

COVID-19 Outbreak Prediction and Analysis using Self Reported Symptoms

Rohan Sukumaran^{*1}, Parth Patwa^{*1}, Sethuraman T V^{*1}, Sheshank Shankar¹,
Rishank Kanaparti¹, Joseph Bae^{1,2}, Yash Mathur¹, Abhishek Singh⁴, Ayush
Chopra⁴, Myungsun Kang¹, Priya Ramaswamy^{1,3}, and Ramesh Raskar^{1,4}

¹ PathCheck Foundation

rohan.sukumaran@pathcheck.org, parth.patwa@pathcheck.org
sethu.ramantv@pathcheck.org

² Stony Brook Medicine

³ University of California San Francisco

⁴ MIT Media Lab

Abstract. It is crucial for policymakers to understand the community prevalence of COVID-19 so combative resources can be effectively allocated and prioritized during the COVID-19 pandemic. Traditionally, community prevalence has been assessed through diagnostic and antibody testing data. However, despite the increasing availability of COVID-19 testing, the required level has not been met in parts of the globe, introducing a need for an alternative method for communities to determine disease prevalence. This is further complicated by the observation that COVID-19 prevalence and spread vary across different spatial, temporal, and demographic verticals. In this study, we study trends in the spread of COVID-19 by utilizing the results of self-reported COVID-19 symptoms surveys as a complement to COVID-19 testing reports. This allows us to assess community disease prevalence, even in areas with low COVID-19 testing ability. Using individually reported symptom data from various populations, our method predicts the likely percentage of the population that tested positive for COVID-19. We achieved a mean absolute error (MAE) of 1.14 and mean relative error (MRE) of 60.40% with 95% confidence interval as [60.12, 60.67]. This implies that our model predicts +/- 1140 cases than the original in a population of 1 million. In addition, we forecast the location-wise percentage of the population testing positive for the next 30 days using self-reported symptoms data from previous days. The MAE for this method is as low as 0.15 (MRE of 11.28% with 95% confidence interval [10.9, 11.6]) for New York. We present an analysis of these results, exposing various clinical attributes of interest across different demographics. Lastly, we qualitatively analyze how various policy enactments (testing, curfew) affect the prevalence of COVID-19 in a community.

* Equal contribution.

Keywords: Machine Learning · COVID-19 · Outbreak Prediction · Time Series

1 Introduction

The rapid progression of the COVID-19 pandemic has provoked large-scale data collection efforts on an international level to study the epidemiology of the virus and inform policies. Various studies have been undertaken to predict the spread, severity, and unique characteristics of the COVID-19 infection, across a broad range of clinical, imaging, and population-level datasets (Gostic, Gomez, Mummah, Kucharski, & Lloyd-Smith, 2020; Liang et al., 2020; Menni et al., 2020; Shi et al., 2020). For instance, Menni et al. (2020) use self-reported data from a mobile app to predict a positive COVID-19 test result based upon symptom presentation. Anosmia was shown to be the strongest predictor of disease presence, and a model for disease detection using symptoms-based predictors was indicated to have a sensitivity of about 65%. Studies like Parma et al. (2020) have shown that ageusia and anosmia are widespread sequelae of COVID-19 pathogenesis. From the onset of COVID-19, there also has been a significant amount of work in mathematical modeling to understand the outbreak under different situations for different demographics (Menni et al., 2020; Saad-Roy et al., 2020; Wilder, Mina, & Tambe, 2020). However, these works primarily focus on the population level. Further, the estimation of different transition probabilities to move between compartments is challenging.

Carnegie Mellon University (CMU) and the University of Maryland (UMD) have built chronologically aggregated datasets of self-reported COVID-19 symptoms by conducting surveys at national and international levels (Delphi group, 2020; Fan et al., 2020). The surveys contain questions regarding whether the respondent has experienced several of the common symptoms of COVID-19 (e.g. anosmia, ageusia, cough, etc.) in addition to various behavioral questions concerning the number of trips a respondent has taken outdoors and whether they have received a COVID-19 test.

In this work, we perform several studies using the CMU (Delphi group, 2020), UMD (Fan et al., 2020), and OxCGRT (Hale, Webster, Petherick, Phillips, & Kira, 2020) datasets. Our experiments examine correlations among variables in the CMU data to determine which symptoms and behaviors are most correlated to high percentages of Covid Like Illness (CLI). We investigate how the different symptoms impact the percentage of populations with CLI across different spatio-temporal and demographic (age, gender) settings. We also predict the percentage of population who got tested positive for COVID-19 and achieve 60% Mean Relative Error. Further, our experiments involve time-series analysis of these datasets to forecast CLI over time. Here, we identify how different spatial window trends vary across different temporal windows. We aim to use the findings from this method to understand the possibilities of modeling CLI for geographic areas in which data collection is sparse or non-existent. Furthermore, results from our experiments can potentially guide public health policies

for COVID-19. Understanding how the disease is progressing can help the policymakers introduce non-pharmaceutical interventions (NPIs) and also help them understand how to distribute critical resources (medicines, doctors, healthcare workers, transportation, and more). This could now be done based on the insights provided by our models, instead of relying completely on clinical testing data. Prediction of outbreaks using self-reported symptoms can also help reduce the load on testing resources. Similar self reported data and survey data have been used by (Rodriguez, Muralidhar, et al., 2020; Rodriguez, Tabassum, et al., 2020; Garcia-Agundez et al., 2021) for understanding the pandemic and drawing actionable insights.

2 Datasets

The **CMU Symptom Survey** aggregates the results of a survey run by CMU (Delphi group, 2020) that was distributed across the US to approx 70k random Facebook users daily. It gives a set of indicators that can inform our reasoning about the pandemic. The indicators include:

- Symptoms related indicators like the percentage of respondents reporting fever and the percentage of respondents reporting sore throat.
- Pre-existing medical condition related indicators like the percentage of respondents having diabetes and the percentage of respondents having Autoimmune Disorder.
- Behavior related indicators like the percentage of respondents who avoid contact with others most of the time and the percentage of respondents who worked outside home.

The data set has a total of 104 columns (in October 2020), including weighted (adjusted for sampling bias), unweighted signals, and demographic information (age, gender, etc.) at county and state level. In this study, we use the state level data from Apr. 4, 2020 to Sep. 11, 2020, which is henceforth referred to as the CMU dataset in the paper.

The **UMD Global Symptom Survey** aggregates the results of a survey conducted by UMD through Facebook (Fan et al., 2020). The survey is available in 56 languages. A representative sample of Facebook users were invited on a daily basis to report on topics including symptoms and social distancing behavior. Facebook provides weights to reduce non-response and coverage bias. Country and region-level statistics are published daily via the public API and dashboards, and micro-data is available for researchers via data use agreements. Over half a million responses were collected daily. We use the data of 968 regions, available from May 1 to September 11, 2020. There are 49 (in October 2020) unweighted signals, as well as their weighted forms (adjusted for sampling bias).

The **Oxford COVID-19 Government Response Tracker (OxCGRT)** (Hale et al., 2020) contains government COVID-19 policy data as a numerical scale value representing the extent of government action. OxCGRT collects publicly available information on 20 indicators of government response. This

information was collected by a team of over 200 volunteers from the Oxford community and was updated continuously. The data set also includes statistics on the number of reported Covid-19 cases and deaths in each country, which were taken from the JHU CSSE (Dong, Du, & Gardner, 2020) data repository for all countries and the US.

The **Prevalence of Self-Reported Obesity by State and Territory, BRFSS, 2019 - CDC** (CDC, 2020) is a dataset published by CDC containing the aggregated self-reported obesity values. The data are at the state level and contain the obesity values and confidence intervals (95%). This dataset contains other information like race, ethnicity, and food habits that can be used for further analysis.

3 Methods and Experiments

Different methods and strategies have been used to analyze the data. Our code used in the analysis is publicly available at <https://github.com/PrivateKit/CovidSymptomChallenge>.

3.1 Correlation Studies

Correlations between features of the datasets provide crucial information about the features and the degree of influence they have over the target value. We conducted correlation analysis on different subgroups like symptomatic and asymptomatic subjects, and varying demographic regions in the CMU dataset to discover relationships among the signals and with the target variable. We also investigated the significance of obesity and population density on the susceptibility to COVID-19 at the state level (CDC, 2020). Refer to the supplementary materials for more information.

3.2 Feature Pruning

We first dropped demographic features such as date, gender, and age. Next, we dropped the unweighted features because their weighted counterparts were used. We also dropped features including the percentage of people who tested negative, the weighted percentage of people who tested positive because they were directly related to testing and would make the prediction trivial. Furthermore, we dropped the derived features such as the estimated percentage of people with influenza-like illness because they were not directly reported by the respondents. Finally, we dropped some features with aggregated information such as the average number of people in respondent's household who have Covid Like Illness. After the entire process, we selected 36 features. The selected feature list is provided in the supplementary materials.

3.3 Outbreak Prediction

We predicted the percentage of the population that tested positive at the state level from the CMU dataset. We ranked these 36 signals using *f*-regression (“sklearn f regression”, 2007-2020) (*f*-statistic of the correlation to the target variable) and predicted the target variable using the top *n* ranked features. We experimented with the *top n features* value from 1 to 36 for various demographic groups. We trained linear regression (Galton, 1886), decision tree (Quinlan, 1986), and gradient boosting (Friedman, 2001) models. All the models were implemented using scikit-learn (Pedregosa et al., 2011). We used 80% of the data for training and the remaining 20% of the data for testing. The data were split randomly.

3.4 Time Series Analysis

We predicted the percentage of people that tested positive using the CMU dataset and the percentage of people with CLI with the UMD dataset. We independently used the top “n” features (according to their ranking obtained in outbreak prediction and empirical evidence combined with human experts ranking) from the CMU (36) and UMD (49) datasets for multivariate multi-step time series forecasting. Given the data spread across different spatial windows (geographies) at the state level, we employed an agglomerative clustering method independently on symptoms and behavioral/external patterns, and sample locations that were not in the same cluster for our analysis. Using the Augmented Dickey-Fuller test (Cheung & Lai, 1995), we found the time series samples for these spatial windows to be stationary. Furthermore, we bucketed the data based on the age and gender of the respondents, to provide granular insights on the model performance on various demographics. With a total of 12 demographic buckets [(age, gender) pairs], we used a vector autoregressive (VAR) (Holden, 1995) model and an LSTM (Gers, Schmidhuber, & Cummins, 1999) model for the experiments. Furthermore, we qualitatively evaluated the impact of government policies, e.g., curfew, on the spread of the virus. We used 80% of the data for training and the remaining 20% of the data for testing.

4 Results and Discussion

4.1 Correlation Studies

The state level analysis revealed a moderate positive correlation, $r = 0.24$ (p -value < 0.001), between the percentage of people tested positive and the statewide obesity level. Here, the obesity is defined as $BMI > 30.0$ (NIH, 2020). The results are consistent with prior clinical studies like (Chan et al., 2020) and indicate that further research is required to investigate if the lack of certain nutrients like Vitamin B, Zinc, Iron, or having a $BMI > 30.0$ could make an individual more susceptible to COVID-19. Figure 1 shows the correlations among multiple self-reported symptoms and the symptoms with significant positive correlations

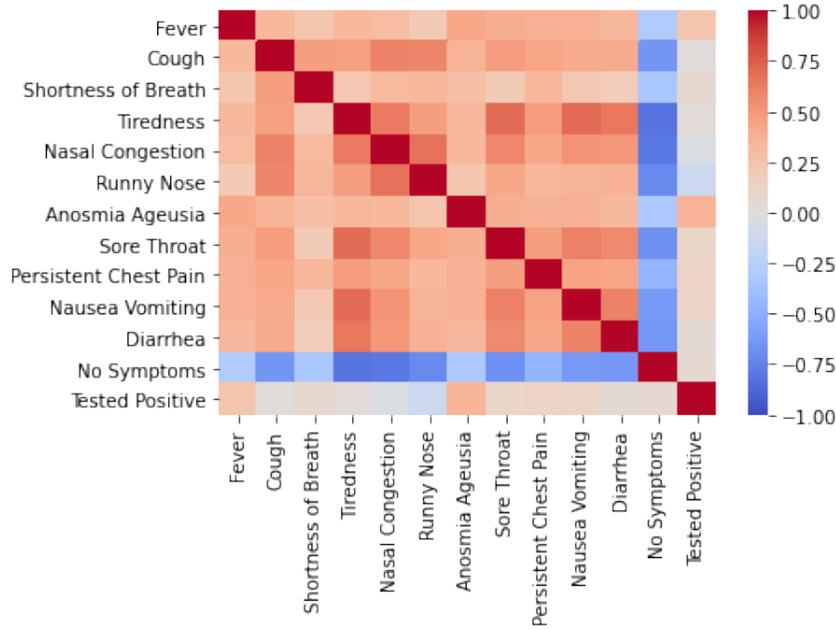


Figure 1. Correlation among self-reported symptoms and the percent of population tested COVID positive.

are highlighted. This clearly reveals that anosmia, ageusia and fever are relatively strong indicators of COVID-19. From Figure 2, we see that contact with a COVID-19 positive individual is strongly correlated with testing COVID-19 positive. Conversely, the percentage of population who avoid outside contact and the percentage of population testing positive for COVID-19 have a negative correlation. We also found a moderate positive correlation between the population density and the percentage of population reporting positive COVID-19, which indicates easier transmission of the virus in a congested environment. These observations reaffirm the highly contagious nature of the virus and the need for social distancing.

The results motivated us to estimate the percentage of people who tested COVID-19 positive based on the percentage of people who had a direct contact with anyone who recently tested positive. In doing so, we achieve a mean relative error (MRE) of 2.33% and a mean absolute error (MAE) of 0.03.

Here, MAE is the absolute value of the difference between the predicted value and the actual value, averaged over all data points:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - x_i|,$$

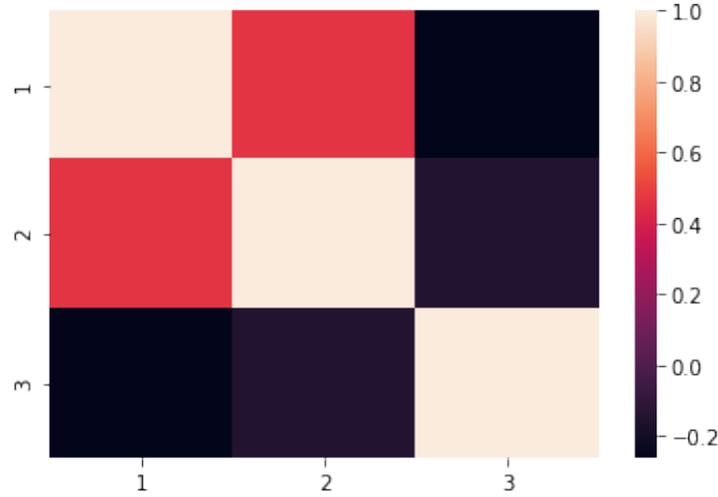


Figure 2. Correlations between the percent of people having contact with someone having CLI and the percent of people who tested positive. Here, the attribute (1) = percentage of people who had contact with someone having COVID-19, (2) = percentage of people tested positive, (3) = percentage of people who avoided contact all/most of the time.

where n is the total data instances, y_i is the predicted value and x_i is the actual value. Relative error is the absolute difference between the predicted value and the actual value, divided by the actual value. MAE is the relative error averaged over all the data points:

$$\text{MRE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - x_i}{x_i + 1} \right|,$$

where 1 is added in the denominator to avoid division by 0.

We found that a low MAE value can be misleading in the case of predicting the spread of the virus. The MAE for the outbreak prediction was low and had a small range (1-1.4) but more than 75% of the target lied between 0-2.6, meaning only a small percentage of the entire population had COVID-19 (if 1% of the entire population was affected and an MAE of 1 indicates the predicted cases could double the actual cases). MRE accounts for even minute changes (errors) in the prediction. Hence, it is a better metric to judge a system.

4.2 Policies vs CLI/Community Sick Impacts

The impacts of different non-pharmaceutical interventions (NPIs) could be analyzed by combining the CMU, UMD, and Oxford data. A particular analysis

from that is reported here, where we noticed that lifting of stay at home restrictions resulted in a sudden spike in the number of cases. This is visualized in figure 3.

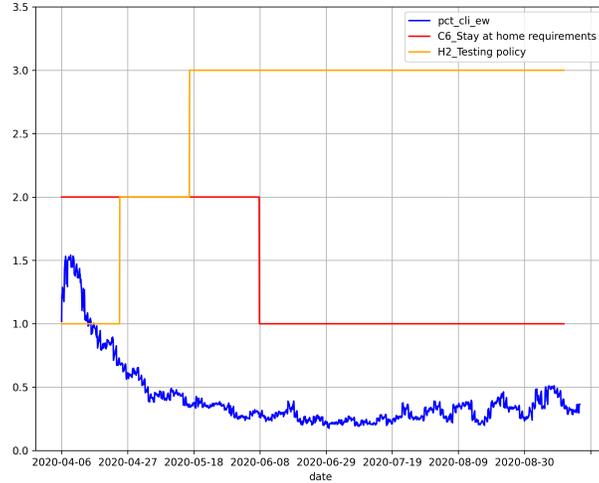


Figure 3. Policy impacts: When Stay at home restrictions were stronger, even with higher testing rates, the percentage of population with CLI (pct_cli_ew) had a downward trend.

4.3 Outbreak prediction on CMU Dataset

Gradient boosting performed the best and considerably better than the next best algorithm in terms of the error metrics for every demographic group. Hence, only the results for Gradient Boosting are presented. Table 1 shows the best accuracy achieved per dataset. For every dataset, the best "n" number of features is about 30. We achieved an MRE of 60.40% for the entire dataset. The performance was better on the female-only data when compared to the male-only data. The performance was slightly better on 55+ age data than other age groups. This can also be observed from figure 4.

4.3.1 Top Features Except for minor reordering, the top 5 features were CLI in community, loss of smell, CLI in house hold (HH), fever in HH, and fever across every data split. The top 6 to 10 features per data split are given in figure 5. We can see that 'worked outside home' and 'avoid contact most time'

Table 1. Results of gradient boosting model for the prediction of the percentage of population tested positive across demographics. The mean relative error (MRE) and mean absolute error (MAE) are average of 20 runs. The 95% confidence interval (CI) for MRE is calculated on 20 runs (data were shuffled randomly each time).

Demographic	best n	MAE	MRE	CI
Entire	35	1.14	60.40	(60.12, 60.67)
Male	34	1.38	78.14	(77.67, 78.62)
Female	36	1.10	56.89	(56.48, 57.30)
Age 18-34	30	1.23	66.35	(65.59, 67.12)
Age 35-54	35	1.29	67.59	(67.13, 68.04)
Age 55+	33	1.20	66.40	(65.86, 66.94)

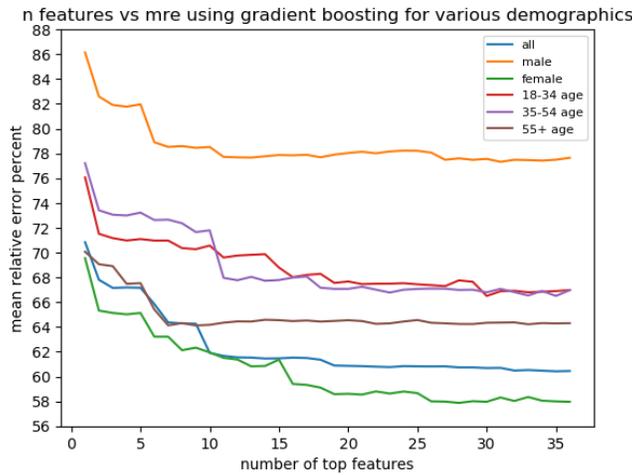


Figure 4. Error vs. the number of top features used for the gradient boosting model. Errors vary across demographics and generally decrease with the increase of the number of features (n). The decrease is not considerable after $n = 20$.

are useful features for male, female, and 55+ age groups. Figure 4 shows MRE vs. the number of features selected for different data splits. Overall, the error decreased as we added more features. However, the decrease in error was not considerable when we went beyond 20 features ($< 1\%$).

4.4 Time Series Analysis

As seen in Tables 2 and 3, we were able to forecast the PCT_CLI with an MRE of 15.31% using just 23 features from the UMD dataset for Lombardia and with an MRE of 42.72% for Northern Ireland. The 23 features (provided in the supplementary materials) were selected with the help of human experts and empirical analysis. We can see that VAR performed better than LSTM on average. This can be explained by the dearth of data available. Furthermore, we can see that

6-10 Ranked Discriminative Features by Demographic

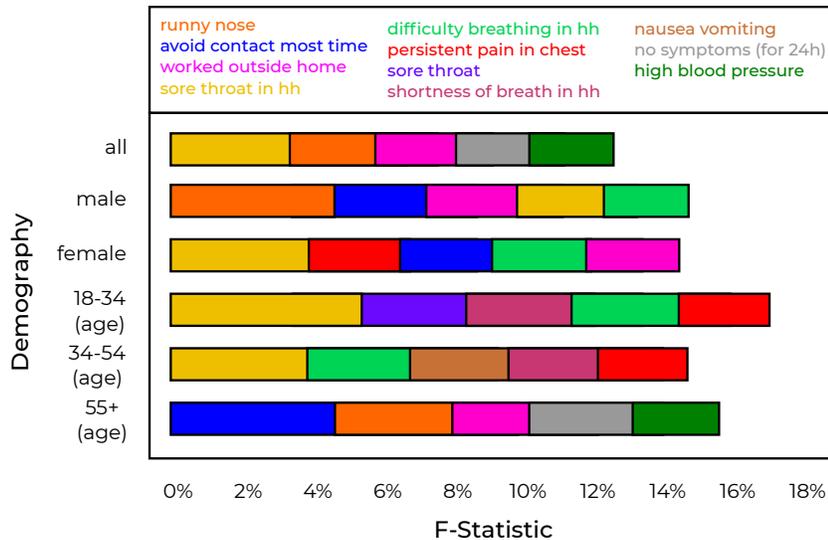


Figure 5. After the top 5 predictive features (which were roughly identical), there are considerable differences between the most predictive features segmented across demographics. For example, for the age 34-55 group, 'sore throat in hh (household)' was the sixth most predictive feature but it is not even in the top 10 most predictive features for the 55+ age group.

the outbreak forecasting for New York achieved 11.28% MRE, making use of only 10 features (these features were selected based on the outbreak prediction results and further empirically identified as well). This might be caused by an inherent bias in the sampling strategy or participant responses. For example, the high correlation noted between anosmia and COVID-19 prevalence suggested several probable causes of confounding relationships between the two. This could also occur if both symptoms were specific and sensitive for COVID-19 infection.

4.5 Symptoms vs CLI overlap

The percentage of population with symptoms like cough, fever, and runny nose was much higher than the percentage of people who suffered from CLI or the percentage of people who were sick in the community. Only 4% of the people in the UMD dataset who reported having CLI did not suffer from chest pain and nausea.

Table 2. The errors of forecasting the outbreak of COVID-19 (the percentage of people who tested positive) for the next 30 days using VAR and LSTM.

Location	<i>MRE</i>	<i>MAE</i>
VAR		
New York	11.28, 95% CI [10.9, 11.6]	0.15
California	13.48, 95% CI [13.4, 13.5]	0.23
Florida	17.49, 95% CI [17.5, 17.5]	0.38
New Jersey	17.93, 95% CI [17.9, 18]	0.26
LSTM		
New York	23.61, 95% CI [23.6, 23.7]	0.36
California	45.06, 95% CI [45, 45.2]	0.91
Florida	64.98, 95% CI [64.8, 65.1]	1.51
New Jersey	15.78, 95% CI [15.7, 15.9]	0.26

Table 3. Results of forecasting the outbreak of COVID-19 (the percentage of people with COVID-19 like illness in the population - PCT_CLI) for the next 30 days using the VAR and LSTM models.

Location	<i>MRE</i>	<i>MAE</i>
VAR		
Tokyo	17.77, 95% CI [17.7, 17.8]	0.28
British Columbia	21.35, 95% CI [21.3, 21.4]	0.34
Northern Ireland	42.72, 95% CI [42.7, 42.8]	0.87
Lombardia	15.31, 95% CI [15.3, 15.4]	0.22
LSTM		
Tokyo	30.00, 95% CI [29.9, 30.1]	0.53
British Columbia	31.11, 95% CI [30.9, 31.3]	0.56
Northern Ireland	42.46, 95% CI [42.1, 42.9]	1.21
Lombardia	16.11, 95% CI [16, 16.2]	0.21

4.6 Ablation Studies

We performed ablation studies to verify and investigate the relative importance of the features that were selected using f_regression feature ranking algorithm (“sklearn f_regression”, 2007-2020). In the following experiments, the top $N = 10$ features obtained from the f_regression algorithm are considered as the subset for evaluation.

4.6.1 All-but-one experiment In this experiment, the target variable which is the percentage of people affected by COVID-19 was estimated by considering $N - 1$ features from a given set of top N features by dropping 1 feature at a time in every iteration in descending order. The results were visualized in figure 6 from which it is clear that there was a considerable increased error when the most significant feature was dropped and the loss in performance was not as drastic when any other feature was dropped. This reaffirms our feature selection method.

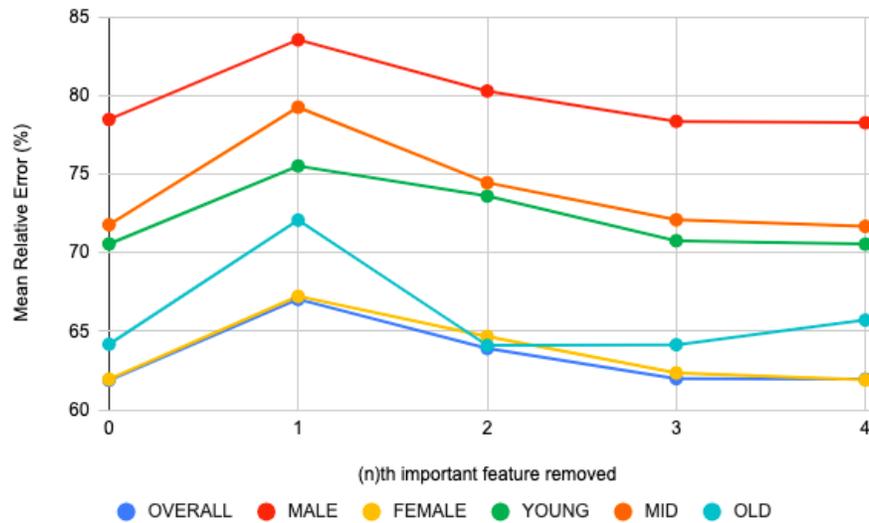


Figure 6. Results of all-but-one experiment (MRE).

4.6.2 Cumulative Feature Dropping In this experiment, we estimated the target variable based on the top $N=10$ features and then carried out the experiment with $N-i$ features in every iteration where i was the iteration count. The features were dropped in descending order. Figure 7 shows the results. The change in slope from the start to the end of the graph shows that the most important feature had a huge significance of the performance. This observation reinforces the inference of the all-but-one experiment and validates our feature selection algorithm.

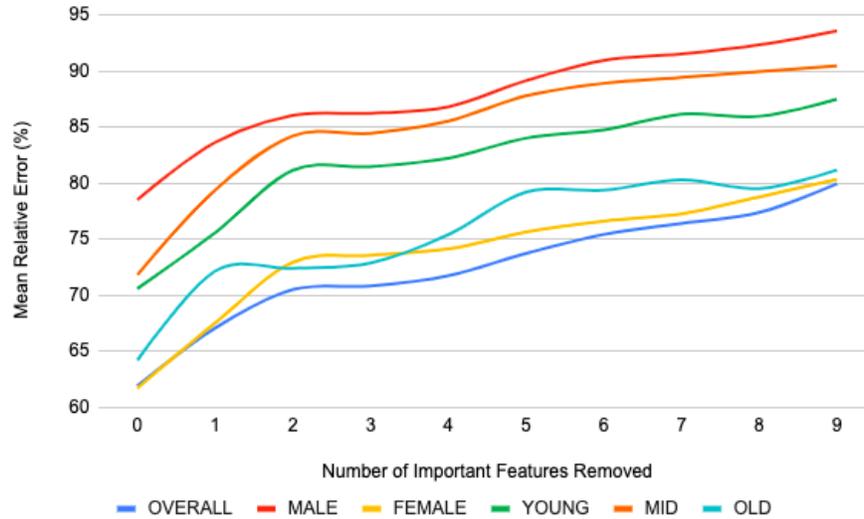


Figure 7. Results of cumulative feature dropping.

5 Conclusion And Future Work

In this work, we analyzed the benefits of the COVID-19 self-reported symptoms presented in the CMU, UMD, and Oxford datasets. We conducted correlation analysis, outbreak prediction, and time series prediction of the percentage of respondents with positive COVID-19 tests and the percentage of respondents who show COVID-like illness. By clustering datasets across different demographics, we revealed micro and macro level insights into the relationship between symptoms and outbreaks of COVID-19. These insights might form the basis for future analysis of the epidemiology and manifestations of COVID-19 in different patient populations. Our correlation and prediction studies identified a small subset of features that can predict measures of COVID-19 prevalence to a high degree of accuracy. Using this, more efficient surveys can be designed to measure only the most relevant features to predict COVID-19 outbreaks. Shorter surveys will increase the likelihood of respondent participation and decrease the chances that respondents provide false (or incorrect) information. We believe that our analysis will be valuable in shaping health policy and in COVID-19 outbreak predictions for areas with low levels of testing by providing prediction models that rely on self-reported symptom data. As shown from our results, the predictions from our models could be reliably used by health officials and policymakers, in order to prioritize resources. Furthermore, having crowd-sourced information as the base helps scale this method at a much higher pace, if and when required in the future, e.g., due to the advent of a new virus or a strain.

In the future, we plan to use advanced deep learning models for predictions. Furthermore, given the promise shown by population level symptoms data, we find more relevant and timely problems that can be solved with individual data. Machine learning systems based on data from mobile/wearable devices can be built to understand users' vitals, sleep behavior, and so on. Having the data shared at an individual level can augment the participatory surveillance dataset and thereby the predictions made. This can be achieved without compromising the privacy of individuals. We also plan to compare the reliability of such survey methods with actual number of cases in the corresponding regions and its generalizability across populations.

Acknowledgement

We thank Seojin Jang, Chirag Samal, Nilay Shrivastava, Shrikant Kanaparti, Darshan Gandhi and Priyanshi Katiyar for their inputs in various stages of this study. We further thank Prof. Manuel Morales (University de Montreal), Morteza Asgari and Hellen Vasques for helping in developing a dashboard to showcase the results. Lastly, we also thank Dr. Thomas C. Kingsley (Mayo Clinic) for his suggestions in the future works.

References

- CDC. (2020). Data and statistics [Computer software manual]. (<https://www.cdc.gov/obesity/data/prevalence-maps.html>)
- Chan, C. C., et al. (2020, Jun 02). Type i interferon sensing unlocks dormant adipocyte inflammatory potential. *Nature Communications*, 11(1). Retrieved from <https://doi.org/10.1038/s41467-020-16571-4>
- Cheung, Y.-W., & Lai, K. S. (1995). Lag order and critical values of the augmented dickey–fuller test. *Journal of Business & Economic Statistics*, 13(3), 277–280.
- Delphi group, C. M. U. (2020). *Delphi's covid-19 surveys*. Retrieved from <https://covidcast.cmu.edu/surveys.html>
- Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. , 20. doi: [https://doi.org/https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/https://doi.org/10.1016/S1473-3099(20)30120-1)
- Fan, J., et al. (2020). *Covid-19 world symptom survey data api*.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189 – 1232. Retrieved from <https://doi.org/10.1214/aos/1013203451> doi: <https://doi.org/10.1214/aos/1013203451>
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. Retrieved from <http://www.jstor.org/stable/2841583>
- Garcia-Agundez, A., Ojo, O., Hernández-Roig, H. A., Baquero, C., Frey, D., Georgiou, C., ... others (2021). Estimating the covid-19 prevalence in

- spain with indirect reporting via open surveys. *Frontiers in Public Health*, 9.
- Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with lstm. *1999 Ninth International Conference on Artificial Neural Networks ICANN 99*.
- Gostic, K., Gomez, A. C., Mummah, R. O., Kucharski, A. J., & Lloyd-Smith, J. O. (2020, February). Estimated effectiveness of symptom and risk screening to prevent the spread of covid-19. *eLife*, 9. Retrieved from <https://europepmc.org/articles/PMC7060038> doi: <https://doi.org/10.7554/elife.55570>
- Hale, T., Webster, S., Petherick, A., Phillips, T., & Kira, B. (2020). *Oxford covid-19 government response tracker blavatnik school of government*.
- Holden, K. (1995). Vector auto regression modeling and forecasting. *Journal of Forecasting*, 14(3), 159–166.
- Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., ... for the China Medical Treatment Expert Group for COVID-19 (2020, 08). Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. *JAMA Internal Medicine*, 180(8), 1081-1089. Retrieved from <https://doi.org/10.1001/jamainternmed.2020.2033> doi: <https://doi.org/10.1001/jamainternmed.2020.2033>
- Menni, C., et al. (2020). Real-time tracking of self-reported symptoms to predict potential covid-19. *Nature medicine*, 1–4.
- Menni, C., Valdes, A. M., Freidin, M. B., Sudre, C. H., Nguyen, L. H., Drew, D. A., ... Spector, T. D. (2020, Jul 01). Real-time tracking of self-reported symptoms to predict potential covid-19. *Nature Medicine*, 26(7), 1037-1040. Retrieved from <https://doi.org/10.1038/s41591-020-0916-2> doi: <https://doi.org/10.1038/s41591-020-0916-2>
- NIH. (2020). Adult body mass index (bmi) [Computer software manual]. (<https://www.nhlbi.nih.gov/health/educational/losewt/BMI/bmicalc.htm>)
- Parma, V., et al. (2020). More than smell. covid-19 is associated with severe impairment of smell, taste, and chemesthesis. *medRxiv*. Retrieved from <https://www.medrxiv.org/content/early/2020/05/24/2020.05.04.20090902> doi: <https://doi.org/10.1101/2020.05.04.20090902>
- Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Quinlan, J. R. (1986, Mar 01). Induction of decision trees. *Machine Learning*, 1(1), 81-106. Retrieved from <https://doi.org/10.1007/BF00116251> doi: <https://doi.org/10.1007/BF00116251>
- Rodriguez, A., Muralidhar, N., Adhikari, B., Tabassum, A., Ramakrishnan, N., & Prakash, B. A. (2020). Steering a historical disease forecasting model under a pandemic: Case of flu and covid-19. *arXiv preprint arXiv:2009.11407*.
- Rodriguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., ... Prakash, B. A. (2020). Deepcovid: An operational deep learning-driven frame-

- work for explainable real-time covid-19 forecasting. *medRxiv*. Retrieved from <https://www.medrxiv.org/content/early/2020/09/29/2020.09.28.20203109> doi: <https://doi.org/10.1101/2020.09.28.20203109>
- Saad-Roy, C. M., et al. (2020). Immune life history, vaccination, and the dynamics of sars-cov-2 over the next 5 years. *Science*.
- Shi, F., et al. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE Reviews in Biomedical Engineering*, 1–1. Retrieved from <http://dx.doi.org/10.1109/RBME.2020.2987975> doi: <https://doi.org/10.1109/rbme.2020.2987975>
- sklearn f regression [Computer software manual]. (2007-2020). (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html)
- Wilder, B., Mina, M. J., & Tambe, M. (2020). Tracking disease outbreaks from sparse data with bayesian inference. *arXiv preprint arXiv:2009.05863*.