COVIDCatcher: Developing A Low-Cost Multimodal Machine-Learning Based App for Detecting COVID-19 Symptoms

Michael Li

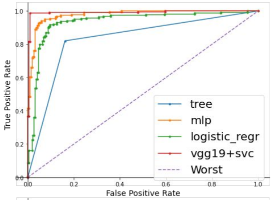
Q1: Question

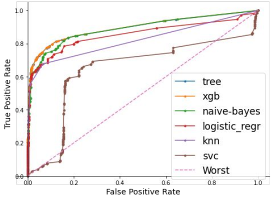
- **Problem**: the elderly and immunocompromised are at risk for COVID-19 transmission when leaving home to take a COVID test; no tool exists to quickly and cheaply detect COVID-19 symptoms at home.
- **Goal**: Develop a cost-effective, multimodal, data-driven tool to detect COVID-19 symptoms

Q2: Framework

1. Identify datasets for symptom and cough detection
2. Clean dataset and extract features for model
3. Build, test and evaluate multiple model and processing methods
4. Deploy top-performing model frameworks
5. Develop symptoms checker app: COVIDCatcher
6. Beta test and collect feedback on COVIDCatcher
7. Iterate and improve models and user experience

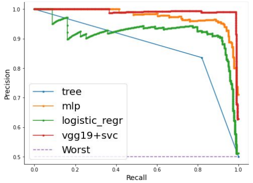
Q3: Findings

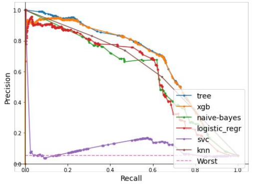




Model	Accuracy	ROC AUC	Avg Precision
Decision Tree	82.94	0.8294	0.7753
Logistic Regression	88.94	0.9355	0.8335
MLP	93.54	0.9778	0.9022
VGG19+SVC	**0.9884**	**0.9909**	**0.9840**

Model	Accuracy	ROC AUC	Avg Precision
Logistic Regression	96.16	0.8527	0.3648
K-Nearest Neighbors	96.11	0.7966	0.3688
Decision Tree	**96.58**	**0.8907**	**0.4419**
XGBoost	**96.62**	**0.8924**	**0.4480**
SVC	93.92	0.6448	0.0749
Gaussian Naive Bayes	94.27	0.8840	0.3275





Q4: Conclusions

1. **COVIDCatcher** is the first multimodal, data-driven approach to evaluate COVID symptoms
2. COVIDCatcher is free and scalable to the public
3. XGBoost and VGG+SVC are effective for COVID symptom and cough detection, respectively, showing **>95%** accuracy

Introduction - Problem

- **54.6 million** elderly and **10 million** immunocompromised people in the U.S.
    - In-person tests present risk of COVID-19 exposure
- At-home COVID-19 tests are expensive (>$100) and limited
- **2.85 million** global deaths from COVID-19, with **555k** U.S. deaths (U.S. Census Bureau & WHO)
- Existing solutions are either not data-driven (CDC), OR lack a human-usable or data-driven application

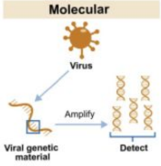
COVID-19 Diagnostics		Advantages	Limitations
Molecular Test (detects piece of viral DNA through PCR testing.) Source: FDA, "A Closer Look at COVID-19 Diagnostic Testing"	 Molecular Virus Amplify Viral genetic material Detect Source: GAO. | GAO-20-584SP	Free to public, accuracy level of 94% https://www.medrxiv.org/content/10.1101/2020.04.05.20053355v1.full.pdf	Risk of exposure when outside home, need to wait 2-3 days for results, long lines, only a few authorized for at home use. Source: FDA, "A Closer Look at COVID-19 Diagnostic Testing"
Antigen test (detects proteins from a virus particle, generally through a nasal swab or nasopharyngeal swab) Source: FDA, "A Closer Look at COVID-19 Diagnostic Testing"	 Source: National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases	Takes within minutes for results, and most are authorized for at home use. Source: FDA, "A Closer Look at COVID-19 Diagnostic Testing"	Higher false positive rate than molecular test, lower sensitivity than molecular test; risk of exposure when tested outside
At-home COVID-19 tests (collect your own sample and test it with RT-PCR or NAAT)	 Source: National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases	Can take test from home; no need for human contact since the test is mail-in	Takes time to mail/mail back tests, expensive: costs >$100 for single use, can only buy 1 at a time because limited in quantity
CDC Coronavirus Self-Checker	CORONAVIRUS DISEASE 2019 (COVID-19) CORONAVIRUS SELF-CHECKER CDC Hi, I'm Clara. I'm here to guide you through the Coronavirus Self-Checker. CDC cdc.gov/coronavirus Source: CDC Coronavirus Self-Checker	Free and easy to find on the CDC website	Uses simple logic that does not take into account asymptomatic carriers and is tedious to fill out

Figure 1. COVID-19 detection methods currently available to the American public

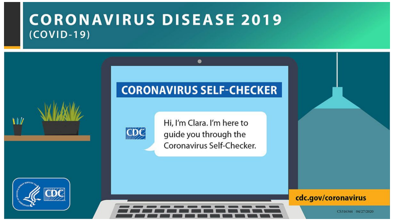
Introduction - Objective & Literature Review

<u>Goal</u>: To develop a **cost-effective**, **multimodal**, **data-driven tool** to help individuals, especially the elderly and immunocompromised, identify COVID-19 symptoms at home

Existing solutions are limited by **expensive** costs, **delays**, or **lack** of a real, **usable** application and deployment to society

***COVID-19 Antibody Tests and Their Limitations (Liu, 2021)*[1]**

- Molecular PCR tests have high <u>false-negative</u> rate, high <u>cost</u>, need <u>skilled</u> workers
- Low-cost rapid antigen tests have <u>poor sensitivity</u>, or require more <u>research</u> validation



Source: CDC Website

***CDC COVID-19 Health Bot (CDC, 2020)*[2]**

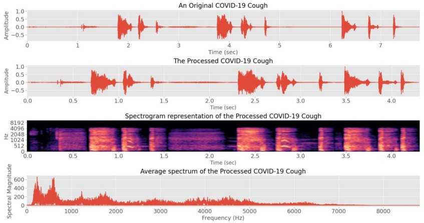
- Open-source COVID-19 symptom checker; <u>no guarantees</u> on accuracy
- Simple <u>rule-based</u> boolean logic using handcrafted flow chart, <u>not data-driven</u>

***Machine learning-based prediction of COVID-19 diagnosis based on symptoms (Zoabi, 2021)*[3]**

- Gradient-boosting predictor using LightGBM Python package
- Limited by <u>small</u> dataset, <u>self-reported</u> symptoms and no edge cases for <u>asymptomatic</u>

***COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings (Pahar, 2020)*[4]**

- ResNet50 discriminated between COVID-19 negative/positive coughs.
- <u>Imbalanced dataset</u>: only 92 COVID-19 positive vs. 1079 healthy subjects



Source: Pahar[4]

Framework - Concepts & Definitions

- **Data processing.** Aggregate and clean data; extract important features and labels.
- **Model development.** Machine learning models were built and tested on the data. ROC AUC, recall and precision were analyzed to select the top performing model.
- **Hyperparameter tuning.** A grid search of model parameters was performed to find the optimal combination of parameters for model performance.
- **XGBoost algorithm** - a popular open-source implementation of the gradient boosted trees algorithm that uses multiple trees to increase robustness.[5]
- **Gradient boosting** - classification technique that utilizes an ensemble of weak prediction models[5]
- **VGG19** - a state-of-the-art convolutional neural network, 19 layers deep[6]
- **Linear SVM** - Finds the hyperplane with best margin of separation for binary classification, used for cough classification.
- **Spectrogram** - a visual representation of the spectrum of frequencies of a signal as it varies with time
- **Logistic Regression** - predictive linear algorithm for binary classification
- **Decision Tree Classifier** - predictive model that uses decision tree for classification
- **Web App Development -** Models were saved via Pickle and loaded to a web app in Heroku with remote hosting.

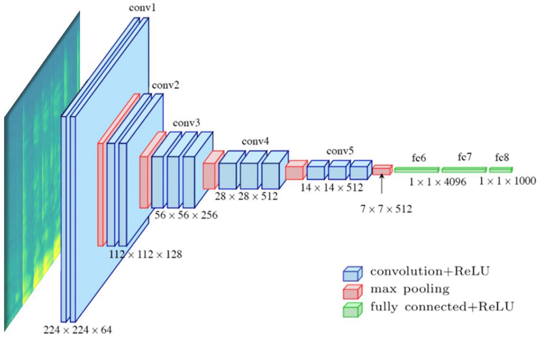


Figure 2. VGG19 Structure[6]

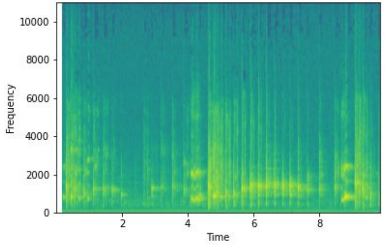
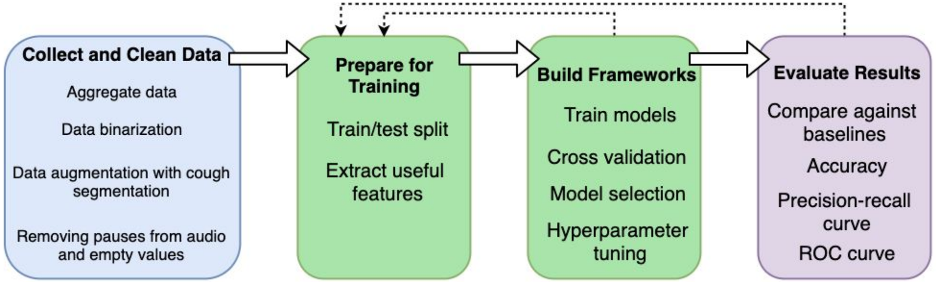


Figure 3. Spectrogram generated using Python

Framework - Methodology

Data & Backend Model Development

1. Identify, aggregate, process training data
   a. 2.7 million Israeli COVID symptoms dataset (COVID-: 2,521,621, COVID+: 220,975)
   b. 1,400 aggregated coughs: Virufy, Coswara, EPFL
2. **Symptom Detection**: Build + test XGBoost, Naive Bayes, Decision Tree, KNN, SVC, Logistic Regression
3. **Cough Detection**: Design + test framework: spectrogram pre-processing, VGG feature extraction, and SVC classification; compared results with baseline models



Collect and Clean Data

Aggregate data

Data binarization

Data augmentation with cough segmentation

Removing pauses from audio and empty values

Prepare for Training

Train/test split

Extract useful features

Build Frameworks

Train models

Cross validation

Model selection

Hyperparameter tuning

Evaluate Results

Compare against baselines

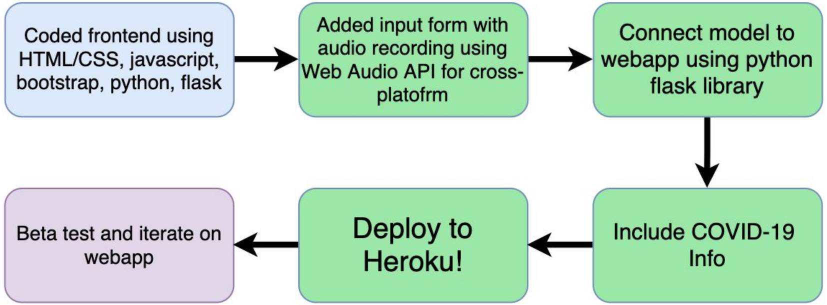
Accuracy

Precision-recall curve

ROC curve

Figure 4. Backend development workflow.

Front-End COVIDCatcher Development



Coded frontend using HTML/CSS, javascript, bootstrap, python, flask

Added input form with audio recording using Web Audio API for cross-platofrm

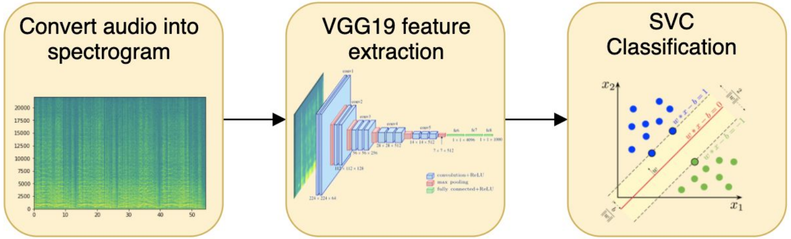
Connect model to webapp using python flask library

Include COVID-19 Info

Deploy to Heroku!

Beta test and iterate on webapp

Figure 5. Web App. Models were saved via Pickle and loaded to a web app in Heroku with remote hosting.



Convert audio into spectrogram

VGG19 feature extraction

SVC Classification

Figure 6. Custom cough detection workflow.

Results

Symptom Detection

- **Task**: Given a set of patient symptoms, classify a patient as COVID-positive or negative
- XGBoost showed top performance for COVID-19 symptom detection, with **96.62%** accuracy
- Symptom examples: Cough, fever, headache, shortness of breath, sore throat, contact with COVID, and elderly

(a)

Model	Accuracy	ROC AUC	Avg Precision
Logistic Regression	96.16	0.8527	0.3648
K-Nearest Neighbors	96.11	0.7966	0.3688
Decision Tree	**96.58**	**0.8907**	**0.4419**
XGBoost	**96.62**	**0.8924**	**0.4480**
SVC	93.92	0.6448	0.0749
Gaussian Naive Bayes	94.27	0.8840	0.3275

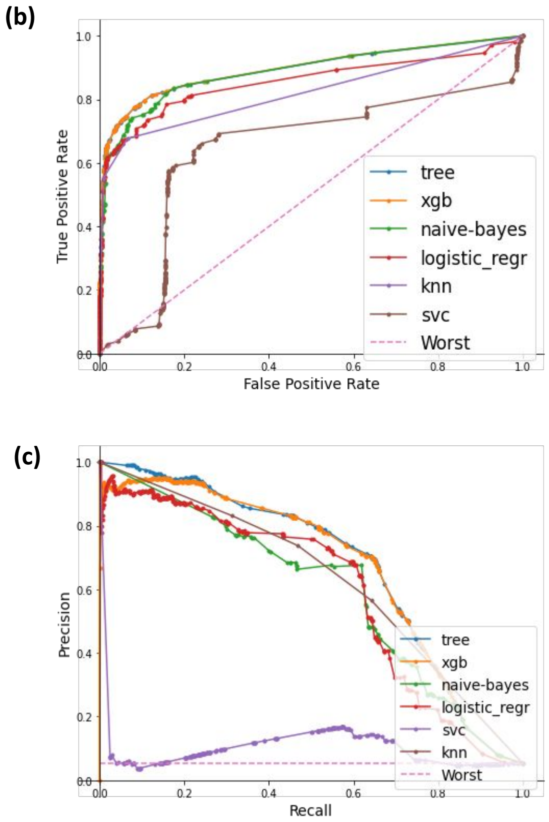


Figure 7. (a) Table of symptom classification models, (b) ROC, (c) Precision-Recall of candidate models, with XGBoost as top performer.

All images were generated by code I wrote.

Results

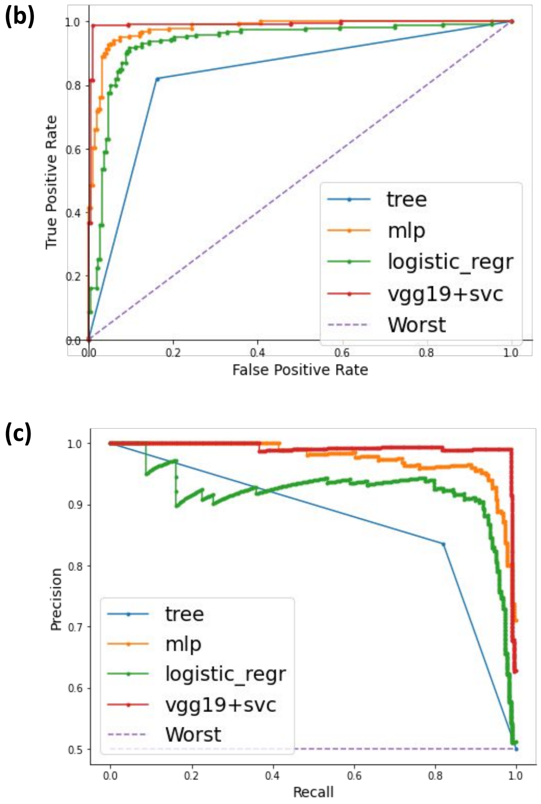
Cough Detection

- **Task:** Given a cough, identify if the cough is COVID-positive or negative
- Spectrogram-VGG19-SVC framework outperformed baselines, with high accuracy of **98.84%**

(a)

Model	Accuracy	ROC AUC	Avg Precision
Decision Tree	82.94	0.8294	0.7753
Logistic Regression	88.94	0.9355	0.8335
MLP	93.54	0.9778	0.9022
VGG19+SVC	**0.9884**	**0.9909**	**0.9840**

Figure 8. (a) Table of symptom classification models, (b) ROC, (c) Precision-Recall of candidate models, with VGG19+SVC as top performer.



All images were generated by code I wrote.

Results - Model Interpretability

XGBoost Model Interpretability

- Kernel SHapley Additive exPlanations (SHAP)
  - A permutation-based explainability method that measures the impact of features across the dataset.
  - Plot ranks features by overall importance (y-axis) and arranges data instances as points along the the x-axis by the impact the feature had on prediction
  - Fever and contact identified as having the largest impacts on prediction

VGG19 + SVC Model Interpretability

- Deep SHapley Additive exPlanations (SHAP)
  - Regions of pixels that contributed to COVID-19 predictions = red, and blue = healthy predictions.
  - Identifies regions of cough instrumental for COVID-positive cough classification

All images were generated by code I wrote

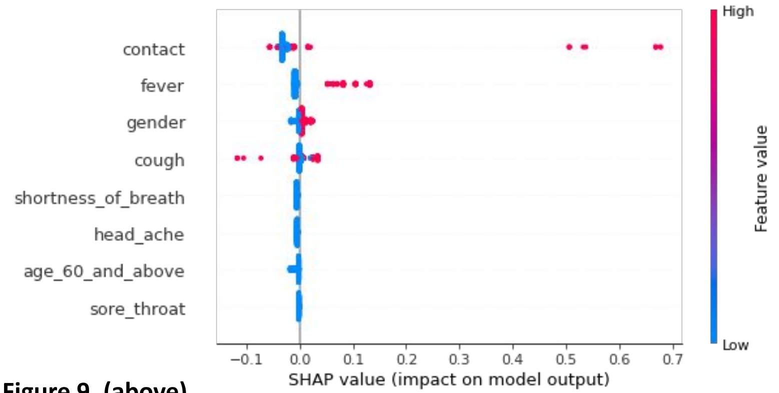


Figure 9. (above)
Kernel SHAP plot for the XGBoost symptom model reflecting the large impacts positive contact and fever have on prediction.

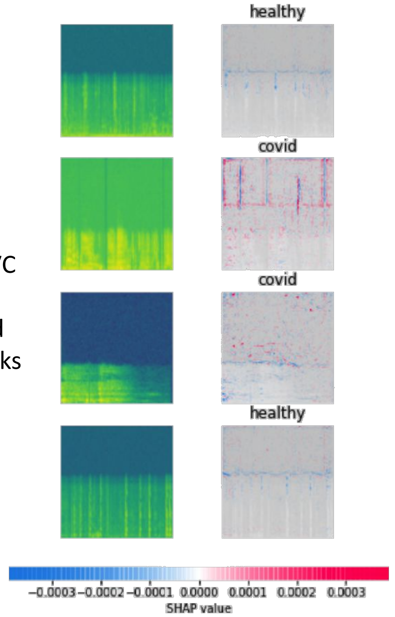
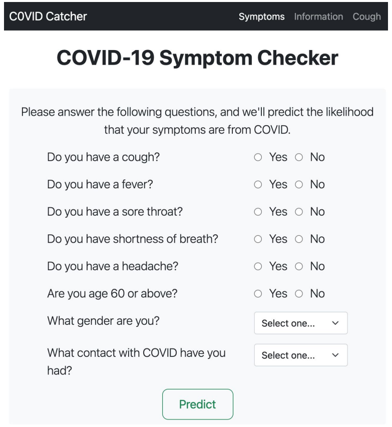


Figure 10. (right)
Deep SHAP plot for the VGG19 + SVC cough classifier. The plot show the model not only examining expected regions of the spectrogram like peaks and valleys, but also regions not immediately visible to human eye.

Findings

Beta-testing and improving results

- A **survey** was conducted to beta-testers to better understand limitations and iterate
- Feedback
  - "This is something that I would ***use every week*** or if I'm ***feeling sick***"
  - "COVID-Catcher is ***creative*** and ***intuitive*** to use. Saves me money and time, and ***reduces transmission risk*** of me going outside"
  - "I have ***peace of mind*** in checking my elderly parents' symptoms with a ***few simple clicks***, without even leaving the house"
- Screenshots of **www.c0vidcatcher.org** on the right



COVID-19 Symptom Checker

Please answer the following questions, and we'll predict the likelihood that your symptoms are from COVID.

Do you have a cough? ○ Yes ○ No

Do you have a fever? ○ Yes ○ No

Do you have a sore throat? ○ Yes ○ No

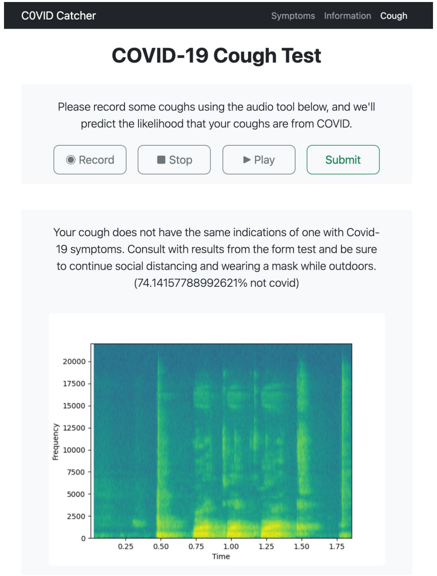
Do you have shortness of breath? ○ Yes ○ No

Do you have a headache? ○ Yes ○ No

Are you age 60 or above? ○ Yes ○ No

What gender are you? Select one...

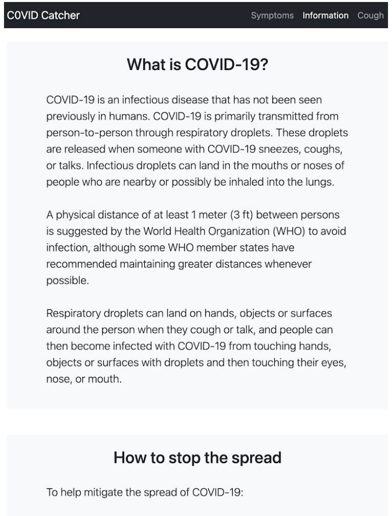
What contact with COVID have you had? Select one...



COVID-19 Cough Test

Please record some coughs using the audio tool below, and we'll predict the likelihood that your coughs are from COVID.

Your cough does not have the same indications of one with Covid-19 symptoms. Consult with results from the form test and be sure to continue social distancing and wearing a mask while outdoors. (74.14157788992621% not covid)



What is COVID-19?

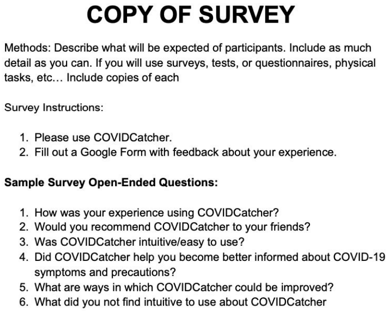
COVID-19 is an infectious disease that has not been seen previously in humans. COVID-19 is primarily transmitted from person-to-person through respiratory droplets. These droplets are released when someone with COVID-19 sneezes, coughs, or talks. Infectious droplets can land in the mouths or noses of people who are nearby or possibly be inhaled into the lungs.

A physical distance of at least 1 meter (3 ft) between persons is suggested by the World Health Organization (WHO) to avoid infection, although some WHO member states have recommended maintaining greater distances whenever possible.

Respiratory droplets can land on hands, objects or surfaces around the person when they cough or talk, and people can then become infected with COVID-19 from touching hands, objects or surfaces with droplets and then touching their eyes, nose, or mouth.

How to stop the spread

To help mitigate the spread of COVID-19:



COPY OF SURVEY

Methods: Describe what will be expected of participants. Include as much detail as you can. If you will use surveys, tests, or questionnaires, physical tasks, etc… Include copies of each

Survey Instructions:

1. Please use COVIDCatcher.
2. Fill out a Google Form with feedback about your experience.

Sample Survey Open-Ended Questions:

1. How was your experience using COVIDCatcher?
2. Would you recommend COVIDCatcher to your friends?
3. Was COVIDCatcher intuitive/easy to use?
4. Did COVIDCatcher help you become better informed about COVID-19 symptoms and precautions?
5. What are ways in which COVIDCatcher could be improved?
6. What did you not find intuitive to use about COVIDCatcher

Conclusions

Direct Biomedical Applications

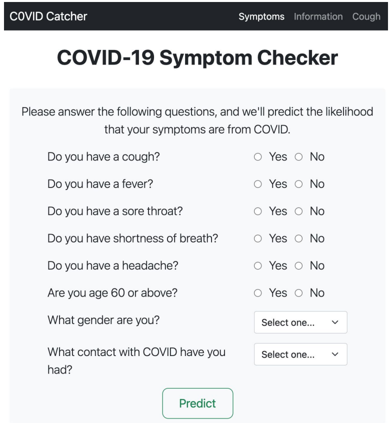
- **A novel diagnostic that is free and scalable for elderly and immunocompromised people worldwide:**
  - Due to its <u>low-cost</u> and <u>scalability</u> as a software solution, COVIDCatcher can assist the elderly and immunocompromised globally with *no user costs* to understand their health symptoms via models informed by patient datasets.
- **Assist doctors and nurses in triaging COVID-19 patients:**
  - As more privacy-approved COVID <u>symptom datasets</u> are collected and released to the public, COVIDCatcher can continue to improve and become useful as a tool to <u>assist doctors and nurses</u> to <u>quickly triage COVID-19 patients</u>.

Limitations

- Some audio files in the dataset had background noise, which could create false positives
- Microphone quality and audio quality may skew results
- Lack of new data for cough detection; limited # of open-source datasets
- Israeli dataset may not represent of U.S. population; no large scale U.S. data collection + dataset for COVID-19

Conclusion

1. In order to protect **high-risk elderly** and **immunocompromised** people, I developed a **low-cost multimodal** machine learning based app for detecting **COVID-19** symptoms.
2. COVIDCatcher employs **XGBoost** to identify COVID-19 symptoms and a custom **Spectrogram+SVC+VGG** framework to detect COVID-19 coughs.
3. XGBoost detects COVID-19 symptoms with **96.62%** accuracy, and SVC+VGG detects COVID coughs with **98.84%** accuracy
4. To date, COVIDCatcher is the **first app** that uses a **multimodal**, **data-driven** approach to evaluate COVID-19 symptoms.
5. **COVIDCatcher** is simple to use and scalable to the public at large, deployed to use on both mobile and computer browsers. Results take less than a minute, and can be used at https://www.c0vidcatcher.org



COVID-19 Symptom Checker

Please answer the following questions, and we'll predict the likelihood that your symptoms are from COVID.

Do you have a cough? ○ Yes ○ No

Do you have a fever? ○ Yes ○ No

Do you have a sore throat? ○ Yes ○ No

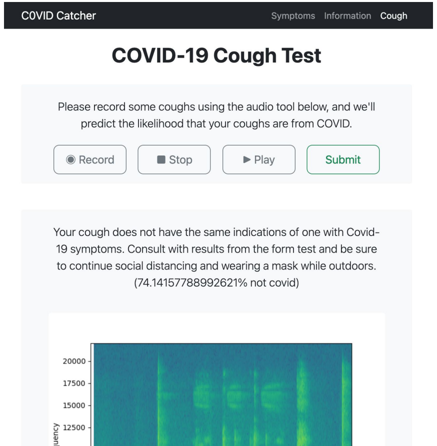
Do you have shortness of breath? ○ Yes ○ No

Do you have a headache? ○ Yes ○ No

Are you age 60 or above? ○ Yes ○ No

What gender are you? Select one...

What contact with COVID have you had? Select one...



COVID-19 Cough Test

Please record some coughs using the audio tool below, and we'll predict the likelihood that your coughs are from COVID.

Your cough does not have the same indications of one with Covid-19 symptoms. Consult with results from the form test and be sure to continue social distancing and wearing a mask while outdoors. (74.14157788992621% not covid)

References

[1] Liu G., Rusling F. J., "COVID-19 Antibody Tests and Their Limitations" 2021

[2] CDC Covid-19 Health Bot, https://github.com/CDCgov/covid19healthbot 2020.

[3] Zoabi Y et al., "Machine learning-based prediction of COVID-19 diagnosis based on symptoms." 2021.

[4] Pahar, M. et al., "COVID-19 cough classification using machine learning and global Smartphone recordings" 2020.

[5] Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System." 2016

[6] Ferguson et al., "Automatic localization of casting defects with convolutional neural networks." 2017