Computer Methods and Programs in Biomedicine 241 (2023) 107743

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

A semi-supervised algorithm for improving the consistency of crowdsourced datasets: The COVID-19 case study on respiratory disorder classification

Lara Orlandic [a,*], Tomas Teijeiro [a,b], David Atienza [a]

[a] Embedded Systems Laboratory (ESL), EPFL, Lausanne, Switzerland
[b] Department of Mathematics, University of the Basque Country (UPV/EHU), Spain

ARTICLE INFO

Keywords:
Semi-supervised learning
Audio signal classification
Automatic respiratory disorder diagnosis
Machine learning
COVID-19

ABSTRACT

Background and Objective: Cough audio signal classification is a potentially useful tool in screening for respiratory disorders, such as COVID-19. Since it is dangerous to collect data from patients with contagious diseases, many research teams have turned to crowdsourcing to quickly gather cough sound data. The COUGHVID dataset enlisted expert physicians to annotate and diagnose the underlying diseases present in a limited number of recordings. However, this approach suffers from potential cough mislabeling, as well as disagreement between experts.

Methods: In this work, we use a semi-supervised learning (SSL) approach – based on audio signal processing tools and interpretable machine learning models – to improve the labeling consistency of the COUGHVID dataset for 1) COVID-19 versus healthy cough sound classification 2) distinguishing wet from dry coughs, and 3) assessing cough severity. First, we leverage SSL expert knowledge aggregation techniques to overcome the labeling inconsistencies and label sparsity in the dataset. Next, our SSL approach is used to identify a subsample of re-labeled COUGHVID audio samples that can be used to train or augment future cough classifiers.

Results: The consistency of the re-labeled COVID-19 and healthy data is demonstrated in that it exhibits a high degree of inter-class feature separability: 3x higher than that of the user-labeled data. Similarly, the SSL method increases this separability by 11.3x for cough type and 5.1x for severity classifications. Furthermore, the spectral differences in the user-labeled audio segments are amplified in the re-labeled data, resulting in significantly different power spectral densities between healthy and COVID-19 coughs in the 1-1.5 kHz range ($p = 1.2 \times 10^{-64}$), which demonstrates both the increased consistency of the new dataset and its explainability from an acoustic perspective. Finally, we demonstrate how the re-labeled dataset can be used to train a COVID-19 classifier, achieving an AUC of 0.797.

Conclusions: We propose a SSL expert knowledge aggregation technique for the field of cough sound classification for the first time, and demonstrate how it can be used to combine the medical knowledge of multiple experts in an explainable fashion, thus providing abundant, consistent data for cough classification tasks.

1. Introduction

Audio signal processing and Machine Learning (ML) can be used to automatically screen for various respiratory pathologies, thus enabling ubiquitous diagnosis of these disorders in resource-limited settings [1]. At the onset of the COVID-19 pandemic, several research teams investigated whether the disease could be detected easily and noninvasively using ML to classify between infected and healthy cough sounds [2–4]. Several teams gathered extensive datasets of crowdsourced COVID-19 cough sounds, many of which were open-sourced to enable rapid model prototyping [2,3]. As a result, researchers have developed ML models for COVID-19 cough classification using diverse methods such as temporal decision trees [5], multi-criteria decision making [6], and transfer learning [7]. These teams report promising results, such as an Area Under the Receiver Operator Characteristics Curve (AUC) of 0.95 on unseen testing data [6], as well as a functioning cough diagnosis mobile application [8]. Despite these promising results, many of these works share the shortcoming that they are trained on crowdsourced samples,

* Corresponding author.
E-mail address: lara.orlandic@epfl.ch (L. Orlandic).

https://doi.org/10.1016/j.cmpb.2023.107743
Received 16 November 2022; Received in revised form 12 July 2023; Accepted 2 August 2023
Available online 9 August 2023

which are recorded in uncontrolled environments and may contain mislabeled data. Moreover, Xiong et al. reported that ML architectures trained with PCR-confirmed cough data were more accurate, required less training data, and employed fewer and more stable features than those trained with crowdsourced cough recordings [9]. However, since the disease is a highly contagious airborne pathogen [10], collecting PCR-confirmed cough sound data from COVID-19 positive individuals requires significant effort and sanitary precautions to ensure the safety of those involved.

In order to enhance the labeling validity of the crowdsourced recordings, the COUGHVID dataset enlisted four expert physicians to listen to a number of cough recordings and diagnose any audible respiratory disorders (ex. COVID-19, upper and lower respiratory infections), as well as characteristics of the cough including its type and severity [2]. However, the general trend was that the four experts did not agree on the COVID-19 diagnosis nor the other cough attributes. Disagreement between physicians is common in the medical field; a study of medical referrals noted that only 12% of final diagnoses agreed with the initial diagnoses, and 21% of final diagnoses significantly differed from the initial ones [11]. Therefore, extra care must be taken to overcome the label ambiguity of any crowdsourced cough audio databases, as the expert disagreement and user mislabeling can lead to erroneous classification.

The winners of the 2017 PhysioNet/CinC Challenge on ECG signal classification observed that expert annotation inconsistencies in physiological data can be alleviated through manual re-labeling, thus leading to significant improvements in classifier performance on unseen data [12]. However, manual re-labeling of cough sounds is difficult to perform without medical training. Furthermore, manually re-labeling such an extensive dataset would require significant time and effort. Semi-supervised learning (SSL) is a ML paradigm that can be used to automate the re-labeling process of biomedical signals [13,14]. While SSL is most often used in conjunction with Deep Learning models for medical inference [13,15,16], it can also be used with classical ML approaches in which the extracted features leverage domain knowledge to shed light on the inner-workings of the classifier.

While most cough classification algorithms focus on fully supervised ML approaches [3,4], several groups have leveraged SSL [17,18]. In this paradigm, unlabeled data is exploited in augmenting the dataset to enhance the performance of the classifier, thus providing ample training data and overcoming the issue of label sparsity [19,20]. Furthermore, Han et al. found that incorporating SSL into sound classification enabled a reduction of 52.2% in human annotations necessary to achieve comparable results to fully-supervised methods [19].

In addition to overcoming label scarcity, SSL approaches have also proven successful in alleviating the burden of inconsistent, ambiguous, and erroneous labels on ML classification tasks [21]. Considering the example of 3D image segmentation tasks, semi-supervised models have been shown to outperform fully supervised ones both in the presence of human mislabeling and added random noise [22]. Furthermore, SSL has been widely utilized in speech emotion recognition, a field that suffers from sparse, inconsistent labeling by multiple untrained annotators [13,14]. In particular, Zhu et al. devised an iterative, semi-supervised scheme using the ambiguous emotion annotations of six to twelve annotators and concluded that sufficient training data and moderately reliable labels at the onset of training can significantly improve the classification performance with respect to fully supervised training [13].

Recent works leveraging SSL to overcome inconsistencies in expert physicians' labels utilize an approach in which each expert is modeled by a Deep Neural Network, and then the outputs of these expert models are combined to generate a final label for each sample [16,23]. For example, Li et al. applied this approach to electronic medical record entity recognition by training five distinct models, expanding them to the whole dataset by generating pseudo-labels with each model, and then using a majority voting algorithm to generate the final labels [23]. Furthermore, Guan et al. used individual expert modeling for diabetic retinopathy classification [16], and found that this approach outperformed the Expectation-Maximization algorithm traditionally used for weighing the accuracies of multiple raters [24]. This promising SSL expert modeling technique has not yet been applied to the field of cough audio signal classification.

In this work, we utilize a state-of-the-art SSL technique based on explainable ML models that integrates knowledge of a variable number of human annotators. Unlike other works that solely rely on noisy crowdsourced labels to generate SSL models to distinguish COVID-19 from healthy cough sounds [17,18], we combine the user labels with expert medical labeling schemes to generate more reliable and consistent pseudo-labels of COVID-19 status, cough type, and cough severity. Similarly to [13], we analyze the trade-offs between label consistency and training data size when selecting the final SSL approach. As opposed to previous works that rely on Deep Learning [13,14,16,18,23], which may be difficult to interpret and therefore not amenable to sensitive medical classification tasks, we rely on classical ML algorithms using state-of-the-art audio feature computation.

This technique uses the cough sound recordings of the COUGHVID dataset that were labeled by expert physicians to train three classifiers, where each one models the medical knowledge of a different expert. Next, we overcome the issue of expert label scarcity by generating pseudo-labels on the entire database using each expert model. Then, the outcomes of these models were compared alongside crowdsourced user labels to identify a subset of cough recordings with the highest probability of originating from either COVID-19 positive or healthy individuals. Thus, we overcome the issues of crowdsourced data mislabeling and expert label inconsistency by identifying a high-quality subsample of datapoints – with a threefold increase in feature separability compared to the user-labeled data, as well as a more significant difference in the power spectral densities of the two cough classes ($p = 1.2 \times 10^{-64}$) – which can be used to train future cough classifiers. Furthermore, we determine the importance of the various features to the classification outcome of the SSL approach versus the user or expert label based models to assess the similarities and differences between the approaches. Finally, we use our SSL expert modeling approach to train classifiers for wet vs dry cough and cough severity classification, both of which are pathology-independent and can thus be used in a variety of cough diagnostic tasks beyond COVID-19.

The subsample of cough audio recordings identified through our SSL approach was subsequently made available to the public for further classifier development and ML exploration. To assess the intra-class consistency of this data, we quantify the class separability of standard audio features extracted from COVID-19 versus healthy coughs in the SSL labeling scheme compared to that of the expert labels and crowdsourcing labels of the COUGHVID dataset. Finally, we demonstrate how this data can be used to train cough audio signal classifiers by training final classifiers for COVID-19 screening, cough type detection, and cough severity analysis and comparing their classification accuracies to those of the fully supervised models. As a result, this work aims to provide an automated approach for increasing the labeling quality of biosignal datasets, which can be applied to many other pathologies.

2. Methods

2.1. COVID-19 classification methodology overview

One of the challenges of performing COVID-19 classification based on user-labeled and expert-annotated data is label ambiguity. Since the COUGHVID dataset is crowdsourced, it cannot be known with absolute certainty if the cough recordings labeled as COVID-19 or healthy truly originated from people with the condition or lack thereof. Furthermore, the experts' cough diagnoses exhibited a Fleiss' Kappa score of 0.07 [25], meaning that there was only a slight agreement between the four experts about the cough diagnoses [2].

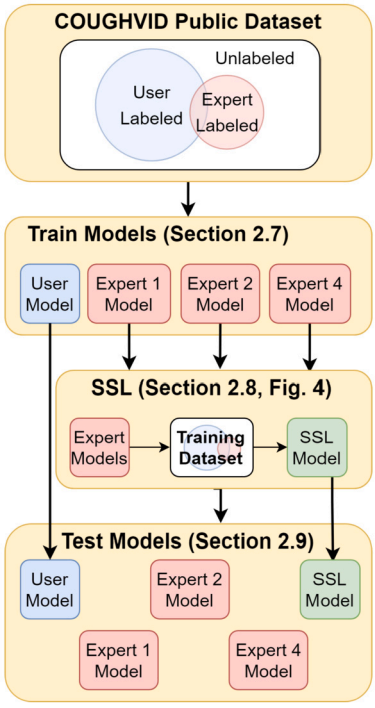


Fig. 1. An illustration of the COVID-19 screening model development methodology, showing the different subsets of the COUGHVID dataset used at each stage. The supervised training, semi-supervised learning, and testing procedures are described in Sections 2.7, 2.8, and 2.9, respectively.

Table 1
COUGHVID public dataset label counts.

Label origin	COVID-19 status		Cough type		Cough severity	
	Healthy	COVID-19	Dry	Wet	Mild	Severe
Users	15,476	1,315	N/A	N/A	N/A	N/A
Expert 1	259	279	425	72	521	34
Expert 2	67	285	600	133	572	59
Expert 3	199	1	358	291	447	105
Expert 4	221	84	654	120	472	69

As shown in Fig. 1, we assessed the label consistency in each of the COVID-vs-healthy classification schemes provided by the dataset (i.e., users, experts) by extracting audio signal features and training ML models based on each set of labels. Then, the semi-supervised learning (SSL) approach was employed to produce a final classifier. Therefore, the following ML models were developed and compared in terms of various classification accuracy metrics on their respective labeling schemes:

1. *User Crowdsourcing Model:* In this classifier, the recordings were self-labeled as "COVID-19" or "healthy" by the users who uploaded the crowdsourced recordings.
2. *Expert [1,2,4] Model:* Three separate models were developed, corresponding to the labels of Experts 1, 2, and 4. Since we see from Table 1 that Expert 3 only labeled one recording as COVID-19, this is not enough information for a ML model to reliably perform generalization. Therefore, this expert's labels are omitted from consideration in further analysis. The positive class was made up of recordings labeled by each expert as "COVID-19", and the corresponding negative class was labeled as "healthy_cough".
3. *SSL Model:* In order to combine the knowledge from both the users and the experts into one model, semi-supervised learning was used. In this approach, the expert models and user labels were used to filter the dataset and determine the subset of coughs with the highest probability of being COVID-19 positive and healthy. The details of the implementation are described in Section 2.8.

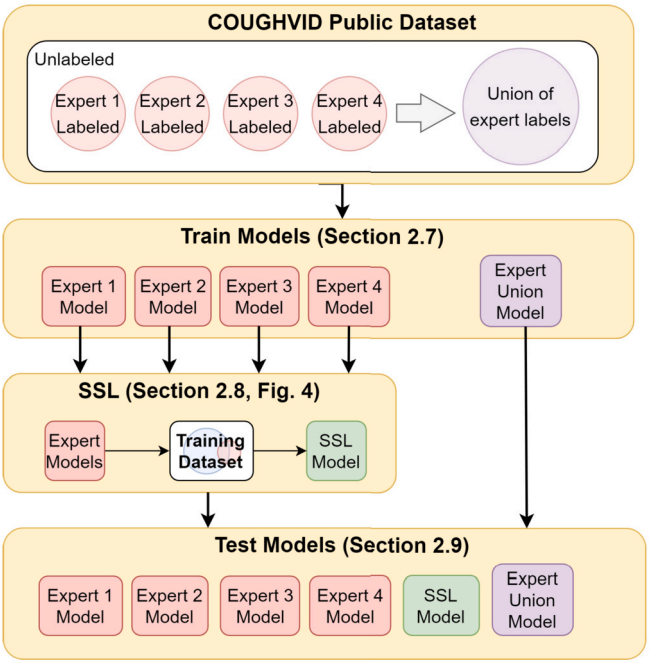


Fig. 2. An illustration of the model development methodology for the cough type and cough severity classifiers.

2.2. Cough type and severity classification methodology overview

The medical experts who annotated the COUGHVID dataset provided two additional labels that are important parameters of the cough: its type and severity. Cough type refers to whether the cough is wet, meaning that mucus is expelled from the lungs, versus a dry cough that does not produce mucus. The severity of the cough indicates the degree of obstruction the cough causes to the subject's normal respiratory function. Unlike the COVID-19 status, the cough type and severity are not known by the users and therefore are only labeled by the experts.

The model development overview is visualized in Fig. 2. First, a baseline model is obtained by combining the expert knowledge in a simple, supervised fashion: we take the union of all recordings labeled by at least one expert as belonging to a given class (ex. wet or dry), and the recordings in both the positive and negative classes are discarded. These labels are then used to train and test a baseline ML model. Then, similarly to Fig. 1, each expert's labels are used to train an expert model, whose labels are then expanded on all unlabeled data to generate an SSL model for the classification task. Unlike the COVID-19 classification, though, the cough type and severity classifiers include the labels of Expert 3, which were much more balanced for these tasks. Finally, each expert's model, the baseline expert union model, and the final SSL model are evaluated on a private test set.

2.3. Dataset description

This analysis uses the COUGHVID crowdsourcing dataset, which is a vast repository of cough audio samples originating from diverse participants located across the globe [2]. The dataset is made up of user-uploaded cough recordings, many of which contain a status label indicating whether the user claimed to be diagnosed with COVID-19, exhibiting symptoms, or healthy at the time of recording. As an additional validation step, four expert physicians each labeled 1,000 cough recordings to diagnose potential respiratory disorders (i.e., COVID-19, upper respiratory infection) that are audible in the recordings, as well as any audible respiratory malfunctions, the cough type, and severity of the cough. Each expert reported spending approximately 10 hours to label the cough sounds, which exemplifies the significant time and effort human labeling takes for such a task.

An expanded version of the training dataset was used, containing recordings uploaded from April 2020 to October 2021. There are about 34,500 recordings in this dataset, 20,644 of which contain user status labels. Both the expert labels and the testing dataset described in [2] are unchanged in this work. Table 1 displays the value counts of the COVID-19 status, cough type, and cough severity as labeled by the users and expert physicians. Although the dataset contains more ambiguous labels, such as the "symptomatic" user label, or the "unknown" cough type expert label, these recordings are not used in the model development and each classification task is considered binary. It should be noted that the coughs labeled by the users and experts are not mutually exclusive, and 150 coughs were annotated by all experts to assess the level of agreement between the physicians.

2.4. Cough audio signal pre-processing

Since the COUGHVID dataset contains some recordings that do not capture cough audio, the cough classifier developed in [2] was used to remove non-cough recordings from consideration. Only recordings with a cough classifier output greater than 0.8 were used in this work.

As an initial pre-processing step, all of the cough recordings were normalized to their maximum absolute value such that the signal values range from -1 to 1. This enables a fair comparison of the RMS power of different signal segments, ensuring that the feature provides meaningful cough amplitude information that is not biased by inherent amplitude differences between recordings due to the subject's proximity to the microphone or differences in recording hardware. Furthermore, this normalization provides numerical stability for subsequent feature computation. Next, a 4th order Butterworth lowpass filter with a cutoff frequency of 6 kHz was applied. Consequently, the recordings were downsampled to 12 kHz. This filtering was performed to reduce high-frequency noise and increase the computational efficiency of all further signal processing and feature extraction algorithms. The cutoff frequency was chosen because visual analysis of the cough signal spectra revealed that most of the signal power lies below 6 kHz. Furthermore, past cough sound classification algorithms used cutoff frequencies ranging from 4 Hz to 8 Hz [26,27], so an intermediate value was chosen.

2.5. Cough segmentation

Once the recordings were pre-processed, a custom cough segmentation algorithm was employed to isolate each individual cough event present in a given recording. The segmentation algorithm exploits cough physiology to divide each recording into its constituent cough sounds. This algorithm enables feature extraction on each cough, thus suppressing silence and extraneous low-amplitude sounds like breathing. Furthermore, the algorithm can be used to perform a simple Signal-to-Noise Ratio (SNR) calculation, as well as aggregation of the ML classifier labels of all coughs originating from the same recording.

The algorithm is depicted in Fig. 3 on a recording of a breath, two coughs, another breath, and two more coughs. First, the signal is squared to compute its power. Next, a hysteresis comparator is applied to extract the sudden bursts in sound amplitude that arise from coughing. This means that potential cough candidates are determined to be regions started by the signal exceeding the upper threshold and ended by the signal going below the lower threshold. A tolerance of 10 ms is applied to the thresholds, meaning that the signal should either exceed the upper threshold or go below the lower threshold for at least 10 ms for a cough onset or offset to be recorded. The lower and upper hysteresis thresholds were set to 0.1 and 2 times the RMS signal power, respectively. These multipliers were empirically determined through analysis of a variety of cough audio signals.

Next, the cough segments were analyzed and discarded based on the physiological limitations of cough duration. The cough is composed of three segments: inspiration, compression, and expiration. The latter two have known timing constraints: the compressive phase, during which inhaled air is compressed in the lungs to increase lung pressure, typically lasts 200 ms [28]. The expiratory phase is initiated by a brief opening of the glottis (30-50 ms), causing the loudest phase of the cough sound and rapid airflow, followed by 200-500 ms of lower respiratory airflow [28]. Therefore, the minimum possible cough sound length is approximately 230 ms. We consequently discard any cough sound candidates shorter than 200 ms, and we include the 200 ms before and after the cough candidate in each segmented cough to capture any low-amplitude noise caused during the compressive and expiratory phases.

The cough segmentation algorithm was subsequently used to eliminate cough recordings with significant background noise. An estimate of the SNR was calculated for each signal as described in [2] by comparing the RMS signal power of the cough segments of a recording to that of the non-cough segments. The training and testing datasets were further filtered by retaining only the recordings with a SNR greater than 5, above which cough sounds were clearly more prominent than background noise. The cough segmentation and SNR estimation algorithms, as well as all feature extraction functions needed to reproduce the feature vectors of each cough sample are available in the COUGHVID public git repository[1] to foster reproducibility.

2.6. Feature extraction

The set of 60 features extracted from each cough audio segment are the same as those computed in the cough classification algorithm in [2]. These features are a mixture of time, frequency, and Mel frequency domain computations that were chosen due to their previous implementation in automatic cough sound classification ML algorithms [26,27]. These features provide both general information about the signal spectra, as well as detailed computations regarding specific frequency bands, thus allowing the feature elimination step to select only the relevant features to each classification task. The code used for feature extraction is available in the COUGHVID repository [2].

For each classification task described in Section 2.7, an additional set of Power Spectral Density (PSD) features were selected by inspecting the averaged PSD of each class in the training dataset, divided by the total average signal power such that the PSD curve is normalized to a unit area. The frequency bands displaying a large variation between the average normalized PSDs of the two classes were noted, and the bandpowers within these frequency ranges were added as features. This produced a variable number of PSD features for each classifier.

Some of the user-labeled data in the COUGHVID dataset contains user metadata information, such as their reported age, gender, and presence of respiratory disorders. In order to provide the model with some user-specific information to assist in classification, the binary gender value was added to the feature set. In case no gender information was provided, a gender identification model was developed using the ML model optimization procedure described in Section 2.7. The model was trained using the training data subset containing gender labels and resulted in a classifier with an AUC of 0.8 on the testing dataset described in Section 2.9. This classifier was used to assign a gender of "male" or "female" to any cough in the dataset for which this was not provided.

2.7. Model comparison and optimization

For each classification task described in Sections 2.1 and 2.2, a ML model was trained to distinguish COVID-19 from healthy coughs, wet versus dry coughs, and severe versus mild coughs based on the feature vectors of the training dataset. Prior to optimization, the features were standardized by removing the mean and scaling to unit variance.

As shown in Table 1, there is a significant class imbalance in each of the classification tasks. This issue was addressed using the Synthetic Minority Over-Sampling Technique (SMOTE) [29], which was employed

[1] https://c4science.ch/diffusion/10770/.

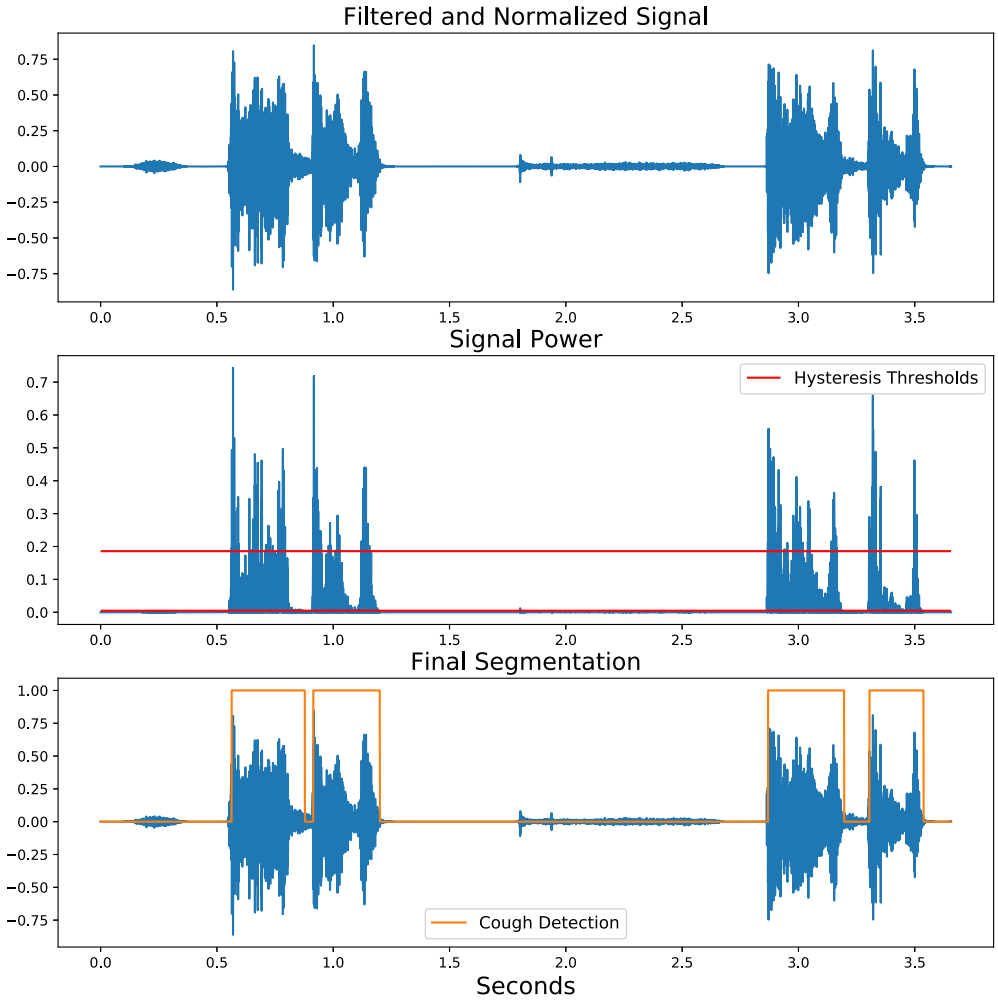


Fig. 3. A step-by-step illustration of the cough segmentation procedure. A hysteresis comparator was applied to the signal power to detect the sound bursts, which enables us to consistently discard the segments without relevant inputs for the subsequent ML process.

to generate synthetic training samples. This technique generates a balanced dataset in each CV fold by interpolating between existing feature vectors of the minority class, thus generating synthetic feature vectors. In theory, samples in the same class should have similar features, so this method exploits these similarities to generate new feature vectors of the minority class. SMOTE is considered the standard framework for learning from imbalanced datasets and has previously been employed in semi-supervised learning scenarios [30].

Next, we compared the efficacy of seven different state-of-the-art binary classification ML algorithms: Logistic Regression (LR), K Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Gaussian Naive Bayes (GNB), Random Forests (RF), eXtreme Gradient Boosting (XGB), and Linear Discriminant Analysis (LDA). These models were implemented using the scikit-learn and XGBoost Python libraries [31,32]. To ensure a fair comparison between the different algorithms, the hyperparameters of each model were tuned simultaneously using Tree-structured Parzen Estimates (TPE) [33], implemented using the Python `hyperopt` package [34]. The objective of the TPE procedure was to find the combination of hyperparameters that produced the highest mean AUC across 5 cross-validation (CV) folds. The hyperparameters tuned for each model, as well as their possible values and the method of TPE choice, are listed in Appendix A.

The utilized CV procedure was a 5-fold GroupShuffleSplit [35]; in each CV fold, 20% of the recordings were randomly selected and used for validation, and the remaining recordings were used for training. The segmented coughs that comprised these recordings were correspondingly assigned to training or validation. This ensured that no coughs originating from the same recording were included in both the training and validation sets of each fold, thereby maintaining the generalizability of our results to unseen cough recordings.

Following TPE, the final mean and standard deviation AUC scores of all of the optimized models were analyzed. The model with the highest mean AUC was chosen, and its learning curve was analyzed to determine if the model was underfitting or overfitting, and whether or not the results converged to a consistent performance with the amount of data available. In the case of overfitting, Recursive Feature Elimination with Cross-Validation (RFECV) was performed on the optimized model to recursively remove the weakest features of the model. This technique has the potential to reduce the variance of the model through the elimination of weak features, but risks increasing the bias of the model by potentially eliminating important features [36]. Finally, the same TPE procedure was used to re-optimize the hyperparameters of the model with a reduced feature set.

An advantage of cough segmentation is that it enables aggregation of the classifier outputs of coughs originating from the same recording, which potentially enhances the accuracy of the classifier. Each recording was segmented into N cough sounds, and each cough was processed separately by the trained classifier. This resulted in a series of classifier output probabilities $[p_1, p_2, ..., p_N]$, corresponding to the probability that each cough signal is COVID-19 positive. Since this diagnosis cannot change from one cough to the next, the probabilities can be combined

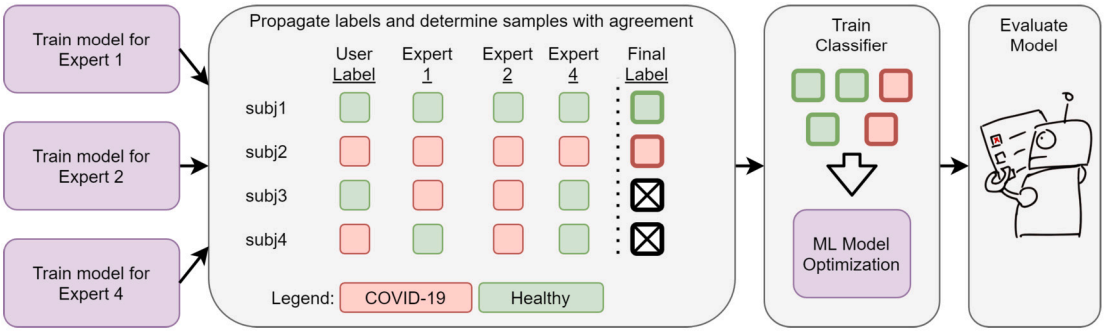


Fig. 4. The SSL approach consists of 1) training three separate ML models based on each expert's labels, 2) using the models to classify the unlabeled samples, 3) training a new classifier with the samples exhibiting a significant agreement between the user labels and expert models, and 4) testing the final model.

to form one classifier output per recording, p_{total}. We combined these probabilities by computing their logit mean:

$$p_{total} = \frac{1}{N}\sum_{i=n}^{N} log(\frac{p_i}{1-p_i}). \quad (1)$$

Once the optimized model was selected, one final 80%-20% CV split was generated to form a training and validation dataset. The model was trained on this reduced training dataset, and the ROC curve was plotted using the logit aggregation method. The optimal classifier decision threshold was determined for computing further accuracy metrics by selecting the aggregated logit threshold with the highest geometric mean between the model's sensitivity and specificity. These final hyperparameters were noted for use in testing, and the model was re-trained using the full training dataset. The final model was tested on the private, unseen COUGHVID test set, as described in Section 2.9.

2.8. Semi-supervised learning (SSL)

Instead of relying solely on the potentially noisy user labels or the often contradictory expert labels described in Section 2.3, an SSL approach was used to overcome the issues of label inconsistency and ambiguity by identifying a subset of consistent training data with a high probability of belonging to COVID-19-positive or healthy subjects, exhibiting a dry or wet cough, or having a known cough severity. At a high level, this method aims to distill the knowledge of each expert onto samples that the expert did not annotate, similarly to what is done in the state-of-the-art Pseudo-Label method [20]. Then, the agreement between the experts' models is used to identify a set of recordings with high label confidence, similarly to previous work on SSL applied to medical datasets with inconsistent labels [16]. This is similar to the cross-voting methodology utilized by Li et al. [23], except that instead of randomly partitioning the data to train each model, the data is divided based on the expert that annotated it, thereby modeling each expert's medical expertise using ML.

The semi-supervised learning methodology of the COVID-19 vs healthy classification task is illustrated in Fig. 4. First, three distinct ML models were trained based on each expert's COVID-19 versus healthy cough labels and optimized based on the procedure in Section 2.7. Then, the optimal classification threshold of each model was applied on these scores to produce a binary COVID-19 or healthy label for each recording. This procedure resulted in three to four labels per recording: three labels from the expert models, and one user label for the recordings for which this information was provided. In the event that an expert labeled a given recording, the expert's original COVID-19 or healthy diagnosis was maintained rather than the output of the model corresponding to that expert.

Next, a subset of high-confidence samples was identified by comparing the agreement of the expert models and user labels. The recordings with a high degree of agreement for being either COVID-19 positive or healthy were used to train one final classifier, and the rest were discarded. In order to select the final dataset, three different agreement schemes were tested and assessed in terms of database size and class separation:

1. *Universal Agreement:* All three expert models have the same label as the user label. This scheme limits the analysis to only the user-labeled datapoints.
2. *Expert Agreement:* All three expert models have the same label. This scheme bypasses the user label and can thus be applied on unlabeled samples in the dataset.
3. *Majority Agreement:* Either all three expert models have the same label, or two expert models have the same label as a user. This is the least conservative scheme as it allows one disagreement or missing user label.

In the case of the dry-vs-wet and mild-vs-severe cough classification tasks, in which no user labels were available, the same Pseudo-Label approach was performed for each of the four experts. The recordings in which at least three of the four expert models – or the original expert label if available – agreed were used to train the final model.

2.9. Model evaluation

Once all of the models described in Section 2.7 were trained and optimized, they were tested on the COUGHVID private test set to determine their generalization capabilities on unseen recordings [2]. This is a set of 625 recordings that have been labeled by at least one expert, and most recordings contain a user COVID-19 status label. The utilized success metric is AUC because it is a fair metric in terms of class imbalance. For each classification task, the training data labeling scheme was also used for testing. For example, when a model was trained using the annotations of Expert 2, it was tested only on the subset of testing data that had been labeled by Expert 2. In the case of the semi-supervised learning approach, the expert model label propagation procedure described in Section 2.8 was also performed on the testing set, and the final testing samples were identified using the same agreement scheme. Although the labels of the testing data may change between classification tasks, all data is drawn from the same set of recordings.

Once the various success metrics of the classifiers were computed, we assessed the most important features contributing to each classifier's outcome using the Shapley additive explanation (SHAP) values. These are measures of the relative importance of each feature, indicating which feature domains and specific measures had the greatest influence on the model's decision [37].

3. Results

3.1. COVID-19 classification SSL agreement scheme selection

First, we evaluate which agreement scheme among the expert models and user labels, described in Section 2.8, strikes the optimal trade-off

Table 2
Agreement scheme dataset coverage.

Label Scheme	Training Recs. (+)	Training Coughs (+)	Testing Recs. (+)	Testing Coughs (+)	Jensen-Shannon Divergence
User	10,850 (720)	25,227 (1,716)	287 (163)	1,098 (637)	0.00877
SSL Universal	2325 (14)	5515 (45)	28 (2)	104 (10)	0.0954
SSL Expert	4128 (98)	9583 (295)	141 (12)	501 (62)	0.0519
SSL Majority	8331 (285)	20337 (848)	240 (53)	876 (239)	0.0284

between training dataset coverage and label consistency. The three schemes were applied on the entire database and the number of remaining samples, both recordings and segmented coughs, of each class is reported in Table 2. This analysis provides an idea of how much training and testing data is maintained, as insufficient data may result in significant overfitting in the final model. Furthermore, the features described in Section 2.6 were computed for all of the segmented cough signals in each scheme to assess how the agreement scheme affects the class separation, which is used as a proxy measure of the label consistency across expert models. To quantify this class separation, the Jensen-Shannon divergence of each feature distribution in the COVID-19 and healthy cough classes was computed and averaged across all of the computed features of the training data. This metric ranges from 0 to 1, with higher values corresponding to a larger class separation. These results are displayed in Table 2, along with the same metrics computed on the user-labeled data subset.

As Table 2 shows, in the universal agreement scheme, there were only 14 COVID-19 positive recordings remaining in the training dataset. This amount is insufficient for the model to perform generalization. As expected, the majority agreement scheme produces the largest number of training samples. Intuitively, the Jensen-Shannon divergence decreases as the agreement scheme gets less conservative, meaning that the universal agreement scheme exhibits the largest class separation across features while the majority scheme has less pronounced differences between features. However, this increase in class separation comes at the expense of a decrease in dataset coverage, so the method that conserves the most data is maintained. This decision is in line with the findings of Zhu et al., which noted that SSL schemes prioritizing a larger initial dataset with moderately consistent labels performed better than small datasets with very reliable labels [13].

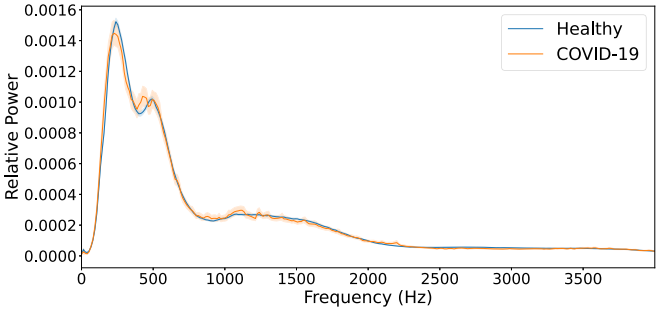
The majority agreement scheme was selected to identify the final COVID-19 and healthy cough samples. While this scheme has the smallest class separation of the other semi-supervised learning schemes, its Jensen-Shannon divergence is still more than three times higher than that of the user-labeled scheme. Furthermore, the number of training samples used in this agreement scheme is only 23% smaller than that of the user-labeled scheme, meaning that the increase in class separability does not sacrifice much of the data coverage. The percentage of COVID-19-labeled coughs in the majority agreement scheme is 3.3%, which is lower than the 6.2% in the user-labeled dataset. However, this class imbalance is handled in training by applying the SMOTE method described in Section 2.7.

3.2. Intra-class consistency analysis

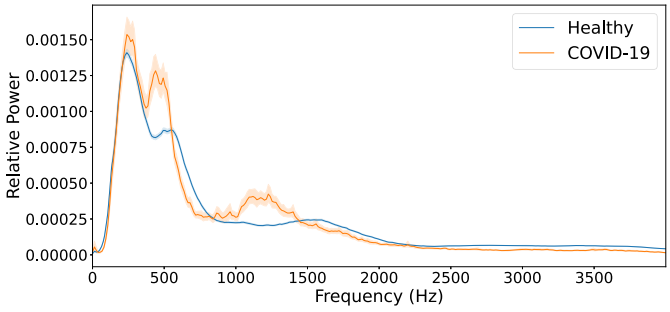
Once all three expert models were trained and optimized using the procedure in Section 2.7, these labels were propagated onto both the training and testing datasets. By selecting the subset of recordings for which the majority of labels were in agreement, we expanded the expert knowledge, combined with user self-report labels, to identify training and testing samples that had a high probability of having correct labels.

Table 3
Final COVID-19 vs healthy model selection.

Label Type	Model Used	Hyperparameters
User	LDA	None
Expert 1	LR	C = 18.31, class_weight = None, solver = "newton-cg"
Expert 2	LR	C = 0.01038, class_weight = "balanced", solver = "newton-cg"
Expert 4	LR	C = 0.3306, class_weight = None, solver = "lbfgs"
SSL	LR	C = 0.01009, class_weight = "balanced", solver = "newton-cg"



(a) User labeling scheme



(b) SSL labeling scheme

Fig. 5. Average normalized PSD of all cough signals in the training dataset belonging to each class according to the user labels and SSL labels.

The final optimized models evaluated in this work are displayed in Table 3, complete with their respective hyperparameters selected through TPE.

To assess the difference in audio properties of coughs labeled as COVID-19 and healthy in this new dataset compared to those of the user labels, the average normalized PSD curves of cough signals belonging to each class are plotted in Figs. 5a and 5b. The figures show a solid line, indicating the average PSD, as well as the 95% confidence interval across all segmented cough audio samples of a given class in each labeling scheme. Fig. 5a shows very few differences between the spectra of the user-labeled COVID-19 and healthy coughs, with a slight variation in the 400-550 Hz and 1000-1500 Hz ranges. In comparison, Fig. 5b depicts a similarly shaped spectrum of healthy coughs as the user-labeled healthy coughs, but there are much more pronounced differences between the COVID-19 coughs and healthy ones. The bandpowers of the COVID-19 coughs are significantly higher in the 400-550 Hz and 1000-1500 Hz ranges than those of healthy coughs, with p-values of

Table 4
SHAP feature ranking across COVID-19 classifiers.

Feature Ranking (SHAP)	User	Expert 1	Expert 2	Expert 4	SSL
1	EEPD 350-400	Spectral Centroid	Gender	Crest Factor	MFCC Std. 0
2	EEPD 600-650	RMS Power	MFCC Mean 7	MFCC Mean 0	MFCC Mean 1
3	RMS Power	MFCC Std. 0	PSD 550-800	Spectral Slope	MFCC Mean 9
4	EEPD 900-950	Spectral Spread	MFCC Std. 5	Spectral Rolloff	Gender
5	Dominant Frequency	Spectral Skewness	EEPD 400-450	MFCC Mean 9	MFCC Mean 7

1.4×10^{-36} and 1.2×10^{-64}, respectively. This analysis highlights the substantial difference in spectral features of COVID-19 and healthy coughs identified through the SSL approach. It is also consistent with the findings of Table 2, which shows that on average, the chosen SSL dataset exhibits over 3x more class separability than the user-labeled data in terms of the average Jensen-Shannon divergence across all extracted audio features.

To expand on the feature analysis, the top five most important features for each classifier, determined by their SHAP values, are displayed in Table 4. When we analyze the three expert models, it is clear that there are few features in common between the classifiers and they each weigh features of different domains (i.e., time, frequency, and Mel frequency) with varying importance. The semi-supervised learning classifier, on the other hand, has features in common with several expert models (MFCC standard deviation 0, MFCC mean 7, and gender). Furthermore, the majority of its top features are in the Mel frequency domain, which is meant to model how the human auditory system processes sound signals.

3.3. Open-sourced SSL dataset

In order to contribute to further research in the field of COVID-19 cough sound diagnosis, we have added the training labels obtained through our SSL majority agreement scheme to the latest version of the COUGHVID dataset public Zenodo repository. This version has been expanded to include all of the crowdsourced recordings obtained through October 2021, whereas the original dataset only contained recordings uploaded through December 2020. These new labels can be found in the newly added `status_SSL` column of the `metadata_ compiled` CSV file.

The new SSL scheme provides labels for 1,018 recordings that were previously unlabeled by users or experts, which demonstrates the utility of SSL in utilizing data that had previously been unusable. Furthermore, there are 581 recordings that the users labeled with the ambiguous "symptomatic" label, but the SSL model provides a "COVID-19" or "healthy" label. A mere 32 of these coughs were labeled by the SSL model as COVID-19 positive, which is feasible considering the COVID-19 infection rates during the period of recording.

Users of the COUGHVID dataset can use these new labels and corresponding data samples to augment their COVID-19 cough classification models with highly consistent training data. The same SSL label expansion procedure was conducted for the private testing dataset described in 2.9, so users are welcome to test their models against these labels as ground-truth, but must acknowledge that these labels are not confirmed by RT-PCR tests.

3.4. Cough type and severity SSL

As described in Section 2.2, a similar SSL approach to that of the COVID-19 detection was employed for the cough type (wet vs. dry) and severity (severe vs. mild) classification tasks. Table 5 lists the dataset coverage, CV AUC, and average J-S divergence across extracted features of each expert model, as well as the baseline expert union and SSL models. The detailed implementations of these models, including which specific classifiers were used and their respective hyperparameters, are discussed in Appendix B.

Table 5
Cough type and severity model training.

Label scheme	Dry (-) wet (+) coughs	Mild (-) severe (+) coughs	Wet vs dry AUC	Severe vs mild AUC	Wet vs dry J-S Div.	Severe vs mild J-S Div.
Exp.1	1,093 - 203 +	1,360 - 122 +	0.76	0.65	0.032	0.038
Exp.2	1,506 - 395 +	1,520 - 234 +	0.55	0.6	0.011	0.024
Exp.3	834 - 897 +	1,130 - 442 +	0.66	0.61	0.019	0.018
Exp.4	1,591 - 361 +	1,260 - 295 +	0.69	0.6	0.021	0.024
Union	4,393 - 1725 +	4,635 - 1,070 +	0.61	0.62	0.0053	0.0084
SSL	28,771 - 1,192 +	12,558 - 2,160 +	0.9	0.9	0.06	0.034

Table 6
COVID-19 classification model testing results.

Model	CV AUC	Test AUC (Not Agg.)	Test AUC (Agg.)
User	0.591	0.564	0.562
Exp. 1	0.653	0.652	0.681
Exp. 2	0.669	0.663	0.743
Exp. 4	0.644	0.561	0.593
SSL	0.883	0.763	0.797

For the wet vs. dry cough classification task, the SSL approach increases the training dataset coverage by a factor of 5. We can see from the average J-S divergences that the SSL model has the highest feature separability between the two classes of all other labeling schemes, including those of individual experts. Furthermore, the feature separability of the SSL-relabeled samples increases by 11.3x compared to the baseline expert union labeling scheme. These results are reflective of the low agreement between each of the experts, and the SSL relabeling approach produces relabeled samples with distinct acoustic differences compared to simply combining all of the experts' labels. As a result of the increased feature separability, the CV AUC of the SSL model is 47.5% higher than that of the expert union model.

Next, for the severe vs. mild cough classifier, the dataset coverage and average J-S divergence are 3x and 5.1x higher in the SSL approach than the expert union modeling approach. The CV AUC of the SSL model is 45.2% higher than that of the expert union model.

3.5. ML model evaluation

To demonstrate how the SSL dataset can be used to train cough analysis ML models for the three classification tasks, the final model was developed using the procedure in Section 2.7 using the SSL labels. The final testing results of each model of the COVID-19 classifier are displayed in Table 6, which shows all of the accuracy metrics, as well as the AUC obtained during cross-validation, non-aggregated testing (i.e., testing on every individual cough sound), and aggregated testing on each recording using Equation (1). We observe an average 5.26% increase in accuracy between non-aggregated and aggregated testing. This implies that testing each cough separately and combining the results for each recording enhances the model's performance. Aggregating the probabilities of each cough sound in a recording may exploit the correlations between the coughs and diminish the effects of outlier cough sounds, thus providing a more robust classification than predicting each cough sound separately.

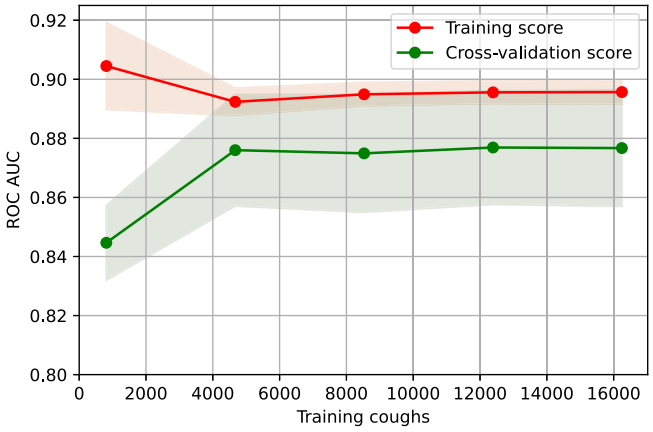


Fig. 6. Learning curve of the final optimized COVID-19-vs-healthy SSL classifier displaying the training and cross-validation accuracy using varying sizes of the training dataset.

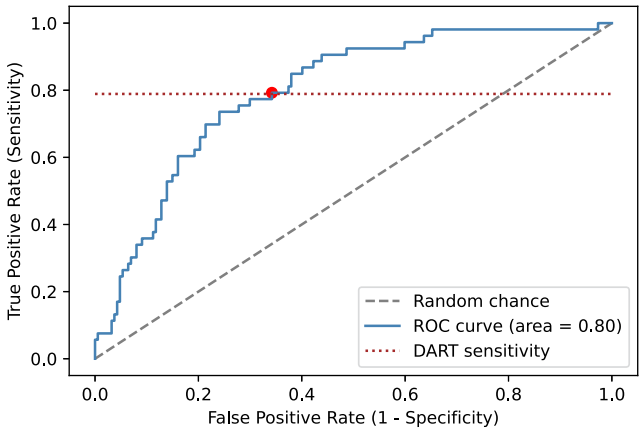


Fig. 7. The final ROC curve of the semi-supervised classifier on the aggregated testing set. At the red point, the model achieves a sensitivity of 79.2%, which is comparable to that of DART COVID-19 tests, and a specificity of 65.8%.

The model trained on self-reported user data exhibited the worst performance. Reaching only a testing AUC of 0.562, it was scarcely better than a random classifier. We observe a high variance in the success of the expert models, with Expert 2 having the highest AUC and Expert 4 the lowest. The AUC of the semi-supervised classification method is, on average, 15.6% higher than that of the expert models, and 29.5% higher than that of the user model.

The final SSL model utilized 32 of the available 66 features, which were selected through RFECV. The learning curve of the final SSL model is displayed in Fig. 6, which shows the effect of varying the training data size (in terms of number of segmented coughs) on the training and validation accuracy. The solid line depicts the mean scores across the five CV folds, and the shading around the line indicates one standard deviation from the mean. We can see that the model converges to a validation accuracy at around 4,000 training coughs, indicating that the model has sufficient training samples to gain insights from the features. Furthermore, the relatively small 2% gap between the training and validation scores indicates that the variance of the model is low, meaning that it is not over-fitting on the training samples.

The ROC curve of the semi-supervised learning classifier is displayed in Fig. 7. The model had an AUC of 0.797 on the private testing set. To evaluate the classification accuracy of such a model, we compare its sensitivity and specificity to those of commonly-used at-home COVID-19 tests. The Direct Antigen Rapid Test (DART) for COVID-19 screening was reported to have a sensitivity of 78.9% and specificity of 97.1% within 0 to 12 days from symptom onset [38]. To achieve a comparable sensitivity, our classifier exhibits a specificity of 65.8%, which is a 32.8% decrease from that of DART tests. This decrease in specificity could be justified by the inexpensive, ubiquitous, and non-invasive nature of an audio-based screening tool versus the traditional nasal swab.

Finally, we evaluate the final testing AUC of all of the classifiers trained throughout the study and report the performance in Fig. 8. These AUC results were obtained for each model using the private test set described in Section 2.9 and the aggregation method in Equation (1) was used. We can observe that regardless of the classification task, the SSL re-labeling led to a higher testing score than any of the individual classifiers. The baseline expert union models in both the cough type and severity models performed no better than a random classifier. The SSL models out-performed the baselines by 65% and 44% in the cough type and severity classifiers, respectively.

4. Discussion

Labeling medical data requires significant time and effort from expert annotators. Moreover, this tedious process often leads to inconsistencies due to a lack of agreement between experts. This situation is a key drawback to confront new viruses, as it has happened with COVID-19. In this work, we have shown that using an SSL model development method, it is possible to overcome expert label scarcity and inconsistency – as well as user mislabeling of crowdsourced medical datasets – to identify a subset of data points with a high-class separability. The re-labeled data was then publicly provided to the research community to assist in COVID-19 classification from cough sounds.

By integrating the knowledge from three medical experts with the user labels through the SSL approach, we identified a subset of cough recordings that had a high probability of belonging to COVID-19 vs healthy individuals, with or without mucus secretion, exhibiting mild or severe cough pathologies. We see from the averaged PSD curves of COVID-19 and healthy coughs in Fig. 5b that the majority-voting approach between the expert pseudo-labels successfully identified two classes of coughs with significantly different spectral characteristics and a high class separation in the extracted audio features. As each expert model and user labels aimed to separate COVID-19 versus healthy coughs, it can be postulated that these spectral characteristics are present in the underlying distributions of the two classes of cough sounds. Furthermore, these spectral differences were much less pronounced when the same analysis was performed for the user-labeled data in Fig. 5a. These figures illustrate the fact that the class separation was over 3x higher in the SSL training data than of the user labeling in terms of the Jensen-Shannon divergence of the feature distributions. This increase in class separation did not come at a significant cost to the data coverage, as the training data size only decreased by 23%. When comparing the SSL approach to the union of experts' labels, the training dataset size of the wet-vs-dry classification task increased 5x, while the average separability between features increased by 11.3x. Similarly for the severe-vs-mild classification task, the SSL approach increased the dataset size 3x with a 5.1x increase in feature separability.

A SHAP analysis of the most important features of the expert and SSL COVID-19 classifiers in Table 4 revealed that there were few important features common among all expert classifiers, which may be reflective of the lack of expert agreement observed in [2]. This analysis also showed that three of the five most important features of the final SSL classifier were also significant in each expert classifier, which implies that the model successfully integrated each expert's medical knowledge. Furthermore, the fact that the SSL model relies almost entirely on MFCC features might imply that the classifier is learning to model the human auditory system, since these features model how humans perceive sound.

Finally, an analysis of the model testing results in Fig. 6 reveals the drawbacks of classic supervised learning approaches, as well as the improved performance of SSL. First, we note that the model that was trained and tested on crowdsourced user labels achieved a testing AUC

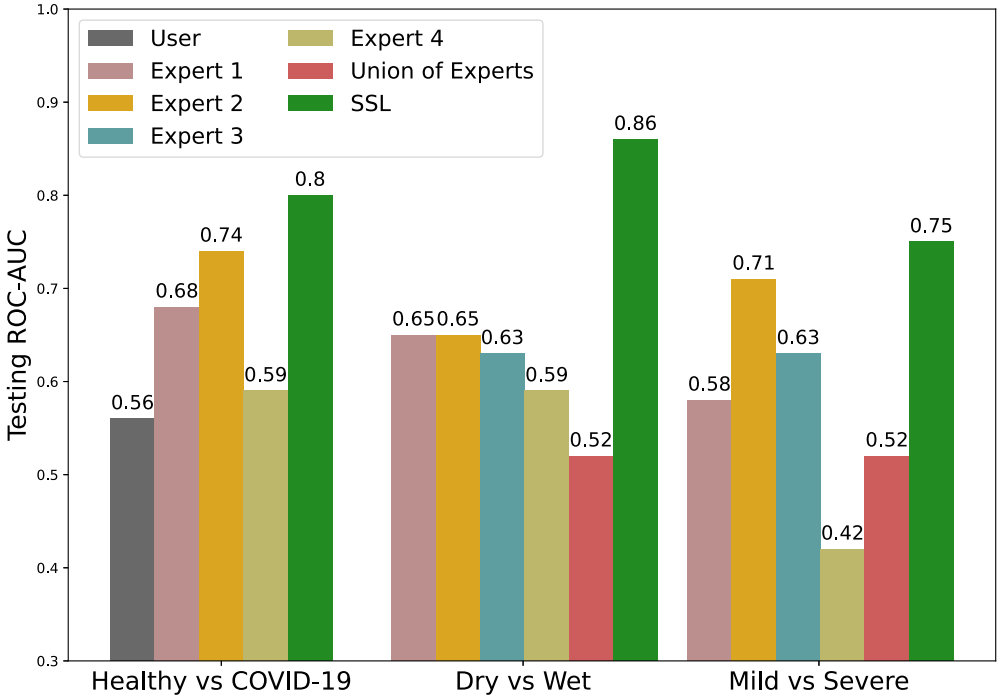


Fig. 8. Final AUC results on the private test set of the various classifiers on each inference task.

of 0.562, and such a low score implies that there is significant mislabeling present in the dataset. Additionally, taking the union of all experts' labels resulted in testing AUC scores of 0.52 for both the cough type and severity classifiers, which is reflective of the poor agreement between the experts for these tasks.

Next, we note a wide variance between the success of each expert model, with the COVID-19 classifier AUC scores ranging from 0.593 to 0.743. This means that the expert labels tend to be inconsistent between and even within each expert's labels. Similar trends were observed in the four expert models of both the cough type and cough severity classifiers. However, despite the setback of label ambiguity, the semi-supervised modeling approach achieved a high final AUC score of 0.797 for the COVID-19 classifier, which leads to a sensitivity of 79.2% and specificity of 65.8%. The AUC of the SSL model was at least 7.3% higher than any of the expert models. This highlights the increase in labeling consistency and consequent model success.

Finally, the cough type and severity classifiers exhibited final AUC scores of 0.86 and 0.75, respectively, which are significantly higher than any of the expert models or expert union model. These results indicate that integrating the medical knowledge of multiple experts in a semi-supervised fashion results in a more robust, consistent classifier than supervised learning based on any of the individual experts' labels. Furthermore, as the cough type and severity are not specific to any pathology in particular, this modeling approach and its consequent re-labeled data can be used to develop classifiers for other conditions beyond COVID-19.

The primary limitation of our study is a lack of clinically-validated labels on which to test the final models. Although the proposed SSL method showed an increased model performance on the COUGHVID hidden test set, this approach must be thoroughly validated on PCR-confirmed cough samples. Such ground-truth data would also be necessary to determine whether the frequency-domain differences between SSL-relabeled COVID-19 and healthy coughs displayed in Fig. 5 truly correspond to the underlying distributions of the two cough classes. Furthermore, the algorithm is unable to account for concept drift due to the varying symptomologies of the different COVID-19 virus variants. The data used in training was obtained through October 2021, whereas the Omicron variant – which had a significantly lower rate of respiratory symptoms than previous variants – was first reported in November 2021 [39]. Moreover, the cough type and severity were not validated by any clinical tests, so it cannot be known from this analysis how well these models perform on patients. Finally, the cough type classifier would require more fine-grained ground-truth labeling to work on a per-cough basis, as there may be both dry and wet coughs present in a given recording. Therefore, these issues must be addressed in future work to accurately assess the clinical usefulness of such respiratory disorder classification models. However, in the absence of an extensive, RT-PCR-validated dataset, our proposed approach can be used to improve the quality of large, crowdsourced cough databases and identify samples with consistent patterns in the cough recordings of each class. These re-labeled coughs can then be used to augment datasets of medically confirmed cough sounds to enhance the training data size and potentially improve the classification accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017915 (DIGIPREDICT). T. Teijeiro is supported by the grant RYC2021-032853-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Appendix A. Optimized hyperparameters

The hyperparameters of each model that are optimized through TPE are displayed in Table A.7. The table lists the possible values or ranges of values of each hyperparameter, as well as the way that the parameter was selected. "Choice" means that discrete values were chosen from the given range, "Uniform" means that a value was selected on a uni-

Table A.7
Hyperparameter optimization.

Model name	Hyperparameter	Possible values	Selection
LR	C	$[10^{-2}, 10^{6}]$	Log
	class_weight	None, "balanced"	Choice
	solver	"newton-cg", "lbfgs"	Choice
SVM	C	$[10^{-2}, 10^{6}]$	Log
KNN	n_neighbors	[1,n_features/2]	Choice
	weights	"uniform","distance"	Choice
NB	var_smoothing	$[10^{-10}, 10^{2}]$	Log
DTC	min_samples_split	[2,5]	Choice
	min_samples_leaf	[1,n_features]	Choice
	max_features	[1,n_features]	Choice
RF	n_estimators	[10,300]	Choice
	max_features	[1,n_features]	Choice
	criterion	"gini", "entropy"	Choice
XGB	max_depth	[1,10]	Choice
	max_delta_step	[1,10]	Choice
	gamma	[0,1]	Uniform
	subsample	[0,1]	Uniform
	reg_lambda	[0.5,2]	Uniform
	eta	[0,1]	Uniform
	sampling_method	"uniform","gradient-based"	Choice
LDA	N/A	N/A	N/A

Table B.8
Wet vs dry cough model configurations.

Label type	Model used	Hyperparameters
Expert 1	RF	criterion: "entropy", max_features: 4, n_estimators: 299
Expert 2	LR	C: 1.1, class_weight: "balanced", solver: "lbfgs"
Expert 3	LR	C: 0.01, class_weight: "balanced", solver: "newton-cg"
Expert 4	LR	C: 0.01, class_weight: None, solver: "newton-cg"
Union	RF	criterion: "entropy", max_features: 44, n_estimators: 297
SSL	RF	criterion: "gini", max_features: 8 n_estimators: 100

Table B.9
Severe vs mild cough model configurations.

Label type	Model used	Hyperparameters
Expert 1	LDA	None
Expert 2	RF	criterion: "gini", max_features: 47, n_estimators: 135
Expert 3	RF	criterion: "entropy", max_features: 46, n_estimators: 135
Expert 4	LR	C: 0.01, class_weight: None, solver: "newton-cg"
Union	RF	criterion: "gini", max_features: 41, n_estimators: 187
SSL	XGB	eta: 0.12, gamma: 0.81, max_delta_step: 9, max_depth: 8, reg_lambda: 1, sampling_method: "uniform", subsample: 0.82

form distribution within the range, and "Log" means that the logarithms within the range were sampled at a random distribution.

Appendix B. Cough type and severity model implementations

For each model implementation, the hyperparameters listed in Appendix A were optimized using TPE, as described in Section 2.7. Tables B.8 and B.9 list the specific model implementations, complete with their lists of optimized hyperparameters, for the cough type and cough severity classifications, respectively.

References

[1] J. Heitmann, A. Glangetas, J. Doenz, J. Dervaux, D.M. Shama, D.H. Garcia, M.R. Benissa, A. Cantais, A. Perez, D. Müller, T. Chavdarova, I. Ruchonnet-Metrailler, J.N. Siebert, L. Lacroix, M. Jaggi, A. Gervaix, M.-A. Hartley, DeepBreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries, npj Digit. Med. 6 (1) (2023) 1–12.

[2] L. Orlandic, T. Teijeiro, D. Atienza, The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms, Sci. Data 8 (1) (2021) 156.

[3] T. Xia, D. Spathis, C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, E. Bondareva, T. Dang, A. Floto, P. Cicuta, C. Mascolo, COVID-19 sounds: a large-scale audio dataset for digital respiratory screening, in: 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 2021.

[4] J. Laguarta, F. Hueto, B. Subirana, COVID-19 artificial intelligence diagnosis using only cough recordings, IEEE Open. J. Eng. Medicine Biol. 1 (2020) 275–281.

[5] F. Manzella, G. Pagliarini, G. Sciavicco, I.E. Stan, The voice of COVID-19: breath and cough recording classification with temporal decision trees and random forests, Artif. Intell. Med. 137 (Mar. 2023).

[6] N.K. Chowdhury, M.A. Kabir, M.M. Rahman, S.M.S. Islam, Machine learning for detecting COVID-19 from cough sounds: an ensemble-based MCDM method, Comput. Biol. Med. 145 (Jun. 2022).

[7] Y. Chang, X. Jing, Z. Ren, B.W. Schuller, CovNet: a transfer learning framework for automatic COVID-19 detection from crowd-sourced cough sounds, Frontiers Digit. Health 3 (2022).

[8] A. Ponomarchuk, I. Burenko, E. Malkin, I. Nazarov, V. Kokh, M. Avetisian, L. Zhukov, Project achoo: a practical model and application for COVID-19 detection from recordings of breath, voice, and cough, IEEE J. Sel. Top. Signal Process. 16 (2) (2022) 175–187.

[9] H. Xiong, S. Berkovsky, M.A. Kâafar, A. Jaffe, E. Coiera, R.V. Sharan, Reliability of crowdsourced data and patient-reported outcome measures in cough-based COVID-19 screening, Sci. Rep. 12 (1) (Dec. 2022).

[10] M. Lotfi, M.R. Hamblin, N. Rezaei, COVID-19: transmission, prevention, and potential therapeutic opportunities, Clin. Chim. Acta 508 (2020) 254–266.

[11] M. Van Such, R. Lohr, T. Beckman, J.M. Naessens, Extent of diagnostic agreement among medical referrals, J. Eval. Clin. Pract. 23 (4) (2017) 870–874.

[12] T. Teijeiro, C.A. García, D. Castro, P. Félix, Abductive reasoning as a basis to reproduce expert criteria in ECG atrial fibrillation identification, Physiol. Meas. 39 (8) (Aug. 2018).

[13] Z. Zhu, Y. Sato, Speech emotion recognition using semi-supervised learning with efficient labeling strategies, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 358–365.

[14] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Semisupervised autoencoders for speech emotion recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (1) (2018) 31–43.

[15] A. Inés, C. Domínguez, J. Heras, E. Mata, V. Pascual, Biomedical image classification made easier thanks to transfer and semi-supervised learning, Comput. Methods Programs Biomed. 198 (Jan. 2021).

[16] M. Guan, V. Gulshan, A. Dai, G. Hinton, Who said what: modeling individual labelers improves classification, Proc. AAAI Conf. Artif. Intell. 32 (1) (Apr. 2018).

[17] H. Xue, F.D. Salim, Exploring self-supervised representation ensembles for COVID-19 cough classification, in: 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1944–1952.

[18] T. Dang, T. Quinnell, C. Mascolo, Exploring semi-supervised learning for audio-based COVID-19 detection using FixMatch, in: Interspeech 2022, ISCA, 2022, pp. 2468–2472.

[19] W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, X. Zhu, Semi-supervised active learning for sound classification in hybrid learning environments, PLoS ONE 11 (9) (Sep. 2016).

[20] D.-H. Lee, Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, vol. 3, 2013, p. 896.

[21] Z.-H. Zhou, A brief introduction to weakly supervised learning, Nat. Sci. Rev. 5 (1) (2018) 44–53.
[22] J. Lv, X. Chen, J. Huang, H. Bao, Semi-supervised mesh segmentation and labeling, Comput. Graph. Forum 31 (7) (2012) 2241–2248.
[23] Z. Li, Z. Gan, B. Zhang, Y. Chen, J. Wan, K. Liu, J. Zhao, S. Liu, Semi-supervised noisy label learning for Chinese clinical named entity recognition, Data Intell. 3 (3) (2021) 389–401.
[24] A.P. Dawid, A.M. Skene, Maximum likelihood estimation of observer error-rates using the EM algorithm, J. R. Stat. Soc., Ser. C, Appl. Stat. 28 (1) (1979) 20–28.
[25] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (5) (1971) 378–382.
[26] R.X.A. Pramono, S.A. Imtiaz, E. Rodriguez-Villegas, A cough-based algorithm for automatic diagnosis of pertussis, PLoS ONE 11 (9) (2016).
[27] H. Chatrzarrin, A. Arcelus, R. Goubran, F. Knoefel, Feature extraction for the differentiation of dry and wet cough sounds, in: MeMeA 2011 - 2011 IEEE International Symposium on Medical Measurements and Applications, Proceedings, 2011.
[28] A.B. Chang, The physiology of cough, Paediatr. Respir. Rev. 7 (1) (2006) 2–8.
[29] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[30] A. Fernandez, S. Garcia, F. Herrera, N.V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, J. Artif. Intell. Res. 61 (2018) 863–905.
[31] scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation, https://scikit-learn.org/stable/.
[32] Python Package Introduction — xgboost 2.0.0, https://xgboost.readthedocs.io/en/latest/python/python_intro.html.
[33] J.S. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, in: Advances in Neural Information Processing Systems, vol. 24, 2011, pp. 2546–2554.
[34] Hyperopt Documentation, http://hyperopt.github.io/hyperopt/.
[35] Group shuffle split, https://scikit-learn/stable/modules/generated/sklearn.model_selection.GroupShuffleSplit.html.
[36] M.A. Munson, R. Caruana, On feature selection, bias-variance, and bagging, in: Machine Learning and Knowledge Discovery in Databases, vol. 5782, 2009, pp. 144–159.
[37] E. Štrumbelj, I. Kononenko, An efficient explanation of individual classifications using game theory, J. Mach. Learn. Res. 11 (2010) 1–18.
[38] A. Harmon, C. Chang, N. Salcedo, B. Sena, B.B. Herrera, I. Bosch, L.E. Holberger, Validation of an at-home direct antigen rapid test for COVID-19, JAMA Netw. Open 4 (8) (Aug. 2021).
[39] D. Bouzid, B. Visseaux, C. Kassasseya, A. Daoud, F. Fémy, C. Hermand, J. Truchot, S. Beaune, N. Javaud, O. Peyrony, A. Chauvin, P. Vaittinada Ayar, A. Bourg, B. Riou, S. Marot, B. Bloom, M. Cachanado, T. Simon, Y. Freund, Comparison of patients infected with delta versus omicron COVID-19 variants presenting to Paris emergency departments, Ann. Intern. Med. 175 (6) (2022) 831–837.