



# OPEN Transformers for rapid detection of airway stenosis and stridor

James Anibal<sup>1,2✉</sup>, Rebecca Doctor<sup>3</sup>, Micah Boyer<sup>3</sup>, Karlee Newberry<sup>3</sup>, Iris De Santiago<sup>3</sup>, Shaheen Awan<sup>4</sup>, Yassmeen Abdel-Aty<sup>3</sup>, Gregory Dion<sup>5</sup>, Veronica Daoud<sup>6</sup>, Hannah Huth<sup>1</sup>, Stephanie Watts<sup>3</sup>, Bradford J. Wood<sup>1</sup>, David Clifton<sup>2</sup>, Alexander Gelbard<sup>7</sup>, Maria Powell<sup>8</sup>, Jamie Toghranegar<sup>3</sup>, the Bridge2AI Voice Consortium\* & Yael Bensoussan<sup>3</sup>

Upper airway stenosis is a potentially life-threatening condition involving the narrowing of the airway. In more severe cases, airway stenosis may be accompanied by stridor, a type of disordered breathing caused by turbulent airflow. Patients with airway stenosis have a higher risk of airway failure and additional precautions must be taken before medical interventions like intubation. However, stenosis and stridor are often misdiagnosed as other respiratory conditions like asthma/wheezing, worsening outcomes. This report presents a unified dataset containing recorded breathing tasks from patients with stridor and airway stenosis. Customized transformer-based models were also trained to perform stenosis and stridor detection tasks using low-cost data from multiple acoustic prompts recorded on common devices. These methods achieved AUC scores of 0.875 for stenosis detection and 0.864 for stridor detection, demonstrating the potential to add value as screening tools in real-world clinical workflows, particularly in high-volume settings like emergency departments.

Upper airway stenosis is defined as a narrowing of the airway which can occur at various levels within the laryngo-tracheal apparatus and can be caused by factors ranging from abnormal scar tissue, traumatic injuries, malignancy, or neurological paralysis of the vocal folds. When the narrowing is significant, airway stenosis often results in stridor, a respiratory bio-acoustic signal caused by turbulent airflow and tissue vibration at different levels within the upper airway<sup>1,2</sup>. Vibrating tissue and turbulent airflow through the human airway create unique acoustic signals.

In previous literature, stridor has been characterized as inspiratory, expiratory, or biphasic, with some proposed correlation to the level of the obstruction<sup>4</sup>. Stridor has also been classified by anatomical location, including supraglottic, glottic, subglottic, extra-thoracic, and intrathoracic<sup>5</sup>. Recognizing stridor is an important clinical skill, as trained experts will listen to stridor and patient respiratory effort to assess the severity of different airway pathologies and determine the appropriate intervention. For example, the presence of inspiratory stridor is used to differentiate severity and treatment for patients with parainfluenza virus causing subglottic swelling<sup>6</sup>. Acute onset stridor may indicate critical decompensation of a patient's upper airway patency, which can be life-threatening if not urgently identified and treated, such as in cases of epiglottitis or anaphylaxis<sup>7</sup>. Stridor can also act as a clinical indicator for intubation, or in assessment of post-extubation laryngeal edema<sup>8</sup>.

The misdiagnosis of airway stenosis can delay onset of critical treatment interventions. An informal poll conducted on an international Facebook group for over 8200 patients with Idiopathic Subglottic Stenosis found that 96% of the participants reported being initially misdiagnosed, causing extended periods of unresolved health challenges/risk<sup>9</sup>. Stridor is a common symptom of airway stenosis, but is often mistaken as wheezing in clinical practice, resulting in frequent misdiagnosis as asthma<sup>10</sup>. Failure to recognize stridor as a sign of severe airway stenosis can lead to important morbidity and mortality. Patients with stridor may need emergent interventions or, if undergoing procedures, customized intubation with technologies like specialized fiberoptic laryngoscopes.

The gold standard for investigating pathologies causing airway stenosis and potentially stridor is through direct visualization by laryngoscopy and bronchoscopy<sup>11</sup>. Unfortunately, many misdiagnosed patients are not referred to expert otolaryngologists promptly, which can result in significant, high-risk delays in the onset of

<sup>1</sup>Center for Interventional Oncology, NIH Clinical Center, National Institutes of Health, Bethesda, USA.

<sup>2</sup>Computational Health Informatics Lab, Institute of Biomedical Engineering, University of Oxford, Oxford, England.

<sup>3</sup>Department of Otolaryngology-Head & Neck Surgery, USF Health Voice Center, University of South Florida, Tampa, USA.

<sup>4</sup>School of Communication Sciences & Disorders, University of Central Florida, Orlando, USA. <sup>5</sup>Department of

Otolaryngology-Head and Neck Surgery, University of Cincinnati, Cincinnati, USA. <sup>6</sup>USF Health Morsani College of Medicine, University of South Florida, Tampa, USA. <sup>7</sup>Department of Otolaryngology-Head and Neck Surgery,

Vanderbilt University, Nashville, USA. <sup>8</sup>Department of Otolaryngology-Head & Neck Surgery, Vanderbilt University Medical Center, Nashville, USA. \*A list of authors and their affiliations appears at the end of the paper. ✉email:

anibaljt@nih.gov

care. Moreover, many otolaryngologists do not perform in-office laryngoscopies below the level of the glottis and may miss subglottic and tracheal stenosis.

Recent advances in machine learning offer the means to potentially recognize unique acoustic signals from different pathologies. Machine learning models may be more sensitive than the human ear and might be trained to identify airway stenosis and stridor, enabling non-invasive screening methods or decision-support tools for healthcare professionals. However, due to challenges in data acquisition and annotation, limited work has been completed in this space compared to areas such as cough assessment with AI models<sup>12</sup>. This work makes the following contributions to the field of automated airway health evaluation:

**A unified dataset to train AI models for detection of airway stenosis and stridor**

The dataset used in this study contained audio recordings from patients with upper airway stenosis, stridor, other respiratory/voice conditions, and general controls. The recordings in this dataset were collected non-invasively from low-cost point-of-care devices via tasks which can be rapidly completed and are not dependent on literacy.

**Airway stenosis detection with deep learning**

A customized transformer model was trained to perform stenosis detection using YAMNet representations of forced inhales and normal deep breaths, offering a potential solution to the challenging problem of misdiagnosis.

**Stridor detection with deep learning**

A second transformer model was trained to identify stridor from YAMNet embedding matrices. This model may be used with the airway stenosis model in a hierarchical system for additional severity assessments (e.g., pre-intubation) or independently for the detection of stridor in cases which may not involve upper airway stenosis but could still be life-threatening.

While the AI models introduced in this report will not replace an expert otolaryngologist with a trained ear and access to imaging technologies, these preliminary tools could accelerate the initial identification of airway stenosis or stridor. This may enhance the delivery of care and ensure that patients are promptly referred to specialists.

**Related work**

There have been relatively few successful attempts to classify airway stenosis with machine learning. Past experiments were mainly centered around small datasets collected from invasive or high-cost modalities such as electromyography, spirometry, and imaging studies<sup>13–16</sup>. Similarly, there has been limited work on the use of AI to detect stridor. One recent effort utilized video-polysomnography data and few-shot learning to detect stridor in a small number of patients (n = 18)<sup>17</sup>. This approach achieved over 96% detection accuracy using data from only eight stridor patients for training the model, outperforming a state-of-the-art AI baseline by 4–13%<sup>17</sup>. A different study attempted to use power spectrums and neural networks to classify nine different categories of respiratory sounds, one of which included stridor<sup>18</sup>. This was done on a dataset of 36 patients collected via online sources and a high-cost electronic stethoscope, with only two patients per class in the training set<sup>18</sup>. Although progress has been made in advancing stridor detection with AI, further advancements are necessary to classify stridor using more scalable data collection techniques like acoustic recordings from smartphones or other low-cost devices.

**Methods**

In this section, datasets, preprocessing methods, machine learning techniques, and validation strategies are described in detail. Table 1 provides definitions for key clinical terminology used throughout this report. Experiments performed in this study were carried out in accordance with relevant guidelines/regulations and were approved by the Institutional Review Board (IRB) of the University of South Florida.

**Datasets**

The data used in this study comes from two main sources: a dedicated stenosis/stridor dataset (“USF Stridor”) collected from the USF Health Voice Center at the University of South Florida (USF) and the Bridge2AI Voice Dataset from the Bridge2AI Voice Data Generation Project (“Bridge2AI”) <sup>19</sup>. In this study, the Bridge2AI dataset was used for additional airway stenosis records and the diversification of the control cohort. The unified dataset

| Name                  | Description                                                                                                 | Conditions or causes                                                                                                                        |
|-----------------------|-------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Upper airway stenosis | Upper airway narrowing including narrowing at any level of the supraglottis, glottis, subglottis or trachea | Supraglottic stenosis<br>Posterior glottic stenosis<br>Subglottic stenosis<br>Tracheal stenosis<br>Bilateral vocal fold paralysis           |
| Stridor               | Audible respiratory sound due to turbulent airflow in the upper airway caused by airway narrowing           | Upper airway stenosis<br>Epiglottitis<br>Hypopharyngeal/oropharyngeal masses<br>Laryngeal edema<br>Trauma<br>Inhalation injury<br>Infection |

**Table 1.** Definition of clinical terms referenced in this study.

| Cohort                      | Male | Female | Other<br>N/A | 18–34 | 34–50 | 50–64 | 65+ | N/A |
|-----------------------------|------|--------|--------------|-------|-------|-------|-----|-----|
| Airway stenosis (pathology) | 24   | 69     | 1            | 14    | 8     | 41    | 29  | 2   |
| Stridor (sound)             | 9    | 36     | 0            | 7     | 2     | 20    | 15  | 1   |
| Other respiratory/voice     | 20   | 30     | 1            | 4     | 6     | 15    | 24  | 2   |
| General controls            | 30   | 41     | 6            | 28    | 11    | 10    | 28  | 0   |

**Table 2.** Demographic statistics from the unified dataset (USF Stridor and Bridge2AI).

| Task                       | Abbreviation | Description                                                   |
|----------------------------|--------------|---------------------------------------------------------------|
| Forced inhale mouth opened | FIMO         | Patients are asked to take a quick inhale through their mouth |
| Deep breath                | DB           | Patients are asked to take a deep breath                      |

**Table 3.** Acoustic tasks used to collect stridor sounds.

contained 94 patients in the airway stenosis cohort, 45 patients in the stridor cohort, 51 patients in the cohort of other disordered voice/respiratory disease controls, and 77 patients in the general control cohort (no medical conditions or unknown health history). Informed consent was obtained prior to the collection of any data.

The disordered voice and respiratory disease control cohort contains patients with the following conditions: recurrent respiratory papillomatosis (RRP), COPD, asthma, spasmodic dysphonia, obstructive sleep apnea (OSA), chronic cough, benign vocal cord lesion, laryngeal cancer, and vocal fold paralysis. Table 2 contains demographic information about each cohort in the dataset (there is an overlap between the stenosis and stridor cohorts).

*Acoustic tasks*

In this study, AI models were trained using audio data collected from two different acoustic tasks (Table 3): forced inhale with the mouth open (FIMO) and normal deep breaths (DB). The tasks were chosen by expert practitioners and laryngologists for multiple reasons, including simplicity, no dependency on literacy or English fluency, elimination of nasal interference due to nasal turbulent airflow (often caused by seasonal allergies or other benign factors), and increased sensitivity to mild stridor due to the forcible pushing of air over the vocal cords (FIMO), increasing the likelihood of audible disordered sound. Before completing these tasks, participants were given the following instructions:

Deep breath data Take five big breaths in and out through your mouth.

Forced inhales through the mouth Exhale, then inhale quickly through your mouth, as if you are trying to catch your breath. Record three of these breaths in a single recording.

Data collection devices FIMO and DB recordings were collected from various technologies depending on the study (USF stridor or Bridge2AI) and the availability of devices. These included an At2035 AudioTechnica microphone placed at 12 inches from the mouth, a previously validated low-cost Avid AE-36 headset microphone placed at 2–3 inches from the mouth, and an internal iPad microphone placed at 6 inches from the mouth<sup>20,21</sup>. To develop generalizable AI methods that can be deployed in real-world settings, compatibility with multiple devices is essential due to the differences in technological ecosystems across different healthcare centers.

For patients in the USF Stridor study, audio data was collected from the full range of recording technologies described above (when possible). However, due to limitations related to the availability of devices and patient compliance, this was not always feasible, demonstrating the need for scalable solutions which can accommodate different types of data. If data from multiple devices was available for one patient, all the data was retained to ensure the maximum number of audio samples were available for model training. These different devices produced similar (but not identical) data points, which may be comparable to audio data augmentation techniques like spectrogram masking or the addition of random noise<sup>22</sup>. For the controls in the Bridge2AI Voice dataset, audio data was collected using only the Avid headset microphone placed 2–3 inches from the mouth. Table 4 includes the quantities of data collected from different microphones.

Data annotations Clinical metadata (patient and physician reporting) was then used to determine if the patients had upper airway stenosis or any other respiratory conditions (the latter being a key consideration in the curation of robust control cohorts). The presence or absence of stridor was determined by a laryngologist with expertise in airway stenosis. These annotations were based on respiratory phase (inspiration/expiration) and the presence of audible stridor (stridor or no stridor). Due to the use of a single annotator, which may more directly resemble scalable clinical data collection processes for real-world environments (particularly in resource-constrained settings), this data was considered weakly supervised<sup>23</sup>. The FIMO data was used as the basis for annotation due to the higher likelihood of stridor detection during this type of breathing compared to other tasks. For the stridor detection experiments, patients with airway stenosis but no audible stridor were included in the cohort of controls with other respiratory conditions.

| Cohort                      | Device       | FIMO | DB   |
|-----------------------------|--------------|------|------|
| Airway stenosis (pathology) | At2O35       | 314  | 0    |
|                             | Avid headset | 1970 | 4590 |
|                             | iPad         | 575  | 755  |
| Stridor (sound)             | At2O35       | 208  | 0    |
|                             | Avid headset | 1134 | 1768 |
|                             | iPad         | 499  | 594  |
| Respiratory controls        | At2O35       | 0    | 0    |
|                             | Avid headset | 1141 | 2989 |
|                             | iPad         | 0    | 0    |
| General controls            | At2O35       | 94   | 0    |
|                             | Avid headset | 1596 | 4608 |
|                             | iPad         | 299  | 379  |

**Table 4.** Statistics on microphone types and acoustic tasks within the cohorts included in the training datasets (there was overlap between the stenosis and stridor cohorts).

Deep breathing recordings collected during the same recording session as the FIMO task were assumed to have the same stridor status as the FIMO data. This resulted in weakly supervised “inferred annotations” which may present an additional source of regularization for training deep learning models, possibly enhancing downstream performance<sup>23</sup>. However, only the FIMO data was used for validation and testing of the stridor detection models. Other tasks were not directly assessed for the presence of stridor, and the inferred annotations were not considered sufficiently robust to be included in the reporting of model performance metrics.

*Data quality control*

Acoustic data was excluded from the study if the patient was fully non-compliant with the instructions (e.g., phonated an elongated vowel instead of deep breathing). For the stenosis and stridor cohorts, data was also removed if any errors were made in the data collection protocol (e.g., breathing through the nose) that could have obscured true biomarkers of disease. The control cohorts were not subjected to this second phase of filtering, potentially introducing noise into the dataset which may be similar to disordered breathing. These cohort-specific quality control pipelines were applied to increase the likelihood that more challenging negative cases would be shown to the model during training. For example, a healthy patient may phonate a gasping sound when asked to forcibly inhale. This can result in a stridor-like sound unrelated to airway obstruction. The presence of such data in the training set may enhance downstream generalization by encouraging the model to learn nuanced, disease-specific representations which enable separation between actual pathological signals and benign variations.

**Data preprocessing**

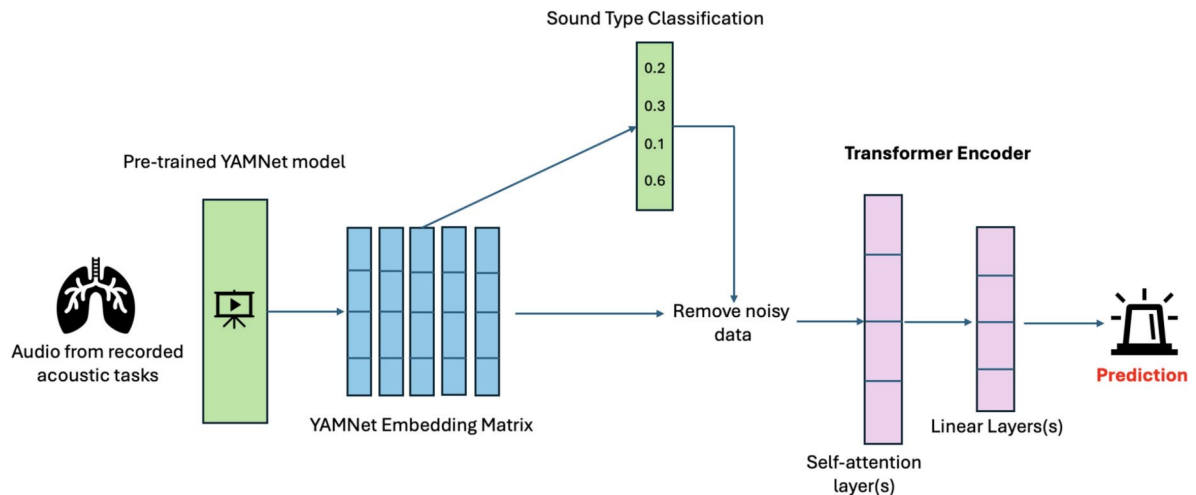
The FIMO and DB audio signals were broken into 2.5-second segments of raw audio waveforms, with 80% overlap. The segment length and overlap percentage were based on the timing of breathing cycles to ensure that some component of the inhale signal was present in all the data points: exhalations are much less likely to contain stridor or other biomarkers of airway narrowing during FIMO and DB acoustic exercises. The 80% overlap was further used as a form of data augmentation, to expand the number of “different” audio segments available for training AI models. The preprocessing steps resulted in a final training dataset of 19,310 audio segments in total. The number of segments per microphone, task, and cohort can be found in Table 4.

**YAMNet embeddings**

After initial preprocessing, the 2.5 s audio segments were standardized to a sampling rate of 16.1 kHz, normalized to the range of [−1,1], converted into Mel spectrograms, and encoded by the YAMNet model<sup>24,25</sup>. YAMNet was chosen due to extensive pre-training on millions of YouTube audio segments from 521 categories of sound<sup>24,25</sup>. This domain knowledge ensures that the model can encode nuanced differences between sounds and may enhance performance on tasks involving breathing audio data with subtle acoustic biomarkers of disease. The YAMNet model returns a matrix containing one embedding vector for each 0.5 s window of data (5×1024 dimensions for a 2.5 s segment). For this study, YAMNet was also used as part of the data filtering process. After generating the embedding matrix, the predicted probabilities for the sound type were output by the classification head of the model. If more than 50% of the segment was placed in a category which indicated a potentially significant source of external noise or microphone malfunction, such as electronic buzzing or humming, the segment was removed from the dataset.

**Model training**

A transformer model was trained on the YAMNet embedding matrices to detect airway stenosis or stridor (Fig. 1)<sup>26</sup>. Transformers rely on global self-attention mechanisms to capture the key relationships (between embedding vectors in the matrix) to which the model should “pay attention” and have demonstrated robust performance on advanced tasks with similar data structures (e.g., matrices of word or image embeddings)<sup>26–28</sup>.



**Fig. 1.** The data processing and modeling pipeline for stenosis and stridor detection.

Unlike conventional transformers for natural language processing or computer vision, positional encodings were not used within the model architecture. This minor adaption of the original transformer algorithm was done to ensure translational equivariance for audio data points with similar features at different positions in the matrix of YAMNet embedding vectors. Due to the application of data chunking with overlap, absolute position would not be relevant and may cause model confusion.

An adaptive movement estimation function was used to train the model based on the cross-entropy loss calculated from minibatches of data. Early stopping was used to optimize the duration of training based on the validation loss. During training, no distinctions were made between data from the different recording devices included in the experiments (Avid, AtO235, iPad). PyTorch was used for the development of all AI software described in this report<sup>29</sup>.

The workflow shown in Fig. 1 includes (1) the generation of YAMNet embedding matrices from Mel Spectrograms, (2) the filtering of noisy data based on sound type classification, and (3) the use of a transformer encoder to predict the label of the input data.

### Validation

Nested  $k$ -fold cross validation was used to determine the performance of the transformer models, ensuring a credible evaluation of generalization potential. The prediction threshold for the test data was determined by aggregating results from the  $k$  validation sets. If a test patient had data available from multiple recording systems, the prediction was made based on the best available device in terms of representation within the training dataset (by number of samples). For this study, data from either the Avid headset, the iPad, or the AtO235 AudioTechnica, respectively, were considered in the final prediction. Here, the aim was to simulate the real-world case wherein only a single device (presumably the best device available at that time) would be used to collect data from a patient – multiple devices would not be used due to time/resource limitations. For stenosis detection, if a test patient had available data from multiple acoustic exercises (FIMO, DB), inference was run separately, and the mean probabilities were calculated for each task. The average of these probabilities was used for decision-making purposes. In the experiments involving stridor detection, only FIMO data was used in the validation/test sets due to the nature of the manual annotation (DB was not checked for stridor).

### Results

In this section, results are presented for transformer models that were trained to complete two tasks from recorded breathing data: (1) the detection of airway stenosis and (2) the detection of stridor. Potential use cases for these models are described in Table 5. All performance metrics are reported as the mean result across multiple experimental iterations, accounting for variability due to the random initialization of transformer weights.

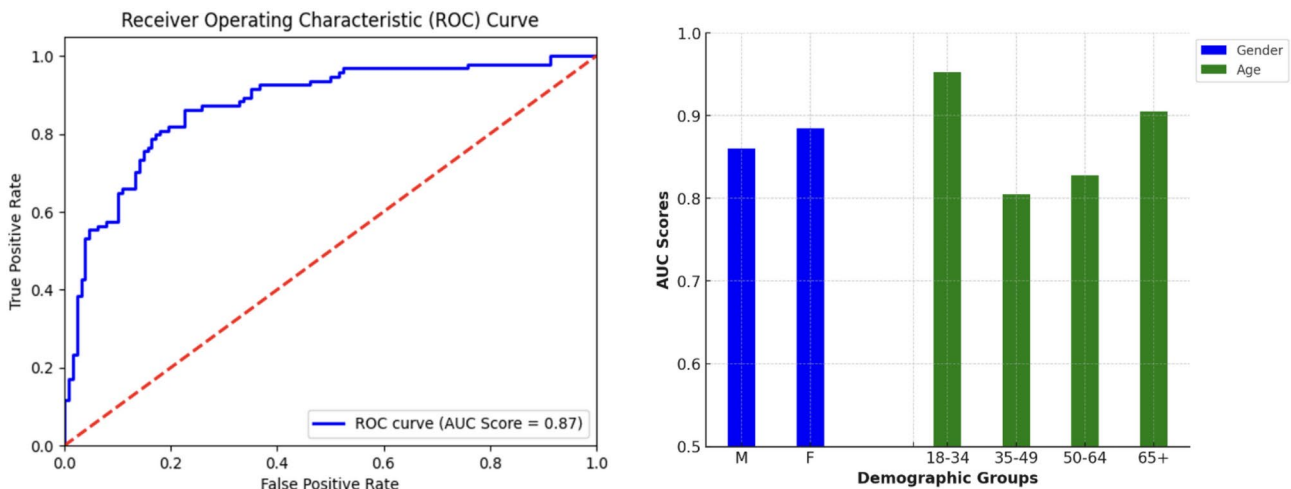
Table 6 shows the sensitivity (calibrated to a value of 0.85), specificity (at 0.85 sensitivity) and the AUC score. The life-threatening nature of airway obstruction justifies the prioritization of sensitivity in this case. Results are presented using data from the “best available” recording device for a test patient (as described in Section 3.5), mirroring the likely scenario where multiple sites have deployed the same AI model, but different devices are used depending on technology resources. If, for a patient in the test set, audio from the Avid headset was available, results are reported based on predictions made from this data. Otherwise, results are reported from the iPad or AtO235 recordings (in that order), favoring optimal data while accounting for potential technology limitations in real-world clinical environments. For specificity, three values are reported: overall, general control cohort, and respiratory control cohort (as defined in Sect. 3.1).

| Task                                                                                                                                | Use cases                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Detection of airway stenosis (in cases with and without stridor)                                                                    | Primary care settings for complaints related to the airway (e.g., to reduce the misdiagnosis of stenosis as asthma) |
| Detection of the stridor sound as a symptom of airway narrowing, which is often caused by airway stenosis but may have other causes | ER/Urgent care settings, for detection of serious airway problems requiring urgent attention, regardless of cause   |

**Table 5.** Potential use cases for AI models trained within this study.

| Task                                                          | AUC   | Sensitivity | Specificity @0.85 sensitivity |
|---------------------------------------------------------------|-------|-------------|-------------------------------|
| Stenosis detection                                            | 0.875 | 0.851       | 0.725 (0.792/0.624)           |
| Stridor detection                                             | 0.864 | 0.856       | 0.684 (0.742/0.624)           |
| Stridor detection (FIMO only, no DB data in the training set) | 0.846 | 0.852       | 0.661 (0.703/0.617)           |

**Table 6.** Results from airway stenosis detection and stridor detection tasks.



**Fig. 2.** Left: mean ROC curve for the airway stenosis detection model. Right: Model performance (mean AUC score) across age and gender identity groups.

**Detection of airway stenosis**

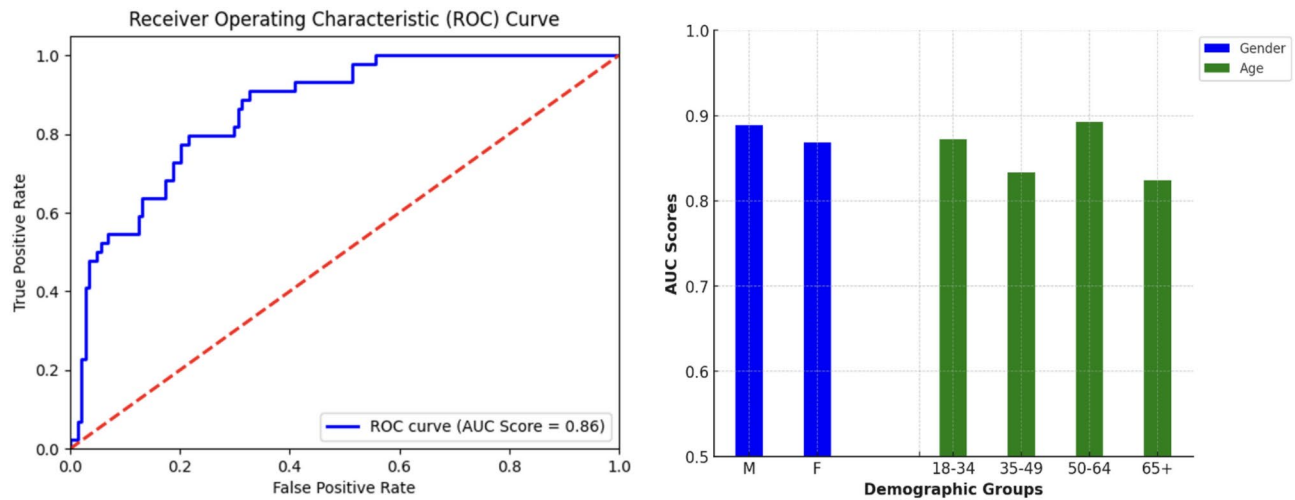
The results of this study show that the transformer model trained on YAMNet embeddings may have learned a moderately robust signal for the detection of airway stenosis (AUC of 0.875), with an overall specificity of 0.725 when sensitivity is calibrated to 0.85. Compared to other types of errors, the model is more likely to incorrectly classify other respiratory controls as having airway stenosis. Given the life-threatening implications of intubating a patient with airway stenosis and the 96+ % misdiagnosis rate, this model may still represent a deployable technology with potential for real-world impact<sup>9</sup>.

Figure 2 shows the difference in AUC score between different demographic groups within the dataset, addressing possible sources of bias in the training data. While the model was reasonably consistent, there were noticeable performance differences between age groups, possibly due to the differing age distributions between the control cohorts (Table 2). The general controls may be less challenging for the model compared to those with other respiratory conditions. The model also obtained a higher AUC score for female patients, which aligns with the distribution of upper airway stenosis cases in the dataset (Table 2) and in the real-world (the disease most often affects middle-aged women)<sup>30</sup>.

**Detection of stridor as a general symptom of airway failure**

A transformer model trained on YAMNet embedding matrices showed promising results on the stridor detection task as well, achieving an AUC score of 0.864 with an overall specificity of 0.684 (at 0.85 sensitivity). Expectedly, the performance of the model was weakest on the cohort of controls with other respiratory conditions, which included upper airway stenosis without audible stridor. Figure 3 (right) shows the difference in AUC score between different demographic groups in the dataset.

Moreover, while only FIMO data was used in the validation/test sets, the transformer model which was trained on both FIMO and DB data achieved the best results. This was despite the weak “inferred” DB annotations in the



**Fig. 3.** Left: mean ROC-AUC curve for the stridor detection model. Right: Model performance (mean AUC score) across age and gender identity groups.

training data. The transformer model trained without the DB data resulted in a lower AUC score of 0.846, with nearly a 4% difference in specificity for the general control cohort.

## Discussion

In this report, a customized transformer model was trained to screen for airway stenosis and stridor using non-invasive data from multiple devices/acoustic tasks. A unified dataset was introduced for training the AI models, containing FIMO and deep breath records from two sources – the USF Stridor dataset and the Bridge2AI Voice dataset. Upon completion of training, resultant AI models showed that there may be viable signal within low-cost, weakly annotated respiratory recordings, facilitating the identification of airway stenosis and stridor cases. These findings may have relevance in the development of new digital health technology leveraging voice as a biomarker of health, with potential use cases in the ER, primary care/point of care settings, and voice health/laryngology clinics (Table 5).

## Limitations

Promising results were obtained by this study, but there were multiple limitations that should be noted in the interpretation of the work. First, while larger than previous studies involving stridor and airway stenosis, the dataset was still limited in scope. The stridor cohort included 45 adult patients (no children or neonates) with data that was labeled by a single expert annotator, which may have implications for the generalizability of the weakly supervised AI system. Pediatric and neonatal patients may exhibit stridor as a symptom of airway narrowing for a variety of different reasons, such as croup or laryngeal defects. The stridor annotations, while performed by an expert clinician, were done based on audio recordings alone, rather than multimodal data involving laryngoscope videos and EHR, which may have resulted in some instances of mislabeled data. The annotations were also binary, without information on the severity of the airway stenosis or stridor cases. Limited variations in audio recording equipment were used for data acquisition, and evolution of audio hardware may lead to distribution shifts affecting model performance.

In terms of performance, the models showed differences in AUC score between age groups and a slight skew towards female gender identity (though the conditions themselves more often impact female patients). Other potential biases like race/ethnicity, smoking status, and type of stridor (i.e., cases not caused by airway stenosis) must still be considered in future work. Moreover, the two datasets used in this study were collected in different settings: USF Stridor data was recorded in a sound isolation booth, whereas the Bridge2AI data was recorded in a small, quiet room. This is not likely to cause a significant difference in results but may have introduced some false signal into the model due to the imbalance between the datasets (many of the controls came from the B2AI dataset). Finally, while the data and devices were scalable, the environment of the study may not directly replicate real-world conditions like an emergency room, which is likely to introduce additional sources of noise and more acute presentations of the disorders compared to the dataset used in this study. Specific recording devices, such as lapel or lavalier microphones, may be required to capture usable audio data from patients. Given these limitations, the early-stage work presented in this study is not intended to perform the role of a domain expert. Instead, this AI technology is designed to enhance screening and detection tasks in high-volume or low-resource settings, helping patients to ultimately receive necessary care from otolaryngologists.

## Future work

Future work will involve the design of a mobile application for deploying trained AI models on a smartphone/tablet, allowing healthcare professionals to test the viability of these tools in real-world settings. Robust performance in retrospective experimentation does not equate to utility in a clinical environment: implementation studies

are essential to the improvement of the AI methods. Moreover, such an application may facilitate the collection of additional data across multiple types of airway disorders and different healthcare settings, including the emergency room.

Future work is also needed to correlate acoustic features with severity of the disease, including the percentage of stenosis in terms of airway circumference, the vertical length of stenosis, and severity of dyspnea. To achieve these results, acoustic signals must be correlated with measurements on CT scan imaging and video endoscopies/bronchoscopies or measurements collected during procedures (in the OR). A hierarchical implementation of the described AI methods may also add future value as a simultaneous predictor of both the presence and severity of airway stenosis (with stridor as indicator of severity). The work described here may also have potential for testing and deployment in low- and middle- income countries (LMICs), where airway failure is a significant challenge to healthcare providers in terms of technology and expertise<sup>31,32</sup>. Rapid and low-cost AI-assisted support for airway stenosis or stridor detection in this context represents a novel opportunity.

## Conclusion

Airway stenosis, and the stridor which often accompanies moderate to severe cases (or other life-threatening airway problems) is a challenging clinical issue related to misdiagnosis in healthcare. Misdiagnoses are common, which may worsen patient outcomes. Deep learning, which has proven capabilities in capturing nonlinear biomarkers of health, may be a solution for early identification and screening by general healthcare providers. In this study, transformer models were used to demonstrate the potential of clinically viable signal in the task of rapidly screening for airway stenosis and stridor using low-cost data. Future development of this work is essential to the downstream deployment of affordable, non-invasive tools for airway assessments in diverse, high-volume settings.

## Data availability

The data that support the findings of this study are available from the University of South Florida, but restrictions apply to the availability of these data, and so are not publicly available. Data are however available from the corresponding authors upon reasonable request, with raw voice data requiring the execution of a Data Use Agreement (DUA) with the University of South Florida.

Received: 26 October 2024; Accepted: 18 April 2025

Published online: 02 May 2025

## References

- Baughman, R. P. & Loudon, R. G. Stridor: Differentiation from asthma or upper airway Noise1-3. *Am. Rev. Respir. Dis.* **139**, 1407–1409 (1989).
- Cortelli, P. et al. Stridor in multiple system atrophy: consensus statement on diagnosis, prognosis, and treatment. *Neurology* **93**, 630–639 (2019).
- Ng, A. K. et al. Role of upper airway dimensions in snore production: acoustical and perceptual findings. *Ann. Biomed. Eng.* **37**, 1807–1817 (2009).
- Koo, D. L. et al. Acoustic characteristics of stridor in multiple system atrophy. *PLoS ONE* **11**(4), e0153935 (2016).
- Malone, M. et al. Acoustic analysis of infantile stridor: a review. *Med. Biol. Eng. Comput.* **31**, 85–96 (1993).
- Johnson, D. W. Croup. *BMJ Clin. Evid.* **2014**, 0321 (2014).
- Pluijms, W. A., van Mook, W. N., Wittekamp, B. H. & Bergmans, D. C. Postextubation laryngeal edema and stridor resulting in respiratory failure in critically ill adult patients: Updated review. *Crit. Care* **19**(1), 295. <https://doi.org/10.1186/s13054-015-1018-2> (2015).
- Sideris, A. et al. A systematic review and meta-analysis of predictors of airway intervention in adult epiglottitis. *Laryngosc.* **130**(2), 465–473 (2020).
- Living with Idiopathic Subglottic Stenosis. Facebook. Retrieved August, 2024, from <https://www.facebook.com/groups/152394673762>
- Chang, S. S. & Williams, M. H. Relationship of wheezing to the severity of obstruction in asthma. *Arch. Intern. Med.* **143**, 890–892 (1983).
- Van der Velden, W. C. P. et al. Acoustic simulation of a patient's obstructed airway. *Comput. Methods Biomech. Biomed. Eng.* **19**(2), 144–158 (2016).
- Ijaz, A. et al. Towards using cough for respiratory disease diagnosis by leveraging artificial intelligence: A survey. *Inf. Med. Unlocked* **29**, 100832 (2022).
- Eitan, D. N., Nikolaus, E. W. & Patrick, S. Using machine learning for endoscopic detection of low-grade subglottic stenosis: A proof of principle. *Int. J. Otorhinolaryngol. Head Neck Surg.* **171**(4), 1254–1256 (2024).
- Volk, O. et al. Classification of tracheal stenosis with asymmetric misclassification errors from EMG signals using an adaptive cost-sensitive learning method. *Biomed. Signal Process. Control* **85**, 104962 (2023).
- Wang, Y. et al. Deep learning for automatic upper airway obstruction detection by analysis of flow-volume curve. *Respiration* **101**, 841–850 (2022).
- Tekatli, H. et al. Artificial intelligence-assisted quantitative CT analysis of airway changes following SABR for central lung tumors. *Radiother. Oncol.* **198**, 110376 (2024).
- Ahn, J. H. et al. Automatic stridor detection using small training set via patch-wise few-shot learning for diagnosis of multiple system atrophy. *Sci. Rep.* **13**(1), 10899 (2023).
- Dokur, Z. Respiratory sound classification by using an incremental supervised neural network. *Pattern Anal. Appl.* **12**, 309–319 (2009).
- Bridge2AI-Voice Consortium. Flagship Voice Dataset from the Bridge2AI-Voice Project (1.0.0). (2024).
- Awan, S. N., Shaikh, M. A., Desjardins, M., Feinstein, H. & Abbott, K. V. The effect of microphone frequency response on spectral and cepstral measures of voice: An examination of low-cost electret headset microphones. *Am. J. Speech Lang. Pathol.* **31**(2), 959–973. [https://doi.org/10.1044/2021\\_AJSLP-21-00156](https://doi.org/10.1044/2021_AJSLP-21-00156) (2022).
- Awan, S., Bahr, R., Watts, S., Boyer, M. & Bensoussan, Y. Validity of acoustic measures obtained using various recording methods incl. smartphones with and without headset microphones. *J. Speech Lang. Hear. Res.* **67**, 1712–1730 (2024).
- Park DS et al. SpecAugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*. (2019).

23. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**(1), 44–53 (2018).
24. Google. YAMNet: A Pretrained Audio Event Classifier. TensorFlow Hub, <https://tfhub.dev/google/yamnet/1>. Accessed June 2024.
25. Transfer Learning with YAMNet for Environmental Sound Classification. TensorFlow, [https://www.tensorflow.org/tutorials/audio/transfer\\_learning\\_audio](https://www.tensorflow.org/tutorials/audio/transfer_learning_audio). Accessed June 2024.
26. Ashish, V. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1 (2017).
27. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. (2020).
28. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. (2018).
29. Paszke, A. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).
30. Wang, H. et al. Idiopathic subglottic stenosis: factors affecting outcome after single-stage repair. *Ann. Thorac. Surg.* **100**(5), 1804–1811 (2015).
31. Kumar, H. P. Airway management in low resource settings. In *The Airway Manual: Practical Approach to Airway Management* 749–762 (Springer Nature, 2023).
32. Armstrong, L. et al. An international survey of airway management education in 61 countries. *Br. J. Anaesth.* **125**(1), e54–e60 (2020).

## Acknowledgements

Bridge2AI-Voice is funded by the NIH Common Fund through the Bridge2AI program, grant number OT2OD032720. The involvement of JA and BW in this work was also supported by the NIH Center for Interventional Oncology and the Intramural Research Program of the National Institutes of Health, National Cancer Institute, the National Institute of Biomedical Imaging and Bioengineering, via intramural NIH Grants Z1A CL040015 and 1ZIDBC011242, and the NIH Intramural Targeted Anti-COVID-19 (ITAC) Program, funded by the National Institute of Allergy and Infectious Diseases. The participation of HH was made possible through the NIH Medical Research Scholars Program, a public-private partnership supported jointly by the NIH and contributions to the Foundation for the NIH from the Doris Duke Charitable Foundation, Genentech, the American Association for Dental Research, the Colgate-Palmolive Company, and other private donors. DAC was supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Wellcome Trust funded VITAL project (grant 204904/Z/16/Z); the EPSRC (grant EP/W031744/1); and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE).

## Author contributions

J.A. wrote the main manuscript text and designed data processing software, machine learning models, and performed experiments. H.H. contributed to the design of methods/experiments. J.A., R.D., M.B., G.D., and Y.B. designed the data collection protocol. K.N., I.D., S.A., Y.A., S.W., A.G., M.P., and the Bridge2AI Voice Consortium supported and optimized the data collection process. V.D. performed literature reviews related to the topic. B.W., D.C., J.T., and Y.B. supervised the research. All authors reviewed the manuscript.

## Funding

Open access funding provided by the National Institutes of Health

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to J.A.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

## the Bridge2AI Voice Consortium

Olivier Elemento<sup>9</sup>, Olivier Elemento<sup>9</sup>, Anais Rameau<sup>9</sup>, Alexandros Sigaras<sup>9</sup>, Satrajit Ghosh<sup>10</sup>, Maria Powell<sup>8</sup>, Vardit Ravitsky<sup>11</sup>, Jean Christophe Belisle-Pipon<sup>12</sup>, David Dorr<sup>13</sup>, Phillip Payne<sup>14</sup>, Alistair Johnson<sup>15</sup>, Ruth Bahr<sup>16</sup>, Donald Bolser<sup>17</sup>, Frank Rudzicz<sup>18</sup>, Jordan Lerner Ellis<sup>15</sup>, Shaheen Awan<sup>4</sup>, Stephanie Watts<sup>3</sup>, Jennifer Sui<sup>19</sup>, Karim Hanna<sup>16</sup>, Theresa Zesiewicz<sup>16</sup>, Robin Zhao<sup>9</sup>, Lochana Jayachandran<sup>20</sup> & Samantha Salvi-Cruz<sup>8</sup>

<sup>9</sup>Weill Cornell Medicine, New York, NY, USA. <sup>10</sup>Massachusetts Institute of Technology, Boston, MA, USA. <sup>11</sup>University of Montreal, Montreal, QC, Canada. <sup>12</sup>Simon Fraser University, Burnaby, BC, Canada. <sup>13</sup>Oregon Health & Science University, Portland, OR, USA. <sup>14</sup>Washington University in St. Louis, St. Louis, MO, USA. <sup>15</sup>University of Toronto, Toronto, ON, Canada. <sup>16</sup>University of South Florida, Tampa, FL, USA. <sup>17</sup>University of Florida, Gainesville, FL, USA. <sup>18</sup>Dalhousie University, Toronto, ON, Canada. <sup>19</sup>Hospital for Sick Children, Toronto, ON, Canada. <sup>20</sup>Mount Sinai Hospital, Toronto, ON, Canada.