2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) | 978-1-6654-0066-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEMCON53756.2021.9623074

An Ensemble Machine Learning Approach For Time Series Forecasting of COVID-19 Cases

Renato R. Maaliw III
College of Engineering
Southern Luzon State University
Lucban, Quezon, Philippines
rmaaliw@slsu.edu.ph

Melvin A. Ballera
Graduate Programs
Technological Institute of the Philippines
Manila, Philippines
melvin.ballera@tip.edu.ph

Zoren P. Mabunga
College of Engineering
Southern Luzon State University
Lucban, Quezon, Philippines
zmabunga@slsu.edu.ph

Aubee T. Mahusay
College of Engineering
Southern Luzon State University
Lucban, Quezon, Philippines
amahusay@slsu.edu.ph

Dhenalyn A. Dejelo
College of Allied Medicine
Southern Luzon State University
Lucban, Quezon, Philippines
ddejelo@slsu.edu.ph

Mariebeth P. Seño
College of Arts and Sciences
Southern Luzon State University
Lucban, Quezon, Philippines
mpseno@slsu.edu.ph

***Abstract*—Forecasting assists governments, epidemiologists, and policymakers make calculated decisions to mitigate the spread of the COVID-19 pandemic, thus saving lives. This paper presents an ensemble machine learning model by combining the distinctive strengths of autoregressive integrated moving averages (ARIMA) and stacked long short-term memory networks (S-LSTM) using extensive training procedures and model integration algorithms. We validated the model's generalization capabilities by analyzing time series data of four countries, such as the Philippines, United States, India, and Brazil spanning 467 days. The quantitative results show that our ensemble model outperforms stand-alone models of ARIMA and S-LSTM for a 15-day forecast accuracy of 93.50% (infected cases) and 87.97% (death cases).**

Keywords—combined time series models, infectious disease, optimization, pandemic, prediction, neural networks

I. INTRODUCTION

The novel coronavirus (COVID-19) pandemic poses a threat and drastically changed different aspects of our lives. To say that it had a devastating and far-reaching global impact more than a year since its declaration as a severe health crisis by the World Health Organization (WHO) is an understatement. As of this writing, the outbreak has stretched out to over 215 countries, claimed 3.9 million lives, and infected 180 million worldwide due to its contagious nature and extended incubation periods without symptoms [1]. Epidemiologists believed that the struggle to contain the pandemic is far from over as it remains active in different parts of the planet. Despite the availability of vaccines, recent discoveries of new variants formed by mutations are adding to new surges [2].

The exponential proliferation of the disease has placed tremendous pressure on healthcare infrastructures and the economy. It has transformed the entire world and put the government's preparedness to the test as an unexpected crisis threatens everyone's safety [3]. At the onset of the pandemic where viable vaccines and medications are not yet available, authorities made numerous efforts to curb the spread of COVID-19 significantly. Non-pharmacological approaches, such as social distancing, lockdowns, travel restrictions, contact tracing, mass testing, workplace and school closures, and mandatory wearing of personal protective equipment such as face masks, were implemented. Experts agree on the importance of obtaining reliable empirical knowledge to decide the next course of action that has the fewest negative consequences for people's health and economic well-being [4].

Today, data are generated at an unprecedented rate. When processed correctly, they can reveal invaluable and in-depth information to solve problems from various perspectives. Forecasting, while not perfect, assists governments and epidemiologists in taking necessary responses to control disease's future incidence. It is a critical component in supporting policymakers in making sound judgments for optimizing guidelines and strategies [5]. Moreover, authorities need reliable estimates of likely cases (infected and deaths) in exploring several 'what if' scenarios in decision-making and ascertain the efficiency of solutions in managing the virus's progression. Numerous forecasting models were implemented to explain observed trends and project future patterns designed to help public healthcare practitioners plan and respond appropriately.

The SIR (susceptible/infected/recovered) and SEIR (with exposed parameter) are the two most often used epidemiological models based on ordinary differential equations in predicting the extent of a disease. While it reflects core epidemiological dynamics such as infection and recovery rates, its parameters rely heavily on assumptions that do not reflect the actual cases [6]. Individual behavior, for example, is exceptionally impossible to quantify during real-world outbreaks, and even a slight change in setting the model's attributes can result in severely inaccurate predictions. This methodology assessed the progression [7 – 9] of dengue, Ebola virus, COVID-19, malaria, tuberculosis, and severe acute respiratory syndrome (SARS).

Judgment is another forecasting approach, but human prejudices mainly influenced them. Expert's collective foresight is well-thought-out but less reliable than quantitative models because of subjectivity. Few studies have shown that it can compete with statistical and epidemiological approaches

[10]. In a study by [11], specialists' forecast accuracy of COVID-19 cases using collective judgment through survey varies. When aggregated, it displayed varying levels of uncertainty. For these reasons, it is only prescribed as a last option when data is incomplete or, in worst cases, not available at all. On the other hand, most data scientists applied time series forecasting, a well-established statistical approach to tracking diseases using exact historical data [12]. Compared to the two preceding models, it forecasts epidemiological incidence with fewer assumptions by not considering the dynamics of transmissions. The model is more straightforward to explain to the masses due to its mathematical simplicity. However, while linear models could adequately describe many fundamental processes, pandemics are intrinsically nonlinear. Numerous studies have used rolling averages, exponential smoothing, and autoregressive integrated moving average (ARIMA) to forecast COVID-19 incidence since the outbreak began [12] [16] [17] [19].

Emerging technology in data science such as Artificial Intelligence (AI) is a breakthrough research area with applications ranging from academia [13] [14], industry and health. It is adaptive to different domains, including time series forecasting of an infectious disease. Advanced machine learning algorithms solve the limitation of epidemiological and statistical models that relies heavily on assumptions to its hyperparameters and linearity features. The works of [15 – 17] showed it could produce substantial prediction accuracy for COVID-19 cases by implementing stand-alone and combined deep learning forecasting models. Based on the literature, limited studies integrate the distinctive strengths of traditional mathematical and deep learning or neural network models to handle seasonality, nonlinear, and complex patterns for time series predictions.

The research's primary contribution was creating an accurate ensemble machine learning model in forecasting COVID-19 cases using a normalized exponential weighted algorithm. Our model was capable of outperforming individual models as it reduces over or under-fitting on forecasting COVID-19 cases. Moreover, the research findings are critical for resource allocation, policy formulation, and streamlining management procedures during a pandemic.

II. Methodology

A. Data Source

The COVID-19 open data were acquired from Johns Hopkins University's Center for System Science and Engineering (CSSE). It is available in a time series format that contains information on confirmed cases, recoveries, and fatalities reported per country. We extracted stable versions of datasets by observing possible changes on record every week from March 8, 2020 to June 18, 2021. According to incidence statistics, the Philippines, India, Brazil, and the United States are the pandemic's most impacted nations. Moreover, choosing different data with unique characteristics can validate the model's forecasting generalization capabilities and remove biases. Figure 1 and Figure 2 show the cases' time series plots.

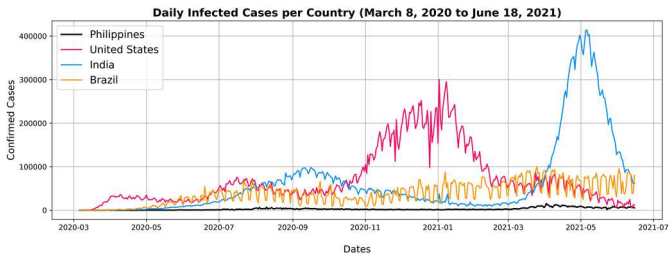

Fig. 1. COVID-19 daily infected cases per country

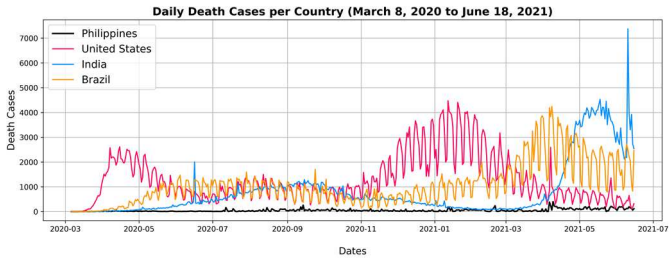

Fig. 2. COVID-19 daily death cases per country

B. Data Transformation, Train-Test Split and Models

Deep learning models are sensitive to extreme value disparities because each layer's inputs change throughout the training phase. Consequently, unscaled variables lead to exploding gradients, resulting in a costly, unstable, and unsuccessful learning process. To overcome the issue, we implemented a pre-processing technique of normalization (*min-max scaling*) to scale the features between 0 and 1. Numerous studies confirm the method's validity in terms of prediction accuracy improvement [18]. We divided the dataset into 80% training and 20% testing ratio for model fitting.

Our model forecasts two associated variables for a 15-day horizon: daily reported infected and death cases. Several models, including ARIMA, long short-term memory (LSTM) networks with varied architectures, and ensemble machine learning models, were constructed to compare forecast accuracy.

C. Stationarity Testing

Time series (TS) data contains three components: trend, seasonality, and residuals. A trend occurs when external factors often cause a specific pattern. Therefore, it is crucial to understand the structure of TS to apply proper forecasting methodologies. A stationary series is easy to analyze because it is not vulnerable to the influence of time components. Such series have a mean and variance that are primarily constant over time. Conversely, a non-stationary TS exhibits trend and seasonality with mean, variance, and standard deviation changing through time.

We have examined each dataset using an augmented Dickey-Fuller (ADF) test to determine the nature of the TS. Based on the metrics, a p-value less than or equal to 0.05 means the series is stationary, while the opposite (greater than 0.05) indicates non-stationary data with a unit root's presence. Table 1 depicts the categorization of each time series based on the stationarity tests results.

TABLE 1
STATIONARITY TEST RESULTS

Time-series data	p-value	Property
Daily infected cases		
Philippines	0.363	Non-stationary
United States	0.469	Non-stationary
India	0.073	Non-stationary
Brazil	0.611	Non-stationary
Daily death cases		
Philippines	0.026	Stationary
United States	0.175	Non-stationary
India	0.970	Non-stationary
Brazil	0.522	Non-stationary

D. Trend and Seasonality Decomposition

Decomposition deconstructs time series into systematic (trend & seasonality) and non-systematic (residual or noise) components to understand the data structure better in improving forecast accuracy. The primary objective is to estimate the trend and seasonality effects on the data. Variability falls into two categories: an additive (linear) model where seasonal variations remain relatively constant across time and a multiplicative (exponential) model where seasonal variations increase over time. The information divulged from the decomposition plots served as the basis for the proper configurations of time series models. Figure 3 and Figure 4 display the seasonal & trend using Loess (STL) decomposition graphs.

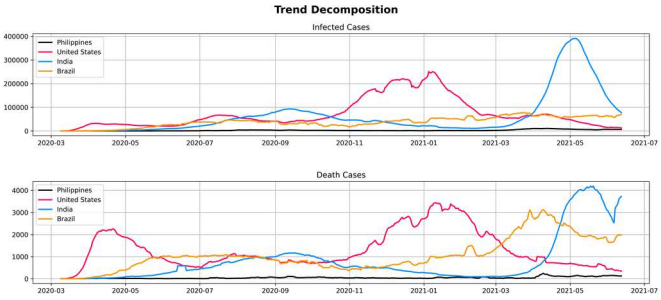

Fig. 3. COVID-19 trend decomposition per country

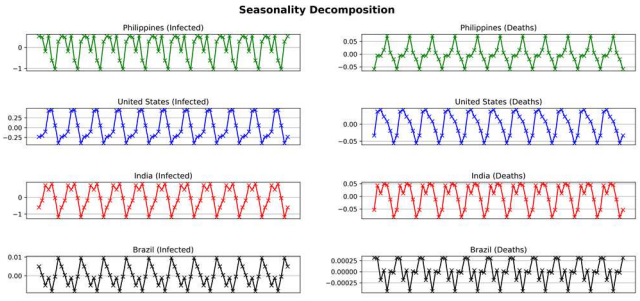

Fig. 4. COVID-19 seasonality decomposition per country

E. Autoregressive Integrated Moving Average

ARIMA is a standard time series forecasting model widely used for decades. While it effectively represents non-stationary data in real-world applications, it is incapable of handling nonlinear relationships between variables basing its concept on linearity. To its advantage, it has a unique strength in seasonal component modeling [19]. The model is denoted by ARIMA*(p, d, q),* where the *AR* (autoregression) uses the dependent relationship of its past and current values or the lag order (*p*). Transformation of the time series data is handled by *I* (integrated). It subtracts observations from previous values at various times to make a time series stationary or the degree of differencing (*d*). The *MA* (moving average model) portion accounts for the dependency between observations and the residual error after averaging is applied or the moving average window (*q*). It indicates that the forecast's outcome is reliant on historical values. An extended version of the model with nomenclature of SARIMA(*p, d, q*)(*P, D, Q*)*m* supports the seasonality element of a time series. It adds three hyperparameters, where *P*, *D*, and *Q* refer to the seasonal component of the model, such as autoregressive, differencing, and moving average orders. The *m* part pertains to the number of periods for each season. Inspection of autocorrelation (ACF) and partial autocorrelation (PACF) plots enables identifying the model's hyperparameters. This manual process was difficult and complex. We solved the dilemma using a grid search technique via statsmodel's *auto-arima* [20] for computing the least Akaike information criterion (AIC) score. It is more intuitive and effective than laborious graph reading. Table 2 shows the different configurations for the model.

TABLE 2
ARIMA MODEL CONFIGURATION

Time-series data	Best configuration	AIC score
Daily infected cases		
Philippines	SARIMA(2,1,4)(1,0,1)[7]	7515.96
United States	SARIMA(5,1,3)(0,0,1)[7]	10104.37
India	SARIMA(1,1,3)(2,0,0)[7]	9309.96
Brazil	SARIMA(5,1,2)(2,0,2)[6]	9885.22
Daily death cases		
Philippines	ARIMA(1,1,1)	4779.18
United States	ARIMA(5,1,2)	6611.84
India	ARIMA(2,1,2)	6634.42
Brazil	ARIMA(1,1,3)	6624.55

F. Long Short-Term Memory Networks

Before drilling down into the intricacies of various LSTM architectures, it is essential to understand the fundamental dynamics of its memory units and why they are good at extracting patterns and the best suitable for long sequence predictions. LSTMs belong to the family of recurrent neural networks (RNN) that incorporates the principle of memory lines, surpassing the conventional feed-forward neural networks in many ways. Additionally, it can circumvent the constraints of ordinary sequence forecasting methods by adjusting to the nonlinear characteristics of the data, hence increasing accuracy. They generate output sequences by iteratively traversing a range of series lengths and are also aware of the temporal constructs of their inputs due to their internal states. Its memory block is the most significant network component because it conserves the parameters that prevent vanishing gradients over long periods. Using different gates in LSTM allows data processing through an activation function to yield a consistent output by feeding positive values to the subsequent gates. Figure 5 illustrates the composition of an LSTM unit with forget, input, and output gates serving as filters.

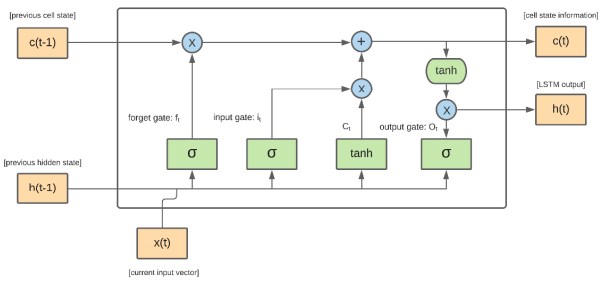

Fig. 5. Structure of a LSTM network unit

The forget gate decides which information is useful, needs attention, and what to disregard. An activation (sigmoid) function utilizes data from the current input *X(t)* and hidden state *h(t-1)* to create values between 0 and 1. The update of the cell state begins with the transfer of the current state *X(t)* and a hidden state *h(t-1)* to a different activation function. The new memory network is a *tanh*-activated neural network that has learned to produce a new update vector by combining the prior hidden state *c(t-1)* with a new input data. This vector effectively includes information about the incoming inputs in light of the last hidden state's context. It indicates how much each component of the network's long-term memory should be updated in response to new data. The final network's prediction *h(t)* is computed using the hidden state's stored information *c(t)*. The deep learning model with the complexities of the process is a robust algorithm to create a sequential model for time series forecasting.

G. Bidirectional LSTM Networks

Bidirectional LSTMs (BDLSTM) extends the capabilities of regular LSTMs in improving model performance by stepping across input sequences in both forward and backward directions [21]. This architectural design involves replicating the first recurrent layer to create two layers adjacent to each other. The first sequence is unaltered, while the second is a reversed copy of the first, resulting in a faster and effective learning process. Figure 6 exhibits the architecture of the BDLSTM.

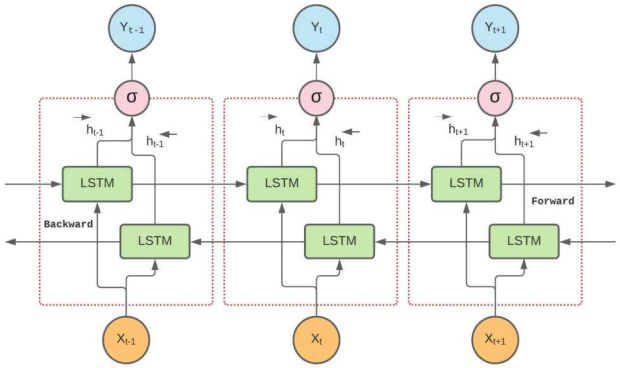

Fig. 6. Structure of a bidirectional LSTM networks

We performed the bidirectional LSTM approach in Python using the *keras* package, allowing for the forward and backward outputs to be merged with different modes such as summation, averaging, and concatenation functions before passing them to the next layer. The output vector (Y_t) is derived using the equation:

$$Y_t = \sigma(\vec{h}_t, \overleftarrow{h}_t) \quad (1)$$

where σ function synthesized two output sequences while the outcome of the final layer's output is described by a vector of $Y_t = [Y_{t-n}, \ldots, Y_{t-1}]$, in which the last element is the predicted sequences for the successive iterations.

H. Stacked Bidirectional and Unidirectional LSTM Networks

In neural network modeling, layer stacking is mainly used to solve complex nonlinear sequences to construct higher levels of data representation. Through the addition of layers, input observations are substantially abstracted by recombining previously acquired learned patterns. The framework works as the output of an LSTM layer is fed as input to other succeeding LSTM layers as an enhancement mechanism. In this sense, the BDLSTM is an ideal candidate for the initial primary layers of the deep learning model, which is responsible for learning valuable information from a time series. For term abbreviation, we refer to the architecture as a stacked LSTM (S-LSTM).

We purposely design the last layer of the stack as a unidirectional LSTM because it only needs to compute the final prediction forward, as data characteristics were already learned from the preceding layers in forecasting unseen sequences. Figure 7 illustrates the architecture of our deep learning model.

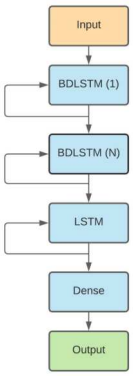

Fig. 7. Stacked bidirectional LSTM and unidirectional LSTM architecture

I. Neural Network's Hyperparameter Optimizations

Hyperparameter tuning is vital for machine learning to perform optimally on a test or validation set. In contrast to model parameters, before training, these configurations are specified by the user. To date, it is one of the most challenging issues to address when implementing machine learning algorithms. It is also the most neglected yet crucial step in deep learning modeling. Due to the complexity of manually defining and testing hyperparameters for neural networks, we utilized a sequential model-based optimization technique [22] via the *scikit-optimize* library in Python. Table 3 shows the mean collective tuned settings based on extensive runs. It yielded an average root mean squared error (RMSE) of 4369.95 and 316.80 for the infected and death cases test sets, respectively.

TABLE 3
OPTIMIZED HYPERPARAMETER CONFIGURATIONS

Time-series data	Hyperparameter	Value
Daily infected cases		
Philippines	Learning Rate	0.005
	BDLSTM-1 Neurons	94
	BDLSTM-2 Neurons	81
United States	UDLSTM Neurons	62
	Dropouts	0.2
	Dense Layer	1
India	Inputs	7
	Batch Size	1
	Activation Function	RELU
Brazil	Loss Function	MSE
	Optimizer	ADAM
	Training Epochs	400
Daily death cases		
Philippines	Learning Rate	0.005
	BDLSTM-1 Neurons	82
	BDLSTM-2 Neurons	67
United States	UDLSTM Neurons	46
	Dropouts	0.2
	Dense Layer	1
India	Inputs	7
	Batch Size	1
	Activation Function	RELU
Brazil	Loss Function	MSE
	Optimizer	ADAM
	Training Epochs	400

Note: BDLSTM (bidirectional LSTM), UDLSTM (unidirectional LSTM)

Along with fine-tuning hyperparameters, we analyzed and compared convergence loss plots to establish the optimal number of training runs (epochs) and assess the effects of stacking different types of LSTM layers. Figure 8 shows the plots.

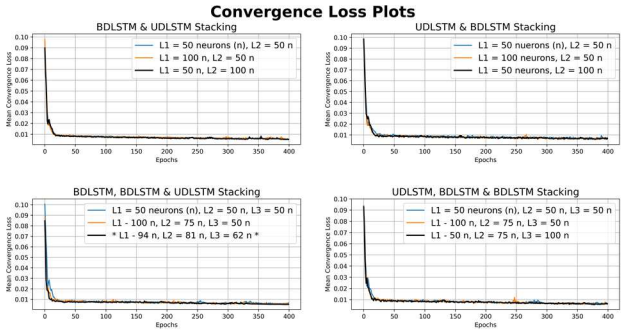

Fig. 8. Convergence loss plots of stacking different LSTM layers

J. Combining Model Forecasts

Traditional ensemble methods such as mode or average in combining multiple models' prediction decrease bias and variance compared to single model forecasting. However, these linear combination approaches, on the other hand, disregard the inherent connections between the participating models. The ensemble's forecasting accuracy suffers because of this restriction, especially when the constituent forecasts are highly linked. To address the limitations of traditional methods, we employ a weighted average ensemble, in which the best predictions of several individual models are dynamically weighted [23]. A normalized exponential weighted algorithm calibrates the influence associated with the best prediction of ARIMA and S-LSTMs. Figure 9 explains the algorithms.

1. INPUT: A model i predicting outcome f_i^t for round t and a free parameter λ
2. $w_i^1 \longleftarrow \frac{1}{N}$ i = 1, …., N models (initial weights of each model's prediction to 1/N that sum of weights is equal to 1)
3. For training samples $t = 1$ to T and for each model $i = 1$ to N do
4. $l(f_i^t y_t) = (f_i^t y_t)^2$ (calculation of squared error where y_t is the actual value)
5. $w_i^{t+1} \longleftarrow w_i^t e^{\lambda l(f_i^t y_t)}$
6. $w_i^{t+1} \longleftarrow w_i^{t+1} / \sum_{i=1}^{N} w_i^{t+1}$ (normalized weights)

Fig. 9. Normalized exponential weigthing algorithm

The strategy begins by providing equal weights (line 1) to all forecasts given a set of *N* predictions from the training data (i.e., ½ weights for ARIMA and S-LSTM forecast; line 2). Assuming every weight to be *1/N*, at every stage of the training, the total of all models' weights equals 1. Next, the squared error of each forecast and actual values are calculated for the training samples (line 4). Each model's weights are updated based on the squared error for various forecasts using a free learning-rate λ parameter (line 5). Finally, normalized weights are computed for each prediction by dividing them with the summation of the weights for all models resulting in values between 0 and 1 (line 6). The procedure is repeated until all training samples have been completed. We deliberately calibrated the λ parameter by iteratively supplying values from 0 to 1 with 0.1 steps. The ARIMA and S-LSTM's optimized weights are selected based on the minimum computed RMSE values.

K. Forecast Evaluation Metrics

We assessed the accuracy of various forecasting models using three commonly used error metrics: root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean bias error (MBE). Equations 2, 3, and 4 specified the complete formulas:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(a_i - p_i)^2}{N}} \quad (2)$$

$$MAPE = \frac{\sum_{i=1}^{N}\left|\frac{a_i - p_i}{p_i}\right| \text{x } 100}{N} \quad (3)$$

$$MBE = \frac{\sum_{i=1}^{N}\left|\frac{a_i - p_i}{p_i}\right|}{N} \quad (4)$$

where a_i is the actual value, p_i is the forecast, and *N* is the number of testing samples. The RMSE is an accurate measure to compare prediction errors of multiple models or model configurations for a specific variable, but not between variables, because it is scale-dependent. In contrast to the RMSE, MAPE is not scale-dependent used to compare forecasts across time series data with varying scales. Finally, the MBE reflects the model's mean forecast error to determine if the prediction is under or over-estimated.

III. Results

This section analyzes and evaluates the fit and accuracy of different time series forecasting models such as ARIMA, S-LSTM, and the ensemble model. The following subsections also detail the outcomes for a 15-day COVID-19 case forecast to test each model's generalization capabilities.

A. The Effects of Layer Stacking on Learning Equilibrium

Figure 8 exhibits the convergence loss function plots of multiple variations of LSTM stacks against the number of epochs. Due to the computational characteristics of a 3-stacked LSTM configuration, the networks took longer to attain equilibrium compared to 2-stacked settings, but it proved to attain a stabilized results. It demonstrates that a BDLSTM at the first and second tiers of the stacks instead of a UDLSTM improves the network's performance. Additionally, a decreasing number of neurons for each layer is advantageous because the first layer must understand the overall structure of the input before passing it to the succeeding layers, rather than the opposite (see Table 3).

B. Ensemble Model

Table 4 displays the optimal ensemble weights derived from the normalized weighted algorithm, including the learning rate parameter (λ). Based on the calculations, the stacked LSTMs have the highest model weights for the ensemble model except for the Philippines' (0.53) death cases. This is due to the robustness of neural networks to learn the underlying structure of series data when exhibiting nonlinearities.

TABLE 4
ENSEMBLE MODEL OPTIMAL WEIGHTS (TEST DATA)

Time-series data	Weights for Ensemble Model		Learning Rate (λ)
	ARIMA	Stacked LSTMs	
Daily infected cases			
Philippines	0.48	0.52	0.4
United States	0.47	0.53	0.5
India	0.44	0.56	0.5
Brazil	0.42	0.58	0.6
Daily death cases			
Philippines	0.53	0.47	0.7
United States	0.45	0.55	0.6
India	0.34	0.66	0.9
Brazil	0.41	0.59	0.5

C. Model Fitness Test

Figure 10 and Figure 11 show the results of the goodness-of-fit plots based on the test data of different models. Furthermore, Table 5 shows that the ensemble model outperforms single models with RMSE values of 679.46, 3438.45, 4824.23, and 6008.514 for infected cases. Lower values are also obtained for death cases with 39.77, 158.25, 435.53, and 487.26.

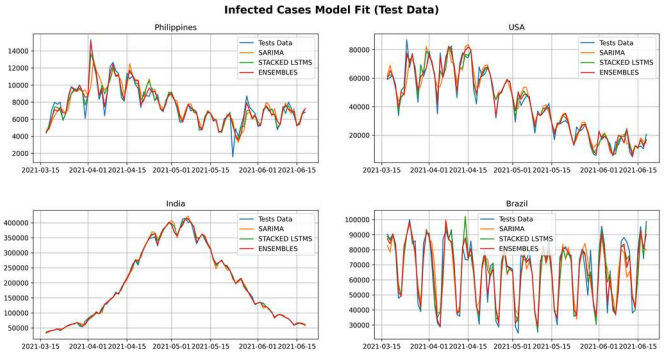

Fig. 10. Graphical plot of model fitness (infected cases)

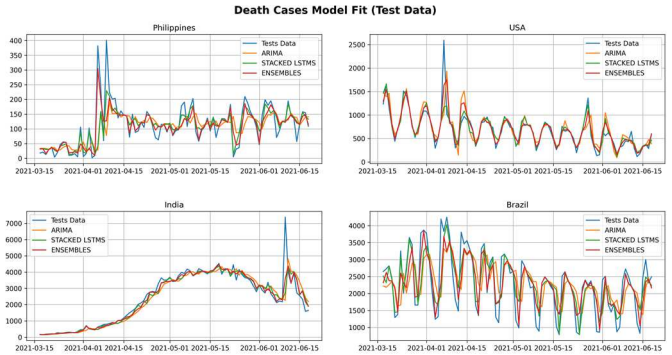

Fig. 11. Graphical plot of model fitness (death cases)

TABLE 5
MODEL'S FIT BASED ON RMSE SCORES (TEST DATA, N = 94 (20%))

Time-series data	ARIMA	Stacked LSTMs	Ensemble Model
Daily infected cases			
Philippines	1230.60	718.31	**679.46**
United States	7630.89	4344.69	**3438.45**
India	9198.68	5343.35	**4824.23**
Brazil	11035.60	7073.48	**6008.514**
Daily death cases			
Philippines	65.57	46.32	**39.77**
United States	261.01	174.72	**158.25**
India	631.97	530.16	**435.53**
Brazil	803.85	515.60	**487.66**

D. Infected Cases Forecast

Table 6 and Table 7 show the evaluation ratings of the 15-day forecast for infected cases. It is found that the ensemble model exceeds the performance of the single models with lower RMSE. The results indicated the superiority of the ensemble model with obtained MAPE values of 4.13%, 10.51%, 4.39%, and 6.63% for COVID-19 projected cases from the Philippines, United States, India, and Brazil. S-LSTM performs better than the ARIMA in single model comparisons as it reduces MAPE values by 9.53% on average. Figure 12 illustrates the infected cases forecast.

E. Death Cases Forecast

Table 8 and Table 9 detail the predicted 15-day outcomes of death instances. The tables demonstrate that the ensemble method approach produces an accurate forecast compared with other models by obtaining lower RMSE. In addition, the ensemble model achieved better MAPE values of 11.90%, 17.22%, 7.57%, and 11.41% from the Philippines, United States, India, and Brazil. The S-LSTM performed relatively better than the ARIMA as it decreases MAPE metrics by an average of 13.15%. Figure 13 exhibits the death cases forecasts.

F. Bias Estimation

Table 6 and Table 7 shows that for infected cases, different models over forecast the prediction except for ARIMA and S-LSTM for the USA (-2859.31 & -2787.73); ARIMA, S-LSTM, and ensemble model for India (-8639.41, -5881.86 & -1606.80). Table 8 and Table 9 revealed that all models overshoot the death case forecasts. Still, the MBE results of the ensemble model are relatively closer to the actual case values.

TABLE 6
EVALUATION METRICS FOR A 15-DAY FORECAST OF INFECTED CASES FOR PHILIPPINES & USA (JUNE 19 – JULY 3, 2021)

Philippines				United States			
Evaluation Metrics	Forecast			Evaluation Metrics	Forecast		
	ARIMA	Stacked LSTMs	Ensemble Model		ARIMA	Stacked LSTMs	Ensemble Model
RMSE	733.048	418.876	**273.765**	RMSE	6075.473	5575.300	**2298.177**
MAPE	12.606	6.264	**4.132**	MAPE	25.034	17.019	**10.514**
MBE	625.95	304.2	**218.933**	MBE	-2859.313	-2787.733	**1049.400**

TABLE 7
EVALUATION METRICS FOR A 15-DAY FORECAST OF INFECTED CASES FOR INDIA & BRAZIL (JUNE 19 – JULY 3, 2021)

India				Brazil			
Evaluation Metrics	Forecast			Evaluation Metrics	Forecast		
	ARIMA	Stacked LSTMs	Ensemble Model		ARIMA	Stacked LSTMs	Ensemble Model
RMSE	9819.826	7342.331	**2814.611**	RMSE	19671.881	8945.567	**4217.296**
MAPE	18.582	12.826	**4.393**	MAPE	32.178	14.159	**6.639**
MBE	-8639.419	-5881.866	**-1606.8**	MBE	10289.030	4584.866	**2630.266**

TABLE 8
EVALUATION METRICS FOR A 15-DAY FORECAST OF DEATH CASES FOR PHILIPPINES & USA (JUNE 19 – JULY 3, 2021)

Philippines				United States			
Evaluation Metrics	Forecast			Evaluation Metrics	Forecast		
	ARIMA	Stacked LSTMs	Ensemble Model		ARIMA	Stacked LSTMs	Ensemble Model
RMSE	32.417	21.628	**16.439**	RMSE	83.742	55.818	**34.707**
MAPE	28.552	17.738	**11.901**	MAPE	45.980	25.620	**17.220**
MBE	13.989	11.266	**9.866**	MBE	44.139	23.400	**15.800**

TABLE 9
EVALUATION METRICS FOR A 15-DAY FORECAST OF DEATH CASES FOR INDIA & BRAZIL (JUNE 19 – JULY 3, 2021)

India				Brazil			
Evaluation Metrics	Forecast			Evaluation Metrics	Forecast		
	ARIMA	Stacked LSTMs	Ensemble Model		ARIMA	Stacked LSTMs	Ensemble Model
RMSE	201.274	140.752	**106.252**	RMSE	490.688	292.981	**196.034**
MAPE	15.347	9.198	**7.575**	MAPE	31.613	16.324	**11.410**
MBE	141.596	84.866	**64.330**	MBE	64.324	59.466	**35.933**

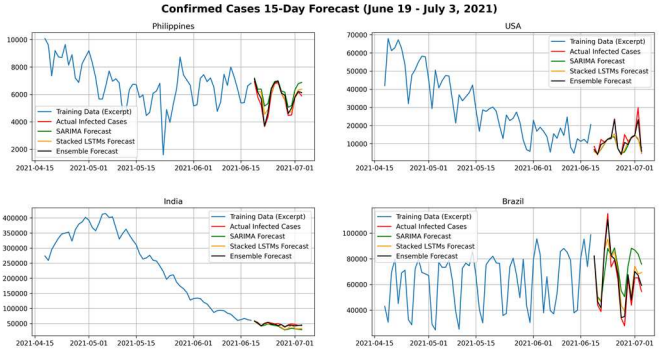

Fig. 12. 15-day forecast plots (infected cases)

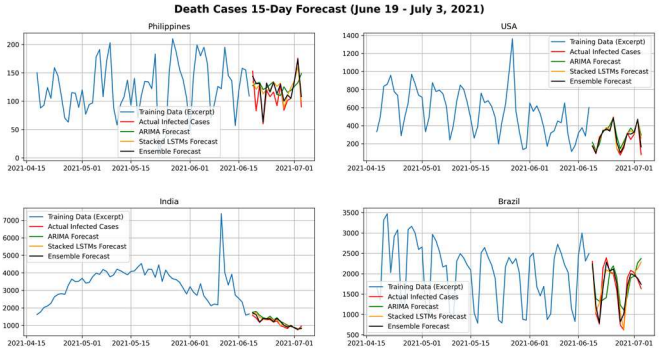

Fig. 13. 15-day forecast plots (death cases)

IV. DISCUSSIONS

This study demonstrated the ensemble machine learning approach's robustness in improving forecast accuracy beyond the capabilities of stand-alone models through extensive training procedures. The empirical results show that the ensemble model outperforms both ARIMA and S-LSTM with a collective mean prediction precision for unseen data of 93.50% (infected cases) and 87.97% (death cases). Moreover, the model reduces forecast error rates for afflicted and mortality incidence by a combined average of 10.91% and 11.76%, respectively. Our results clearly illustrate that integration of the unique strengths of various models coupled with comprehensive optimization strategies (e.g., stationarity tests, decomposition, grid searches, hyperparameter tuning & neural network layer stacking configurations) during the training phase and a normalized weighted algorithm for combining each model provides a significant boost in forecast accuracy. Fine-tuning models are computationally expensive and time-consuming; however, we firmly believe that the benefits outweigh their downside. The findings of this study are at par with the works of [21] [23] [24] [25] in ensemble forecasting and we firmly believe that our model is applicable to other domains. Figure 12 and Figure 13 suggested that the pandemic is still progressing based on the projected trends of cases per country due to the discovery of highly transmissible variants of

the virus such as beta (B.1.351), gamma (P.1), alpha (B.1.1.7), and delta (B.1.617.2). Strict implementations of governmental policies, the practice of minimum health and safety protocols, and active vaccination rollouts are still recommended. Like any forecasting model, a sharp increase or decline in the number of cases will force our model to over or underestimate the prediction (see Figure 10, Philippines; Figure 11, Philippines, USA & India) due to factors such as lockdowns, vaccinations, and other related measures.

V. Conclusions and Future work

Humanity is confronting an unprecedented emergency health crisis in its modern history as it battles the onslaught of the COVID-19 pandemic, which requires an extraordinary response from multidisciplinary collaborations. Data science and machine learning are powerful technologies that significantly solve these enormous challenges by anticipating the disease's progressions. Therefore, this advanced inferential knowledge can save millions of lives. It gives governments, healthcare experts, policymakers, and the general public a broader perspective on establishing appropriate plans to slow down or eradicate the virus. Our research discovered the potential of the ensemble machine learning approach in creating an accurate prediction model through a combination of well-established mathematical and sophisticated deep learning models. We evaluated data from four highly impacted nations encompassing 467 days between March 8, 2020, and June 18, 2021. Empirical results show that our ensemble model outclasses the single models of ARIMA and S-LSTM in forecasting unseen data of COVID-19 infected and death cases with an average forecast accuracy of 90.73%. While we have made significant modeling progress, hurdles persist due to the inherent difficulties in capturing human behaviors, environmental factors, and related social aspects during a pandemic. The authors plan to improve the model's forecasting accuracy for future work by exploring different deep learning algorithms and considering essential data like transmission, recovery, and vaccination rates.

Acknowledgment

The authors would like to thank the Southern Luzon State University for supporting this research.

References

[1] World Health Organization, "Weekly epidemiological update on COVID-19", WHO Bulletin, Available at: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---29-june-2021.

[2] S. Otto, T. Day, J. Arino et al., "The origins and potential future of SARS-COV-2 variants of concern in the evolving COVID-19 pandemic", Current Biology, volume 31, issue 14, pp. 918-929, 2021.

[3] H. Feyisa, "The world economy at COVID-19 quarantine: a contemporary review", International Journal of Economics, Finance and Management Sciences, volume 8, issue 2, pp. 63-74, 2020.

[4] S. Nabi and V. Mishra, "Analysis and impact of COVID-19 on economy and organization", International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), pp. 219-224, 2021.

[5] A. Hasan, E. Putri, H. Susanto and N. Nuraini, "Data-driven modeling and forecasting of COVID-19 outbreak for public policy making", ISA Transactions, 2021.

[6] S. Moein, N. Nickaeen, A. Roointan et al., "Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan", Scientific Reports, volume 11, issue 1, pp. 1-9, 2021.

[7] E. Piccolomini and F. Zama, "Monitoring Italian COVID-19 spread by a forced SEIRD model", Plose One, volume 15, issue 8, 2020.

[8] A. Jajarmi, S. Arshad and D. Baleanu, "A new fractional modeling and control strategy for the outbreak of dengue fever", Physica A: Statistical Mechanics and Its Applications, 535(1), 2019.

[9] C. Viboud, K. Sun, R. Gaffey, M. Alelli, L. Fumanelli, S. Merler and A. Vespignani, "The RAPIDD ebola forecasting challenge: synthesis and lessons learnt", Epidemics, volume 22, pp. 13-21, 2018.

[10] D. Farrow, L. Brooks, S. Hyun, R. Tibshirani, D. Burke and R. Rosenfeld, "A human judgment apporach to epidemiological forecasting", PLOS Computational Biology, volume 13, issue 3, 2017.

[11] Y. Qi, C. Du, T. Liu, X. Zhao and C. Dong, "Experts' conservative judgment and containment of COVID-19 in early outbreak", Journal of Chinese Governance, volume 5, issue 2, pp. 140-159, 2020.

[12] M. Maleki, M. Mahmoudi, D. Wraith and K. Pho, "Time Series modeling to forecast the confirmed and recovered cases of COVID-19", Travel Medicine and Infectious Disease, volume 37, 2020.

[13] R. Maaliw III and M. Ballera, "Classification of learning styles in virtual learning environment using J48 decision tree", 14th International Conference on Cognition and Exploratory Learning in Digital Age, pp. 146-156, 2017.

[14] R. Maaliw, "Early prediction of electronics engineering licensure examination performance using random forest", IEEE World Artificial Intelligence and Internet of Things Congress (AIIOT), pp. 41-47, 2021.

[15] A. Ashofteh and J. Bravo, "Life table forecasting in COVID-19 times: an ensemble learning approach, 16th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-6, 2021.

[16] N. Talkhi, N. Fatemi, Z. Ataei and M. Nooghabi, "Modeling and forecasting number of confirmed and death caused by COVID-19 in Iran: a comparison of time series forecasting methods", Biomedical Signal Processing and Control, volume 66, 2021.

[17] R. Maaliw, Z. Mabunga and F. Villa, "Time series forecasting of COVID-19 cases using stacked long short-term memory networks", IEEE International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021.

[18] Y. Geng and S. Li, "A LSTM based campus network traffic prediction system", IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), pp. 327-330, 2019.

[19] D. Benvenuto, M. Giovanetti, L. Vassalo, S. Angeletti and M. Ciccozzi, "Application of the ARIMA model on COVID-19 epidemic dataset", Data in Brief, volume 29, 2020.

[20] L. Felizardo, R. Oliveira, E. Hernandez and F. Cozman, "Comparative study of bitcoin price prediction using wavenets, recurrent neural networks and other machine learning methods", IEEE 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), pp. 1-6, 2019.

[21] A. Saeed et al., "Hybrid bidirectional LSTM model for short-term wind speed interval prediction", IEEE Access, volume 8, pp. 182283-182294, 2020.

[22] T. Xia, R. Shu, X. Shen and T. Menzies, "Sequential model optimization for software effort estimation", IEEE Transactions on Software Engineering, pp. 1-5, 2020.

[23] R. Adhikari and R. Agrawal, "Performance evaluation of weights selection schemes for linear combination of multiple forecasts", Artificial Intelligence, volume 42, pp. 529–548, 2015.

[24] S. Lahmiri, R. Saade, D. Morin and F. Nebebe, "An artificial neural networks based on ensemble system to forecast bitcoin daily trading volume", IEEE 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), pp. 1-4, 2020.

[25] B. Alaskar et al., "Next-day electricity demand forecast: a new ensemble recommendation system using peak and valley," IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1-5, 2021.