**Project 4 - Stridor Screening**

This is an interactive document to support the project team focused on stridor screening using the stridor dataset and the Bridge2AI dataset (J. Wang, B. Seerapu, V. Dauod).

**Important note:** These dates are tentative, and may change based on the specific needs of your project. AI research is a dynamic and unpredictable process.

**Introduction:**

For this project, the main area of focus will be the optimization of an AI system to improve detection and characterization of stridor, a life-threatening respiratory condition often mis-diagnosed in real-world healthcare settings. Integration of the B2AI dataset as controls for the stridor would be very advantageous for generalizability, but may result in challenges such as significant device imbalances between stridor positive and negative datasets. The B2AI dataset is collected via an Avid headset microphone, while the stridor dataset spans 3 devices.

**Literature Review**

This section contains (1) required reading for team members, (2) keywords for a subsequent literature search, and (3) a table for the input of relevant manuscripts found via literature search.

**Required Reading:**

1. [**“Everyone wants to do the model work, not the data work”**](https://dl.acm.org/doi/abs/10.1145/3411764.3445518?casa_token=9KaD00XUxq4AAAAA:VlsqCe_P1U_pU2Z3D-YnN7zy7q39DueJsL0NBA9m1F6vyAchK6hHRhQG4cadZ5JMkXynO7OnJ2w)
2. [**“Voice as an AI biomarker of health”**](https://jamanetwork.com/journals/jamaotolaryngology/article-abstract/2815136?casa_token=JfOZhigOUNAAAAAA:2-alOLVqKzRbaCL8ok4UGOj6TWeZZPRHF-16ROiKH8LlrOGwAL6eqEoc3lbmYrlhS1Nf1KC4)
3. [**“Assessment and causes of stridor”**](https://drive.google.com/file/d/1j80oPqlNub5JulP2zbP5Wrt2SMKsoxqg/view?usp=sharing)

**Potential Search Terms for Literature Review:**

1. stridor
2. stridor AI
3. Automated Stridor detection
4. Stridor classification
5. AI for respiratory recordings
6. AI for breathing data
7. AI for airway assessment
8. AI for respiratory care
9. AI screening
10. AI in the ER

**Manuscripts:**

| **Citation** | **Keywords** | **Notes** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Datasets:** In this section, information on datasets relevant to this project are presented and described,

including an interactive section for the identification of similar public datasets which can be downloaded is used within the experiments (if applicable).

**Public Data:**

Here, the team should perform a search of publicly available datasets related to studies involving the use of AI for assessing airway conditions, not necessarily limited to stridor.

| **Citation** | **Keywords** | **Notes** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Phase 1: Setup and Data Understanding (Week 1)**

Teams will seek to achieve the following objectives in addition to the literature/data review:

**1a. Team building**

**Not Started**

**Team structure:** teams should identify strategies for effective collaboration when completing the project tasks, to include communication platforms (Slack recommended) and roles.

**1b. Azure Labs**

**Not Started Familiarity with the Azure labs platform** used for the B2AI summer school: teams should understand how to use the system, to include writing/running/testing python code, performing basic tasks (e.g., visualization) with synthetic Bridge2AI data or public datasets, and accessing/using OpenAI large language models (LLMs).

**1c. Github**

**Not Started**   **Creation of a** [**Github**](http://www.github.com) **account** for version control and downstream code access. All project GitHub repositories should be shared with all team members and the following faculty: James Anibal (jamesanibal), Mohamed Ebraheem () , and Isaac Bevers (ibevers). Datasets should not be uploaded to Github.

**1d. Bridge2AI Data Collection Application**

**Not Started**

Teamsshould take some time to complete the app survey at least once, understanding the system for acquiring data and outlining any potential challenges for the user which could impact the quality of data.

**1e. Synthetic Data**

**Not Started**

Isaac Bevers, the Bridge2AI summer school TA, has prepared a small sample data which very closely resembles the actual B2AI dataset, which can be found at the links below:

**Compressed file:**  
<https://usf.box.com/s/djur0v89cx0rztzkstr2ikns9jb1evx0>

**Uncompressed file:**

<https://usf.box.com/s/2orruh093w2gj5gclc3u8ey7cq8phgtz>

**1f. Data Familiarity**

**Not Started**

**Initial Exploration of the Bridge2AI dataset:** On Friday, June 21st, the Bridge2AI dataset will become available for use by the teams. Friday afternoon should be dedicated to initial exploration of the data on Azure labs, including loading in the audio data, visualizing the data, and linking the audio data to correlated metadata using functions which can be found at the Github link [here.](https://github.com/sensein/b2aiprep/tree/alistair/streamlit_dashboard)

Questions to Consider - General:

1. How could this data collection process be improved to benefit
   1. Patients/Participants
   2. Healthcare providers/caretakers

Are there small ways (low regulatory complexity) in which the experience of using the app/collection data could actually provide value in a clinical setting or for a patient using the app at home, and potentially reduce biases within this initial process?

**Phase 2: Data Preparation (Week 2)**

In phase 2 of the stridor project, the team will become familiar with the two datasets and consider the best methods to utilize the B2AI dataset as a control cohort.

**2a. Data Preparation - Stridor**

**Not Started**

The stridor dataset which has been shared with you on box contains the following information:

* Audio data files for each patient (including various different tasks across different devices). The core tasks which we will focus on are forced inhale - mouth open (FIMO), deep breath, and rainbow passage.
* Basic demographic information and clinical context for each patient
* Expert annotations by Dr. Yael Benoussan, which indicates both the presence of stridor and the nature of that stridor as phonatory or non-phonatory. There are also controls from the research team which have no stridor.
  + **Key definition:** Phonatory Stridor - stridor defined by vibration of the vocal folds in a state of due to obstruction at the level of the glottis, resulting in a raspy, high-pitched phonation indicating severe airway obstruction.
  + **Key definition:** Non-Phonatory Stridor - an audible sound indicative of airway obstruction but lacking the raspy, high-pitched component of phonatory stridor which is an indicator of severity.

You should familiarize yourself with this external dataset, which includes listening to the different examples of phonatory/non-phonatory stridor and no stridor, read the data into python, and visualize the different tasks/clinical categories.

**2b. Control Data Preparation - Bridge2AI**

**Not Started**

With the tools provided in the link [here](https://github.com/sensein/b2aiprep/tree/alistair/streamlit_dashboard), prepare the dataset for AI modeling by associating each audio file with the corresponding patient/participant information, including race/ethnicity, socioeconomic status, and other metadata included in the dataset. For this project, teams should identify the different clinical categories associated with the data point, particularly the presence or absence of stridor. If there is no stridor, the presence of another airway disorder indicates that this data point may be a challenging control for the model and is of particular interest in this study. This is a key first step which will make future tasks/processes significantly easier. At this point in the project, the team should identify which audio files from the Bridge2AI dataset will be included as controls in the stridor screening experiments, and why these points will be included **(development of inclusion criteria for controls from the Bridge2AI dataset).**

Question to Consider - 2b:

The key tasks outlined here are FIMO, deep breath, and rainbow passage. However, given that there are some stridor patients who completed the app, which is more comprehensive, and that there are many controls from the app as well, it may be worthwhile to ask the following question: are there other tasks on the app which are similar enough to the core stridor tasks such that inclusion in the training dataset would improve performance of the stridor screening system via transfer of knowledge/enhanced generalization? For example, is there a task which has breathing exercises similar enough to FIMO that the model could conceivably use that information to capture stridor from a slightly different perspective, increasing the likelihood of generalization to downstream users.

**Phase 3: Preprocessing and Initial Methods Testing (Weeks 3-5)**

In this section, steps are outlined for data preparation and the development of initial AI models for results assessment.

**3a. Preprocessing**

**Not Started**

Following the characterization of the data, data should be made AI-ready using common data processing methods, including some combination of the following: noise reduction methods, normalization, conversion to spectrograms (testing on raw audio data may also have value), standardization of length via padding.

**3b. Basic Methods**

**Not Started**

The initial step of the “AI modeling” phase of the project will be a simple binary classification task to answering the following question:

* Is there a detectable signal between audio samples from patients with stridor and those without stridor?

Teams should begin the “AI modeling” section with a “simpler is always better” approach, in which they build the simplest possible pipeline to complete the task. For team #1, that may look something like the following:

1. Create a dataloader using all spectrograms and corresponding labels for the task of stridor screening.
2. Train a lightweight CNN model (e.g.,2 layers) using the Adam optimization function and a cross-entropy loss function. Use early stopping based on the validation loss (~5% of the data).
3. Test the model on ~10% of the data which was excluded from the training dataset.
4. Repeat for multiple randomly selected test sets.

#1 - Question to Consider - Section 3b:

Here might be an ideal time to consider some simple **ablation studies** to understand the factors which affect the performance of your basic model. For example, how do different tasks, different devices, and different sets of control data points in the training set impact performance?

#2 - Question to Consider - Section 3b:

As a related question, how should you structure testing if one test patient has multiple tasks from multiple devices? Some experimentation may be very helpful in this case. Consider how that patient might realistically donate data in a real-world setting.

**3c. Device Bias**

**Not Started**

The 40+ patients in the stridor dataset donated data through a combination of various devices, which included an Avid headset, an iPad microphone, and an At2O35 Audio Technica microphone. The Bridge2AI dataset from the app was collected only with the avid headset, which may cause a false signal, given that the Bridge2AI dataset is also the source of the control cohort (a significant majority). The team should explore this further, determining the following:

1. Can you train an AI model to separate data points based only on the device used to collect the audio recording? (No clinical information should be considered).
2. If so, can you use domain adversarial learning (for example, as described in this [manuscript](https://www.jmlr.org/papers/volume17/15-239/15-239.pdf)), to remove the false signal while still performing robust stridor screening.

Answers to these questions may help facilitate the future integration of the B2AI voice dataset with other datasets generated by other researchers, where these types of minor differences may compromise interoperability.

**Phase 4 (BONUS): Additional Experimentation (Week 3-Hackathon)**

By this point, you will have a general understanding of the performance of your model(s) on the basic project tasks. This section is less structured, because now the goal is to optimize the performance of the method on your task(s), and ensure that all steps have been taken to maximize the generalizability of the methods.

**4a. Optimization/Testing**

**Not Started**

In this section, the key question is “how can the stridor screening model improve?” Some questions to consider:

* Would a pre-trained model be beneficial in this case?
* Is there a way to integrate other datasets to improve generalization/performance?
* Is there a benefit of considering both raw audio and spectrograms? What about tabular acoustic features?
* Is there a role for multimodality? For example, metadata from the participant or images from a laryngoscope.
* Are there other external test sets or ways of testing the methods which would more faithfully simulate real-world conditions?

**Phase 5 (BONUS): Interpretation and LLM Modeling (Week 3 - Hackathon)**

In this final stage of the project, the goal is to use existing large language models (e.g., GPT-4o) to operationalize the progress you have made in the previous stages, helping to ensure that a deployed software system can, in one example, be useful for ED clinicians who may benefit from the stridor screening system.

**5a. Prompt Design**

**Not Started**

Prompt engineering must be performed and validated using a selection of examples from the dataset. Here, the following components may go into your prompt as context:

* Model predictions about stridor
* Participant metadata (other clinical conditions, etc.)
* Extract acoustic features (e.g., amplitude) from the sound and input these into the prompt as well.

The ultimate goal of this prompt engineering task is twofold:

1. **If the user is a clinician who may not have experience dealing with stridor:**
   1. Can the LLM take the prediction, acoustic features, and participant metadata and explain the likely significance of the stridor in this case, potentially identifying specific online sources to improve physician knowledge.
2. **If the user is a patient using the system at home to monitor airway function:**
   1. Can the LLM take the prediction, acoustic features, and participant metadata and explain the likely significance of the stridor to the patient, potentially providing online resources which help guide the patient to act appropriately to manage the condition or seek medical attention.

Potentially helpful resource (OpenAI API): <https://openai.com/index/openai-api/>