Predictive Information Workflow of Forecasting Number of COVID-19 Confirmed Cases

Yi Cong Areeve Gam
School of Computing Science, University of Glasgow
Singapore 567739
2508451G@student.gla.ac.uk

Qi Cao
School of Computing Science, University of Glasgow
Glasgow, Scotland, UK
qi.cao@glasgow.ac.uk
ORCID: 0000-0003-3243-5693

Chee Kiat Seow
School of Computing Science, University of Glasgow
Glasgow, Scotland, UK
CheeKiat.Seow@glasgow.ac.uk
ORCID: 0000-0002-6499-9410

***Abstract*—There are questions about how to accurately prepare with the correct number of resources for distribution in order to properly manage the healthcare resources (e.g., healthcare workers, Masks, ART-19 TestKit) required to tighten the grip on the COVID-19 pandemic. Mathematical and computational forecasting models have well served the means to address these questions, as well as the resulting advisories to governments. A workflow is proposed in this research, aiming to develop a forecasting simulation that makes accurate predictions on COVID-19 confirmed cases in Singapore. According to the analysis of the prior works, six candidate forecasting models are evaluated and compared in the workflow: polynomial regression, linear regression, SVM, Prophet, Holt's linear, and LSTM models. The study's goal is to determine the most suitable forecasting model for COVID-19 cases in Singapore. Two algorithms are also proposed to better compute the performance of two models: the order algorithm to determine optimal degree order for the polynomial regression model , and the optimizing algorithm for the Holt's linear model to calculate the optimal smoothing parameters. Observed from the experiment results with the COVID-19 dataset, the Prophet method model achieves the best performance with the lowest Root Mean Square Error (RMSE) score of 1557.744836 and Mean Absolute Percentage Error (MAPE) score of 0.468827, compared to the other five models. The Prophet method model achieving average accuracy range of 90% when forecasting the number of confirmed COVID-19 cases in Singapore for the next 87 days ahead. is chosen and recommended to be used as a system model for forecast the COVID-19 confirm cases in Singapore. The developed workflow will greatly assist the authorities in taking timely actions and making decisions to contain the COVID-19 pandemic.**

Keywords—Workflow simulation, forecasting models, forecast COVID-19 cases, healthcare, decision making

I. INTRODUCTION

The spread of COVID-19 in countries around the world has posed significant challenges to global public health security. As of July 2022, the cumulative number of confirmed diagnoses worldwide had surpassed 549 million, resulting in more than 6 million deaths [1], having a significant impact on people's daily lives. Singapore reported their first case of COVID-19 on January 23, 2020. Singapore's proactive measures during the early months of the COVID-19 outbreak had a significant impact on virus containment. However, the change in trajectory of new COVID-19 infections through April 2020 raised concerns about Singapore's mitigation strategies and the social, political, and economic impacts on the country.

Despite government-enforced sealing measures and stringent movements across the country, the evolution of new COVID cases is nearly exhausting available healthcare resources. Demand projections for healthcare resources are fraught with uncertainty in a rapidly evolving COVID-19 pandemic situation with many unknowns. The most significant challenge disrupting traditional healthcare operations is the unexpected and unanticipated surge in demand for healthcare facilities by healthcare professionals. As the pandemic progresses, hospitals have postponed routine elective services and reduced the number of healthcare workers attending to non-COVID-19-related needs. This may make maintaining continuity in the delivery of healthcare to medically vulnerable populations such as chronically ill patients and functionally impaired older adults difficult.

Research and studies on COVID-19 have reported on medical factors as well as retrospective studies to help the general public combat COVID-19. Given the ongoing pandemic, a study on forecasting the future outcome of COVID-19, specifically in Singapore, would be beneficial. Forecasting is a mathematical modelling technique for predicting future outcomes based on historical data and trends [2]. It is widely used not only in the business sector, but also in the healthcare sector. Forecasting models are one of many tools used by businesses to predict sales, consumer behaviours, supply and demand, etc.. These models are particularly useful in the fields of sales and marketing [3]. In the healthcare sector, forecasting is critical to an organization's ability to plan and execute strategies to meet the demands of a rapidly changing health environment. The use of appropriate forecasting tools can aid in combating future health events or situations such as demand for health services and healthcare needs, as well as facilitating preventative health strategies [4]. As a result, the purpose of this study is to explore the most suitable model for forecasting the number of COVID infections that Singapore will face over the next two months. The project outcomes will be a nod not only to the healthcare sector, but also to the country that will play a long-term role in protecting citizens from the pandemic.

The main contributions of this project are to compare the results of Machine Learning and Time Series forecasting models, and then propose the best forecasting method to

predict the number of infections that Singapore will face two months in advance using data from the World Health Organization website. With performance and accuracy being the relevant metrics used to select the best model, the project's final deliverable is to provide the best performance model with the most accurate prediction for the next two months of confirmed COVID cases. The selected model will be used as a simulation and framework to make accurate predictions on COVID active cases to assist the government in promoting policies. It can prevent the spread of the COVID-19 disease outbreak in the country and describe the response that the health system should adopt in light of the dynamic evolution of the COVID-19 pandemic.

The organising of the remaining parts of this paper is as follows. Section 2 presents the related works in the literature. Section 3 presents the methodology and proposed design of the system. Section 4 presents the results and analysis of the implemented system. Section 5 presents the conclusion for this capstone study.

II. Related Works

This section is to analyse existing solutions from the literature in order for the healthcare industry to embrace in consideration of the dynamic evolution of the COVID-19 pandemic. General findings derived from the literature publications conclude a consensus in solution where mathematical modelling and simulation are utilised to describe the response that health system should embrace in consideration of the dynamic evolution of the COVID-19 pandemic. Reported models include regression models and time series forecasting models. These models have a common goal, to enable the estimation of healthcare resources needed to deal with rapidly evolving outbreak scenarios. it allows for the fast adaptation of new structural and behavioural assumptions on both the demand and supply scenarios.

With the variety of models used in achieving the common goals, a comparison will be made among the models reported by each of the prior works in the literature, to identify the limitations and derive the most appropriate model to be used for implementation.

Rustam *et al.* [7] compare four regression models, i.e., Linear Regression (LR), LASSO Regression, Support Vector Machine (SVM), and Exponential Smoothing (ES), to cater for COVID19 future forecasting. The prediction accuracy is determined using Root Mean Square Error (RSME). The results illustrate that the ES outperforms all other models. While SVM performs poorly in all prediction scenarios given the available dataset. S. Maurya and S. Singh [8] analyse the data with the time series manner to forecast the future effect of COVID-19 on a global or individual scale. This analysis is carried out using four methods for prediction: NAIVE, Holt's linear trend method, Holt's Winter seasonal method, and Autoregressive Integrated Moving Average method (ARIMA). The NAIVE model is the best model to be chosen, because its error is lower than the other three models. A prophet method for time series data is introduced to forecast affected, recovered, and death COVID cases [9]. It captures the trend, the short-term repeating cycle in the time series, and the holiday effects. However, the prophet method relies on some underlying assumptions, which makes the forecasting results unreliable [10]. A methodology analyzes the Inda's dataset for COVID-19 using two regression model namely linear and polynomial [11]. It shows that performance of the polynomial regression is better than the linear regression. However, the error rate in the results tends to be high, due to the small dataset.

A model is reported based on hybrid Self-Organizing Map (SOM) and fuzzy time series (SOMFTS) to forecast COVID-19 spread [12]. This forecasting model is applied to three datasets Confirmed cases, Cured cases, and Death cases due to COVID-19 in Delhi. The SOMFTS forecasting method is compared with other models in the presence of more than one conflicting performance measure. Experimental results show that the SOMFTS technique is advantageous for future forecasting of COVID19 cases. A linear regression model is implemented on the clinical dataset of COVID19 patients in Ukraine that is capable of making accurate forecasts of COVID-19 future cases [13]. It provides the time-series prediction of confirmed, deaths, and recovered cases in Ukraine. An ELM-based forecasting model is reported which can forecast new and active cases each day [14]. The experiments are conducted using public COVID datasets. However, the prediction results of this method are affected by network connection parameters and the number of hidden layer nodes.

Dash *et al.* [15] use ARIMA machine learning model to predict the daily-confirmed cases for 90 days' future values of six worst-hit countries of the world and six high incidence states of India. A Spatiotemporal Long Short-Term Memory (ST-LSTM) based Multi-Agent Deep Reinforcement Learning (MADRL) simulation model is implemented [16], that is useful in predicting the growth trend of COVID-19 cases for a geographic area. The input data for the simulation model consist of confirmed, recovered, tested and deceased cases. However, this method requires intense data pre-processing and further normalizing of data to ensure prediction accuracy.

The performances of the univariate modelling techniques and Box-Jenkins methodology are compared to estimate and validate forecasting models [17], which are then used to forecast the future number of COVID-19 cases, from the best model selected from four univariate models: Average Percent Change (APC), Single Exponential Smoothing (SES), Double Exponential Smoothing (DES), and Holt's method models. The 30-day forecast from the best model, Holt's method, reveal that the pandemic trend would substantially increase. Cheng *et al.* [18] compare and evaluate between two prediction models, ARIMA and SVM, to predict the COVID-19 trend. The prediction performance of ARIMA model is worse as compared to the SVM model in the experiments, which show that ARIMA model is more suitable for regular and stable sample data. However, despite the SVM model having a better performance score, the constraint for a single prediction model is restricted by its internal conditions and

lacks comprehensive monitoring of data factors, which reduce the prediction performance of the model.

Investigations of recurrent LSTM are carried out for medium-term forecasting of covid-19 pandemic indicators at the example of Ukraine [19]. The models of recurrent LSTM of two architectures (basic and extended) on different data sets are explored, after which the best models are found and corresponding forecast results are obtained. A comparative analysis of forecasting efficiency of different architectures is performed. It is shown that the basic LSTM model is better for predicting daily indicators, and the extended model is better for predicting absolute indicators. The value of LSTM networks is compared to several other methods (i.e., Box-Jenkins method, Prophet method, and Holt-Winters Additive method with Dampe Trend) in forecasting the total number of COVID-19 cases in Turkey [20]. The COVID-19 data for 30 days are used to estimate the next fifteen days' predictions. Investigation in ensembling techniques is carried to forecast and examine the potential for use in nonseasonal time-series similar to those in the early days of the COVID-19 pandemic [21]. Four forecasting modelling techniques to make prediction on COVID-19 cases. The four models are ARIMA, Holt Winters, Prophet and LSTM. The forecasting results have been quantified and compared using performance metrics like Mean Absolute Percentage Error (MAPE). Achieved scores indicate that Prophet has outperformed the others three models in a 14 days forecasting.

III. Proposed Methodology

Observed from the previous section, regression analysis techniques, time series model and RNN for time series forecasting approaches are among the commonly discussed and implemented methods in the literature. The work in this research will consist of building and comparing some of these methods to determine the best model to be used for forecasting COVID cases.

The linear regression obtains better performance when comparing with methods of LASSO Regression, SVM, and Exponential Smoothing in [7]. The linear Regression model is compared to the Polynomial Regression model in [11] to identify the optimum strategy for forecasting, while the Polynomial regression outperforms linear regression. In publication [20], A comparison is made between ARIMA and SVM models forecasting accuracy in [20]. It is observed that SVM has better forecasting accuracy as compared to ARIMA model. Therefore, the linear regression, polynomial regression and SVM models will be incorporated as the candidates in this research.

The model performance comparison is made among models of APC, SES, DES, and Holt's linear method in [18]. It shows that the Holt's linear method performs better than others. The comparison is made among models of ARIMA, Holt's Winters, Prophet and LSTM in [21]. Achieved scores indicate that Prophet outperforms others three models in a 14-day forecasting. Therefore, the Holt's linear method and Prophet model will be selected as the candidates in this research.

The recurrent LSTM are used for COVID-19 forecasting in [16][19][20] with good results. Compared to the Box-Jenkins method and Prophet method, the LSTM gives superior results in forecasting the total number of COVID-19 cases. As such, the LSTM model will be utilized to predict COVID situations in this research.

In this research, a workflow is proposed to evaluate and identify the most suitable forecasting model for Singapore COVID-19 confirmed cases among six well performed models in the literature: polynomial regression, linear regression, SVM, Holt's linear method, Prophet model and LSTM model. Two algorithms are also proposed to determine the optimal degree order for the polynomial regression model, and the optimal smoothing parameters for the Holt's linear model. A forecasting system for Singapore COVID-19 confirmed cases will be built with the selected most suitable forecasting model with the best performance.

An overview of the proposed workflow design to identify the most suitable forecasting model is shown in Fig. 1. The six forecasting models will be compared in terms of accuracy, performance and error rate. The model with the best performance results calculated from the performance comparison metrics will be selected as the most suitable model to create a forecasting system for Singapore COVID-19 confirmed cases.

As shown in Fig. 1, the proposed workflow will first involve a collection of raw COVID-19 data from various sources, which will then be stored in a Data Warehouse. The Microsoft Excel platform will be used as the Data Warehouse, where the raw data will be pre-processed, cleaned, and filtered in order to reduce data redundancy and achieve an informative set of processed data. The processed data will then be analysed with these six forecasting models to convert a descriptive set of data to predictive data. The performance of each forecasting model will be evaluated and compared. The forecasting model with the best performance will be selected by the proposed workflow to derive the prescriptive data. The prescriptive data is information that helps healthcare professionals make better decisions and keep the system running smoothly, to forecast future Singapore COVID-19 confirmed cases.

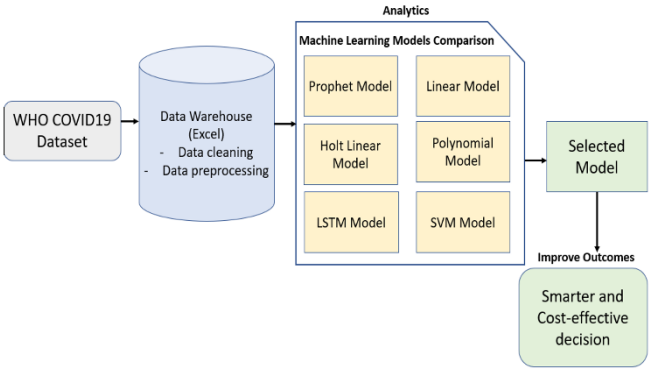

Fig. 1. Overall proposed workflow

A. Data Pre-processing

The datasets for this research are a set of COVID-19 data exported from World Health Organization (WHO) Dashboard

[32]. The dataset consists of information about the Observation Date, Country/Region, Last Update, the number of Confirmed, Deaths and Recovered cases. The dataset contains a total number of 98,252 entries of records, for 223 unique countries, 725 unique provinces with the timestamp between each entry varies by seconds. The first five entries of records in the dataset are shown in TABLE I.

Data cleaning and pre-processing is conducted to extract the relevant information of the dataset. As the focus of this research is to predict the COVID-19 cases pandemic in Singapore, the current dataset which includes details for all Countries and regions that are not necessary for the system implementation. Thus, the "Province/State" column will be removed from the dataset as only Singapore details that are necessary for the later part of processing and training of the models. The "Observation Date" will be used to keep track on the progression of the COVID-19 cases. Therefore the "Last Update" column will also be removed. Taking into factor that when a csv (comma-separated values) file is imported and a Data Frame is made, the Date time objects in the file are read as a string object rather a Date Time object. Hence it is tough to perform operations like time difference on a string rather a Date Time object. A Python package, Pandas to_datetime() method helps to convert string Date time (Observation Date) into Python Date time object [33].

TABLE I. COVID 19 Dataset Retrieved From WHO

SNo	Observation Date	Province/ State	Country/Region	Last Update	Confirmed	Deaths	Recovered
1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0

The "Deaths" and "Recovered" details are deemed redundant as the main purpose of the implementation is to predict and forecast number of "Confirmed" cases in the coming days, thus these columns will also be removed.

B. Building and Training the Models

To train these six candidate forecasting models, 95% data in the dataset after cleaning will be used as the train data, and 5% of the data will be used for test/validation. The period where data is collected range between 23 Jan 2020 to 29 Aug 2020, a total 219 days of records being used for training and testing. The reason using such percentage of data for training is to give as much data as possible for the models to learn on. Such that it can capture recent trends and changes in number of COVID-19 confirmed cases, as COVID-19 has constantly changed its trend over time. The six candidate forecasting models adopted in the proposed workflow will be discussed next one by one.

1) Polynomial Regression

The polynomial regression is a regression algorithm that models the relationship between a dependent variable *y*, and independent variable *x* as the n^{th} degree polynomial. The Polynomial Regression equation is given below in Eq. (1).

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n \quad (1)$$

where *y* is the response variable we want to predict; *x* is the feature; β_0 is the *y* intercept; *n* is the degree of the polynomial. The higher *n* is, the more complex curved lines can be created.

In order to derive the optimal degree order for the polynomial regression model with the least Root Mean Square Error (RMSE) values to ensure accurate predicted results, an order algorithm is proposed and created in this research. The flowchart of the proposed order algorithm is shown in Fig. 2. It goes through a *for loop* with a degree span between the ranges of 1 to 10. In every iteration, the implementation of the polynomial regression model will be fitted with the degree from 1 to 10 respectively. The RMSE will be calculated in each iteration with the respective degree. The degree with the lowest RMSE will be chosen to train the polynomial regression model fitted with the train data.

After deriving the optimal degree order, the polynomial model regression will be trained with the optimal degree and fit with the train data (X-axis: 95th percentile Subsequent Days; Y-axis: 95th percentile No. of Confirmed Covid Cases) to achieve the model best fit line.

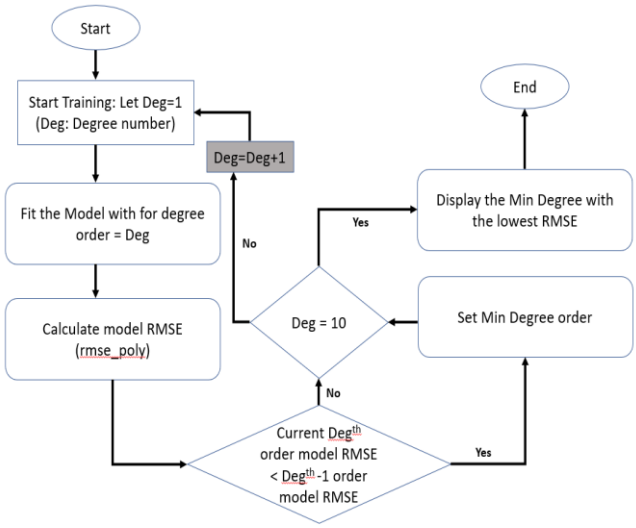

Fig. 2. Flowchart of the proposed order algorithm to determine optimal degree order

2) Linear Regression

Linear Regression is a supervised machine learning model in which the model finds the best fit linear line between the independent and dependent variables. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous, real or numeric variables. The Linear Regression equation is given in Eq. (2).

$$y = a \times x + b \quad (2)$$

where the *a* variable is the regression coefficient; The *b* variable is the *intercept*, the value where the plotted line

intersects the *y*-axis; The *x* variable is the input variable; The y is the output variable.

Utilizing the open source Linear Regression Model imported from Scikit-Learn, the model is trained and fitted upon with training dataset (X-axis: 95th percentile Subsequent Days; Y-axis: 95th percentile No of Confirmed Covid Cases) to achieve the model best fit line.

3) Support Vector Machine

SVM is a machine learning technique that can be used to solve regression as well as classification problems. It builds a hyperplane in multidimensional space to best separate a dataset into different classes. An SVM's task is to find the optimal plane that best divides the dataset into two classes, that is, the hyperplane with the highest margin. As a result, the kernel to use is solely determined by hyperparameter tuning to derive the most optimal hyperplane for finding a decision boundary that clearly divides the data points. Therefore, the appropriate kernel to use for SVM initialization must be evaluated and determined.

A kernel is a function that places a low dimensional plane to a higher dimensional space. This allows the projection of data onto a higher dimensional space where it can be separated using a plane. There are 3 main types of kernels used by SVMs: polynomial kernel, linear kernel, and radial basis function (RBF) Kernel.

TABLE II. RESULTS COMPARISONS OF THREE SVM KERNELS

Kernel Name	RMSE	MAE
Polynomial	34123.870066	26404.613209
RBF	96181.856928	85160.785115
Linear	134328.086295	126057.709867

The results of three kernels pf SVM have been compared in this research. As shown in TABLE II, based on the results, the polynomial model performs the best with the given dataset. By looking at the performance of these models, the dataset used follows a non-linear pattern as it performs best with the non-linear kernels.

As such, the polynomial kernel will be used for the of initialising the SVM model. Based on the polynomial kernel algorithm, the best hyper-parameters values calculated is with "C" = 100, "degree' = 2, 'epsilon' = 0.0001 and "coef0" = 0.1.

4) Prophet Model

Prophet is a nonlinear regression model for forecasting daily data with weekly and yearly seasonality, plus holiday effects. It works best with time series that have strong seasonality and several seasons of historical data. The Prophet equation is given below in Eq. (6).

$$Yt = g(t) + s(t) + h(t) + \varepsilon t \qquad (6)$$

where *g(t)* describes a piecewise-linear trend; *s(t)* describes the various seasonal patterns; *h(t)* captures the holiday effects; εt is a white noise error term.

The dataset is made up of daily observation data, "weekly_seasonality" and "yearly_seasonality" are set True in the instantiation of Prophet. As it assumes the pattern of daily seasonality is the same throughout the year. Dataframe period are set to 89 as the algorithm is set to forecast the *y* value (Confirmed Covid Cases) for the next 89 days.

5) Holt's Linear Model

Holt technique also known as Double Exponential Smoothing is used for forecasting with trending data. In order to derive the most optimal smoothing parameters for the Holt's linear model with the least RMSE value to ensure that predicted results will be the upmost accurate, an optimizing algorithm for the Holt's linear model is proposed to calculate the optimal smoothing parameters.

The parameters "alphas" and "betas" are the smoothing level and smoothing slope respectively. The parameters of the proposed optimizing algorithm will search for the suitable parameters of "alphas" and "betas". The proposed optimizing algorithm will loop through a range of 0.01 to 1, with an increment of 0.05 for each iteration for both "alphas" and "betas". It is to derive the optimal smoothing parameters with the lowest MAE for the forecasting of the next 13 days COVID-19 confirmed cases. The results of the proposed optimizing algorithm are illustrated in Table III. Fig. 3 shows the flowchart on an overview LSTM model.

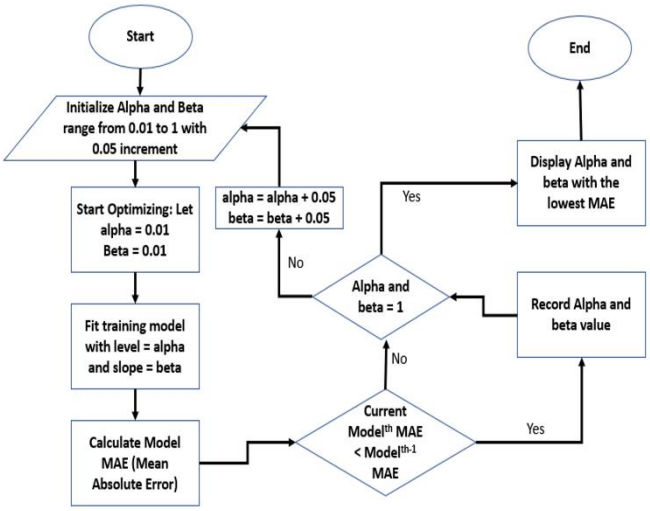

Fig. 3. Flowchart of proposed optimizing algorithm for Holt's linear model.

TABLE III. RESULTS OF THE PROPOSED OPTIMIZING ALGORITHM FOR HOLT'S LINEAR MODEL

Iteration	Alpha	Beta	MAE
379	0.91	0.96	2992.656876
398	0.96	0.91	3058.694186
397	0.96	0.86	3586.286096
378	0.91	0.91	3840.112011
359	0.86	0.96	4322.795931
.. ..	.. ..	.. ..	.. ..
0	0.01	0.01	244760.431574

As seen in Table III, the derived alpha (i.e., representing the smoothing level) value of 0.91 and beta (i.e., representing the smoothing slope) value of 0.96 will be used as the smoothing parameters for the initializing of the Holt's linear model, which give the lowest MAE value in optimization.

6) LSTM Model

LSTM cells are used in RNN that learn to predict the future from sequences of variable lengths. In this research, the number of hidden layers is set as two; with 50 neurons set in each hidden layer. Fig. 4 shows the flowchart on an overview LSTM model.

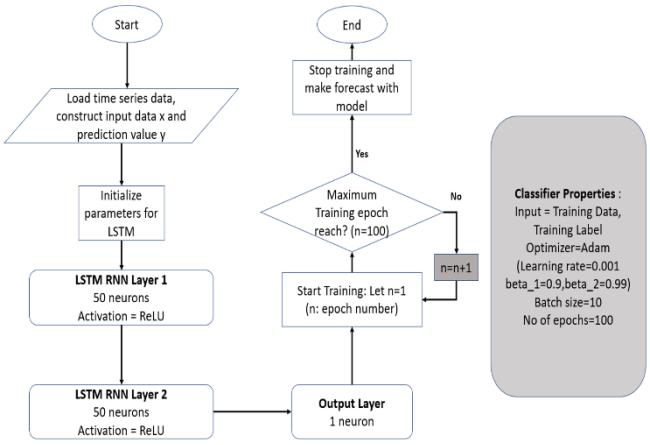

Fig. 4. Flowchart of the LSTM model

The window size is more of a tool for saving memory or computing requirements. Picking a good window size is important, but fine-tuning is not necessary. Therefore, window size is default at 10 which is large enough to learn longer dependencies.

The LSTM network has a visible layer with 1 input, two hidden layer with 50 LSTM blocks or neurons, and an output layer that makes a single value prediction. The Rectified Linear Unit (ReLU) activation function is used for the LSTM blocks. The model is fit using the efficient Adam version of stochastic gradient descent. Adam optimization is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. The arguments of the Adam algorithm; Learning Rate, beta_1, beta_2 are set at a default value of 0.001, 0.9 and 0.999 respectively. The network is trained for 100 epochs and a batch size of 10 is used.

IV. Experiment Results and Analysis

This section discusses and analyses the results achieved from training the six forecasting models. The performance metrices, RMSE and MAPE are incorporated to evaluate model performance and accuracy. The model with the best result of the performance metrices will be one of the factors to be chosen as the best forecasting model for Singapore confirmed COVID-19 cases.

Accuracy of each model will be an important factor in the evaluation to accurately predict the number of confirmed COVID cases in the next two months. The accuracy range in the proposed workflow is set between 80% - 95% to account for external constraints that may affect the accuracy of the prediction on the number of confirmed cases. The external factor includes constraints such as not requiring public personnel who have tested positive via Antigen Rapid Test (ART) self-test to report their positive status to the government authority [5] or personnel who may be positive but are unaware of their status. The goal to reach the accuracy ranges between 80% - 95% is reasonable and realistic. This is also consistent with industry standards [6].

A. Results Visualization

Fig. 5 shows an overview of the prediction lines by the six models against the actual data. Visualization analysis of these graph models suggest that the polynomial regression, Prophet and Holt's linear models yield better predicted result against the actual Confirmed COVID-19 cases where the best predicted line is more aligned, compared to those of the linear regression, SVM and LSTM models. While the linear regression model has the worst performance as the predictions are either overshooting or really lower than what is expected.

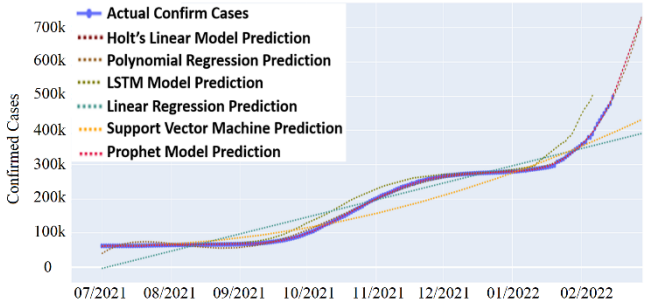

Fig. 5. Six models' Prediction Results against Actual Results

The visualisation analysis on each model performance can be further substantiate by the comparison on the performance metrices of RMSE and MAPE in Table IV. It shows that the Prophet model emerges in the first rank for having the best performance where RMSE = 1557.744836 and MAPE = 0.468827 indicating that predicted results by the Prophet model will be more accurate compared to the other five models. While the linear regression model has RMSE = 248669.028170 and MAPE = 38.127555 which yield the worst performance in terms of prediction accuracy as compared to the other five models.

TABLE IV. MODELS RMSE AND MAPE PERFORMANCE

Ranking	Model Name	RMSE	MAPE
1	Prophet Model	1557.744836	0.468827
2	Holt's Linear Model	4224.459153	0.475848
3	LSTM Model	7754.998209	11.709110
4	Polynomial Regression	15220.701223	2.304378
5	SVM	216168.791190	32.972292
6	Linear Regression	248669.028170	38.127555

Shown in Table IV, the top three performing models, i.e., Prophet linear, Holt's Linear and LSTM Models will be used

to further evaluate the actual accuracy performance by analysing the accuracy rate when predicting future trends.

B. Future Trends Analysis

Fig. 6 depicts the accuracy rate when predicting for future COVID-19 confirmed cases between 1st March – 28th May 2022 using three good performing models: prophet, Holt's linear and LSTM models. The accuracy trends using the Holt's linear and LSTM models are generally decreasing as the days advances with the LSTM model having lower accuracy rate. Whereas when using the prophet model for prediction, the accuracy trend fluctuates throughout the prediction range, but still maintains its accuracy at 70% when reaching the end of the prediction range. The gradual decrease in accuracy is probably due to certain policies implemented by the government to capturing and recording the actual confirmed COVID-19 cases. As of 10th April 2022, Ministry of Health Singapore announced that it is not necessary for the public to report their COVID positive status results. It will no longer log confirmed cases if the infected do not have severe symptoms. This thus affects the model capability to capture the trend as the dataset that was previously used to train the model is consistent reporting of COVID-19 cases.

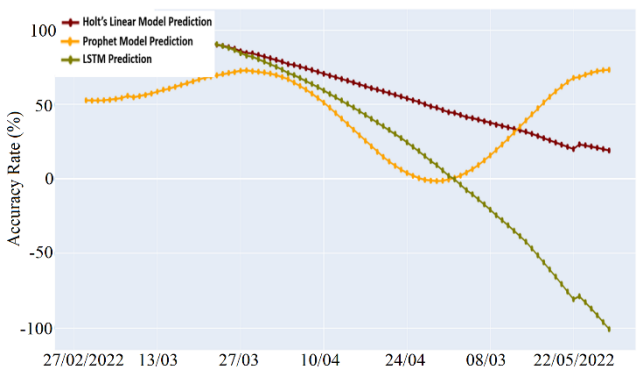

Fig. 6. Analyzing future trends accuracy of three good performing models

TABLE V. PREDICTIONS OF PROPHET MODEL FOR NEXT 87 DAYS

Day	Prophet Prediction	Actual Confirmed Cases	Model Accuracy
1	749994	748504	99.80%
2	768060	767663	99.94%
3	786126	785825	99.96%
4	804192	803389	99.90%
5	822258	819663	99.68%
…	…	…	
87	2267540	1278113	21.75%

Despite the external factor affecting the model accuracy, in comparison of all the six models, the prophet model is chosen by the proposed workflow as the most suitable model for COVID-19 prediction given that it has an accuracy rate of 70% for forecasting and prediction. TABLE V shows the predicted COVID-19 Confirmed cases for the next 87 days using the prophet model against the actual number of Confirmed Cases. From the general analysis of the Table V, the accuracy rate of the trend for prediction using the Prophet model shows very high accuracy of the 99% for the next 2 months when compared against the actual confirm cases. But accuracy decreases when the prediction hits over the 3rd month period indicating that such model is more accurate for short-term forecasting.

V. CONCLUSION

This study aims to develop a workflow to select the forecasting model that makes accurate predictions on COVID-19 confirmed cases in Singapore, in order to assist the government in promoting policies that can prevent the spread of the COVID-19 disease outbreak in the country.

In this study, a workflow is proposed to find a most suitable forecasting model for the COVID-19 confirmed cases in Singapore. The six candidate forecasting models are trained using the COVID-19 dataset exported from WHO website. In order to evaluate the performance of these models, two algorithms have been developed in this research. One proposed algorithm is the order algorithm to determine optimal degree order for the polynomial regression model. The other proposed algorithm is the optimizing algorithm for the Holt's linear model to calculate the optimal smoothing parameters.

The performance of these six datasets are compared in the workflow, including polynomial regression, linear regression, SVM, Prophet, Holt's Linear, and LSTM models. The prophet method model shows the best performances on RSME and MAPE, compared to other five models. It also obtains about 90% accuracy in forecasting the confirmed COVID-19 cases in for the next 87 days (about 3 months) ahead. It helps better prepare healthcare industry officials to plan resources and be ready to deal with the upcoming situation.

REFERENCES

[1] Wikimedia Foundation. "Template: COVID-19 pandemic data," Accessed: July 12, 2022. [online]. Available: https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data

[2] C. Trinidad, "Forecasting. Corporate Finance Institute," Accessed: July 12, 2022. [online]. Available: https://corporatefinanceinstitute.com/resources/knowledge/finance/forecasting/

[3] "Forecasting methods," Corporate Finance Institute. Accessed: July 19, 2022, [online]. Available: https://corporatefinanceinstitute.com/resources/knowledge/modeling/forecasting-methods/

[4] "Data-driven approaches to pandemic forecasting," Frontiers. (n.d.). Accessed: July 19, 2022, [online]. Available: https://www.frontiersin.org/research-topics/38630/data-driven-approaches-to-pandemic-forecasting

[5] S. Khalik, "Experts agree with MOH move to stop daily covid-19 updates to media, as 'they are no longer meaningful'," The Straits Times. https://www.straitstimes.com/singapore/health/moh-to-stop-issuing-daily-covid-19-updates-to-media-experts-say-reports-no-longer (Accessed: July 12, 2022).

[6] "How to know if your machine learning model has good performance: Obviously AI," Data Science without Code. (n.d.). Accessed: July 12, 2022, [online]. Available: https://www.obviously.ai/post/machine-learning-

modelperformance#:~:text=Good%20accuracy%20in%20machine%20learning,not%20only%20ideal%2C%20it's%20realistic.

[7] F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi:10.1109/ACCESS.2020.2997311.

[8] S. Maurya and S. Singh, "Time Series Analysis of the Covid-19 Datasets," IEEE International Conference for Innovation in Technology, 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298390.

[9] S. Kanimozhi, K. S. Kalaivani and S.K.C. Abinayaa, "An Analysis and Forecasting of Covid-19 Spread in Tamil Nadu based on Additive Model for Non-Linear Pattern," 6th International Conference on Inventive Computation Technologies, 2021, pp. 1279-1283, doi: 10.1109/ICICT50816.2021.9358564.

[10] S. Goled, "Why are people bashing Facebook Prophet," Analytics India Magazine. Accessed: July 19, 2022, [online]. Available: https://analyticsindiamag.com/why-are-people-bashing-facebook prophet/

[11] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara and M. Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting," 11th International Conference on Cloud Computing, Data Science & Engineering, 2021, pp. 989-995, doi: 10.1109/Confluence51648.2021.9377137.

[12] A. Kumar and K. Kaur, "A hybrid SOM-Fuzzy time series (SOMFTS) technique for future forecasting of COVID-19 cases and MCDM based evaluation of COVID-19 forecasting models," International Conference on Computing, Communication, and Intelligent Systems, 2021, pp. 612-617, doi: 10.1109/ICCCIS51004.2021.9397216.

[13] A. Mohammadi, D. Chumachenko and T. Chumachenko, "Machine Learning Model of COVID-19 Forecasting in Ukraine Based on the Linear Regression," IEEE 12th International Conference on Electronics and Information Technologies, 2021, pp. 149-153, doi: 10.1109/ELIT53502.2021.9501122.

[14] J. P. Devarajan, A. Manimuthu and V. R. Sreedharan, "Healthcare Operations and Black Swan Event for COVID-19 Pandemic: A Predictive Analytics," in IEEE Transactions on Engineering Management, doi: 10.1109/TEM.2021.3076603.

[15] S. Dash, C. Chakraborty, S. K. Giri, S. K. Pani and J. Frnda, "BIFM: Big-Data Driven Intelligent Forecasting Model for COVID-19," in IEEE Access, vol. 9, pp. 97505-97517, 2021, doi: 10.1109/ACCESS.2021.3094658.

[16] G. Srikanth, N. Nukavarapu and S. Durbha, "Deep Reinforcement Learning Interdependent Healthcare Critical Infrastructure Simulation model for Dynamically Varying COVID-19 scenario - A case study of a Metro City," IEEE International Geoscience and Remote Sensing Symposium, 2021, pp. 8499-8502, doi: 10.1109/IGARSS47720.2021.9554875.

[17] H. A. Abd Rahman, M. A. Rahman, A. N. Rozaimi and I. B. Zulnahar, "Forecasting of COVID-19 in Malaysia: Comparison of Models," IEEE International Conference on Computing, 2021, pp. 324-329, doi: 10.1109/ICOCO53166.2021.9673498.

[18] Y. Cheng, Y. Bai, T. Xu and H. Hu, "Global COVID-19 development trend forecast based on machine learning," International Conference on Big Data, Information and Computer Network, 2022, pp. 128-131, doi: 10.1109/BDICN55575.2022.00032.

[19] Y. Zaychenko, H. Zaichenko and G. Hamidov, "Investigation of recurrent networks LSTM in the problem of Covid-19 forecasting," IEEE 16th International Conference on Computer Sciences and Information Technologies, 2021, pp. 9-12, doi: 10.1109/CSIT52700.2021.9648696.

[20] S. S. HELLİ, Ç. DEMİRCİ, O. ÇOBAN and A. HAMAMCI, "Short-Term Forecasting COVID-19 Cases In Turkey Using Long Short-Term Memory Network," Medical Technologies Congress, 2020, pp. 1-4, doi: 10.1109/TIPTEKNO50054.2020.9299235.

[21] P. Cawood and T. L. van Zyl, "Feature-Weighted Stacking for Nonseasonal Time Series Forecasts: A Case Study of the COVID-19 Epidemic Curves," 8th International Conference on Soft Computing & Machine Intelligence, 2021, pp. 53-59, doi: 10.1109/ISCMI53840.2021.9654809.

[22] World Health Organization. (n.d.). "WHO Coronavirus (COVID-19) Dashboard," Accessed: July 12, 2022, [online]. Available: https://covid19.who.int/data