COMP3115 Exploratory Data Analysis and Visualization

Topic 1: Open Covid-19 Data Analytics

Introduction

Recent years, COVID-19 has been affecting many aspects of our life. Therefore, related data analysis plays a critical role in understanding the impact of the pandemic and making informed decisions to mitigate its spread. What's more, it plays a vital role in managing the pandemic, protecting public health, and saving lives.

This project aims to understand the spread and impact of the virus, inform public health policies and interventions, optimize resource allocation, and support research and development efforts. According to the suggested data source, I identify four tasks to fully understand and dig deep data information.

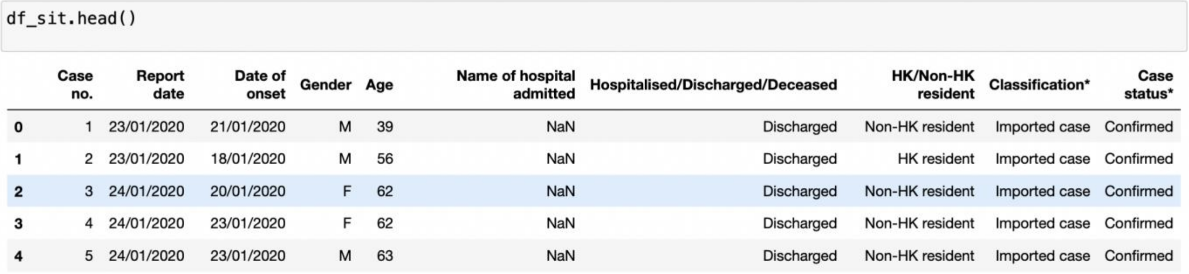
- Task1: Visualize the latest local situation data, including daily trends in cases and gender distribution of cases.
- Task2: Predict the trend of the number of cases in mainland China in the following days and visualize the result.
- Task3: Analyze the high-frequency areas visited by COVID-19 cases, including counting the COVID-19 case distribution by district and clustering analysis of buildings visited by COVID-19 cases.
- Task4: Analyze the COVID-19 situation of areas outside China through the vehicle data and cases data.

Data processing

In this project, the suggested data sources are from:

https://data.gov.hk/en-data/dataset/hk-dh-chpsebcddr-novel-infectious-agent

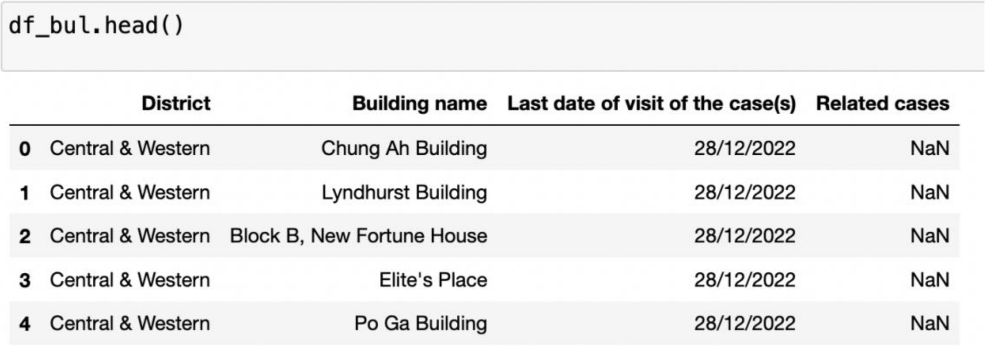
Firstly, I identify the data sources relevant to the four tasks above and download them. And then, I tabulate the acquired data and their statistics to gain better understanding of them.



```
df_sit.head()
```

	Case no.	Report date	Date of onset	Gender	Age	Name of hospital admitted	Hospitalised/Discharged/Deceased	HK/Non-HK resident	Classification*	Case status*
0	1	23/01/2020	21/01/2020	M	39	NaN	Discharged	Non-HK resident	Imported case	Confirmed
1	2	23/01/2020	18/01/2020	M	56	NaN	Discharged	HK resident	Imported case	Confirmed
2	3	24/01/2020	20/01/2020	F	62	NaN	Discharged	Non-HK resident	Imported case	Confirmed
3	4	24/01/2020	23/01/2020	F	62	NaN	Discharged	Non-HK resident	Imported case	Confirmed
4	5	24/01/2020	23/01/2020	M	63	NaN	Discharged	Non-HK resident	Imported case	Confirmed

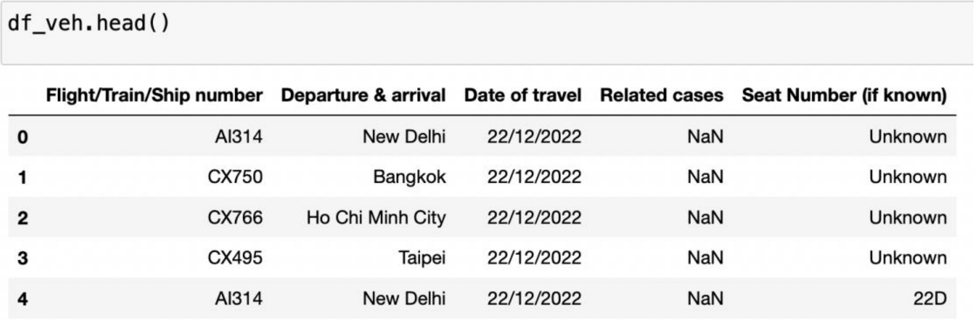
The latest local situation of COVID-19



```
df_bul.head()
```

	District	Building name	Last date of visit of the case(s)	Related cases
0	Central & Western	Chung Ah Building	28/12/2022	NaN
1	Central & Western	Lyndhurst Building	28/12/2022	NaN
2	Central & Western	Block B, New Fortune House	28/12/2022	NaN
3	Central & Western	Elite's Place	28/12/2022	NaN
4	Central & Western	Po Ga Building	28/12/2022	NaN

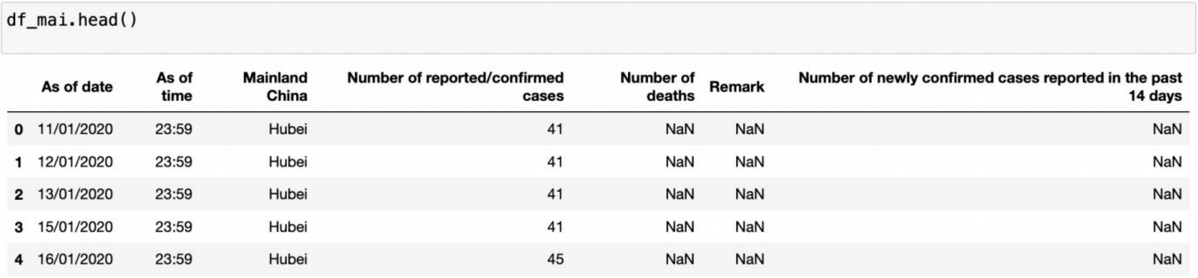
List of buildings visited by cases tested positive for SARS-CoV-2 virus in the past 7 days



```
df_veh.head()
```

	Flight/Train/Ship number	Departure & arrival	Date of travel	Related cases	Seat Number (if known)
0	AI314	New Delhi	22/12/2022	NaN	Unknown
1	CX750	Bangkok	22/12/2022	NaN	Unknown
2	CX766	Ho Chi Minh City	22/12/2022	NaN	Unknown
3	CX495	Taipei	22/12/2022	NaN	Unknown
4	AI314	New Delhi	22/12/2022	NaN	22D

List of flights/trains/ships/vehicles taken by cases tested positive for SARS-CoV-2 virus in the past 7 days



```
df_mai.head()
```

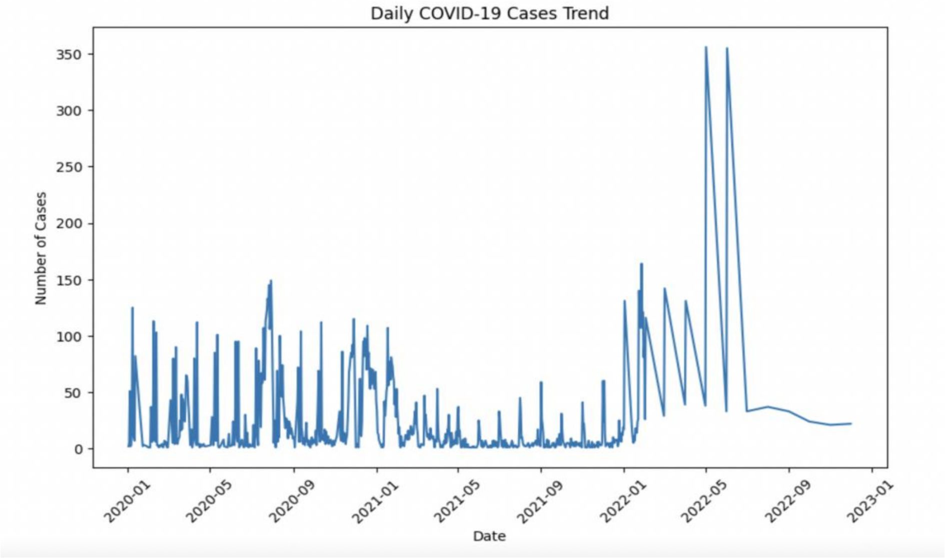
	As of date	As of time	Mainland China	Number of reported/confirmed cases	Number of deaths	Remark	Number of newly confirmed cases reported in the past 14 days
0	11/01/2020	23:59	Hubei	41	NaN	NaN	NaN
1	12/01/2020	23:59	Hubei	41	NaN	NaN	NaN
2	13/01/2020	23:59	Hubei	41	NaN	NaN	NaN
3	15/01/2020	23:59	Hubei	41	NaN	NaN	NaN
4	16/01/2020	23:59	Hubei	45	NaN	NaN	NaN

Countries/areas outside Mainland China have reported cases of COVID-19

Task1: Visualize the latest local situation data

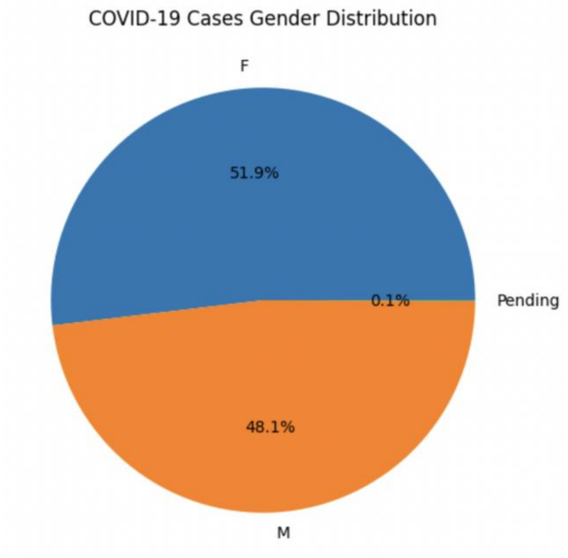
To complete this task, the data of *latest local situation of COVID-19* is necessary.

Firstly, the 'Date' should be convert to datetime format, and count the number of new cases by date. I visualize trends in daily COVID-19 cases by the 'Date' and 'Number of Cases' data. The trend chart is as follows.



From the figure, we can see that the daily number of cases peaked from May to September 2022. After September 2022, the situation of COVID-19 flatten out.

Secondly, I analyze the gender distribution of cases to estimate whether the COVID-19 is related to human gender. The pie chart is used to visualize the result and show it more intuitively.



As shown in the figure above, the ratio of men to women cases is almost equal. To a certain extent, this disease is not associated with gender. Therefore, when researchers make pathological inference for COVID-19, the difference of physiological function between men and women can get less attention. This conclusion is important for the research of the treatment of the COVID-19.

Task2: Predict the trend of the number of cases in mainland China

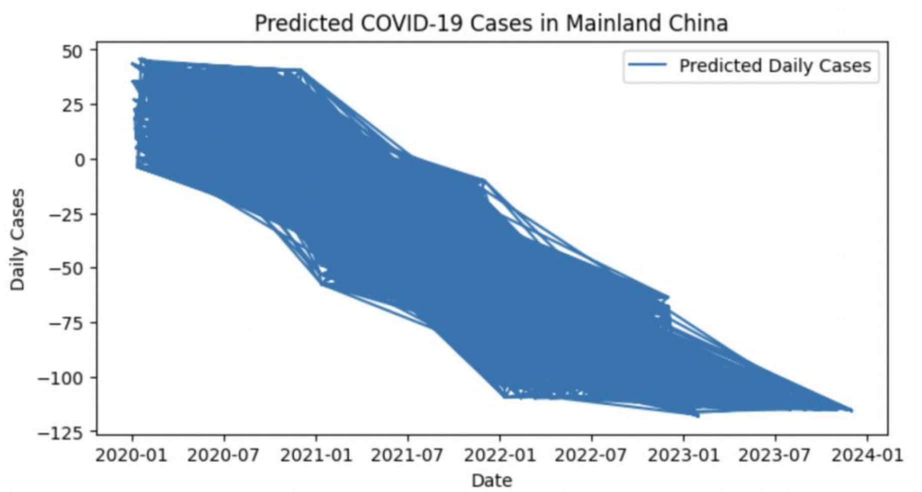
For this task, the data of *latest situation of reported cases of COVID-19 in Mainland China* is necessary.

After data reading, the 'As of date’ data should be converted to datetime format. The 'Number of reported/confirmed cases' data need to be used to calculate the number of new confirmed cases per day. And then, combined ‘Date’ data to split the training and test set. And then, using the machine learning algorithms to train models.

Two models have been chosen, which is linear regression model and MLP classifier model.

- Linear regression is a widely used statistical model. It is a classical statistical method used to establish the linear relationship between independent variables and dependent variables. Its basic idea is to find an optimal line, which can best fit the sample data, and use this line to predict the new independent variables. It is also the simplest data prediction model (Van Nguyen et al., 2021).
- The Multi-Layer Perceptron (MLP) Classifier is a type of artificial neural network that is used for supervised classification tasks. The MLP Classifier is trained using labeled data, where the input features and their corresponding class labels are used to adjust the weights and biases of the model during an iterative optimization process, typically using gradient descent or other optimization algorithms. It is a versatile and powerful model that can be used for a wide range of classification tasks, including image recognition, text classification, and speech recognition (Dimitsaki et al., 2023).

The training result is as follows.

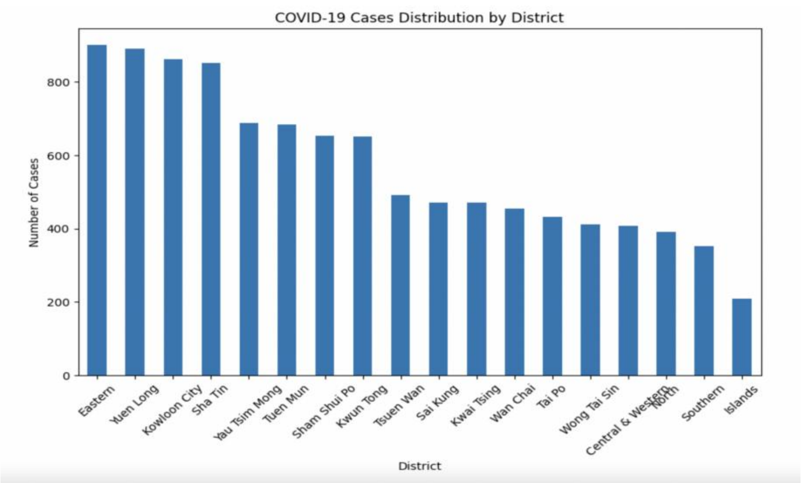


The figure above shows the prediction of COVID-19 cases in Mainland China. Predicting trends in the number of cases is important for managing pandemics and making informed decisions (Tuli et al., 2020). It helps in the early detection of potential outbreaks or case surges. By analyzing historical data and applying predictive models, the risk of an increase in cases can be identified. This can trigger timely interventions to prevent further spread of the virus and mitigate its impact. At the same time, case prediction allows for advance planning of healthcare resource allocation and facilitates optimal use of resources.

Task3: Analyze the high-frequency areas visited by COVID-19 cases

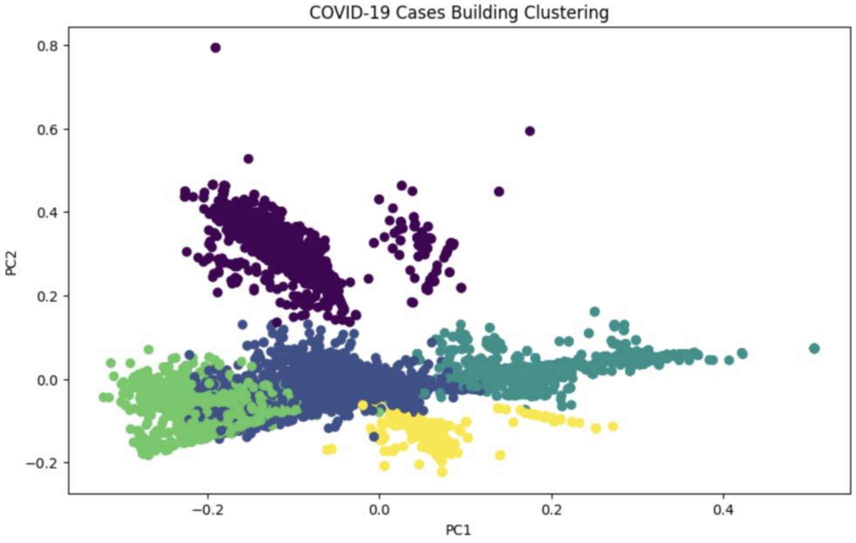
For this task, the data of *list of buildings visited by cases tested positive for SARS-CoV-2 virus* is necessary.

Firstly, I count the COVID-19 cases distribution by district and visualize the result by bar chart. The figure is as follows.



The figure above shows the cases distribution in eighteen districts of Hong Kong. Cases have appeared most frequently in Eastern district, and also the top four districts have similar proportions. It may be caused by population density and regional construction differences.

Secondly, I cluster analysis of buildings visited by COVID-19 cases. Because of the name of building is not a recognizable feature, I convert textual features to numerical features. And then use the way of K-means to cluster. Consider for the data differences of cases distribution figure, I set five clusters. At last, I use PCA to reduce the data to two dimensions to visualize the result. The figure is as follows.

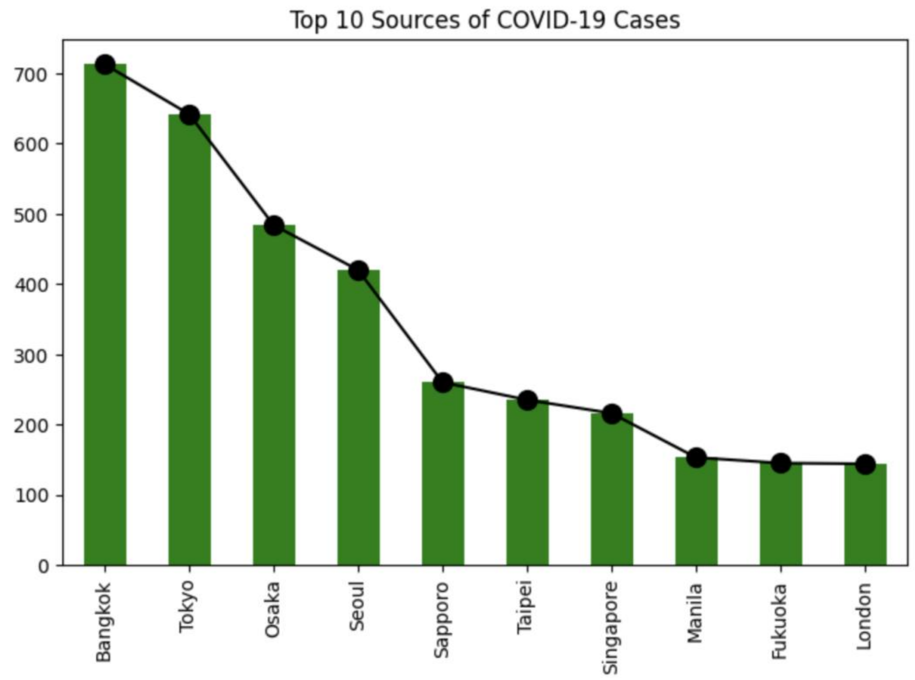


In conclusion, it is important to understand the high-frequency areas visited by COVID-19 cases to predict the situation of disease, which also helps in the allocation of health care resources and the adjustment of government control policies.

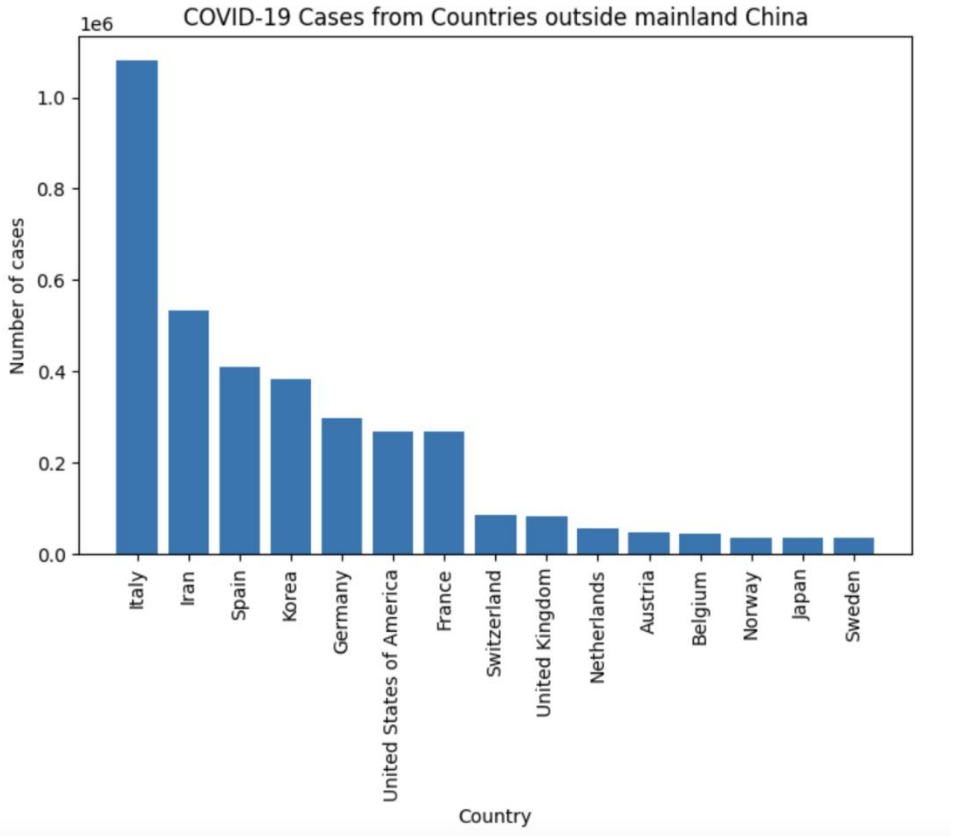
Task4: Analyze the COVID-19 situation of areas outside China

For this task, the data of *list of flights/trains/ships/vehicles taken by cases tested positive for SARS-CoV-2 virus* and *countries/areas outside Mainland China have reported cases of COVID-19* is necessary.

After analyzing the COVID-19 cases data in China, knowing more about the situation of areas outside China is also important. I count the 'Departure & arrival' of vehicles data and list the top 10 countries/areas where COVID-19 cases occurred during one week (22/12/2022-28/12/2022). The result is as follows.



For further information, I calculate the number of cases from different countries/areas outside Mainland China. The result is as follows.



From the two figures above, it can directly reflect the severity of the COVID-19 situation in the country during the time period in which the data were collected.

Conclusion

The COVID-19 is spreading fast and widely. In the past two years, it has had a great impact on the world. This project digs into relevant data through influencing factors, case trend prediction, travel trajectory clustering and domestic and foreign status analysis of the COVID-19. The supervised and unsupervised algorithms such as K-means clustering, Random Forest, CNN, Auto-Encoder, and Regression approaches are fully applied in the process of data analysis to obtain effective analysis results (Chahar et al., 2022).

Prospect

The analysis of COVID-19 is not only based on data, but also on images. Especially in disease prediction, multi-modality analysis combined data and images such as X-ray images and pathological sections will make the analysis results more useful (Ghoshal et al., 2020).

Reference

[1] Van Nguyen, Q., Cao, D. A., & Nghiem, S. H. (2021). Spread of COVID-19 and policy responses in Vietnam: An overview. International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases, 103, 157–161. https://doi.org/10.1016/j.ijid.2020.11.154

[2] Dimitsaki, S., Gavriilidis, G. I., Dimitriadis, V. K., & Natsiavas, P. (2023). Benchmarking of Machine Learning classifiers on plasma proteomic for COVID-19 severity prediction through interpretable artificial intelligence. Artificial intelligence in medicine, 137, 102490. https://doi.org/10.1016/j.artmed.2023.102490

[3] Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet of Things, 11, 100222. https://doi.org/10.1016/j.iot.2020.100222

[4] Chahar, S., & Roy, P. K. (2022). COVID-19: A Comprehensive Review of Learning Models. Archives of computational methods in engineering : state of the art reviews, 29(3), 1915–1940. https://doi.org/10.1007/s11831-021-09641-3

[5] Ghoshal B, Tucker A (2020) "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," arXiv:2003.10769