author:
Mohamed Ebraheem, MSc
Bellini College of Artificial Intelligence, Cybersecurity and Computing
3220 USF Banyan Circle
Tampa, FL, 33620
United States
Phone: 1 8135850780
Email: mohamedusama@usf.edu

Abstract

Background: Driven by recent advances in artificial intelligence, particularly in medicine, audio-based voice and speech biomarkers are increasingly investigated for various medical applications as a complementary or even alternative modality to traditional medical devices. The adoption of deep learning techniques in recent literature is motivated by their superior performance compared to classical machine learning (ML) methods. However, ethical and regulatory concerns regarding the black-box nature of these models have limited their integration into clinical workflows. Consequently, Explainable AI (XAI) has recently been employed to address this issue by generating explanations for opaque model output. Ideally, medical XAI systems aim to provide human-understandable, clinically grounded explanations essential for enhanced AI trustworthiness and, thereby, facilitated adoption into real-world clinical settings.

Objective: We conduct a systematic literature review of XAI methods applied for explaining deep learning techniques in audio-based voice and speech clinical applications. We present a taxonomy of XAI methods in the literature and discuss the limitations of these methods, particularly for their application to clinical audio, evaluation of XAI outputs, and stakeholder relevance of generated explanation. Then, we identify opportunities and recommendations for future clinical audio XAI design.

Methods: This review follows the Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Six databases (IEEEXplore, ACM, Scopus, PubMed, Web of Science, and Nature) were searched for articles between January 2015 and February 2025. Included studies applied explainability and/or interpretability methods to deep learning techniques for clinical voice and speech audio.

Results: A taxonomy of XAI methods is presented for 30 eligible studies. These methods are grouped into four categories: visualization-based techniques, feature-importance and attribution methods, attention-based explanations, and concept detectors and model intrinsic approaches. We

find that current XAI methods and implementations lack rigorous evaluation and validation, are not suitable for the unique nature of clinical audio, and do not align with stakeholder expectations and needs.

Conclusions: This survey presents a categorization of XAI techniques employed for voice and speech AI. We discuss several gaps and considerations and identify several opportunities for future clinical audio XAI design.

KEYWORDS

Explainable artificial intelligence; clinical voice analysis; speech biomarkers; deep learning; interpretability; medical decision support; trustworthy AI

Introduction

Voice is a rich modality that has garnered the interest of the research community for its potential in numerous medical applications [1, 2]. Voice and speech biomarkers have recently been used for voice pathology detection [3-5], voice quality assessment [6, 7], neurodegenerative disease diagnosis [8-13], mental health monitoring [14, 15], cardio-respiratory condition classification [16, 17], as well as automatic speech recognition for disordered speech [18-20]. The main reason behind this interest in audio-based voice and speech biomarkers is the costliness and invasiveness of traditional voice evaluation techniques, such as laryngoscopy, stroboscopy, laryngeal electromyography, and imaging technologies like MRI and CT scans, which limit accessibility for many. Alternatively, artificial intelligence (AI)-driven audio-based medical systems pave the way for broader access to medical services for marginalized and underprivileged populations [21].

Yet, AI integration into real clinical settings is limited, in part due to the scarcity of high-quality data needed to train reliable and fair models [22]. Consequently, myriad large-scale projects are underway for the purpose of collecting extensive and representative voice audio datasets. NIH-funded Bridge2AI is a multi-institution, large-scale project that aims to collect standardized, AI-ready, and ethically sourced voice data across various health conditions, where voice and speech samples have been collected from 442 participants [23]. Similarly, AphasiaBank is another NIH-funded endeavor that had amassed multimodal data from 306 persons with aphasia [24]. Launched in 2023 and funded through 2028, SpeechDx is a global initiative dedicated to creating an extensive dataset of Alzheimer's speech biomarkers and has recruited about 2000 participants [25]. These efforts demonstrate the general recognition of the potential of voice and speech AI in routine clinical practice.

Recently, trustworthiness of AI systems has been a central issue for clinical integration of AI, particularly the obscurity of the decision-making processes of deep learning models to end-users [22, 26-29]. Known as the "black-box" problem, this issue is especially critical in medicine, where decisions directly impact patient safety. Yet regulatory frameworks have struggled to keep pace with AI's rapid development, leaving unresolved questions of liability and accountability [30]. Clinicians and regulators must understand how models operate, how reliable they are, and under what conditions they fail, before AI system are integrated into clinical workflows in a meaningful way [31]. Only then can patients be guaranteed safe, high-quality medical care. On one end, there are opinions against the use of black-box models at all for high-stakes environments such as medicine [32]. While white-box, classical machine learning models such as decision trees or support vector machines trained on interpretable, hand-crafted features offer greater transparency, deep learning models consistently achieve superior performance; this is known as the interpretability-accuracy tradeoff [33].

Explainable AI (XAI) has emerged to ameliorate this challenge, aiming to make black-box models more transparent [22, 29, 31, 34]. However, many XAI techniques have been developed for image-based domains, such as overlaying saliency maps on brain MRI or chest CT scans.

In voice analysis, raw audio is often transformed into time–frequency representations such as spectrograms before being input into deep learning models. Mapping a region on a spectrogram back to an intuitively understandable auditory event (e.g., a tremor on a specific syllable) is inherently more complex than identifying a visible tumor on an X-ray. The relationship between spectral features and perceived voice quality or pathology is often highly non-linear, making clinical interpretation challenging.

Furthermore, the intuitiveness of model explanations varies substantially with respect to end-users [35-37]. Computer scientists and ML researchers may be able to interpret technical visualizations such as activation maps or feature attribution plots, whereas clinicians, patients, and regulators have different knowledge bases, priorities, and constraints. A “one-size-fits-all” approach to XAI design is therefore inappropriate, especially when explainability is positioned as a pathway to increasing trust. Thereby, multi-disciplinary effort in the design of XAI methods is vital for providing appropriate presentations of explanations suited for the diverse backgrounds and needs of the respective stakeholders [37, 38].

In this context, this paper presents a systematic literature review of XAI methods applied to deep learning models for clinical voice and speech analysis. The review addresses the following research questions:

- What XAI methods have been used to explain the decisions of deep learning voice and speech AI systems in healthcare?
- What are the limitations of these XAI methods in the context of clinical audio in terms of clinical applicability and stakeholder relevance?

The remainder of this paper is organized as follows. Section II presents background for XAI concepts and different perspectives regarding the definitions of explainability and interpretability. We also outline the objectives of explainable AI, provide a discussion on the broad application of XAI in diverse medical domains. Section III describes the systematic review methodology. In Section IV, a taxonomy of explainability approaches is presented. Section V discusses the limitations of current approaches, stakeholder alignment, and the impact of audio representation on interpretability, upon which future directions are presented in Section VI. Finally, we conclude in Section VII.

Background

Explainability Vs. Interpretability

There is no clear consensus in the literature on the definitions of explainability and interpretability in the context of AI, and the two terms are often used interchangeably. Nonetheless, various works have attempted to provide distinctions between them.

Linardatos et al. [39] highlights the persistent ambiguity surrounding these concepts and reviews the different ways they have been differentiated in literature. They describe interpretability as relating to the intuition behind a model’s outputs, such that a more interpretable model makes it easier to identify causal relationships between inputs and outputs. One the other hand, explainability concerns the internal logic and mechanics of the model. They conclude that interpretability is the broader term, and that a model can be interpretable without being explainable.

Gilpin et al.[40] approach the distinction differently, defining explainability as a means of answering the questions “why” or “why not” a system behaves in a particular way. In contrast, they describe interpretability as the ability to represent the model’s internal processes in a human-understandable form, emphasizing that this is dependent on the knowledge and needs of the target user.

Das and Rad [41] define interpretability as a quality of a system in which its expressions convey human-understandable insights into how it works. They differentiate explanations as additional metadata—produced either by the model or an external algorithm—that clarify the relationship between inputs and outputs.

The National Institute of Standards and Technology (NIST) also provides guidance on

explainability, outlining four principles for explainable systems, most important of which is that they should deliver accompanying justifications or reasons for model outputs [42]. The work further distinguishes between self-interpretable models, which are inherently understandable to humans, and post-hoc explanations, which are generated by an explainability algorithm to provide insight into otherwise opaque models.

In summary, interpretability relates to the inherent transparency of the model itself; white-box models, such as decision trees and linear regression, are generally considered interpretable, while explainability refers to the generation of additional information (i.e., explanations) that clarify the reasoning behind a model's decision, regardless of the model's inherent transparency. In this work, we adopt the latter definition.

Objectives of Explainable AI

Explainable AI serves multiple partially overlapping objectives, which vary according to on the target application and the perspective of the stakeholder [43, 44]. We summarize them into five overarching categories that capture the core purposes of XAI

Debugging and Monitoring

Explanations help AI developers identify model errors, biases, and spurious correlations, thereby revealing opportunities for performance improvement. XAI can also be used to monitor for performance drift during deployment, ensuring the model continues to operate as intended over time.

Evaluation and Validation

XAI enables stakeholders to assess whether a model is appropriate, reliable, and clinically valid for a given application. It supports both pre-deployment evaluation and ongoing validation, ensuring that model decisions remain aligned with intended clinical outcomes.

Justification and Transparency

XAI fosters trust among stakeholders by providing context-relevant rationales for model decisions. This includes justifying individual predictions and improving transparency in the decision-making process, allowing clinicians, patients, and regulators to audit and verify challenge the system's outputs.

Improvement and Learning

Explanations support iterative model refinement through collaboration with domain experts, enhancing alignment with clinical reasoning. XAI can also contribute to the discovery of new domain knowledge, such as identifying previously unknown biomarkers.

Governance and Compliance

In high-stakes domains such as medicine, XAI facilitates compliance with ethical, legal, and regulatory requirements, including provisions like the European General Data Protection Regulation (GDPR) "right to explanation" [45, 46]. It enables model auditability, supports liability attribution, and strengthens governance processes.

Characteristics of Explainable AI Methods

Traditionally, XAI methods are categorized according to their scope of application (i.e., model-specific and model-agnostic). Classical taxonomy groups XAI methods into intrinsic methods, where the model contains native interpretable or explainable components, and post-hoc techniques, where a secondary system generates explanations. Resultant explanations provide either local insight for individual sample instances or globally justify model behavior. Accordingly, methods in the literature are categorized in Table I

Model-Specific vs. Model-Agnostic

Model-specific methods exploit structural aspects particular to model type or architecture to, in theory, produce higher-quality explanations. For example, Gradient-weighted Class Activation Mapping (Grad-CAM) are applicable to convolutional neural networks (CNNs) and take advantage of their spatial feature maps to highlight input features that are most influential for the output.

In contrast, model-agnostic methods, such as SHapley Additive exPlanations (SHAP), are generalizable to any machine learning model regardless of architecture. This universality enables versatile implementation across many applications. Unlike model-specific techniques that benefit from internal model architecture, model-agnostic methods often exhibit lower fidelity and efficiency, especially when dealing with high-dimensional data.

Local vs. Global

Local explainability methods aim to clarify the reasoning behind a model's prediction for a single input instance. These explanations are instance-specific and do not describe the model's overall decision-making process. For example, visualization techniques such as Class Activation Maps can generate heatmaps that indicate the regions of an input spectrogram most relevant to the model's prediction. Local explanations are particularly valuable in clinical decision support scenarios where individual case justification is critical.

Global explainability methods, on the other hand, aim to describe the model's general behavior across the entire dataset. This can involve identifying the most influential features for distinguishing between classes, mapping decision boundaries, or summarizing feature interactions. Such methods are crucial for understanding systematic model biases, validating clinical relevance, and ensuring that the model's logic aligns with domain knowledge.

Intrinsic vs. Post-hoc

As mentioned earlier, intrinsic (or self-) interpretability refers to the degree to which a model's decision-making process is transparent and human-understandable by design. White-box models are examples of intrinsically interpretable models. However, they often underperform compared to more complex black-box architectures like deep neural networks.

Post-hoc methods, in contrast, are applied after training process is completed to explain the model behavior without altering its internal mechanics. These techniques, ranging from saliency maps to perturbation-based analyses, are particularly prevalent for explaining black-box models.

Explainable AI in Medicine

With recent advances in medical AI, the implementation of XAI has become increasingly critical to ensure that AI models are reliable, ethical, legally compliant, and clinically aligned, particularly in high-stakes environments such as healthcare. The literature contains numerous surveys discussing the role of XAI across diverse medical applications.

Several recent reviews have broadly examined the adoption of XAI in medicine, emphasizing its potential to improve transparency, foster trust, and facilitate regulatory compliance [47-51]. These works cover a variety of data modalities, including medical imaging, electronic health records (EHR), genomics, and time-series analysis. Commonly reported methods include model-agnostic techniques, most notably LIME and SHAP, alongside post-hoc visualization approaches. Nonetheless, these reviews consistently highlight persistent challenges relating to faithfulness, stability, and standardized evaluation of explanations.

Domain-specific surveys further illustrate these trends. For example, Van der Velden et al. [52] reviewed over 200 studies applying XAI to medical imaging, noting the predominance of visualization techniques, followed by perturbation-based, textual, and example-based approaches. While visual explanations have been extensively validated, the authors stress the need for equivalent validation of textual and example-based methods. Muhammad and Bendechache [53] similarly highlight the interpretive

ambiguity and perturbation sensitivity of visual methods. In the context of medical time-series data, Caterson et al. [54] presents a scoping review of XAI for EHR, finding feature attribution methods to be the most widely used. Salih et al. [55] review XAI in cardiology for ECG and EHR, reporting the frequent use of SHAP and Grad-CAM, followed by LIME, but also note that nearly half (47%) of the studies did not employ any formal evaluation of the explanations produced. In the mental health and psychiatry domain, Joyce et al. [56] reviewed XAI applied to neuroimaging, interview transcripts, and physiological data, underscoring the importance of human-centered design for producing clinically useful explanations beyond raw saliency maps. With respect to clinical audio, Chen et al. [16] reviewed XAI approaches for vocal biomarker-based lung disease detection, discussing several issues including utility of explanations defined in terms of informativeness and user understanding as being an important criterion for evaluating explanations.

Despite the breadth of these surveys, to the best of our knowledge, no systematic reviews have comprehensively examined the use of XAI for clinical voice and speech biomarkers more broadly. This gap highlights the need for a dedicated synthesis in this domain, particularly given the unique challenges and interpretive demands of audio-based clinical decision support systems.

Unique Challenges of Clinical Audio for XAI

Audio Abstraction and Representation

In the time domain, audio is represented as waveforms that fully capture the acoustic signal but are difficult for humans to visually interpret directly. Consequently, audio is often transformed into spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs). While spectrograms provide a visually structured time–frequency representation, interpreting a highlighted “hotspot” is far less intuitive than interpreting a highlighted region on an X-ray. This is a result of human perception of sounds through their ears, not through vision [57]. MFCCs, a compact nonlinear “spectrum-of-the-spectrum” representation of audio, are even more obscure to clinicians, limiting their clinical interpretability [58, 59]. Additionally, speaker characteristics are usually spread across multiple frequency bands,

making the problem of localizing relevant frequency information too broad to solve effectively with traditional vision-based XAI methods [60]. Nonetheless, assuming relevant spatio-temporal regions are identifiable through traditional vision- based methods, understanding why these regions are important remains obscure, at least partially, and requires further analysis to answer such questions. Ultimately, the non-visual perceptual nature of voice and speech makes current visual explanations a non-ideal solution to the overarching problem of explainability and trustworthy clinical AI.

Table 1. Classical XAI taxonomy and subjective ease of perception.

XAI Methods	Model Agnostic/ Specific	Global/ Loca	Post-hoc/ Intrinsic	Explanation Modality	Designed for Audio	Ease of Perception			
						AI Scientist	Policy Makers	Clinicians	Patients
Grad-CAM	Specific	Local	Post-hoc	Visual (heatmap)	No	High	Medium	Medium	Low
Guided Backpropagation	Specific	Local	Post-hoc	Visual (overlay)	No	High	Low	Low	Low
Saliency Maps	Specific	Local	Post-hoc	Visual	No	High	Low	Low	Low
Eigen-CAM	Specific	Local	Post-hoc	Visual (heatmap)	No	Medium	Low	Medium	Low
SHAP	Agnostic	Local	Post-hoc	Tabular	No	High	Medium	Medium	Low
GradientSHAP	Specific	Local	Post-hoc	Tabular/Visual	No	High	Medium	Medium	Low
LIME	Agnostic	Local	Post-hoc	Tabular/Word-Level	No	High	High	High	Medium

xDMFCC	Specific	Local	Post-hoc	Feature-based/ Tabular	Yes	Medium	Low	Medium	Low
Ablation Studies	Specific	Global	Post-hoc	Conceptual/ Numerical	Partial	Medium	Low	Low	Low
Simple Attention	Specific	Local	Intrinsic	Visual/Textual	Partial	Medium	Low	Medium	Low
Attention Rollout	Specific	Local	Post-hoc	Visual	Partial	Medium	Low	Medium	Low
Concept Detectors Network	Specific	Global	Intrinsic	Concept-level Attribution	Yes	High	Medium	High	Medium
Sinc Filter Analysis	Specific	Global	Intrinsic	Conceptual (filter shape)	Yes	Medium	Low	Medium	Low
Feature Map Analysis	Specific	Global	Post-hoc	Visual (CNN filters)	No	Medium	Low	Medium	Low
t-SNE	Agnostic	Global	Post-hoc	Latent Space Visualization	No	Medium	Medium	Medium	Low

interpret directly. Consequently, audio is often transformed into spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs). While spectrograms provide a visually structured time–frequency representation, interpreting a highlighted “hotspot” is far less intuitive than interpreting a highlighted region on an X-ray. MFCCs, a compact nonlinear “spectrum-of-the-spectrum” representation of audio, are even more obscure to clinicians, limiting their clinical interpretability [56]–[58].

Temporal Dynamics

Unlike medical images, where both axes carry spatial meaning, audio has an inherently temporal structure. Identifying when a clinically relevant event occurs (e.g., a stutter or pause) is as important as identifying which acoustic features are involved. This mismatch makes it difficult for visualization-based methods such as saliency maps or Grad-CAM, designed for spatial data, to yield clinically actionable explanations in audio [57].

Annotation Scarcity

Phonetic, prosodic, disfluency, and voice quality annotations are essential for aligning model explanations with known biomarkers; however, such granular annotations, typically performed by multiple trained experts, are resource-intensive and, consequently, scarce. Without such annotations, validation of model explanations alignment to medically grounded features become increasingly daunting, limiting their utility in practice.

In summary, these challenges highlight that explainability methods developed for domains such as imaging or EHR are not directly transferable to clinical audio. Audio-native evaluation frameworks and human-centered approaches are needed to ensure that XAI methods drive actionable clinical insights.

Methods

Systematic Review Search Strategy

We conducted our systematic review according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Our goal was to survey articles that perform explainability or interpretability analyses of deep learning models applied to clinical voice and speech audio. A comprehensive search was carried out across IEEE Xplore, ACM, Scopus, PubMed, Web of Science, and Nature.

The following search terms were used to create a query, which was adapted for each database:

- Explainability and Interpretability: (”Explainable AI” OR ”interpretable AI” OR ”XAI” OR ”explainable machine learning” OR ”interpretable machine learning” OR ”feature attribution” OR ”model interpretability” OR ”post-hoc explainability” OR ”SHAP” OR ”LIME” OR ”Grad-CAM” OR ”attention” OR ”saliency maps” OR ”Layer-wise Relevance Propagation” OR ”LRP”) AND
- Speech and Acoustic Analysis: (”voice biomarker” OR ”speech biomarker” OR ”speech analysis” OR ”acoustic analysis” OR ”phonetic analysis” OR ”speech processing” OR ”vocal assessment”) AND
- Clinical and Health Context: (”healthcare” OR ”medical diagnosis” OR ”neurological disorder” OR ”neurodegenerative disease” OR ”mental health” OR ”biomarker” OR ”voice disorder” OR ”speech disorder” OR ”voice quality” OR ”vocal quality” OR ”clinical voice analysis” OR ”disordered speech”)

The search was conducted across all databases on February 7, 2025 except ACM, which was search on February 11, 2025. The search included studies published between January 1, 2015 and date of search. In total, 1,426 records were retrieved, which was reduced to 1,348 after removing 78 duplicates.

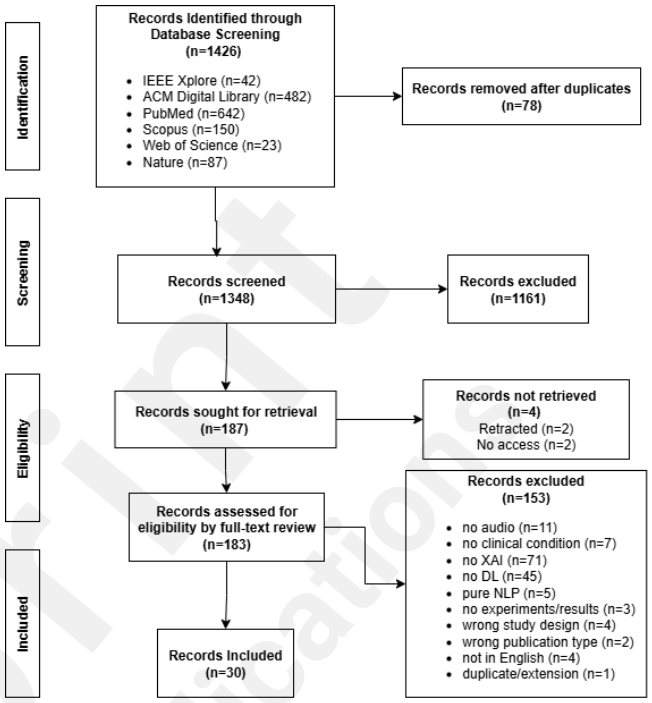
Eligibility Criteria

Articles were selected if they met the following criteria:

- Focused on deep learning models applied to clinical voice and/or speech data.
- Applied explainability or interpretability techniques to the model.
- Reported empirical results or experimental validation.
- Published in peer-reviewed journals or conferences.

Articles were excluded for any of the following reasons:

- Explainability and/or interpretability analysis performed for non-deep learning models.
- Used audio transcripts only (i.e., purely natural language processing methods).
- Used non-clinical audio datasets.
- Not published in English.
- Reviews, theses, or non–peer-reviewed articles.



Study Selection

Results of the search strategy were exported to RAYYAN tool [61] for screening. One reviewer conducted the initial screening of titles and abstracts against the eligibility criteria, and a second reviewer independently verified the filtering decisions. Articles were excluded only if there was clear evidence in the title or abstract that the paper did not meet the criteria; otherwise, they proceeded to full-text review. The screening process resulted in 187 articles deferred for full-text review, of which two were retracted and two were inaccessible. The final set for full-text review was 183 articles.

Full-text screening was performed by two reviewers, with deliberation and discussion used to resolve disagreements. Reasons for exclusion at this stage were documented. The final included set consisted of 30 articles that met all eligibility criteria. A PRISMA flow diagram summarizing the selection process is shown in Figure 1.

Data Extraction

Data extraction was carried out by a single reviewer using a standardized spreadsheet. A second reviewer reviewed the extracted data and deliberated on any uncertainties or discrepancies to ensure accuracy and completeness. The following information was extracted from the final set of included articles:

- Bibliographic information (authors, year),
- Hypothesis/goal,
- Dataset information (clinical condition, number of subjects, acoustic tasks),
- Deep learning methodology (model, hyperparameters, training/validation strategy),
- Model performance,
- Explainability/interpretability strategy,
- Insights gained from explainability/interpretability analysis,
- Validation or support for the explainability/interpretability results.

Results

In this section, we present a taxonomy of XAI approaches for clinical audio from articles identified in our review, categorizing explainability and interpretability methods into four groups: visual explanation methods, feature importance and attribution methods, attention-based methods, and concept detectors or model-intrinsic approaches. We summarize included articles in Table II.

Visual Explanation Methods

Saliency Maps

Saliency maps are a set of XAI visualization techniques that highlight features (e.g. pixels) that are important for the neural network prediction by calculating gradients of the model output (i.e., logits) with respect to the input features. The magnitude of the gradient at each feature indicates its importance to the model's decision. In the resulting heatmap, higher intensity values correspond to high sensitivity features, if sightly perturbed will cause the most significant impact to the model's decision.

Shaikh et al. [4] attempted to explain decision process of 1D-CNN model for multiclass voice disorder classification through saliency maps overlaid on input spectrograms and MFCCs. Although the authors state that the model focuses on high frequency, low amplitude regions, they concede that no conclusive pattern can be derived. that saliency maps were also calculated for an MLP model trained on OpenSMILE ComParE2016 feature set, resulting in the top-5 features for each group.

Figure 1. PRISMA flow diagram

Gupta et al. [62] used a variation of saliency maps called Guided Backpropagation [63] to explain decisions of a 14-layer ResNet trained for dysarthria severity classification. Guided Backpropagation differs from vanilla saliency [64] maps in that gradients are only propagated when both the forward activation and the backward gradient are positive, resulting in cleaner, less noisy, and more visually interpretable maps. The authors concluded that the model focuses on high-energy vowel regions, noting that patients with lower severity of dysarthria exhibit better articulation and clearer phonemes compared to high-severity patients, where the model struggles to delineate phoneme boundaries. This distinction is also illustrated in Figure 2, where vanilla saliency highlights regions with strong gradient responses but little structural detail, whereas Guided Backpropagation produces sharper, edge-like visualizations that reveal more localized spectral patterns.

Gradient-Weighted Class Activation Map

Grad-CAM [65] is visualization technique that produces coarse localization maps (heatmaps) emphasizing important regions in an input image for a given class prediction for CNN-based models.

Grad-CAM is a model-agnostic, post-hoc XAI method such that it does not require any model modifications. The gradients of the logits of

the target class with respect to the final CNN layer feature maps are calculated. These backpass gradients are then global-average-pooled to obtain the neuron importance coefficients or weights which represent the importance of each feature map in the CNN layer for the target class. These coefficients are used to calculate the weighted sum of the forward pass activation maps followed by Rectified Linear Unit operation to keep only feature maps that have a positive influence on the target class. The output is then upscaled to the dimensions of the input image.

Fu et al. [66] leveraged Grad-CAM overlays for visual explanation of schizophrenia classification. The Grad-CAMs highlights reduced high frequency concentration in schizophrenic speech and shows that the model focuses on formants in low-frequency bands below 2000 Hz. These visualizations, as interpreted by the authors, suggest that the Sch-net model is able to identify incorrect articulator placement attributed to schizophrenic speech.

Lee et al. [6] trained an EfficientNet-B4 model to analyze postoperative voice recovery three months after thyroid surgery. Grad-CAM was used as a post-hoc technique to explain the model decisions. The authors simply note that for severe poor voice quality, Grad-CAM exhibits high magnitude activations across a wide frequency range, while for high voice quality, heatmaps showed much less activation. No further clinically relevant interpretation was given.

Peng et al. [3] performed multiclass voice disorder classification OpenL3-SVM model trained on spectrograms of sustained /a/ phonation. Grad-CAM was used to analyze the features learned by the OpenL3 model, where the model seemed to exhibit different activation patterns for each condition. For healthy voice the model focused on low frequency region of the spectrogram, while activations in low- and high-frequency regions were evident for hyperkinetic dysphonia. High-frequency band was deemed important by Grad-CAM for reflux laryngitis. Lastly, heatmaps of hypokinetic dysphonia showed weak activation in a narrow mid-frequency band. Similar to Lee et al., the authors do not attempt to further interpret the results.

Shen and Zhang [67] used a variation of Grad-CAM called time-related (or temporal) Grad-CAM to verify that the model focuses on stutter disfluencies in speech using Wav2Vec2.0 embeddings. Time related Grad-CAM is an adaptation of Grad-CAM design for time-series data, such that it generates localization maps highlighting significant temporal segments in time-series data. Being one of the only works to quantitatively verify outputs of XAI methods, Shen and Zhang verified the accuracy of time-related Grad-CAM output using a temporally annotated dataset where segments are annotated as normal or disfluent [67].

Rojas et al. [59] employed Grad-CAM to explain their ResNet model's decision for classifying mild traumatic brain injury (mTBI). Grad-CAM heatmap overlaid on mel-spectrograms showed the for mTBI speech the model focuses on higher frequency features.

Eigen Class Activation Map

Eigen-CAM [68] is XAI method that produces activation maps comparable to Grad-CAM. However, unlike Grad-CAM, Eigen-CAM is a gradient-free method that calculates saliency maps by taking the first principal component of the fattened feature maps. Additionally, Eigen-CAM is not limited to CNN layers. The main advantage behind this method is that it avoids the noise and potential instability inherent in gradient-based methods, as shown in Figure 2.

Jeong et al. [10] trained and Audio Spectrogram Transformer (AST) on spectrograms of sustained vowels, consonant, and diadochokinesis (DDK) phonations for Parkinson's (PD) disease detection. Eigen-CAM heatmaps showed that the model recognizes high-frequency bands as important for PD detection. This high frequency region is associated with muffled speech and degraded speech clarity exhibited by PD patients.

Feature Importance and Attribution Methods

SHapley Additive exPlanations

SHAP [69] is a model-agnostic explanation algorithm that is grounded in cooperative game theory. It is a unified framework for interpreting model predictions by attributing each feature a contribution score called Shapley values. Shapley values are the average marginal contribution of a feature value across all possible subsets of features. In other words, the SHAP value for each feature quantifies how much the presence/absence of that feature changes the model's output. SHAP is considered one of the popular methods for explaining DNN models due to its intuitive output.

Figure 2. Activation maps show varying patterns for different vision-based methods. Red corresponds to high activation and blue signifies low neural activation or low region influence.

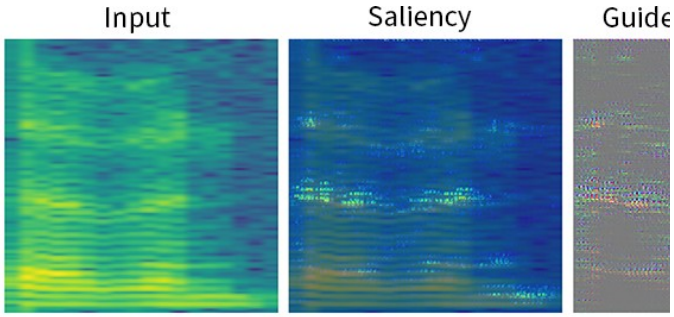
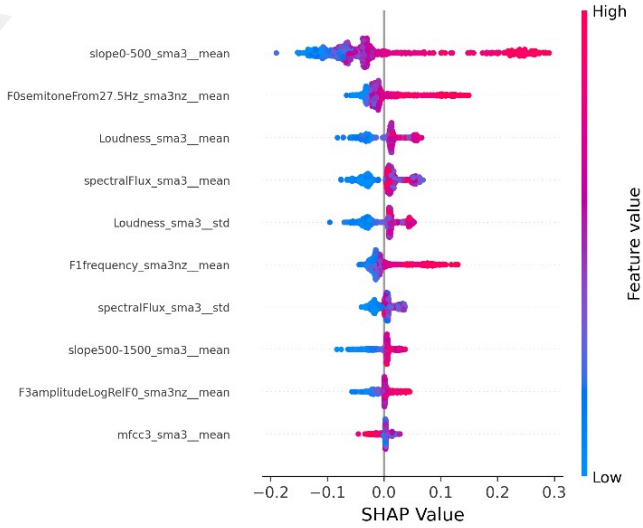


Figure 3 presents a SHAP summary plot showing the contribution of the top acoustic features to the model's predictions. Each dot represents a single sample, with position along the x-axis indicating the SHAP value (i.e., the feature's impact on the prediction). Features are ranked by importance from top to bottom. Color encodes the feature value (blue = low, red = high). For example, higher values of slope0–500_ma3__mean and F0semitoneFrom27.5Hz_ma3nz__mean consistently increase the prediction score, while variability in features like spectralFlux_sma3__std and mfcc3_sma3__mean shows mixed effects depending on their value.

Schultebraucks et al. [70] employed SHAP to explain predictions of a multimodal (audio, video, transcripts) DNN for the purpose of post-traumatic stress disorder (PTSD) and major depressive disorder (MDD) assessment. NLP features dominated the top 20 most influential features. Audio intensity features proved important for predictions of PTSD and MDD, while pitch was among the top features for PTSD status.

Zhang et al. [9] found that noun phrase rate and word rate are the two top positive speech features for dementia, while empty word phrase was the top negative feature. SHAP value of the standard deviation of fundamental frequency indicated that higher pitch variation in speech correlates with higher cognitive function. Moreover, dementia patients showed slower speech with more pauses and higher hesitation ratio.

GradientSHAP [71] is an adaptation of SHAP designed for DNNs that builds on Integrated Gradients (IG). IG attributes feature importance by computing gradients along a straight path from a baseline (or reference input) to the actual input, effectively measuring how much replacing an input feature with its baseline value affects the model's output. A key limitation of IG, however, is its strong sensitivity to the choice



of baseline (e.g., silence, noise, or a mean spectrogram), which can heavily bias the resulting attributions. GradientSHAP mitigates this issue by averaging IG across many random baselines and noisy perturbations, sampling from a distribution of references. This produces feature attributions that are both more stable and less dependent on a single baseline choice, thereby offering a more reliable interpretation of model behavior.

Ditthapron et al. [72] applied GradientSHAP to interpret their cascading GRU model for traumatic brain injury (TBI) classification. At the lexical level, GradientSHAP attributions averaged over word boundaries revealed that filler words such as “um,” “uh,” and “the” showed statistically significant differences between TBI and control groups. In contrast, spectro-temporal analyses indicated that high-frequency bands were highlighted as potentially informative for TBI, though no statistically significant group differences were observed in these features.

Local Interpretable Model-agnostic Explanations

LIME [73] is a model-agnostic, perturbation-based explainability method that learns a simple, interpretable surrogate model, such as a linear models or decision trees, that approximates the predictions of a black-box model locally. It generates slightly perturbed versions of the original data sample and obtains the corresponding predictions from the black-box model. These perturbed instances are weighted based on their proximity to the original input, where closer samples receive higher weights. This weighting ensures that the surrogate model focuses on the local neighborhood of the input, thereby maintaining local fidelity in the explanation.

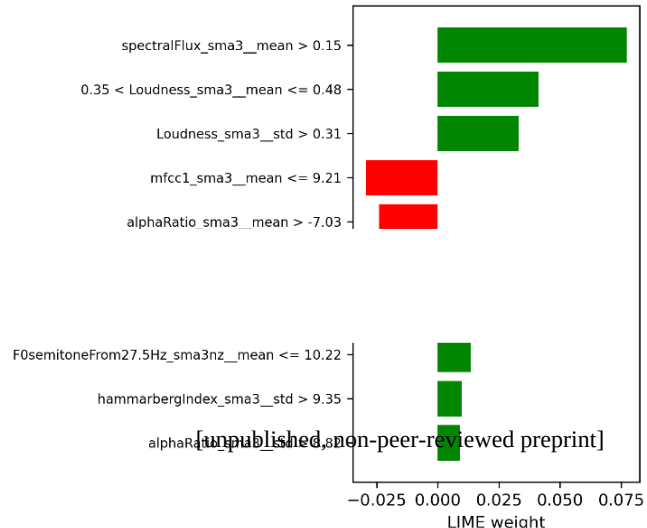
Figure 4 shows a LIME explanation highlighting the contribution of specific acoustic features to a single model prediction. Features with positive weights (green bars) increase the likelihood of the predicted class, while features with negative weights (red bars) decrease it. For this instance, features such as spectralFlux_sma3__mean, Loudness sma3 mean, and Loudness sma3 std strongly support the prediction, whereas mfcc1 sma3 mean and alphaRatio_sma3__mean provide evidence against it.

The work in [9] provided local explanations of the text modality for Alzheimer’s disease assessment. A thousand perturbed samples are generated for each sample in the original dataset. Then, the resulting perturbed dataset is used to train lasso regression surrogate model that locally approximates the predictions of the complex deep learning model. Weight coefficients of the proxy model showed that model focused on meaningful words for healthy controls, while pause fillers, such as “mhm” and “hm”, were highly indicative of Alzheimer’s disease.

While the original LIME paper [73] provided example use cases of the algorithm with text and image data, Wullenweber et al. [74] extended basic LIME for explaining CIdeR [75] model predictions on COVID-19 cough audio, being one of the only XAI methods specifically designed for clinical audio.

CoughLIME [74] uses a loudness decomposition algorithm to decompose the cough audio into interpretable components, where each component contains an individual cough sound. Perturbed versions of the sample are created by selecting subsets of these components. A sparse linear surrogate model is then trained on binary presence/absence vectors of these components, weighted by their similarity to the original sample. CoughLIME reconstructs audio using the most influential components to create sonified or listenable explanations.

Similarly, Gutiérrez-Serafín et al. [76] proposed an adaptation of LIME for audio



called eXplainable Deep Learning MFCCs (xDMFCCs). In this approach, 13 MFCCs are extracted from fixed-length audio segments of one second. The resulting MFCC spectrogram is divided into 13 frequency and 11 time segments, where each superpixel represents a single coefficient over a 90 ms window. Perturbations are performed by randomly generating binary masks to remove superpixels, and as in the original LIME framework, a surrogate regression model is trained on the perturbed data to derive superpixel attributions for each sample. Significant superpixels are then aggregated across the dataset to produce global explanations. Although the authors did not formally evaluate the faithfulness of their method, they applied xDMFCCs to a CNN trained for brain lesion identification. Their results suggested that healthy controls exhibited greater discriminative energy in higher-order coefficients, with crisper articulation and faster phoneme transitions compared to subjects with brain lesions.

Ablation Studies

An ablation study is an experimental framework used to evaluate the contribution of individual components, modules, features, or layers within a complex model. The process usually involves systematic removal of specific parts of the model or input features and observing the resulting impact on model performance. This methodology is especially relevant to understanding the complex and obscure nature of deep learning models. Unlike many other XAI methods, ablation studies do not have a specific explainability modality (e.g., visual or textual) but rather provide insight through performance-based comparisons.

For dysarthria speech recognition, Liu et al. [18] found that audio speed perturbation significantly improved performance compared to tempo and vocal tract perturbation. They further concluded that, among the speaker adaptation techniques investigated, learning hidden unit contributions (LHUC) performed best in significantly lowering word error rates.

In dysarthria severity classification, Joshy and Rajan [77] compared multiple acoustic feature sets under speaker-dependent (SD) and speaker-independent (SID) conditions. MFCCs yielded the best performance in the SD setup, while Constant Q Cepstral Coefficients (CQCCs) were more discriminative in the SID setting. The authors also evaluated speech-disorder–specific feature sets (articulatory, prosodic, glottal, and phonation), finding articulatory features most effective in the SD case, though less discriminative in the SID scenario. Finally, they proposed a two-level learning approach with i-vectors (derived from MFCCs and CQCCs) followed by DNN classification, which achieved the best overall results, outperforming other feature–model combinations.

Yue et al. [20] likewise examined various feature sets for automatic dysarthria speech recognition, combining MFCCs, spectrograms, mel-filterbanks, vocal tract and source excitation features, and raw waveforms. The combination of vocal tract and source excitation signals with spectrogram magnitude yielded the best performance. Additionally, they evaluated speed perturbation as a data augmentation strategy, both with and without fundamental frequency (F0) fixing. Results showed that speed perturbation without F0 fixing significantly lowered WER, while applying F0 fixing offered no performance improvement.

Lahoti et al. [11] proposed a multi-head attention BiLSTM model for Parkinson’s disease detection and evaluated its performance using MFCCs, single frequency filtering cepstral coefficients (SFFCCs), and shifted delta cepstra (SDC). They found that augmenting MFCCs and SFFCCs with SDC significantly enhanced discriminative ability, underscoring the importance of modeling long-term temporal dependencies in speech for PD detection.

Similarly, for Alzheimer’s disease detection, Laguarta and Subirana [12] introduced the Open Voice Brain Model (OVBM), a framework that integrates heterogeneous speech-based biomarkers. They evaluated model performance when fine-tuned on each individual biomarker and found that each one

separated a distinct subset of patients, highlighting their complementary and orthogonal nature. Specifically, memory biomarkers capture deficits in recall and articulation, sentiment biomarkers reflect fluctuations in affect and mood, cough biomarkers convey information about neuromotor alterations, and wake word biomarkers capture attention and speech initiation. Importantly, while each biomarker was informative on its own, their combination provided complementary information, enabling the multimodal model to achieve the highest overall accuracy.

For aphasia severity classification, Herath et al. [78] found MFCC to significantly outperform chroma features, zero crossing rate (ZCR), and mel-spectrogram. Additionally, 20-second audio samples slightly improved performance compared to 10second clips.

In mental health applications, Huang et al. [14] developed a multi-task schizophrenia assessment model based on the Thought, Language and Communication (TLC) Scale and Positive and Negative Syndrome Scale (PANSS), in which they performed feature ablation studies to understand the importance of semantic (BERT), syntactic (ELECTRA), and acoustic (TERA) features. The authors found that BERT semantic features were the most impactful for all scales except PANSS General; the acoustic features had low contribution for the TLC scale but high contribution for PANSS, suggesting that acoustic features are more relevant for emotion-related PANSS items; and syntactic features exhibited moderate contribution. They also evaluated the learning scheme to validate the effectiveness of multi-task (i.e., TLC and PANSS) learning and concluded its superiority to single-task learning.

In contrast, He et al. [79] introduced a dual-branch axial-attention network for schizophrenia detection, WNSA-Net, which integrates wideband and narrowband spectrograms. The dual-branch setup outperformed single-branch variants, underscoring the complementary role of fine temporal resolution (wideband) and detailed frequency resolution (narrowband) in capturing schizophrenia-related speech anomalies. Ablation studies further showed that dilated convolutions were critical for multi-scale feature extraction: smaller dilations captured micro-level acoustic cues such as pitch period, formant structure, and energy, while larger dilations modeled macrolevel patterns such as speech rate and prosodic contours. These findings highlight the importance of modeling speech abnormalities at both fine-grained and long-range levels in schizophrenia detection.

In the context of depression, Zhang et al. [15] conducted ablation studies on pooling strategies, self-attention mechanisms, and input audio length. They found that attention pooling outperformed both max and average pooling in terms of classification performance. Additionally, incorporating self-attention significantly enhanced model accuracy. Regarding input length, performance improved with longer audio clips and plateaued around 7 seconds, suggesting that this duration captures the most informative speech patterns for depression detection.

Attention Mechanisms

Attention mechanisms [80], particularly in transformer-based architectures, have been instrumental in achieving state-of-the-art performance across various domains. Beyond their modeling power, attention provides an intrinsic form of explainability by revealing how the model distributes its focus across different parts of the input sequence. This makes attention a distinct XAI approach: while its outputs can be visualized as heatmaps (creating some overlap with visualization-based methods), attention is fundamentally part of the model's computation rather than an external post-hoc tool.

The self-attention mechanism, which is the most basic form of attention, process an input sequence by representing each token (frame, word, or segment) in three vectors: Query (Q), Key (K), and Value (V) vectors. To calculate the attention for given to a specific token, the Query vector is multiplied by the Key vectors of all other tokens. This captures the

dependencies and relationships between the Query token and all other tokens, the Keys. This score is then normalized and scaled by a softmax function to produce a set of attention weights. These weights are then used to create a weighted sum of all the Value vectors. This transforms the Value vectors, which capture the original semantic meaning of each token, into a context-aware representation. This process is described by the following formula

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

,

where the denominator is a scaling factor and the softmax operation normalizes the dot product to a range of $[0,1]$.

Attention Weights Visualization

A direct approach to using attention for explainability is to visualize the attention weights computed during the forward pass. The resulting heatmaps show how strongly each token attends to others. However, in multi-layered models, these weights reflect only a single step in the computation and may not represent the complete causal information flow.

For example, Wang et al. [81] developed a multimodal DNN with self-attention for classifying Auditory Verbal Hallucination (AVH) severity from diary recordings (audio, text, and sensor data). They found textual transcripts to be more informative than audio and used attention heatmaps overlaid on transcripts to highlight words linked to high AVH severity, such as clauses describing auditory distress or interference.

Attention Rollout

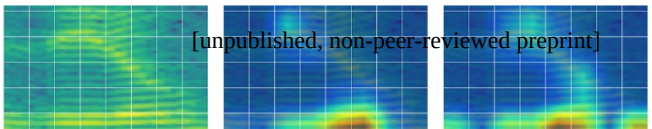
Attention rollout [82] addresses the limitation of single-layer visualization by aggregating attention weights across all layers and heads to produce a relevance matrix which is a holistic representation of how information flows through the model. This is achieved by recursively multiplying attention matrices layer-by-layer, propagating influence scores from the input to deeper layers. Figure 5 shows an input spectrogram with attention-based explanations. Simple attention highlights localized regions but produces fragmented focus, whereas attention rollout provides smoother, more coherent attribution, aligning with formant-rich regions of the signal.

Lau et al. [83] applied attention rollout to an Audio Spectrogram Transformer for voice disorder classification (organic, inorganic, and healthy controls). Their analysis showed that the model attended to specific phonemes rather than merely high-energy spectral regions (e.g., fundamental frequency and formants). However, after fine-tuning, consistent patterns were harder to detect, underscoring that attention interpretability can vary across training stages.

Concept-Based and Model-Intrinsic Approaches

Concept-Detectors

This class of XAI methods enhances the interpretability of neural network models either by identifying whether specific neurons or groups of neurons respond to high-level, human-understandable concepts or by embedding architectural components that offer inherent transparency. Unlike post-hoc methods such as visual explanations or feature attribution techniques, concept-based approaches incorporate interpretability directly into the model or establish links between internal model structures (e.g., neurons, layers) and semantically meaningful constructs. This is known as concept alignment, where predefined domain knowledge is mapped to neurons or



subnetworks that consistently activate in the presence or absence of the clinical concepts, thus providing clinicians with explanations grounded in domain knowledge.

For example, Abderrazek et al. [84] examined neuron activations in the fully connected layers of a CNN-based French phone classifier trained on healthy speech. For each phonetic feature, the authors computed a histogram of normalized neuron activations. They then defined a scoring metric that measured the difference between the median activation of a neuron when the phonetic feature was present and when it was absent. Neurons with scores above an empirically defined threshold were considered concept detectors for that phonetic feature. This phone classifier was subsequently applied to speech from individuals with head and neck cancer, enabling intelligibility assessment based on the deviation of phonetic feature activations from the healthy reference patterns.

Mathad et al. [7] extended the concept-detector approach to hypernasality assessment in children with cleft palate. They trained a DNN nasality model on healthy speech to classify nasal consonants (NC), oral consonants (OC), nasalized vowels (NV), and oral vowels (OV). Unlike Abderrazek et al. [84], this work did not analyze neuron-level activations for individual features but instead relied on the model's output probabilities to derive the hypernasality score. These probabilities were combined into an Objective Hypernasality Measure (OHM). Subjects read predefined text samples in which the expected occurrence of nasal consonants and vowels was known. The OHM quantified the level of detected nasality by comparing nasal and oral categories against the expected nasality for each phrase and was validated against perceptual ratings from expert clinicians.

Model-Intrinsic Representations

Unlike concept detectors, model-intrinsic representation methods explain neural network behavior by interpreting internal structures or components—such as weights, filters, or intermediate layer outputs, without relying on predefined, domain-specific concepts. These methods aim to provide insight into how input data is transformed as it propagates through the model, typically through architectural constraints (e.g., filter parameterization, sparsity), statistical analysis of activations, or visualization of internal feature maps.

While the work in [5] does not implement a strict concept detector (i.e., it does not identify predefined semantic clinical features), the SincNet architecture used in the study learns physically interpretable frequency filters—specifically, low and high cut-off frequencies of bandpass filters. Unlike conventional CNN filters, SincNet filters are parameterized to directly model frequency bands, allowing them to more effectively emphasize formant frequencies, which are critical acoustic biomarkers for voice disorder classification. By preserving the structure of these formants, SincNet contributes to improved model interpretability and relevance in clinical acoustic analysis.

Vasquez-Correa et al. [13] examined onsets and offsets from´ diadochokinesis speech tasks for Parkinson's disease detection. Motivated by the observation that PD speech exhibits slower voiced and unvoiced transitions, the authors analyzed feature maps from the second and fourth CNN layers using Kruskal–Wallis H-tests, finding significant group differences. Results suggest that the CNN learned internal representations aligned with disease-relevant speech impairments.

Representation Visualization with t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) [85] is a nonlinear dimensionality reduction technique widely used to project high-dimensional learned representations (e.g., embeddings) into an intuitive 2D or 3D visualization. t-SNE maps high-dimensional points into a low-dimensional space while preserving local neighborhood relationships, such that points close in the original space remain so in the projection. The algorithm

computes pairwise similarities of data points in the high-dimensional space, expressed as probability distributions where closer points receive higher probabilities. In the lowdimensional embedding, a similar distribution is computed using a Student's t-distribution to alleviate the "crowding problem" and separate clusters more effectively. t-SNE then arranges points to minimize the Kullback–Leibler divergence between the high- and low-dimensional distributions. This technique is used pervasively in the literature for representation-level interpretability, as it reveals latent structures learned by deep models.

Lau et al. [83] leveraged t-SNE to visualize the latent space of two model configurations: a pretrained Audio Spectrogram Transformer with a linear classifier and an AST fully-finetuned on the Saarbrucken Voice Dataset. The t-SNE plots revealed¨ that in the pretrained AST, embeddings clustered primarily by gender rather than pathology. In contrast, finetuning on pathological data re-oriented the model's latent space toward disease-relevant information. Similarly, et al. Hiresˇ [84] demonstrated the effectiveness of their proposed multiple finetuning strategy, where t-SNE showed better clustering of Parkinson's and healthy voice samples. Joshy and Rajan [75] used t-SNE to examine latent spaces learned from MFCC-based and CQCC-based i-vectors, finding that MFCC-based i-vectors produced clearer clustering of dysarthric severity levels. Likewise, in Kim et al. [85], t-SNE plots for pooled CNN features for laryngeal disease classification showed that benign disease features overlapped with those of cancer and paralysis. Moreover, Geng et al. [19] showed that their proposed singular value decomposition (SVD)-based spectro-temporal deep features yielded superior separation of nonaged and elderly subjects compared to i-vectors and x-vectors. Yue et al. [20] examined latent representations across different layers of their automatic dysarthric speech recognition (ADSR) model, finding that layers closer to the output were more invariant to speech pathology. Their t-SNE analysis also demonstrated that the model became progressively less sensitive to speaker variations (e.g., speaker ID and gender) as training epochs increased. Finally, Lee et al. [86] used t-SNE to compare feature spaces derived from eGeMAPS acoustic descriptors and autoencoder embeddings of infant vocalizations for autism spectrum disorder (ASD) detection. Whereas eGeMAPS features showed heavy overlap between ASD and typically developing (TD) groups, the autoencoder embeddings redistributed the latent space to yield clearer separation between ASD and TD vocalizations.

In conclusion, t-SNE does not provide explainability in the traditional sense of XAI; however, it offers visual representation-level interpretability by revealing how model layers encode clinical and speaker-related information.

Figure 3. Example of SHAP explanation beeswarm plot. Every data sample is represented by a point, where color represents the feature value. Positive SHAP value indicates that the feature increases the likelihood of target class, while negative SHAP value signifies a decrease in likelihood of the target class.

Figure 4. Example of LIME explanation. Positive LIME weight (green) indicates that the feature pushes the model toward predicting the target class, while a negative weight pushes the model away from the target class.

Figure 5. Attention rollout accumulation of attention weights across layers results in smoother activation for formants compared to simple attention visualization.

Table 2. Summary of explainability and interpretability studies in healthcare with voice and speech audio.

Work	Application	Model	XAI Method	XAI Insight	XAI Output Validation/ Support
Shaikh et al. [4]	Voice disorder classification	MLP, 1D-CNN	Saliency maps	Top five LLD features extracted per condition; saliency maps emphasized high-frequency and high-amplitude spectrogram regions, though no consistent pattern emerged.	Cross-dataset evaluation.
Gupta et al. [62]	Multiclass dysarthria severity classification	ResNet-14	Guided Backpropagation	For low-severity cases the model focused on high-energy vowel regions (clearer phoneme boundaries); for high-severity cases activations were diffuse, consistent with phonatory instability and temporal smearing.	Referenced supporting studies.
Fu et al.	Schizophrenia	Sch-Net (CNN	Grad-CAM	Reduced high-frequency energy focus	Referenced

[66]	vs. healthy (binary)	with skip connections, CBAM)		(<5kHz) and emphasis on low-frequency formant stripes (< 2kHz), suggesting articulation errors (voiced instead of unvoiced consonants) aligned with blunted affect.	supporting studies.
Lee et al. [6]	Postoperative vocal recovery (GRBAS)	EfficientNet-B4 (CNN) + LSTM	Grad-CAM	Different activation patterns by GRBAS level: attention to low bands (0–2 kHz, formants), mid bands (2–4 kHz, harmonics/noise), and temporal regions (pauses/breathy segments).	—
Peng et al. [3]	Multiclass voice disorder classification	OpenL3 + SVM	Grad-CAM	Disorder-specific band focus: healthy in low-frequency regions; hyperkinetic dysphonia in both high and low bands; reflux laryngitis in high bands; hypokinetic dysphonia weak across bands.	—
Rojas et al. [59]	Mild traumatic brain injury detection	ResNet	Grad-CAM	Model down-weighted low frequencies and emphasized high-frequency regions for mTBI predictions.	Notes the need for clinical expertise to interpret heatmaps.
Shen and Zhang [67]	Speech disfluency detection	Multi-adversarial neural network	Time-related Grad-CAM	Highlighted frames aligned with annotated disfluencies; distinct disfluency types exhibited different temporal Grad-CAM patterns; results supported wav2vec2 capturing meaningful temporal cues.	Disfluencies annotated by non-clinicians.
Jeong et al. [10]	Parkinson's disease classification	AST, EfficientNet-based	Eigen-CAM	Emphasis on higher-frequency bands associated with muffled/degraded speech in PD.	Referenced supporting studies.
Schulte-braucks et al. [70]	PTSD/MDD (binary & severity) at 1 month	Multimodal DNN	SHAP	Key predictors: reduced pitch/intensity (voice), negative affect/self-focus (language), flat affect (face).	Referenced supporting studies.
Ditthapron et al. [72]	TBI vs. healthy (binary)	pSinc + cascading GRU (cGRU)	GradientSHAP	High attribution to filler words and high-frequency spectral patterns; formants were salient for TBI detection.	—
Zhang et al. [9]	Dementia detection	BiLSTM + multi-head attention (audio); DistilBERT + 1D-CNN + cross-modal attention	SHAP (LLDs), LIME (text/audio)	Linguistic features (e.g., noun phrase rate, word rate) most predictive among LLDs; AD speech showed more fillers/pronouns/function words, lower energy and slower rate; attention emphasized disfluent/low-energy segments.	Referenced supporting studies.
Gutiérrez-Serafíın et al. [76]	Brain lesion detection	CNN	xDMFCCs (LIME)	MFCC-1/2 (energy/clarity) most important; controls showed earlier/clearer articulation with more discriminative energy in higher-order MFCCs; patients showed delayed onset, slower articulation, and longer phoneme duration.	Referenced supporting studies.

Table 2. Summary of explainability and interpretability studies in healthcare with voice and speech audio.

Work	Application	Model	XAI Method	XAI Insight	XAI Output Validation/ Support
Liu et al. [18]	Automatic dysarthric speech recognition	TDNN-HMM; CTC; LAS encoder–decoder	Ablation studies	TDNN outperformed CTC on moderate–severe dysarthria (temporal dependencies matter); speaker adaptation substantially reduced WER (per-speaker customization is beneficial).	—
Huang et al. [14]	Schizophrenia severity (classification/regression)	Transformer embeddings (BERT, ELECTRA, TERA) + BiLSTM + FC	Ablation studies	BERT most crucial for TLC and for PANSS scales except PANSS-General; ELECTRA contributed moderately; TERA low for TLC but important for PANSS, especially positive symptoms and general psychopathology.	—
Herath et al. [78]	Aphasia severity classification	DNN	Ablation studies	MFCC-DNN performed best; ZCR-DNN performed worst.	—
He et al. [79]	Schizophrenia detection	WNSA-Net	Ablation studies	Wideband and narrowband spectrograms provided complementary information; dilated convolutions captured micro-level (pitch, formants) and macro-level (rate, prosody) cues.	Cross-dataset evaluation; referenced studies.
Lahoti et al. [11]	Parkinson's disease detection	Multi-head attention BiLSTM	Ablation studies	Augmenting cepstral features with shifted delta cepstra improved performance over SFFCCs alone, highlighting long-term temporal dependencies for PD detection.	—
Zhang et al. [15]	Depression detection	Wav2Vec + 1D-CNN + LSTM	Ablation studies	Models with self-attention performed better; Wav2Vec embeddings were superior; 7-second segments worked best, suggesting emotion is concentrated in short spans.	—
Laguarta and Subirana [12]	Alzheimer's disease detection	Open Voice Brain Model (GNN)	Ablation studies	Memory/fluency features dominated early-stage AD detection, with sentiment/prosody also contributing; AD often showed high saliency in respiratory control, disfluency, or memory-related patterns.	Referenced supporting studies.
Joshy and Rajan [77]	Dysarthria severity classification	DNN, CNN, GRU	Ablation studies; t-SNE	MFCCs best in speaker-dependent setup; CQCCs generalized better to unseen speakers; articulatory features strongest among disorder-specific sets but weaker in speaker-independent setup; MFCC-based i-vectors showed clearer class clustering in t-SNE.	Multi-dataset evaluation; referenced studies.
Yue et al. [20]	Automatic dysarthric speech recognition	Multi-stream CNN + LiGRU	Ablation; CNN filter analysis; t-SNE	Best WER from combining spectrogram magnitude with vocal tract and excitation streams; speed perturbation without F0 fixing improved WER; filters fed with vocal-tract signals emphasized low quefrencies, whereas excitation filters suppressed them; t-SNE showed progressive dysarthric/typical separation and reduced gender clustering over training.	—
Wang et al. [81]	Auditory verbal hallucination detection	Uni-modal BiGRU; multimodal self-attention DNN	Simple attention visualization	Attention prioritized clauses describing distress, influence, or interference, aligning higher weights with higher AVH severity.	Referenced supporting studies.

Table 2. (continued) Summary of explainability and interpretability studies in healthcare with voice and speech audio.

Work	Application	Model	XAI Method	XAI Insight	XAI Output Validation/ Support
Lau et al. [83]	Voice disorder detection	AST	t-SNE; attention rollout	Model focused on specific phonemes (e.g., /ɔ/ and the segment "/e/ /s/ /i/ /n/") rather than merely high-energy regions.	Referenced supporting studies.
Abderrazek et al. [84]	Head and neck cancer intelligibility	CNN	Concept detector network	In FC1 no neurons detected phonetic features; from subsequent layers to output, the number of phonetic feature detectors increased by a factor of 1.75 for subsequent layers layer.	—
Mathad et al. [7]	Hypernasality assessment (children with cleft palate)	DNN	Concept detector network	A DNN nasality model estimated posterior probabilities for nasal consonants, oral consonants, nasalized vowels, and oral vowels; these were combined into an Objective Hypernasality Measure that quantified detected nasality against expected nasality per phrase.	Cross-dataset evaluation.
Hung et al. [5]	Voice disorder classification	SincNet (CNN-based)	Sinc filter analysis	SincNet filters emphasized F1/F2 more clearly than standard CNN filters, preserving formant structure and energy in 500–3000 Hz bands.	—
Vasquez-´ Correa et al. [13]	Parkinson's disease assessment	CNN	Feature map analysis	Feature maps showed filters highlighting speech transitions (syllable onsets/offsets); many filters in layers 2 and 4 differentiated PD from healthy controls.	Referenced supporting studies.
Lee et al. [86]	ASD detection in infants	BiLSTM with autoencoder	t-SNE	Autoencoder embeddings yielded clearer ASD vs. TD separation than eGeMAPS with BLSTM.	—
Geng et al. [19]	Automatic dysarthric speech recognition	TDNN; Conformer	t-SNE	SVD-based spectro-temporal deep embeddings showed better separation of dysarthric vs. typical speech than i-vectors/x-vectors.	Cross-language, multi-dataset evaluation.
Kim et al. [87]	Laryngeal disease classification	ResNet-50	t-SNE	Pooled CNN features for benign disease overlapped with cancer and vocal cord paralysis, explaining reduced multi-class performance.	—

Discussion

Interpretation of Explanations

Although current XAI methods provide preliminary insight into the inner workings of clinical audio models, the interpretation of these explanations is often not rigorously validated. Many studies rely on a handful of samples to interpret local explanations (e.g., saliency maps or attention weights), without accompanying quantitative or statistical analysis. This raises concerns about over-interpretation, where claims of clinical relevance are drawn from anecdotal or visually compelling results that may not generalize. Such practices risk misleading downstream applications or clinical adoption.

Some studies attempt to corroborate their interpretation of model explanations with clinically established findings; however, these interpretations are not always accurate. For instance, Figure 5 shows attention visualization where the model seems to focus on formants; however, several works have demonstrated that attention weights are not inherently faithful explanations of model behavior and often lack correlation with gradient-based importance measures which may result in misleading interpretations [88-91]. Such instances risk producing plausible but incorrect interpretations. Furthermore, the tendency to interpret explanations in line with existing medical literature may lead to confirmation bias, where researchers overlook alternative explanations or novel insights that deviate from conventional understanding [35]. In the absence of rigorous validation, explanations that merely appear consistent with known clinical markers are more

likely to be accepted uncritically, hindering the discovery of new medical knowledge or deeper understanding of model behavior.

Other works are unable to draw conclusive assessments about their findings. Some other works do not relate the results to current medical knowledge all together, missing an opportunity to contextualize or validate their findings in a medically relevant framework. Ultimately, there is a need to standardize the evaluation and validation of XAI explanations both qualitatively and quantitatively to ensure their reliability and clinical utility.

One approach to validating XAI explanations in clinical contexts involves incorporating clinician expertise to assess the plausibility and medical relevance of the model's reasoning. For example, Rojas et al. [59] emphasized the necessity of clinical expertise to interpret Grad-CAM heatmaps in the context of mild traumatic brain injury, suggesting that comparison with expert annotations could improve the interpretability and reliability of the explanations.

Another approach for XAI validation can be done by comparing XAI explanations with expert annotations. Shen and Zhang [67] adopted a similar validation method, although annotators were not clinically trained. Another important strategy is to evaluate the consistency of explanations across inputs from similar subjects and, perhaps, across datasets. Shaikh et al. [4] evaluated the saliency analysis results on multiple datasets and achieved consistent results. Similarly, Geng et al. [19] evaluated the efficiency of their proposed method across English and Chinese dysarthric datasets.

While perturbation-based techniques are widely used in the literature as a means of generating explanations (e.g., LIME, SHAP, xDMFCC), perturbation-based validation is far less common. It involves masking or altering the features identified by the XAI method as important, then observing the resulting changes in model output [92]. If performance drops significantly, this supports the relevance of the identified features, thereby providing stronger evidence of faithfulness. Additionally, fidelity and faithfulness metrics offer complementary, quantitative means to assess explanation validity by analyzing changes in prediction performance after removing or modifying top-ranked features. For example, Intersection over Union (IoU) is often used in image-based XAI to measure the overlap between highlighted regions in an explanation and ground-truth annotations. Similarly, insertion and deletion tests evaluate whether the model behaves consistently with the explanation. Refer to Coroama et al. [93] for a comprehensive overview of XAI evaluation metrics.

Complexity-Transparency Trade-off

There is an inherent trade-off between model complexity, input representation, and explainability in clinical audio-based AI systems. In particular, the interpretability of model explanations is closely linked to the degree of semantic transparency in the input features. We categorize these representations into four levels: (1) low-level acoustic descriptors (e.g., jitter, shimmer, harmonic-to-noise ratio), (2) raw time-frequency representations (e.g., spectrograms, mel-spectrograms), (3) coefficient-based features (e.g., MFCCs), and (4) transformer-based embeddings (e.g., Wav2Vec, HuBERT).

Models trained on manually extracted low-level descriptors (LLDs) offer the highest degree of interpretability, as these features capture directly established clinical biomarkers or constructs. Explanations from models trained on LLDs can be relatively easily understood using feature attribution methods like SHAP or LIME, which rank the influence of clinically meaningful variables. This alignment between model explanations and clinical knowledge contributes to the popularity of such methods in the medical AI domain.

While MFCCs are also hand-engineered features, they present interpretability challenges due to their abstract and decorrelated nature. The coefficients do not correspond directly to intuitive phonetic or physiological phenomena. Although recent techniques like xDMFCC attempt to provide coefficient-level explanations across time, and Tracey et al. [94] sought to demystify MFCCs as vocal biomarkers by correlating them with known LLDs, the clinical relevance of these explanations remains limited. Despite efforts to position MFCCs as vocal biomarkers, they largely obscure the internal

workings of deep models and contribute to the opacity of clinical audio systems.

Spectrograms offer a more interpretable alternative. As visual time-frequency representations, they capture dynamic changes in energy that are often physically and clinically meaningful, such as vocal formants or pauses. Trained speech pathologists and clinicians can interpret spectrograms directly, making visualization-based explanations (e.g., saliency maps) more accessible and potentially clinically actionable compared to MFCCs. Although, visual explanations of spectrograms might inform about the time segment most important for detection of a condition and can highlight relevant frequency bands, more granular visual explanations are typically not feasible.

Finally, state-of-the-art systems increasingly rely on transformer-based embeddings such as Wav2Vec and HuBERT, which yield task-optimized, high-dimensional representations learned from raw waveforms. While these embeddings deliver substantial performance gains, they are the most difficult to interpret due to high-dimensional, nonlinear abstraction from both low-level acoustic features and clinically grounded descriptors. As a result, explanations tend to be less granular, less transparent, and harder to align with clinical reasoning.

Misalignment of Explanations with Stakeholder Needs

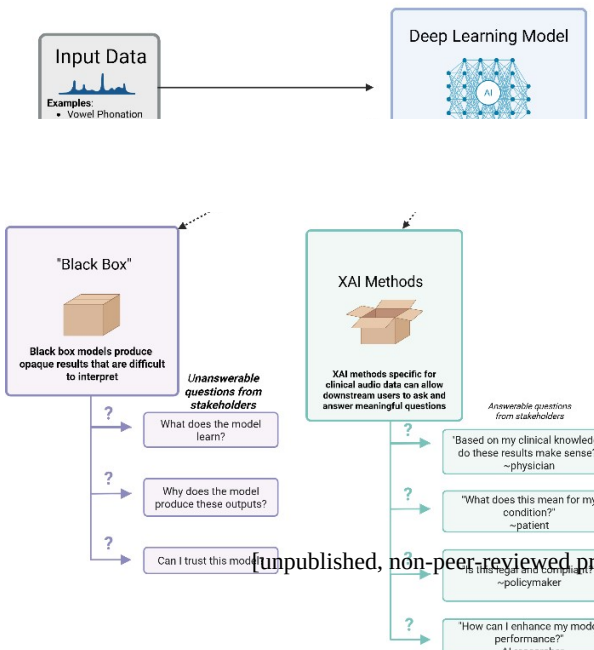
A recurring theme in the literature is the misalignment of the explanation form or modality and the needs of the end-users or stakeholders. Many works produce explanations targeted for highly technical audiences (i.e., AI researchers and developers), without taking into consideration the interpretive frameworks of clinicians, patients, and regulators and policy makers. While such explanations offer insight to AI developers and researchers, critical for understanding model behavior, enhancing performance, and ensuring reliability, these explanations are often too technical, abstract, or detached from domain-specific language to be directly actionable in clinical decision making. Figure 6 highlights the central challenge of stakeholder–explanation misalignment in clinical voice and speech audio AI, illustrating how diverse stakeholders hold distinct expectations and informational needs, and emphasizing the necessity of tailoring explanation strategies accordingly.

Explanations that focus on algorithmic mechanisms poorly convey information in terms of established diagnostic criteria or clinical reasoning processes [35]. Patients, with their highly limited technical and clinical knowledge, are less likely to draw relevant insights from such technical explanation modalities such as activation maps and feature attribution plots [95]. Similarly, regulators and policy makers prefer explanations that offer transparent, auditable decision pathways to assess compliance, fairness, and accountability [96].

This gap is partly a result of explanation design being driven primarily by XAI method availability rather than user requirements. The literature discussed here adapts generic explainability techniques to clinical audio tasks without taking into consideration the information needs, domain expertise, or cognitive constraints of their target users. This highlights the imperative need for human-centered, context-aware design of clinical audio explainability methods in which concerned stakeholder participate provide valuable input for suitable explanation modalities and content. For this reason, in the remainder of this section, we provide a detailed discussion into the goals and needs of these stakeholders.

Clinicians

Trust from clinicians and healthcare professionals (HCPs) is critical for the successful



adoption of AI systems, particularly in high-stakes settings such as intensive care units or voice pathology assessment [97-101]. To support informed decision-making, HCPs require explanations that are grounded in established medical knowledge and aligned with clinical reasoning [97, 102]. They are particularly interested in understanding which features most influenced a model's recommendation and in evaluating the associated uncertainty or confidence [35]. Such explanations not only assist experienced clinicians in resolving ambiguous cases but also support less experienced HCPs in building domain knowledge [35]. Rather than receiving opaque AI-generated actions, HCPs prefer evidence-based suggestions that integrate transparently with their workflow [100]. This interaction also enables clinicians to identify potential errors, biases, or inconsistencies in model reasoning or training data, thereby ensuring high quality patient care. Furthermore, given that HCPs often work in fast-paced and cognitively demanding environments, explanations must be presented in a clear, concise manner and be seamlessly integrated into existing clinical systems [100, 103].

Patients

Patients, as lay users with limited technical and medical domain knowledge, have a specific set of requirements for interacting with AI systems. Primarily, patients' trust in AI is paramount, particularly when its recommendations directly impact their health [37] [[101]. XAI can foster trust between patients and AI systems, as well as between patients and healthcare providers, by offering clear and understandable explanations for AI-generated decisions [96]. Moreover, patients' trust in AI is often influenced by their healthcare provider's confidence in the system [95]. At the same time, patients must be made aware of the capabilities and limitations of AI to prevent over-reliance [97]. Furthermore, overly detailed or technical explanations may lead to cognitive overload and are generally not preferred. Instead, explanations should be intuitive and tailored to the patient's level of understanding, providing concise, relevant information about their health conditions and potential treatment options [104].

Regulators and Policy Makers

Regulatory bodies and policy makers (RPMs) are critical stakeholders who play a pivotal role in ensuring the responsible, ethical, and safe deployment of AI in the healthcare sector [37]. Their focus lies primarily on overarching concerns such as legal compliance, transparency, and societal implications, distinct from the operational needs of clinicians or the personal needs of patients. One of RPMs' main objectives regarding AI-based medical systems is to assess reliability, fairness, and bias mitigation [95, 105]. This involves a thorough evaluation of model specifications, data, and performance testing, as well as analysis of potential failure scenarios and the risk of misleading or harmful outputs [95, 105]. These assessments are essential for weighing the benefits and risks prior to regulatory approval. RPMs emphasize the need for transparency, interpretability, and explainability in AI systems and incorporate these principles into regulatory frameworks for trustworthy AI [96, 106, 107]. For example, the European General Data Protection Regulation (GDPR) includes provisions such as the "right to explanation" [45, 46]. However, they also acknowledge that the level of required explainability may vary depending on the clinical context and associated risk [106].

AI Researchers and Computer Scientists

AI researchers and developers (AI-RnDs) are foundational stakeholders responsible for designing, building, and maintaining legally compliant, reliable, and responsible AI systems, along with explainability and transparency techniques tailored to the needs of various stakeholder groups [108-110]. AIRnDs primarily use such techniques for debugging and monitoring model behavior, enhancing reliability and performance [108, 111], and identifying and mitigating bias [98], thereby ensuring fairness. Some of the methods discussed earlier, such as ablation studies and t-SNE, are mainly used by AI-RnDs for these technical purposes rather than for direct clinical interpretation.

However, AI-RnDs face numerous challenges in designing XAI systems. They must balance the trade-off between model performance and explainability [112], a task further complicated by the lack of a unified, clearly defined understanding of explainability across policy documents and academic disciplines [45]. As a result, they must navigate diverse and often conflicting expectations from multiple stakeholder groups. To address these challenges, AI-RnDs increasingly adopt iterative, human-centered design approaches that involve continuous engagement with clinicians, patients, and regulators to ensure that XAI solutions align with varied backgrounds, goals, and informational needs [96, 113].

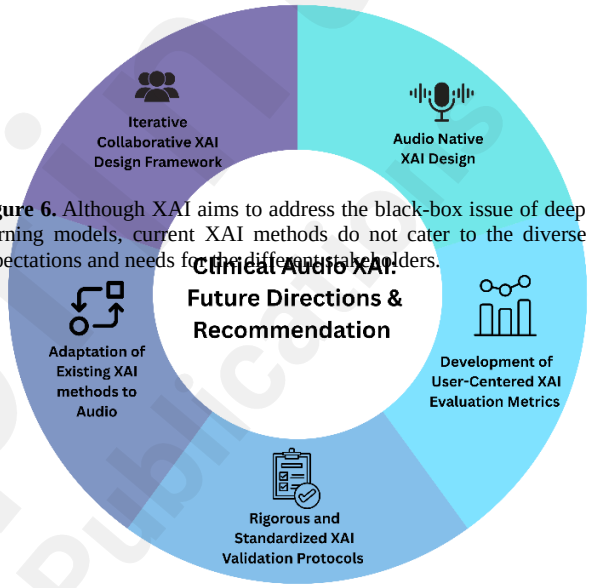


Figure 6. Although XAI aims to address the black-box issue of deep learning models, current XAI methods do not cater to the diverse expectations and needs for the different stakeholders.

Future Directions

In this work, we have discussed several limitations of current methods and opportunities for advancing interpretability and explainability in clinical audio-based deep learning systems. Therefore, in this section we suggest future directions and recommendations for advancing XAI for clinical audio, summarized in Figure 7

Most works in the literature implemented explainability and interpretability methods that fall into one of the four previously discussed groups. However, there are other techniques popular in related domains that have rarely been adapted for clinical audio-based systems. For example, example-based and counterexample-based explainability methods are
widely used in areas such as audio-based emotion recognition [57]. Example-based explanations present representative cases from the training set to contextualize predictions, while counterfactual-based methods explain model behavior by minimally altering a sample so that it belongs to a different class. Additionally, leveraging large language models to produce textual explanations (e.g. question-and-answer) has been an area of interest for medical imaging [114] and potentially transferable to clinical audio.

Figure 7. Key future directions and recommendations for clinical audio XAI.

Our work also identified a lack of rigorous validation of XAI explanations. Many studies draw conclusions based on only a small number of samples, and explanations for misclassified samples are often not evaluated at all. Future work should employ both quantitative measures (i.e., objective metrics), such as fidelity, sensitivity, perturbation-based testing, and cross-dataset explanation consistency, and qualitative measures (i.e., human-centered evaluation), such as expert annotation comparisons and inter-rater agreement.

Moreover, most works in the literature used XAI

techniques originally designed for image or tabular data, as shown in Table I. There is a clear need for domain-specific approaches suited to the temporal–spectral nature of audio data and the nuanced manifestations of particular diseases or conditions. Such methods should aim to map abstract features, such as deep transformer-based embeddings, to established clinical constructs, thereby bridging the gap between model outputs and actionable clinical insights.

Additionally, the misalignment of explanations for different stakeholders underscores the importance of an iterative, collaborative design process in which AI developers engage with clinicians, patients, and regulators from the outset to develop explanations that are understandable, clinically relevant, and operationally feasible. Such a design framework ensures that clinicians can verify explanations against established clinical knowledge, patients can provide feedback on clarity and the appropriate level of detail, and regulators can confirm that outputs meet auditability, transparency, and liability attribution requirements.

Finally, throughout this systematic review, we encountered particular difficulty in quantifying the degree of explainability and practical utility of the explainability methods discussed, especially in relation to different stakeholders. Table I provides our subjective classification of ease of perception of the presented methods for the stakeholders. The review in [44] found only six studies that evaluated the effectiveness of XAI in clinical applications. Future work should explore composite evaluation frameworks that integrate objective indicators (e.g., performance improvement, error detection rates) with subjective measures (e.g., perceived clarity, intuitiveness, trust, and satisfaction). Such metrics would allow for the comparison of explanation methods not only in terms of model alignment but also in terms of real-world usability and impact.

Conclusion

In this literature review, we examined the explainability and interpretability of deep learning models for clinical voice and speech, emphasizing their role in fostering trust for integration into clinical workflows. We categorized the most prevalent approaches into four groups: visualization-based methods, feature attribution techniques, attention-based explanations, and concept-detectors/model-intrinsic approaches. We also outlined challenges unique to clinical audio explainability, including the complexity of audio representations, temporal dependencies, and nuisance variability.

Across the literature, we highlighted key limitations: the lack of rigorous validation of explanations, the trade-off between data interpretability and model performance, and frequent misalignment between explanation modalities and stakeholder needs. For future design and implementation of explainable AI for clinical voice and speech audio, we stress the importance of developing validation frameworks, balancing performance with interpretability, and designing human-centered explanations that align explanations with clinical reasoning and decision making. Ultimately, progress in explainable AI for clinical audio will require sustained collaboration across technical, clinical, regulatory, and lay stakeholders to ensure explanations are both trustworthy and clinically meaningful.

Acknowledgements

This project is part of the Bridge2AI-Voice program funded by the NIH Common Fund #**3OT2OD032720-01S2.** Ebraheem was responsible for conception, design, and main manuscript preparation. Torghranegar, Bensoussan, and Templeton were responsible for supplemental materials as well as overall review and editing. All authors reviewed and approved the final version of the manuscript.

Conflict of Interest

None declared.

References

1. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for Health: The Use of Vocal Biomarkers

from Research to Clinical Practice. Digital Biomarkers. 2021;5(1):78-88. doi: 10.1159/000515346.
2. Ramanarayanan V, Lammert AC, Rowe HP, Quatieri TF, Green JR. Speech as a Biomarker: Opportunities, Interpretability, and Challenges. Perspectives of the ASHA Special Interest Groups. 2022;7(1):276-83. doi: doi:10.1044/2021_PERSP-21-00174.
3. Peng X, Xu H, Liu J, Wang J, He C. Voice disorder classification using convolutional neural network based on deep transfer learning. Scientific Reports. 2023 2023/05/04;13(1):7264. doi: 10.1038/s41598-023-34461-9.
4. Shaikh AAS, Bhargavi MS, Naik GR. Unraveling the complexities of pathological voice through saliency analysis. Computers in Biology and Medicine. 2023 2023/11/01/;166:107566. doi: https://doi.org/10.1016/j.compbiomed.2023.107566.
5. Hung C-H, Wang S-S, Wang C-T, Fang S-H. Using SincNet for Learning Pathological Voice Disorders. Sensors [Internet]. 2022; 22(17).
6. Lee JH, Lee CY, Eom JS, Pak M, Jeong HS, Son HY. Predictions for Three-Month Postoperative Vocal Recovery after Thyroid Surgery from Spectrograms with Deep Neural Network. Sensors [Internet]. 2022; 22(17).
7. Mathad VC, Scherer N, Chapman K, Liss JM, Berisha V. A Deep Learning Algorithm for Objective Assessment of Hypernasality in Children With Cleft Palate. IEEE Transactions on Biomedical Engineering. 2021;68(10):2986-96. doi: 10.1109/TBME.2021.3058424.
8. Liu N, Yuan Z, Tang Q. Improving Alzheimer's Disease Detection for Speech Based on Feature Purification Network. Frontiers in Public Health. 2022 2022-March-03;Volume 9 - 2021. doi: 10.3389/fpubh.2021.835960.
9. Zhang Z, Wang T, Hu Z, Yang LZ, Li H. DEMENTIA: A Hybrid Attention-Based Multimodal and Multi-Task Learning Framework With Expert Knowledge for Alzheimer's Disease Assessment From Speech. IEEE Journal of Biomedical and Health Informatics. 2025;29(4):2957-68. doi: 10.1109/JBHI.2024.3509620.
10. Jeong S-M, Kim S, Lee EC, Kim HJ. Exploring Spectrogram-Based Audio Classification for Parkinson’s Disease: A Study on Speech Classification and Qualitative Reliability Verification. Sensors [Internet]. 2024; 24(14).
11. Lahoti A, Gurugubelli K, Arroyave JRO, Vuppala AK. Shifted Delta Cepstral Coefficients with RNN to Improve the Detection of Parkinson’s Disease from the Speech. Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing; Noida, India: Association for Computing Machinery; 2022. p. 284–8.
12. Laguarta J, Subirana B. Longitudinal Speech Biomarkers for Automated Alzheimer's Detection. Frontiers in Computer Science. 2021 2021-April-08;Volume 3 - 2021. doi: 10.3389/fcomp.2021.624694.
13. Vásquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Nöth E. Multimodal Assessment of Parkinson's Disease: A Deep Learning Approach. IEEE Journal of Biomedical and Health Informatics. 2019;23(4):1618-30. doi: 10.1109/JBHI.2018.2866873.
14. Huang YJ, Lin YT, Liu CC, Lee LE, Hung SH, Lo JK, et al. Assessing Schizophrenia Patients Through Linguistic and Acoustic Features Using Deep Learning Techniques. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2022;30:947-56. doi: 10.1109/TNSRE.2022.3163777.
15. Zhang X, Zhang X, Chen W, Li C, Yu C. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Scientific Reports. 2024 2024/04/25;14(1):9543. doi: 10.1038/s41598-024-60278-1.
16. Chen Z, Liang N, Li H, Zhang H, Li H, Yan L, et al. Exploring explainable AI features in the vocal biomarkers of lung disease. Computers in Biology and Medicine. 2024 2024/09/01/;179:108844. doi: https://doi.org/10.1016/j.compbiomed.2024.108844.
17. Bauser M, Kraus F, Koehler F, Rak K, Pryss R, Weiß C, et al. Voice Assessment and Vocal Biomarkers in Heart Failure: A Systematic Review. Circulation: Heart Failure. 2025;18(8):e012303. doi: doi:10.1161/CIRCHEARTFAILURE.124.012303.
18. Liu S, Geng M, Hu S, Xie X, Cui M, Yu J, et al. Recent Progress in the CUHK Dysarthric Speech Recognition System. IEEE/ACM Trans Audio, Speech and Lang Proc. 2021;29:2267–81. doi:

10.1109/taslp.2021.3091805.
19. Geng M, Xie X, Ye Z, Wang T, Li G, Hu S, et al. Speaker Adaptation Using Spectro-Temporal Deep Features for Dysarthric and Elderly Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022;30:2597-611. doi: 10.1109/TASLP.2022.3195113.
20. Yue Z, Loweimi E, Christensen H, Barker J, Cvetkovic Z. Acoustic Modelling From Raw Source and Filter Components for Dysarthric Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022;30:2968-80. doi: 10.1109/TASLP.2022.3205766.
21. Di Cesare MG, Perpetuini D, Cardone D, Merla A. Assessment of Voice Disorders Using Machine Learning and Vocal Analysis of Voice Samples Recorded through Smartphones. BioMedInformatics. 2024;4(1):549-65. PMID: doi:10.3390/biomedinformatics4010031.
22. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature Medicine. 2022 2022/01/01;28(1):31-8. doi: 10.1038/s41591-021-01614-0.
23. Bensoussan Y SA, Rameau A, Elemento O, Powell M, Dorr D, Payne P, Ravitsky V, Bélisle-Pipon J, Johnson A, Bahr R, Watts S, Bolser D, Siu J, Lerner-Ellis J, Rudzicz F, Boyer M, Salvi Cruz S, Abdel-Aty Y, Ahmed Syed T, Anibal J, Aradi S, Martinez A S, Awan S, Bedrick S, Bernier A, Bevers I, Brito R, Casalino S, Costello J, De Santiago I, Diaz-Ocampo E, Ebraheem M, Eiseman E, Elmahdy M, Evangelista E, Fletcher K, Gallois H, Gelbard A, Goldenberg A, Hanna K, Hersh W, Jayachandran L, Jenney K, Jenkins K, Jo S, Kalia A, Krussel A, Lapadula E, Loewith C, Mahajan R, Maharaj V, Miao S, Mifsud M, Mikhael M, Moothedan E, Nafii Y, Neal T, Newberry K, Ng E, Nickel C, Urbano M, Pharr T, Pontell M, Premi-Bortolotto C, Rahman J, Rohde S, Russell L, Shah S, Shawkat A, Silberholz E, Sutherland D, Swarna Mukhi V, Tang J, Toghranegar J, Vinson K, Wilson C, Zanin M, Zeng X, Zesiewicz T, Zhao R, Zisimopoulos P, Ghosh S. Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information. 2.0.1 ed: PhysioNet; 2025.
24. MacWhinney B, Fromm D, Forbes M, Holland A. AphasiaBank: Methods for studying discourse. Aphasiology. 2011 2011/11/01;25(11):1286-307. doi: 10.1080/02687038.2011.589893.
25. (DxA) AsDDFDA. SpeechDx: A Longitudinal Harmonized Speech Dataset for Alzheimer's Disease Biomarker Development. 2023 [cited 2025 August, 15]; Available from: https://www.alzdiscovery.org/research-and-grants/speechdx.
26. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of Medical Ethics. 2021;47(5):329-35. doi: 10.1136/medethics-2020-106820.
27. Poon AIF, Sung JJY. Opening the black box of AI-Medicine. Journal of Gastroenterology and Hepatology. 2021;36(3):581-4. doi: https://doi.org/10.1111/jgh.15384.
28. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report. 2019;49(1):15-21. doi: https://doi.org/10.1002/hast.973.
29. Raposo VL. The fifty shades of black: about black box AI and explainability in healthcare. Medical Law Review. 2025;33(1). doi: 10.1093/medlaw/fwaf005.
30. Aravazhi PS, Gunasekaran P, Benjamin NZY, Thai A, Chandrasekar KK, Kolanu ND, et al. The integration of artificial intelligence into clinical medicine: Trends, challenges, and future directions. Disease-a-Month. 2025 2025/06/01/;71(6):101882. doi: https://doi.org/10.1016/j.disamonth.2025.101882.
31. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature Communications. 2020 2020/07/31;11(1):3852. doi: 10.1038/s41467-020-17431-x.
32. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019 2019/05/01;1(5):206-15. doi: 10.1038/s42256-019-0048-x.
33. Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems. 2022 2022/12/01;28(6):2335-55. doi: 10.1007/s00530-022-00960-4.
34. Marey A, Arjmand P, Alerab ADS, Eslami MJ, Saad AM, Sanchez N, et al. Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. Egyptian Journal of Radiology and Nuclear Medicine. 2024 2024/09/13;55(1):183. doi:

10.1186/s43055-024-01356-2.
35. Wysocki O, Davies JK, Vigo M, Armstrong AC, Landers D, Lee R, et al. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. Artificial Intelligence. 2023 2023/03/01/;316:103839. doi: https://doi.org/10.1016/j.artint.2022.103839.
36. Bienefeld N, Boss JM, Lüthy R, Brodbeck D, Azzati J, Blaser M, et al. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. npj Digital Medicine. 2023 2023/05/22;6(1):94. doi: 10.1038/s41746-023-00837-4.
37. Kim M, Kim S, Kim J, Song T-J, Kim Y. Do stakeholder needs differ? - Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. International Journal of Human-Computer Studies. 2024 2024/01/01/;181:103160. doi: https://doi.org/10.1016/j.ijhcs.2023.103160.
38. Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. Journal of Imaging. 2020;6(6):52. PMID: doi:10.3390/jimaging6060052.
39. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2021;23(1):18. PMID: doi:10.3390/e23010018.
40. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L, editors. Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); 2018 1-3 Oct. 2018.
41. Das A, Rad P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey2020 June 01, 2020:[arXiv:2006.11371 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2020arXiv200611371D.
42. Phillips PJ, Phillips PJ, Hahn CA, Fontana PC, Yates AN, Greene K, et al. Four principles of explainable artificial intelligence. National Institute of Standards and Technology, Commerce USDo; 2021 2021. Report No.: NISTIR 8312.
43. Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, et al. Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Barcelona, Spain: Association for Computing Machinery; 2020. p. 648–57.
44. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon. 2023;9(5). doi: 10.1016/j.heliyon.2023.e16110.
45. Nannini L, Balayn A, Smith AL. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; Chicago, IL, USA: Association for Computing Machinery; 2023. p. 1198–212.
46. UK GDPR guidance and resources. Information Commissioner's Office; [updated 2025; cited 2025 September, 6]; Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/.
47. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems. 2021;32(11):4793-813. doi: 10.1109/TNNLS.2020.3027314.
48. Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. Sensors. 2023;23(2):634. PMID: doi:10.3390/s23020634.
49. Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, et al. A review of Explainable Artificial Intelligence in healthcare. Computers and Electrical Engineering. 2024 2024/08/01/;118:109370. doi: https://doi.org/10.1016/j.compeleceng.2024.109370.
50. Mienye ID, Obaido G, Jere N, Mienye E, Aruleba K, Emmanuel ID, et al. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. Informatics in Medicine Unlocked. 2024 2024/01/01/;51:101587. doi: https://doi.org/10.1016/j.imu.2024.101587.
51. Sheu R-K, Pardeshi MS. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. Sensors. 2022;22(20):8068. PMID: doi:10.3390/s22208068.
52. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence

(XAI) in deep learning-based medical image analysis. Medical Image Analysis. 2022 2022/07/01/;79:102470. doi: https://doi.org/10.1016/j.media.2022.102470.
53. Muhammad D, Bendechache M. Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. Computational and Structural Biotechnology Journal. 2024 2024/12/01/;24:542-60. doi: https://doi.org/10.1016/j.csbj.2024.08.005.
54. Caterson J, Lewin A, Williamson E. The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review. DIGITAL HEALTH. 2024;10:20552076241272657. doi: 10.1177/20552076241272657.
55. Salih AM, Galazzo IB, Gkontra P, Rauseo E, Lee AM, Lekadir K, et al. A review of evaluation approaches for explainable AI with applications in cardiology. Artificial Intelligence Review. 2024 2024/08/09;57(9):240. doi: 10.1007/s10462-024-10852-w.
56. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. npj Digital Medicine. 2023 2023/01/18;6(1):6. doi: 10.1038/s41746-023-00751-9.
57. Zhang W, Lim BY. Towards Relatable Explainable AI with the Perceptual Process. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; New Orleans, LA, USA: Association for Computing Machinery; 2022. p. Article 181.
58. Petti U, Nyrup R, Skopek JM, Korhonen A. Ethical considerations in the early detection of Alzheimer's disease using speech and AI. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; Chicago, IL, USA: Association for Computing Machinery; 2023. p. 1062–75.
59. Rojas F, Madanian S, Templeton JM, Poellabauer C, Schneider SL, editors. Exploring Deep Learning and Grad-CAM for Speech-Based Detection of Mild Traumatic Brain Injury. 2024 IEEE International Conference on Big Data (BigData); 2024 15-18 Dec. 2024.
60. Li P, Li L, Hamdulla A, Wang D. Reliable Visualization for Deep Speaker Recognition2022 April 01, 2022:[arXiv:2204.03852 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2022arXiv220403852L.
61. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews. 2016 2016/12/05;5(1):210. doi: 10.1186/s13643-016-0384-4.
62. Gupta S, Patil AT, Purohit M, Parmar M, Patel M, Patil HA, et al. Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments. Neural Networks. 2021 2021/07/01/;139:105-17. doi: https://doi.org/10.1016/j.neunet.2021.02.008.
63. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net2014 December 01, 2014:[arXiv:1412.6806 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6806S.
64. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps2013 December 01, 2013:[arXiv:1312.6034 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2013arXiv1312.6034S.
65. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2020 2020/02/01;128(2):336-59. doi: 10.1007/s11263-019-01228-7.
66. Fu J, Yang S, He F, He L, Li Y, Zhang J, et al. Sch-net: a deep learning architecture for automatic detection of schizophrenia. BioMedical Engineering OnLine. 2021 2021/08/03;20(1):75. doi: 10.1186/s12938-021-00915-2.
67. Shen J, Zhang X. Individual-independent and cross-language detection of speech disfluencies in stuttering based on multi-adversarial tasks and self-training. Biomedical Signal Processing and Control. 2025 2025/02/01/;100:107051. doi: https://doi.org/10.1016/j.bspc.2024.107051.
68. Muhammad MB, Yeasin M, editors. Eigen-CAM: Class Activation Map using Principal Components. 2020 International Joint Conference on Neural Networks (IJCNN); 2020 19-24 July 2020.
69. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA:

Curran Associates Inc.; 2017. p. 4768–77.
70. Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. Psychological Medicine. 2022;52(5):957-67. doi: 10.1017/S0033291720002718.
71. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: A unified and generic model interpretability library for PyTorch2020 September 01, 2020:[arXiv:2009.07896 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2020arXiv200907896K.
72. Ditthapron A, Lammert AC, Agu EO. Continuous TBI Monitoring From Spontaneous Speech Using Parametrized Sinc Filters and a Cascading GRU. IEEE Journal of Biomedical and Health Informatics. 2022;26(7):3517-28. doi: 10.1109/JBHI.2022.3158840.
73. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 1135–44.
74. Wullenweber A, Akman A, Schuller BW, editors. CoughLIME: Sonified Explanations for the Predictions of COVID-19 Cough Classifiers. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2022 11-15 July 2022.
75. Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller B. End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study. BMJ Innovations. 2021;7(2):356-62. doi: 10.1136/bmjinnov-2021-000668.
76. Gutiérrez-Serafín B, Andreu-Perez J, Pérez-Espinosa H, Paulmann S, Ding W. Toward assessment of human voice biomarkers of brain lesions through explainable deep learning. Biomedical Signal Processing and Control. 2024 2024/01/01/;87:105457. doi: https://doi.org/10.1016/j.bspc.2023.105457.
77. Joshy AA, Rajan R. Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2022;30:1147-57. doi: 10.1109/TNSRE.2022.3169814.
78. Herath HM, Weraniyagoda WA, Rajapaksha RT, Wijesekara PA, Sudheera KL, Chong PH. Automatic Assessment of Aphasic Speech Sensed by Audio Sensors for Classification into Aphasia Severity Levels to Recommend Speech Therapies. Sensors [Internet]. 2022; 22(18).
79. He L, Fu J, Li Y, Xiong X, Zhang J. WNSA-Net: An Axial-Attention-Based Network for Schizophrenia Detection Using Wideband and Narrowband Spectrograms. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023;31:721-33. doi: 10.1109/TASLP.2022.3209941.
80. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 6000–10.
81. Wang W, Xu W, Chander A, Nepal S, Buck B, Pakhomov S, et al. The Power of Speech in the Wild: Discriminative Power of Daily Voice Diaries in Understanding Auditory Verbal Hallucinations Using Deep Learning. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2023;7(3):Article 133. doi: 10.1145/3610890.
82. Abnar S, Zuidema W, editors. Quantifying Attention Flow in Transformers. 2020 July; Online: Association for Computational Linguistics.
83. Lau HS, Huntly M, Morgan N, Iyenoma A, Zeng B, Bashford T, editors. Interpreting Pretrained Speech Models for Automatic Speech Assessment of Voice Disorders. Artificial Intelligence in Healthcare; 2024 2024//; Cham: Springer Nature Switzerland.
84. Abderrazek S, Fredouille C, Ghio A, Lalain M, Meunier C, Woisard V. Interpreting Deep Representations of Phonetic Features via Neuro-Based Concept Detector: Application to Speech Disorders Due to Head and Neck Cancer. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023;31:200-14. doi: 10.1109/TASLP.2022.3221039.
85. Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9(86):2579-605.

86. Lee JH, Lee GW, Bong G, Yoo HJ, Kim HK. Deep-Learning-Based Detection of Infants with Autism Spectrum Disorder Using Auto-Encoder Feature Representation. Sensors. 2020;20(23):6762. PMID: doi:10.3390/s20236762.
87. Kim H-B, Song J, Park S, Lee YO. Classification of laryngeal diseases including laryngeal cancer, benign mucosal disease, and vocal cord paralysis by artificial intelligence using voice analysis. Scientific Reports. 2024 2024/04/23;14(1):9297. doi: 10.1038/s41598-024-58817-x.
88. Jain S, Wallace BC. Attention is not Explanation2019 February 01, 2019:[arXiv:1902.10186 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190210186J.
89. Liu Y, Li H, Guo Y, Kong C, Li J, Wang S. Rethinking Attention-Model Explainability through Faithfulness Violation Test. In: Kamalika C, Stefanie J, Le S, Csaba S, Gang N, Sivan S, editors. Proceedings of the 39th International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR; 2022. p. 13807--24.
90. Lopardo G, Precioso F, Garreau D. Attention Meets Post-hoc Interpretability: A Mathematical Perspective2024 February 01, 2024:[arXiv:2402.03485 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv240203485L.
91. Bibal A, Cardon R, Alfter D, Wilkens R, Wang X, François T, et al., editors. Is Attention Explanation? An Introduction to the Debate. 2022 May; Dublin, Ireland: Association for Computational Linguistics.
92. Schlegel U, Keim DA. A Deep Dive into Perturbations as Evaluation Technique for Time Series XAI2023 July 01, 2023:[arXiv:2307.05104 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2023arXiv230705104S.
93. Coroama L, Groza A, editors. Evaluation Metrics in Explainable Artificial Intelligence (XAI). 2022; Cham: Springer Nature Switzerland.
94. Tracey B, Volfson D, Glass J, Haulcy Rm, Kostrzebski M, Adams J, et al. Towards interpretable speech biomarkers: exploring MFCCs. Scientific Reports. 2023 2023/12/21;13(1):22787. doi: 10.1038/s41598-023-49352-2.
95. Aranovich TdC, Matulionyte R. Ensuring AI explainability in healthcare: problems and possible policy solutions. Information & Communications Technology Law. 2023 2023/05/04;32(2):259-75. doi: 10.1080/13600834.2022.2146395.
96. Moorthy UMK, Muthukumaran AMJ, Kaliyaperumal V, Jayakumar S, Vijayaraghavan KA. Explainability and Regulatory Compliance in Healthcare. Explainable Artificial Intelligence in the Healthcare Industry2025. p. 521-61.
97. Chauhan R. A Review on Explainable Artificial Intelligence for Healthcare. Explainable Artificial Intelligence in the Healthcare Industry2025. p. 1-15.
98. Manek AS, Vashist S, Tripathi G, Sindhu S. Explainable Artificial Intelligence (XAI) in Healthcare. Explainable Artificial Intelligence in the Healthcare Industry2025. p. 17-37.
99. Ahamed SM, Jabez J. Bridging the Gap. Explainable Artificial Intelligence in the Healthcare Industry2025. p. 349-74.
100. Clark JN, Wragg M, Nielsen E, Perello-Nieto M, Keshtmand N, Ambler M, et al. Exploring the Requirements of Clinicians for Explainable AI Decision Support Systems in Intensive Care2024 November 01, 2024:[arXiv:2411.11774 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv241111774C.
101. Alam MN, Kaur M, Kabir MS. Explainable AI in healthcare: enhancing transparency and trust upon legal and ethical consideration. Int Res J Eng Technol. 2023;10(6):1-9.
102. Hou J, Wang LL. Explainable AI for Clinical Outcome Prediction: A Survey of Clinician Perceptions and Preferences. AMIA Jt Summits Transl Sci Proc. 2025;2025:215-24. PMID: 40502256.
103. Rosenbacke R, Melhus Å, McKee M, Stuckler D. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. JMIR AI. 2024;3:e53207. doi: 10.2196/53207.
104. He X, Hong Y, Zheng X, Zhang Y. What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System. International Journal of Human–Computer

Interaction. 2023 2023/04/21;39(7):1519-42. doi: 10.1080/10447318.2022.2095093.
105. Hoffman RR, Mueller ST, Klein G, Jalaeian M, Tate C. Explainable AI: roles and stakeholders, desirements and challenges. Frontiers in Computer Science. 2023 2023-August-17;Volume 5 - 2023. doi: 10.3389/fcomp.2023.1117848.
106. Farah L, Murris JM, Borget I, Guilloux A, Martelli NM, Katsahian SIM. Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence–Based Health Technologies: What Healthcare Stakeholders Need to Know. Mayo Clinic Proceedings: Digital Health. 2023 2023/06/01/;1(2):120-38. doi: https://doi.org/10.1016/j.mcpdig.2023.02.004.
107. Veeramani M, Karthick P, Venkateswaran S, Sriman B, Bhanu ST, Devi VS. Transparency in Text. Explainable Artificial Intelligence in the Healthcare Industry2025. p. 131-60.
108. Dhanorkar S, Wolf CT, Qian K, Xu A, Popa L, Li Y. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. Proceedings of the 2021 ACM Designing Interactive Systems Conference; Virtual Event, USA: Association for Computing Machinery; 2021. p. 1591–602.
109. Srinivasan K, Ramamurthy CKV, Matheswaran S, Shamsudheen S. Introduction to Explainable AI in Healthcare. Explainable Artificial Intelligence in the Healthcare Industry2025. p. 161-83.
110. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. Frontiers in Artificial Intelligence. 2022 2022-May-30;Volume 5 - 2022. doi: 10.3389/frai.2022.879603.
111. Chakrapani K, Safa MI, Moorthy SG, Kumaraswamy M, Parimala G. Envisioning Explainable AI. Explainable Artificial Intelligence in the Healthcare Industry2025. p. 563-91.
112. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?2017 December 01, 2017:[arXiv:1712.09923 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2017arXiv171209923H.
113. Sporsem T, Rasdal Finserås S, Strümke I. Clinicians don't know what explanations they need: A case study on eliciting AI software explainability requirements2025 January 01, 2025:[arXiv:2501.09592 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2025arXiv250109592S.
114. Wang P, Lu W, Lu C, Zhou R, Li M, Qin L. Large Language Model for Medical Images: A Survey of Taxonomy, Systematic Review, and Future Trends. Big Data Mining and Analytics. 2025;8(2):496-517. doi: 10.26599/BDMA.2024.9020090.

Abbreviations

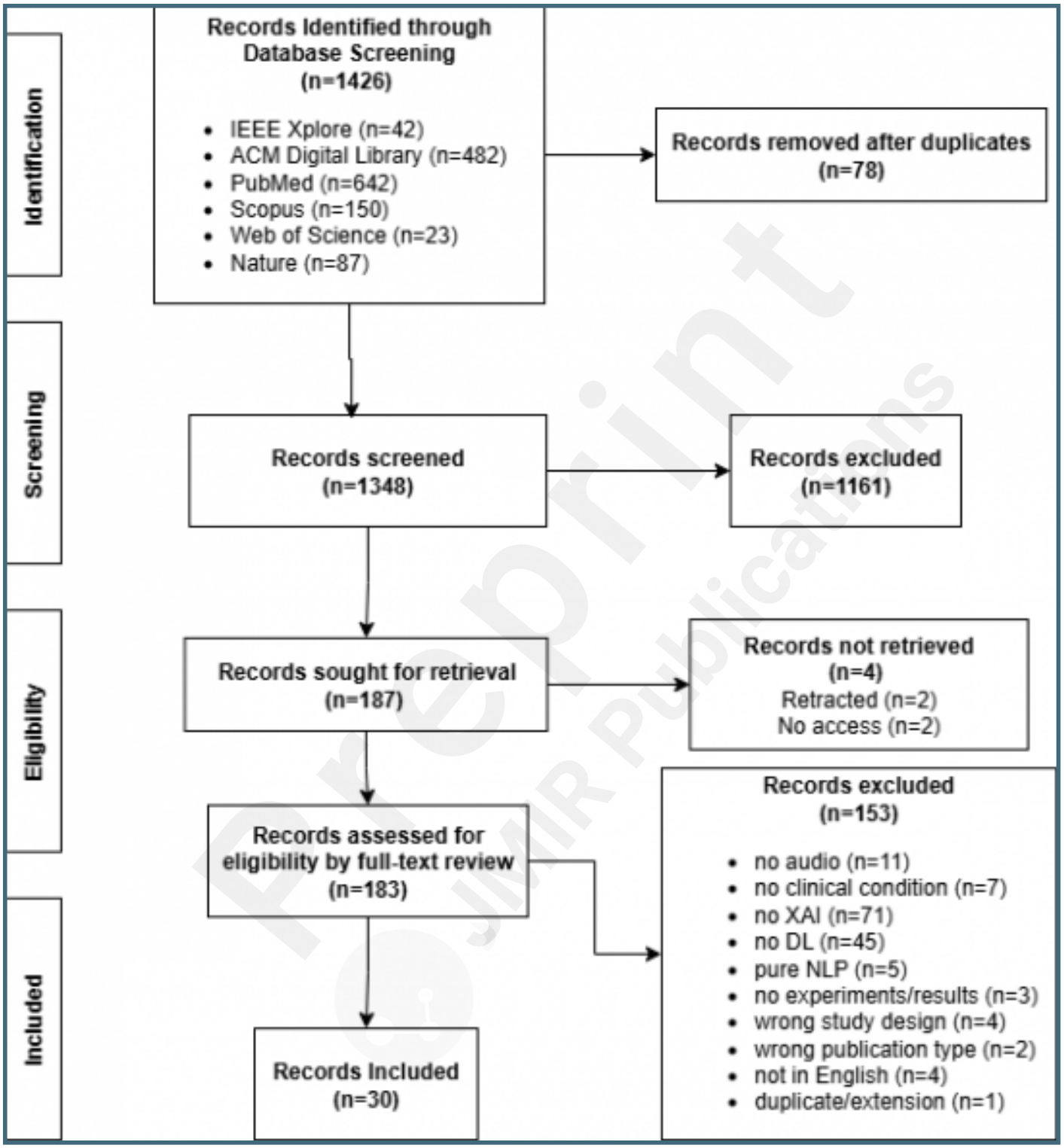
AD: Alzheimer’s disease
AI: Artificial Intelligence
ASD: Autism Spectrum Disorder
AST: Audio Spectrogram Transformer
AVH: Auditory Verbal Hallucination
BERT: Bidirectional Encoder Representations from Transformers
BLSTM: Bidirectional Long Short-Term Memory
CBAM: Convolutional Block Attention Module
CNN: Convolutional Neural Network
CQCC: Constant Q Cepstral Coefficients
cGRU: Cascading Gated Recurrent Unit
CTC: Connectionist Temporal Classification
DDK: Diadochokinesis
DNN: Deep Neural Network
EHR: Electronic Health Record
ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately
FC: Fully Connected (layer)
F0: Fundamental Frequency
GRBAS: Grade, Roughness, Breathiness, Asthenia, Strain (voice scale)

GRU: Gated Recurrent Unit
HuBERT: Hidden-Unit BERT
LiGRU: Light Gated Recurrent Unit
LIME: Local Interpretable Model-Agnostic Explanations
LLD: Low-Level Descriptor
LSTM: Long Short-Term Memory
MFCC: Mel-Frequency Cepstral Coefficients
MDD: Major Depressive Disorder
ML: Machine Learning
MLP: Multi-Layer Perceptron
mTBI: Mild Traumatic Brain Injury
NC: Nasal Consonant
NV: Nasalized Vowel
OHM: Objective Hypernasality Measure
OC: Oral Consonant
OV: Oral Vowel
PANSS: Positive and Negative Syndrome Scale
PD: Parkinson's Disease
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PTSD: Post-Traumatic Stress Disorder
ReLU: Rectified Linear Unit
ResNet: Residual Network
RPM: Regulatory and Policy Maker
SFFCC: Single Frequency Filtering Cepstral Coefficients
SHAP: SHapley Additive exPlanations
SincNet: Sinc Convolutional Neural Network
SVD: Singular Value Decomposition
TDNN: Time Delay Neural Network
TLC: Thought, Language, and Communication scale
t-SNE: t-Distributed Stochastic Neighbor Embedding
Wav2Vec: (Self-supervised representation learning model)
WER: Word Error Rate
WNSA-Net: Wideband and Narrowband Spectrogram Attention Network
XAI: Explainable Artificial Intelligence
xDMFCC: eXplainable Deep Learning MFCCs
ZCR: Zero Crossing Rate

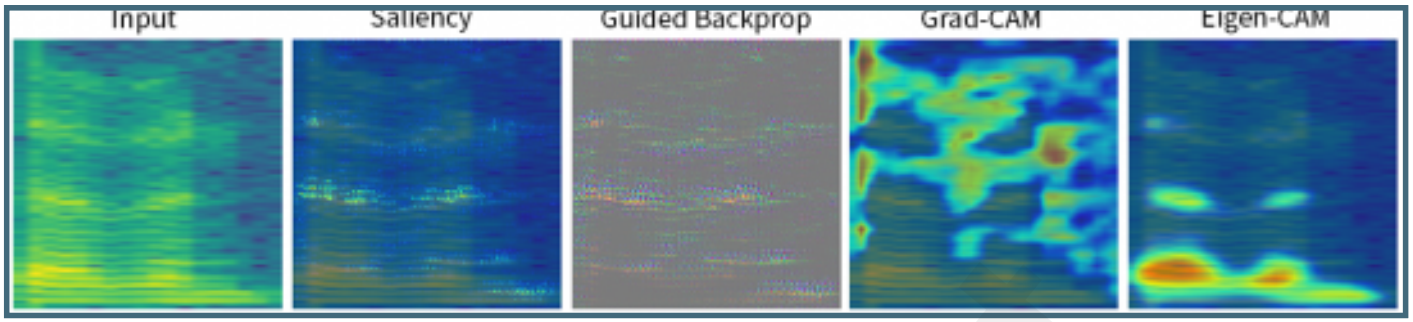
Supplementary Files

Figures

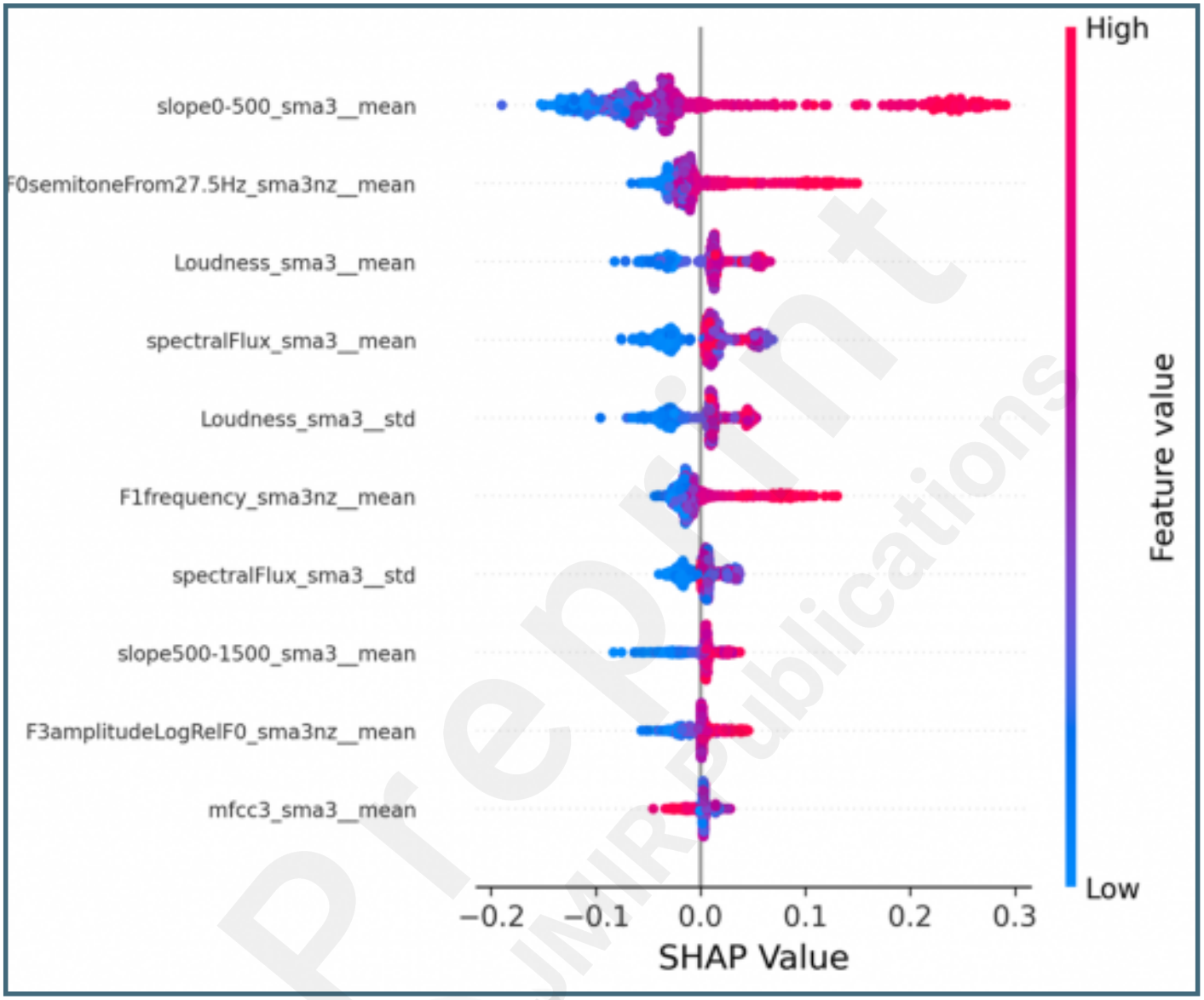
Prisma flowchart.



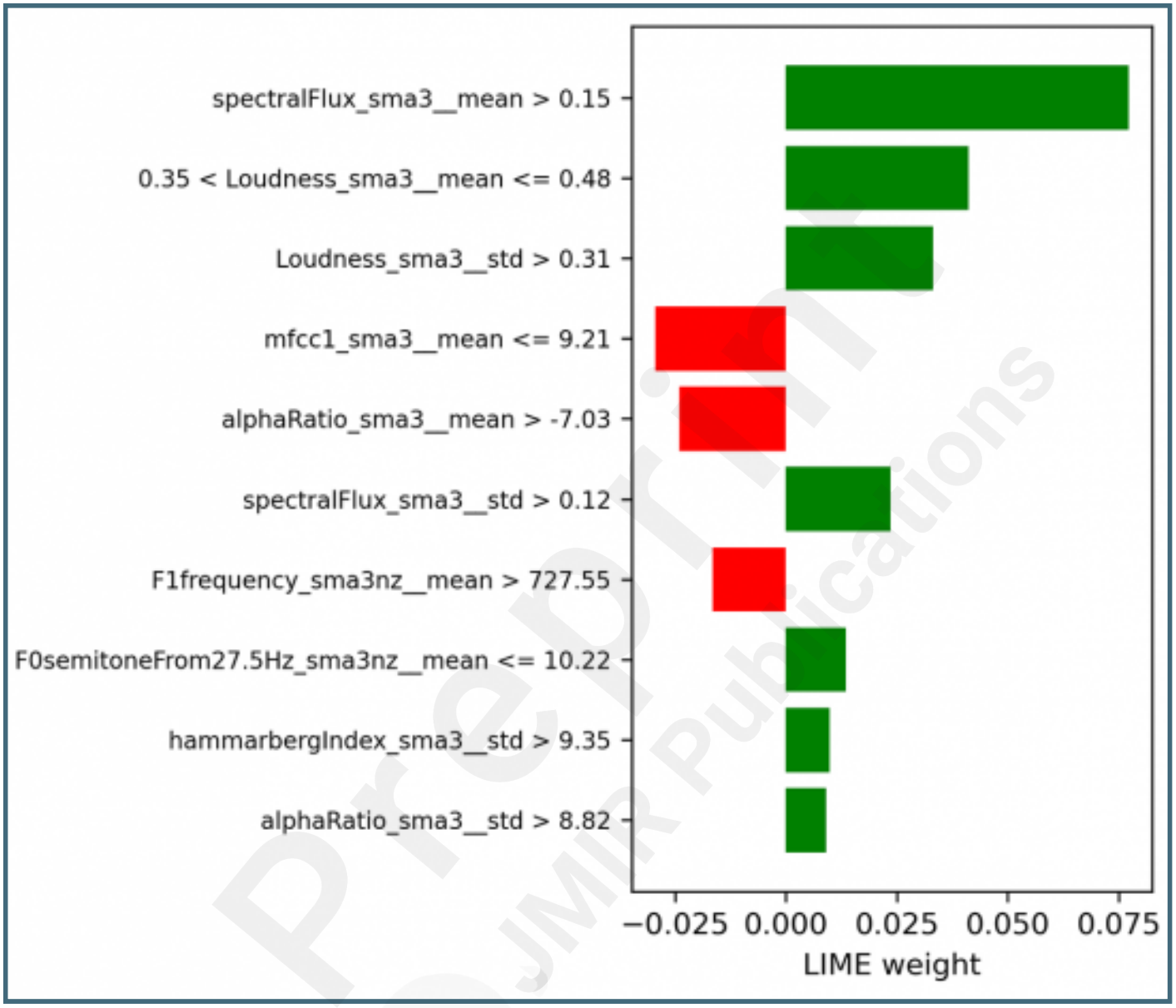
Activation maps show varying patterns for different vision-based methods. Red corresponds to high activation and blue signifies low neural activation or low region influence.



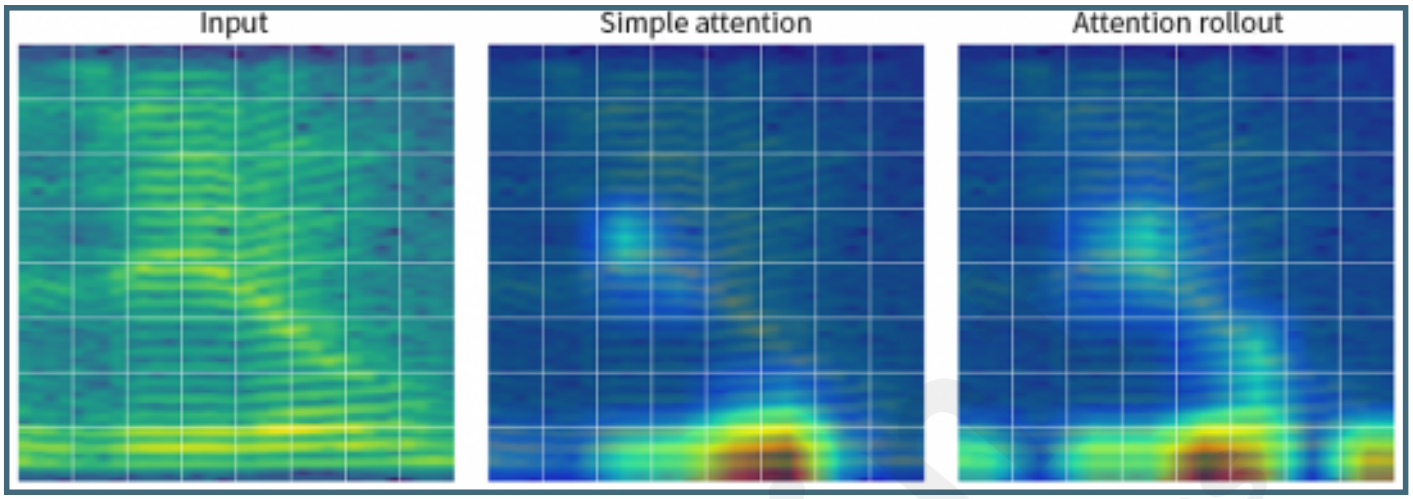
Example of SHAP explanation beeswarm plot. Every data sample is represented by a point, where color represents the feature value. Positive SHAP value indicates that the feature increases the likelihood of target class, while negative SHAP value signifies a decrease in likelihood of the target class.



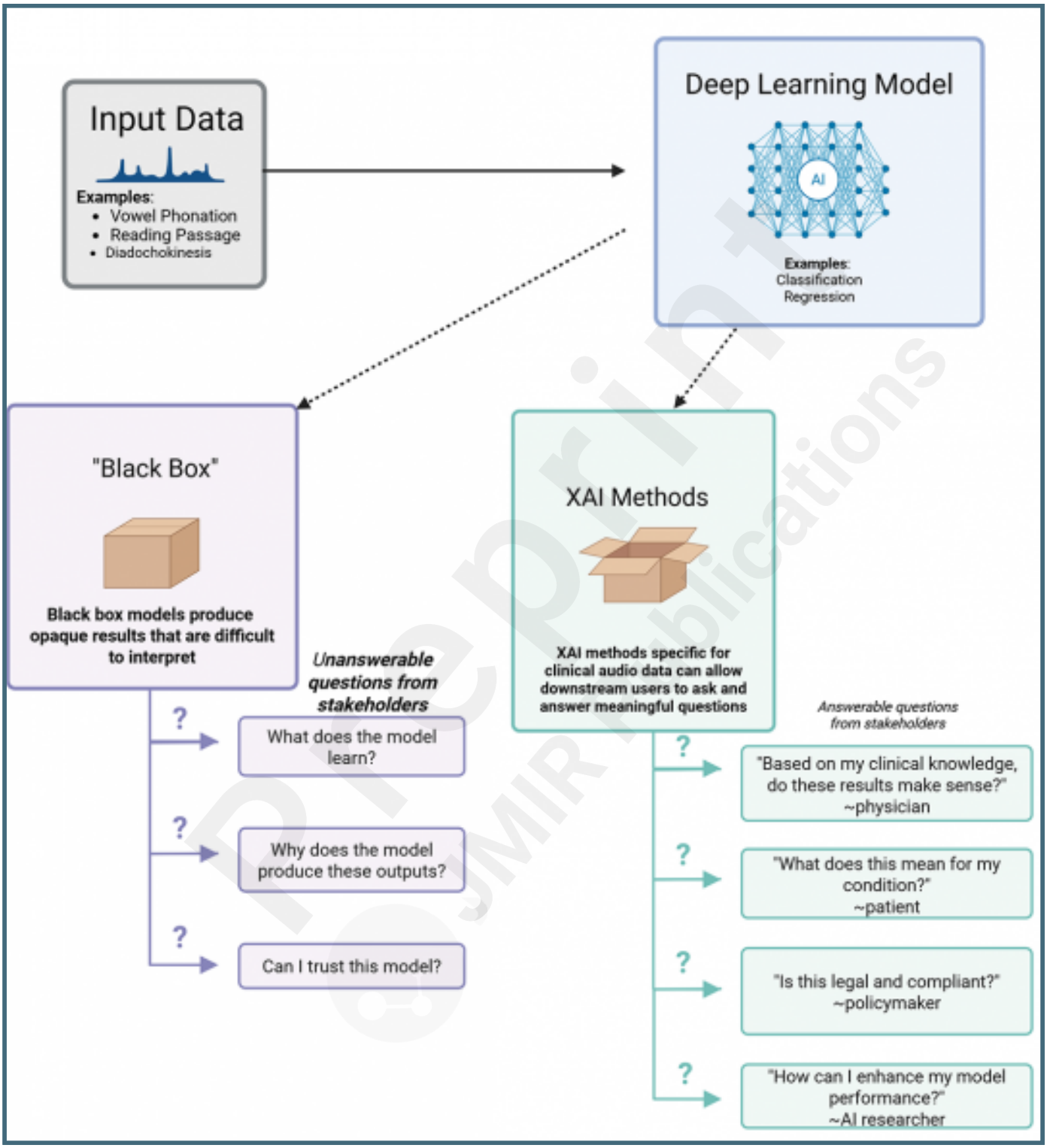
Example of LIME explanation. Positive LIME weight (green) indicates that the feature pushes the model toward predicting the target class, while a negative weight pushes the model away from the target.



Attention rollout accumulation of attention weights across layers results in smoother activation for formants compared to simple attention visualization.



Although XAI aims to address the black-box issue of deep learning models, current XAI methods do not cater to the diverse expectations and needs for the different stakeholders.



Key future directions and recommendations for clinical audio XAI.