Development of COVID-19 Regional Forecast Models Using Machine Learning Approaches

Md Mamunur Rashid, Ben D. Sawyer
Department of Industrial Engineering and Management Systems, University of
Central Florida,
Orlando, Florida - 32816, United States
mdmamunur.rashid2@ucf.edu, sawyer@ucf.edu

Abstract

The present work predicts different parameters key to a proper COVID-19 response, through machine learning methods. The number of patients affected, hospitalizations, and deaths in the context of the state of Florida is considered to provide a blueprint for regional pandemic response. To predict these variables, twenty inputs in four categories (type of tests performed, gender, race, and age group) were collected. Official data from the Florida State Department of Health were collected and submitted to a linear regression model, fuzzy logic model, and long short-term memory (LSTM) deep learning model with the intent of producing predictions as close as possible to the actual numbers. The mean absolute percentage error (MAPE) was calculated to measure the deviation of the predicted result from the actual values. In addition, a one-way analysis of variance (ANOVA) models was developed for each output parameter to statistically assess the results of these models. The LSTM deep learning model outperformed the fuzzy model, which in turn outperformed linear regression, in terms of the MAPE for the 'number of Florida residents affected'. For the 'number of patients hospitalized', the LSTM deep learning model again outperformed the regression model, while the regression model and the fuzzy model were not significantly different. For 'the number of patient deaths', however, no model significantly outperformed any other. These findings suggest that in regional data, the fuzzy model outperforms linear regression, and the LSTM deep learning model outperforms both the fuzzy model and the linear regression model. Implications of this work include a better understanding of opportunities and appropriate tools for short-term prediction of future trends when variability is high, as well as replacement strategies for datasets with missing values.

Keywords
COVID-19; LSTM deep learning; Fuzzy logic; Linear regression; MAPE; ANOVA

1 Introduction

On March 11, 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a global pandemic (Arora et al.2020). Epidemics in one region of the world can quickly spread, especially given the high population density and high mobility of the present day (Aristarán et al. 2021). Due to the global pandemic, the world is facing a severe and acute public health emergence (Arencibia et al. 2020). The demand for healthcare services has rapidly escalated with the spread of COVID-19 (Austin et al. 2020). The economies of all nations of the world are facing adverse effects due to COVID-19 (Barbara et al. 2020, Cadogan and Hughe 2021), so the allocation of resources to healthcare must be careful and deliberate. Indeed, modeling public health epidemics is important to understand the potential impacts and encumbrances of the outbreak (Castillo and Melin 2020). Scantamburlo et al. 2021, believe that health authorities and policymakers have been helped by new technology tools to plan and modify containment efforts. To be better prepared to manage the impact of this outbreak, mathematical modeling is of prime importance.

The government and health service providers need accurate predictions of the spread of the virus to design and plan their service responses for different counties and different geographic regions (Fanelli and Piazza et al. 2020) Accurate forecasting of the spread of confirmed cases, as well as the analysis of the number of recoveries and deaths, is required to predict the global impact of COVID-19. To ensure better healthcare service responses, this study aims to develop mathematical forecast models using machine learning approaches to predict the number of confirmed cases, the number of hospitalizations, and the number of deaths per day in Florida based on available and authentic data.

Health crisis management can be guided by mathematical models. In healthcare policy making, forecasting has become the cutting-edge position due to the COVID-19 pandemic (Hamzah et al. 2020). To anticipate the spread of the COVID-19 in the UAE, Seasonal autoregressive moving average and autoregressive conditional heteroscedasticity models which shows trustworthy result. Liu used a growth model to depict the cumulative number of COVID-19 infections in China (Pandey et al. 2020). His model showed that the residual model is not a null plot and could not find a distribution function that fit the actual frequency. The key properties of the blockchain can enable the successful implementation of various use situations, block chain technology can successfully aid in the battle in fighting back the coronavirus, described by Chimmula and Zhang (2020). Ye and Yang (2020) developed an uncertain time series model for predicting the evolution of confirmed cases in China, excluding imported cases. Their model passed the stationarity test and the white noise test, and the result was no paragon. Hamzah et al. used the SEIR model to forecast the COVID-19 outbreak within and outside of China based on daily data and these data indicated that each country's policies and social responsibilities play influential roles in the spread of coronavirus. Therefore, the health crisis must be carefully managed, but none of the developed mathematical models provide sufficiently exact results.

COVID can be modeled well, and various approaches have been tested. Yadaw et al. used machine learning techniques to predict the mortality of patients with COVID-19 treated at the Mount Sinai Health System in New York City (2020), and this model showed high accuracy in terms of the area under the curve (AUC) considering three clinical features: the patient’s age, minimum oxygen saturation, and type of patient encounter. Prediction models combined with disease and patient characteristics help in clinical decision-making in the case of treating many patients who require intensive treatment for multiple weeks. They also recommended collaboration among different researchers, clinicians, and institutes and sharing to help develop better prediction models.

Forecast modeling can help decision-makers understand the pattern of COVID-19 spread to take the necessary precautions to fight the pandemic. Takefuji [19] recommended, to assess the success of particular policy, divide the number of deaths attributable to COVID-19 by the population (in million) and the most successful policy against the covid-19 is built on a strong digital fence, while the physical isolation approach demonstrates the pandemic's extraordinary success. Pandey et al. [20] used the SEIR and regression models for COVID-19 outbreak prediction and suggested that continuous monitoring of weekly forecasts is necessary to prepare for medical services for COVID-19 patients. Chimmula and Zhang developed an automated artificial intelligence and deep learning LSTM network model for forecasting COVID-19 transmission, and the results of the study help to monitor the pattern of the spread and take preventive measures. The authors claimed that the coronavirus spread could be linear instead of exponential if the public health authority closely monitors the affecting sequence and acts accordingly. The logistic and prophet model was proposed for the prediction of epidemic trends at the global and country levels by Wang et al. (2020). The proposed model successfully predicted the epidemic peak, fastest growth point, and turning point of recovery, which helped public policymakers develop appropriate intervention plans. Singer showed that the country-specific infection rate follows power-law growth behavior and scaling exponents by analyzing infection data from the top 25 infected countries and concluded that monitoring the prediction of power-law growth can help for health care system planning (Yadaw et al, 2020). Bidirectional LSTM, convolutional LSTM, and deep LSTM models achieved high accuracy for predicting the number of positive COVID-19 cases in 25 states of India. Based on the results of one study, the authors categorized Indian states as mild, moderate, and severe risk zones and proposed statewide lockdowns based on the categories rather than a full country lockdown. The authors also suggested that prediction models could help authorities and planners make better plans for medical services. Fanelli and Piazza applied a simple mean-field model and the susceptible-infected-recovered-death (SIR) model to forecast the spread of COVID-19 in China, Italy, and France. They concluded that the infection rate of COVID-19 depends on many cultural factors, which is why the infection rate and death rate differ from country to country. The authors also expect the models to be useful for political and healthcare authorities during the pandemic. Therefore, forecasting models with sufficient accuracy will surely play a pivotal role in helping policymakers plan health care system responses to mitigate the adverse effects of the pandemic.

Modeling regional-level data is not the same as modeling federal-level aggregations. For example, in the United States, the pattern of detecting affected people can vary greatly on a regional level. Prediction models based on countrywide or larger-scale data might conceal the local dynamics of each state. To combat the pandemic properly, it is important to consider the smallest subsections of the population. Moreover, the federal government empowered the state government to take necessary actions to control the spread of COVID-19 in that state. Therefore, understanding the COVID-19-affected trend of a region is necessary. This study is developed to meet this necessity.

In the present study, we compare several models that we believe will be effective in predicting output parameters based on our regional-level data. Which technique is most effective depends on the purpose? Many techniques have been developed to forecast the number of affected participants (confirmed cases) of COVID-19, e.g., machine learning model by Johns Hopkins University, Center for Systems Science and Engineering, metapopulation SEIR model by Johns Hopkins University, Infectious Disease Dynamic Lab, an ensemble of combined linear and exponential predictors (CLEP) by the Berkeley Yu Group, autoregressive time-series model by Carnegie Mellon University, and the statistical survival-convolutional model by Columbia University and University of North Carolina. However, fuzzy logic modeling has not been tested. In real-world data modeling applications, fuzzy logic has outperformed classical statistical and mathematical modeling techniques [28]. Because of its capability of approximate reasoning and the precise logic of uncertainness, fuzzy logic modeling is applicable for modeling time series data. To make immediate action decisions within a short time window, a combination of fractal theory and a fuzzy logic model was applied to forecast COVID-19 time-series data from 10 different countries with 98% accuracy by Castillo & Melin (2020).

Our choice of fuzzy logic in this study is advantageous due to two noteworthy human capabilities: the capability to converse in an environment of uncertainty and the capability to perform mental and physical tasks without any computations with actual data. Therefore, fuzzy logic modeling is an advantageous and widely used technique for forecasting. In this study, a fuzzy logic model was developed to forecast the number of confirmed cases, the number of people admitted to the hospital and the number of patient deaths due to COVID-19 in Florida. This research tests the compatibility of the developed fuzzy logic model with the output of two other models: a linear regression model and an LSTM deep learning model. Our chosen models' effectiveness is tested on our regional data because different models work differently with different sample sizes and due to other considerations.

Thus, the research question addressed in this study is to determine 'which of the chosen models is most effective for our regional data'. Our null hypothesis is that a common linear regression model will produce the most accurate predictions. Our alternative hypotheses are, first, that our fuzzy logic model will produce the most accurate predictions and, second, that our LSTM deep learning model will produce the most accurate predictions. Therefore, we compare two modeling approaches against linear regression for predicting the number of confirmed cases, the number of patients hospitalized, and the number of patient deaths in Florida.

The models assume that the number of affected participants, the number of hospitalized participants, and the number of patients who died depend only on the factors for which data are available on the official website of the Department of Health of Florida. We expect that our fuzzy logic model, the novel approach that has remained untested for COVID-19 forecasting, will predict the number of affected individuals, the number of hospitalized individuals, and the number of deaths in Florida with accuracy exceeding that of the other two models. The accuracy of the developed model and the comparator models is assessed via K-fold cross-validation.

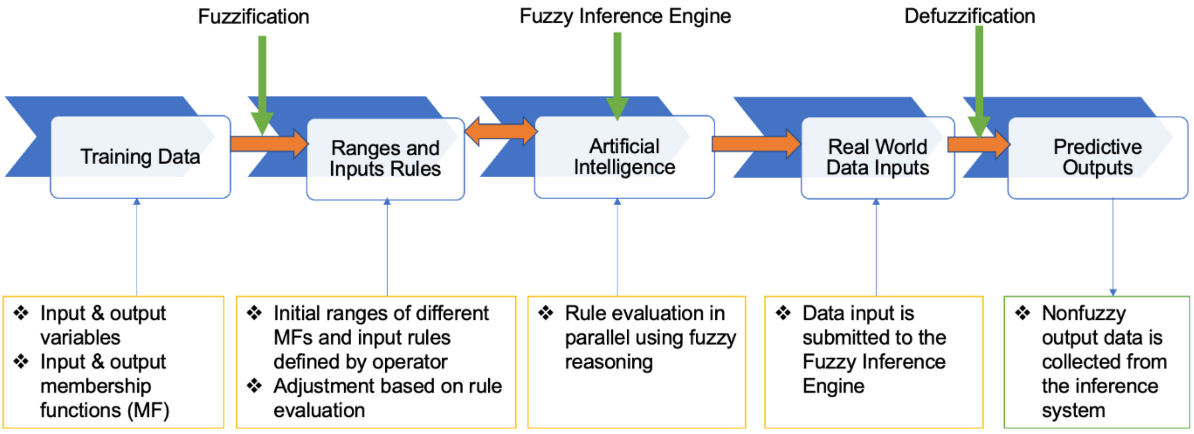
Fuzzification
Fuzzy Inference Engine
Defuzzification
Training Data
Ranges and Inputs Rules
Artificial Intelligence
Real World Data Inputs
Predictive Outputs
❖ Input & output variables
❖ Input & output membership functions (MF)
❖ Initial ranges of different MFs and input rules defined by operator
❖ Adjustment based on rule evaluation
❖ Rule evaluation in parallel using fuzzy reasoning
❖ Data input is submitted to the Fuzzy Inference Engine
❖ Nonfuzzy output data is collected from the inference system

Figure 1. A fuzzy inference system (above) and evaluate data based upon input-output parameters and rule-based relationships

In the system architecture summary of fuzzy inference systems (FIS), shown in Figure 1, real-life data are considered crisp inputs/outputs. The initial step is to define the input and output parameters. Then, the membership functions of the input and output parameters are defined. Every parameter has a membership function. The range of each membership function is selected based on the overall data of each parameter. Based on the training data set, different rules are created to act as the fuzzy system input. FIS evaluates the provided input-output parameters and their relationships based on the rules. Finally, the testing input data are provided, and the FIS provides appropriate specific output values relating to the testing input.

2 Methodology

To determine the number of affected patients, the number of hospitalized patients and the number of deaths, official data from the Department of Health in Florida State were collected. The input parameters were selected as input variables to develop the model. In the fuzzy inference system, the input and output variables were defined, and triangular membership functions were considered with three ranges (High, Moderate, and Low). The parameters and ranges of the input and output functions were categorized with appropriate numerical values. The data set was separated into training and testing data for K-fold cross-validation. A total of 13 folds with 12 days in each fold were considered to develop the model. Input and output variables were interlinked by creating different rules based on the collected training data, and the appropriate formula produced the correct output on the test data. Mean absolute percentage error (MAPE) values were calculated to validate the models and compare the results of the fuzzy logic model, linear regression, and the LSTM deep learning model. Finally, the results were statistically analyzed to test the hypotheses.

2.1 Data collection

Our data reflect the fact that managing healthcare systems for a large population is challenging. Florida is the 3rd most populous state in the United States. The current confirmed cases in Florida ranked 3rd, just after California and Texas, as of August 6, 2021. Therefore, we are interested in determining whether the models are sufficiently accurate to help in planning the health care system response in Florida.

To make decisions and implement public health measures, it is critical to understand the spread of COVID-19 based on reliable data [31]. Data were collected from the website of the Florida Department of Health. (http://ww11.doh.state.fl.us/comm/) from the 'county report' folder under 'case monitoring and pui information' under the 'COVID-19 report archive' in the 'partner' folder of the website. Data were post-processed by the open-source and free offline software Tabula Extractor [32] in a java environment. The data sets were collected in noneditable PDF format, and these data were extracted by Tabula Extractor. Tabula does not work with image-based pdf, only text-based pdf. The software was used to extract data from four tables. Figure 2 shows the sequential used to extract the data. The first step is to run the software, which opens in the browser without installing it on the users' machine. The required pdf is then uploaded, which stores the file into the users' local host of the machine that is being used. The next step is to highlight the data to be extracted. Then, Tabula shows a preview of the extracted data to show whether the data are aligned. After necessary adjustments, the user can save the data into editable.csv or.xls format.

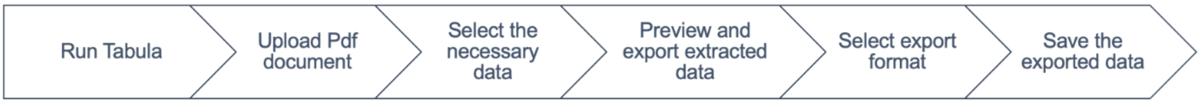

Figure 2. Tabula Extractor to recover data from the only source released by the Florida Department of Health: noneditable PDFs

The resultant dataset includes 20 input parameters in 4 categories: type of test (PCR, Antigen), gender (male, female, unknown), race (white, black, other, unknown), and age group (0-4, 5-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+, unknown). The outputs are the number of affected Florida resident patients, the number of Florida resident patients hospitalized, and the number of Florida resident deaths. Many websites have regularly updated COVID-19 data, but these sources are neither responsible for collecting the raw data nor reliable. To make decisions in a pandemic situation, the accuracy of the data is among the top concerns. As the data were collected from the official website of the Florida Department of Health (FDH), the data we used is the most reliable available data. Over time, as the pandemic continues, the types of data sets available on the official websites of the FDH varied significantly. Although the data are available from April 4, 2020, to the current date, not all the data from the beginning of the pandemic until July 1, 2020, could be used as the format of the data was occasionally modified by the FDH.

2.2 Data segregation for analysis

Our toolset for this study included MATLAB R2018b, which was used to develop the Mamdani FIS [33] for the fuzzy model of this study. Sequential steps were followed in MATLAB to build the model. First, the FIS was designed with 20 inputs and 3 outputs. The input and output parameter data were collected from July 1, 2020 to November 30, 2020. The data are cumulatively updated on the website; therefore, the input-output parameter data were calculated as the difference in cumulative values of two adjacent days. All the data of this study are subdivided into 13 folds with 12 days each, as shown in Table 1.

Table 1. Time spans of the 13 folds in the data set

Fold	Start	End	Fold	Start	End
1	1-Jul	12-Jul	8	23-Sep	4-Oct
2	13-Jul	24-Jul	9	5-Oct	16-Oct
3	25-Jul	5-Aug	10	17-Oct	28-Oct
4	6-Aug	17-Aug	11	29-Oct	9-Nov
5	18-Aug	29-Aug	12	10-Nov	21-Nov
6	30-Aug	10-Sep	13	22-Nov	30-Nov
7	11-Sep	22-Sep			

Every 3rd value was held out as testing data, and the rest were used as training data. Therefore, in a particular fold, the 1st, 2nd, 4th, 5th, 7th, 8th, 10th and 11th days are kept as training data set and the 3rd, 6th, 9th and 12th days are used as testing data.

2.3 Missing data

After the data collection and segregation, there are some anomalies found in the data. There were some days when the data was not uploaded on the website of the Florida Department of Health. There were some days when the value of some parameters was not updated on the website. To overcome this problem, the missing data are filled in the following ways and then validated with hypothesis testing.

2.3.1 Filling of the missing data

1) Missing the data of the whole day: The data set contains some missing values on which dates there is no record available on the FDH website. These dates are September 3, 13 and 28, 2020; 10, 23 and October 25, 2020; and November 26, 2020. To impute the missing data of the above days, it is assumed that the data of the following day include these missing data. Based on this assumption, the missing values are imputed as the average of the adjacent 2 dates (e.g., to provide the data of September 3rd, the values of September 2nd and 4th are averaged).
The following points were excluded from the testing data set due to missing data:

There are four days with missing data in the testing data set. Additionally, these output values of the testing dataset vary substantially from the mean value. Therefore, September 13 & 28, 2020 and October 10 & 25, 2020 were excluded as outliers.

2) Missing data of some parameters of a particular day: There are also some missing input values for parameters on several days; for example, as the values of PCR-positive affected and antigen-positive affected people are missing from July 1, 2020, to August 14, 2020 and age group data are not available from July 1, 2020 to July 11, 2020. These missing values are imputed based on the ratio of the data of affected Florida residents. The following formula is used to impute the values of these input parameters on different days.

$$x_{inew} = \frac{\overline{x_i} / \overline{x_a}}{\sum(\overline{x_i} / \overline{x_a})} \cdot x_a$$

where, x_a = number of Florida resident affected
x_i = number of affected for a particular input on missing date where i = 1,2,3 ……, n

2.3.2 Validation the filling of the missing data

Hypothesis testing (F-Test) is conducted to assess the validity of the method of imputing these missing values. For PCR-positive patients, the P-values of the normality test of the number of Florida residents affected and PCR-positive affected patients were 0.288 and 0.287, respectively. Therefore, the data are normal at a significance level of α = 0.05. Therefore, the means of the formulated values and actual values are equal and normal. To assess the variances of the formulated data set and the available data set, two standard deviations are considered: σ_1 = standard deviation of imputed values of PCR-positive patients, σ_2 = standard deviation of total Florida residents affected. The null hypothesis is that the two variances are equal, whereas the alternative hypothesis is that the variances are not equal. Minitab [34] was used for the statistical analysis: the F-statistic was 0.66, the degrees of freedom for both samples were 44, and the P-value was 0.179. The P-value is greater level of significance of 0.05; therefore, based on the two-sample F-test, the variances are equal. Thus, the imputation of the missing cells (July

1, 2020 to August 14, 2020) for PCR-positive patients is justified by the evidence of the hypothesis test, and the missing data of the age groups were also imputed for July 1, 2020 to July 11, 2020 with a similar procedure.

2.4 Range setup for input and output parameter

There is an advantage of the fuzzy logic approach in that the forecasting accuracy can be improved by changing the values of the range of the fuzzy parameters [28]. To set the ranges of input and output parameters, collected data of each day were categorized into three categories (low, moderate, and high) for each parameter based on the entire data set (1st July to 30th November 2020). Table 2 shows the different ranges of the input and output parameters selected for the fuzzy logic model. For example, for the first row of the table, the minimum and maximum numbers of Florida residents affected by COVID-19 within the time frame of the data set were 364 and 15135. Therefore, to account for some tolerance, the range was set from 0 to 15500. Similarly, for the second row, the minimum and maximum values Florida resident hospitalizations were 36 and 621, respectively; thus, the range was set from 0 to 800. The ranges of the other parameters were set in a similar manner. These ranges are crucial for constructing the fuzzy logic model, as these rules act as the fuzzy input in the FIS. The center point was calculated based on the average value of each parameter within the time period of the data collected for this study.

Table 2. Ranges used for each input and output parameter

Ranges		Low			Moderate			High		
		min	center	max	min	center	max	min	center	max
Output	No of affected Florida Residents	0	1500	3500	2500	5500	8500	6000	9000	15500
	Hospitalizations	0	150	300	200	350	550	400	600	800
	Deaths	0	50	120	70	150	250	170	300	450
Test	PCR positive	0	2500	5000	3000	5000	7500	6000	8000	12000
	Antigen positive	0	200	400	350	750	1100	1000	1500	3600
Gender	Male	0	1300	2000	1500	2500	3500	2500	4000	7500
	Female	0	1300	2000	1500	3000	4000	2500	4000	8000
	Unknown Gender	0	100	200	100	250	350	300	500	1500
Race	White	0	1000	2000	1700	3000	4000	3500	4500	6000
	Black	0	300	600	350	700	1200	850	1300	2000
	Other	0	600	1000	700	1100	1700	1500	2200	3000
	Unknown	0	350	700	400	1200	2000	1500	4000	11000
Age Group	0-4 years	0	50	120	70	150	250	170	300	450
	5-14 years	0	100	250	200	300	400	300	450	650
	15-24 years	0	300	500	400	900	1500	1000	2000	3000
	25-34 years	0	350	650	500	1000	1250	1000	1500	3500
	35-44 years	0	300	500	400	800	1000	800	1200	3000
	45-54 years	0	300	500	400	800	1000	800	1200	2500
	55-64 years	0	300	500	400	700	1000	800	1200	2000

	65-74 years	0	200	350	250	500	650	500	750	1100
	75-84 years	0	100	250	200	300	400	300	450	600
	85+ years	0	50	100	50	150	200	150	220	350
	Unknown	0	40	80	50	90	140	100	150	250

2.5 Rules creation for data analysis

After setting the ranges, various rules were created based on the training data set. Based on the ranges shown in Table 2 above, if the value of a particular variable of a particular day is greater than the center point of the high range, then the condition is set as high; if the value is greater than the center point of the moderate range, the condition is set as moderate; otherwise, the condition is set as low. The fuzzy inference engine then evaluates the rules to generate the output. Different surfaces can be observed relating any one output with any two inputs of the generated model. The values of the input parameters of a particular day were fed into the engine, and the engine provided the value of the output parameters for that day.

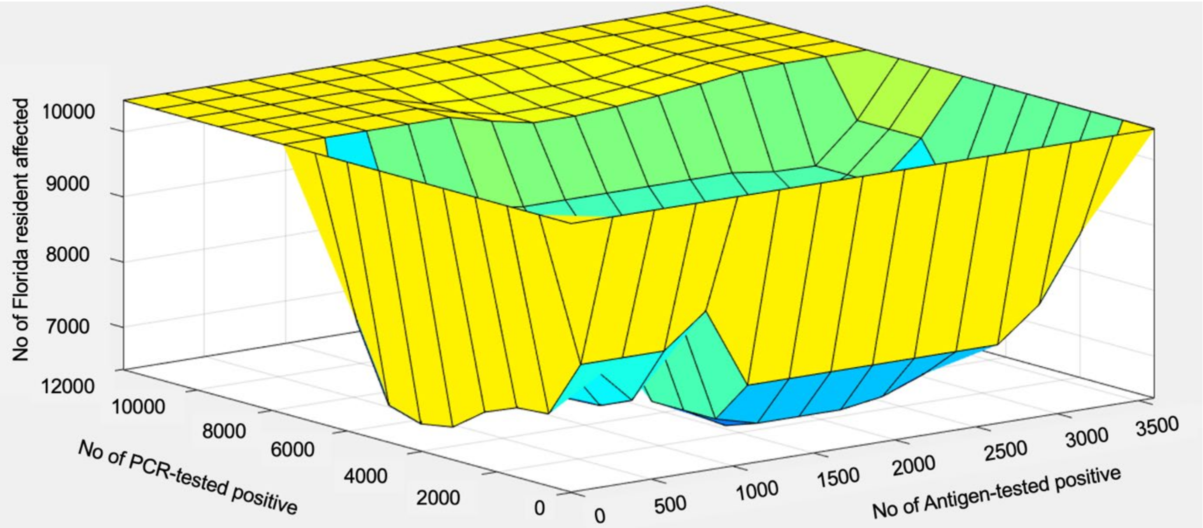

Figure 3. Surface describing the relationship between the Florida residents affected and the number of PCR and Antigen test-positive patients

The entire span of the output surface system of any one output can be observed with the relationship of the entire span of 2 inputs set by the surface viewer. Figure 3 shows the whole span of 2 inputs (number of PCR positive tests and number of antigen-positive tests) for the entire span of 1 output (number of Florida residents affected). Colors closer to yellow indicate more people are being affected per day; colors closer to blue indicate fewer people. The surface shows that the minimum (blue) number of affected patients is at the intersection of the minimum level of PCR and antigen-positive tests on a particular day.

3 Results

Linear Regression Model: Fold by fold regression equations for the training data is calculated to generate the output of the regression model. Table 3 shows the regression equation outcome of the linear regression model from the training data. For the ‘Affected’ column, the x represents the number of Florida residents affected per day, for the ‘Hospitalized’ column, the x represents the number of Florida residents hospitalized per day, and for the ‘Death’ column, the x represents the number of Florida residents’ death per day.

Table 3. Linear regression equations for each fold

Folds	Affected	Hospitalized	Death
1	200.34x + 8374.6	18.674x + 205.83	4.8696x + 28.033
2	104.24x + 9121.2	6.575x + 265.16	5.1118x + 26.033
3	-56.333x + 10785	6.4721x + 269.67	4.805x + 28.301
4	-110.14x + 11531	4.2255x + 302.18	2.7192x + 56.428
5	-142.3x + 12094	1.3185x + 355.32	1.5789x + 77.409
6	-145.38x + 12151	-1.5329x + 416.19	0.5005x + 100.32
7	-127.98x + 11692	-2.8253x + 449.29	0.0295x + 112.67
8	-113.21x + 11249	-3.0746x + 456.52	-0.0521x + 115.34
9	-96.465x + 10681	-2.8385x + 448.44	-0.1316x + 118.01
10	-79.446x + 10037	-2.7494x + 445.05	-0.3172x + 125.17
11	-60.463x + 9239.4	-2.5628x + 437.36	-0.4509x + 130.79
12	-42.904x + 8428.1	-2.0551x + 414.12	-0.4601x + 131.18
13	-32.073x + 7892.3	-1.9496x + 408.91	-0.4362x + 130.01

Fuzzy Logic Model: Fold-by-fold rules are created based on the training data to calculate the output based on the testing data. The method is repeated for each fold to obtain the output.

LSTM Deep Learning Model: Fold-by-fold values of input and output parameters are provided from the training data set. Input values of the testing data set are provided in the MATLAB workspace. After 2000 iterations in each fold, the LSTM model generated predictions of the output parameters of each fold based on the testing data.

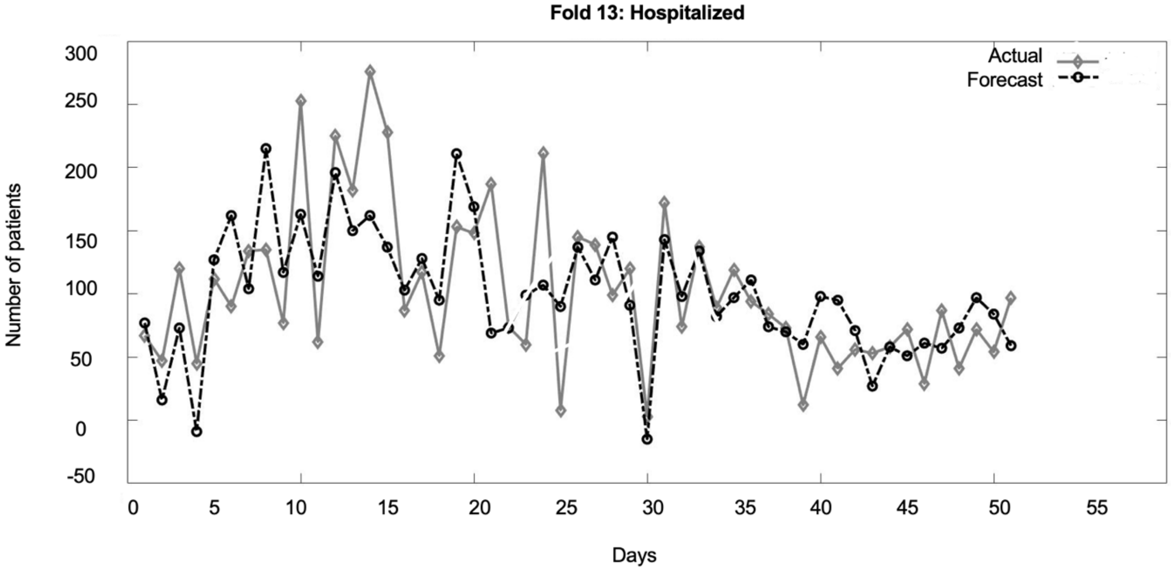

Figure 4. Output (Hospitalization) of the LSTM deep learning model up to Fold – 13

Figure 4 shows a sample of predictions generated by the LSTM deep learning model compared with the actual values. The figure shows the forecast of the number of Florida residents hospitalized per day based on the 13-fold data.

A comparison of the outputs of these three models is shown in Appendix.

The results are strong for both the fuzzy and deep learning models, but clear superiority is achieved by the LSTM deep learning approach. Fold-by-fold averages of the output variables and predicted values were calculated for comparison. Table 4 compares the fold-by-fold predicted values with the actual data.

Table 4. Comparison of fold-to-fold average output of different models

Folds	Affected				Hospitalized				Death			
	Actual	Fuzzy Forecast	Regression Forecast	LSTM Forecast	Actual	Fuzzy Forecast	Regression Forecast	LSTM Forecast	Actual	Fuzzy Forecast	Regression Forecast	LSTM Forecast
1	9931	9933	9877	8945	287	270	346	227	70	58	65	46
2	10503	10243	11154	10533	498	268	393	442	118	119	126	149
3	7791	7631	9011	7827	396	269	474	405	154	118	180	144
4	5736	6395	6740	5815	494	240	486	433	193	104	175	115
5	3397	2405	4196	3424	310	276	428	322	117	122	165	138
6	4023	4093	2338	3972	240	272	313	221	133	121	134	81
7	2682	1699	1518	2660	194	270	225	191	98	120	115	108
8	1952	1716	890	2099	149	271	175	154	92	121	111	91
9	2818	1708	697	2801	190	271	155	222	110	121	104	123
10	3382	2977	861	3350	176	266	127	198	59	118	89	81
11	3958	4607	1530	3946	168	266	111	176	52	118	73	59
12	7807	8255	2443	7836	182	266	127	218	57	118	67	60
13	7642	8364	3081	7682	218	268	116	232	74	118	65	81

Figures 5, 6, and 7 compare the results of the fuzzy logic model with those of the other models: the linear regression model and LSTM deep learning model. The average value of the output of the folds of three variables are compared with the actual values: number of Florida residents affected, number of patients hospitalized and number of patients that died.

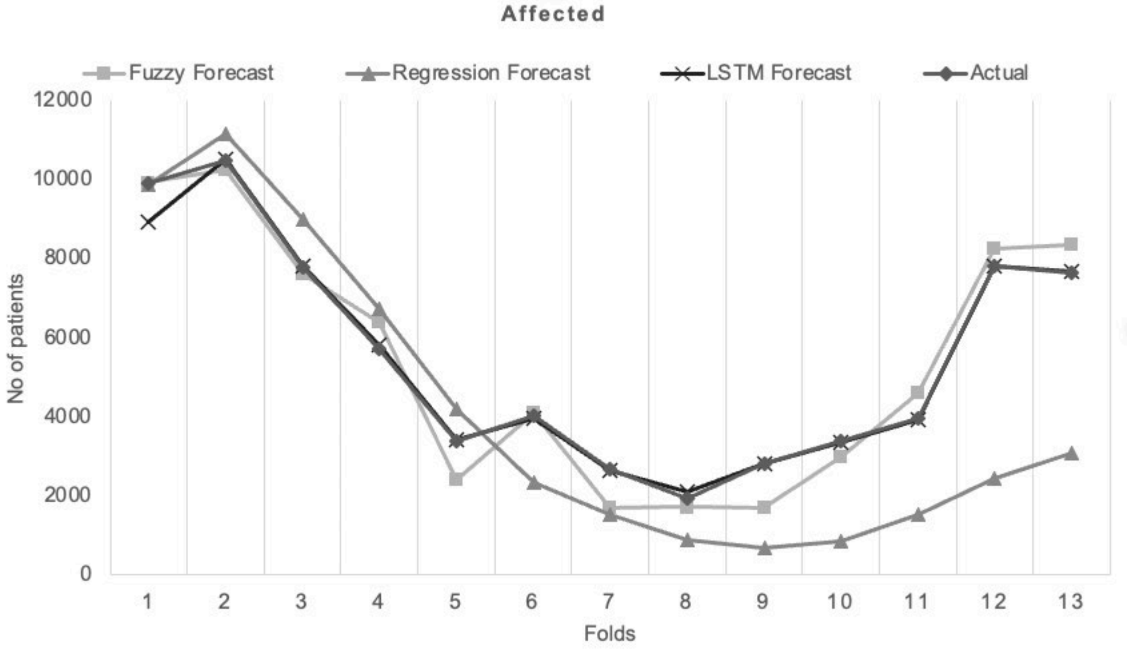

Figure 5. No of people affected for different models

Figure 5 shows that the linear regression model overestimates the results for folds 1-5 and underestimates the results for the remaining folds. The fuzzy logic model predicts values closer to the actual values for folds 1, 2, 3, 6, 8, 10, and 12. For folds 4, 11, and 13, the fuzzy model overestimates the real values, and for folds 5, 7, and 9, it underestimates the real values. The LSTM deep learning model closely estimates the actual number of Florida residents affected by COVID-19 per day.

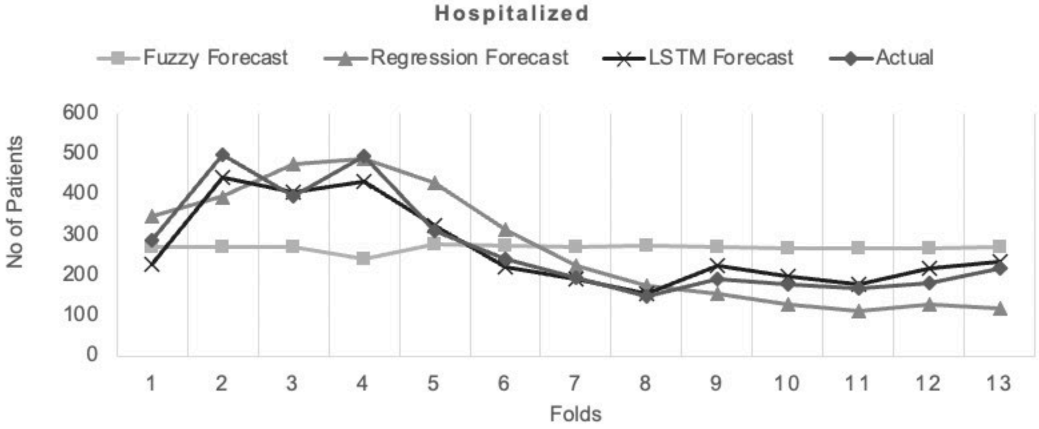

Figure 6. Number of patients hospitalized for different models

Figure 6 shows a fold-by-fold comparison of the forecasted and actual number of hospitalized Florida resident due to COVID-19. The linear regression model underestimates the actual values for folds 2 and 9-13 and overestimates the values for the remaining folds. Except for fold - 4, the fuzzy logic model produced an average forecast in the range between 266 and 276. Again, the LSTM deep learning model produces the best forecasts: in folds 1, 3, 4, and 6, it slightly underestimates the values, whereas for the remaining folds, it slightly overestimates the values.

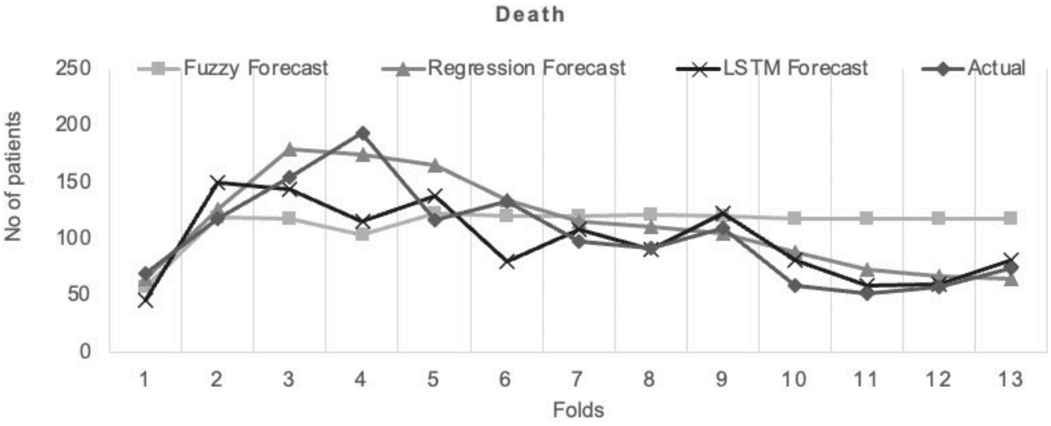

Figure 7. Number of patient deaths for different models

Figure 7 illustrates the average number of Florida resident deaths. The actual number of deaths is the highest for fold 4. The linear regression model shows an almost accurate forecast for folds 2, 6, 9, and 12. For folds 3, 5, 7, 8, 10, 11, and 12, the model overestimates the true values, and for folds 1 and 13, it underestimates the true values. For the fuzzy logic model, except for folds 1 and 4, the forecasts are within the range of 118 to 122. The LSTM model underestimates the actual number of deaths for folds 1, 3, 4, and 6 and overestimates the number of deaths all other folds, except fold 8, where the forecasted value and the actual value are similar.

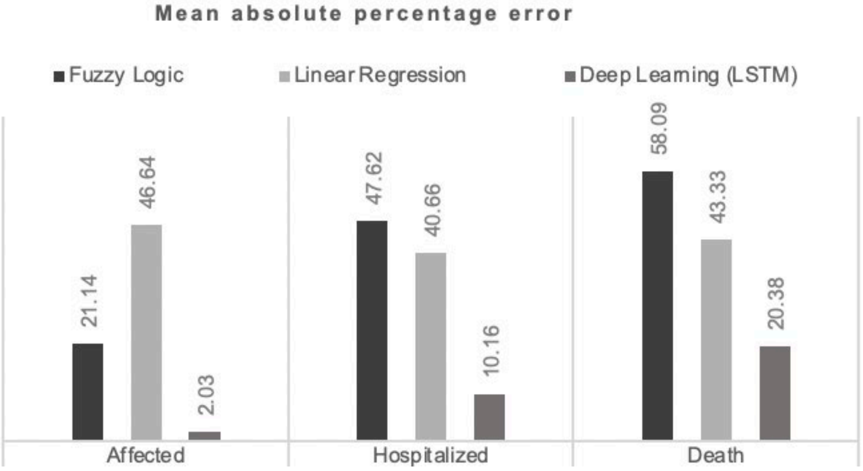

Figure 8. Mean absolute percentage error (MAPE) of the models.

Figure - 8 shows the performance measure of different models. The mean absolute percentage error (MAPE) of the proposed fuzzy model with the other two comparator models is compared in the figure.

The MAPE values indicate that the LSTM deep learning model outperforms the fuzzy model and linear regression model. The MAPE values for death for the deep learning LSTM model and fuzzy model were not significantly different, as the P-value was 0.145 > 0.05. The MAPE of the number of Florida residents affected is 21.14% in the fuzzy model, which is much lower than that of 46.64% for the regression model (P-value < 0.05). The MAPE values for the numbers of Florida resident hospitalizations and deaths are the best for the linear regression model, but the difference is not significant, as the P-values are 0.412 and 0.182. Therefore, the LSTM model outperformed the developed fuzzy model and the linear regression model for these predictions. However, the developed fuzzy logic model is better than the linear regression model at predicting the number of affected patients, and the performance of predicting the number of patients hospitalized and the number of patient deaths are statistically the same for these two models.

4 Discussion

This work has the potential to impact and advance the state of the art in COVID modeling. The FDH updates the numbers of COVID-19 confirmed cases, hospitalizations, and deaths daily. Simply observing these updates is not the prime purpose; this study aims to ensure the better use of the official data and generate prediction models to functionalize the updates. Another contribution of this study is to generate an algorithm to impute the missing information in the official data. The data collection and scraping procedure used in this study is also novel.

Since the start of COVID-19, no fuzzy logic prediction model has been applied to forecast the number of affected people, number of hospitalizations, and number of deaths. Therefore, the developed model is unparalleled. The uniqueness of this study is the development of prediction models based on regional-level data. In most cases, day-to-day operational-level decisions are not made by the federal government but by the state. Therefore, the developed model could help the decision-makers in Florida to take appropriate measures to combat COVID-19. This study will inspire other researchers to conduct additional studies based on state-level data. The developed model is fit to apply for other states also. Future research should address several challenges and useful extensions in this context.

The developed model uses data from July 1, 2020, to November 30, 2020 but data are available for April 25, 2020 to June 3, 2021. The model may have been more accurate if all the data were used. The model was developed based on official data, so only four factors (type of test, gender, race, and age) were considered as input variables. However, other factors, e.g., previous health conditions of the tested individuals, immunizations, and vaccination, that were not included in the official data might be relevant. The official website that we have used in this study has stopped updating daily data of COVID-19 affected, hospitalization and death on response from the decision of the federal government. But that does not affect the credibility of this work due to the availability of the data set of this works exists in the source. Moreover, the developed models are not customized for any particular data rather a generic one. Nonetheless, we believe our results clearly demonstrate opportunities to use the developed model to construct short-term forecasts of future trends where there is large variability.

More broadly, forecasts are rarely perfect. State-by-state prediction models can surely play a pivotal role in obtaining a clear picture of the future scenario of COVID-19 at the state level. The spread of the COVID-19 pandemic depends not only on the considered internal input factors but also on other external factors. Weekdays, weekends, public holidays, strict lockdown-self isolation, and other factors could potentially be considered to increase the accuracy of prediction models. A fuzzy model with different combinations of input-output membership functions (Gaussian, bell, sigmoidal, etc.) may also improve the results.

In this study, the fuzzy model was developed based on real data without normalization. Fuzzy modeling with normalized data is another option to increase the accuracy. The proposed model can also be tested with data from different states or with county-level data to validate the model. The LSTM deep learning model outperformed the proposed model and the linear regression model. We hope that the effectiveness of other LSTM models (convolutional, bidirectional, etc.) can be tested on the same data. The implications of this work include a better understanding of opportunities and appropriate tools for short-term predictions of future trends when variability is high. The regional data used for this effort are highly variable, contain many missing values and are stored in a format that is not conducive for data science (i.e., PDF). As such, the present work also presents data replacement strategies for data sets with missing values. We hope that the approaches demonstrated here will assist in such situations.

References

Arora, H. Kumar, and B. K. Panigrahi, “Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India,” *Chaos Solitons Fractals*, vol. 139, p. 110017, Oct. 2020, doi: 10.1016/j.chaos.2020.110017.

Aristarán, Manuel, Tigas, Mike, and Merrill, Jeremy B., *Tabula: Extract Tables from PDFs*. 2018. Accessed: Jan. 05, 2021. [Online]. Available: https://tabula.technology/

Arencibia, Dagmaris Martínez Cardero, and J. Gulín-González, “Life Between Economics And Politic In The Context Of The Covid 19 Crisis: Readings For Latin America And The Caribbean,” Sep. 2020.

Austin, A. Widyastuti, N. El Jundi, and R. Nagrani, “COVID-19 Surveillance Data and Models: Review and Analysis, Part 1,” *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3695335.

Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner, *Minitab 19*. Minitab, LLC, Penn State University, 2020. Accessed: Mar. 30, 2021Cleve Moler, *MATLAB. (2018).* Natick, Massachusetts:The MathWorks Inc., 2018.

Cadogan and C. M. Hughes, “On the frontline against COVID-19: Community pharmacists’ contribution during a public health crisis,” *Res. Soc. Adm. Pharm.*, vol. 17, no. 1, pp. 2032–2035, Jan. 2021.

Castillo and P. Melin, “Forecasting of COVID-19 time series for countries in the world based on a hybrid approach combining the fractal dimension and fuzzy logic,” *Chaos Solitons Fractals*, vol. 140, p. 110242, Nov. 2020,

Chimmula and L. Zhang, “Time series forecasting of COVID-19 transmission in Canada using LSTM networks,” *Chaos Solitons Fractals*, vol. 135, p. 109864, Jun. 2020,Oh, “From a ‘super spreader of MERS‘ to a ‘super stopper‘ of COVID-19: Explaining the Evolution of South Korea’s Effective Crisis Management System,” *J. Asian Public Policy*, pp. 1–16, Dec. 2020.

Cucinotta and M. Vanelli, “WHO Declares COVID-19 a Pandemic,” *Acta Bio Medica Atenei Parm.*, vol. 91, no. 1, pp. 157–160, Mar. 2020.Walker *et al.*, “Report 12: The global impact of COVID-19 and strategies for mitigation and suppression,” Imperial College London, Mar. 2020.

Fanelli and F. Piazza, “Analysis and forecast of COVID-19 spreading in China, Italy and France,” *Chaos Solitons Fractals*, vol. 134, p. 109761, May 2020, doi: 10.1016/j.chaos.2020.109761.

Hamzah *et al.*, “CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction,” nCoV, preprint, Mar. 2020. doi: 10.2471/BLT.20.255695.

Messner and S. E. Payson, “Variation in COVID-19 outbreaks at the US state and county levels,” *Public Health*, vol. 187, pp. 15–18, Oct. 2020, doi: 10.1016/j.puhe.2020.07.035.

Pandey, P. Chaudhary, R. Gupta, and S. Pal, “SEIR and Regression Model based COVID-19 outbreak predictions in India,” *ArXiv200400958 Q-Bio*, Apr. 2020, Accessed: Sep. 29, 2020. [Online]. Available:

Scantamburlo, A. Cortés, P. Dewitte, D. Van der Eycken, R. De Wolf, and M. Martens, “Covid-19 and tracing methodologies: A lesson for the future society,” *Health Technol.*, Aug. 2021.

Taylor and J. W. Taylor, “A Comparison of Aggregation Methods for Probabilistic Forecasts of COVID-19 Mortality in the United States,” *ArXiv200711103 Stat*, Aug. 2020, Accessed: Oct. 06, 2020. [Online]. Available: http://arxiv.org/abs/2007.11103

Wang, X. Zheng, J. Li, and B. Zhu, “Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics,” *Chaos Solitons Fractals*, vol. 139, p. 110058, Oct. 2020, doi: 10.1016/j.chaos.2020.110058.

White and L. Hébert-Dufresne, “State-level variation of initial COVID-19 dynamics in the United States,” *PLOS ONE*, vol. 15, no. 10, p. e0240648, Oct. 2020, doi: 10.1371/journal.pone.0240648.

Yadaw, Y. Li, S. Bose, R. Iyengar, S. Bunyavanich, and G. Pandey, “Clinical features of COVID-19 mortality: development and validation of a clinical prediction model,” *Lancet Digit. Health*, vol. 2, no. 10, pp. e516–e525, Oct. 2020, doi: 10.1016/S2589-7500(20)30217-X.

Ye and X. Yang, “Analysis and prediction of confirmed COVID-19 cases in China with uncertain time series,” *Fuzzy Optim. Decis. Mak.*, Sep. 2020, doi: 10.1007/s10700-020-09339-4.

Appendix:

Appendix I: Fold-by-fold forecasts generated by different models

Fold	Date	Affected			Hospitalized			Death		
		Regression	Fuzzy	LSTM	Regression	Fuzzy	LSTM	Regression	Fuzzy	LSTM
1	3-Jul	8976	10193	9448	262	268	315	43	57	65
	6-Jul	9577	8682	6337	318	270	91	57	58	26
	9-Jul	10178	10223	8871	374	268	252	72	57	72
	12-Jul	10779	10633	11123	430	274	248	86	60	19
2	15-Jul	10685	10196	10094	364	268	313	103	118	86
	18-Jul	10998	10200	10260	384	268	543	118	118	160
	21-Jul	11310	10202	9275	403	268	326	133	118	84
	24-Jul	11623	10375	12503	423	270	586	149	120	267
3	27-Jul	9264	9120	8877	444	268	347	158	118	154
	30-Jul	9095	8553	9871	464	268	494	172	118	164
	2-Aug	8926	8542	7071	483	269	340	187	119	123
	5-Aug	8757	4307	5490	503	272	438	201	117	135
4	8-Aug	7236	8450	8454	467	268	437	162	118	132
	11-Aug	6905	7231	5770	480	269	539	171	119	126
	14-Aug	6575	8228	6065	492	275	519	179	122	158
	17-Aug	6244	1669	2969	505	150	236	187	57	44
5	20-Aug	4837	4544	4536	423	296	373	158	131	118
	23-Aug	4410	1697	3007	427	270	188	163	120	89
	26-Aug	3983	1708	2987	430	271	423	167	120	213
	29-Aug	3556	1672	3164	434	268	302	172	118	131
6	1-Sep	2992	8561	7425	320	269	299	132	119	127
	4-Sep	2556	1685	1796	315	269	122	133	119	66
	7-Sep	2120	4403	4149	310	281	299	135	125	48
	10-Sep	1684	1724	2516	306	269	162	136	120	82
7	13-Sep	2094	1707	2368	237	270	75	115	120	59
	16-Sep	1710	1681	2694	229	268	227	115	119	122
	19-Sep	1326	1692	3151	220	269	215	115	119	105
	22-Sep	942	1715	2425	212	271	246	115	121	147
8	25-Sep	1400	1716	2791	189	271	171	111	121	98
	28-Sep	1060	1745	470	180	274	25	111	124	-13
	1-Oct	720	1684	2384	171	269	239	110	119	161
	4-Oct	381	1717	2749	161	271	179	110	121	119
9	7-Oct	1131	1704	2608	167	270	288	105	120	142
	10-Oct	842	1724	2663	159	272	171	105	122	79
	13-Oct	552	1728	2594	150	272	184	104	123	108
	16-Oct	263	1675	3339	142	268	245	104	118	163
10	19-Oct	1218	3177	3506	140	269	177	90	119	74
	22-Oct	980	3114	3574	132	269	190	89	119	78
	25-Oct	742	1711	2333	123	271	121	88	121	51
	28-Oct	503	3907	3986	115	256	302	87	114	122
11	31-Oct	1802	1687	2290	122	269	194	75	119	91
	3-Nov	1621	5812	4550	114	269	233	74	119	72
	6-Nov	1440	6887	5144	107	257	150	73	114	18

	9-Nov	1258	4044	3799	99	270	126	71	120	56
12	12-Nov	2636	7683	5485	137	270	176	69	119	44
	15-Nov	2507	8488	9831	131	268	193	68	118	49
	18-Nov	2379	8443	7778	124	267	223	66	118	52
	21-Nov	2250	8406	8249	118	261	279	65	115	94
13	24-Nov	3178	8451	8143	122	268	319	66	118	96
	27-Nov	3081	8531	8455	116	269	236	65	119	88
	30-Nov	2985	8108	6448	111	268	142	63	118	60