Surrogate Parameters Optimization for Data and Model Fusion of COVID-19 Time-series Data

Ogundare Timilehin
Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa.

Terence L. van Zyl
Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa.

Abstract—Our research focuses on developing a computational framework to simulate the transmission dynamics of COVID-19 pandemic. We examine the development of a system named ADRIANA for the simulation using South Africa as a case study. The design of ADRIANA system interconnects three sub-models to establish a computational technique to advise policy regarding lockdown measures to reduce the transmission pattern of COVID-19 in South Africa. Additionally, the output of the ADRIANA can be used by healthcare administration to predict peak demand time for resources needed to treat infected individuals. ABM is suited for our research experiment, but to prevent the computational constraints of using ABM-based framework for this research, we develop an SEIR compartmental model, a discrete event simulator, and an optimized surrogate model to form a system named ADRIANA. We also ensure that the surrogate’s findings are accurate enough to provide optimal solutions. We use the Genetic Algorithm (GA) for the optimization by estimating the optimal hyperparameter configuration for the surrogate. We concluded this study by discussing the solutions presented by the ADRIANA system, which aligns with the primary goal of our study to present an optimal guide to lockdown policy by the government and resource management by the hospital administrators.

Index Terms—Optimization, genetic algorithm, data science, recurrent neural network, time-series, stationary data, COVID-19, forecast, deep learning.

I. INTRODUCTION

Towards the end of 2019, the world experienced the outbreak of coronavirus-2019 (COVID-19) pandemic which has had significant impact on everyday life [1]. COVID-19 disease was identified as a novel disease that emanated from Wuhan, China [2]. The World Health Organization (WHO) made a declaration on the 11th of March 2020, that COVID-19 disease was a pandemic [3], as it was widespread globally.

Agent-Based Modeling (ABM) is an effective tool for simulating the discrete events of a pandemic [4]. However, there are limitations to using an ABM approach for event simulation. These limitations include computational costs with resulting factors such as long simulation run times, high-level complexity of parameters, and complex model evaluation [5].

Because of the constraints of ABMs mentioned above, we developed a surrogate model to substitute the ABM. We could minimize the burden of long simulation time and computational complexity by using the surrogate.

We use the Genetic Algorithm (GA) to parameterize the surrogate model [6]. The GA offers problem-solving strategy by allowing us to determine the optimal hyperparameters for the surrogate. A study by Thengade et al. [7] confirms that the GA is a universal optimizer, which has been applied in various scientific disciplines and has been proven to perform well in the optimization of different machine learning models such as Radial Basis Function Networks (RBFNs) [8], Artificial Neural Networks (ANNs) [9], Kriging Model [10], [11] and Polynomial Regression [12].

We develop a system named ADRIANA. The system alerts policymakers and hospital managements about the future transmission trend of COVID-19. The system will guide policymakers in determining when to increase or decrease the length of lock-downs, as well as recommend the duration of high resource demand for hospital management. ADRIANA is composed from three simulations:

- The SEIR compartmental model;
- The COVID-19 discrete event simulator; and
- An optimized surrogate model.

The above-mentioned sub-models are discussed further below.

In our study, we train and fit the surrogate with a historical COVID-19 dataset for South Africa. We implemented three deep neural network frameworks. They are Long Short-Term Memory (LSTM), Multi-Layer Perception (MLP), and Gated Recurrent Network (GRU), using LSTM as the benchmark model. We also implemented other machine learning architectures such as XGBoost Regression (XGBR), Random Forest Regressor (RFR), Support Vector Regressor (SVR), Linear Regressing (LR), and Decision Tree Regressor (DTR). We evaluate each of the models with prediction-performance metrics, which are model Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared ($R^2$).

II. LITERATURE REVIEW

Since COVID-19 started, researchers across the globe have developed statistical and machine learning models to forecast the transmission rate of COVID-19. Despite advances in research, certain challenges persist, such as unstructured COVID-19 data and the use of statistical models to predict epidemic cases [13], [14]. Therefore, researchers are motivated to develop deep learning model to address these challenges and to predict the future trend of COVID-19 disease [15]–[17].

To forecast a pandemic, the use of statistical models such as the Auto-Regressive (AR), Moving Average (MA) and,
Integrated-ARMA (ARIMA) have been proven to be dependent on invalid assumptions [18], [19]. In most cases, the statistical models are not able to fit the data properly, and evaluating their results proved difficult [18]–[20]. Previous studies forecasting COVID-19 were limited by assumptions made during model development [15], [21]. The COVID-19 time-series data has non-stationary features, and using only epidemiological (such as reproduction number \( R_0 \)) and statistical tools for modeling them can lead to incorrect predictions [15], [16], [22]. Deep learning is less reliant on these assumptions, therefore it is worth exploring whether it is better at detecting patterns from COVID-19 dataset.

Additionally, previous prediction models did not consider patient recovery rate and the impact of government policy as factors in their models [22], this is captured in our simulation experiment. Our data exploration revealed that within the early stages of the pandemic, algorithms tend to allocate out-of-range hyperparameter values and neglect historical information.

Some of the existing models that forecasts the transmission trend of COVID-19 are CHIME [23], Adjutorium [24], Rationizer [13], and Epidemic Calculator [25]. The researchers that developed these models were able to build a surrogate, train the surrogate, and make forecasts about the future transmission of COVID-19 in their respective regions. Some other synonymous study with specific focus are pandemic surge forecast by Andrew Crane-Droesch et al. [23], scientific approach to pandemic control by Mihaela Van Der Schaar et al. [24], simulation of pandemic transmission rate by Douglas et al. [13], hospital capacity planning for epidemic by J. Ashleigh et al. et al. [14], and epidemic forecasting in Canada by Kumar et al. [26].

III. METHODOLOGY AND MODELS

The subsections that follow highlight the step-by-step approaches and workflow that we explored in our study.

A. Data Collection and Pre-processing

The datasets used for our study were extracted from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [27]. The dataset is presented in a time-series format which contains the daily cumulative statistics of COVID-19 confirmed cases, recovered cases, and death cases (CRD) for 191 countries across the globe starting from January 22nd, 2020 up till day\( (X_t - 1) \), where \( X_t \) is the current day. We then derived the daily instance for \( n \) by deducting the cumulative value of \( n \) from the cumulative value of day\( (n - 1) \) from the cumulative value of day\( (n) \).

To build our surrogate model with the datasets, we test for stationarity. The stationarity of the time-series datasets transforms it to a constant mean, constant variance, and constant probability distribution at any chosen point \( t_n \) in the series. In addition, because unwanted patterns in the dataset are removed, the stationarity of the CRD dataset offers a higher level of certainty in the simulation result.

To verify for the stationarity of the CRD dataset, we plotted each of the CRD datasets on an ACF plot. Figure 1 shows the ACF-plot for the confirmed cases dataset and we observed a similar trend of non-stationarity in the recovered and death datasets [28].

In addition, we used the Augmented Dickey-Fuller (ADF) statistical test to evaluate if the unit root exists in the CRD dataset. By default, ADF-test presents the null hypothesis that our dataset has a unit root [29], showing that our CRD dataset is not stationary.

The p-value estimated by the ADF-test for confirmed, recovered and death cases are 0.12, 0.98, and 0.96 respectively, which indicates that we can accept the null hypothesis that the CRD datasets are not stationary. Therefore, we apply the differencing technique to transform the CRD datasets into a stationary series.

By analyzing the PACF-plot in Figure 2, we observe that differencing of order 1 (i.e. \( d = 1 \)) is sufficient to transform each of the CRD datasets to a stationary series.

Additionally, we normalize the CRD datasets before training and fitting the surrogate by re-scaling the data to a range between 0 and 1. We accomplished this to enhance the surrogate’s learning rate and prediction accuracy during model training.
B. SEIR Compartmental Model

The Susceptible, Exposed, Infectious, and Recovered (SEIR) compartmental model is a mathematical model that was developed by mathematics-of-epidemiology experts McKendrick and Kermack for grouping homogenous populations into different SEIR compartments during pandemic outbreak \[30\]. The SEIR model is used to provide computational insight into the dynamics that are involved in the spread of pandemic disease based on observed local factors such as disease prevalence rate and disease transmission pattern.

In this study, we develop the SEIR model using Euler’s method of mathematical differential equation \[31\]. We simulate the number of people that recovered or died from the disease based on factors such as individual risk score, age, disease historical record, etc. Equation \[1\] shows the mathematical simulation of grouping population into the susceptible, exposed, infectious and recovered compartments respectively over a defined period \(t\) during a pandemic.

During model development, we considered internal factors contributing to the SEIR model architecture. Factors include initial disease prevalence rate, population size, disease transmission rate, incubation period, recovery rate, duration of infection, and reproduction number. Generally, the SEIR model offers insight into categorizing South Africa populations into epidemic compartments, as illustrated in Figure 3. The proposed SEIR model takes the form:

\[
\begin{align*}
S'(t) &= -\frac{\beta SI}{N} & \text{Susceptible compartment} \\
E'(t) &= \frac{\beta SI}{N} - \sigma E & \text{Exposed compartment} \\
I'(t) &= \sigma E - \gamma I & \text{Infectious compartment} \\
R'(t) &= \gamma I & \text{Recovered compartment}
\end{align*}
\]

where \(N = S(t) + E(t) + I(t) + R(t)\) is the total population to be grouped into compartments. The internal parameters of the SEIR model includes infection rate \(\beta\), incubation rate \(\sigma\), incubation period \(1/\sigma\), recovery rate \(\gamma\), and reproduction number \(R_0 = \frac{\beta}{\gamma}\).

C. Discrete Event Simulation

As part of our research method and architecture for the ADRIANA system, we developed a Discrete Event Simulation (DES) using SimPy \[32\] to demonstrate how individuals in the exposed compartment of the SEIR model move to the infectious stage and their response to treatment which determines if an infected individual recovered or died during hospital treatment. We simulate patient activity based on defined model variables and individual personality traits, such as time of patient arrival, disease severity level, duration of treatment, time of exit from the hospital, age group, etc. This is comparable with range of population parameters that were enumerated by the World Health Organization (WHO) \[33\] and the National Institute of Communicable Diseases (NICD) \[34\] in South Africa.

The architecture of the discrete event simulation presupposes that the simulated system changes at distinct points in simulation time. The simulation system is asynchronous, which implies that it does not operate on a perpetual clock but on random simulation time intervals \[32\]. Furthermore, we use the DES to simulate the social processes that infected individuals undergo during a pandemic, with a starting population sample \(N\) estimated as the number of infectious population in the SEIR model.

D. Surrogate Model and Evaluation

Recurrent Neural Network (RNN) performed well in modeling sequential data \[35\]. Generally, RNN gives an accurate prediction of time-series sequential data which may be difficult to model with other computational algorithms. Our study implemented three neural network models which are LSTM, GRU, and multi-layer perceptron (MLP).

1) Long Short-Term Memory (LSTM): The LSTM is an archetype of the RNN that can learn long-term dependencies and remember past information to predict into an extended period \(t\) in the future. LSTM is widely used in modeling because of its reliable results in solving varieties of scientific problems. LSTM has three major gates in its architecture. They are input gate, forget gate, and output gate. In our study, we implemented one input layer, three hidden layers, and one output layer for LSTM. We used Distributed Evolutionary Algorithm in Python (DEAP) \[36\] as the optimization engine, by minimizing the Root Mean Squared Error (RMSE) as the loss function in our optimization. Our optimization proposed the best number of LSTM neurons to be 58, and the best window size to be 9. We used ReLU as activation function and ADAM as internal model optimizer for LSTM. To avoid overfitting the LSTM model, we introduced a dropout layer and early stopping criteria. The LSTM surrogate implementation in our

\[\text{Available at https://simpy.readthedocs.io/en/latest/}\]
\[\text{https://www.afro.who.int/}\]
\[\text{Available at https://www.nicd.ac.za/}\]
\[\text{https://deap.readthedocs.io/ and } \underline{http://github.com/DEAP/deap}\]
study has 50,954 trainable internal parameters. More knowledge on the LSTM framework can be accessed at [34].

2) Gated Recurrent Units (GRU): The Gated Recurrent Units (GRU) has a similar network architecture with LSTM. Contrary to LSTM, GRU has only two gates which are the reset gate and update gate. GRU has been proven to be less complex in its operation and quicker for training and execution [33], [34]. This inspired us to use GRU into our study. The GRU surrogate implementation in our study has 38,734 trainable internal parameters.

3) Multi-layer Perceptron (MLP): In this study, we used the back-propagation technique to train a multi-layer feedforward neural network (FFNN). MLP FFNN has an input layer, at least one hidden layer and an output layer [33]. The COVID-19 CRD datasets are a time-series format of data points on a 24-hours interval. We also model our datasets with the statistical autoregression AR and ARIMA model. This was done to compare the performance of the neural network models to the statistical regression models using the appropriate evaluation metrics. Table II shows the performance rank of the models considered in this study. We trained our surrogates and search for optimized hyperparameters by using 10-fold cross-validation.

E. Model Performance Metrics

The performance evaluation of a surrogate model is determined by how accurately it can relate the input parameters to the output sequence [35]. We explore three performance metrics to evaluate the prediction of the surrogate model. They are Mean Absolute Error (MAE), R-squared ($R^2$), and Root Mean Squared Error (RMSE). MAE is calculated by summing the absolute difference between the actual value and the predicted value of the model.

$R^2$ estimates the degree of similarity of data points to a fitted regression line. It indicates the interdependence level of variables in the same computation environment. [36]. The $R^2$ accuracy is estimated by observing the range of its value between 0 and 1. Generally, $R^2$ value close to 1 suggests a better fit of the model, indicating a better correlation between the actual and predicted values.

Root Mean Square Error (RMSE) measures the standard deviation of model prediction errors by showing the distance between a regression line and data points [37]. It generally describes the level of data concentration around an optimal fit. RMSE is widely used as an evaluation metric in forecasting, climatology, and regression analysis [37].

F. Surrogate Optimization

It is essential to optimize the surrogate models. The optimization will allow us to search, identify, and configure the optimal surrogate hyperparameters. We employed the Genetic Algorithm (GA) as our surrogate optimizer because of its wide spectrum of uses in optimization problems across different scientific modeling fields [38].

1) Hyperparameter Selection: As shown in Table I our study demonstrates the process of obtaining an optimal hyperparameter value(s) for each of the models that we implemented. The cost function uses the GA to identify surrogate

| Model | Hyperparameter | C | R | D |
|-------|----------------|---|---|---|
| Deep  | Neurons        | 58| 58| 58|
| Learning | Window Size  | 9 | 9 | 9 |
| XGBR  | Learning Rate  | 1 | 0.1 | 0.1 |
|       | Max Depth      | 1 | 1 | 1 |
|       | Max features   | 10 | 10 | 10 |
|       | Estimators     | 25 | 25 | 25 |
| SVR   | Epsilon        | 0.6 | 0.6 | 0.4 |
|       | Kernel         | Linear | Linear | Linear |
| DTR   | Max Depth      | 4 | 4 | 4 |
|       | Min Split      | 7 | 7 | 7 |
|       | Min Leaf       | 3 | 3 | 3 |
| LR    | Coefficient($\beta$) | 0.91 | 0.56 | 0.41 |
|       | Intercept($\alpha$) | 11.37 | 13.73 | 13.92 |
| ARIMA | (p,d,q)        | (1,1,0) | (1,1,0) | (1,1,0) |
|       | p-value        | 0.90 | 0.96 | 0.94 |
|       | Min AIC        | 4180 | 4714 | 2708 |
| GA (Optimizer) | Population Size  | 10 | 10 | 10 |
|       | Generation     | 45 | 45 | 45 |
|       | Gene Length    | 10 | 10 | 10 |

| Models | RMSE | $R^2$ | MAE |
|--------|------|-------|------|
| LSTM   | 15.19 | 0.52 | 103.67 |
| GRU    | 4.10  | 0.47 | 71.03 |
| MLP    | 17.22 | 0.67 | 818.67 |
| XGBR   | 19.30 | 0.63 | 901.00 |
| RFR    | 17.40 | 0.60 | 841.30 |
| SVR    | 17.67 | 0.62 | 738.67 |
| LR     | 19.60 | 0.62 | 868.01 |
| DTR    | 37.96 | 0.24 | 1931.20 |
| AR     | 51.39 | 0.23 | 2552.66 |
| ARIMA  | 52.52 | 0.55 | 2686.00 |

Fig. 4. Performance Ranking of Models
hyperparameters by minimizing the RMSE during the search process and maximizing accuracy. The mutation rate and crossover function were defined in the architecture of the GA.

In Table I, the heading labeled as C, R, and D represents the confirmed, recovered and death cases respectively.

Table II shows the evaluation metrics of the models considered in our study. Likewise, Figure 4 illustrates the performance rank of the models explored in this study, with a rating of 1 to 10, with 10 being the best performing rank.

The 400-days projection by the benchmark surrogate (LSTM) showing the transmission trend for confirmed, recovered, and death cases of COVID-19 in South Africa from April 24th 2021 till May 22nd 2022 are plotted in Figure 5, Figure 6, and Figure 7 respectively.

IV. RESULTS AND DISCUSSION

Our research is a univariate time series analysis of the COVID-19 datasets of confirmed, recovered, and death cases in South Africa. We ensured that the time-series data are stationary for surrogate simulation as described in Section III. We used 70% of the CRD dataset for training the surrogate, while 30% was used for testing. The results from our research are discussed further in the sub-sections below.

Relating to the third wave of the COVID-19 in South
Africa, Figure 5 shows the transmission dynamics of the confirmed cases, Figure 6 shows the trend of disease recovery by infected individuals, and Figure 7 represents the predicted death pattern.

We discovered that LSTM can remember longer data sequences than GRU, although, GRU has less hyperparameter complexity and is easier to analyze. The relationship between the SEIR model, DES, and surrogate is that, the surrogate offers future predictions based on real-world data, and we use the SEIR results to group a random population sample into epidemiological compartments. Likewise, the DES retrieves the number infectious individuals in the SEIR model as the starting population for its event simulation.

We predict the future trend of COVID-19 CRD cases up to a specified duration $n = 400$. This study does not consider the impact of hospital resource distribution, economic effect, or the administration of personal protective equipment (PPE) required by healthcare professionals.

Our work is significant because it addresses the development of a system with an optimal function capable of making an approximate daily forecast of the pandemic instance in South Africa. In addition, the prediction covers the projected third wave of COVID-19 in South Africa.

V. CONCLUSION

By comparing various machine learning models and statistical models for predicting future transmission trend of COVID-19 in South Africa, we observed that the deep learning models outperform the conventional and statistical models as illustrated in Figure 4. We also demonstrated that our optimized surrogate can forecast the transmission dynamics of COVID-19 or other pandemic disease in the future with the same data presentation and features.

The data pattern from the COVID-19 CRD datasets indicates that significant disease spread can be controlled if South African government authorities adopt timely and functional preventive action at the emergence of the disease.

Following model simulation and data fitting, our surrogate forecasted COVID-19 in South Africa for 400-days (beginning April 24th, 2021). The projection is based on the disease's historical transmission pattern during the last 458-days (January 22nd, 2020 - April 23rd, 2021). The long-term forecast captures the anticipated third-wave occurrence of the COVID-19 pandemic in South Africa.

This study has demonstrated the impact of surrogate models for forecasting pandemic disease for range $[0, n]$. The upper range $n$ is advised to be less or equal to 500, i.e. $n \leq 500$ to avoid skew forecast or early convergence of results. This demonstrates that ADRIANA’s result can guide decisions for adequate resources planning during a pandemic.

For future research, the ADRIANA system can be improved to respond dynamically to different pandemic cases while taking into consideration external factors (such as geolocation, the effect of weather, government policy type, and vaccination distribution to the population), simulate hospital resources for treatment, simulate the allocation of personal protective equipment (PPE) and healthcare personnel to hospitals to allow for more model robustness and flexibility.

For our results to be reproducible, the source file for engineering the ADRIANA system can be accessed through this link.

ACKNOWLEDGMENT

The authors would like to appreciate the Computer Science department at the University of Witwatersrand, and Institute of Intelligent Systems at the University of Johannesburg, South Africa for the interest in providing a computational solution to the COVID-19 outbreak in South Africa. This research is supported by the Nedbank Research Chair.
