COVID-19 Outbreak Prediction and Analysis using Self Reported Symptoms

Rohan Sukumaran[*1] **Parth Patwa**[*1] **Sethuraman T V**[*1] **Sheshank Shankar**[1]
Rishank Kanaparti[1] **Joseph Bae**[1,2] **Yash Mathur**[1] **Abhishek Singh**[1,4] **Ayush Chopra**[1,4]
Myungsun Kang[1] **Priya Ramaswamy**[1,3] **Ramesh Raskar**[1,4]
[1]**PathCheck Foundation** [2]**Stony Brook Medicine**
[3]**University of California San Francisco** [4]**MIT Media Lab**
{rohan.sukumaran, parth.patwa, sethu.ramantv}@pathcheck.org

Abstract

The COVID-19 pandemic has challenged scientists and policy-makers internationally to develop novel approaches to public health policy. Furthermore, it has also been observed that the prevalence and spread of COVID-19 varies across different spatial, temporal and demographics. Despite ramping up testing, we still are not at the required level in most parts of the globe. Therefore, we utilize self-reported symptoms survey data to understand trends in the spread of COVID-19. The aim of this study is to segment populations that are highly susceptible. In order to understand such populations, we perform exploratory data analysis, outbreak prediction, and time-series forecasting using public health and policy datasets. From our studies, we try to predict the likely % of population that tested positive for COVID-19 based on self-reported symptoms. Our findings reaffirm the predictive value of symptoms, such as anosmia and ageusia. And we forecast that the % of population having COVID-19-like illness (CLI) and those tested positive as 0.15% and 1.14% absolute error respectively. These findings could help aid faster development of the public health policy, particularly in areas with low levels of testing and having a greater reliance on self-reported symptoms. Our analysis sheds light on identifying clinical attributes of interest across different demographics. We also provide insights into the effects of various policy enactments on COVID-19 prevalence.

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 et al. 2020; Liang et al. 2020; Menni et al. 2020a; Shi et al. 2020). For instance, (Menni et al. 2020a) uses 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 significant amount of work in mathematical modeling to understand the outbreak under different situations for different demographics (Menni et al. 2020b; Saad-Roy et al. 2020; Wilder, Mina, and Tambe 2020). Although these works primarily focus on population level 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 (Fan et al. 2020; Delphi group 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, UMD and OxCGRT (Fan et al. 2020; Delphi group 2020; Hale et al. 2020) datasets. Our experiments examine correlations among variables in the CMU data to determine which symptoms and behaviors are most correlated to high % of CLI. We see how the different symptoms impact the % of population with CLI across different spatio-temporal and demographic (age, gender) settings. We also predict the % 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 modelling 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.

Using self reported symptoms collected across spatio-temporal windows to understand the prevalence and outbreak of COVID-19 is the first of its kind to the best of our knowledge.

[*] Equal contribution.

2 Datasets

The **CMU Symptom Survey** aggregates the results of a survey run by CMU (Delphi group 2020) which was distributed across the US to ˜70k random Facebook users daily. It has 104 columns, including weighted (adjusted for sampling bias), unweighted signals, demographic columns (age, gender etc) for county and state level data. We use the data from Apr. 4, '20 to Sep. 11, '20.

The **UMD Global Symptom Survey** aggregates the results of a survey conducted by the UMD through Facebook (Fan et al. 2020). We use the data of 968 regions, available from May 01 to September 11. There are 28 unweighted signals provided, as well as a weighted form (adjusted for sampling bias). These signals include self reported symptoms, exposure information, general hygiene etc.

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.

3 Method and Experiments

Correlation Studies: Correlation between features of the dataset provides crucial information about the features and the degree of influence they have over the target value. We conduct correlation studies on different sub groups like symptomatic, asymptomatic and varying demographic regions in the CMU dataset to the discover relationships among the signals and with the target variable. We also investiage the significance of obesity and population density on the susceptibility to COVID-19 at state level (CDC 2020). Please refer to the Appendix for more information.

Outbreak Prediction: For outbreak prediction, we predict the % of the population that tested positive from CMU state data. After pruning unweighted and other signals, we are left with 36 input signals (Refer to the Appendix for details about the signal pruning process). We rank these 36 signals according to their *f_regression* (*f_statistic* of the correlation to the target variable) and predict the target variable using the top *n* ranked features. We experiment *n features* from 1 to 36 for various demographic groups. Linear Regression, Decision Tree and Gradient Boosting models (Pedregosa et al. 2011) are tested. Only the results for the best-performing model (Gradient Boosting) are shown.

Time Series Analysis: We predict the % of people that tested positive using the CMU dataset and % of people with CLI with the UMD dataset, using various combinations of features in the CMU (36) and UMD (56) datasets for multivariate multi-step time series forecasting. Given the data is spread across different spatial windows (geographies) at a state level, we employ an agglomerative clustering method independently on symptoms and behavioural/external patterns, and sample locations which are not in the same cluster for our analysis. Using the Augmented Dickey-Fuller test (Cheung and Lai 1995) we found the time series samples for these spatial windows to be stationary. Furthermore, we bucket 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] available, we use a Vector Auto Regressive (VAR) (Holden 1995) model and an LSTM (Gers, Schmidhuber, and Cummins 1999) model for the experiments. We also look at the impact of government policies (contact tracing, etc) on the spread of the virus.

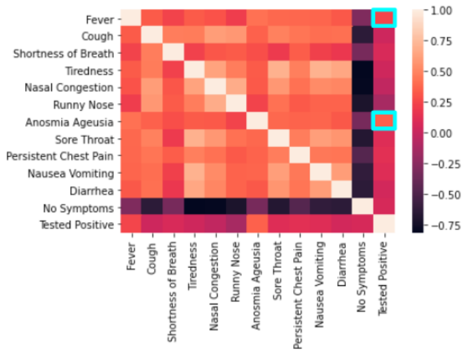

Figure 1: Correlation amongst self reported symptoms and % tested COVID positive.

4 Results and Discussion

Correlation Studies: State level analysis revealed a mild positive correlation between the % of people tested positive and statewide obesity level. Here the obseity is defined as BMI$>$ 30.0 (NIH 2020). These results are consistent with prior clinical studies like (Chan et al. 2020) and indicate that further research required to see if 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 correlation amongst multiple self reported symptoms and the symptoms having a significant positive correlations are highlighted. This clearly reveals that Anosmia , Ageusia and fever are reletively strong indicators of COVID-19. From Figure 5, we see that contact with a COVID-19 positive individual is strongly correlated with testing COVID-19 positive. Conversely, the % of population who avoid outside contact and the % of population testing positive for COVID-19 have a negative correlation. We also find a mild positive correlation between population density and % of population reporting COVID-19 positivity, which indicate easier transmission of the virus in congested environment. These observations reaffirm the highly contagious nature of the virus and the need for social distancing.

The above results motivate us to estimate the % of people tested COVID-19 positive based on % of people who had a direct contact with anyone who recently tested positive. In doing so, we achieve an MRE of 2.33% and MAE of 0.03.

Policies vs CLI/Community Sick Impacts : The impacts

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)

Table 1: Results for prediction of % of population tested positive across demographics. The 95% confidence interval (CI) for MRE is calculated on 20 runs (data shuffled randomly every time). the MRE and MAE are average of 20 runs.

of different non pharmaceutical interventions (NPIs) could be analysed by combining the CMU, UMD data and Oxford data (Hale et al. 2020). A particular analysis from that is reported here, where we notice that lifting of stay at home restrictions resulted in a sudden spike in the number of cases. This can be visualised in figure 4.

Error Metric - We find that a low MAE value is misleading in the case of predicting the spread of the virus; the MAE for outbreak prediction is low and has a small range (1-1.4) but more than 75% of the target lies between 0-2.6, meaning only a small percentage of the entire population has COVID-19. This makes MRE a better metric to use.

Outbreak prediction on CMU Dataset: Table 1 shows best accuracy achieved per dataset. For every dataset, the best ”n” is in 30s. We achieve an MRE of 60.40% for the entire dataset. The performance is better on Female-only data when compared to Male-only data. The performance is slightly better on 55+ age data than other age groups. This can also be observed from figure 2.

Top Features - Except for minor reordering, the top 5 features are - CLI in community, loss of smell, CLI in house hold (HH), fever in HH, fever across every data split. Top 6-10 features per data split are given in table 3. We can see that ’worked outside home’ and ’avoid contact most time’ are useful features for male, female and 55+ age data. Figure 2 shows mre vs number of features selected for different data splits. Overall, the error decreases as we add more features. However, the decrease in error isn’t very considerable when we go beyond 20 features ($<$ 1%).

Time Series Analysis - As seen in Tables 2, 3, 4 and 5, we are able to forecast the PCT_CLI with an MRE of 15.11% using just 23 features from the UMD dataset. We can see that VAR performs better than LSTM on an average. This can be explained by the dearth of data available. Furthermore, we can see that the outbreak forecasting on New York was done with 11.28% MRE, making use of only 10 features. 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 suggests several probable causes of confounding relationships between the two. This could also occur if both symptoms are specific and sensitive for COVID-19 infection.

Location	**VAR (%)**	
	MRE	***MAE***
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

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

Location	**LSTM (%)**	
	MRE	***MAE***
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: The errors of forecasting the outbreak of COVID-19 (% of people tested and positive) for the next 30 days using LSTM model.

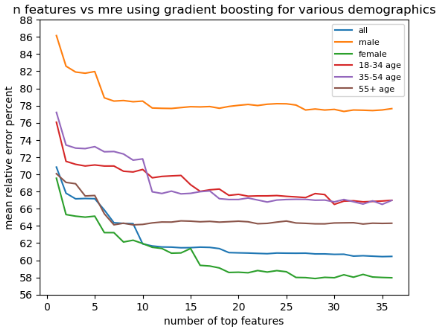

Figure 2: Errors vary across demographics, and generally decrease with increase in ”n”. The decrease is not considerable after n = 20.

Symptoms vs CLI overlap : The % of population with symptoms like cough, fever and runny nose is much higher than the % of people who suffer from CLI or the % of people who are sick in the community. Only 4% of the

Location	**VAR (%)**	
	MRE	***MAE***
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

Table 4: Results of forecasting the outbreak of COVID-19 (% of people with COVID-19 like illness in the population - PCT_CLI) for the next 30 days using VAR model

Location	**LSTM (%)**	
	MRE	***MAE***
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

Table 5: Results of forecasting the outbreak of COVID-19 (% of people with COVID-19 like illness in the population - PCT_CLI) for the next 30 days using LSTM model

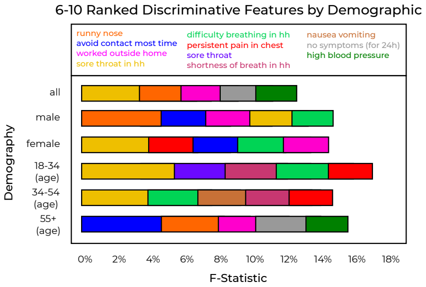

Figure 3: Excluding top 5 discriminative features (which are roughly identical), there are considerable differences between the next 5 features across demographics.

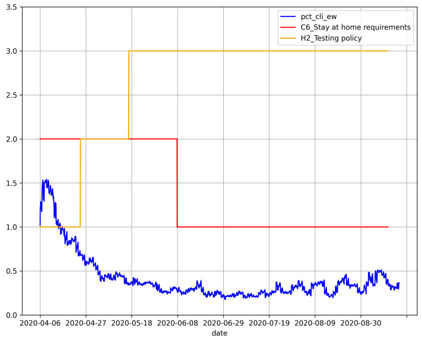

Figure 4: Policy Impacts: when Stay at home restrictions were stronger, even with higher testing rates, the % of population with CLI (pct_cli_ew) was having a downward trend.

people in the UMD dataset who reported to have CLI weren't suffering from chest pain and nausea.

Ablation Studies : Here, we perform ablation studies to verify and investigate the relative importance of the features that were selected using f regression feature ranking algorithm (Fre 2007-2020). In the following experiments the top $N = 10$ features obtained from the f regression algorithm are considered as the subset for evaluation.

All-but-One: In this experiment, the target variable which is the percentage of people affected by COVID 19 is 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 a descending order. The results are visualised in figure 6 from which it is clear that there is a considerable increase error when the most significant feature is dropped and the loss in performance is not as drastic when any other feature is dropped. This reaffirms our feature selection method.

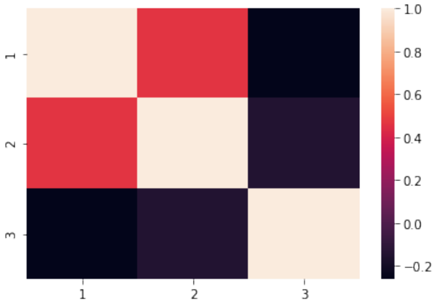

Figure 5: Correlation between the people having contact with someone having CLI and People tested positive. Here the attribute (1) = % of people who had contact with someone having COVID-19, (2) = % of people tested positive, (3) = % of people who avoided contact all/most of the time

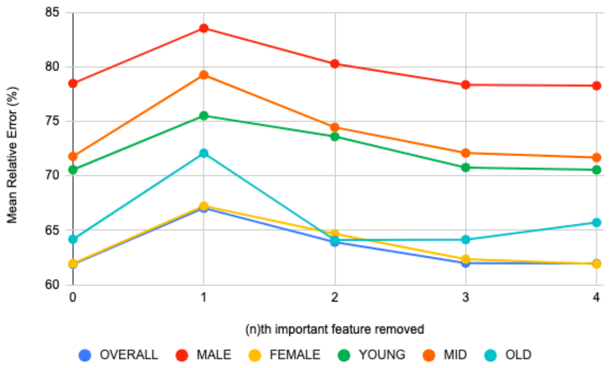

Figure 6: Results of All-but-One experiment (MRE)

Cumulative Feature Dropping: In this experiment, we estimate the target variable based on top N=10 features and then carry out the experiment with $N - i$ features in every iteration where i is the iteration count. The features are dropped in the descending order. Figure 7 shows the results. The change in slope from the start to the end of the graph strongly supports our previous inference that the most important feature has a huge significance on the performance and error rate and reaffirms our features selection algorithm.

5 Conclusion And Future Work

In this work, we analyse the benefits of COVID-19 self reported symptoms present in the CMU, UMD, and Oxford datasets. We present correlation analysis, outbreak prediction, and time series prediction of % of re+spondents with

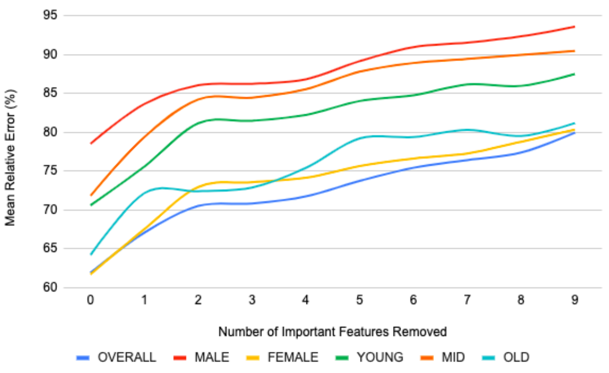

Figure 7: Results of Cumulative Feature Dropping

positive COVID-19 tests and % of respondents who show COVID-like illness. By clustering datasets across different demographics, we reveal 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 identify 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 providing false 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.

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. Building machine learning systems on data from mobile/wearable devices can be built to understand users' vitals, sleep behavior etc., have the data shared at an individual level, can augment the participatory surveillance dataset and thereby the predictions made. This can be achieved without compromising on the privacy of the individual. We also plan to compare the reliability of such participatory surveillance methods with actual number of cases in the corresponding regions and it's generalisability across the population.

6 Acknowledgement

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

References

2007-2020. *sklearn f regression*. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html.

CDC. 2020. *Data and Statistics*. https://www.cdc.gov/obesity/data/index.html.

Chan, C. C.; et al. 2020. Type I interferon sensing unlocks dormant adipocyte inflammatory potential. *Nature Communications* 11(1). ISSN 2041-1723. URL https://doi.org/10.1038/s41467-020-16571-4.

Cheung, Y.-W.; and 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. URL https://covidcast.cmu.edu/surveys.html.

Fan, J.; et al. 2020. COVID-19 World Symptom Survey Data API.

Gers, F. A.; Schmidhuber, J.; and 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.; and Lloyd-Smith, J. O. 2020. Estimated effectiveness of symptom and risk screening to prevent the spread of COVID-19. *eLife* 9. ISSN 2050-084X. doi:10.7554/elife.55570. URL https://europepmc.org/articles/PMC7060038.

Hale, T.; Webster, S.; Petherick, A.; Phillips, T.; and 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.; Li, Y.; Guan, W.; Sang, L.; Lu, J.; Xu, Y.; Chen, G.; Guo, H.; Guo, J.; Chen, Z.; Zhao, Y.; Li, S.; Zhang, N.; Zhong, N.; He, J.; and for the China Medical Treatment Expert Group for COVID-19. 2020. 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. ISSN 2168-6106. doi:10.1001/jamainternmed.2020.2033. URL https://doi.org/10.1001/jamainternmed.2020.2033.

Menni, C.; Valdes, A. M.; Freidin, M. B.; Sudre, C. H.; Nguyen, L. H.; Drew, D. A.; Ganesh, S.; Varsavsky, T.; Cardoso, M. J.; El-Sayed Moustafa, J. S.; Visconti, A.; Hysi, P.; Bowyer, R. C. E.; Mangino, M.; Falchi, M.; Wolf, J.; Ourselin, S.; Chan, A. T.; Steves, C. J.; and Spector, T. D. 2020a. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine* 26(7): 1037–1040. ISSN 1546-170X. doi:10.1038/s41591-020-0916-2. URL https://doi.org/10.1038/s41591-020-0916-2.

Menni, C.; et al. 2020b. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature medicine* 1–4.

NIH. 2020. *Adult Body Mass Index (BMI)*. https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm.

Parma, V.; et al. 2020. More than smell. COVID-19 is associated with severe impairment of smell, taste, and chemesthesis. *medRxiv* doi:10.1101/2020.05.04.20090902. URL https://www.medrxiv.org/content/early/2020/05/24/2020.05.04.20090902.

Pedregosa, F.; et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

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. ISSN 1941-1189. doi:10.1109/rbme.2020.2987975. URL http://dx.doi.org/10.1109/RBME.2020.2987975.

Wilder, B.; Mina, M. J.; and Tambe, M. 2020. Tracking disease outbreaks from sparse data with Bayesian inference. *arXiv preprint arXiv:2009.05863* .

7 Appendix

Dataset	Example Signals
UMD	COVID-like illness symptoms, influenza-like illness symptoms, mask usage
CMU	sore throat, loss of smell/taste, chronic lung disease
OxCGRT	containment and closure policies, economic policies, health system policies

Table 6: Example Signal Information for the Datasets

The sample features present in the datasets can be observed in table 6.

Correlation Studies

The detailed plots of the correlation analysis of the CMU dataset is noted in figure 11.

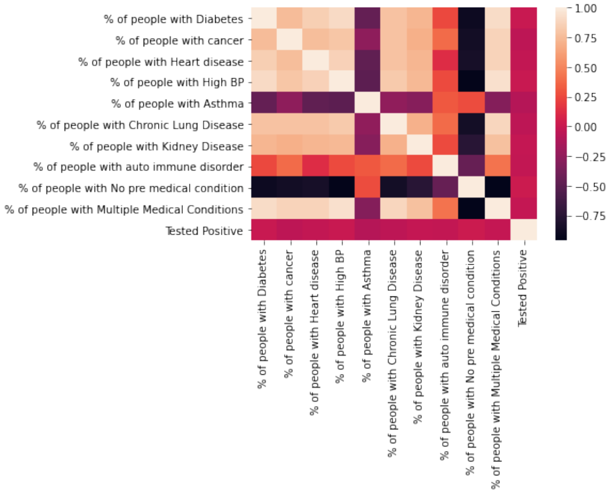

Figure 8: Correlation study: The relationship between the underlying medical condition and percentage People tested COVID positive

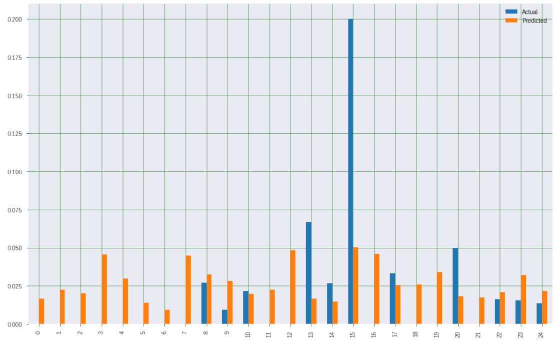

Figure 9: Observed VS Predicted: Prediction made on Percentage of people tested positive using percentage of people who recently had a contact with someone with COVID positive

Feature Pruning

We drop demographic columns such as date, gender, age etc. Next we drop the unweighted columns because their weighted counterparts exist. We also drop features like % of people who got tested negative, weighted % of people who got tested positive etc as these are directly related to testing and would make the prediction trivial. Further, we drop derived features like(t), like estimated % of people with influenza-like illness because they were not directly reported by the respondents. Finally, we drop the features which calculate mean (such as average number of people in respondent's household who have cli) because their range was in the order of 10^{50}. After the entire process we are left with 36 features.

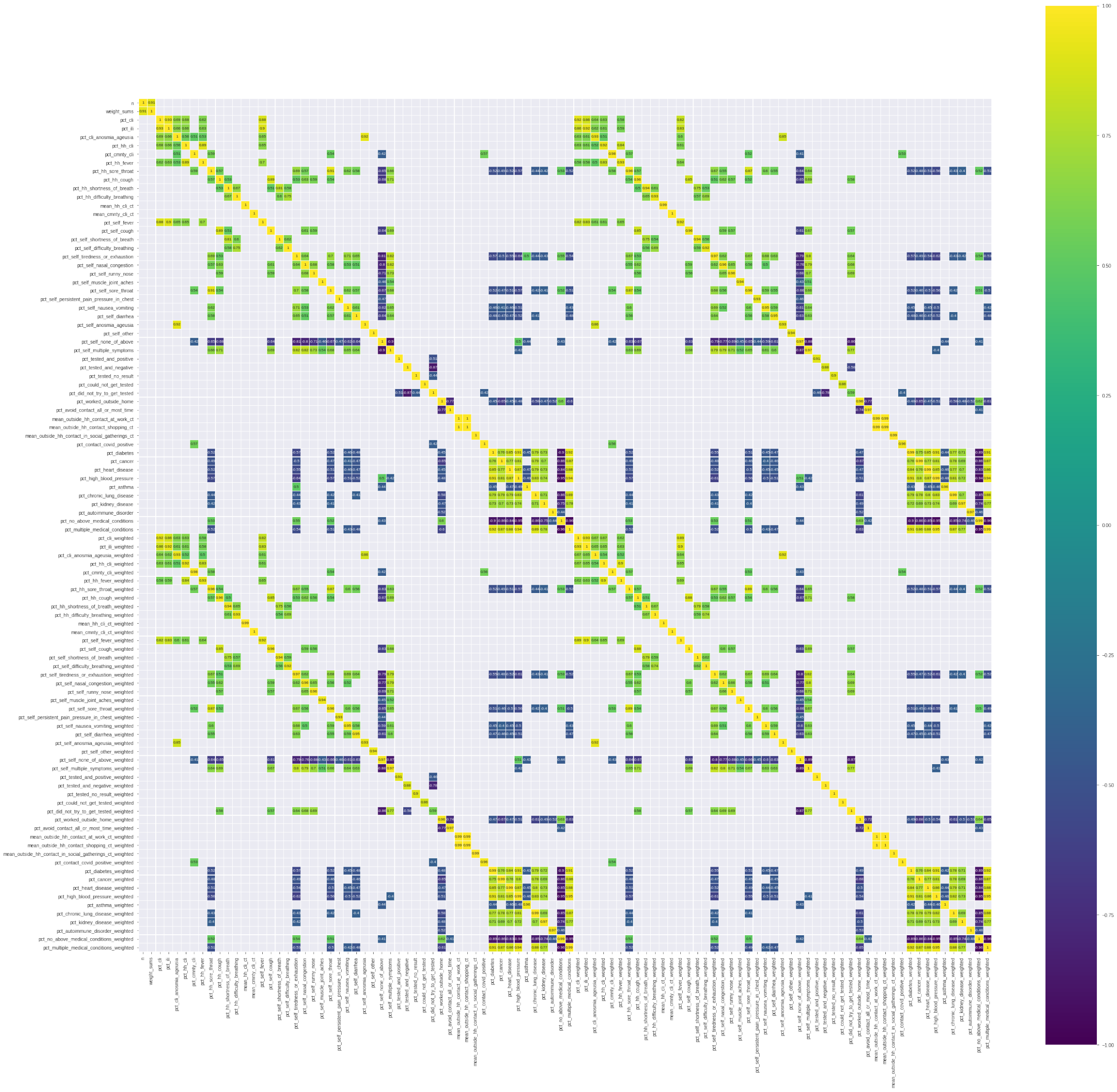

Figure 10: Correlation map depicting the relationship between the features along with the target variable(s)

Time Series

In table 8 we continue to experiment with different spatial windows, like trying to predict PCT_CLI for different locations like ”Tokyo” and ”British Columbia” using different combination of features. Further on table 10 analysis is done on more US states with an LSTM based deep learning model to predict PCT_CLI and we notice that there is no significant gain in using DL models (probably due to lack of data). The pct_community_sick is another variable which we try to predict, and the results can be seen in table 9

In figs [13,15] we do Dynamic Time Warping(DTW) to compare how well our forecasted timeseries curve matches with the original curve. DTW was used due to the flexibility to compare timeseries signals which are of different lengths. This will enable us to compare different temporal windows across different spatial windows to understand the effectiveness of the model at different contexts.

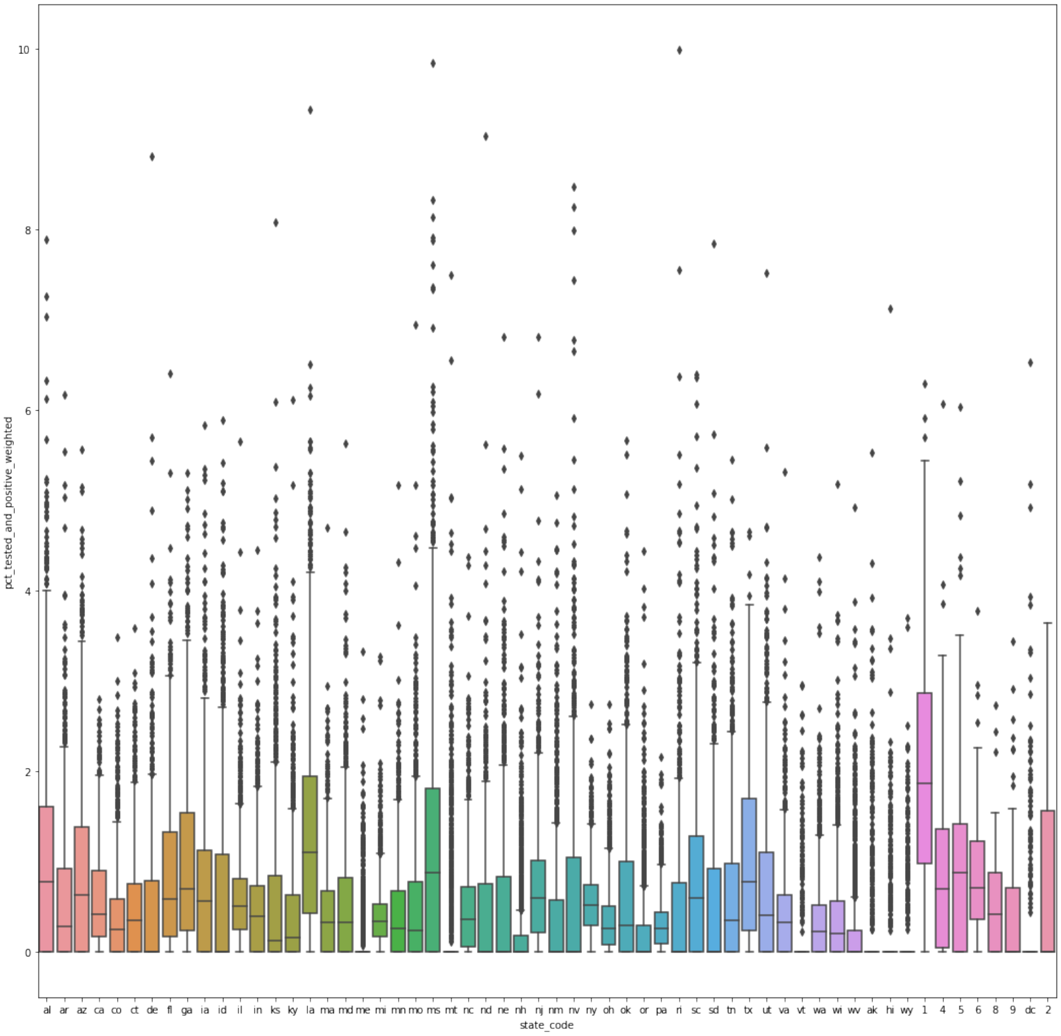

Figure 11: State wise distribution of percentage of people tested COVID positive

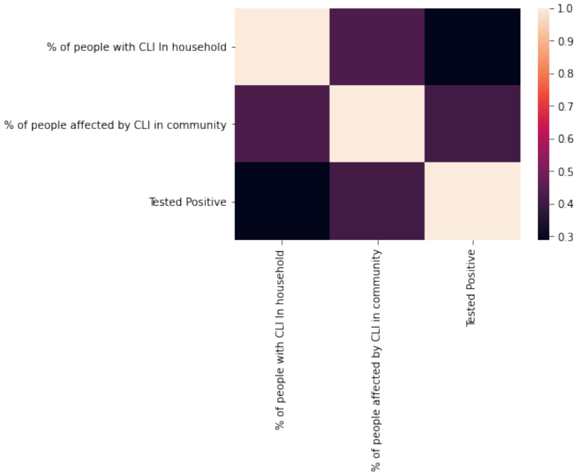

Figure 12: Correlation study: The relationship between the COVID like illness and percentage People tested COVID positive

Rank	**Signal**	**F_Statistic**
1	COVID-like Illness in Community	14938.48816456
2	Loss of smell or taste	9498.89229794
3	COVID-like Illness in Household	6050.88250153
4	Fever in Household	5490.15612527
5	Fever	4388.95759983
6	Sore Throat in Household	1787.42269067
7	Avoid contact with others most of the time	1494.25038393
8	Difficulty breathing in Household	1330.48793481
9	Persistent Pain Pressure in Chest	1257.78331468
10	Runny Nose	1084.84412662
11	Worked outside home	1023.50285601
12	Nausea or Vomiting	1016.94758914
13	Shortness of breath in Household	1004.67944587
14	Sore Throat	975.25614266
15	Difficulty Breathing	723.49150048
16	Asthma	466.91243179
17	Shortness of Breath	440.88344033
18	Cough in Household	322.05679444
19	No symptoms in past 24 hours	241.72819985
20	Diarrhea	228.59465358
21	Chronic Lung Disease	224.24651285
22	Cancer	205.19827073
23	Other Pre-existing Disease	158.31567587
24	Tiredness or Exhaustion	134.36715409
25	Cough	84.66549815
26	No Above Medical Conditions	84.40193799
27	Heart Disease	74.71994609
28	Multiple Medical Conditions	52.61630823
29	Autoimmune Disorder	40.8942176
30	Nasal Congestion	33.60170138
31	Kidney Disease	23.88450351
32	Average people in Household with COVID-like ilness	14.52969291
33	Multiple Symptoms	12.56805547
34	Muscle Joint Aches	1.72398411
35	High Blood Pressure	0.48328156
36	Diabetes	0.24390025

Table 7: Features ranked for entire by *F score*. All signals are represented as percentages of respondents who responded that way.

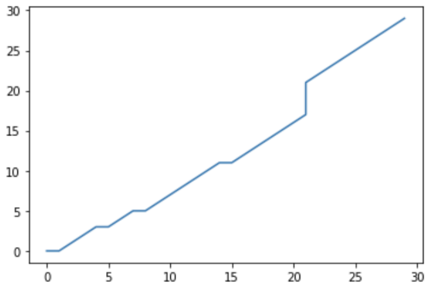

Figure 13: DTW plot analysing the relationship between our forecasted curve vs the original curve for Ohio State

Location	Bucket	RMSE	MAE	MRE (%)	Features Used
Abu Dhabi	male and age 18-34	2.43	2.23	167.86	difficulty breathing + anosmia ageusia (weighted)
Tokyo	female and age 35-54	0.56	0.47	30.16	difficulty breathing + anosmia ageusia (weighted)
British Columbia	male and age 55+	1.09	0.59	28.68	difficulty breathing + anosmia ageusia (weighted)
Lombardia	male and age 55+	0.95	0.67	28.72	difficulty breathing + anosmia ageusia (weighted)
Lombardia	male and age 55+	0.95	0.67	28.72	Behavioural / external features (weighted)
British Columbia	male and age 55+	1.07	0.76	50.17	Behavioural / external features (weighted)
Tokyo	female and age 35-54	0.58	0.49	31.38	Behavioural / external features (weighted)
Abu Dhabi	male and age 18-34	2.91	2.78	207.94	Behavioural / external features (weighted)

Table 8: RMSE and MAE scores for different buckets of interest + Ablation - VAR model - PCT _CLI _weighted

Location	Bucket	RMSE	MAE	MRE (%)
Abu Dhabi	male and age 18-34	9.99	8.94	73.11
Tokyo	female and age 35-54	1.13	1.02	41.67
British Columbia	male and age 55+	3.21	2.65	137.13
Lombardia	male and age 55+	1.25	1.25	24.49

Table 9: RMSE and MAE scores for different buckets of interest - VAR model - PCT _Community _Sick

Location	Bucket	RMSE	MAE	MRE	Model
TX	male and age overall	1.56	1.21	43.00	VAR
CA	male and age overall	1.22	0.93	23.44	VAR
NY	female and age overall	0.7	0.56	21.59	VAR
FL	female and age overall	1.48	1.18	19.35	VAR
TX	male and age overall	6.28	4.06	89.4	LSTM
CA	male and age overall	2.83	2.68	71.24	LSTM
NY	female and age overall	2.02	1.9	68.17	LSTM
FL	female and age overall	4.33	4.19	73.34	LSTM

Table 10: RMSE and MAE scores for different buckets of interest - VAR/LSTM models - PCT_CLI - Here we see that deep learning models aren't performing better than normal statistical models

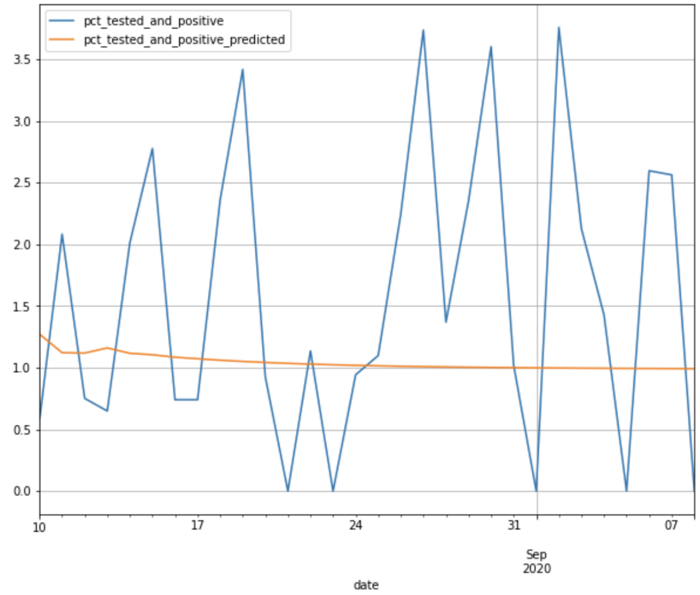

Figure 14: Forecasted curve vs the original curve for Ohio.

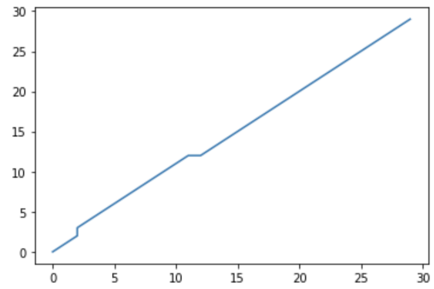

Figure 15: DTW plot analysing the relationship between our forecasted curve vs the original curve

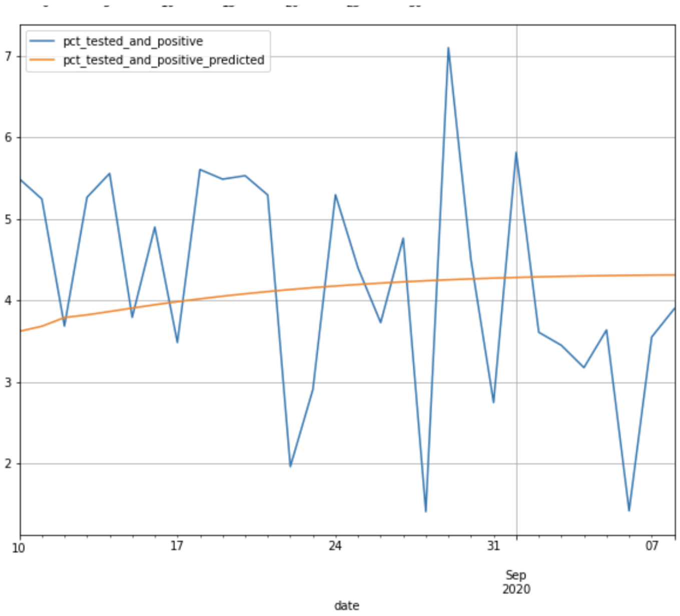

Figure 16: Forecasted curve vs the original curve for Texas.

Demography	**Feature Removed**	**MAE**	**MRE**
Male	no feature removed	1.389806313	77.42367322
	pct_cmnty_cli_weighted	1.470745054	82.97970974
	pct_self_anosmia_ageusia_weighted	1.423361929	79.90430572
	pct_self_none_of_above_weighted	1.410196471	78.62630177
	pct_self_runny_nose_weighted	1.398427829	78.13485192
Female	no feature removed	1.100879926	57.63336087
	pct_cmnty_cli_weighted	1.218554308	64.54253671
	pct_self_anosmia_ageusia_weighted	1.155647687	61.12311515
	pct_self_none_of_above_weighted	1.121811889	58.73118158
	pct_self_runny_nose_weighted	1.104380112	57.92685018
Young	no feature removed	1.231519891	67.07207641
	pct_cmnty_cli_weighted	1.31846811	72.201516
	pct_self_anosmia_ageusia_weighted	1.277138933	70.38556851
	pct_avoid_contact_all_or_most_time_weighted	1.244334089	67.80402144
	pct_self_runny_nose_weighted	1.234101952	67.46623764
Mid	no feature removed	1.276053866	67.05778653
	pct_cmnty_cli_weighted	1.384547554	73.44381028
	pct_self_anosmia_ageusia_weighted	1.326526868	70.22181485
	pct_self_none_of_above_weighted	1.321293709	69.44829708
	pct_avoid_contact_all_or_most_time_weighted	1.285893087	67.62940495
Old	no feature removed	1.172592164	63.98633923
	pct_cmnty_cli_weighted	1.314221647	72.59134309
	pct_avoid_contact_all_or_most_time_weighted	1.191250701	64.98442049
	pct_self_anosmia_ageusia_weighted	1.192677984	65.76644281
	pct_self_multiple_symptoms_weighted	1.186357275	64.7244507

Demography	Feature Removed	MAE	MRE
Overall	no feature removed	1.143995128	60.83421503
	pct_cmnty_cli_weighted	1.248043237	67.08605954
	pct_self_anosmia_ageusia_weighted	1.177417511	63.07033879
	pct_self_none_of_above_weighted	1.169464223	61.67148756
	pct_self_runny_nose_weighted	1.149200232	61.32185068
	pct_hh_cli_weighted	1.14551667	60.93481883
	pct_avoid_contact_all_or_most_time_weighted	1.149772918	61.16631628
	pct_worked_outside_home_weighted	1.147615986	61.0433573
	pct_self_fever_weighted	1.144711565	60.92739832
	pct_hh_fever_weighted	1.143703022	60.7946325
	pct_hh_difficulty_breathing_weighted	1.143007815	60.83204654