Appendix A: Ethical Considerations and Datasets

In this section, we answer the questions available at https://2021.aclweb.org/ethics/Ethics-review-questions/:

- *Does the paper describe the characteristics of the dataset in enough detail for a reader to understand which populations the technology could be expected to work for?*

  In the Introduction section, the paper describes the conditions in which the data were collected, and the limitations regarding the collection in hospitals under a first-wave pandemic situation, in which patients with respiratory insufficiency could only be reached inside isolated COVID-19 wards. This collected data differ from the control group, where data were collected by an application developed for this purpose. To address the characteristics of the noisy environment of the patient data group, we showed that bias can be treated by adding noise samples to both control and patient samples during training. The characteristics and limitations of the dataset is well documented in the paper (Section 3.1), as the method to address this issue (Section 3.

- *Do the claims in the paper match the experimental results, in terms of how far the results can be expected to generalize?*

  The paper goals, as stated, was to demonstrate that speech can be used as a biomarker in the detection of respiratory insufficiency, and we believe that such an original goal was achieved in this work. The experimental results show that there are characteristics in speech that can provide information to detect COVID-19 respiratory insufficiency.

- *Does the paper describe the steps taken to evaluate the quality of the dataset?*

  Section 3.1 presents an analysis concerning spoken utterances, age and sex distribution as well as noise presence in the dataset. A selection was performed by a single researcher in the validation and test sets. Citing the text, "We selected audios with the best signal-noise ratio to use in the test set, and the second best audios were used for validation".

- *Does the paper describe how the technology would be deployed in actual use cases?*

  The paper explains as future work the aim to extend the study to other respiratory illnesses besides COVID-19 and the data collection of both patient and control groups in the same location. Such a collection should be much more feasible in a situation outside the peak of a pandemic and is currently in development. We also have the goal to develop a practical application but this was not the goal of the paper.

- *Does the task carried out by the technology match how it would be deployed?*

  The paper showed 91% of accuracy using test samples without noise samples being added to the test samples. This supports the generalization ability of the model for respiratory insufficiency detection of COVID-19 patients, which can help to address patient triage. Such model could be deployed, for example, embedded in mobile applications in two scenarios. First, as an automated fast triage alternative in highly demand hospital. Second as an cheap alternative for for infrared thermometers were they are not always easy accessible, such as in residences in general.

- *Does the paper address possible harms when the technology is being used as intended and functioning correctly?*

  As our experiment was performed during the pandemic, only COVID-19 patients with respiratory insufficiency conditions were targeted. As a result, we do not known how the system behaves with respiratory insufficiency arising from other causes, such as heart conditions, H1N1, among others.

  Furthermore, the data from patients and control was collected using different methods and a workable system should require data collected from both patients and control in the same acoustic environment. This issue was addressed in the Conclusions Section.

- *Does the paper address possible harms when the technology is being used as intended but gives incorrect results?*

  The main issue with incorrect results is false negatives, in which a patient presenting COVID-19 would not be treated as soon

as desirable. This issue is addressed in the Introduction Section.

- *Does the paper address possible harms following from potential misuse of the technology?*

  We do not foresee a misuse case involving the proposed model.

- *If the system learns from user input once deployed, does the paper describe checks and limitations to the learning process?*

  This question is out of the scope for the proposed method.

- *Does the paper ensure that the harms identified are not likely to fall disproportionately on populations that already experience marginalization or are otherwise vulnerable?*

  This question is out of the scope for the proposed method.

In this section, we also comment on the items requested at https://2021.aclweb.org/calls/reproducibility-checklist/.

- *A clear description of the mathematical setting, algorithm, and/or model*

  This is presented in Section 3.4: Proposed Model.

- *A description of computing infrastructure used*

  It is described in the last paragraph of Section 3.

- *The number of parameters in each model*

  This was detailed in Section 3.4

- *A clear definition of the specific evaluation measure or statistics used to report results.*

  The classical measure accuracy was the main metric used through the paper.

- *For all results involving multiple experiments, such as hyperparameter search, the exact number of training and evaluation runs.*

  It was performed 3 training runs (random seeds) for each experiment, as presented in the paper (Section 3.3). Hyperparameters were adjusted manually.

- *Relevant statistics such as number of examples and label distributions*

  We provide the distribution of ages in Figure 1 and some statistics such as duration, number of people on control/patient sets and number of gender instances in each set of the filtered dataset (Table 1). We also provide the metadata of the dataset.

- *Details of train/validation/test splits*

  A paragraph of Section 3.1 details: "The dataset was divided in training, validation and test, as is usual in statistical learning. We Selected audios with the best signal-noise ratio to use in the test set, and the second best audios were used for validation. The aim of this partitioning is to detect training overfitting. Information of the resulting filtered dataset is presented in Table 1. Dataset metadata is anonymously available".

- *An explanation of any data that were excluded, and all pre-processing steps*

  We explain that some audio samples were excluded from validation and test sets by manually filtering the dataset in Section 3.1. Section 3.2 details the pre-processing steps and Section 3.3 explains how noise samples were added to the original data to prevent bias.

- *For natural language data, the name of the language(s)*

  The title of the paper states the language addressed: Brazilian Portuguese.

- *A link to a downloadable version of the dataset or simulation environment*

  The dataset and the source code were submitted as supplementary material (for review purpose only, as we are still trying to make it publicly available).

  Please do not distribute the dataset as it is still under consideration at the ethical committee for distribution.

- *For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control*

  Section 3.1 details the data collection process.