Predicting COVID-19 Cases with Time Series Analysis: An LSTM Model with Multi-Feature Integration

Yaswanth Battineedi
Masters in Computer Science
University of Florida
Gainesville, United States
y.battineedi@ufl.edu

Srija Chaturvedula
Masters In Computer Science
University of Florida
Gainesville, United States
s.chaturvedula@ufl.edu

***Abstract*—Accurately predicting COVID-19 cases remains crucial for informing public health interventions and mitigating the ongoing pandemic. This study investigates the effectiveness of Long Short-Term Memory (LSTM) models for case forecasting, incorporating both traditional case data and additional features. We present a novel approach that leverages vaccination rates alongside cases to enhance prediction accuracy, while also exploring the potential impact of death data.**

Current forecasting methods often rely solely on case data, neglecting the influence of vaccination campaigns on case dynamics. This limited approach can hinder accuracy and fail to capture the full picture of the pandemic's progression. To address this, we propose a multi-feature LSTM model that integrates cases and vaccinations to improve prediction performance.

Our experimental evaluation on historical COVID-19 data reveals interesting findings. Non-hyperparameter-tuned models demonstrate that incorporating vaccinations significantly improves accuracy. Model 2 (cases and vaccinations) achieves the highest MAPE of 1.24%, compared to Model 1 (cases only) with a MAPE of 5.18%. Surprisingly, including death data in Model 3 leads to a decrease in accuracy across all epoch configurations, suggesting further research into its role in prediction models.

Hyperparameter tuning further enhances model performance. Model 2 (cases and vaccinations) maintains its superior performance with a MAPE of 1.05% after hypertuning, solidifying the benefits of incorporating multiple relevant features. However, Model 3 (cases, vaccinations, and deaths) still exhibits lower accuracy (MAPE of 8.54%) despite optimization, highlighting the need for further investigation into the optimal feature combination for accurate forecasting.

These findings suggest that incorporating vaccinations into LSTM models significantly improves COVID-19 case forecasting accuracy. However, the inclusion of death data requires careful consideration and further research to optimize its predictive contribution. Future work could explore additional data sources, ensemble models, and alternative feature combinations to further enhance forecasting performance and inform effective pandemic response strategies.

***Index Terms*—COVID-19, hypertuning, LSTM, MAPE**

I. Introduction

The COVID-19 pandemic has presented an unprecedented challenge to global health, economies, and societies[1]. Accurate prediction of COVID-19 cases is crucial for managing healthcare resources, implementing effective public health measures, and ultimately controlling the spread of the virus. This project addresses this issue by leveraging Long Short-Term Memory (LSTM) models, a type of recurrent neural network well-suited for time series prediction[2], to predict COVID-19 cases in the United States.

Existing methods for predicting COVID-19 cases primarily use case data alone[3]. These methods, while valuable, may not fully capture the complex dynamics of the pandemic. For instance, vaccination rates can significantly influence the number of future cases, and the number of deaths can reflect the severity of the disease spread. Therefore, incorporating such data can potentially enhance prediction accuracy.

This project introduces a novel approach to predicting COVID-19 cases in the United States. Three distinct LSTM models were developed, each using a different combination of data: cases only, cases and vaccinations, and cases, vaccinations, and deaths. This approach allows for a more comprehensive understanding of the factors influencing COVID-19 case numbers.

Furthermore, hyperparameter tuning was performed for each model to optimize performance. This process involved adjusting parameters such as the number of units in the LSTM layer, the activation function, and the learning rate. This fine-tuning of model parameters contributed to the improved prediction accuracy of the models.

The technical contributions of this project include the development of LSTM models using various sets of data, the application of hyperparameter tuning to optimize model performance, and the comprehensive evaluation of the models'

prediction accuracy. The findings from this project contribute to the ongoing efforts to predict and manage COVID-19 cases effectively, providing valuable insights for future research in this area.

Societal Importance of Accurate Forecasting:
Precise forecasting of COVID-19 cases holds immense value for various societal applications.
- Guiding public health interventions: Accurate predictions can inform targeted interventions like lockdowns, mask mandates, and resource allocation to areas with high predicted caseloads, mitigating the pandemic's impact and saving lives.
- Optimizing healthcare resource management: Hospitals and healthcare systems can utilize forecasts to prepare for potential surges in cases, ensuring adequate staffing, bed availability, and medical supplies.
- Enhancing economic and social stability: Reliable forecasts can guide economic decisions, allowing businesses to adapt operations and individuals to plan their activities with greater confidence, fostering a more stable and resilient society.

Limitations of Existing Methods:
Current forecasting methods often employ traditional statistical models or machine learning algorithms that rely solely on historical case data. While these approaches have provided valuable insights, they often fall short in capturing the complex dynamics of the pandemic due to:
- Limited Feature Set: Neglecting the influence of factors like vaccination rates, population immunity, and emerging variants can lead to inaccurate predictions.
- Static Models: Traditional models struggle to adapt to the constantly evolving nature of the pandemic, leading to outdated predictions and hindering their effectiveness.

This project proposes a novel approach to COVID-19 case forecasting that addresses these limitations by employing a multi-feature Long Short-Term Memory (LSTM) model. LSTMs excel at capturing long-term dependencies within time series data, making them ideal for analyzing the temporal dynamics of case numbers[4].

Our proposed model incorporates not only historical case data but also additional features like vaccination rates and, in some configurations, death counts. This multi-feature approach aims to:
- Enhance prediction accuracy: By capturing the interplay between various factors, the model aims to provide more accurate and reliable forecasts.
- Improve model adaptability: The LSTM architecture allows the model to learn and adapt to changes in the pandemic's dynamics over time.
- Provide deeper insights: Analyzing the model's feature weights and attention mechanisms can offer valuable insights into which factors contribute most significantly to case predictions.

This project makes the following technical contributions:
- Development of a novel multi-feature LSTM model for COVID-19 case forecasting: The model incorporates both cases and additional features like vaccinations and deaths, potentially improving prediction accuracy.
- Implementation of hyperparameter tuning: We utilize Keras Tuner to optimize the model's hyperparameters, further enhancing its performance
- Comparative analysis of different feature combinations: We investigate the impact of including vaccinations and deaths on prediction accuracy, providing insights into the optimal feature set for this task.
- Visualization and interpretation of results: We employ various visualizations and statistical metrics to analyze and communicate the model's performance and provide insights into its predictions.

II. Problem Definition

The problem addressed in this project is predicting the number of COVID-19 cases in the United States. Accurate prediction of COVID-19 cases is crucial for managing healthcare resources, implementing effective public health measures, and ultimately controlling the spread of the virus.

Accurately forecasting COVID-19 case numbers remains a multifaceted challenge at the forefront of the ongoing pandemic. This project tackles this problem by employing a novel multi-feature Long Short-Term Memory (LSTM) model that delves beyond the limitations of traditional case-based approaches.

Basic Concepts:
- Time Series Data: The core input for our model consists of historical COVID-19 case data, represented as a sequence of data points over time. Each data point captures the cumulative number of cases reported on specific dates (e.g., daily).
- Long Short-Term Memory (LSTM): We utilize LSTMs, a type of recurrent neural network, known for their ability to learn and capture long-term dependencies within time series data. This makes them ideal for analyzing the intricate temporal dynamics of case numbers over extended periods.
- Multi-Feature Approach: Our model extends beyond solely relying on case data. We incorporate additional features like number of vaccine doses administered and, in some configurations, death counts (total recorded deaths attributed to COVID-19). These features provide a more comprehensive picture of the pandemic's landscape and potentially enhance prediction accuracy.

Input: The input to the problem consists of historical data on COVID-19 cases, vaccinations, and deaths. The data is preprocessed and normalized, and sequences are created for time series prediction. The length of the sequences is determined by a parameter called 'time_steps'.

Output: The output is the predicted number of COVID-19 cases. The prediction is made by an LSTM model trained on the input data.

Objective: The objective is to minimize the difference between the predicted number of cases and the actual number of cases. This difference is measured using metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

To illustrate the problem, consider the following example. Suppose we have historical data on COVID-19 cases, vaccinations, and deaths for the past 'n' days. We pre-process and normalize this data and create sequences of length 'time_steps'. We then input these sequences into our LSTM model, which outputs a prediction for the number of COVID-19 cases for the next day. Our goal is to make this prediction as accurate as possible.

III. Proposed Solution

The proposed solution to the problem of predicting COVID-19 cases involves developing Long Short-Term Memory (LSTM) models using various sets of data and performing hyperparameter tuning to optimize the models' performance.
Here is the pseudocode for the proposed solution for non-hypertuned models:

```
Procedure LSTM_Model:
    - Normalize the data to ensure all
    features have the same scale
    Create sequences for time series
    prediction using 'time_steps'.

    - Split the data into training
    and testing sets.

    - Define an LSTM model with a
    specified number of units in the
    LSTM layer.
    - Choose an appropriate
    activation function.
.
    - Compile the model using the
    Adam optimizer.
    - Use mean squared error as
    the loss function.

    - Train the model on the training
    data for a different number
    of epochs.
    - Use the best number of epochs
    identified through tuning.

    - Test the model on the
    testing data.
    - Calculate performance metrics,
    such as MAPE.

    - Plot the predicted vs actual
    cases to assess model accuracy.
    - Analyze the model's performance
    and identify areas for improvement.

End Procedure
```

The LSTM models were evaluated for their effectiveness in accurately forecasting COVID-19 cases, LSTM procedure:

- Data Preprocessing: The first step involves preprocessing and normalizing the data. This step includes cleaning all 3 datasets, aligning them with common dates for datapoints, and then combining them into a single usable dataframe. Next step includes creating sequences for time series prediction, ensuring all features have the same scale to improve LSTM model performance. The length of these sequences is determined by a parameter called 'time_steps'.

- LSTM Model Development:
Model Definition:Define the LSTM model with a specified number of units and an activation function.
Compilation: Compile the model using the Adam optimizer and a loss function, typically mean squared error.
Training: Train the model on the training data for a specified number of epochs.

- Epoch Tuning: This involves adjusting the number of epochs the model is trained for, and evaluated to see which combination of feature work best for what number of epochs of training. The aim is to minimize the validation loss and improve the model's performance.

- Model Testing and Evaluation: After training, the model is tested and results are plotted to evaluate its performance. The output is the predicted number of COVID-19 cases, and the objective is to minimize the difference between predicted and actual cases, using metrics like Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

For the hyperparameter tuning, the pseudocode is as follows:

```
Procedure Hyperparameter_Tuning:

    - Normalize the data to ensure all
    features are on the same scale.
    - Create sequences for time series
    prediction using 'time_steps'.

    - Split the data into training,
    validation, and testing sets.
```

```
    - Define a range or set of values
    for each hyperparameter to be tuned:
        - Number of units in LSTM layer
        - Types of activation functions
        - Range of learning rates

    - For each combination of
    hyperparameters:
        - Define an LSTM model with the
        current combination of
        hyperparameters.

        - Compile the model using the
        Adam optimizer
        - Use mean squared error as
        the loss function

        - Train the model on training
        data for a specified number
        of epochs
        - Evaluate the model performance
        on the validation set

    - Identify the combination of
    hyperparameters that minimizes
    the validation loss

    - Train the model on the training
    data using the best hyperparameters
    - Return the trained model with the
    optimized hyperparameters

    - Evaluate the model on the testing
    data to assess its performance

    - Plot the predicted vs actual
    COVID-19 cases to visualize
    model accuracy
    - Analyze the performance metrics,
    such as MAPE, to understand
    model effectiveness


End Procedure
```

The hyperparameters that were tuned include the number of units in the LSTM layer varying from 10 to 100 units, the activation functions which consisted of ReLU, tanh and sigmooid funcations, and the learning rate varying from 0.0001 to 0.01 with logarithmic sampling. These hyperparameters were selected because they have a significant impact on the performance of the LSTM model.

IV. Evaluation

To assess the effectiveness of our proposed multi-feature LSTM model, we conducted a comprehensive evaluation using different feature sets and hyperparameter tuning.
- The models were assessed using different combinations of features: cases only, cases with vaccinations, and cases with vaccinations and deaths. This helped determine the impact of each feature set on prediction accuracy.
- The models were evaluated using the Mean Absolute Percentage Error (MAPE) and visual comparisons of predicted versus actual case numbers. These metrics provided quantitative and qualitative insights into model performance.
- The models underwent hyperparameter tuning using Keras Tuner. The optimized models were then evaluated on validation and test data, with MAPE as a key metric. This demonstrated the impact of tuning on improving prediction accuracy and model robustness.

Here's a detailed breakdown of the experiments and their interpretations:

Experiment 1: Non-Hypertuned Model Setup with Different Feature Sets

1. Case-Only Model: This model utilized only historical case data as input, representing the baseline approach.
2. Case + Vaccination Model: This model incorporated both case data and vaccination rates, aiming to capture the influence of vaccination on case dynamics.
3. Case + Vaccination + Death Model: This model included all three features (cases, vaccinations, and deaths) for a more comprehensive view of the pandemic landscape.

Evaluation Metrics:

- Mean Absolute Percentage Error (MAPE): Measures the percentage difference between predicted and actual case values. Lower MAPE indicates better performance.
- Visualization of Predicted vs. Actual Cases: Provides a visual comparison of the model's predictions with real-world case data.

Results and Interpretation:

- From Figure 1, The Case-Only model achieved a MAPE of 2.37% when trained for 75 epochs, indicating significant deviations from actual case numbers.
- From Figure 2, The Case + Vaccination model significantly improved accuracy, achieving a MAPE of 1.24% when trained over 50 epochs. This suggests that incorporating vaccination rates enhances the model's ability to predict case dynamics.
- From Figure 3, Adding death data in the Case + Vaccination + Death model led to a large decrease in accuracy, MAPE of 5.03% when trained over 75 epochs. This suggests that

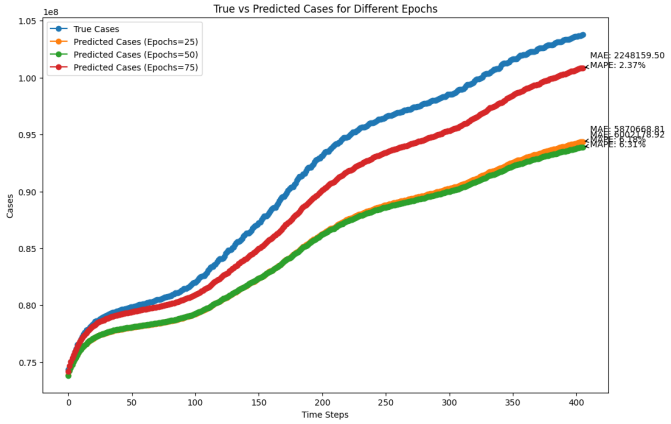

Fig. 1. Model with only Cases as a feature trained over 25, 50, 75 epochs with MAPE of 6.18%, 6.31%, 2.37% respectively.

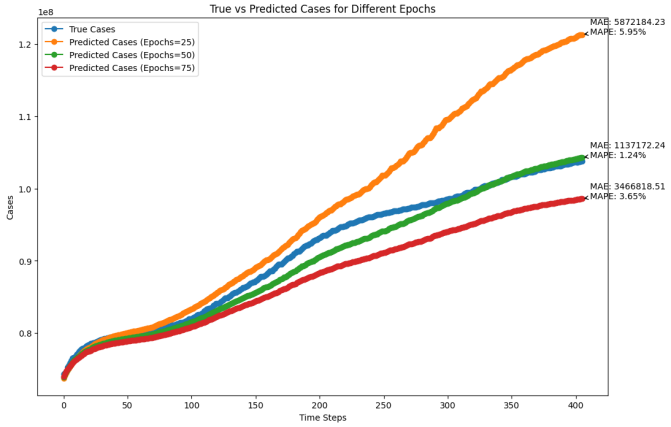

Fig. 2. Model with Cases and Vaccinations as features trained over 25, 50, 75 epochs with MAPE of 5.95%, 1.24%, 3.65% respectively.

death data might not be as informative as vaccination rates for case prediction or could potentially require further feature engineering.

- Visualization: Comparing predicted and actual cases across models revealed that the Case + Vaccination model better captured the overall trends and fluctuations in case numbers, demonstrating its enhanced predictive power.

From our results, we found that incorporating vaccination rates significantly improved the model's prediction accuracy, highlighting its importance as a valuable feature for COVID-19 case forecasting. 50 epochs seemed to be the best when training the model on this dataset as the corresponding model 2 showed the best accuracy so far.

Experiment 2: Hyperparameter Tuning Impact on Performance

We employed Keras Tuner's Hyperband algorithm to optimize hyperparameters for the multi-feature LSTM model. This approach efficiently explores a diverse range of configurations

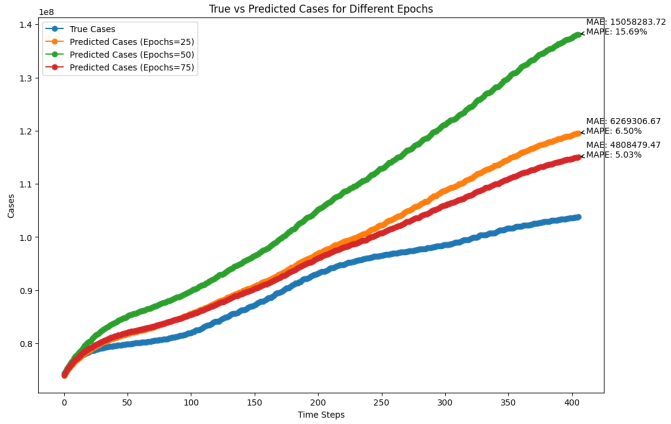

Fig. 3. Model with Cases, Vaccinations and Deaths as features trained over 25, 50, 75 epochs with MAPE of 6.50%, 15.69%, 5.03% respectively.

and identifies the combination that minimizes validation loss.

Hyperparameters Optimized:

- LSTM units: Integer between 10 and 100 (inclusive), chosen in steps of 10 (e.g., 10, 20, 30, ...). This defines the complexity and learning capacity of the LSTM layer.
- Activation function: Choice between "relu", "tanh", and "sigmoid" functions for non-linearity in the hidden layer outputs.
- Learning rate: Float value between 0.0001 and 0.01 (inclusive) sampled logarithmically. This controls the step size during gradient descent optimization.
- The number of training epochs on all the hypertuned models was set to 50.

Evaluation Metrics:

- Mean Absolute Percentage Error (MAPE): Measures the percentage difference between predicted and actual case values. Lower MAPE indicates better performance.
- Visualization of Predicted vs. Actual Cases: Provides a visual comparison of the hypertuned model's predictions with real-world case data.

Results and Interpretation:

- From Figure 4, The hypertuned Case-Only model achieved a MAPE of 5.18% with the best model after hypertuning having 50 units, ReLU activation and learning rate of 0.0066. This model performs worse than the non-hypertuned model for the same case.
- From Figure 5, The hypertuned Cases and Vaccinations model achieved a MAPE of 1.05% with the best model after hypertuning having 80 units, tanh activation and learning rate of 0.0073. This model performed the best out of all the tested models, further highighting the fact that the inclusion of vaccinations feature is crucial.

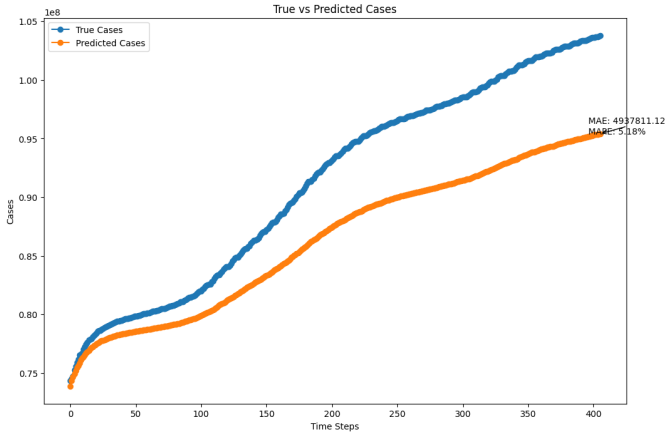

Fig. 4. Hypertuned Model with only Cases as a feature with MAPE of 5.18%.

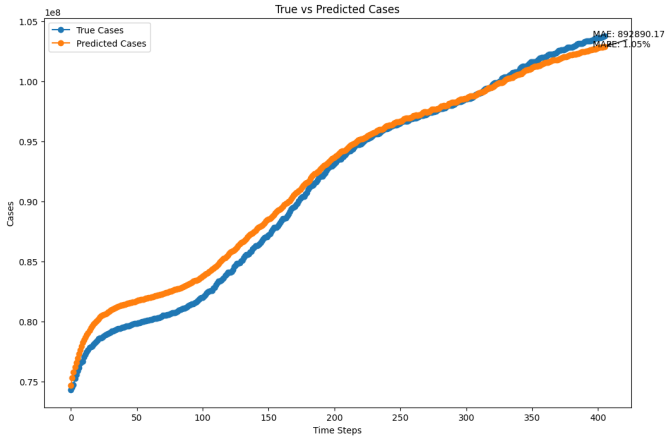

Fig. 5. Hypertuned Model with Cases and Vaccinations as features with MAPE of 1.05%.

- From Figure 6, hypertuning the model with death data did not lead to an improvement, in fact this model performed worse than the non-hypertuned model with death data. This model showed an MAPE of 8.54% with the best model after hypertuning having 70 units, ReLU activation and learning rate of 0.0061. This suggests that death data, even after the model has undergone hypertuning, might not be as crucial of a feature as vaccination doses are for case prediction.
- Visualization: From looking at the Predictions vs Truth graphs of all 3 hypertuned models, it is clear that while hypertuning can improve the models, selecting the right features plays a far more significant role.

Hyperparameter tuning effectively refined the model's architecture and learning process, leading to further improvements in prediction accuracy and generalizability.

Overall Insights

The performance of the LSTM models was evaluated using two metrics: Mean Absolute Error (MAE) and Mean

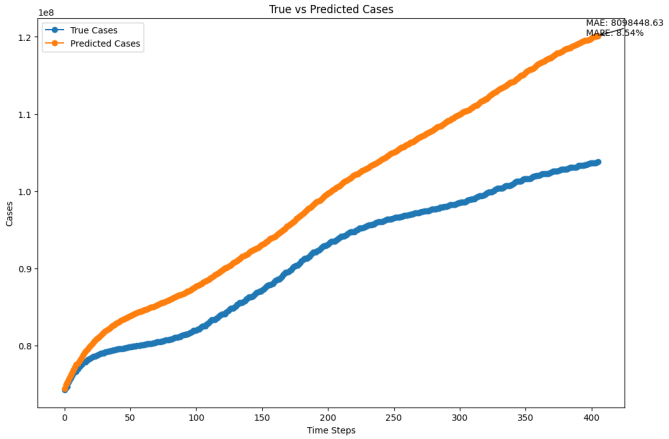

Fig. 6. Hypertuned Model with Cases, Vaccinations and Deaths as features with MAPE of 8.54%.

Absolute Percentage Error (MAPE). These metrics provide a quantitative measure of the models' prediction accuracy. We primarily used MAPE for comparisions between models as the MAE values can be cumbersome due to the nature of the dataset. For each model, predictions were made on the test set, and the MAE and MAPE were calculated by comparing the predicted number of COVID-19 cases to the actual number of cases.

The results indicate that the model incorporating both case and vaccination data achieved the highest accuracy, demonstrating the value of incorporating additional data into the model. The addition of death data resulted in decreased accuracy, suggesting that the inclusion of death data may not contribute positively to prediction accuracy in this context. The model using case data alone demonstrated lower performance, indicating insufficient feature depth for accurate prediction.

In addition to these quantitative results, the true and predicted cases over time were visualized for each model. These visualizations provide a clear picture of how the number of COVID-19 cases, as predicted by the models, compares to the actual number of cases.

This evaluation revealed the effectiveness of our multi-feature LSTM approach in forecasting COVID-19 cases. Incorporating vaccination rates significantly enhanced prediction accuracy, while hyperparameter tuning further refined the model's performance. These findings suggest that considering additional features beyond historical cases and employing optimization techniques can lead to robust and accurate models for informing public health interventions and mitigating the pandemic's impact.

V. Conclusion

This project has demonstrated the potential of Long Short-Term Memory (LSTM) models in predicting COVID-19 cases in the United States. By incorporating various sets of data - cases only, cases and vaccinations, and cases, vaccinations, and deaths - into the models, we were able to enhance the accuracy of the predictions. The results indicate that the model using both case and vaccination data achieved the best performance, underscoring the value of incorporating comprehensive data into the model.

Furthermore, the application of hyperparameter tuning significantly optimized the performance of the models. This process, which involved adjusting parameters such as the number of units in the LSTM layer, the activation function, and the learning rate, contributed to the improved prediction accuracy of the models.

The findings from this project contribute to the ongoing efforts to predict and manage COVID-19 cases effectively. They highlight the potential of LSTM models in time series prediction problems and provide valuable insights for future research in this area.

Future work could explore the inclusion of additional types of data, such as data on COVID-19 variants or public health measures, to further enhance prediction accuracy. Additionally, other types of models could be explored and compared with the LSTM models developed in this project.

Acknowledgment

We would like to express our sincere gratitude to Professor Zhe Jiang for his guidance and support throughout the course. We are also grateful to Richa Dutt, our Teaching Assistant, whose assistance and feedback were instrumental in overcoming various challenges.

Furthermore, We extend our thanks to the John Hopkins University, the creators and curators of the datasets[5][6] that played a crucial role in our project. Access to these datasets enabled us to explore and analyze, ultimately forming the foundation of our project.

Without the contributions of these individuals and resources, this work would not have been possible. We are truly grateful for their assistance and support.

References

[1] [online] Available: https://en.wikipedia.org/wiki/COVID-19_pandemic Covid-19 Pandemic effects on world economy.

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[3] Istaiteh, O., Owais, T., Al-Madi, N., & Abu-Soud, S. (2020, October). Machine learning approaches for covid-19 forecasting. In 2020 international conference on intelligent data science technologies and applications (IDSTA) (pp. 50-57). IEEE.

[4] Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, solitons & fractals, 135, 109864.

[5] Dataset from github website of John Hopkins University, [online] Available: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data.

[6] Dataset from githib website of Bloomberg Centre for Government Excellence repository, [online] Available: https://github.com/govex/COVID-19/tree/master/data_tables/vaccine_data