**Report on Challenges Faced**

**1. Data Collection and Preprocessing Challenges**

* **Challenge**: The COVID-19 dataset provided by Johns Hopkins University is updated daily, leading to data inconsistencies such as missing values, irregular updates, and occasional outliers.
  + **Solution**: Missing values were handled using interpolation techniques, and we carefully reviewed outliers before deciding to either cap or remove them based on their impact on the model’s predictions.

**2. Time-Series Data Preparation**

* **Challenge**: Transforming the dataset into a form suitable for time-series analysis posed a challenge, particularly in capturing daily trends, lags, and seasonality effects.
  + **Solution**: We calculated daily changes in confirmed cases and deaths, and generated lag features to model temporal dependencies effectively. Smoothing techniques were applied to reduce noise, enabling more accurate predictions.

**3. Model Selection and Complexity**

* **Challenge**: Identifying the right models for forecasting proved difficult due to the diverse nature of available algorithms. Simpler models like Linear Regression struggled to capture the non-linearities and dynamics in the data, while more complex models such as LSTM required significant computational power.
  + **Solution**: A comprehensive model comparison approach was adopted, testing models with cross-validation to ensure robustness. We prioritized models based on accuracy and scalability, ultimately selecting LSTM for its superior ability to handle time-dependent data.

**4. Overfitting in Non-Linear Models**

* **Challenge**: Polynomial Regression and high-variance models like Random Forest showed tendencies to overfit the data, especially when using higher-degree polynomial features.
  + **Solution**: To address overfitting, we applied cross-validation and introduced regularization techniques, such as Ridge and Lasso regression, where necessary. These techniques helped control model complexity and reduced overfitting.

**5. Computational Constraints**

* **Challenge**: The LSTM model, while highly accurate, required significant computational resources for training, especially on large datasets with long time sequences.
  + **Solution**: We optimized the LSTM model by reducing the number of epochs, using batch processing, and applying early stopping techniques to avoid unnecessary computations without sacrificing performance. Model training was also parallelized where possible.

**6. External Factors and Prediction Uncertainty**

* **Challenge**: COVID-19 case predictions are inherently uncertain due to external factors such as government interventions, public compliance with safety measures, and the emergence of new virus variants.
  + **Solution**: To account for prediction uncertainty, we generated confidence intervals around the predictions, providing a range of possible outcomes. Additionally, we emphasized the importance of updating the model with new data as it becomes available to improve prediction accuracy over time.

**Conclusion**

The COVID-19 prediction project faced several challenges, from data inconsistencies to model complexity. However, through careful preprocessing, model selection, and tuning, we were able to build a reliable forecasting model. The **LSTM model** emerged as the most effective, although simpler models like **Random Forest** offered a practical alternative in resource-constrained environments. By continuously updating the model with fresh data and applying confidence intervals, we aim to provide actionable insights for public health planning and decision-making.