Forecasting COVID-19 Cases via Multimodal Machine Learning Models

Michael Li

The objective of this research is to explore **several machine learning models** and **methodologies** to determine whether they are effective in forecasting **daily cumulative** COVID-19 case counts at the **county** level. Forecasting COVID-19 cases empowers health-officials to take proactive measures, stopping potential outbreaks. We explore an **ANN**, **TCN** and **multimodal** model (which incorporates county-level characteristics). We also explore several forecasting methodologies, and compare a **county-county** with a **past-future** method. We find that the TCN model and past-future method, when evaluated with MAE and MAPE, perform the best. The TCN model outperforms **existing literature**, and also **outperforms** the **ensemble model** used by the **Centers for Disease Control and Prevention (CDC)** for COVID-19 forecasting. The novel contributions of our research are three-fold: we determine that TCN performs well when forecasting COVID-19 cases, validate the county-county and past-future methods, and provide accessible and high-quality COVID-19 forecasts to county health officials through a dashboard.

Mathematics/Computer Science

Introduction - Objective and Literature Review

<u>Goal</u>: To develop a machine learning model that can forecast daily cumulative COVID-19 cases to inform health officials in preventive measures.

Prediction of daily COVID-19 cases in European countries using automatic ARIMA model (Awan, Aslam 2020) **[1]**

- The <u>ARIMA model</u> is used to forecast <u>daily European</u> COVID-19 cases
- The study finds that ARIMA can satisfactorily predict cases for the <u>next ten days</u>

COVID-19 prediction using LSTM algorithm: GCC case study (Ghany et. al 2021) **[2]**

- The <u>LSTM</u> model is used to predict COVID-19 cases in the <u>Gulf Area</u> (Bahrain, Qatar, etc.)
- Model evaluation is done with <u>root mean square error</u> (RMSE) and <u>mean absolute relative error</u> (MARE).

Machine Learning Approaches for COVID-19 Forecasting (Istaiteh et. al 2020) **[3]**

- This paper uses the <u>ARIMA</u>, <u>ANN</u>, <u>LSTM</u> and <u>CNN</u> models to predict COVID-19 cases across 189 countries.
- CNN performs the best, evaluated with <u>mean absolute percentage error</u> (MAPE), <u>root mean squared logarithmic error</u> (RMSLE) and <u>mean squared logarithmic error</u> (MSLE)

Introduction - Prior Work

Last Year's Work

- Detecting COVID-19 using symptoms and cough test
- Preprocessed **2.7 million** Israeli government symptoms dataset and **1,400** aggregated cough dataset
- **XGBoost** - boosted decision-tree, used to predict based on health form
- **VGG-19** - state of the art CNN, used to predict based on cough recording
- Built webapp for users to check symptoms
    - Developed health form and cough recording flows

This Year's Work

- Modeling COVID-19 time series behavior based on past case counts
    - JHU CSSE dataset contains **cumulative** daily records of COVID-19 case counts across **3112** counties within the United States from January 2020 to March 2022. [4]
    - **Models: ANN**, **TCN**, **Multimodal**
    - Explore **Past-Future** and **County-County** forecasting methods
- Changed UI and built new page on webapp:
    - **Live** COVID-19 dashboard showing **1-month** forecasts for **all** U.S. counties
    - https://www.c0vidcatcher.org/dashboard

Framework - Concepts and Definitions

- **Data processing** - Aggregate and clean data; extract important features and labels.
- **MAE** - Mean absolute error; sum of all errors divided by total number of observations
- **MAPE** - Mean absolute percentage error; sum of all percent errors divided by total number of observations
- **Model development** - Machine learning models were built and tested on the data. Models were validated with MAE and MAPE
- **Feedforward Artificial Neural Network (ANN)** - Baseline model, made of dense layer, dropout layer and output layer
- **Temporal Convolutional Network (TCN)** - TCNs are a new class of temporal models, first proposed by Lea et al. in 2017 [5]
- **Multimodal model**- Multimodal model that combines TCN and ANN, and incorporates county-level characteristics into the ANN

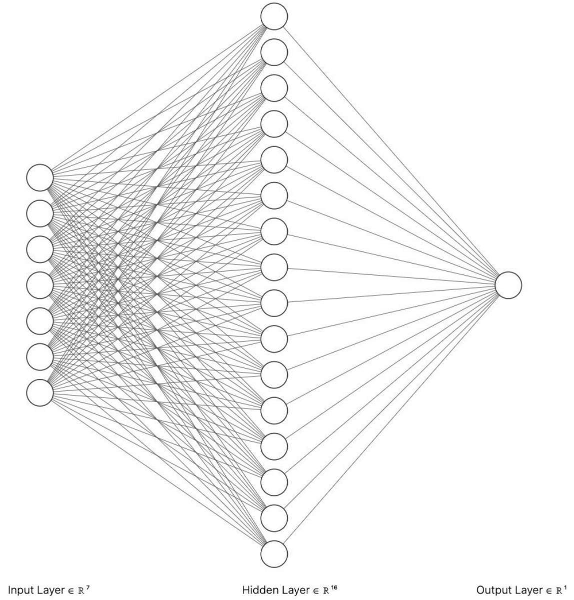


Figure 1. ANN Model Architecture

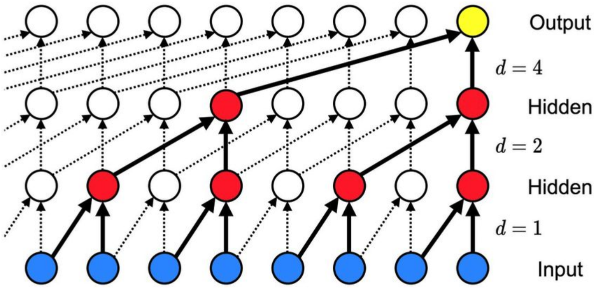


Figure 2. TCN Architecture[4]

Created by student researcher (Figure 1)

Framework - Methodology

Model Development

1. **Process training dataset**
    a. Use data from each county starting from at least 5 total cases
2. **County to County Forecasting**
    a. Forecast the entire COVID-19 case history of a single excluded county using the collective history of all the other counties.
    b. 80/20% split on the counties to separate the dataset into training (2490 counties) and test (622 counties) set
3. **Past to Future Forecasting**
    a. Forecast future COVID-19 cases using the past collective data of all counties.
    b. 2 most recent months excluded from training dataset and used for testing
4. **FPC and Percent Change**
    a. Include dimensional reduction score, and predict percent change instead of cumulative

Front-end COVIDCatcher Development

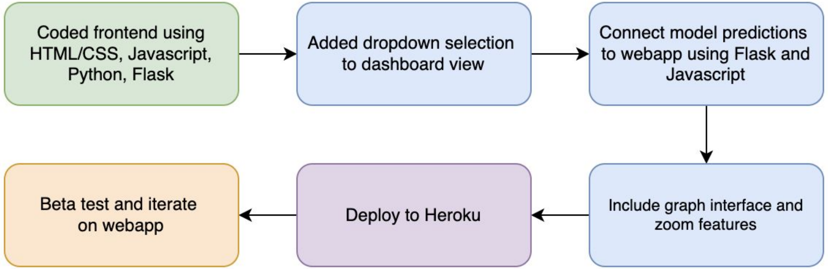


Figure 3. Webapp development flow. Predictions are stored in CSV and loaded to a web app in Heroku with remote hosting

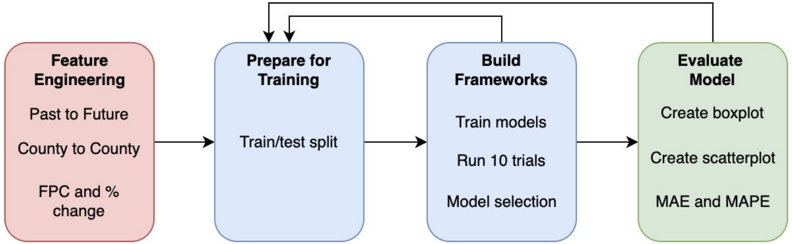


Figure 4. Backend development flow.

Created by student researcher (Figures 3 and 4)

Preliminary Findings

FPC Score and Percent Change

- Using **FPC** score as a feature caused model performance to **significantly decline**
- Using **Percent change** as the label also **decreased** model performance

MAE	
26.68	multi_model_all_county_722
34.62	multi_model_722
48.46	ff_model_7
576.78	multi_model_71_sublayer_22
2131.25	ff_model_30
5252.78	multi_model_71
7756.89	multi_model_71_22
9801.96	multi_model_71_sublayer_all_county_22
19249.86	ff_model_8

Figure 6. Model results for dimensional reduction score. Top 3 models are trained without using the score. The rest use the FPC score.

Figure 7. Predicted vs. Actual (% Change) results for ANN Model on three random counties. Predictions are very unstable.

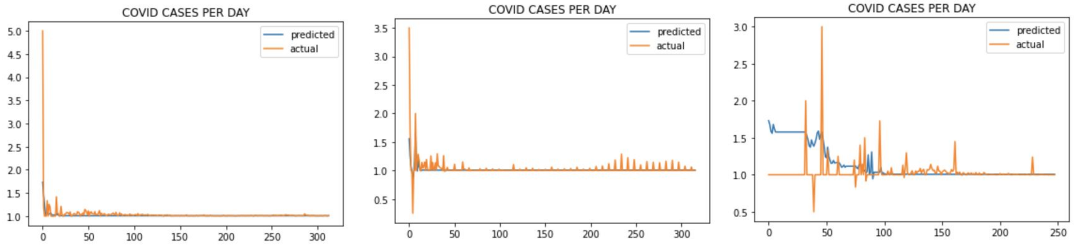
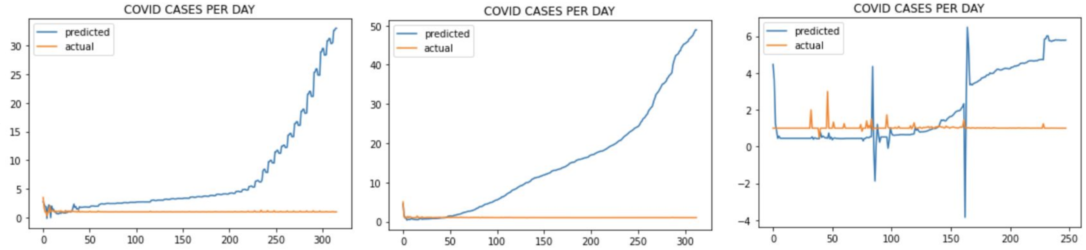


Figure 8. Predicted vs. Actual (% Change) results for TCN Model on three random counties.



Findings

<u>County to County Forecasting</u>

- TCN performs **slightly better** than multimodal

Models	MAE	MAPE
TCN	19.72	0.014%
Multimodal	20.23	0.018%

Table 1. MAE and MAPE for TCN and Multimodal trained on the County to County dataset (train/test split contains 2490 train and 622 test counties)

<u>Past to Future Forecasting</u>

- TCN performs **slightly better** than multimodal
- **Past to Future outperforms** County to County in terms of MAPE

Model	MAE	MAPE
TCN	38.68	0.0076%
Multimodal Model	35.55	0.0078%

Table 2. MAE and MAPE for TCN and Multimodal trained on the Past to Future dataset (train excludes last two months to use for evaluation)

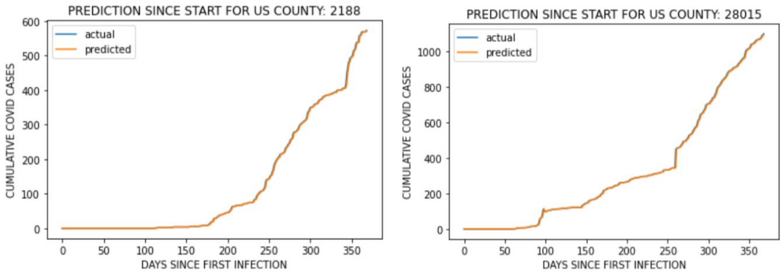


Figure 9. (County to County) TCN Model Graph Actual Cases v. Predicted Cases, entire case history predicted for county

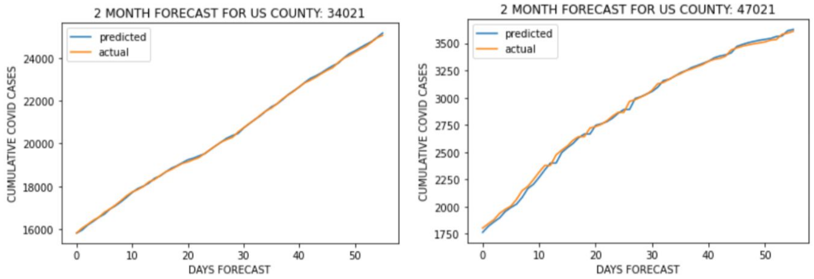


Figure 10. (Past to Future) TCN Model Graph Actual Cases v. Predicted Cases, latest 57 days predicted for each county after being trained on all past history

Created by student researcher (Figures 11 & 12, Tables 3)

Findings

Direct Model Comparison

- We directly compare our two models against each other
- We train TCN and multimodal models for **10 trials**, and on the same train/test split for each of the 10 trials, with new splits made for each trial
- An **independent 2-sample t-test** was performed
- TCN has a **statistically significant** lower mean MAE than the multimodal model

Group	Mean	Standard Deviation
TCN	19.08	4.67%
Multimodal Model	26.45	4.66%

Table 3. Trial variables for performing significance test. TCN has a lower mean MAE.

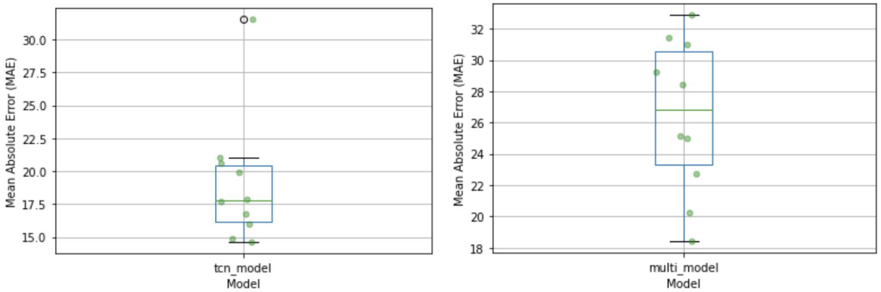


Figure 11. TCN and Multimodal Model Boxplot of MAE. The horizontal lines represents the min, max, 1st quartile, median and 2nd quartile.

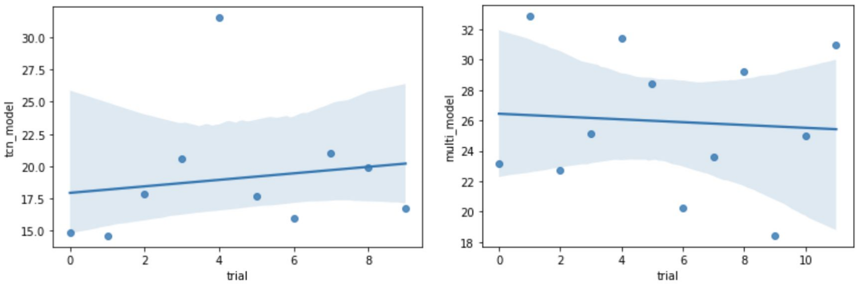


Figure 12. TCN and Multimodal Model Scatterplot of MAE. A best fit line is shown in blue, with a 95% confidence interval shaded in.

Conclusions

Literature Review

- Current literature uses the ARIMA, LSTM, ANN and CNN models to predict COVID-19 cases [1][2][3]
- **Autoregressive Integrated Moving Average (ARIMA) [6]** - time series forecasting algorithm that uses lagged moving averages to analyze it's past values
- **Long Short-term Memory (LSTM) [7]** - artificial recurrent neural network that uses feedback connections
- **Convolutional Neural Network (CNN) -** contains a series of convolutional layers with activation functions intersperses
- **Results**
  - TCN outperforms ARIMA in MAE and MAPE (**55.62** & **0.016%** vs **629.26** & **0.099%**)
  - TCN outperforms the best performing model in literature (ANN) by nearly **2 times** the MAE and MAPE

Model	MAE	MAPE
TCN	55.62	0.016%
ARIMA	629.26	0.099%

Table 4. Performance of TCN in terms of MAE and MAPE errors versus ARIMA.

Model	MAE	MAPE
LSTM	56.84	0.047%
CNN	724.53	0.349%
ANN	49.26	0.035%
TCN	27.82	0.019%
Multimodal	74.23	0.036%

Table 5. Performance of TCN and multimodal model in terms of MAE and MAPE errors versus LSTM, CNN, and ANN.

Created by student researcher (Tables 4 & 5)

Conclusions

Centers for Disease Control and Prevention

- The CDC uses an ensemble model to forecast cases [8][9]
  - Ensemble model outputs average of predictions of 20+ models from institutions like MIT and Caltech
- Our TCN model outperforms the CDC's ensemble model by an average of ~15 cases.
- I will continue to reevaluate my model with the CDC's to validate my initial findings

Model	MAE	MAPE
TCN	94.54	0.0078%
CDC Ensemble	109.45	0.0588%

Table 6. Performance of TCN vs CDC ensemble model in terms of MAE and MAPE errors

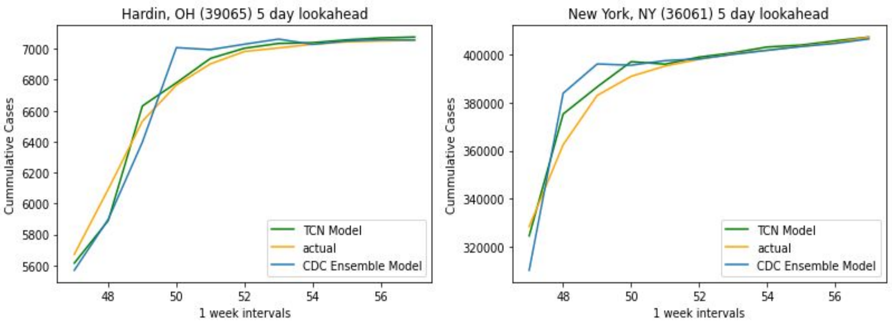


Figure 14. Two graphs of cumulative cases comparing the TCN model's predicted cumulative cases and CDC ensemble model's predicted cumulative cases with the actual cases. TCN follows the real case trend more closely than the CDC ensemble.

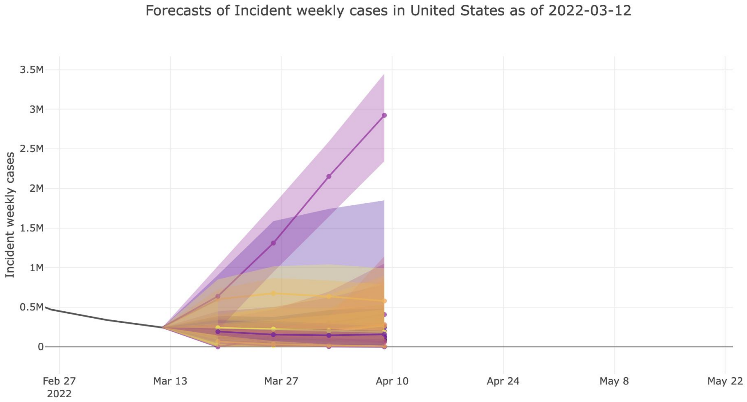
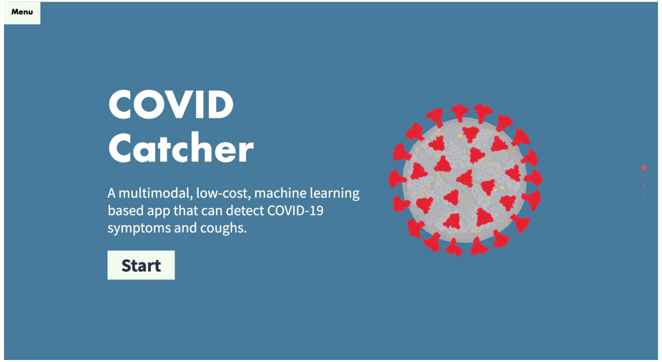


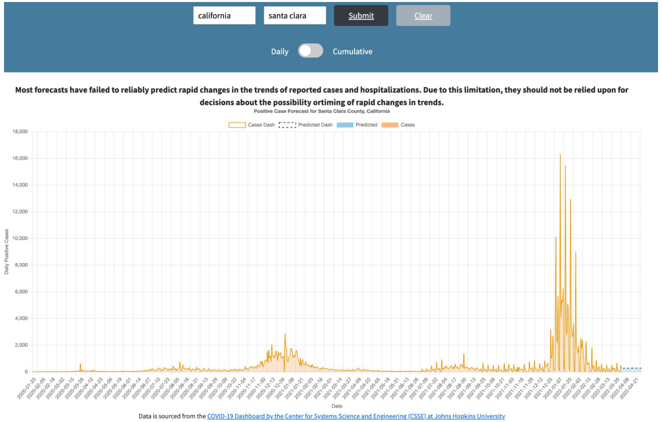
Figure 13. Graph of 20+ model's predictions (each a different color) that are used in the CDC ensemble model. There is a clear outlier in the purple line forecast.

Created by student researcher (Figures 13 & 14, Table 6)

Conclusions

1. To help **health officials** make **data-informed decisions** on lockdown measures, I developed a **free and accessible dashboard** that delivers **high-quality** forecasts of **daily cumulative** COVID-19 cases
2. **TCN** performs the best across our models and models in existing literature
3. COVIDCatcher offers a **free**, machine learning-informed **dashboard**
4. **County to County** and **Past to Future** are viable and accurate methods of prediction.
5. Using a dimensional reduction feature and predicting percent change in daily cumulative cases are not viable optimizations.
6. COVIDCatcher is easy to use, accessible to the general public and health officials, and can be used at **https://www.c0vidcatcher.org**





References

[1] Awan TM, Aslam F., “Prediction of daily COVID-19 cases in European countries using automatic ARIMA model.” 2020

[2] Ghany KKA, Zawbaa HM, Sabri HM. “COVID-19 prediction using LSTM algorithm: GCC case study.” 2021

[3] O. Istaiteh, T. Owais, N. Al-Madi and S. Abu-Soud, "Machine Learning Approaches for COVID-19 Forecasting." 2020

[4] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis; published online Feb 19. https://doi.org/10.1016/S1473-3099(20)30120-1.

[5] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, G. D. Hager, “Temporal convolutional networks for action segmentation and detection.” 2017

[6] Harvey A.C. (1990) ARIMA Models. In: Eatwell J., Milgate M., Newman P. (eds) time-series and Statistics. The New Palgrave. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-20865-4_2

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780.

[8] Cramer EY, Huang Y, Wang Y, et al. “The United States COVID-19 Forecast Hub dataset.” 2021. medRxiv. (preprint)

[9] CDC. “Coronavirus Disease 2019 (COVID-19).” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 11 Feb. 2020, https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html.