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

Michael Li, Amador Valley High School, Pleasanton, CA

Abstract

New at-home COVID-19 tests are expensive, and traditional tests require leaving the safety of one's home, which presents a **danger** to the elderly and immunocompromised. There is a **strong** need for an **easy, quick, and cost-effective** way to understand if a person has **COVID-19 symptoms**. I developed a **COVID-Catcher**: a multimodal, low-cost, machine learning based app that can detect COVID-19 symptoms. For **symptom detection**, a training data set of 2.7 million patients was processed, and several machine learning models were built and compared based on accuracy, recall, and precision. For **cough detection**, a training dataset of ~1445 coughs was processed and used to design a COVID-19 cough detection workflow. The top performing models were selected for use in COVIDCatcher, in which COVID-19 symptoms are detected using XGBoost, and COVID-19 coughs are identified by a spectrogram, VGG, and a support vector machine. To make these models accessible to the public, I built a **web app** and deployed both models for users to check for COVID-19 symptoms and learn about COVID-19 by inputting symptoms. **Beta-testing** of COVIDCatcher showed that users found the app easy-to-use and informative. To date, this is the first app that uses a **multimodal, data-driven** approach to evaluate COVID-19 symptoms.

Objective

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

Background

- **54.6 million elderly** and **10 million immunocompromised people** in the U.S.
  - **In-person** tests present **risk** of exposure for immunocompromised and elderly
- **At-home** COVID-19 tests are **expensive** (>$100) and limited in quantity
- The **CDC**'s Coronavirus Self-Checker has simple rule-based logic, rather than a data-driven framework
- Prior work (Zoabi et al., Pahar et al., Ahamad et al.) shows the theoretical potential of machine learning in detecting COVID-19 symptoms, but a **real**, **human-usable** and **data-driven** application has yet to be researched and developed

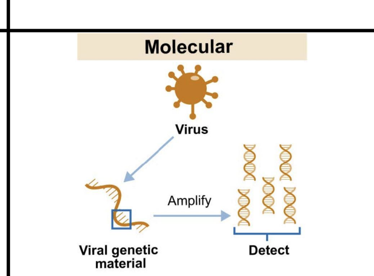
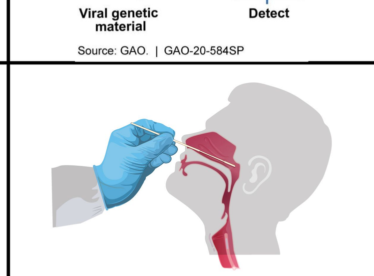
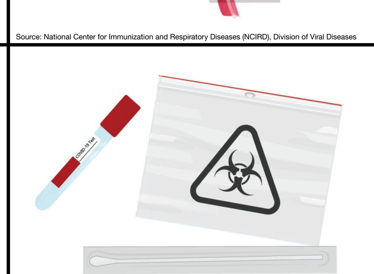
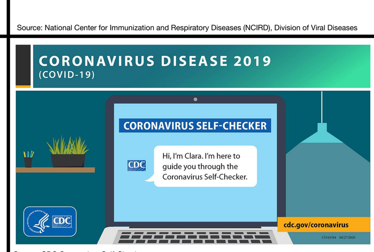
COVID-19 Diagnostics		Advantages	Limitations
Molecular Test (detects piece of viral DNA through PCR testing.)		Free to public, accuracy level of 94% https://www.medrxiv.org/content/10.1101/2020.04.05.20053355v1.full.pdf	Risk of exposure when outside home, need to wait 2-3 days for results, long lines, only a few authorized for at home use.
Antigen test (detects proteins from a virus particle, generally through a nasal swab or nasopharyngeal swab)		Takes within minutes for results, and most are authorized for at home use.	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)		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		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

Materials and Methods

Model selection and comparison. Multiple 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.

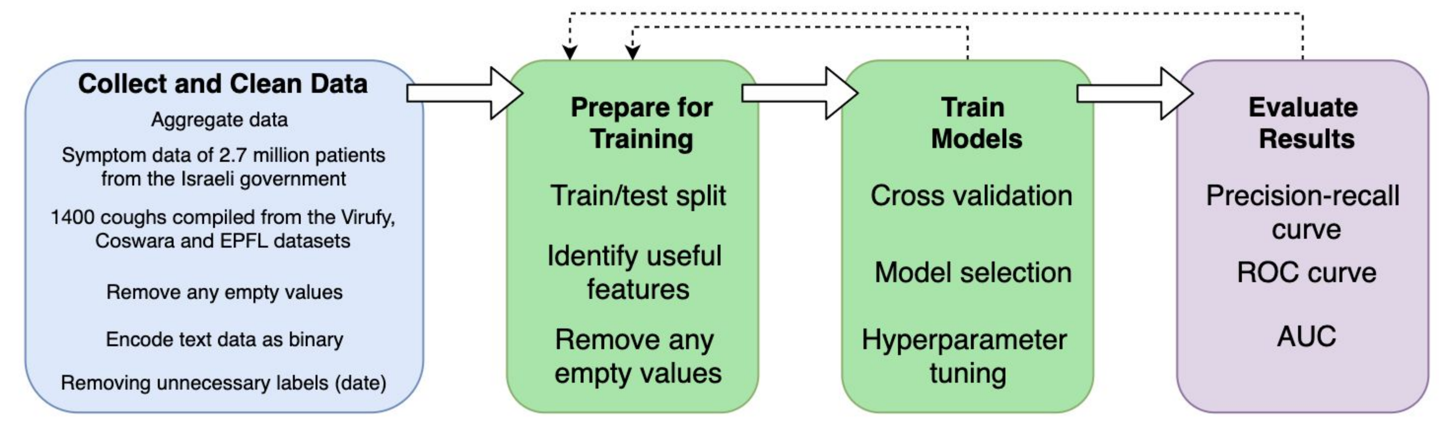
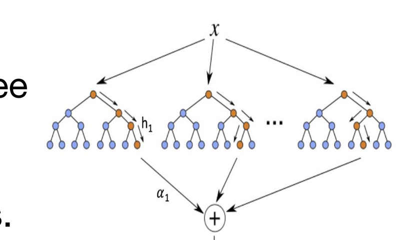
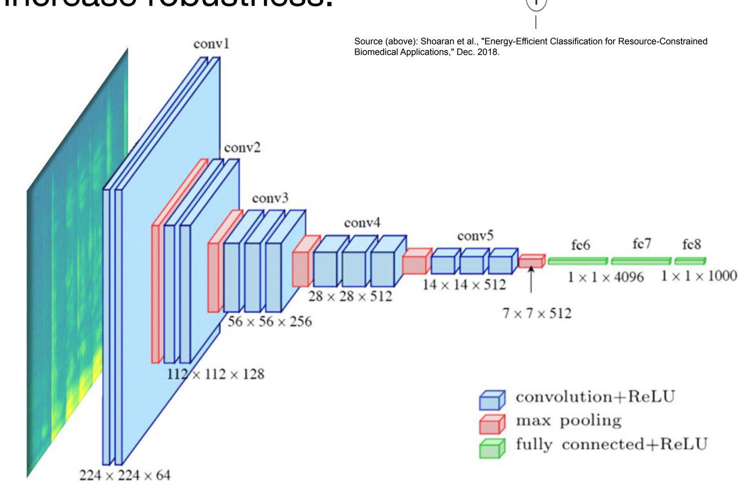


Figure 2. Machine learning model development workflow (above).

XGBoost. Gradient boosted decision tree model that uses multiple trees to increase robustness.



VGG19. A variation of VGG (state of the art convolutional neural network) with 19 layers used to create a feature map from coughs.



Web App. Models were saved via Pickle and loaded to a webapp in Heroku with remote hosting.

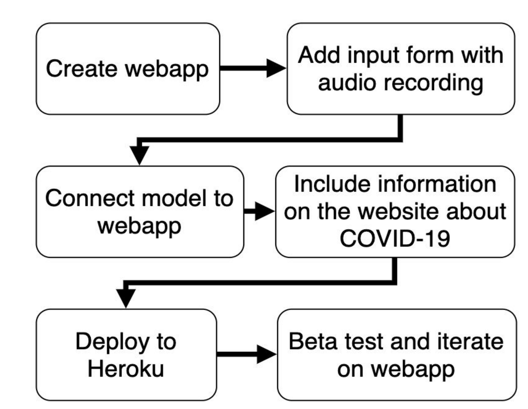
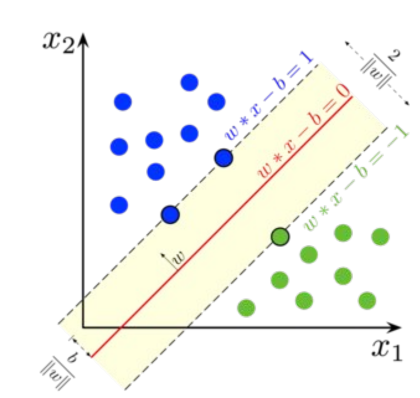


Figure 3. App development flow (above).

Linear SVM. Finds the hyperplane with best margin of separation for binary classification, used for cough classification.



Results and Interpretations

XGBoost showed top performance for COVID-19 symptom detection, with 96.62% accuracy

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

Table 1. Model comparisons for COVID-19 symptom detection.

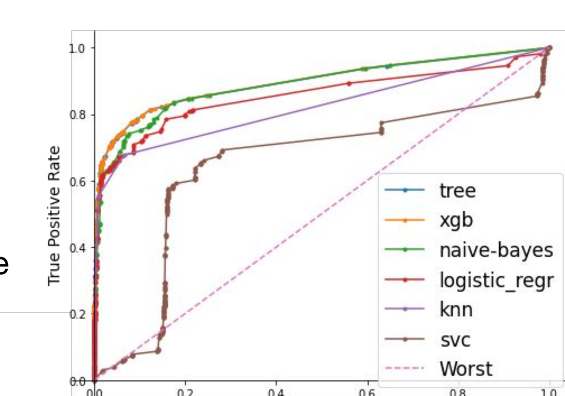


Figure 4. (right) ROC of candidate models showing XGBoost and Decision Tree's high performance

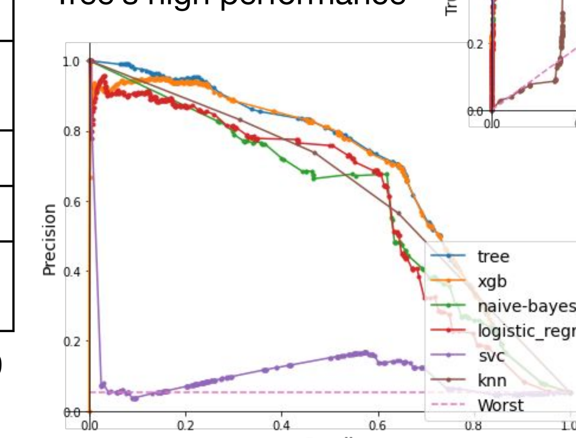


Figure 5. (left) Precision-Recall of candidate models showing XGBoost's and Decision Tree as top performers

VGG19-SVM achieved high performance for COVID-19 cough detection, with 98.84% accuracy

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**

Table 2. Model comparisons for COVID-19 cough detection.

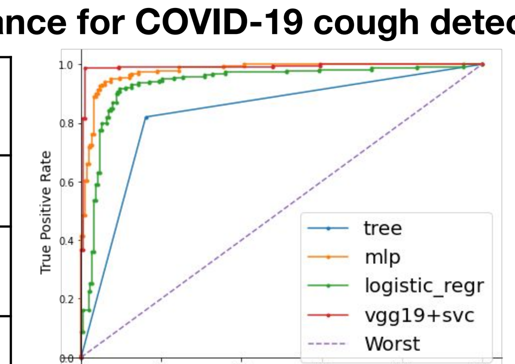


Figure 6. ROC of candidate models showing VGG19+SVC outperforming.

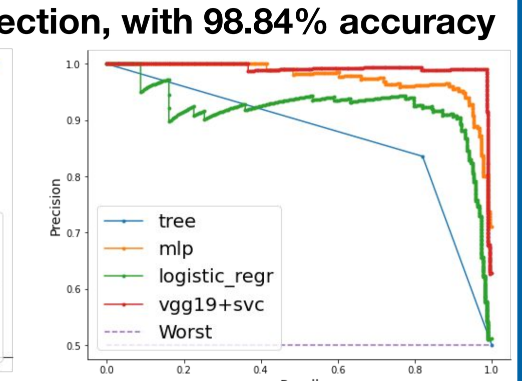
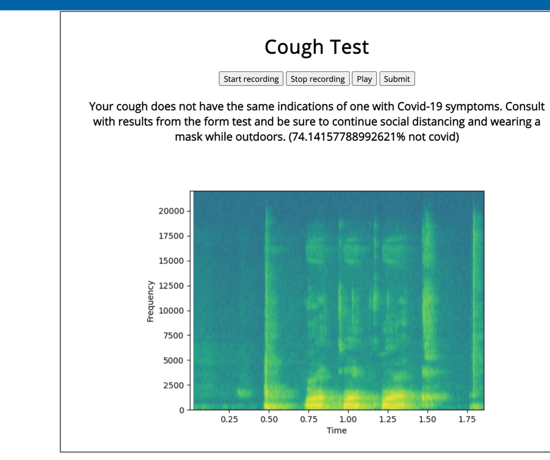
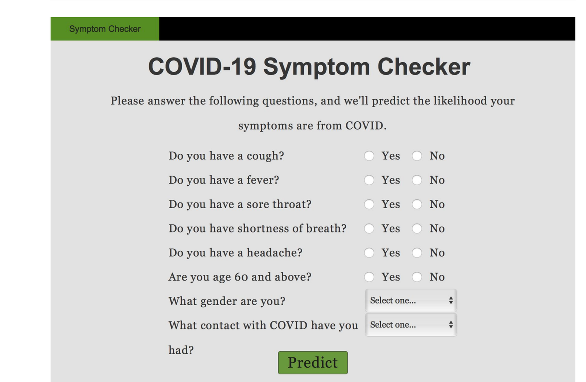


Figure 7. Precision-Recall of candidate models where VGG19+SVC performs best.

Results and Interpretations (cont.)

- **COVIDCatcher**'s website and backend were designed to incorporate the top performing models: **XGBoost** and **VGG-SVM** were successfully deployed
- A survey was conducted to beta-testers to better understand limitations and iterate
  - "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"





Conclusion and Significance

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 **SVC+VGG** to detect COVID-19 coughs.
3. To date, COVIDCatcher is the first app that uses a **multimodal**, **data-driven** approach to evaluate COVID-19 symptoms.
4. COVIDCatcher is **simple to use** and **scalable** to the public at large. Results take less than a minute, and can be used at https://www.c0vidcatcher.org

Relevant Applications to Biotechnology

1. **A novel diagnostic that is free and scalable for elderly and immunocompromised people worldwide:** Due to its low-cost and scalability 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.
2. **A quick and easy-to-use supplement for existing at-home health diagnoses:** COVIDCatcher is easy to use and can be incorporated into existing flows of at-home COVID-19 tests to quickly provide further information to those concerned about symptoms, without a long wait time. A simple user interface and quick results in <1 minute ensures anyone can use it efficiently.
3. **Assist doctors and nurses in triaging COVID-19 patients:** As more privacy-approved COVID symptom datasets are collected and released to the public, COVIDCatcher can continue to improve and become useful as a tool to assist doctors and nurses to quickly triage COVID-19 patients.

Acknowledgements

I would like to thank my mentor Emily Wang for giving me constructive feedback on my project, all the volunteers who helped to beta-test COVIDCatcher, and the following organizations for letting me use their datasets: Virufy, Coswara, EPFL, AudioSet Google and the Israeli Ministry of Health.