International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Enhanced Machine Learning Model for Prediction of COVID-19 Cases in Iraq

Zakarya A Mohamed Zaki[1], Aisha Hassan Abdalla[2]

[1, 2]Department of ECE, Fac. of Engineering, International Islamic Univ. Malaysia (IIUM), Jalan Gombak, 53100 Kuala Lumpur, Malaysia
[1]zaltalib92[at]yahoo.com,
[2]aisha[at]iium.edu.my

Abstract: ***The SARS-CoV-2 virus is responsible for the emergence of the highly contagious illness known as COVID-19. The disease has been classified as a global pandemic, impacting millions of people throughout the globe. It has created a change in the research community's orientations for identification, analysis, and control via the application of different statistical and predictive modelling methodologies. These numerical models are examples of decision-making techniques that depend significantly on data mining and machine learning to create predictions based on historical data. In order to make smart judgments and create strong strategies, policymakers and medical authorities need reliable forecasting techniques. These studies are carried out on a variety of small scale datasets including a few hundreds to thousands of records. This study uses a large dataset consisting of COVID-19 instances recorded on a daily basis in Iraq, together with socio-demographic and health related attributes for the region. The primary goal of the research is to do daily forecasting of Covid-19 instances using time series forecasting. The predictive modeling for daily COVID-19 infection cases involved several neural network architectures, including artificial neural networks, convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and Hybrid CNN-LSTM model. Prior to the modeling, appropriate procedures were used to prepare the data and identify any seasonality, residuals, and trends. The contribution of this work lies in the development of an enhanced forecasting model for COVID-19 infection cases. It utilizes a combination of different neural network models to create an effective forecasting tool. The proposed enhanced hybrid model built using a Convolutional Neural Network and a Long Short-Term Memory network (EH-CNN-LSTM). The model is trained and tested on various subsets of the dataset. It is discovered that the higher the amount of training data, the better the predicted performance. Compared to other models, the proposed EH-CNN-LSTM performs better. Mean Absolute Percentage Error (MAPE), Mean Squared Logarithmic Error (MSLE), and Root Mean Squared Logarithmic Error (RMSLE) are used to evaluate the predictive performance. EH-CNN-LSTM, which was trained on 80% of the data and evaluated on 20% of the data, achieved a MAPE of 5.28, MSLE of 0.00, and RMSLE of 0.02.***

Keywords: COVID-19, Prediction, Forecasting, and Machine Learning

1. Introduction

Throughout the past years, machine learning has been used excessively in several problems including ecommerce (Rath, 2022), sports (Richter et al., 2021), and healthcare (Qayyum et al., 2020). Time series forecasting is also one of the main areas for machine learning algorithms (Ahmed et al., 2010), because efficient forecasting may lead to better trading returns and enhance utilization of healthcare infrastructure. Many researchers now a days are focusing on hybrid approaches such as CNN-LSTM as superior alternative for time series forecasting.

The new coronavirus emerged in late 2019 is the consequence of exposure with the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) (Samuel Lalmuanawma, 2020). It has expanded internationally from late 2019, resulting in a protracted epidemic. COVID-19 dissemination is based on inter-individual physical proximity and breathing particle transfer. Corona virus are a broad virus group that have been linked to illnesses spanning from cold or flu to far more serious illnesses. There have been two more outbreaks caused by coronavirus and the most recent virus discovered in Wuhan, China, is known as SARS-COV-2, and it causes Covid-19 (Samuel Lalmuanawma, 2020).

The first case of an unidentified pneumonia was report in Wuhan, China on December 31, 2020. Since then, the frequency of corona virus kept increasing, along with the mortality rate. It took only thirty days to spread to the whole country (Samuel Lalmuanawma, 2020). Because Covid-19 spreads between person to person, artificial intelligence assisted electronic gadgets (Bhaskar et al., 2020) can perform a critical role in stopping the virus's transmission. As the function of healthcare epidemiologists has grown, so has the prevalence of digital medical records. The growing accessibility of digital clinical information gives a significant potential in medicine including both research and pragmatic implementation to enhance healthcare.

Most technologically advanced deep learning modelling techniques are based on ANNs, particularly CNNs, while they can incorporate probabilistic algorithms or latent constructs organized tier in deep generative designs like the endpoints in deep learning and deep Boltzmann automated systems (Salakhutdinov & Larochelle, 2010). Deep learning techniques can be used to solve unsupervised training problems (Károly et al., 2018). This is a significant advantage since unidentified input is much frequent than classified data. Deep belief networks (Hinton, 2009) are indeed an illustration of a deep architecture that may be learned unsupervised. If breadth of a deep neural network containing ReLU (Agarap, 2018) stimulation is higher than the incoming size, the system may estimate any Lebesgue integrable value (Burkill, 2004); if the dimension is less than or equal to the incoming dimension, the structure is not a universal probabilistic model.

Machine learning is where the probabilistic understanding comes from. It includes the optimization ideas of learning and evaluation, which are linked to matching and generalization, accordingly. The probability approach views the activating nonlinearity as a cdf. Dropout was introduced as a regularizer in neural networks because of the deterministic understanding. The deterministic approach was developed by scholars such as Hopfield, Widrow, and Narendra and promoted in questionnaires such as Bishop's.

2. Related Works

There is a huge number of designs and developed models that have been design to predict different Covid-19 important attributes such as new cases, death status, recovery status and more using different kind of machine learning alorithms such as ANN, CNN, LSTM and etc. Therefore, this section will present previous study and discuss the motivation and drawbacks.

(Niazkar, 2020) made their predictions on the number of cases of Covid-19 in China, Japan, Singapore, Iran, Italy, South Africa, and the United States of America by using ANN-based models. The reports of the World Health Organization (WHO) and the National Health Commission of the People's Republic of China were combed through to gather information that was obtained on verified instances of Covid-19 and fatalities caused by it (NHC). When there was a contradiction between the two sets of data pertaining to China, it was decided that the data from the NHC should be used. The China dataset collection included both confirmed and clinically diagnosed instances of the disease. After performing statistical analysis in Excel, the data were then partitioned into the train and test sets. Several different ANN models were trained, and RMSE, MAE, and R-squared were used to evaluate the models' performance. The results of the different models for virtually all of the nations were not very encouraging, which is probably due to the fact that ANN often produces the most accurate predictions when it is used for forecasting within the same time range as the training set. Inevitably, the data from the test set will go beyond the parameters defined by the training data. To be more specific, ANN models are effective in terms of prediction when they are employed for a short time forecast into the future such as few days. This is because of the way that artificial neural networks work. When used to predictions in the long future, it does not provide very useful results. It is important to note that short-term projections are not very helpful in the Covid-19 situations and would not have a good impact on the management of resources or the treatment of patients. This is something that should be kept in mind.

To make his predictions on the number of fatalities caused by Covid-19, (Jarndal et al., 2020) analyzed data from the WHO and employed Gaussian Process Regression (GPR) (Schulz et al., 2018) and Artificial Neural Networks (ANN) (Yegnanarayana, 2009). The population that was chosen was comprised of persons who were older than 65 years old, as well as smokers and diabetics. The information was obtained from the WHO. The data was updated each day. The selection of patients from each of these three distinct groups was followed by the selection of patients from each of these distinct areas. Europe, the Middle East, and North America were chosen as the areas to focus on. The United States of America, Canada, and Mexico were chosen to represent North America in the selection process. France, Italy, Spain, the United Kingdom, and Germany were chosen to represent Europe, and Qatar, Saudi Arabia, Kuwait, Oman, and the United Arab Emirates were chosen to represent the Middle East. Both a training set and a test set were generated from the dataset. The training phase used 84% of the dataset, while the testing phase employed the remaining portion. Statistical examination revealed that the data had a significant lack of linearity. In addition, the preprocessing method known as normalization was used in order to scale the data between-1 and 1. Due to the high degree of unpredictability in the data, the ANN was unable to reproduce the precise distribution of the data despite the fact that it is a non-linear model. The simulation of the distribution of American regions, on the other hand, was vastly enhanced because to the practically linear nature of the relationships between the input and output variables. The data for the GPR were divided into thirds and sevenths. The GPR model is a probabilistic one that is based on the parametric kernel. Once again, the data were scaled to a normal range between-1 and 1, and then afterwards they were rescaled to the usual range for comparison. Because of its innate capabilities and the assumption that it makes about the Gaussian probability distribution, GPR produced very positive outcomes. In the instance of GPR, the assessment measures showed a considerable improvement when compared to those of ANN.

(Kayode Oshinubi, 2021) investigated the use of deep learning and spectral analysis on epidemiological time series data in research that he carried out. It has resulted in the researchers developing a newfound interest as a result. As the SARS-COV2 is still in the process of changing into the omicron version, which is extremely infectious and for which governments and other stakeholders are accountable in combatting this catastrophe via vaccines and other methods, it is imperative that these measures be taken immediately. This study was carried out by the authors with a new methodology in order to contribute to the cause. They compared Extreme Machine Learning (EML), Multilayer Perceptron (MLP), LSTM, Gated Recurrent Unit (GRU), CNN, and Artificial neural networks (ANN) on time series analysis of the COVID-19 data from the beginning of the pandemic in Turkey, Russia, France, USA, Brazil UK, and India until September 3 of 2021 in order to predict the daily new cases They did this by using a method known as spectrum analysis, which included converting days of time into frequencies in order to study the periodicity and frequency.

(Nahla F. Omran, 2021) conducted research to evaluate the performance of LSTM and GRU in predicting new cases of COVID-19 and mortality in Egypt, Kuwait, and Saudi Arabia. The Novel Corona Virus 2019 dataset was obtained from Kaggle and was given the moniker "The Novel Corona Virus 2019 dataset" for the purposes of the study. This data collection included daily information about the total number of new cases and fatalities around the globe. The dataset was a time series from January 22, 2020, to June 12, 2020. As the training process was supervised the dataset was divided into input and output variables. The split was 80 and 20. The

data was then scaled through MinMaxScaler of the scikit learn library. ReLU activation was used as activation function for both LSTM and GRU. After the training phase was over, the data was re-scaled again by using inverse transform to evaluate the models. Evaluation metrics used for evaluation of the models and to compare them with each other. Three errors were used. Two of three errors are scale dependent i. e., they are used on datasets that are on the same scale and cannot be used on datasets that are on a different scale. These scale dependent errors are MAE and RMSE. The third error used was MAPE. This is a percentage error, and an advantage of this error is that it is not scale dependent. The results obtained from this Scenario was that LSTM achieved better results in cases of confirmed cases in all countries while GRU performed better in case of deaths in Kuwait and Egypt.

(Hafiz Tayyab Rauf, 2021) and colleagues have conducted a study where they compared Recurrent Neural Network (RNN), GRU and LSTM for prediction of severity of COVID-19 Pandemic outbreak in the countries of the Asia pacific including Pakistan, India, Afghanistan, and Bangladesh. The data used in this research work collected from WHO data which is globally available and actively updated. The models were trained on data from January 21, 2020, till June 06, 2020. The dataset was split into training and test sets. ReLU was used as the activation function and batch size set to 1. Adam was used as optimizer and 100 epochs for each model with early stopping criteria. RNN accuracies were 0.9 for Pakistan, 0.94 for India and Bangladesh each and 0.87 for Afghanistan. GRU accuracies calculated were 0.93 for Pakistan, 0.87 for India, 0.9 for Afghanistan and 0.94 for Bangladesh. The accuracies observed for various countries for LSTM ranged from 0.93 to 0.95.

(Batool, 2021) examined the relationship between COVID-19 and the weather using a variety of analytical approaches, including time series analysis, statistical analysis, and deep learning. In this research, the effectiveness of multiple models was evaluated for estimating the number of new cases and fatalities caused by COVID-19 across a variety of locations in Pakistan, in conjunction with the meteorological conditions of those places. This research took into account the temperature and humidity of certain regions, performed calculations, and made predictions for new cases and fatalities in those regions. The author came to the conclusion that forecasting the dynamics of COVID-19 using just meteorological conditions is more successful than using additional parameters like age, smoking, diabetes, and cardiac morbidities. The information that was used in this investigation was obtained from the National Institute of Health, which is situated in Islamabad, the nation's capital city. This dataset included all of the information regarding cumulative cases and deaths, as well as daily fresh cases and deaths. Along with other demographic information, the dataset also included the individual's history of receiving vaccinations. On the NIH website you may see the overall number of recoveries, as well as the daily diagnostic test results. The data was gathered from March 10, 2020, all the way through December 20, 2020, and then it was separated into two categories: train and test. The train set included entries from the beginning of the dataset up through November 15, 2020, whereas the test set included observations beginning on November 15 and continuing until the conclusion of the dataset. ARIMA, linear regression, SVM, MLP, RNN, LSTM, and GRU were the models that were used over the course of the investigation. In comparison to the other models, it was discovered that LSTM was the most effective one even without the incorporation of the meteorological data. It is possible that this is due to its capacity to learn both short-term and long-term patterns. Because the dataset was so tiny, the findings from the other models, which need more extensive data, were not very impressive. When further data on the weather was added during the learning phase, it was discovered that LSTM maintained a large performance advantage over the other models. In addition, the findings led researchers to the conclusion that the weather, namely temperature and humidity, play a significant part in the dynamics of the COVID-19 system. As the temperature rises, there is a corresponding drop in the number of newly diagnosed cases as well as a reduction in the overall number of fatalities.

Most of the literature have only applied statistical learning models in the prediction of COVID-19 dynamics. These models perform well when there is linear relationship between the input and output features and are incapable to learn deep hidden relationship between the variables. In case of COVID-19 complex inter-related features, it looks quite impossible for simple statistical methods to learn the pattern.

In order to address problems involving complex patterns between inputs and target variables or where multiple causative factors exist, advanced methods such as based on deep learning come into play. There is a need for a method that is efficient in learning the hidden correlation between the variables and this research takes hybrid CNN-LSTM into special account as this is a combination of the feature extraction capabilities of CNN and long-term memory of LSTM.

3. Proposed Solution

This research mainly concentrates on the prediction about the number of confirmed instances of COVID-19 cases in Iraq. An Enhanced Hybrid CNN-LSTM was selected to be compared against conventional models. The architecture of the Enhanced CNN-LSTM model contains a CNN base which perform feature learning with convolutional layers which are provided to an LSTM network to perform sequence learning. The proposed model incorporates layer tuning which is performed via grid search and weighted concatenation of layers. The model is provided data after performing preprocessing such data scaling and normalization. As illustrated in Figure 1, data is split up between training set used to build model and test set for testing the performance of the model. The model is then evaluated in comparison with ANN, CNN, and plain LSTM by using defined evaluation metrics.

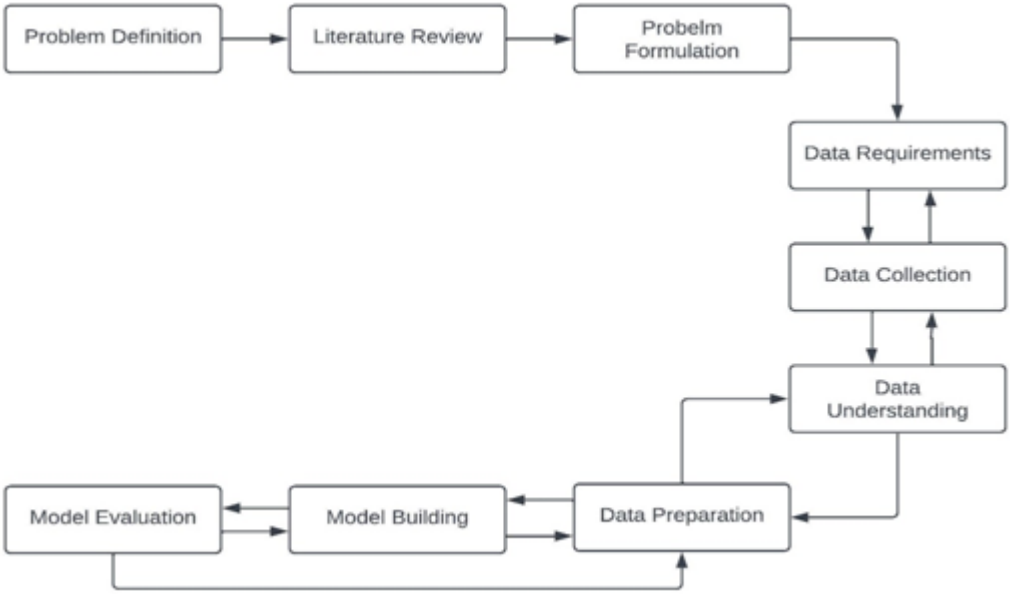

Figure 1: Flow Chart of the modelling building stages

The design approach of the proposed deep learning algorithms in the prediction of COVID-19 new cases are illustrated in Figure 2.

The following are the summarized steps of the design approach:
Step 1: Filtering the dataset and selecting only rows for Iraq.
Step 2: Dropping all the columns except date and new cases.
Step 3: Converting the dataset into a time series.
Step 4: Splitting the dataset into train and test sets.
Step 5: Scaling the output feature before feeding it to the models.
Step 6: Training various models on the training set.
Step 7: Test the model on test set.
Step 8: Evaluating the performance of the models using various metrics.
Step 9: Predicting further into the future.

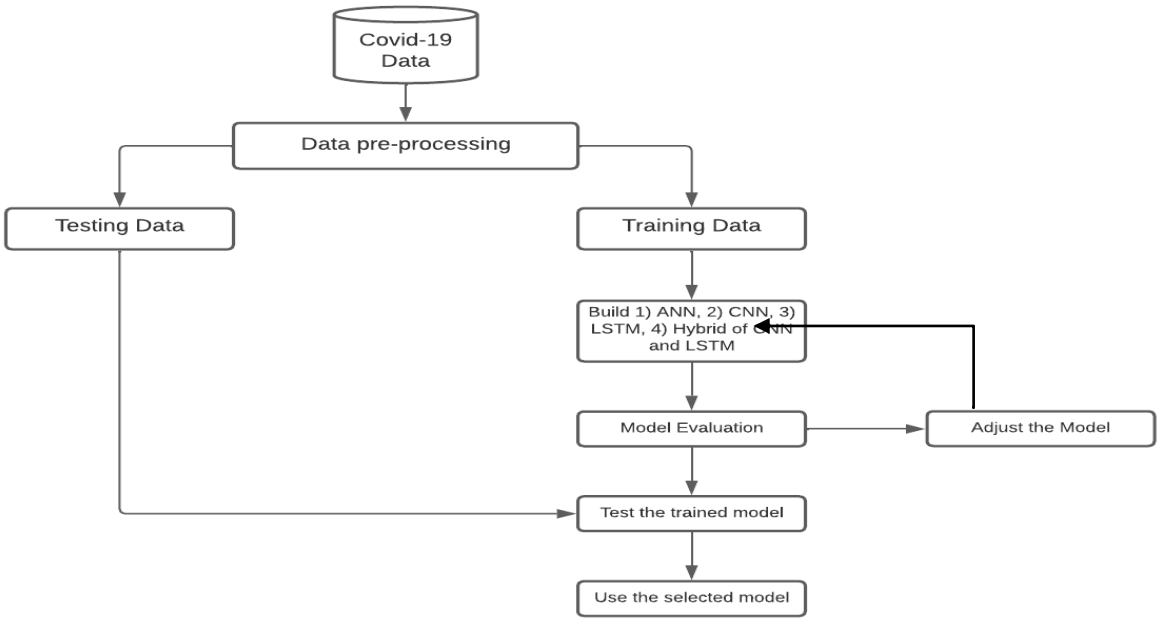

Figure 2: Design approach

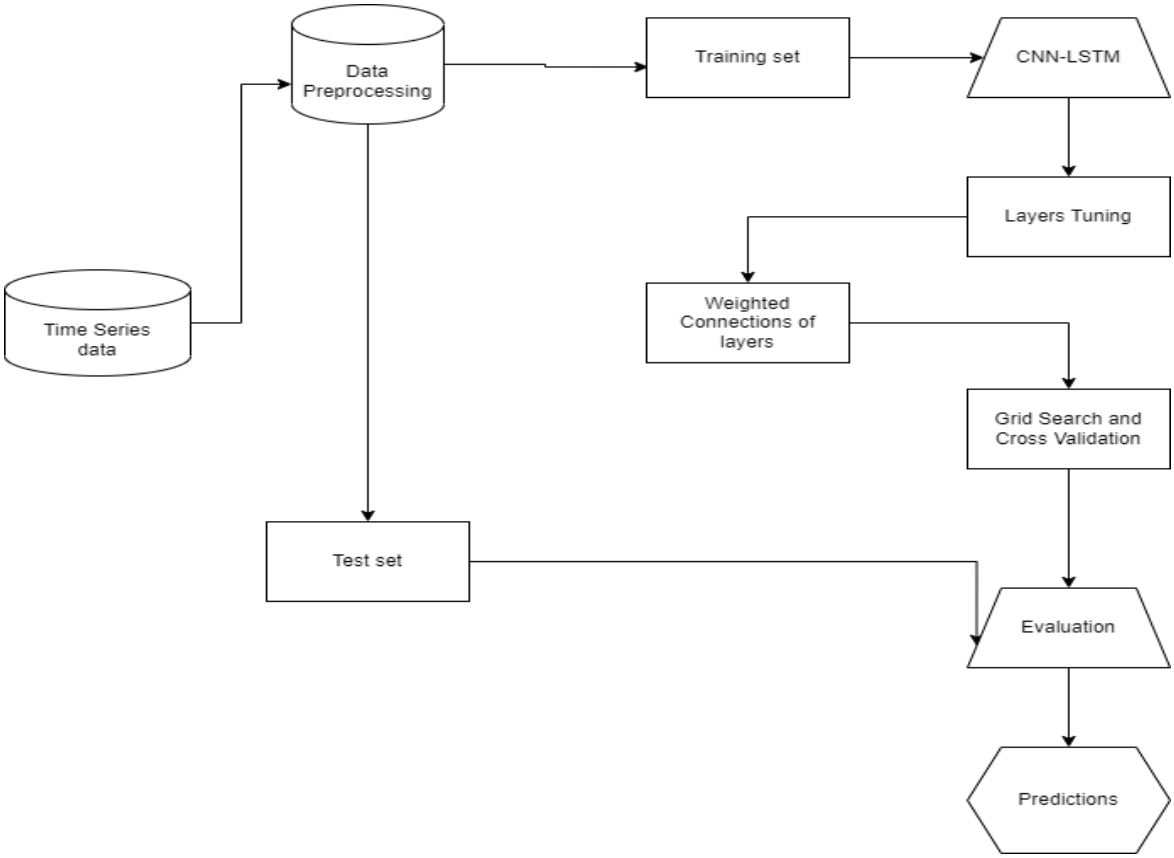

Figure 3: Block diagram of enhanced hybrid CNN-LSTM model

In order to perform the hyperparameter optimization of the proposed CNN-LSTM architecture, we have performed layers tuning. Deep neural networks are large and complex networks with many interlinked parameters and it is therefore difficult to perform the selection of optimal parameters which maximize the performance of a neural network. In layers tuning, the learning rate of each layer is adaptively selected based optimization strategies.

The learning of deep neural network may render some layers of the neural networks to achieve a value of zero which result in setting some inputs of the neural networks to have no effect on the output values. In order to eliminate this situation, as setting the value of an input to zero may result loss of information coming from certain layers or inputs. Layer weighting is used to overcome this scenario which resultantly assigns a small positive value to these inputs/layers so there is no true value assigned to a layer. This scenario always takes a contribution from each input and only minimizes the effect of an input if it is redundant or irrelevant.

As a last step, to perform the optimization of several hyperparameters, cross-validation grid-search is performed. In this strategy, the data is partitioned into cross-validation splits where the value of k (number of splits) is provided based on the dataset size. In order to perform selection of the optimal hyperparameters, a set of discrete values is assigned to each value and iterated via grid search which is checked for its suitability via cross-validation testing.

The contribution of the proposed approach lies in the CNN-LSTM architecture, which is optimized via layers tuning, weighted connections of layers and cross-validation grid search for an enhanced hybrid model. The performance of the model is further validated through train-test Scenarios and superior performance is demonstrated.

4. Method

This section comprises of implementation, testing, and validation steps of the applied deep learning techniques in this research and has a deep dive into all the phases of the workflow as well. The dataset used was filtered through pandas python library. All the rows were filtered out except the rows for Iraq. The features of the dataset were dropped using the same pandas library except for dates and new cases, as these were needed for building, testing and evaluating models.

The data was then converted to matrices using and NumPy arrays and split up between training and test sets. The features were then scaled using MinMaxScaler from scikit learn. Different models under evaluation in this work were trained using various classes and interfaces from scikit learn library. And then evaluated with various performance metrics provided by the same scikit learn library.

The dataset used is from WHO, COVID-19 data, and 2022. This dataset includes daily entries from 24-02-2020 to 01-08-2022. The dataset includes description in the form of various columns about the country, region, and continent. Moreover, it has data for country wise new cases, deaths, cumulative cases, and cumulative deaths. It contains records for 216 countries and 890 days.

The dataset was read using pandas python library through its pandas. read_csv (Pandas, 2022) function and readily converted to time series. As only date, new cases of country Iraq were required to proceed with the research workflow, all the irrelevant data was dropped using the same pandas library. Below Figure 4 has the code snippet for reading and filtering dataset.

```
df = pd.read_csv('Iraq daily dataset.csv', usecols = ['date','new_cases', 'location'], header=0, infer_datetime_format=True,
parse_dates=['date'], index_col=['date'])
df = df.loc[df['location'] == 'Iraq']
```

Figure 4: Code for Reading and Filtering Data

Every machine and deep learning algorithm require data to be processed in some way to have better training and performance in terms of validation and prediction. Most of the machine and deep learning scientists consider this step to be even more crucial than building the model itself. Thus, it has been considered the most important step in the direction of model development.

The first step was converting the dataset into matrices with the help NumPy by using its numpy. array (Numpy, 2022). Figure 5 is how it is done by defining a function that returns two arrays.

```
def convert2matrix(data_arr, look_back):
  X, Y =[], []
  for i in range(len(data_arr)-look_back):
    d=i+look_back
    X.append(data_arr[i:d,0])
    Y.append(data_arr[d,0])
  return np.array(X), np.array(Y)
```

Figure 5: Converting data to matrices

Dataset was then split 80/20 with 80% in the training set and 20% in the test set for one set of models. Next, the train and test features were scaled using MinMaxScaler (Scikit) from scikit learn library as shown in Figure 6.

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0, 1))
trainX = scaler.fit_transform(trainX)
testX = scaler.transform(testX)
```

Figure 6: Feature Scaling

The following tools were used to carry out the research work swiftly with Python as the programming language.

a) Lenovo Ideapad, Windows 10 Pro, Processor Intel (R) Core (TM) i5-6200U, CPU[at]2.30GHz 2.40 GHz, RAM 20GB
b) Anaconda for python distribution.
c) Jupyter Notebooks hosted locally on server.
d) Pandas for data reading and processing.
e) NumPy for computations and matrices operations.
f) Scikit learn for feature scaling.
g) TensorFlow and Keras for deep learning.
h) Matplotlib and Seaborn for visualization.

5. Results

Eighty percent of the data is used to train the model, while the remaining twenty percent is used to evaluate the model's performance. When modeling using the train-test split, each of the four models are trained and evaluated for prediction.

5.1 Artificial Neural Networks

The training and testing loss of artificial neural networks training is depicted in Figure 7 whereas the observed and predicted values are plotted in Figure 8. Table 1 provides the evaluation results for training and testing data for three evaluation metrics.

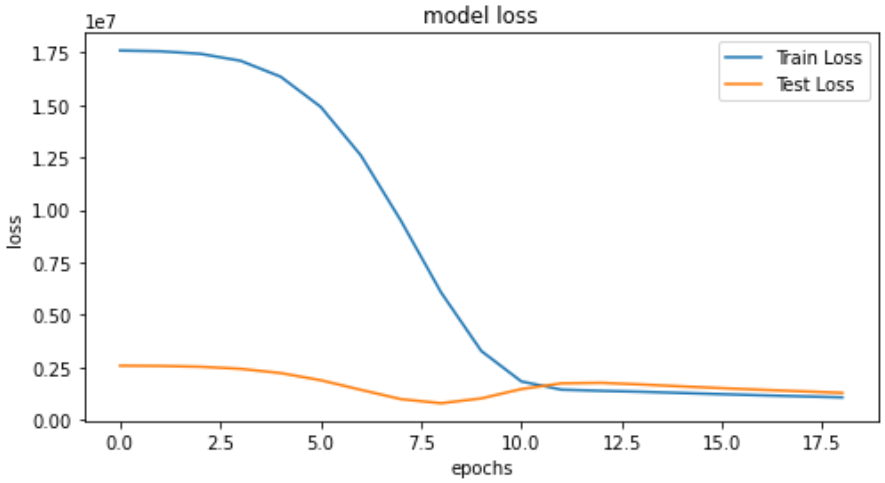

Figure 5: Training and testing loss of the artificial neural networks model

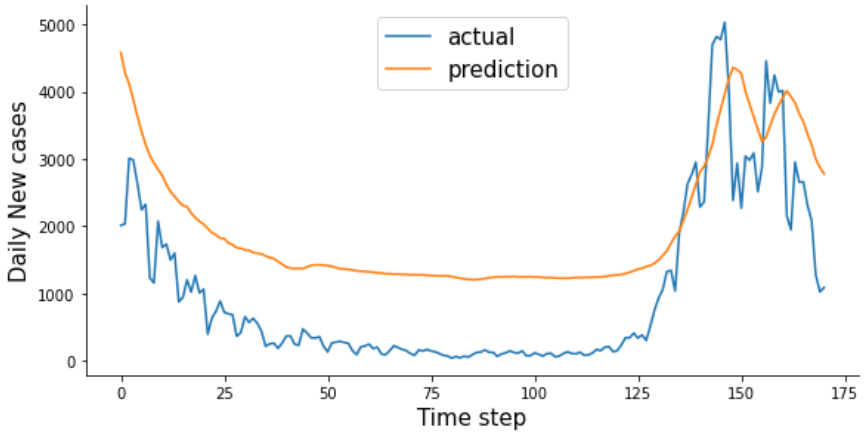

Figure 8: Plot of observed and predicted cases

Table 1: Performance of three evaluation metrics for training and testing sets

Evaluation metric	Training set	Testing set
Mean Absolute Percentage Error	79505.86	21.51
Mean Squared Logarithmic Error	1.85	2.54
Root Mean Squared Logarithmic Error	0.78	1.37

5.2 Long Short-Term Memory (LSTM)

The training and testing loss of long short-term memory networks training is depicted in Figure 9 whereas the observed and predicted values are plotted in Figure 10. Table 2 provides the evaluation results for training and testing data for three evaluation metrics.

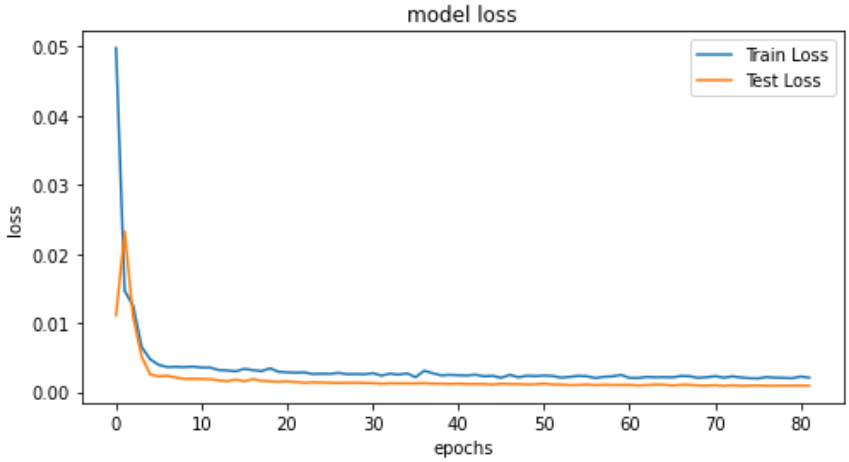

Figure 5: Training and testing loss of the LSTM networks model

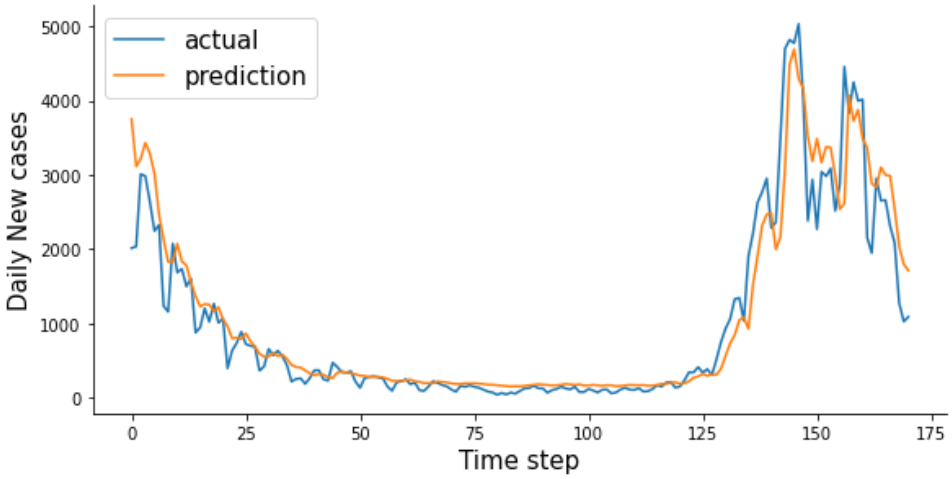

Figure 5: Plot of observed and predicted cases

Table 2: Performance of three evaluation metrics for training and testing sets

Evaluation metric	Training set	Testing set
Mean Absolute Percentage Error	242.57	6.45
Mean Squared Logarithmic Error	0.00	0.00
Root Mean Squared Logarithmic Error	0.02	0.02

5.3 Convolutional Neural Networks

The training and testing loss of convolutional neural networks training is depicted in Figure 11 whereas the observed and predicted values are plotted in Figure 12. Table 3 provides the evaluation results for training and testing data for three evaluation metrics.

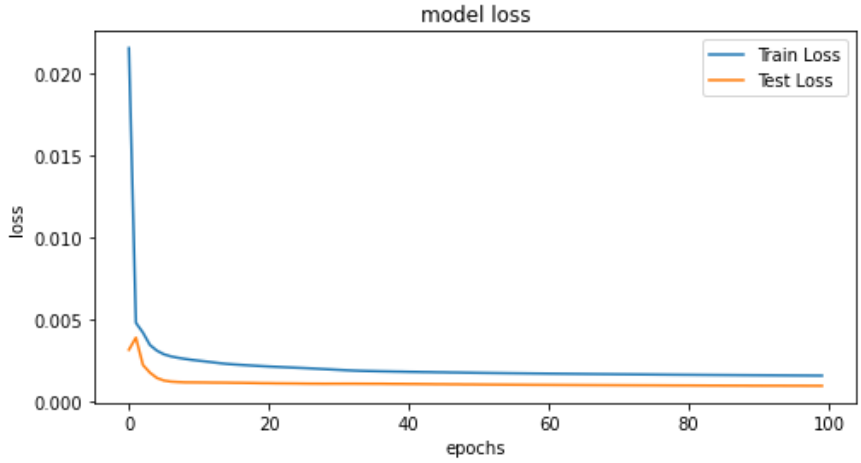

Figure 11: Training and testing loss of the CNN model

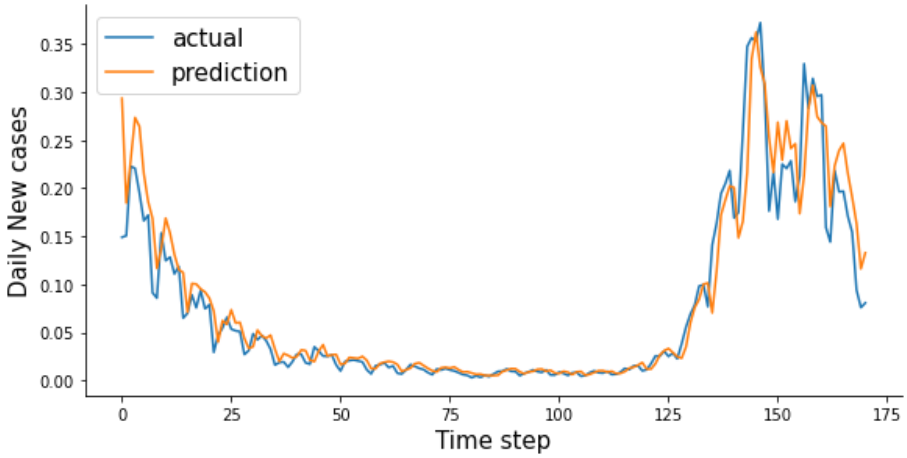

Figure 5: Plot of observed and predicted cases

Table 3: Performance of three evaluation metrics for training and testing sets

Evaluation metric	Training set	Testing set
Mean Absolute Percentage Error	31.53	5.52
Mean Squared Logarithmic Error	0.00	0.00
Root Mean Squared Logarithmic Error	0.02	0.02

5.4 Enhanced Hybrid Model (EH-CNN-LSTM)

The training and testing loss of hybrid model's training is depicted in Figure 13 whereas the observed and predicted values are plotted in Figure 14. Table 4 provides the evaluation results for training and testing data for three evaluation metrics.

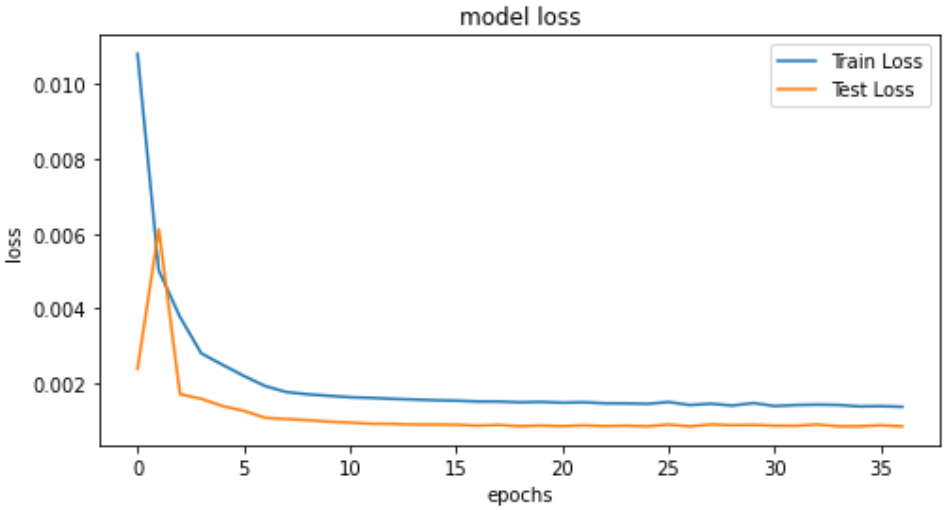

Figure 5: Training and testing loss of the hybrid model

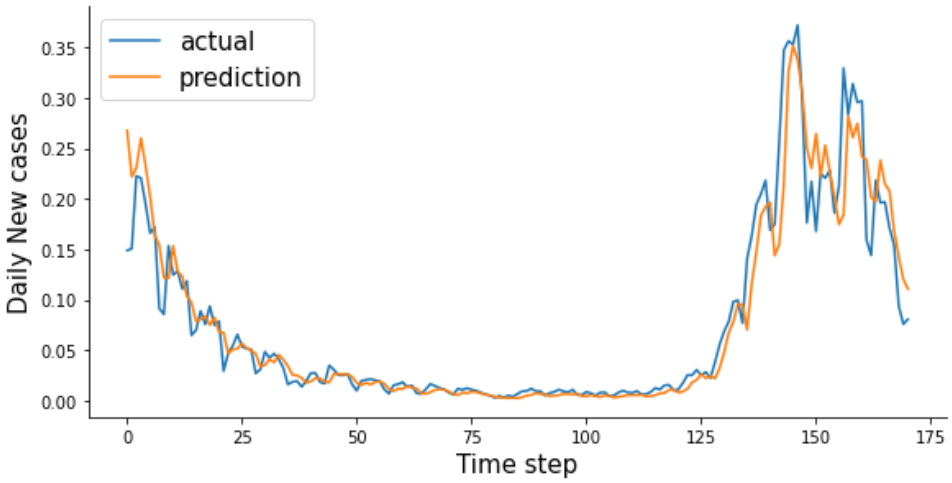

Figure 14: Plot of observed and predicted cases

Table 4: Performance of three evaluation metrics for training and testing sets

Evaluation metric	Training set	Testing set
Mean Absolute Percentage Error	70.06	5.28
Mean Squared Logarithmic Error	0.00	0.00
Root Mean Squared Logarithmic Error	0.03	0.02

5.5 Comparison of all Models

Table 5 summarizes the results and compares all of the potential models. The findings show that the enhanced hybrid model CNN-LSTM performs better than the other candidate models in predicting new instances of COVID-19. Figures 15-18 show a similar pattern of predicted and observed instances for one month's forecast. Figure 19 shows the pattern of predicted and observed instances for the period 24-02-2020 to 01-08-2022 using EH-CNN-LSTM.

Table 5: Comparison of prediction models

Model	Mean Absolute Percentage Error	Mean Squared Logarithmic Error	Root Mean Squared Logarithmic Error
ANN	21.51	2.54	1.37
LSTM	6.45	0.00	0.02
CNN	5.52	0.00	0.02
EH-CNN-LSTM	5.28	0.00	0.02

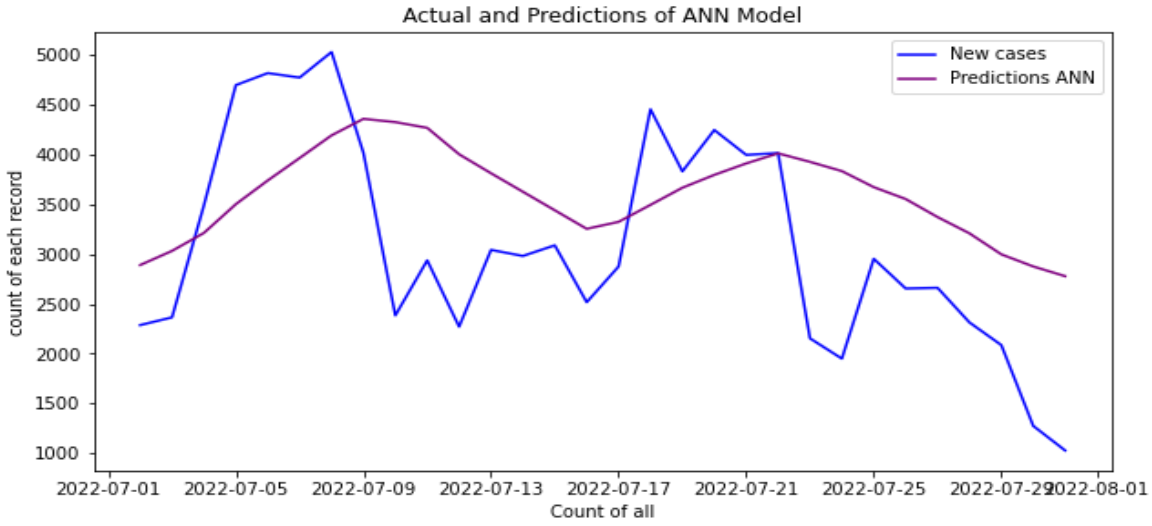

Figure 15: Plot of predictions and observations for ANN model

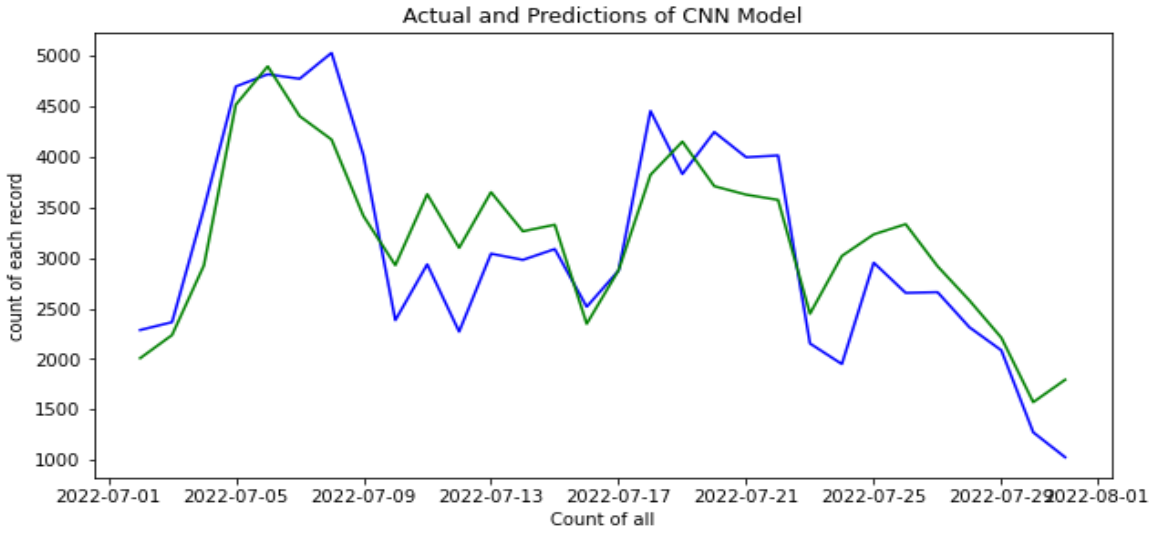

Figure 5: Plot of predictions and observations for CNN model

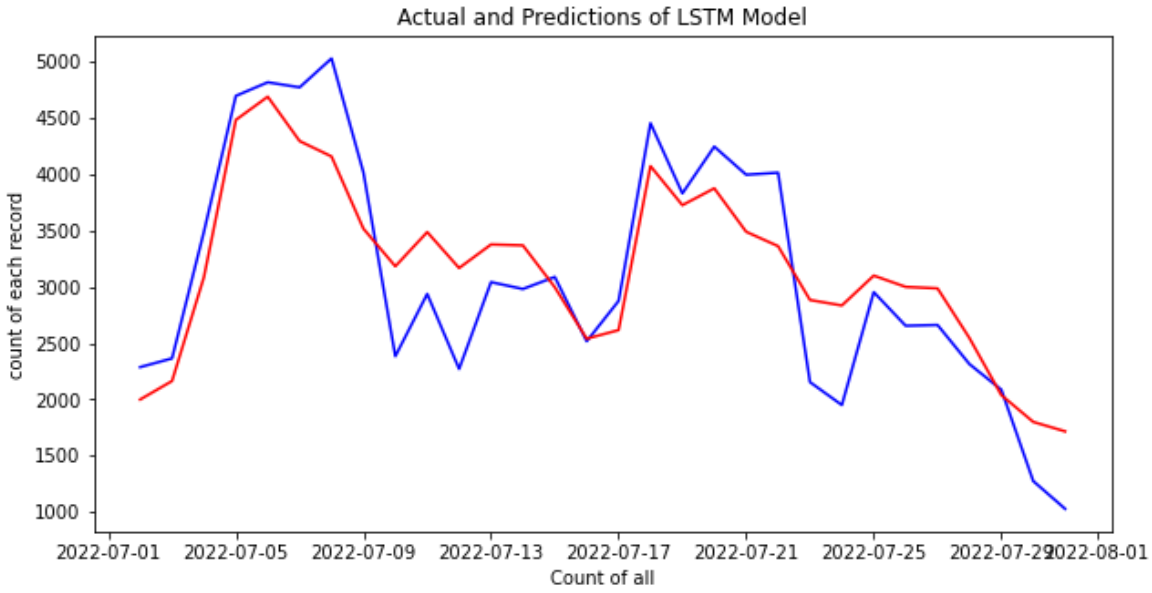

Figure 17: Plot of predictions and observations for LSTM model

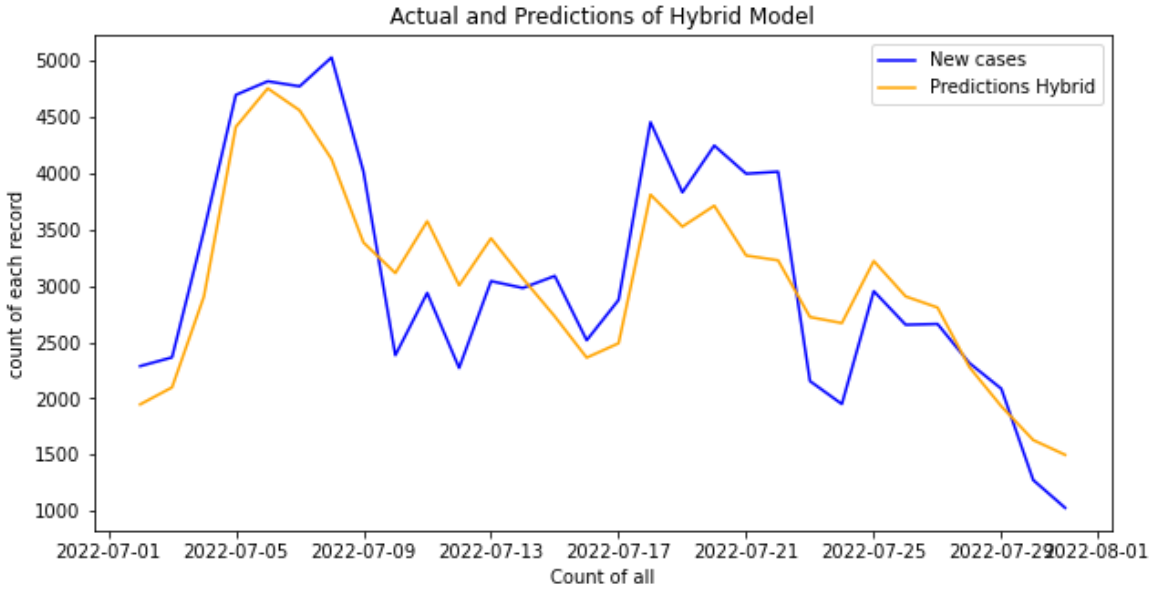

Figure 5: Plot of predictions and observations for EH-CNN-LSTM model

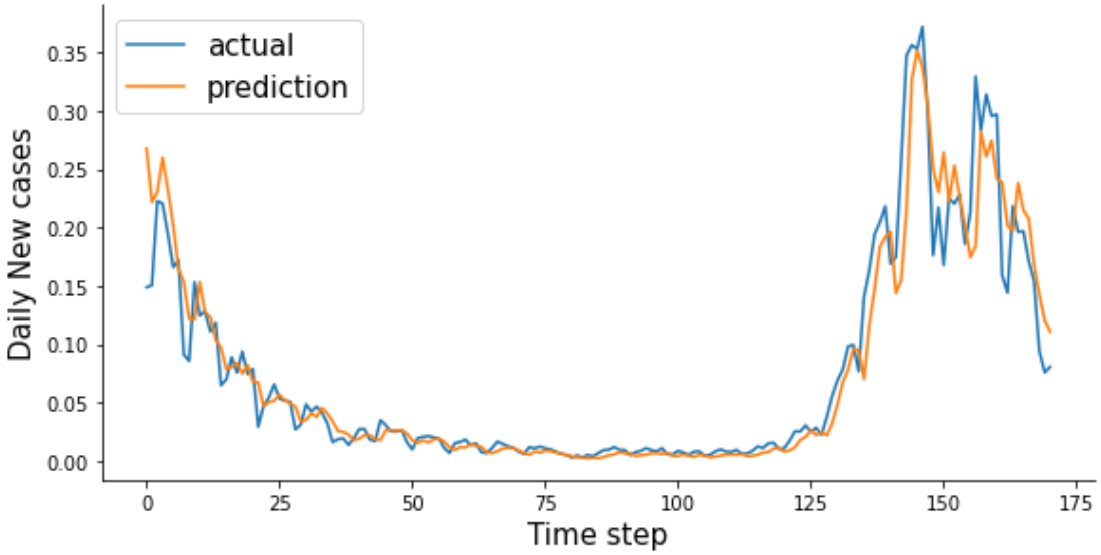

Figure 5: Plot of observed and predicted cases for the period 24-02-2020 to 01-08-2022

6. Discussion

In order to get the Scenario results for the four different regression methods, train-test split of data was created for the scenario. The results using this data split are reported in the form of scenario, and they are helpful for evaluating the utility of a larger dataset size for improved prediction performance. The performance of various prediction algorithms changed as a result of a change in the train-test split, and the enhanced hybrid CNN-LSTM (EH-CNN-LSTM) method produced the best performance overall. EH-CNN-LSTM, which was trained on 80% of the data and evaluated on 20% of the data, achieved a MAPE of 5.28, MSLE of 0.00, and RMSLE of 0.02. In addition, ANN scores the lowest in terms of predictive performance, and both its loss plots and its larger proportion of error point to the possibility that it has been overfit. Both CNN and LSTM models provided an intermediate level of performance, although the exact level varied depending on the Scenario.

7. Conclusion

Disturbance to international commerce and economies all over the globe have resulted from the Covid-19 pandemic. Governments and healthcare institutions are responding by taking aggressive actions to prevent the spread of disease. Each of these interventions may have a different effect on the spread and prevalence of Covid-19 and the number of deaths caused by this virus, so it is essential to evaluate their relative effectiveness. In addition, the ability to foresee the influx of new cases enables policymakers and health experts to implement preventative measures in a timely manner. Prediction algorithms that can estimate the daily occurrence of instances can be developed with the use of machine learning, therefore facilitating the achievement of this goal.

The analysis of the dataset and the research requirements indicate that a time series forecasting algorithm can be trained to predict new cases of COVID-19. A systematic methodology consisting of six stages is followed for research development. The data preparation is carried out

after exploratory data analysis to transform it into a suitable for predictive modeling. To train a prediction algorithm, four candidate models are trained and evaluated on the prepared dataset. Deep neural network model which is a feed forward neural network is observed to be a poor choice for modeling this problem. CNN and LSTM provided intermediate performance and their predictive performance varied with the amount of training data where CNN was more suitable when the amount of training data was increased, and LSTM proved to be more successful in the event of lesser training data. The main contribution of this work is to propose an enhanced hybrid model consisting of CNN and LSTM connected in sequential manner is used to model the problem and superior predictive performance is achieved. The predictive performance of the model is especially effective when 80% of the data is used for model training and 20% is used for evaluation. The best performing model provided an MAPE of 5.28, MSLE of 0.00 and RMSLE of 0.02. The prediction of the model on new cases is performed and provided in the form of plot indicating trend of new cases and close prediction of the proposed model.

References

[1] Rath, M. (2022). Machine learning and its use in e-commerce and e-business. In *Research Anthology on Machine Learning Techniques, Methods, and Applications* (pp.1193-1209). IGI Global.

[2] Richter, C., O'Reilly, M., & Delahunt, E. (2021). Machine learning in sports science: challenges and opportunities. *Sports Biomechanics*, 1-7.

[3] Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, *14*, 156-180.

[4] Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric reviews*, *29* (5-6), 594-621.

[5] Samuel Lalmuanawma, J. H., LalrinfelaChhakchhuak. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Science Direct*.

[6] Bhaskar, S., Bradley, S., Sakhamuri, S., Moguilner, S., Chattu, V. K., Pandya, S., Schroeder, S., Ray, D., & Banach, M. (2020). Designing futuristic telemedicine using artificial intelligence and robotics in the COVID-19 era. *Frontiers in public health*, 708.

[7] Salakhutdinov, R., & Larochelle, H. (2010). Efficient learning of deep Boltzmann machines. Proceedings of the thirteenth international conference on artificial intelligence and statistics,

[8] Károly, A. I., Fullér, R., & Galambos, P. (2018). Unsupervised clustering for deep learning: A tutorial survey. *Acta PolytechnicaHungarica*, *15* (8), 29-53.

[9] Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, *4* (5), 5947.

[10] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv: 1803.08375*.

[11] Burkill, J. C. (2004). *The lebesgue integral*. Cambridge University Press.

[12] Niazkar, H. R. (2020). Application of artificial neural networks to predict the COVID-19 outbreak. *BMC*.

[13] Jarndal, A., Husain, S., Zaatar, O., Gumaei, T. A., & Hamadeh, A. (2020). GPR and ANN based Prediction Models for COVID-19 Death Cases. *IEEE Xplore*.

[14] Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1-16.

[15] Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.

[16] Kayode Oshinubi, A. A., Olumuyiwa James Peter, Mustapha Rachdi and. (2021). Approach to COVID-19 time series data using deep learning and. *AIMS Bioengineering*, 1–21.

[17] Nahla F. Omran, S. F. A.-e. G. (2021). Applying Deep Learning Methods on Time-Series Data for Forecasting COVID-19 in Egypt, Kuwait, and Saudi Arabia. *Hindawi*.

[18] Hafiz Tayyab Rauf, M. I. U. L., Muhammad Attique Khan, Seifedine Kadry, Hanan Alolaiyan, Abdul Razaq & Rizwana Irfan. (2021). Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks. *Springer Link*.

[19] Batool, H. (2021). Correlation Determination between COVID-19 and Weather Parameters Using Time Series Forecasting: A Case Study in Pakistan. *Hindawi*.

[20] Pandas. (2022). *docs/reference/api/pandas. read_csv. html*. https: //pandas. pydata. org/docs/reference/api/pandas. read_csv. html

[21] Numpy. (2022). *docs*. https: //numpy. org/doc/stable/reference/generated/numpy. array. html