Can you hear me now? Clinical applications of audio recordings

Anish Kumar[1$], Theo Jaquenoud[2$], Jacqueline Helcer Becker[3], Dayeon Cho[4], Monica Rivera Mindt[5,6], Alex Federman[7], Gaurav Pandey[8]*

[1] MD Program, Icahn School of Medicine at Mount Sinai, New York, NY 10029, anish.kumar@icahn.mssm.edu
[2] ME Program, The Cooper Union, New York, NY, 10003, jaquenou@cooper.edu
[3] Department of Medicine, Division of General Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, jacqueline.becker@mountsinai.org
[4] Department of Medicine, Division of General Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, dayeon.cho@mountsinai.org
[5] Department of Psychology, Latin American and Latino Studies Institute, and African and African American Studies, Fordham University, New York, NY 10023, riveramindt@fordham.edu
[6] Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY 10029
[7] Department of Medicine, Division of General Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, alex.federman@mountsinai.org
[8] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, gaurav.pandey@mssm.edu

[$]These authors contributed equally.

*Corresponding author.

Abstract

Audio and speech have several implicit characteristics that have the potential for the identification and quantification of clinical disorders. This PRISMA-guided review is designed to provide an overview of the landscape of automated clinical audio processing to build data-driven predictive models and infer phenotypes of a variety of neuropsychiatric, cardiac, respiratory and other disorders. We detail the important components of this processing workflow, specifically data acquisition and processing, algorithms used and their customization for clinical applications, commonly used tools and software, and benchmarking and evaluation methodologies. Finally, we discuss important open challenges for the field, and potential strategies for addressing them.

Introduction

Audio has long served as a rich source of information for clinicians to evaluate the health of their patients.[1,2] Routine physical exams involve auscultation or listening to various sounds (respiratory, cardiac, gastrointestinal, etc.), which indicate the health of various organs and physiologic systems.[2] Many clinicians also consider patients' speech patterns to assess mood, cognition, and other neurological functions.[1,3–9]

Audio and speech also have implicit characteristics that human experts are unable to hear and quantify.[10] However, with improvements in computational power and machine learning algorithms,[11,12] data scientists have recently been able to unlock novel dimensions of audio and provide clinicians with more information to support decision-making.[10] We present an overview of this work on developing novel biomarkers, predictive models and other data-driven inferences from clinical audio for a variety of disorders. We detail various components of the audio processing workflow (**Figure 1**), beginning with a description of the considerations and tools for collecting and processing clinical audio data. We then present an overview of statistical, traditional machine learning,[11] and deep learning[13] methods that have been leveraged for the analysis of clinical audio for diagnostic and prognostic purposes. We end with a discussion of the challenges and opportunities for automated analyses of clinical audio data.

Methods

Our literature search was conducted using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology[14] on the PubMed, Web of Science, and Google Scholar databases using the following keywords: "clinical", "speech", "automatic analysis", "health", "computational", "machine learning", "deep learning" and "audio". Given the rapid developments in clinical audio processing, we focused on research published after 2010 to emphasize the most recent trends in this field. Furthermore, studies that involved analyses of text generated from speech were included only if they were done in conjunction with an analysis of the native audio. For details of studies from before 2010 and/or focusing on clinical speech, we recommend other excellent reviews. [1,3–5,9]

The PRISMA flow chart in **Figure 2** details the search results and inclusion/exclusion selection criteria. From the initial 1812 records retrieved from the databases, 561 duplicates were automatically removed, and a further 880 were removed after manual screening of titles and abstracts based on specific inclusion/exclusion criteria, leaving 371 in-scope articles. Studies were included if they explicitly used audio or audio-derived features as clinical biomarkers or were reviews of relevant clinical uses of audio signals; studies that did not use audio signals or used them only for obtaining textual transcription, as well as those that considered ultrasound or echocardiogram only for imaging, were excluded. Among these in-scope articles, this review discusses the 69 most recent and least redundant representative papers on clinical audio processing. These studies were considered representative if they proposed a novel solution or a

solution to a new problem, or if they published a new clinical audio dataset. Studies that offered redundant methodologies were excluded in favor of newer or more frequently cited papers. The screening and reviewing of the articles was principally conducted by the first authors.

Preliminaries of audio data

Effective data collection and pre-processing are crucial steps in the clinical audio processing workflow (**Figure 1**). Below, we describe several aspects of these tasks.

Collection of audio data

A variety of devices were used to record audio in the studies reviewed, ranging from studio-quality microphones to wearable lavaliers, lapel microphones and smartphones. The choice of recording device and its placement relative to the participant may significantly affect data quality.[15] However, few studies have reported on optimal microphone placement configurations. Some studies suggested using wearable microphones, such as lavaliers, since they are less susceptible to changes in proximity, orientation and posture of the person being recorded.[5]

Other technical factors that can also affect the quality of the recording include the range of frequencies that a microphone can capture. For speech, the recommended upper limit of this range is at least 10 kilohertz (kHz).[15,16] The choice of sample rate (number of times a data point is recorded per second) and bit depth (number of binary digits used to represent each point) can also affect audio quality. Modern systems generally record at a sample rate of 44.1 kHz and a bit depth of 16 or 24, although some applications require a lower sample rate to improve computational efficiency.[15]

Pre-processing audio data

Denoising is the most basic pre-processing of audio data.[17] While there are sophisticated solutions to this problem,[18] it is often sufficient to filter out frequencies below 60 Hz, which removes most of the noise induced by common electronic devices.[15]

Normalization is the process of making audio signals from multiple sources, e.g., patients, compatible with each other.[16] A popular normalization method is min-max scaling, which linearly scales all data from a single source to a given range, typically 0 to 1.[19–21] Normalization can be performed on the raw audio signal, or more commonly, on features derived from it (next subsection).

When recording a conversation between multiple individuals, e.g., a patient and their physician, it may be necessary to separate the audio signal of each speaker, a task referred to as diarization.[22] Another useful pre-processing operation is forced alignment, which aligns the audio signals with the transcribed text of the recording.[23]

Table 1 lists several commonly used tools for audio pre-processing.

Extraction of features from audio

In data analysis, a *feature* is defined as a quantitative variable that can be used to describe a data entity.[11] In the case of (pre-processed) audio, a variety of features can be derived, or *engineered*, as described below.

Audio is natively represented as unstructured waveforms that correspond to variations in sound pressure over time. The features derived from this representation can be broadly related to time, pitch, or energy of the signal, although these categories often overlap (**Table 2**). An example of a time-related feature is the pause rate, which measures how often a speaker takes pauses of a prescribed length during a time period.[24] An informative pitch-related feature is the Mel-frequency cepstrum coefficient (MFCC), which rescales frequencies of the audio signal logarithmically to mimic human auditory perception.[25] Finally, an energy-related feature is the loudness of a sound, typically measured in decibels.[25] Several tools are available for feature extraction from audio signals (**Table 1**), of which, the open-source Praat[26] and OpenSMILE[27] the most widely used.

It can often be unfavorable for the computational efficiency and performance of data analysis methods to use too many features due to the *curse of dimensionality* problem.[28] Thus, the Geneva Minimalist Acoustic Parameter Set (GeMAPS) has been proposed as a minimalist set of 62 features that are expected to be effective for audio.[25] GeMAPS is widely used due to its inclusion in openSMILE.[27]

Prediction methods for clinical audio

Automated audio processing methods have been widely used for predicting disease indications, especially in neuropsychiatry. We describe these efforts in the following subsections.

Statistical and traditional machine learning (ML) methods

Statistical methods, such as hypothesis testing, have been used routinely to analyze biomedical data.[29] Traditional ML methods, e.g., Support Vector Machine (SVM), Random Forest (RF) and k-Nearest Neighbor (kNN), have also been widely applied in biomedicine.[11] These methods are designed to sift through large amounts of data, structured or unstructured, without any particular guiding (biomedical) hypothesis, to discover potentially actionable knowledge. A *predictive model* encapsulates a mathematical relationship between the data describing an entity of interest, say a patient, and an outcome or label, say the disease status, of the entity. The purpose of this model is to make predictions of this outcome or label that are not yet known for other entities.

Early applications of these methods aimed to understand which features of clinical audio had explanatory or predictive power.[30–33] A semi-automated approach assessed speech differences between children with cerebral palsy and controls by analyzing data from speech elicitation tasks.[30] Trained listeners transcribed speech recordings, which were used to determine word counts and what proportion of the words uttered matched the target elicitation (intelligibility). The study found that speech rate (words uttered per minute) and intelligibility classified normally developing children and those with cerebral palsy.

A similar study analyzed speech from picture-describing and sentence-repeating tasks to distinguish between patients with Alzheimer's dementia (AD) and those with mild cognitive impairment (MCI).[31] The study found that the duration of speech and the increased likelihood of inserting or deleting to words in prompted sentences differentiated AD and MCI patients. Other studies used statistical tests like Mann-Whitney U to evaluate the association of pitch features (**Table 2**) with neuropsychiatric conditions.[31–33]

Several studies also used audio features and ML methods (**Table 2**) to classify patients into classes corresponding to neuropsychiatric conditions.[31,34–36] Most of these studies use variations of the SVM algorithm, which finds an optimal boundary separating two classes of data points. One study used multiple SVM models to identify a motor speech disorder by determining the severity of unintelligible speech.[35] A sentence-level SVM trained on energy features and a phoneme-level SVM trained on pitch and time features yielded accuracies of 79.8% and 77.3% respectively. This accuracy increased to 84.8% when the SVMs were combined, an approach known as ensemble learning.[11] Other studies have found success in similar tasks using other traditional ML algorithms like RF and kNN.[37,38]

Several approaches also combined audio features with linguistic characteristics derived from textual transcripts of the audio.[32,36,38,39] One study performed 3-way classification of levels of cognitive impairment (control, mild, and early stage Alzhiemer's disease) using an SVM trained on acoustic features derived using Praat[26] and an automatic speech recognition (ASR) system (**Table 1**).[40] This classifier had an accuracy of 60%, which improved to 66.7% after combining the acoustic features with a variety of linguistic ones.

Deep learning

More recently, deep learning (DL) techniques[13] have been used to characterize clinical conditions from patient audio.[10] These techniques typically utilize much larger, but not always explicitly specified, feature sets than statistical and traditional ML techniques. DL techniques generally utilize multi-layer neural networks to build implicit representations of datasets and enable various analysis tasks, including predictive modeling (**Figure 3**).[12] Due to this architecture, DL techniques are capable of building predictive models directly from native audio recordings.

Convolutional neural networks (CNNs) (**Figure 4**) are among the most prevalent DL techniques for audio and other unstructured data, especially due to their ability to represent contextual

information in data.[13] Another established architecture is the recurrent neural network (RNN) (**Figure 5**), where sequentially structured inputs pass through functional transformations at consecutively connected layers of the network.[13]

Most clinical audio processing that utilizes DL inputs pre-computed features into a CNN, which can then predict the presence or degree of a condition.[41,42] One investigation built a CNN model using GeMAPS[25] and other feature sets to classify depression severity.[43] A related approach trained parallel CNN models for the different categories of audio features in **Table 2.**[43] The last dense layers of these CNNs were then concatenated to predict depression severity. Multiple investigations found that an ensemble of individual CNNs built from different data modalities (e.g., audio, text, and video) can predict depression severity even more accurately.[41,42,44]

Other approaches have leveraged the sequential or temporal nature of audio. One AD detection effort employed a Time-Delayed CNN.[45] Instead of the entire recording, this approach applied the convolutional filter to utterances (segments of speech separated by silence) over all preceding time frames (hence the "delay"). This allowed them to extract local features from different temporal segments of the recording.[45] Another study used a Long-Short Term Memory architecture (LSTM, a sophisticated implementation of an RNN)[12] to screen for depression.[46] MFCCs (**Table 2**) were extracted from different temporal segments of the audio and input to the LSTM. The outputs of the recurrent layers were then fed to the fully connected layers of the network to predict depression scores.[46]

DL methods have also been used in situations of insufficient audio data. One study[46] employed transfer learning[12] to classify eight types of emotions from the relatively small RAVDESS dataset[47] (**Table 3**). This approach repurposed an RNN trained on a data-rich task (depression score classification), and fine-tuned it with RAVDESS for a related task (emotion recognition). The approach achieved a validation accuracy of 76.3%, an increase of 8.7% from a baseline RNN trained solely on the RAVDESS.[46] Another useful DL architecture for data augmentation is a Generative Adversarial Network (GAN),[12] which consists of two competing neural networks, a generator and a discriminator, to generate new data samples. In an approach to diagnose childhood autism,[48] the generator of a GAN was trained using GeMAPS[25] and other feature sets extracted from the Child Pathological Speech Database[34] (**Table 3**), and a discriminator was trained to help the model generate more realistic data points. Learned representations of the data were then extracted from the intermediate layers of the discriminator, and used to train an SVM model to classify four levels of pathology related to developmental disorders and autism.

Some approaches use *x-vectors,* which are DL-based representations trained for speaker identification in a conversation.[49,50] These *x-vectors* can then be used with ML or DL methods to classify speakers with and without a pathology.[51] Several studies reported better performance of this approach for Alzheimer's and Parkinson's disease diagnosis compared to feature-based methods .[51–54] Another study utilized an alternative Active Data Representation (ADR).[55] built for speakers from the Pitt Corpus (**Table 3**), and used them to classify AD, performing better than traditional ML methods.

Finally, recent DL methods learn which characteristics of a native audio signal are useful for classification. Zhao and colleagues used a hierarchical attention transfer network that reconstructed segments of patient speech using an autoencoder, and integrated them with LSTM representations from a speech recognition model to screen for depression.[56] Another study compared using the raw audio signal, and its various filtered versions, with a CNN.[57] Filtering the signal boosted the CNN's ability to accurately classify levels of depression from patient speech, illustrating the advantages of effective pre-processing.

Other clinical applications

While most of the research in this area has been on cognitive health, similar work is emerging in other realms, as described below.

Cardiac conditions

High-fidelity recordings of heart sounds can be collected using digital stethoscopes or phonocardiograms (PCGs).[58] After pre-processing to remove ambient noise, the recordings are generally segmented to isolate each beat and its components, traditionally using thresholds of acoustic features defined by clinical experts, or Gaussian probabilistic models.[59] A recent study used a Hidden Markov Model (HMM) to segment PCG recordings into four stages of a heartbeat, namely S1, systole, S2, and diastole.[59]

Several studies have also developed classifiers of cardiovascular disease from segmented PCGs. The 2016 PhysioNet Computing in Cardiology Challenge (**Table 3**) produced several accurate classifiers for these diseases.[11,60] The best-performing method used an ensemble of a feature-based adaptive boosting model and a CNN.[61] The runner-up used an ensemble of 20 neural networks.[62] More recent studies have found similar success using DL techniques.[63,64]

Respiratory diseases

The most common use of audio for respiratory diseases has been the identification and classification of coughs. Several studies classified coughs, most commonly as "wet" or "dry", to help distinguish lower respiratory tract infections like pneumonia, bronchitis, and tuberculosis from other diseases.[65,66] These studies used features extracted from audio (**Table 2**) and classifiers like logistic regression.

Similar traditional ML and some DL methods have been used to directly classify specific respiratory diseases, such as pneumonia, asthma, and whooping cough.[65–70] A representative study compared traditional ML and DL methods for this task using the Respiratory Sound Database (**Table 3**)[71], and found that the best-performing models were CNN, LSTM, and an ensemble of CNN and RNN.

Finally, a number of studies have attempted to diagnose COVID-19 from recordings of patients' coughs.[72] Researchers crowdsourced the COUGHVID dataset (**Table 3**) of cough recordings using a web-app accessible from mobile devices, a subset of which were labeled by pulmonologists in terms of the type of cough and other characteristics.[17] The same effort proposed an algorithm based on XGBoost[73] to filter out recordings containing no coughs. DL methods using extracted audio features with CNN, LSTM, and ResNet architectures have found the most success for classifying coughs and diagnosing COVID-19.[74,75]

Sleep disorders

Studies on sleep disorders attempted to quantify snoring and identify conditions like obstructive sleep apnea and sleep-disordered breathing from data typically recorded using bedside[76] or overhead[76] microphones and smartphones[77] in a controlled environment. Several of these studies employed DL methods with some success, although the quality and quantity of available data remains a problem.[54,78]

Other Applications

Audio processing and ML continue to be applied to a variety of new clinical problems, such as automatically detecting seizures,[79] improving the quality of hearing aids,[80] and identifying out-of-hospital cardiac arrests from emergency dispatchers.[81]

Evaluations methodologies, benchmarking and open issues

Evaluation methodologies

As in general ML, rigorously evaluating audio-based predictive models is as important as building them. Several of the studies reviewed above evaluated these models on a separate test and validation set. If a test set was not available, studies often used bootstrapping or cross-validation,[11] which repeatedly use different randomly chosen partitions of the training set to develop and evaluate the models.

Several evaluation measures have been proposed to quantify the performance of classification-oriented predictive models (**Figure 6**).[82] Accuracy and the area under the Receiver Operating Characteristic (ROC) curve (AUC score) were the most routinely used evaluation measures. However, these measures can be misleading in cases where the classes are imbalanced, which is common in biomedical sciences.[82,83] In these cases, it is recommended to use the class-specific Precision, Recall and F-measure metrics (**Figure 6**) for evaluation.

Benchmarking using crowdsourced challenges and public datasets

It can be difficult to gauge the general effectiveness of a new ML tool for audio analysis unless it is evaluated against an established benchmark dataset.[84] Since it is laborious for individual studies or laboratories to collect large datasets, *crowdsourced challenges* have been organized to address this important need.[85] **Table 3** provides the details of several prominent challenges and publicly available datasets.

Issues facing the field

Although substantial progress has been made in automated analysis of clinical audio, the field still faces several interrelated issues that should be addressed to enable further progress.

Insufficient data

Often, the most significant problem for any data-driven study is insufficient data. In particular, recording clinical audio introduces a number of challenges. These include the cost of a recording device, distraction from providing care, and the need for patient consent. As a result, several studies reviewed above relied on relatively small datasets. This issue is particularly challenging for DL, which typically requires large training datasets.[13] Although benchmark datasets (**Table 3**) and data augmentation[12,20] methods can address this problem to some extent, a general paucity of sufficiently large datasets in multiple clinical areas poses a significant issue for the area.

Incompatibility of data from multiple sources

The problem of insufficient data is partially caused by the incompatibility of audio collected from multiple sources. Even in studies on the same disease, the respective datasets may be incompatible due to differing recording protocols and methodologies. For instance, both the Pitt Corpus[86] and the 2020 ADReSS Challenge dataset[87] (**Table 3**) include recordings of dementia patients and controls. However, due to different questionnaires, recording hardware, and pre-processing, the two datasets are quite incompatible. As a result, a model trained on one is unlikely to perform well on the other. The absence of a standard approach to data collection in the field hinders progress.

Reproducibility of features

An issue with audio feature-based approaches is that the features themselves are not always reproducible. Some studies found that even features extracted with established software (**Table 1**) do not produce sufficiently consistent results when compared across recordings.[26,27] While robust ML and DL approaches are able to address this problem to some extent by focusing on the important information in the features, it is difficult to complete resolve this data reliability problem algorithmically.

Biases and confounding factors

As in many applications of ML, the performance of audio analysis tools may be limited by the characteristics of their training data, raising concerns about bias, diversity and inclusion. For

instance, the vast majority of the work in the field is conducted with recordings of English-speaking individuals, raising questions about their effectiveness for other languages. Some tools may also be affected by confounding factors like dialect, accent, age, ethnocultural status (e.g., race/ethnicity), gender and medication regimen.[88,89] Few studies reviewed here reported participants' demographic characteristics, further exacerbating this problem. Future studies should use as inclusive and representative datasets and algorithms as possible, and consider confounding factors.

Conclusion

This review described a broad spectrum of work in automated audio processing that aimed to accurately characterize the clinical status of patients. These studies utilized as large datasets of audio as possible, and a wide array of statistical, traditional ML and DL techniques. Most of this work focused on neuropsychological conditions, but interesting work in other clinical areas (e.g., cardiac, infectious and sleep disorders) is emerging. The review also discussed several important challenges facing the field, as well as potential ways to address them.

The ultimate goal of this area is to clinically operationalize these algorithms to aid in monitoring and diagnosing diverse conditions. However, most of the efforts to date have been in laboratory settings, focusing on refining analytic processes. While there has been some movement towards standardization and benchmarking , there has been limited deployment of these methods. Some software systems that can be shared among clinicians have been developed. NeuroSpeech is one such system that can help analyze dysarthric speech and diagnose Parkinson's disease.[90] Another group deployed a smartphone application to recognize phonemic boundaries and interpret a patient's speaking rate for real-time speech therapy.[91]

Future work in this direction will have to address several challenges. First, an infrastructure for the recording, processing and storage of high-quality patient audio will have to be established and standardized.[8,11] These data must be obtained with patients' consent, and be stored securely to maintain privacy and confidentiality.[13] Finally, institutions may also need high-performance computing facilities for this deployment, especially to develop predictive models for their patient populations.[92] As progress is made in these directions, we expect that audio-based tools will be deployed in the clinic much more extensively in the future.

Acknowledgements

This work was supported by NIH grant #s R01 AG066471 and R01 HG011407-01A1.

Author Contributions

A.K., A.F., G.P., and T.J. conceived this review. A.K. and T.J. analyzed the literature, collected the data and drafted the manuscript under G.P.'s supervision. All the authors contributed to the drafting of the manuscript and approved the final version.

Competing interests statement

The authors declare no competing interests.

References

1. Voleti, R., Liss, J. M. & Berisha, V. A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders. *ArXiv190601157 Cs Eess* (2019).
2. Bickley, L. S., Szilagyi, P. G. & Hoffman, R. M. *Bates' guide to physical examination and history taking*. (Wolters Kluwer, 2017).
3. Pulido, M. L. B. *et al.* Alzheimer's disease and automatic speech analysis: A review. *Expert Syst. Appl.* **150**, 113213 (2020).
4. Li, Y., Lin, Y., Ding, H. & Li, C. Speech databases for mental disorders: A systematic review. *Gen. Psychiatry* **32**, e100022 (2019).
5. Cummins, N. *et al.* A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **71**, 10–49 (2015).
6. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment: MOCA: A BRIEF SCREENING TOOL FOR MCI. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).

7. Folstein, M. F. The Mini-Mental State Examination. *Arch. Gen. Psychiatry* **40**, 812 (1983).

8. Galvin, J. E. *et al.* The AD8: A brief informant interview to detect dementia. *Neurology* **65**, 559–564 (2005).

9. Percha, B. Modern Clinical Text Mining: A Guide and Review. *Annu. Rev. Biomed. Data Sci.* **4**, 165–187 (2021).

10. Cummins, N., Baird, A. & Schuller, B. W. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods* **151**, 41–54 (2018).

11. Alpaydin, E. *Machine learning*. (The MIT Press, 2021).

12. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (The MIT Press, 2016).

13. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).

14. Page, M. J. *et al.* PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* n160 (2021).

15. Bartek Plichta. Best Practices in the Acquisition, Processing, and Analysis of Acoustic Speech Signals. *Univ. Pa. Work. Pap. Linguist.* **8**, (2002).

16. Rabiner, L. R. & Schafer, R. W. *Introduction to Digital Speech Processing*. vol. 1 (now Publishers Inc., 2007).

17. Orlandic, L., Teijeiro, T. & Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Sci. Data* **8**, 156 (2021).

18. Xu, Y., Du, J., Dai, L.-R. & Lee, C.-H. A Regression Approach to Speech Enhancement Based on Deep Neural Networks. *IEEEACM Trans. Audio Speech Lang. Process.* **23**, 7–19 (2015).

19. Patro, S. G. K. & Sahu, K. K. Normalization: A Preprocessing Stage. *ArXiv E-Prints* arXiv:2002.11102 (2015).

20. Li, B., Wu, F., Lim, S.-N., Belongie, S. & Weinberger, K. Q. On Feature Normalization and Data Augmentation. *ArXiv E-Prints* arXiv:2002.11102 (2020).

21. Alam, M. J., Ouellet, P., Kenny, P. & O'Shaughnessy, D. Comparative Evaluation of Feature Normalization Techniques for Speaker Verification. in *Advances in Nonlinear Speech Processing* (eds. Travieso-González, C. M. & Alonso-Hernández, J. B.) 246–253 (Springer Berlin Heidelberg, 2011).

22. Tranter, S. E. & Reynolds, D. A. An overview of automatic speaker diarization systems. *IEEE Trans. Audio Speech Lang. Process.* **14**, 1557–1565 (2006).

23. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. in *Proc. Interspeech 2017* 498–502 (2017).

24. Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K. & Kaye, J. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE Trans. Audio Speech Lang. Process.* **19**, 2081–2090 (2011).

25. Eyben, F. *et al.* The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **7**, 190–202 (2016).

26. Boersma, P. & Weenink, D. *Praat*. (Phonetic Sciences, University of Amsterdam, 2021).

27. Eyben, F., Wöllmer, M. & Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. in *Proceedings of the 18th ACM International Conference on Multimedia* 1459–1462 (Association for Computing Machinery, 2010).

28. Zhai, Y., Ong, Y.-S. & Tsang, I. W. The Emerging 'Big Dimensionality'. *IEEE Comput. Intell. Mag.* **9**, 14–26 (2014).

29. Zar, J. H. *Biostatistical analysis*. (Prentice-Hall/Pearson, 2010).

30. Hustad Katherine C., Sakash Ashley, Broman Aimee Teo, & Rathouz Paul J. Differentiating Typical From Atypical Speech Production in 5-Year-Old Children With Cerebral Palsy: A Comparative Analysis. *Am. J. Speech Lang. Pathol.* **28**, 807–817 (2019).

31. König, A. *et al.* Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **1**,

112–124 (2015).

32. Beltrami, D. *et al.* Speech Analysis by Natural Language Processing Techniques: A Possible Tool for Very Early Detection of Cognitive Decline? *Front. Aging Neurosci.* **10**, 369 (2018).

33. Zhang, J. *et al.* Analysis on speech signal features of manic patients. *J. Psychiatr. Res.* **98**, 59–63 (2018).

34. Schuller, B. *et al.* The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proc. Interspeech* 148–152 (2013).

35. Berisha, V., Utianski, R. & Liss, J. Towards A Clinical Tool For Automatic Intelligibility Assessment. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Spons. Inst. Electr. Electron. Eng. Signal Process. Soc. ICASSP Conf.* 2825–2828 (2013).

36. Arevian, A. C. *et al.* Clinical state tracking in serious mental illness through computational analysis of speech. *PLoS ONE* **15**, 1-17 (2020).

37. Lopez-de-Ipiña, K. *et al.* On Automatic Diagnosis of Alzheimer's Disease Based on Spontaneous Speech Analysis and Emotional Temperature. *Cogn. Comput.* **7**, **44-45** (2013).

38. Jarrold, W. *et al.* Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* 27–37 (Association for Computational Linguistics, 2014).

39. Stegmann, G. M. *et al.* Repeatability of Commonly Used Speech and Language Features for Clinical Applications. *Digit. Biomark.* **4**, 109–122 (2020).

40. Gosztolya, G. *et al.* Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Comput. Speech Lang.* **53**, 181–197 (2019).

41. Yang, L. *et al.* Multimodal Measurement of Depression Using Deep Learning Models. in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* 53–59 (ACM,

2017).

42. Yang, L. *et al.* Hybrid Depression Classification and Estimation from Audio Video and Text Information. in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* 45–51 (Association for Computing Machinery, 2017).

43. He, L. & Cao, C. Automated depression analysis using convolutional neural networks from speech. *J. Biomed. Inform.* **83**, 103–111 (2018).

44. Yang, P., Hwa Yang, Y., B. Zhou, B. & Y. Zomaya, A. A Review of Ensemble Methods in Bioinformatics. *Curr. Bioinforma.* **5**, 296–308 (2010).

45. Warnita, T., Inoue, N. & Shinoda, K. Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data. *ArXiv180311344 Cs Eess* (2018).

46. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S. & Othmani, A. MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech. *ArXiv190907208 Cs Eess* (2020).

47. Livingstone, S. R. & Russo, F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* **13**, 1–35 (2018).

48. Deng, J. *et al.* Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations. in *Proceedings of the 2017 International Conference on Digital Health* 53–57 (Association for Computing Machinery, 2017).

49. Snyder, D., Garcia-Romero, D., Povey, D. & Khudanpur, S. Deep Neural Network Embeddings for Text-Independent Speaker Verification. in *Interspeech 2017* 999–1003 (ISCA, 2017).

50. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. & Ouellet, P. Front-End Factor Analysis for Speaker Verification. *IEEE Trans. Audio Speech Lang. Process.* **19**, 788–798 (2011).

51. Botelho, C., Teixeira, F., Rolland, T., Abad, A. & Trancoso, I. *Pathological speech*

detection using x-vector embeddings. arXiv preprint arXiv:2003.00864 (2020).

52. Pompili, A., Rolland, T. & Abad, A. The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge. *ArXiv200514646 Eess* (2020).

53. Zargarbashi, S. S. H. & Babaali, B. A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language. *ArXiv191000330 Cs Eess Stat* (2019).

54. Perero-Codosero, J. M., Espinoza-Cuadros, F., Antón-Martín, J., Barbero-Álvarez, M. A. & Hernández-Gómez, L. A. Modeling Obstructive Sleep Apnea Voices Using Deep Neural Network Embeddings and Domain-Adversarial Training. *IEEE J. Sel. Top. Signal Process.* **14**, 240–250 (2020).

55. Haider, F., de la Fuente, S. & Luz, S. An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech. *IEEE J. Sel. Top. Signal Process.* **14**, 272–281 (2020).

56. Zhao, Z. *et al.* Automatic Assessment of Depression From Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders. *IEEE J. Sel. Top. Signal Process.* **14**, 423–434 (2020).

57. Ma, X., Yang, H., Chen, Q., Huang, D. & Wang, Y. DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* 35–42 (ACM, 2016).

58. Kao, W.-C. & Wei, C.-C. Automatic phonocardiograph signal analysis for detecting heart valve disorders. *Expert Syst. Appl.* **38**, 6458–6468 (2011).

59. Springer, D. B., Tarassenko, L. & Clifford, G. D. Logistic Regression-HSMM-Based Heart Sound Segmentation. *IEEE Trans. Biomed. Eng.* **63**, 822–832 (2016).

60. Clifford, G. D. *et al.* Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. in *2016 Computing in Cardiology Conference (CinC)* 609–612 (2016).

61. Potes, C., Parvaneh, S., Rahman, A. & Conroy, B. Ensemble of feature-based and deep

learning-based classifiers for detection of abnormal heart sounds. in *2016 Computing in Cardiology Conference (CinC)* 621–624 (2016).

62. Zabihi, M., Rad, A. B., Kiranyaz, S., Gabbouj, M. & Katsaggelos, A. K. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. in *2016 Computing in Cardiology Conference (CinC)* 613–616 (2016).

63. Li, F. *et al.* Feature extraction and classification of heart sound using 1D convolutional neural networks. *EURASIP J. Adv. Signal Process.* **2019**, 59 (2019).

64. Raza, A. *et al.* Heartbeat Sound Signal Classification Using Deep Learning. *Sensors* **19**, 4819 (2019).

65. Swarnkar, V. *et al.* Neural network based algorithm for automatic identification of cough sounds. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2013**, 1764–1767 (2013).

66. Swarnkar, V. *et al.* Automatic identification of wet and dry cough in pediatric patients with respiratory diseases. *Ann. Biomed. Eng.* **41**, 1016–1028 (2013).

67. Kosasih, K., Abeyratne, U. R., Swarnkar, V. & Triasih, R. Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis. *IEEE Trans. Biomed. Eng.* **62**, 1185–1194 (2015).

68. Pramono, R. X. A., Imtiaz, S. A. & Rodriguez-Villegas, E. A Cough-Based Algorithm for Automatic Diagnosis of Pertussis. *PloS One* **11**, e0162128 (2016).

69. Parker, D. *et al.* Detecting paroxysmal coughing from pertussis cases using voice recognition technology. *PloS One* **8**, e82971 (2013).

70. Pham, L. *et al.* Robust Deep Learning Framework For Predicting Respiratory Anomalies and Diseases. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)* 164–167 (2020).

71. Rocha BM et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiol. Meas.* **40 035001**, (2019).

72. Zheng, Q. *et al.* HIT-COVID, a global database tracking public health interventions to COVID-19. *Sci. Data* **7**, 286 (2020).

73. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).

74. Pahar, M., Klopper, M., Warren, R. & Niesler, T. COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings. *ArXiv201201926 Cs Eess* (2020).

75. Bansal, V., Pahwa, G. & Kannan, N. Cough Classification for COVID-19 based on audio mfcc features using Convolutional Neural Networks. in *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)* 604–608 (2020).

76. Akhter, S., Abeyratne, U. R., Swarnkar, V. & Hukins, C. Snore Sound Analysis Can Detect the Presence of Obstructive Sleep Apnea Specific to NREM or REM Sleep. *J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med.* **14**, 991–1003 (2018).

77. Kim, T., Kim, J.-W. & Lee, K. Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques. *Biomed. Eng. OnLine* **17**, 16 (2018).

78. Nakano, H., Furukawa, T. & Tanigawa, T. Tracheal Sound Analysis Using a Deep Neural Network to Detect Sleep Apnea. *J. Clin. Sleep Med. JCSM Off. Publ. Am. Acad. Sleep Med.* **15**, 1125–1133 (2019).

79. Arends, J. *et al.* Diagnostic accuracy of audio-based seizure detection in patients with severe epilepsy and an intellectual disability. *Epilepsy Behav.* **62**, 180–185 (2016).

80. Zhang, T., Mustiere, F. & Micheyl, C. Intelligent hearing aids: the next revolution. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2016**, 72–76 (2016).

81. Blomberg, S. N. *et al.* Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Netw. Open* **4**, e2032320 (2021).

82. Altman, N. & Krzywinski, M. Graphical assessment of tests and classifiers. *Nat. Methods* **18**, 839–839 (2021).

83. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* **10**, e0118432 (2015).

84. Rotemberg, V., Halpern, A., Dusza, S. & Codella, N. C. The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice. *Semin. Cutan. Med. Surg.* **38**, E38–E42 (2019).

85. Saez-Rodriguez, J. *et al.* Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat. Rev. Genet.* **17**, 470–486 (2016).

86. Becker, J. T., Boller, F., Lopez, O., Saxton, J. & McGonigle, K. L. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* **51(6)**, 585–594 (1994).

87. Luz, S., Haider, F., de la Fuente, S., Fromm, D. & MacWhinney, B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. in *Proceedings of INTERSPEECH 2020* 2172–2176 (2020).

88. Markl, N. & Lai, C. Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. *Proc. First Workshop Bridg. Hum.-Comput. Interact. Nat. Lang. Process.* 34–40 (2021).

89. Cummins, N. *et al.* A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **71**, 10–49 (2015).

90. Orozco-Arroyave, J. R. *et al.* NeuroSpeech: An open-source software for Parkinson's speech analysis. *Digit. Signal Process.* **77**, 207–221 (2018).

91. Aharonson, V. *et al.* A real-time phoneme counting algorithm and application for speech rate monitoring. *J. Fluen. Disord.* **51**, 60–68 (2017).

92. Hofer, I. S., Burns, M., Kendale, S. & Wanderer, J. P. Realistically Integrating Machine

Learning Into Clinical Practice: A Road Map of Opportunities, Challenges, and a Potential Future. *Anesth. Analg.* **130**, 1115–1118 (2020).

93. McFee, B. *et al.* librosa: Audio and music signal analysis in python. in *Proceedings of the 14th python in science conference* vol. 8 18–25 (2015).

94. The MathWorks, Inc. *MATLAB Audio Toolbox*. (2021).

95. The MathWorks, Inc. *MATLAB DSP System Toolbox*. (2021).

96. Jianjing Kuang & Danni Ma. *Penn Phonetics Lab Forced Aligner for English*. (Penn Phonetics Laboratory Department of Linguistics, 2012).

97. Bredin, H. *et al.* pyannote.audio: neural building blocks for speaker diarization. in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing.* 7124-7128 (2020).

98. Zhang, A., Wang, Q., Zhu, Z., Paisley, J. & Wang, C. Fully Supervised Speaker Diarization. in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 6301–6305 (2019).

99. Reubold, U., Harrington, J. & Kleber, F. Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Commun.* **52**, 638–651 (2010).

100. Gratch et al. The Distress Analysis Interview Corpus of human and computer interviews. *Proc. Ninth Int. Conf. Lang. Resour. Eval. LREC14* 3123–3128 (2014).

Figure Legends

Figure 1: A general workflow for audio processing for clinical use. The process begins with acquisition and compilation of audio recordings from patients. The audio almost always has to be pre-processed to prepare it for downstream modeling. In a feature-based approach, features are extracted from the audio and analyzed. If working with patient speech, an optional step is to perform a speech-to-text transcription and extract textual or linguistic features. Deep learning approaches are often able to work directly with the native audio signal, without a need to explicitly extract features. In the most common applications of clinical audio, a machine learning model is then trained, evaluated and tuned from either the extracted features or the native audio signal. The model can then be integrated into an application and deployed for clinical use. Attributions: Icons in this image were taken from open-source websites, namely www.flaticon.com, www.iconfinder.com, www.clipart-library.com, www.nicepng.com, www.typiccor.com, www.kissclipart.com, www.creazilla.com, www.pngall.com and www.pinclipart.com.

Figure 2: A flowchart detailing the process used to conduct the literature search following PRISMA guidelines. First, records published after 2010 were identified from PubMed, Web of Science, and Google Scholar using the listed keywords. Duplicates were automatically removed. Through manual screening, records are retained/removed according to the specified inclusion/exclusion criteria. The main goal was to only include articles that used or described the use of audio in a clinical setting for purposes other than medical imaging or obtaining transcriptions and performing textual analysis. Of the remaining records, a representative subset were included in this review to accurately portray the scope of clinical applications of audio signals.

Figure 3: An illustration of a deep learning-based neural network model that can be used to represent audio at various levels of abstraction. The raw audio is input to the network at the waveform level, and at each successive “layer” of the network, a more abstract representation is learned. At the final layer of the network, the model predicts the most likely class that the audio originated from.

Figure 4: The architecture of a Convolutional Neural Network (CNN). Input data are passed through a convolutional filter to extract important contextual information embedded in the data through a mathematical function. The intuition behind convolutions is to mimic human neuronal processes, wherein certain neurons attend to specific parts of sensory stimuli.[12] These convolved inputs are then pooled to reduce dimensionality, and eventually passed through fully connected layers to generate classifier outputs.

Figure 5: The architecture of a Recurrent Neural Network (RNN). Ordered inputs are fed sequentially into a neural network. The “unfolding” of the network is designed to infer hidden states h_n, which are a function of the current input x_n and of the previous hidden state h_{n-1} , thus capturing the “recurrent” nature of the network. The learned representation embedded in these states is then used to produce the desired output at different points in time or in the sequence. This sequential structure of RNNs makes them particularly suitable for time series data like language, audio, and speech.

Figure 6: A binary confusion matrix for classification models that can be used to evaluate several evaluation measures. The definition of and relationships between these measures are summarized. This matrix and these measures can be similarly defined for more than two classes.

Tables

Table 1: Widely used audio pre-processing and feature extraction tools.

Tool	**Purpose**	**URL**
librosa[93]	Pre-processing, feature extraction and analysis	librosa.org
MATLAB Audio Toolbox[94]	Pre-processing, feature extraction and analysis	mathworks.com/products/audio.html
MATLAB DSP System Toolbox[95]	Pre-processing, feature extraction and analysis	mathworks.com/products/dsp-system.html
Praat[26]	Pre-processing, feature extraction and analysis	fon.hum.uva.nl/praat
openSMILE[27]	Pre-processing, feature extraction and analysis	audeering.com/opensmile
NeuroSpeech[90]	Pre-processing, feature extraction and analysis	github.com/jcvasquezc/NeuroSpeech
Penn Phonetics Lab Forced Aligner[96]	Forced Alignment	web.sas.upenn.edu/phonetics-lab/facilities
Montreal Forced Aligner[23]	Forced Alignment	montreal-forced-aligner.readthedocs.io
PyAnnote-Audio[97]	Diarization	github.com/pyannote/pyannote-audio
Fully Supervised Speaker Diarization[98]	Diarization	github.com/google/uis-rnn

Table 2: Features commonly used in audio analysis.

Category	**Feature**	**Definition**
Time	Pause rate[24]	Rate (instances per unit of time) at which an audio pauses for a duration exceeding a predefined threshold.
(features mostly related to how the signal changes over time)	Phonation rate[24]	Total phonation time (duration of real signal) divided by total duration of the audio.
	Fundamental frequency trajectory[25]	Changes in the fundamental frequency (lowest frequency) emitted over the audio
Pitch	Jitter[25]	Changes in the length of consecutive occurrences of the fundamental frequency
(features mostly related to the frequency content of the signal)	Formants (F_1, F_2, ...)[99]	Frequencies of resonances caused by the shape of the vocal tract, typically during the phonation of a vowel[16]
	Mel-frequency cepstrum coefficients (MFCC)[25]	Coefficients derived from filtering and scaling frequencies according to how human hearing typically perceives them
Energy	Loudness[25]	Perceived intensity of an acoustic signal
(features mostly related to the energy of the signal)	Harmonic Difference[25]	Ratio of energies contained in different pairs of harmonics*, often consecutive
	Harmonic to Noise Ratio (HNR)[25]	Ratio of harmonic components of a signal to the underlying signal noise
* naturally occurring frequencies at integer multiples of a fundamental frequency[16]		

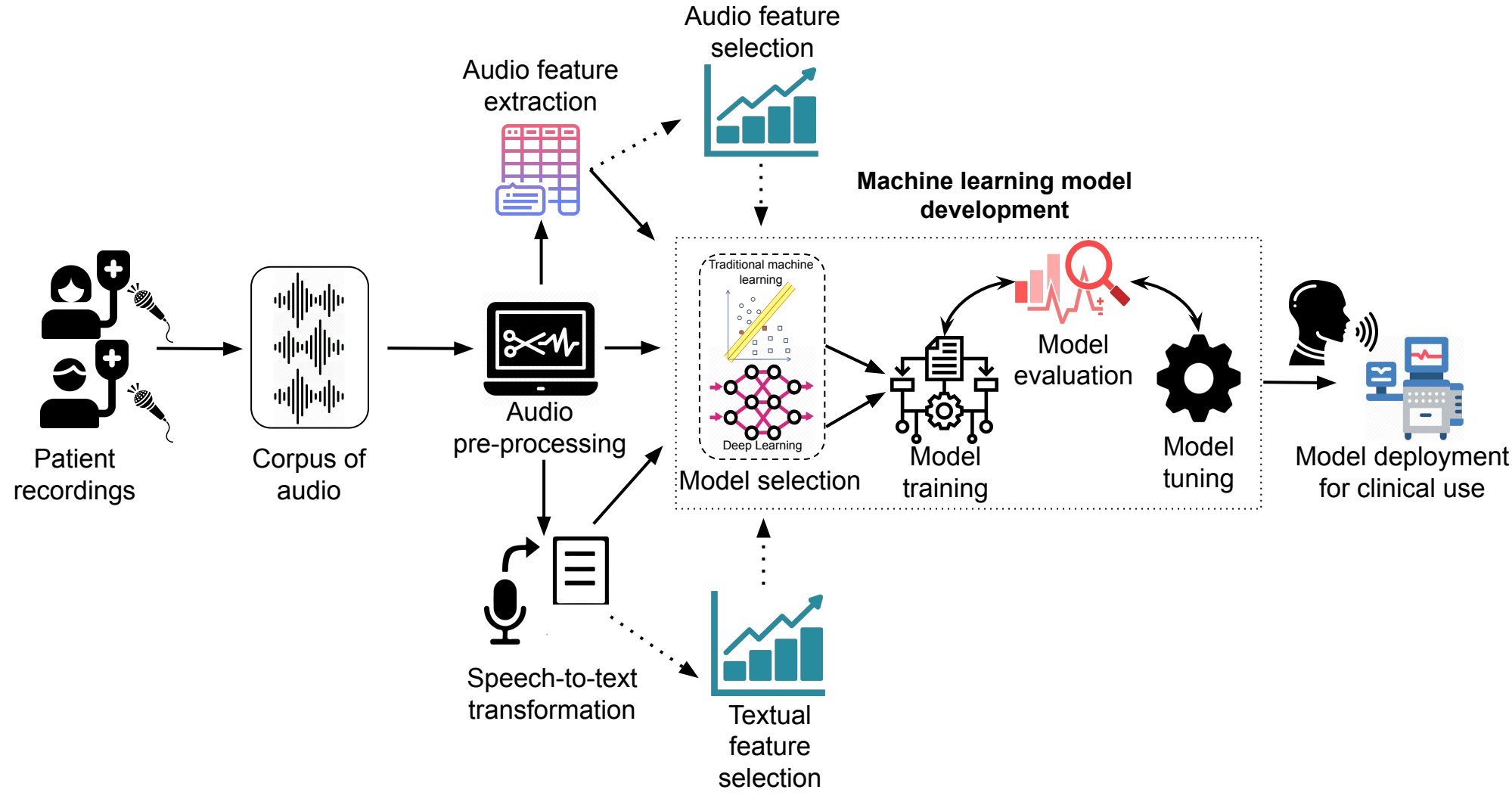
Table 3: Details of prominent crowdsourced challenges and publicly available datasets.

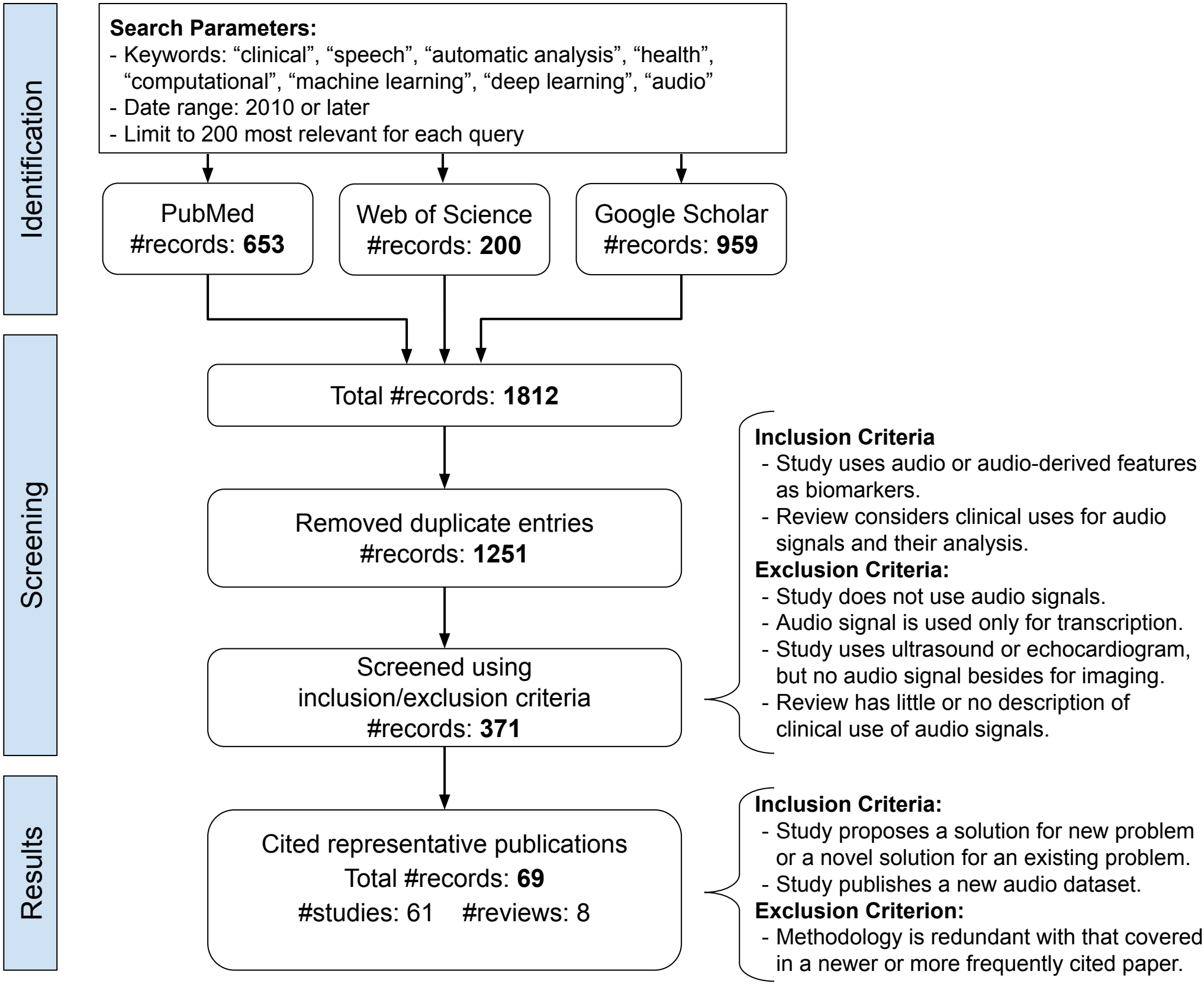
Name	Source	Clinical area	Dataset size	
			# of recordings	# of participants
SSPNet Vocalisation Corpus (SVC)[34]	INTERSPEECH 2013 ComParE	Emotional psychology	2,763	120
Geneva Multimodal Emotion Portrayals (GEMEP)[34]	INTERSPEECH 2013 ComParE	Emotional psychology	1,200	10*
Child Pathological Speech Database (CPSD)[34]	INTERSPEECH 2013 ComParE	Neurocognitive disorders	2,500	99
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[47]	SMART Lab, Ryerson University	Emotional psychology	7,356	24
Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ)[100]	USC Institute for Creative Technologies	Psychiatric disorders	189	189
ADReSS Challenge Database[87]	INTERSPEECH 2020 Challenge	Neurocognitive disorders	156	156
The Pitt Corpus[86]	University of Pittsburgh School of Medicine	Neurocognitive disorders	397	397
The PhysioNet Computing in Cardiology Challenge 2016 Dataset[60]	The PhysioNet Computing in Cardiology Challenge 2016	Cardiology	4,430	1,072†
The Respiratory Sound Database[71]	Internal Conference on Biomedical Health Informatics (ICBHI) 2017 challenge	Pulmonology	6,898	120

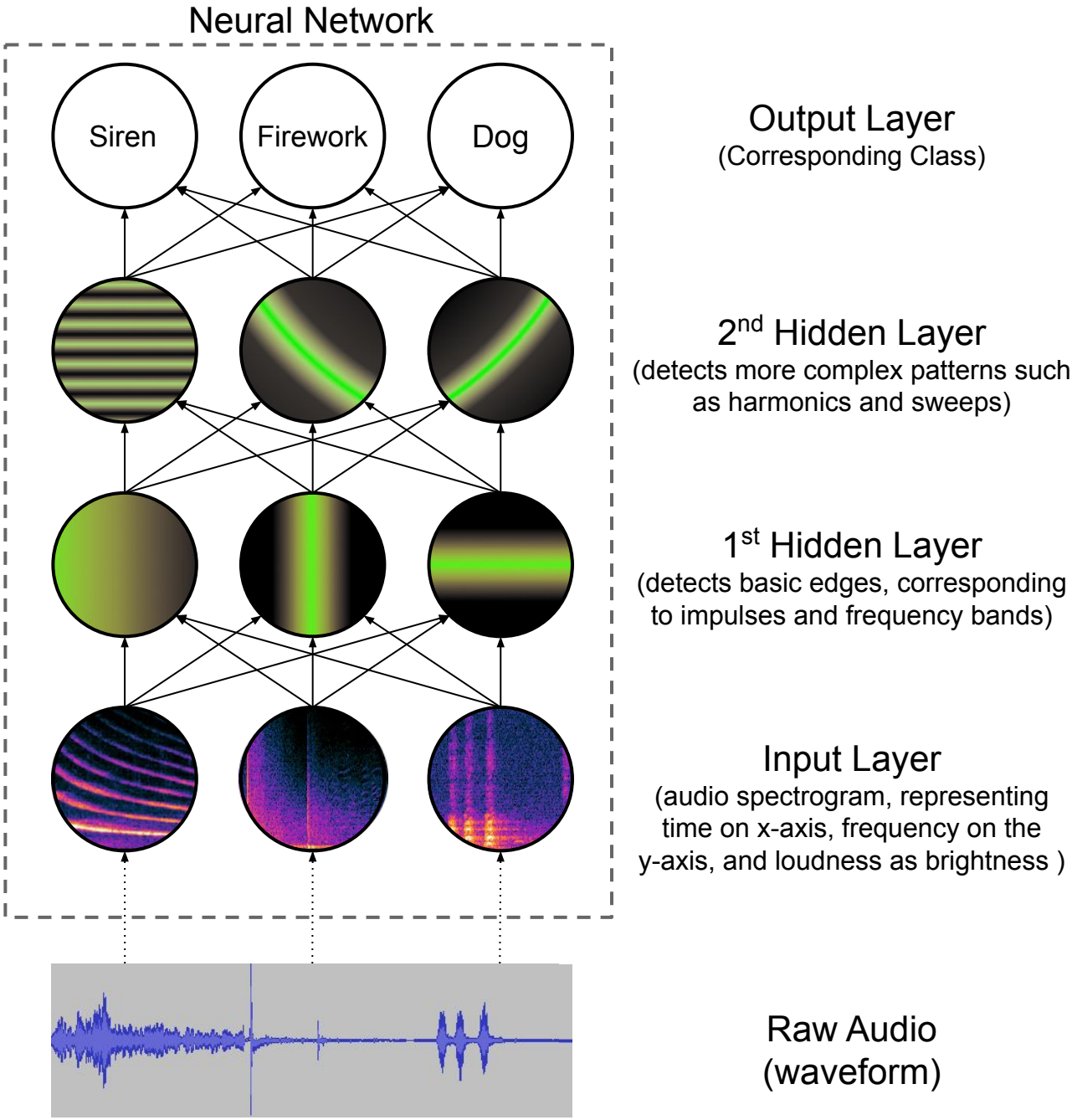
COUGHVID Crowdsourced COVID16 database[17]	Swiss Federal Institute of Technology Lausanne (EPFL)	Pulmonology	>25,000	>2800†

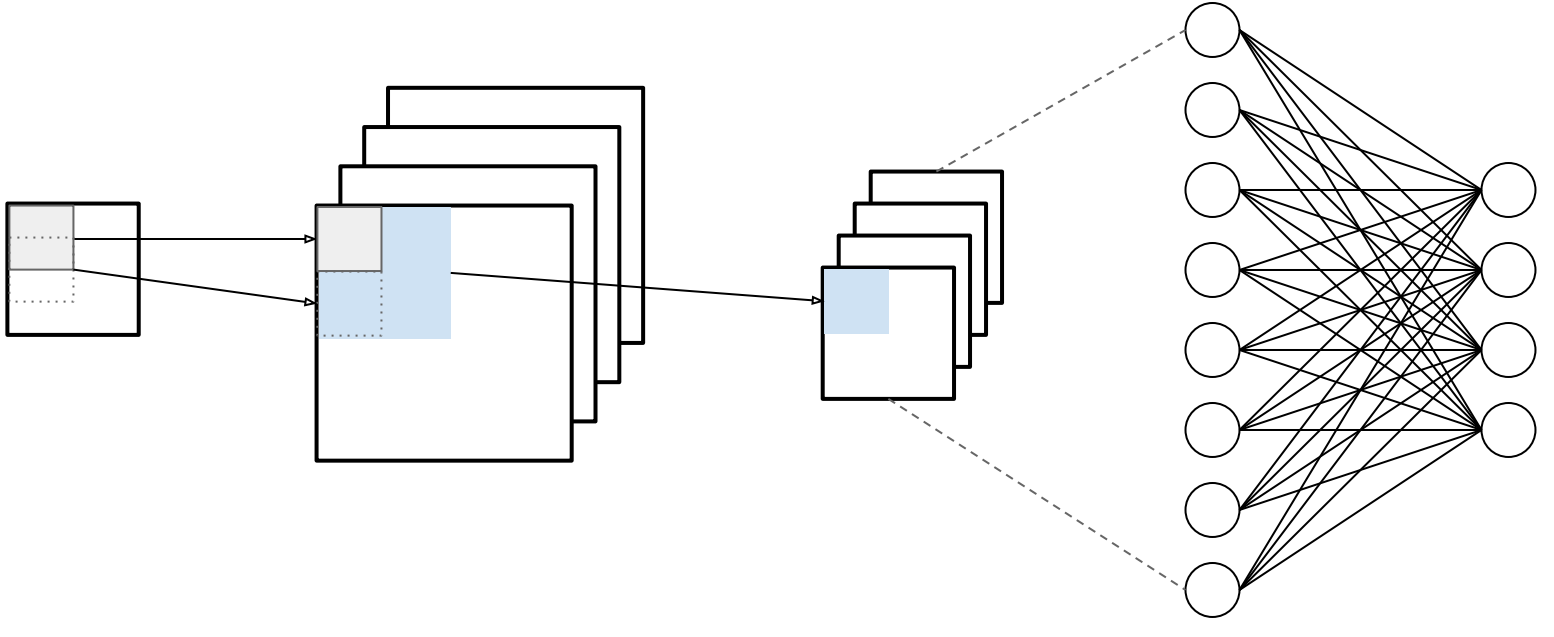
* Study used 10 actors, 5 identified as male and 5 as female, to recreate a number of emotional scenarios.
† Study is ongoing and the dataset continues to grow. These figures were taken from the most recent publication cited for this dataset.

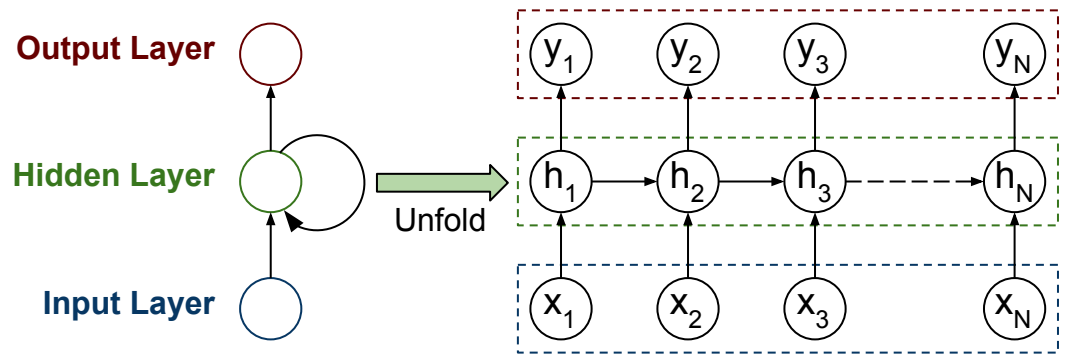
Clinical Audio Processing Workflow




Identification
Search Parameters:
- Keywords: “clinical”, “speech”, “automatic analysis”, “health”, “computational”, “machine learning”, “deep learning”, “audio”
- Date range: 2010 or later
- Limit to 200 most relevant for each query
PubMed #records: 653
Web of Science #records: 200
Google Scholar #records: 959
Screening
Total #records: 1812
Removed duplicate entries #records: 1251
Screened using inclusion/exclusion criteria #records: 371
Inclusion Criteria
- Study uses audio or audio-derived features as biomarkers.
- Review considers clinical uses for audio signals and their analysis.
Exclusion Criteria:
- Study does not use audio signals.
- Audio signal is used only for transcription.
- Study uses ultrasound or echocardiogram, but no audio signal besides for imaging.
- Review has little or no description of clinical use of audio signals.
Results
Cited representative publications
Total #records: 69
#studies: 61 #reviews: 8
Inclusion Criteria:
- Study proposes a solution for new problem or a novel solution for an existing problem.
- Study publishes a new audio dataset.
Exclusion Criterion:
- Methodology is redundant with that covered in a newer or more frequently cited paper.


Neural Network
Siren
Firework
Dog
Output Layer
(Corresponding Class)
2nd Hidden Layer
(detects more complex patterns such as harmonics and sweeps)
1st Hidden Layer
(detects basic edges, corresponding to impulses and frequency bands)
Input Layer
(audio spectrogram, representing time on x-axis, frequency on the y-axis, and loudness as brightness)
Raw Audio
(waveform)


Input
Convolutional Layer
Pooling Layer
Hidden Layer
Output


Output Layer
Hidden Layer
Input Layer
Unfold
y1
y2
y3
yN
h1
h2
h3
hN
x1
x2
x3
xN

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$Accuracy = \frac{a + d}{a + b + c + d}$$

$$Precision_{Yes}(P_{Yes}) = PPV = \frac{a}{a + c} \qquad Recall_{Yes}(R_{Yes}) = Sensitivity_{Yes} = TPR = \frac{a}{a + b}$$

$$Precision_{No}(P_{No}) = NPV = \frac{d}{b + d} \qquad Recall_{No}(R_{No}) = Sensitivity_{No} = 1 - FPR = \frac{d}{c + d}$$

$$F - measure_{Yes}(F_{Yes}) = \frac{2P_{Yes}R_{Yes}}{P_{Yes} + R_{Yes}} = \frac{2a}{2a + b + c}$$

$$F - measure_{No}(F_{No}) = \frac{2P_{No}R_{No}}{P_{No} + R_{No}} = \frac{2d}{2d + b + c}$$

$ROC\ Curve = Variation\ of\ TPR\ against\ FPR\ across\ all\ classification\ threshold$

$AUC\ score = Area\ under\ the\ ROC\ curve$