Cough event detection algorithm validation

insubiq.com
igor@insubiq.com

May 13, 2021

Abstract

In this paper, we describe one of our experiments in technical validation of the algorithm according to our regulatory strategy for clinical validation. We used a balanced dataset of 260 sound samples to validate the algorithm. These sound samples have not been used for AI training.

Respiratory experts listened to each sound sample and labeled them. Their aggregated opinions were marked as the Gold Standard [1]. Next, we compared the algorithm's predictions with the Gold Standard. In this way, we tried to detect the difference between the level of cough recognition by the human ear and the Cough Tracker algorithm. We have got the following results: Gold Standard - 0.985 Accuracy. Not 1.0 because experts gave conflicting opinions in several cases of cough-like sounds. Cough Tracker algorithm - 0.96 Accuracy.

1 Introduction

Cough is a congenital protective unconditional reflex, which is part of the body's immune system. Its function is to clear the airways from particles inhaled or generated in the airways. It may be elicited by chemical or mechanical stimuli arising in the larynx, trachea, carina, or main bronchi (Lamb, AB. Nunn's Applied Physiology, 7th edition. Churchill Livingston-Elsevier eds. 2010).

There are many physiological characteristics of the cough that can be identified by sound: cough can be dry or wet, barking, hacking, loud or soft. To date, however, there is no reliable way to collect this kind of data since the information provided by a patient is subjective, episodic, and could be misrepresented. Healthcare providers have only limited access to the cough episode during a patient visit and have to rely on the patient's observations.

The ubiquity of smartphones and their presence near the user transforms this device into a remote cough detection and monitoring tool. It can be helpful for patients with chronic diseases such as Asthma (330M patients worldwide) [3], [4], COPD (220M patients worldwide) [2], and other respiratory conditions. Such algorithms will will be able to detect deterioration of cough symptoms and, based on the dynamics, predict potential Asthma attacks that could lead to respiratory arrest and death [4].

Our team at Insubiq has developed an Artificial Intelligence-powered algorithm to automatically detect cough sounds from surrounding audio streams via the user's smartphone (Cough Tracker app, iOS and Android version). The app is highly sensitive and capable of selecting sounds specific only to cough and filtering out speech and background sounds, which also helps protect user privacy. To make sure that the algorithm can work in a 24/7 environment with all the noise that may be present in everyday life, we have validated the algorithm.

2 Objectives

We set the following objectivies for the experiment:

1. Identify the human ear level of cough detection.
2. Identify the accuracy of a cough detecting algorithm compared with the human ear.
3. Validate the accuracy of cough detection by the Cough Tracker app algorithm in real-life settings environment.

3 Data collection

We used the balanced dataset for this experiment. It consisted of 130 cough sounds and 130 non-cough

sounds. This dataset was not used to train the algorithm but only for validation. Usage of a small dataset allowed us to shortly perform basic technical validation of the algorithm to select the validation strategy. We mixed sounds collected by the Cough Tracker app in experiments studies and sounds collected from open sources.

Cough sounds were collected through the participants of the experiment with different smartphone models (Table 3 [1]). All the devices in the population were based on Android 6 (or higher).

Phone Model	Count
ASUS-ZB602KL	7
HUAWEI-JAT	2
HUAWEI-SNE	1
LENOVO-P1A42	1
OPPO-CPH1931	1
SAMSUNG-SM	41
SONY-G3412	3
SONY-H4311	8
XIAOMI-MI	5
XIAOMI-REDMI	30

Table 1: Smartphone models and their counts that were used in the experiments

4 Experiment design

We have recruited three independent respiratory experts with various backgrounds in assessing coughing respiratory patients from different countries. They had to mark up a prepared dataset with randomized and blinded audio files containing cough and non-cough sounds. Below is the list of experts.

- Expert 1: University Hospital, Head of Pulmonary lab,
- Expert 2: Medical center for COVID patients, General physician
- Expert 3: University Hospital, Oncologist

These experts were provided with a file (xls) with links to sounds. The experts had to listen to each audio sample independently of each other and label these sounds by the category: Cough, No-cough, or Unknown. Reviewers were allowed to listen to every sound as many times as they considered necessary to make a decision. Next, text labels were replaced with numerical values in the following way: Cough-1, Non-cough-0, Unknown-0.5.

The averaged results of the experts' labels were compared with the threshold of 0.5. If the average opinion of the experts was higher than 0.5 (the majority determined that the sound was a cough), then the sound was considered a Cough, otherwise No-cough. Next, these results were compared with each other and with the Cough Tracker algorithm predictions.

5 Experiment Result

We track the threshold-dependent and threshold-independent metrics. To estimate the algorithm by threshold-dependent metrics, necessary to convert the probability predicted by the algorithm into a class label (Cough or No-cough). The decision for converting a predicted probability or scoring into a class label is defined by a parameter referred to as the 'decision threshold', 'discrimination threshold', or simply the 'threshold'. The default value for the threshold is 0.5 for normalized predicted probabilities or scores in the range between 0 or 1. In this validation study, we used a standard threshold of 0.5.

Each of the 260 sounds was predicted by the algorithm. For each sound, the algorithm predicted the probability of coughing. Then the algorithm's predictions are compared with the threshold of 0.5. If the possibility is higher than 0.5, it is considered as a Cough, contrarily No-cough. Next, the predicted labels were compared with the Gold Standard.

For threshold-dependent metrics, we are tracking Accuracy and F1. Table 2 [2] shows a detailed classification report which we obtained after comparing algorithms prediction and the Gold Standard.

	Precision	Recall	F1	Accuracy
Cough	1	0.91	0.96	0.96
No cough	0.92	1.0	0.96	0.96
Average	0.96	0.96	0.96	0.96

Table 2: Classification report

Precision(1) quantifies the number of positive class predictions that actually belong to the positive class. In simple words, what proportion of the cough sounds detected by the model are really cough sounds.

Recall(2) quantifies the number of positive class predictions made out of all positive examples in the dataset. In simple words, what proportion of cough sounds did the model detect out of all cough sounds in the dataset.

F1(3) - The F1 score is the harmonic mean of the precision and recall. The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero.

Accuracy(4) - Accuracy is the fraction of correctly predicted samples out of all the samples.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

$$Accuracy = \frac{TP * TN}{TP + TN + FP + FN} \tag{4}$$

F1 measure has an intuitive meaning. It tells how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). With high precision but low recall, the classifier is extremely accurate, but it misses a significant number of instances that are difficult to classify. This is not very useful. The math behind the F1 metric is quite simple, but first, need to define the following terms:

- Cough - Positive class.
- No cough - Negative class.
- True Positive(TP) - Sounds which actually cough and predicted as cough
- True Negative(TN) - Sounds which actually not-cough and predicted not-cough
- False Positive(FP) - Sounds which actually not-cough and predicted as cough
- False Negative(FN) - Sounds which actually cough and predicted as not-cough

According to the Classification Report in Table 1 [2], the algorithm has 0.96 Accuracy and 0.96 F1. The Confusion matrix (Figure 1 [1]) shows only 11 False Negative errors and no False Positive errors. It means that if the algorithm makes a mistake, it will be more likely a False Negative error than a False Positive.

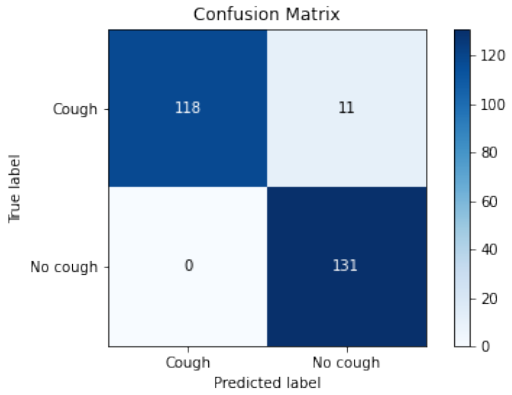

Figure 1: The Confusion Matrix is a summary of prediction results on a classification. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

The clinical meaning of these findings is that an app based on the algorithm is a reliable tool for cough counting, as it filters out environmental noises. If we got a low accuracy it could lead to an overestimation of the number of coughs and cause false alarms, harming users' confidence in the app. One of the disadvantages of these metrics is that there is a need to set the threshold manually by the algorithm developer. Setting the threshold manually is usually based on prior knowledge of the distribution of real data or the algorithm's error cost.

For many tasks, the misclassification costs are unknown or variable. In that case, the overall accuracy is often fairly meaningless, and the AUROC is a better indicator of performance. That is why we also track threshold-independent metrics such as AUC-ROC. In this method, there is no need to compare the probabilities predicted by the algorithm with the threshold, AUC-ROC is based on probabilities.

We evaluated the algorithm's predictions using the AUC-ROC metric. Figure 2 [2] shows a chart of the ROC curve and the area under the ROC curve. We got the AUC-ROC metric at 0.99.

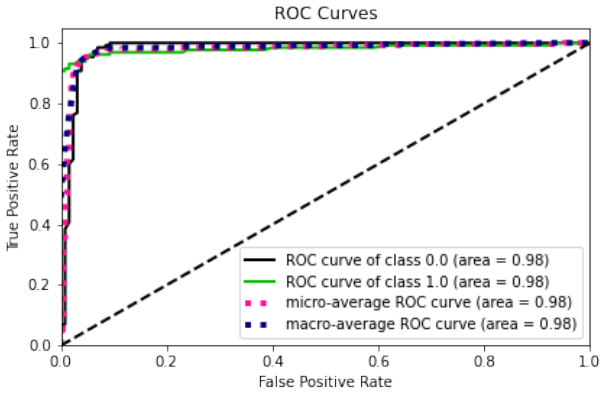

Figure 2: AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC (Area Under The Curve) represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

The Higher the AUC, the better the model is at distinguishing between cough and no cough sounds. In a ROC curve, a higher X-axis value indicates a higher number of False positives than True negatives. While a higher Y-axis value indicates a higher number of True positives than False negatives. So, the choice of the threshold depends on balancing between False positives and False negatives. The AUC at 0.99 means that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative ex-

ample. It measures the classifiers skill in ranking a set of patterns according to the degree to which they belong to the positive class, but without actually assigning patterns to classes.

In addition, we found that some sound samples were marked differently by experts. We found 12 sounds in which one of the experts disagreed with the other two experts. We interpret this as an error. On average, each expert disagrees with the others in 3 sounds so every expert has 1.5% of errors it means the average experts accuracy is 98.5%. Probably that this discrepancy will always occur as a result of the effects of background sounds and noises. We were unable to find any research on the accuracy of cough recognition in noise environments by the human ear, and we assume that this average accuracy should be at a level of 0.98±0.1.

The expert accuracy allows us to estimate cough sounds recognition in a noisy environment by humans and determine Human-Level Performance (HLP) in this task. So now we have a quality metric that we aim for.

We measured the level of agreement of the experts through the kappa coefficient [5] and obtained a very high agreement between reviewers when defining the presence or absence of cough in a sound recording. Table 3 [3] shows the criterion of agreement of experts between each other and their average value at the level of 0.94. We also calculated the 95% confidence interval of the kappa coefficient and obtained a range from 0.90 to 0.98.

Kappa coefficient [5] measures the degree of agreement between a pair of experts. The values of range lie in [-1, 1] with 1 presenting complete agreement and 0 meaning no agreement or independence. A negative statistic implies that the agreement is worse than random.

Expert accuracy is not the same as kappa coefficient. The kappa coefficient shows the probability that two experts were agree when determining the presence or absence of cough in a sound recording. The expert accuracy is the fraction of sounds in which at least two experts were agree when determining the presence or absence of cough in a sound recording.

	E1 vs. E2	E2 vs. E3	E1 vs. E3	Average
N° of agreements	255	251	250	252
% of agreement	98.08	96.54	96.15	96.9
N° of agreements by chance	127.5	127	126.6	127
% of agreements by chance	49.05	48.86	48.68	48.86
Kappa	0.962	0.932	0.925	0.94
SE kappa	0.016	0.022	0.023	0.02
95% CI	0.93-0.994	0.89-0.975	0.881-0.969	0.90 - 0.98
Qualification	Very good	Very good	Very good	Very good

Table 3: Agreement between Experts on validate dataset.
E1, E2 and E3 are experts 1, 2 and 3 respectively

6 Errors Analysis

We manually listened to all the sounds where the algorithm made a mistake and tried to understand the error cause or find any patterns in the errors. After analyzing the sounds on which the algorithm was wrong, we found the following.

Some of the sounds were similar to the sound of sneezing (two sound samples). Sometimes it is hard to distinguish a cough from a sneeze, even for a human. These specific sounds the algorithm predicted as Cough with an average probability of 0.35. A prediction of 0.35 may mean that the model still tends to think that there is a Cough, but not enough. The algorithm made an error on the one cough sound, which the experts commented as Barking cough. The model predicted this cough sound with a probability of 0.25. This type of cough has a specific pattern unusual for a usual cough. These two cases reduced the Accuracy level down by 0.8.

In general, there seems to be a problem with non-standard cough patterns. Either more such coughs are needed, or we need to learn how to simulate them in the training process. Another reason for the errors is that the sound of cough may not be completely on the recorded sound segment. For example, part of the cough may be on the edge of the recorded sound. Such a problem can be fixed by simulation in training.

We also found that it is essential not only to detect coughs. It is crucial to train the algorithm to distinguish the cough from the extraneous noise of real life, and the dataset for this training must be much larger than the dataset with coughs itself.

7 Conclusion and Further Research.

The difference between the Gold Standard (0.985) and Algorithm Accuracy (0.96) was 2%. Additional investigations are needed to assess the significance of

this difference and its impact on clinical use. However, our preliminary opinion is that this difference is not significant compared to the benefits that this technology can deliver to chronic patients.

We believe that by iteratively increasing the size of the training dataset by manually labeling and comparing Gold Standard and algorithm prediction, we would find a plateau level of the algorithm accuracy. According to this idea, we develop a strategy to get closer to human-level performance (HLP):

1. Collect new dataset for labeling by experts
2. Define experts labels as the Gold standard
3. Validate the algorithm on the collected dataset
4. Calculate the gap between algorithm accuracy and the Gold standard
5. If we have not reached a plateau
   (a) Add the Gold standard dataset to the training dataset
   (b) Retrain the algorithm
   (c) Go to 1

Reaching this plateau means that adding data no longer decreases the difference between algorithm accuracy and Gold standard. Also, it means we need to look for other ways to improve quality. One such way could be to change the architecture of the model or the sound processing methods, but changing these characteristics can also affect the speed of the app work. There is always a trade-off between algorithm accuracy and execution performance. After finding the plateau, we can more effectively find the compromise between speed and accuracy.

Based on the validation results, we can assume that the algorithm can correctly predict cough sounds, even in the presence of extraneous noises in the environment. The quality level of the algorithm allows it to use it for cough monitoring 24/7. The use of an artificial intelligence-powered algorithm is instrumental for this purpose since a subject can't count his coughs manually. An independent observer could count a subject's coughs, but it is impractical since it needs the observer to be present 24 hours a day. A validated algorithm can decide which of those sounds must be analyzed in detail and define the presence or absence of a cough. Most importantly, the present analysis shows that this process can be conducted automatically, in real-time, and without compromising the subject's privacy. Cough detection and counting could be essential tools in the follow-up of patients with chronic respiratory diseases.

References

[1] Jefferson Rosa Cardoso et al. "What is gold standard and what is ground truth?" In: Dental press journal of orthodontics 19.5 (2014), pp. 27–30.

[2] Chronic obstructive pulmonary disease (COPD). https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

[3] Shyamali C Dharmage, Jennifer L Perret, and Adnan Custovic. "Epidemiology of asthma in children and adults". In: Frontiers in pediatrics 7 (2019), p. 246.

[4] Oladunni Enilari and Sumita Sinha. "The global impact of asthma in adult populations". In: Annals of global health 85.1 (2019).

[5] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data". In: biometrics (1977), pp. 159–174.