DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting

Alexander Rodriguez[*], Anika Tabassum[*†], Jiaming Cui[*], Jiajia Xie[*], Javen Ho[*], Pulak Agarwal[*], Bijaya Adhikari[‡], B. Aditya Prakash[*]

[*]College of Computing, Georgia Institute of Technology, Atlanta, USA
[†]Department of Computer Science, Virginia Tech, Arlington, USA
[‡]Department of Computer Science, University of Iowa, Iowa City, USA
{arodriguezc, jiamingcui1997, jxie, jho67, pulak.agarwal, badityap}@gatech.edu, anikat1@vt.edu, bijaya-adhikari@uiowa.edu

Abstract

How do we forecast an emerging pandemic in real time in a purely data-driven manner? How to leverage rich heterogeneous data based on various signals such as mobility, testing, and/or disease exposure for forecasting? How to handle noisy data and generate uncertainties in the forecast? In this paper, we present DEEPCOVID, an operational deep learning framework designed for real-time COVID-19 forecasting. DEEPCOVID works well with sparse data and can handle noisy heterogeneous data signals by propagating the uncertainty from the data in a principled manner resulting in meaningful uncertainties in the forecast. The framework also consists of modules for both real-time and retrospective exploratory analysis to enable interpretation of the forecasts. Results from real-time predictions (featured on the CDC website and FiveThirtyEight.com) since April 2020 indicates that our approach is competitive among the methods in the COVID-19 Forecast Hub, especially for short-term predictions.

1 Introduction

Motivation. The devastating impact of the currently unfolding global COVID-19 pandemic has sharply illustrated our enormous vulnerability to emerging infectious diseases. Forecasting disease trajectories is a non-trivial and important task. Estimating various measures related to the epidemic gives policymakers valuable lead time to plan interventions and optimize supply chain decisions. Hence, accurate forecasts are critical in combating epidemic outbreaks including the current pandemic (Holmdahl and Buckee 2020).

To encourage research in this space and to provide a unified forecasting platform, the US Centers for Disease Control and Prevention (CDC) is currently organizing a collaborative forecasting task under the umbrella of the COVID-19 Forecast Hub. The forecasting targets include various COVID-related metrics including mortality and hospitalizations at various temporal and spatial resolutions. The initiative has attracted submissions from more than 58 teams from industry and academia (as of Sept 2020).

The majority of the participating approaches can be classified into two categories (i) statistical and (ii) mechanistic. The mechanistic approaches model the underlying disease transmission over the population contact network under various assumptions using either ordinary differential equations (Zhang et al. 2017) and/or agent based models (Venkatramanan et al. 2018). While very valuable for long-term 'what-if' scenario generation, mechanistic approaches have several challenges in real-time forecasting of an emerging pandemic. The first drawback is that it the complexity of the model rises with the number of data sources being used. Hence, to manage the complexity, most models often use only one or two data signals (such as weather) along with the observed disease incidence (Shaman, Goldstein, and Lipsitch 2010). Hence it is not trivial to extend these models to include various other data signals (like social media data).

On the other hand, statistical approaches are fairly new in this space. They exploit correlations between various data sources and the forecast targets to learn a functional dependence between the two and use the learnt function to make predictions (Yuan et al. 2013). There are several challenges in designing statistical approaches as well as we discuss later (see Sec 4). Most of the existing statistical approaches for epidemic forecasting (usually developed for influenza forecasting, including work by the authors (Adhikari et al. 2019)) cannot be readily adapted for COVID forecasting due to

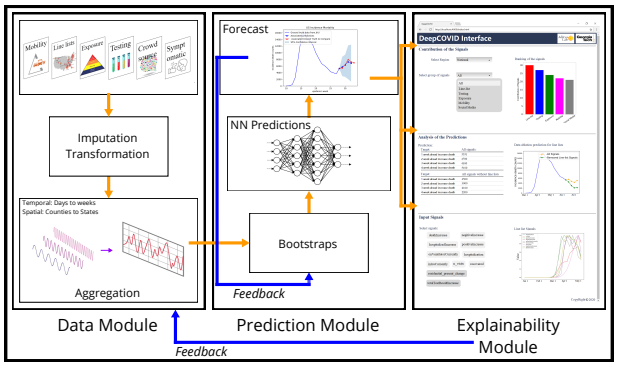


Figure 1: Schematic of DEEPCOVID framework for real-time COVID-19 forecasting. The data module is dedicated to data pre-processing including imputation of missing values and aggregating at the right temporal and spatial resolution. The prediction module generates probabilistic forecasts based on the curated data. Finally, the explainability module (with interface) allows both the real-time and retrospective analysis of forecasts to an build intuitive explanation of forecasts.

several issues such as lack of data and poor data quality. Indeed there is hardly (if any) work on describing a completely data-driven approach for real time forecasting.

Our goals. In this paper, we describe our framework DEEPCOVID (see Fig.1), the *first* purely data-driven deep learning (DL) model for real time pandemic forecasting in the COVID-19 Forecast Hub (Reich et al. 2020)[1] and the official ensemble. Our real-time forecasts are showcased in the hub, the CDC website (CDC 2020) and the popular FiveThirtyEight website (538 2020). By our work, we aim to address a gap in the literature pertaining to using purely data driven approaches for emerging pandemics, with the following goals: *G1. Coping with heterogeneous, scarce and noisy data:* A DL-based model can ingest many heterogeneous signals that are more sensitive to what is happening on the ground, without laborious feature engineering. To fully take advantage of this, our framework is designed with careful consideration of data and modeling challenges faced in robust real-time forecasting with principled uncertainty estimation. *G2. Bring a complementary forecasting perspective:* A good ensemble needs diverse perspectives (Reich et al. 2019; Ray et al. 2020). The overwhelming majority of the teams in the ensemble perform mechanistic modeling. By being the first data-driven DL method in the ensemble, DEEPCOVID brings a unique perspective closer to the observed data signals with minimal assumptions. *G3. Excellence in short-term forecasting:* The utility of statistical models in short-term forecasting is well-known (Holmdahl and Buckee 2020), which are useful to plan intervention and allocate resources. We demonstrate DEEPCOVID excels in this task by comparing to a strong baseline. *G4. Enable communication to domain-experts:* This is to ensure our framework gives explanations for its forecasts, which are very important for communication and interpretation by both the public and decision makers.

Contributions. Our contributions are as follows:

- We propose DEEPCOVID, one of the first deep learning based real-time COVID forecasting frameworks, whose performance in the CDC COVID-19 Forecast Hub (since April 2020) demonstrates that it is consistently competitive, especially at short term forecasting.
- We address several challenges of such real time forecasting including interpretability and uncertainty estimation in a principled fashion.
- We provide valuable observations and 'lessons learnt' from our experience for modeling emerging infectious diseases using purely data-driven methods.

2 Related Work

Epidemic Forecasting: Modeling approaches for epidemic forecasting can be broadly categorized in mechanistic (Zhang et al. 2017; Shaman and Karspeck 2012) and statistical (Tizzoni et al. 2012; Osthus et al. 2019; Brooks et al. 2018), the latter closely related to time series analysis (Box et al. 2015; Jha et al. 2015). Past work has explored leveraging multiple sources of data, from search engine (Ginsberg et al. 2009; Yuan et al. 2013) to weather data (Shaman, Goldstein, and Lipsitch 2010; Tamerius et al. 2013; Volkova et al. 2017) for epidemic forecasting. There has been a recent spate of work in deep learning approaches that tackle the data sparsity problem by leveraging seasonal time series intra- and inter- similarity (Adhikari et al. 2019) and principled simulation (Wang, Chen, and Marathe 2019).

COVID-19 Forecasting. Other approaches adopted by the contributing models are mechanistic (Zou et al. 2020; Chinazzi et al. 2020; Baek et al. 2020) and statistical (Altieri et al. 2020; Murray et al. 2020). The official hub ensemble (Ray et al. 2020) combines forecasts from ours and other models.

3 Background

Forecasting requirements from CDC. Starting in April 2020, CDC requests probabilistic forecasts for COVID-19 associated mortality and hospitalizations at various temporal and spatial resolutions to be used by policymakers to plan intervention and allocate resources. As mentioned before, our model has been part of the ensemble since its inception.

Forecasting Targets. The forecasting targets are: *T1. Incidence and cumulative weekly deaths:* Reported incidence (new) and cumulative deaths for US states and the US overall. The data reported by Johns Hopkins University (JHU) (Dong et al. 2020) serves as gold standard for the CDC. *T2. Incidence daily hospitalizations:* Reported new hospitalizations for US states and the US overall. CDC does not fix a gold standard for this but we found the data are provided by the COVID Tracking Project (cov 2020c) to be the closest.

Problem formulation. We can state our real-time forecasting problem for a specific geography as follows.

Given: an observed multivariate time series of COVID-related signals $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ and corresponding values for the forecasting target $\mathcal{Y} = \{y_i\}_{i=1}^{N}$, where N is the size of the sequence until the current date.

Predict: next k values of forecasting target, i.e. $\{\hat{y}_i\}_{i=N+1}^{N+k}$, where $k = 4$ for T1 (4 weeks ahead) and $k = 28$ for T2 (28 days ahead).

4 Our Framework: DeepCOVID

DeepCOVID is our framework for explainable real-time COVID-19 forecasting, which contains three modules depicted in Figure 1: data module, prediction module, and explainability module. By separating data and prediction modules, our goal is to differentiate between the handling of noisy data from the learning process. For the predictive module, we use deep learning because it is a flexible, scalable, and efficient technology, and an excellent choice to model non-linearities. As mentioned before, explainability is a challenge in data-driven models. However, we want to understand and connect forecasts with epidemiological reasons. Once we have insights about our predictions, we will have a feedback loop to improve performance. Therefore, we have an explicit module for explainability which helps to shed some light on our predictions in a dynamic situation.

[1] Indeed, when we started participating, there were only 11 teams in the hub. We were the first two teams for predicting hospitalizations.

4.1 Data Module

In this section, we describe our data collection and pre-processing for input to our deep neural network.

Challenges. Data pre-processing for real-time COVID forecasting brings several challenges primarily due to a novel emerging scenario. *C1. Data collection:* Collecting data in such a chaotic scenario is challenging because it comes from multiple sources (often collected by volunteers), is in different formats, some even changing over time (e.g. reported deaths from JHU), and unexpected hiatus. *C2. Selection of signals:* We carefully select signals that describe the different facets of the disease spread. To enable epidemiological observations about our predictions, this selection has to be driven by appropriate rationale. *C3. Temporal misalignment:* Since the signals are collected from multiple sources, they often have temporal misalignment. Some signals presented 1-2 weeks of lag due to delays in reporting from hospitals, public records and government officials; others have different temporal resolutions (days vs weeks). *C4. Spatial misalignment:* Some records are reported at specific spatial granularity and their conversion to higher geographical levels is non-trivial. *C5. Missing values:* Most prominently, our forecasting target, i.e., incidence hospitalizations, has not been reported in 11 states (CA, DC, TX, IL, LA, PA, MI, MO, NC, NV, DE).

Our approach. To address *C1*, we developed a data extraction program personalized for each data format to convert them all to a standard format. For *C2*, we extensively searched for meaningful signals from an epidemiological perspective. The outcome of our search is summarized in Table 1 for the signals consistently used in our predictions. We collected the signals for 52 geographical regions: US National, the 50 US states and Washington D.C. We address the data misalignment in *C3* as follows. For the lags in reporting, we downshift the signals. Our idea is to align all the signals based on their latest records since it is safe to assume that the latest records are more indicative for future targets. To address weekly/daily inconsistency, signals require different treatment depending on their nature. For example, since weekly hospitalization rate and CLI% ER visits have been recorded as percentages, we choose to consider the same value for daily incidence because it is not meaningful to transform it. For *C4*, signal 13 in Table 1 is recorded only at county-level, while other signals (6 and 7) contain records at the HHS region-level[2]. To provide a state-wise forecast model, we transformed the signal record state-wise by aggregating and de-aggregating, respectively. For this, it is crucial to consider the population of such geographic regions. Last, it is important to address *C5* in a meaningful way. For 11 states with missing values, we had related signals such as hospitalization and ICU patients. We inferred missing values by leveraging these signals and making reasonable assumptions (such as effect of people's recovery and death in one week).

More details about how we handle these challenges and detailed description of our data signals are in our appendix.

[2]https://www.hhs.gov/about/agencies/iea/regional-offices/index.html

Algorithm 1 Predictive module training

1: **Input:** Observed multivariate time series $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$; target labels $\mathcal{Y}^k = \{y_i^k\}_{i=1}^{N-k}$ for each days/weeks ahead to predict $k = 1, \ldots, K$; no. of bootstrap samples M
2: **Output:** $\{Z^k\}_{k=1}^{K}$, where Z^k is a set of M predictions
3: **for** $k = 1$ to K **do**
4: **if** $k > 1$ **then**
5: {Sample prev. forecasts and add to next train data}
6: Uniform sample $y_{N-j+1}^k \sim \mathcal{Z}^{k-j}$ for $j = 1$ to $k-1$, and include each of them in the $(N-j+1)$-th position of $\mathcal{Y}^k$
7: **end if**
8: $Z^k \leftarrow$ Fit/predict M models, each trained with a bootstrap sample from $(\mathcal{X}, \mathcal{Y}^k)$.
9: **end for**
10: Return $\mathcal{Z}^k$ for $k = 1, \ldots, K$

4.2 Prediction Module

Challenges. Real-time COVID forecasting is a difficult problem with challenges originating from data and CDC requirements. Some of the challenges we encountered in designing the prediction module are: *C1. Data Sparsity:* One of the major challenges in forecasting emerging disease, specially at an early stage is data sparsity. Extracting enough information from the few available data points to ensure generalizable forecasts is a challenging problem. *C2. Robust point and probabilistic forecasting:* Forecasting disease in real-time is already challenging, and the problem is even more difficult in presence of data issues mentioned above. Hence, the questions we tackle in designing our framework are (1) how do we principally translate data noise to forecast uncertainty?; and (2) how do we ensure that our approach is robust to the noise and other data issues such that ensures reliable point and probabilistic forecasts? *C3. Temporal consistency between the forecasts:* Due to data sparsity we cannot hope to train deep networks enforcing temporal consistency such as LSTM and GRU. So the question is, how to design a neural architecture, which has few enough parameters to train from sparse data while ensuring temporal consistency between the forecasts?

Our approach. See Alg. 1. When we started to participate in the CDC task (April 2020), there was 1 month of meaningful (non-zero) data; so it was unclear how to effectively train a recurrent neural network. Hence we opted to use a feedforward network with autoregressive inputs to incorporate short-term dependencies in the time series. To fully address *C1*, we also need to avoid overfitting, for which we empirically evaluated several number of layers & sizes for them.

To address *C2*, we want capture uncertainty from data in a principled way. Note that we want data-based uncertainty, not model-based uncertainty (e.g. MC dropout). Thus, for each future incidence target, we obtain one single prediction per bootstrap sample so that we obtain a set of M predictions representing uncertainty in data (lines 7-10 in Alg. 1). Robustness of our point and probabilistic predictions are closely related to optimization. Optimizing the parameters of a neural network with sparse, noisy and heterogeneous data is challenging. In fact, we found that our optimization sometimes

Table 1: Overview of data signals used in DEEPCOVID. (ILI=Influenza like Illness; CLI=COVID like Illness)

Type of signal	Description	Signals	Rationale
(DS1) *Line list*	Directly related to the disease spread (derived from records of who, when, and where a person got infected) (cov 2020c; cdc 2020b; jhu 2020)	1. Confirmed cases; 2. UCI beds currently occupied; 3. People on ventilation; 4. Recovered; 5. Hospitalization rate (COVID-Net); 6. ILI ER visits; 7. CLI ER visits; 8. Excess Deaths;	Traditional surveillance for tracking patients and symptoms (CDC 2020)
(DS2) *Testing*	Related to social policy and behavioral considerations (cov 2020c; cdc 2020b)	9. People tested; 10. Negative cases; 11. Emergency facilities reporting; 12. Number of providers;	To capture changing screening and diagnostic artifacts on surveillance
(DS3) *Crowdsourced symptomatic*	Collected from individuals using Kinsa digital thermometers at home (Miller et al. 2018)	13. Digital thermometer readings;	Syndromic symptomatic surveillance
(DS4) *Mobility*	Indicate people visits in multiple POIs (cov 2020a),(cov 2020d)	16. Retail and recreation; 17. Grocery and pharmacy; 18. Parks; 19. Transit stations; 20. Residential 21. Workplaces; 22. Overall-region-based	Evidence of changing spatio-temporal contact patterns due to non pharma. interventions and behavior changes
(DS5) *Exposure based*	Collected from tracking overlapping location of distinct smartphones in commercial venues (Couture et al. 2020).	23-24. Device exposures (normal & adjusted);	Measure social contacts and direct potential exposures
(DS6) *Social Surveys*	Facebook symptomatic survey (cov 2020b)	35. CLI (%); 36. ILI (%)	Measuring related symptomatic burden

fell into local optima that returned absurd predictions. Hence, to improve robustness, we used batch normalization (Ioffe and Szegedy 2015) to alleviate initialization problems. Besides that, we run the same optimization for several times and select the one that leads to lowest loss in the (training) data.

To address *C3*, we capture temporal correlations between consecutive forecasts by using predictions from week k as part of the training data for week $k+1$. In this way, we are also propagating uncertainty to future predictions (lines 4-5 in Alg. 1). This process can be regarded as a semi-supervised learning procedure because we use our model to create new labels that are used in training data for future targets.

Note: Cumulative counts is non-decreasing. Adding this restriction to our optimization problem would make it even more prone to get stuck in local optima. Therefore, for predicting cumulative deaths, once we get incidence predictions, we convert them to cumulative ones by aggregating the point predictions. We found this gives consistent and stable performance.

4.3 Explainability Module

Challenges. Policymakers have been constantly placing and lifting travel restrictions and quarantine orders to balance the trade-off between the pandemic burden and economic costs. Moreover, these policies differ in each administrative region. Hence, the data signals that we collect have different meaning and usability at different times and regions. Hence, the challenge here is to ensure that the our framework is able to generate forecasts explainable with respect to the data signals and enable analysis of signal strengths over different geographies and periods.

Our approach. See Alg. 2. We selected data ablation accompanied with an interface as means for this purpose. It is a systematic and simple way to quantify the contribution of signals in the predictions. In Alg. 2, we train models with every group of signals removed and quantify the deviation with respect to a reference point. Then, we rank the groups of signals by this deviation, and check if the contribution of the signal is statistically significant with respect to the model with all signals (this model is denoted as $s = \emptyset$). Our

Algorithm 2 Explanations

1: **Input:** Group of signals to analyze $\mathcal{A}$; statistical significance α; days/weeks for analysis $\mathcal{W}$
2: **Output:** Groups of signals ranked by impact
3: **for** $s \in \mathcal{A} \cup \{\emptyset\}$ **do** $\{s = \emptyset$ is to have a model with all signals$\}$
4: $\quad I(s) \leftarrow \emptyset$ {Initialization}
5: $\quad$ **for** $w \in \mathcal{W}$ **do**
6: $\quad\quad$ Remove signals in s from dataset
7: $\quad\quad$ Train as per Alg. 1 and obtain predictions $\left\{Z^k(s)\right\}_{k=1}^{K}$
8: $\quad\quad$ **if** ground truth available **then**
9: $\quad\quad\quad R_k \leftarrow y_{N+k}$
10: $\quad\quad$ **else**
11: $\quad\quad\quad$ { Average predictions of model with all signals }
12: $\quad\quad\quad R_k \leftarrow \sum z^k / M$ where $z^k \in Z^k(s = \emptyset)$
13: $\quad\quad$ **end if**
14: $\quad\quad I(s) = I(s) \cup \left|Z^k(s) - R_k\right|, \forall k \in \{1, \ldots, K\}$
15: $\quad$ **end for**
16: **end for**
17: Rank mean of $I(s)$ for all $s \in \mathcal{A}$ whose means are different from mean of $I(\emptyset)$ with statistically significance α
18: Return ranking of signals based on $I(s)$

reference can be of two types: (1) available ground truth values for target y_{N+k} until time N (lines 8-9), useful to explain strengths of signals in the past; *or* (2) mean of predictions of $s = \emptyset$ (lines 11-12), useful to enable understanding on of signals in the current predictions. To test *statistical significance*, we use two-sample t-test with null hypothesis $H_0 : \mathbb{E}[I(s)] = \mathbb{E}[I(\emptyset)]$ (line 17). This test will tell if removing signals in s will truly change predictions. If we fail to reject H_0, then s is removed from the ranking output.

Interface: To enable interaction in the process of understanding the predictions, we constructed a web-based graphical user interface using Vega, a tool for creating web-based interactive visualization designs. The user interacts with our interface (see Fig. 1) by setting which region to analyze. The system retrieves a ranking of the signals whose removal impact the most to the predictions (output of Alg. 2). Then, the user can visualize the actual predictions along with the input predictors for a selected geographical region.

Enabled Analysis: Together with Alg 2 and the interface, our module enables the following two-fold analysis, both driving insights that can inform our data module (feedback loop in Fig. 1). *Real-time insights:* With our interface, a user can understand which signals are driving the predictive behavior (e.g. trends, slope). In addition, we can reason about what is happening with this group of signals. Important insights can be that an important group of signals is displaying erratic or unreliable behavior. If such signals are found, then we might choose not to include it in our model or we might want to see if there are any issues with the collection, cleaning or transformations (i.e. send them back to the data module). *Retrospective insights:* Our interface also enables analysis of signal strengths in past predictions. We can evaluate how we could have done in the past given that we had removed some signals (using the training data available in that week). Therefore, we can understand which signals have a positive contribution to our performance and which others have a negative contribution, which ultimately informs our data module.

5 Empirical Results

We first present the metrics used to evaluate predictive performance, and then make quantitative and qualitative observations about our performance and properties of our forecasts.

Setup. All experiments are conducted with a Linux machine of 40 processors Intel Xeon CPU E5-2698 v4 @ 2.20GHz, with 252 GB of RAM. Training and obtaining predictions is fast, for a single target takes about 1.5 min. All the results are based on the real-time forecasts submitted during three months (May to August 2020).

Metrics. In epidemic forecasting, predictive performance is usually measured for both point estimates and confidence interval of our probabilistic distribution of predictions (Tabataba et al. 2017). Hence we utilize the one metric to evaluate each aspect of the predictive power of our method.

For measuring performance of our point estimates, we use the mean absolute percentage error measures the average of absolute percentage error, i.e., $\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N} \mid \frac{e_i}{y_i} \mid$ and its value describes how large on average the error is, compared with the actual value.

For measuring performance from a probabilistic perspective, we adopt the probabilistic interval performance metric used in the COVID-19 Forecast Hub introduced in (Bracher et al. 2020). Given the central $1-\alpha$ prediction interval, the interval score Γ_α is computed as follows:

$$\Gamma\alpha = (u-l) + \frac{2}{\alpha}(l-y)\mathbb{1}(y<l) + \frac{2}{\alpha}(y-u)\mathbb{1}(y>u)$$

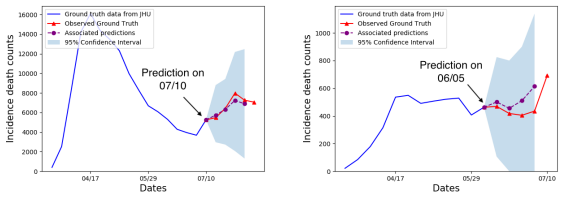
where y is ground truth, $\mathbb{1}(\cdot)$ is the indicator function, l corresponds to the $\frac{\alpha}{2}$ confidence interval number, and u corresponds to the $1-\frac{\alpha}{2}$ confidence interval number.

Questions. With the observations in this section, we aim to demonstrate the our framework DEEPCOVID achieves our goals introduced in Section 1. Specifically, our questions are:

Q1. Is DEEPCOVID able to anticipate trend changes? (**G1**: first goal from Section 1 , **G2**)

Q2. Does DEEPCOVID capture finer grain patterns? (**G1, G2**)

Q3. How does DEEPCOVID perform in US National short-term forecasting? (**G1, G3**)



(a) US peak prediction (b) CA uptrend prediction

Figure 2: Two examples when we were able to anticipate the upcoming change of trends.

Q4. Does DEEPCOVID's emphasis on short-term forecasting sacrifice longer-term performance? (**G3**)

Q5. Can DEEPCOVID explain its predictions to epidemiological experts for interpretation? (**G4**)

Remind that the main goal of the paper is to explore the utility of purely data-driven models for emerging pandemics. Questions Q1-4 directly address this, and Q5-6 are related to communication, an important asset in a forecasting model.

5.1 Observations

Observation 1 *[Q1]* DEEPCOVID *is able to anticipate important changes in trends several weeks ahead.*

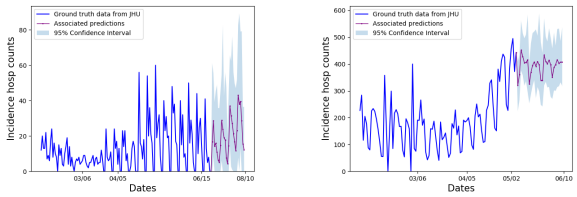
We are able to anticipate important changes in trends several weeks ahead, which suggests that we achieve goals **G1** and **G2** thanks to our capability of exploiting many heterogeneous signals that are sensitive to what is happening in the ground. Most notably, in Fig. 2(a), we can see we predicted the second peak value and time for US National three weeks early. Our method also predicted with three weeks of anticipation that California, after a stable period, was going to suffer a new increase on deaths (epidemic onset) (Fig. 2(b)).

Observation 2 *[Q2]* DEEPCOVID *is able to capture finer grained reporting patterns.*

One of the advantages of purely data-driven is that they are able to capture micro patterns in time series data, which brings a complementary forecasting perspective (goals **G1** and **G2**). In forecasting daily hospitalizations, we noticed many regions have the following patterns: P1: drop during weekends; P2: rise on Monday, and continues stable during weekdays. DEEPCOVID is able to capture them as noted in Fig. 3. In fact, our model is the only approach submitting forecasts to the CDC that captures these patterns (cdc 2020a).

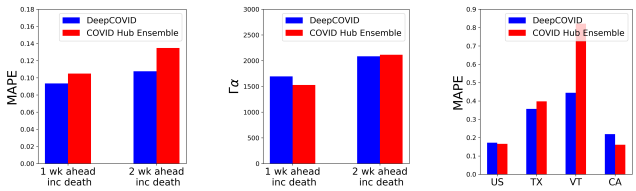
Observation 3 *[Q3]* DEEPCOVID *excels in US National short-term forecasting.*

Here we compare against the official ensemble of all contributing models in the COVID-19 Forecast Hub (including ours). The ensemble has been regarded as one of the best performing models by different independent assessments published on the Web. Needless to say, national-level forecasts are crucial for federal decision makers and are the most visible forecasts in national media. In Fig. 4(a), DEEPCOVID clearly outperforms this very strong baseline in 1- and 2-week



(a) Oregon (b) Florida

Figure 3: Two examples of finer grained reporting patterns captured by DEEPCOVID. Note the dips in reporting for weekends.



(a) 1-2 wk ahead (b) 1-2 wk ahead (c) 1-4 wk ahead

Figure 4: (a) DEEPCOVID outperforms the official ensemble in US National short-term (1-2 wk ahead) forecasting in MAPE. (b) Our in US National short-term confidence intervals are close probabilistic metric $\Gamma\alpha$ with $\alpha = 0.7$. (c) Our focus on short-term predictions does not compromise longer-term (1-4 wk ahead) performance in multiple geographical regions.

ahead across three months. Fig. 4(b) indicates probabilistic performance of our confidence intervals. The fact that we are close suggests we are propagating the uncertainties in the right way. These results show our success in goals **G1** and **G3**.

Observation 4 *[Q4]* DEEPCOVID *does not compromise longer-term performance.*

In Fig. 4(c), we can notice that our excellence in short-term forecasting does not compromise when we consider longer-term predictions in the evaluation. We are very close to the ensemble in US National but at state-level our performance is mixed. States such as Texas (TX), Vermont (VT), but in some other states such as California (CA) we are far from this strong baseline. This is indicative that there are still open questions for forecasting in lower-level granularities, where some signals are more bursty. This is also a good example of where mechanistic models can help. This complements our success for goal **G3**.

Observation 5 *[Q5]* DEEPCOVID *explains its predictions to epidemiological experts for interpretation.*

Our explainability module has been key to enable the previously shown quantitative and qualitative performance in real-time forecasting, meeting goal **G4**.

For example, to predict the second US peak in Fig. 2(a), we were able to understand that mobility was the main signal driving this prediction with clear statistically significance ($\alpha = 0.05$), followed by testing, whose predictions were partially statistically significant. By July, it was unclear if mobility patterns were still capturing social distancing measures, however, the fact testing signals (which were rapidly increasing in the US) were also contributing to predict the same peak time allowed us to have more confidence in this peak prediction. For predicting the uptrend in Fig. 2(b), using our explainability module, we found that impact of line-list, mobility and exposure were statistically significant (with $\alpha = 0.05$) and removing them one at a time did not change the uptrend, which suggests that signals from different groups are being utilized to predict the uptrend. This gave us confidence in our predictions of such important characteristic of the epidemic trajectory.

6 Discussions and Conclusion

In this paper, we introduced DEEPCOVID, an operational DL driven framework for real-time COVID forecasting, whose predictions have been submitted to the CDC via COVID-19 Forecast Hub on a weekly basis since April 2020. This was the first of purely data driven and DL approach to be submitted to the COVID-19 Forecast Hub. DEEPCOVID exhibits interpretability, encouraging short-term and trend performance, principled uncertainty estimation, correlation between forecasts, and ingestion of several data sources despite the chaotic and fast-moving pandemic scenario which naturally brings several modeling and data challenges.

There are several lessons learnt from our experience apart from the observations we have already noted above. First is the lack of standard data reporting, even among traditional sources (e.g. several states do not report new hospitalizations, or report deaths in different ways causing different lags). This means that while some artifacts can be handled statistically (lags), some mild assumptions are also needed to do meaningful predictions (like for hospitalizations) which are still useful. Second, due to the chaotic situation, a purely data driven model needs to be regularly (weekly) updated to reflect the changing dynamics and data quality, to ensure good performance. Third, explainability is very important, as it can serve as a sanity check based on domain knowledge and also highlight *why* we are making forecasts at different points of the pandemic e.g. we found that testing and mobility data signals were most important for predictive performance on average. Finally, we also found that data revisions are pervasive, and even the *ground truth* can be revised (see appendix), which implies measures of performance may be unreliable till data stabilizes. As future work we plan to extend this methodology to handle such revisions, and also work more robustly at smaller scales (e.g. county level).

Acknowledgments

This paper is based on work supported by the NSF (Expeditions CCF-1918770, CAREER IIS-1750407, RAPID IIS-2027862, Medium IIS-1955883, NRT DGE-1545362), CDC MInD-Healthcare U01CK000531-Supplement, funds from Georgia Tech Research Institute (GTRI) and funds/computing resources from Georgia Tech.

References

2020a. CDC Consolidated Forecasts COVID-19 Hospitalizations 2020-09-07. URL https://www.cdc.gov/coronavirus/2019-ncov/downloads/cases-updates/Consolidated-Forecasts-Hosp-2020-09-07.pdf.

2020a. *COVID-19 Community Mobility Report*. URL https://www.google.com/covid19/mobility?hl=en.

2020b. *COVID Symptom Survey*. URL https://cmu-delphi.github.io/delphi-epidata/.

2020c. *The COVID Tracking Project*. URL https://covidtracking.com.

2020d. *COVID-19 - Mobility Trends Reports*. URL https://www.apple.com/covid19/mobility.

2020. *JHU COVID Dashboard*. URL https://coronavirus.jhu.edu/map.html.

2020b. *Weekly U.S. Influenza Surveillance Report*. URL https://cdc.gov/flu/weekly/index.html.

2020. Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree. URL https://projects.fivethirtyeight.com/covid-forecasts/.

Adhikari, B.; Xu, X.; Ramakrishnan, N.; and Prakash, B. A. 2019. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD*, 577–586.

Altieri, N.; et al. 2020. *Curating a COVID-19 data repository and forecasting county-level death counts in the United States*.

Baek, J.; et al. 2020. *The Limits to Learning an SIR Process: Granular Forecasting for Covid-19*.

Box, G. E.; Jenkins, G. M.; Reinsel, G. C.; and Ljung, G. M. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.

Bracher, J.; et al. 2020. Evaluating epidemic forecasts in an interval format. *arXiv preprint arXiv:2005.12881* .

Brooks, L. C.; et al. 2015. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS computational biology* 11(8): e1004382.

Brooks, L. C.; et al. 2018. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLOS Computational Biology* 14(6): e1006134.

CDC. 2020. *Coronavirus Disease 2019 (COVID-19)*. URL https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/about-epidemiology/monitoring-and-tracking.html.

CDC. 2020. *COVID-19 Forecasting in the U.S.* URL https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasting.html.

Chinazzi, M.; et al. 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 368(6489): 395–400.

Couture, V.; Dingel, J. I.; Green, A. E.; Handbury, J.; and Williams, K. R. 2020. Measuring movement and social contact with smartphone data: a real-time application to COVID-19. Technical report, National Bureau of Economic Research.

Dong, E.; et al. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20(5): 533–534.

Ginsberg, J.; et al. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232): 1012.

Holmdahl, I.; and Buckee, C. 2020. Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us. *New England Journal of Medicine* 383(4): 303–305.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. volume 37 of *Proceedings of Machine Learning Research*, 448–456. PMLR.

Jha, A.; Ray, S.; Seaman, B.; and Dhillon, I. S. 2015. Clustering to forecast sparse time-series data. In *ICDE, 2015*, 1388–1399. IEEE.

Miller, A. C.; et al. 2018. A smartphone-driven thermometer application for real-time population-and individual-level influenza surveillance. *Clinical Infectious Diseases* 67(3): 388–397.

Murray, C.; et al. 2020. Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries. *medRxiv* .

Osthus, D.; et al. 2019. Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy (with discussion). *Bayesian Analysis* 14(1): 261–312.

Ray, E. L.; et al. 2020. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. *medRxiv* .

Reich, N. G.; Niemi, J.; House, K.; d Hannan, A.; Cramer, E.; Horstman, S.; et al. 2020. *covid19-Forecast-Hub: Pre-Publication Snapshot*.

Reich, N. G.; et al. 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences* 201812594.

Shaman, J.; Goldstein, E.; and Lipsitch, M. 2010. Absolute humidity and pandemic versus epidemic influenza. *American journal of epidemiology* 173(2): 127–135.

Shaman, J.; and Karspeck, A. 2012. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences* 109(50): 20425–20430.

Tabataba, F. S.; et al. 2017. A framework for evaluating epidemic forecasts. *BMC Infectious Diseases* 17(1): 345.

Tamerius, J. D.; et al. 2013. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS pathogens* 9(3): e1003194.

Tizzoni, M.; et al. 2012. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC medicine* 10(1): 165.

Venkatramanan, S.; et al. 2018. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics* 22: 43–49.

Volkova, S.; Ayton, E.; Porterfield, K.; and Corley, C. D. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one* 12(12): e0188941.

Wang, L.; Chen, J.; and Marathe, M. 2019. DEFSI: Deep learning based epidemic forecasting with synthetic information. In *Proceedings of AAAI*, volume 33, 9607–9612.

Yuan, Q.; et al. 2013. Monitoring influenza epidemics in china with search query from baidu. *PloS one* 8(5): e64323.

Zhang, Q.; et al. 2017. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of WWW*, 311–319.

Zou, D.; et al. 2020. Epidemic Model Guided Machine Learning for COVID-19 Forecasts in the United States. *medRxiv* .

Appendix

A1: Data Module (More Details)

Detailed Description of Signals (C2)

Next, we give a brief description about what each type of signal represents, its time and spatial granularity, and rationale behind using such signals.

1. *Line list* signals (DS1): These signals are derived from records that report who, when and where facets of an infected person; thus, they are directly related to the disease spread. This type includes number of persons infected, hospitalized, and recovered (or deceased). Others report visits of emergency room (ER) due to influenza-like illness (ILI) symptoms and COVID-like illness (CLI). Signal 8 represents the difference between the observed deaths (overall, not only COVID-related) from the expected deaths in during specific time periods.
2. *Testing* signals (DS2): These type of signals reflect social policy (local leader's efforts for escalating testing) and social behavior, e.g. people with fever caused by any other disease may request to be tested for COVID. Here, we have signals that report number of tests (total and negative), number of emergency facilities and health care providers reporting data in *DS1*. The latter two are measures of the perceived importance of the disease by health care providers, as it has been previously observed in influenza forecasting (Brooks et al. 2015).
3. *Crowdsourced symptomatic* signals (DS3): This signal is collected from individuals using Kinsa digital thermometers at home who present fever and influenza-like symptoms, which have a significant overlap with COVID-like symptoms. This signal is potentially an indirect measure of both reported and unreported COVID cases.
4. *Mobility* signals (DS4): These signals are collected from the record of people visits to points of interest (POI) in different regions. Google collects the daily change of visits in multiple categories of POI compared with the period January 3-February 6, 2020 (signals 16-21 in Table 1). We also collected a daily change of visits from Apple (signal 22), which indicates the relative volume of direction requests in the geographic map compared to January 13 across different US states. Mobility signals implicitly show the impact of different non-pharmaceutical interventions (NPI) and change of policies adopted by different states.
5. *Exposure based* signals (DS6): The signals measure social contacts in different groups of individuals. The records have been collected from tracking the overlapping location of distinct smartphone devices in commercial venues. They estimate the signals considering standard sample (signal 23) and removing biased samples (signal 24), i.e., removing the devices from sample which have less movement due to stay-at-home order. The signals implicitly indicate the impact of social contacts and NPI interventions on COVID-19 cases.
6. *Social Media* signals (DS6): Facebook collects the daily percentage of people with covid-like-illness (%CLI) and influenza-like-illness (% wILI)(see signals 35-36 in Table 1) across national level and different states of US. They estimate this percentage from the records of voluntary surveys based on disease symptoms.

Addressing missing values in incident hospitalizations (C5)

While collecting incident hospitalizations for forecasting target *T2*, we observe that daily records of 11 states (CA, DC, TX, IL, LA, PA, MI, MO, NC, NV, DE) are missing. Instead, there are other signals such as the current records of hospitalizations and ICU patients. However, assuming current hospitalizations as target *T2* would affect our forecast by over-estimating incidences. Hence, our goal was to find a way to estimate *T2* from the current hospitalizations for these 11 states. We tried a compartmental model based on recovered cases and deaths to address our goal. However, we found this challenging to do with the available data because recovered cases are not only considering people that was hospitalized, and the number of deaths is being reported with a different delay than hospitalizations (because they come from different sources). Therefore, as we could not include recoveries nor deaths, we assumed that after a week only a fraction β of the current hospitalizations at time $t-7$ would stay in the hospital. Therefore, we used

$$hosp_{inc}(t) = hosp_{cur}(t) - \beta \cdot hosp_{cur}(t-7)$$

where t is time in days, $hosp_{inc}$ are the incident hospitalizations, and $hosp_{cur}$ are the current hospitalizations. We used grid-search and found $\beta = 0.5$ closely matches with the ground-truth incident hospitalizations of the states that do not have missing values.

A2: Additional Empirical Results

Observation 6 *Mobility and testing signals were found as the most important for predictive performance.*

Our explainability module also allow us to communicate domain-experts which signals are helping the most (goal **G4**), and provide feedback to the data module for selection of signals.

By analyzing several geographical regions over the span of three months, we noticed that the largest negative effect comes from removing mobility and testing, while removing line list had positive effect for some states. However, each geographical region requires its own optimized set of data signals. In Table 2, when looking at individual geographical regions, contribution on performance varies. For example, testing has a large positive contribution in US National, but has a negative contribution in Texas. Therefore, each region requires a different treatment to optimize its performance based on the input data signals.

Observation 7 *Importance of signals change as the spread of the disease progresses.*

DEEPCOVID With our explainability module, we can communicate temporal insights about the importance of signals (goal **G4**).

Table 2: Contribution of signals in 1-4 wk ahead forecasting performance (retrospective analysis) for US National and four states. We present the t-statistic to measure the contribution (higher t-stat, higher contribution). Green indicates positive contributions, red negative contributions, and black non-statistically significant contributions.

	1-4 wk ahead			
	US	CA	GA	TX
Line-list	3.51	6.09	1.34	-14.17
Mobility	10.02	9.40	3.28	2.85
Testing	6.10	8.77	2.66	2.63
Exposure	3.02	1.11	-0.82	0.04

For instance, we analyzed mobility in predictions for US National and California, two regions where mobility was found to have a positive contribution (see Table2). We considered two periods: (1) May to June, when stay-at-home orders were lift in most states, businesses reopened, and mobility signals increased; (2) July to August, a period when most mobility signals have already stabilized. We noticed a high contribution of mobility during (1) in both US National and California, and low or non-existing in (2). This observation suggests that data driven model needs to be regularly updated to reflect the changing dynamics of the disease spread.

A3: Data Revisions

Several of our data sources, especially disease surveillance data from health agencies, report an initial value that undergoes several rounds of revisions to reach a stable value, a process which typically lasts several weeks. As also noted by (Reich et al. 2019), our experiments suggest that these data revisions have an impact into real-time forecasting performance.

To illustrate this issue, we computed the revision error $|v_w - v_s|/v_s$, where v_w is the value at revision week w ($w = 0$ denotes first release) and v_s is the stable value. When we average this error over several past observations, we get a time series over revision weeks, which allow us to see how many revision weeks in average it takes for a signal to stabilize. In Fig. 5, we can notice that even our ground truth target, reported death incidences (JHU), exhibit this issue, which implies measures of performance may be unreliable till data stabilizes.

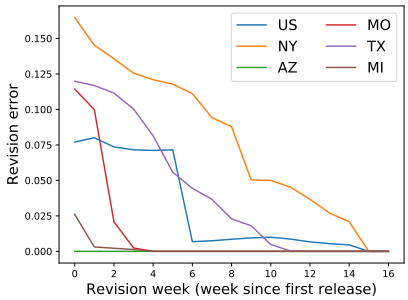


Figure 5: Average revision error for ground truth reported deaths from JHU (our forecasting target). We can see there are a few clusters of geographical regions. Arizona (AZ) and Michigan (MI) have none to minor data revisions problems, which are resolved promptly. Montana (MO) has large revision error, but it stabilizes rapidly. US National and Texas (TX), and New York have large revision error and are slow in stabilizing, taking up to 15 weeks.