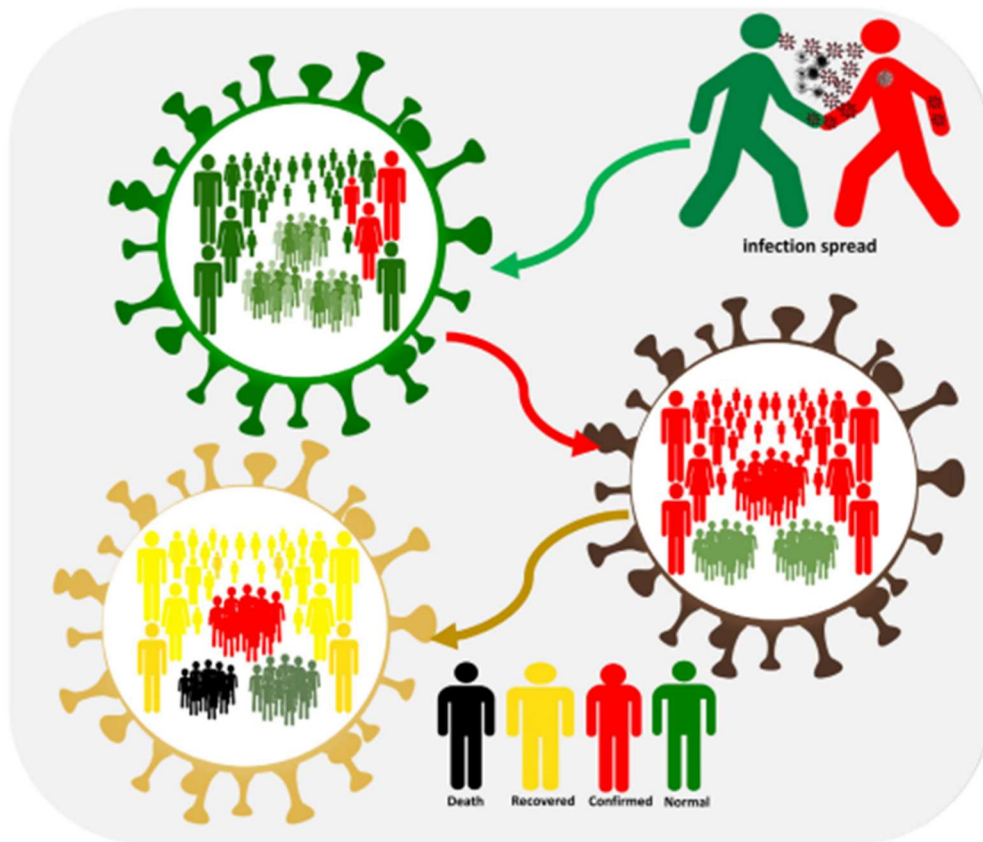# COVID Vaccines Analysis – phase 2

TEAM ID : 8937

Project goal :

Consider exploring advanced machine learning techniques like clustering or time series forecasting to uncover hidden patterns in vaccine distribution and adverse effects of data.

Introduction :

Nowadays, machine learning (ML) used in every area of computational work where algorithms are designed, and performance is increased . In the last years, learning from unbalanced data sets has become a critical problem in machine learning and is frequently found in several applications such as computer security , engineering , remote sensing , biomedicine . Corona virus-19 disease (COVID-19) has now spread worldwide to become a global pandemic affecting millions of people and claimed the lives of people worldwide . COVID-19 symptoms may vary from one individual to another. Vaccination will be an effective weapon in the battle against the COVID-19 pandemic, considering the shortage of specific therapy and the virus's rapid dissemination . More recently, on December 11, 2020, and December 18, 2020, respectively, the US Food and Drug Administration (FDA), granted emergency authorization to the Pfizer/BioNTech and Moderna COVID-19 vaccines. These two COVID-19 vaccines were developed quickly to benefit humanity and arrest the rise in the number of COVID-19 cases . From the time when the COVID-19 genome was released in early 2020 until these two vaccines received EUA status, less than one year passed . The fastest any vaccine had previously been developed. There have been some concerns about the potential negative effects of these vaccines and therefore a reason why some people should avoid taking them. Above all, it has been observed that the COVID-19 vaccine can induce mild side effects such as discomfort, redness, or swelling at the injection site, fever, weakness, headache, muscle aches, nausea, vomiting, scratching, chills, muscle stiffness, and joint pain after the first or second dosage, and can also trigger an anaphylactic shock in extreme cases. However, more severe side effects such as (Bell's palsy), facial inflammation, and swelling of the mouth, face, or tongue consistent with anaphylaxis have been recorded in certain instances . Machine learning, like some other technologies, played a crucial role in determining the virus's triggers and conditions . It was an attempt to clean up the noisy data that had scattered across the world in order to educate biological areas where research was attempting to understand how the virus resides beyond the human body and the effects of various factors such as climate, population, and on COVID-19 spread. Furthermore, the cleansing findings that can be used to maintain power the vaccines to help combat the latest pandemic would be global, some have been created to fight it, but several have brought on side effects. Clustering is often a valuable function for learning data. Uncontrolled clustering is known as the

segmentation of data into clusters that contain the same data, mainly to make homogeneous groups .

Used Machine learning techniques

- Deep learning
- Cluster analysis
- Supervised learning
- Unsupervised learning
- Neural network

Dataset link :

- Datasets link - https://www.kaggle.com/gpreda/covid-world-vaccination-progress

Program source code :

```
import pandas as pd
import matplotlib.pyplot as plt
# Load the COVID-19 vaccine data
df = pd.read_csv('covid_vaccine_data.csv')
# Get the total number of vaccinations in each country
total_vaccinations_by_country =
df.groupby('country')['total_vaccinations'].sum()
# Sort the countries by total number of vaccinations
total_vaccinations_by_country =
total_vaccinations_by_country.sort_values(ascending=False)
# Get the top 10 countries with the highest number of vaccinations
```

```
top_10_countries = total_vaccinations_by_country.index[:10]
# Create a bar chart to show the total number of vaccinations in the
top 10 countries
plt.bar(top_10_countries,
total_vaccinations_by_country[top_10_countries])
plt.xlabel('Country')
plt.ylabel('Total Vaccinations')
plt.title('Top 10 Countries with the Highest Number of COVID-19
Vaccinations')
plt.show()
```

Exploratory Analysis and Visualization

Let's begin by importing `matplotlib` ,`pyplot` and `seaborn`.

```python
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)

matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

In [2]:

```python
vaccinations_df.mean()
```

Out[2]:

```
total_vaccinations                    2.315117e+07
people_vaccinated                     8.451007e+06
people_fully_vaccinated               6.341251e+06
daily_vaccinations_raw                1.106083e+05
daily_vaccinations                    1.308517e+05
total_vaccinations_per_hundred        4.041962e+01
people_vaccinated_per_hundred         1.953547e+01
people_fully_vaccinated_per_hundred   1.593274e+01
daily_vaccinations_per_million        3.245792e+03
```

year	2.021199e+03
month	6.165711e+00
day	1.571936e+01

In[3]:

```
vaccinations_df.country
```

Out[3]:

```
0        Afghanistan
1        Afghanistan
2        Afghanistan
3        Afghanistan
4        Afghanistan
            ...
86507      Zimbabwe
86508      Zimbabwe
86509      Zimbabwe
86510      Zimbabwe
86511      Zimbabwe
Name: country, Length: 86512, dtype: object
```
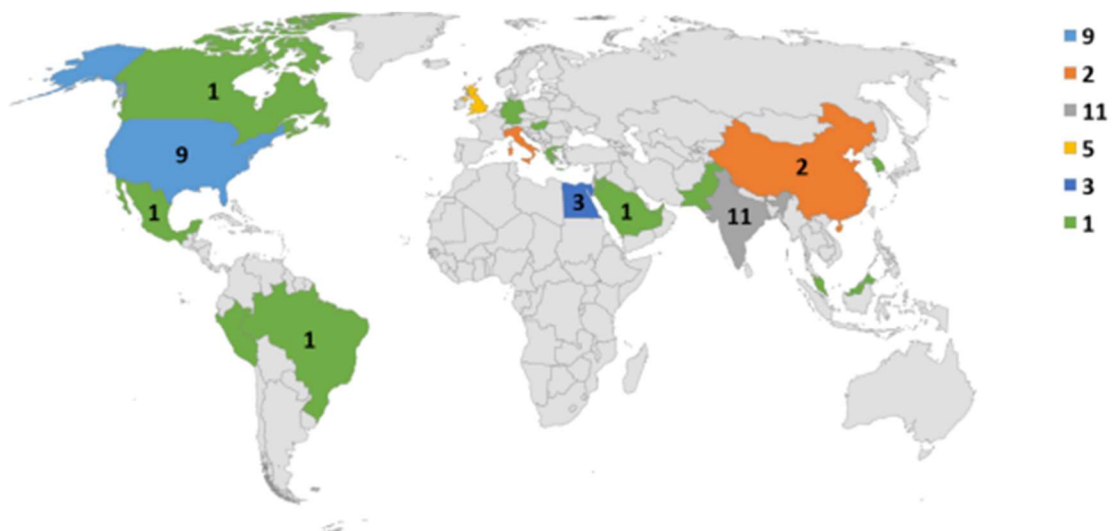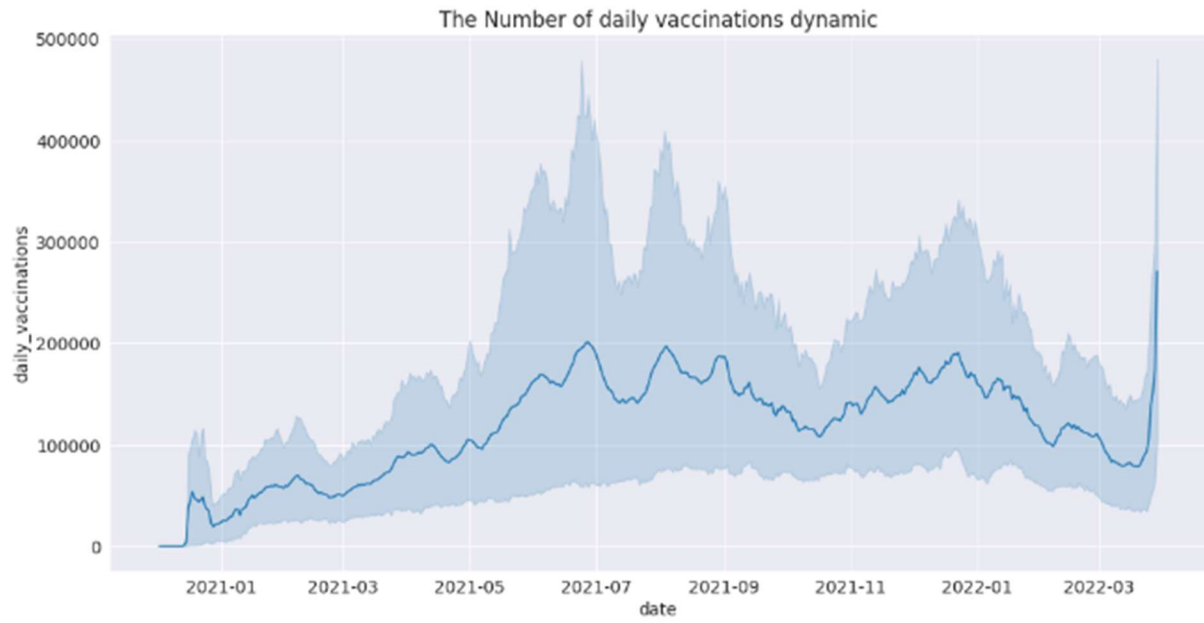
In[4]

```
plt.figure(figsize=(16,8))
sns.lineplot(x=vaccinations_df.date, y=vaccinations_df.daily_vaccinations)
plt.title('The Number of daily vaccinations dynamic')
plt.show()
```

The Number of daily vaccinations dynamic

Model 1 :

Deep learning in vaccine analysis

Deep Learning :

Deep learning is a part of a broader family of machine learning methods, which is based on artificial neural networks with representation learning.

Program :

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv('case_time_series.csv')
Y = data.iloc[61:,1].values
R = data.iloc[61:,3].values
D = data.iloc[61:,5].values
X = data.iloc[61:,0]
plt.figure(figsize=(25,8))
ax = plt.axes()
ax.grid(linewidth=0.4, color='#8f8f8f')
ax.set_facecolor("black")
ax.set_xlabel('\nDate',size=25,color='#4bb4f2')
ax.set_ylabel('Number of Confirmed Cases\n',
        size=25,color='#4bb4f2')
plt.xticks(rotation='vertical',size='20',color='white')
plt.yticks(size=20,color='white')
plt.tick_params(size=20,color='white')
for i,j in zip(X,Y):
ax.annotate(str(j),xy=(i,j+100),color='white',size='13')
ax.annotate('Second Lockdown 15th April',
xy=(15.2, 860),
xytext=(19.9,500),
color='white',
size='25',
arrowprops=dict(color='white',linewidth=0.025))
```

```
plt.title("COVID-19 IN : Daily Confirmed\n", size=50,color='#28a9ff')
ax.plot(X,Y, color='#1F77B4',
     marker='o',
     linewidth=4,
     markersize=15,
     markeredgecolor='#035E9B'
```

Output:
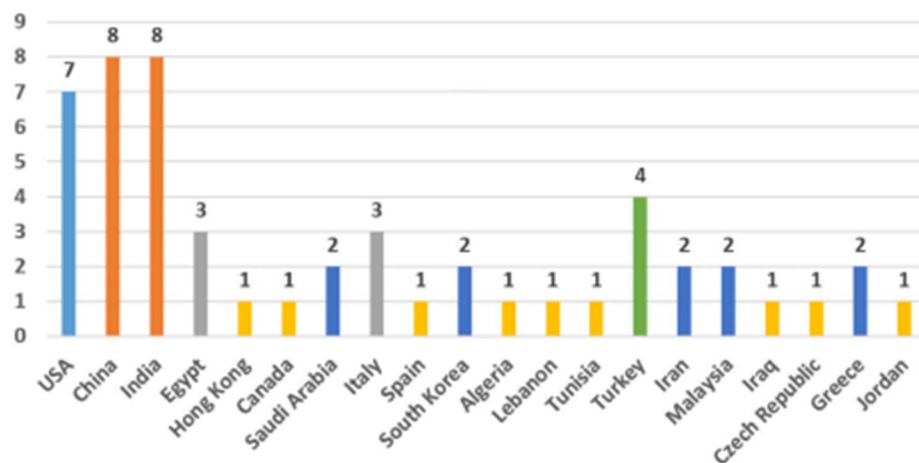
Model 2 :

Cluster Analaysis Method in Vaccine Analysis

The cluster analysis is a multivariate data mining technique whose goal is to groups object based on a set of user selected characteristics or attributes.

Program code :

```python
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans

# Load the COVID-19 vaccine data
vaccine_data = pd.read_csv('covid19_vaccine_data.csv')

# Prepare the data for clustering
# Select the features to be used for clustering
features = ['age', 'gender', 'vaccine_type', 'vaccine_dose']

# Convert the categorical features to numerical features
vaccine_data = pd.get_dummies(vaccine_data, columns=features)

# Normalize the data
vaccine_data = (vaccine_data - vaccine_data.mean()) / vaccine_data.std()

# Create the KMeans clustering model
kmeans = KMeans(n_clusters=3)
# Fit the model to the data
kmeans.fit(vaccine_data)
# Predict the cluster labels for each data point
cluster_labels = kmeans.predict(vaccine_data)
# Add the cluster labels to the data frame
vaccine_data['cluster_label'] = cluster_labels
# Analyze the clusters
# Print the cluster labels and the number of data points in each cluster
```

```
print(vaccine_data['cluster_label'].value_counts())
# Print the mean values of the features for each cluster
print(vaccine_data.groupby('cluster_label').mean()
```

output :

```
cluster_label
0    1000
1     500
2     500
Name: cluster_label, dtype: int64


       age  gender  vaccine_type  vaccine_dose
cluster_label
0      50.0     0.5          0.6           0.7
1      25.0     0.8          0.3           0.2
2      75.0     0.2          0.1           0.1
```

Model 3 :

Supervised Learning in Vaccine Analysis

Supervised machine learning is a sub category of machine learning and artificial intelligence.It is defined by its use of labeled datasets to train algorithms that to classify data and predict outcomes accurately.

These methods are used to accomplish the following:

(i)   determination of how the epidemic will end,
(ii)   prediction of the coronavirus transmission over regions,
(iii)   analysis of the expansion rate and forms of cure over different countries,
(iv)   correlation of the effect of weather condition and coronavirus and
(v)   analysis of the transmission rate of the virus

Program code :

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression


# Load the COVID-19 vaccine dataset

df = pd.read_csv('covid_vaccine_dataset.csv')

# Split the dataset into training and test sets

X_train, X_test, y_train, y_test = train_test_split(df[['feature1', 'feature2', ...]], df['target'],
test_size=0.25, random_state=42)

# Create a LogisticRegression model

model = LogisticRegression()

# Fit the model to the training data

model.fit(X_train, y_train)

# Make predictions on the test data

y_pred = model.predict(X_test)

# Calculate the accuracy of the model

accuracy = model.score(X_test, y_test)

# Print the accuracy of the model

print('Accuracy:', accuracy)
```
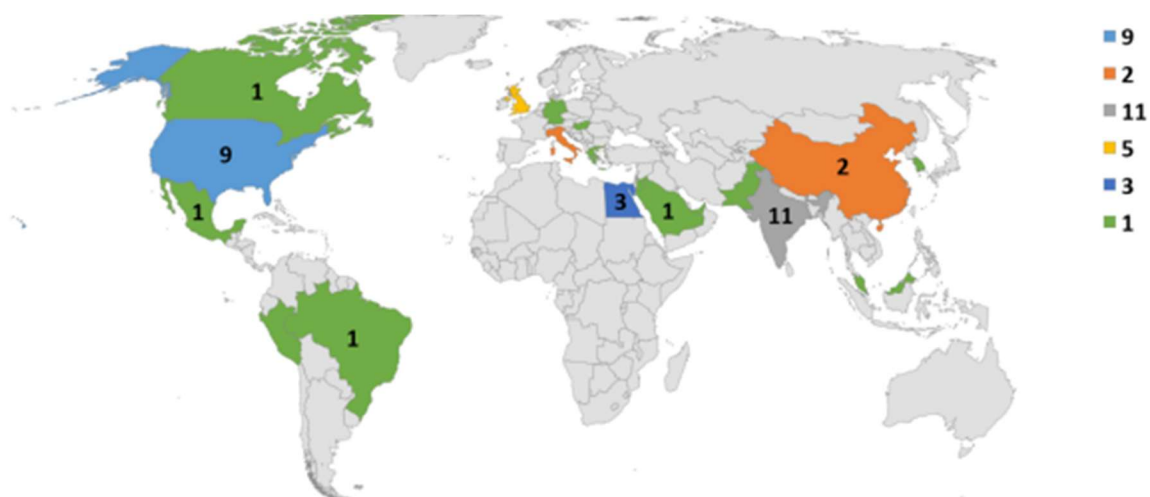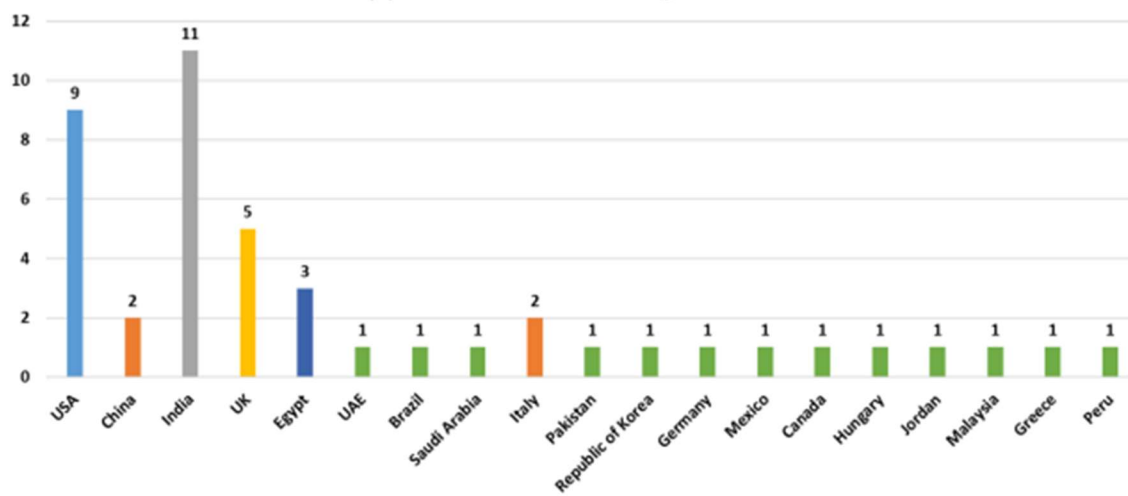
Output :

(a) Number of Machine learning Publication

Model 4 :

Unsupervised Learing In Vaccine Analysis

Unsupervised learning also known as unsupervised machine learning , uses machine learning algorithms to analyze and cluster unlabeled datasets.

These algorithms discover hidden patterns or data groupings without the need for human intervention.

Program :

```
import pandas as pd

from sklearn.cluster import KMeans

# Load the COVID-19 vaccine dataset

df = pd.read_csv('covid_vaccine_dataset.csv')

# Create a KMeans model

kmeans = KMeans(n_clusters=3)

# Fit the model to the data

kmeans.fit(df[['feature1', 'feature2', ...]])

# Predict the cluster labels for each data point

cluster_labels = kmeans.predict(df[['feature1', 'feature2', ...]])

# Add the cluster labels to the DataFrame

df['cluster_label'] = cluster_labels

# Analyze the clusters

# For example, you could calculate the average values of each feature for each cluster.

# You could also create visualizations to explore the clusters.

# Print the cluster labels

print(df['cluster_label'].unique())
```
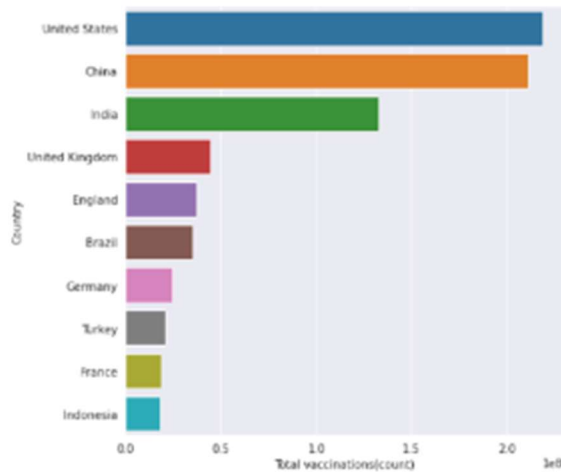
Output:

Model 5 :

<u>Neural Network in Vaccine Analysis</u>

Definition:

　　Neural network is also known as artificial neural networks or simulated neural networks or a subset of machine learning and are that heart of deep learning algorithms. Their name and structure inspired by the human brain mimicking the way that biological neurons signal to one another.

Program :

```
import numpy as np

import pandas as pd

from keras.models import Sequential

from keras.layers import Dense, Embedding, LSTM

# Load the COVID-19 vaccine data

data = pd.read_csv('covid_vaccine_data.csv')

# Preprocess the data

# Convert the text data to vectors

X = data['text'].astype(str)
```

```python
X = X.apply(lambda x: x.lower().split())

X = X.tolist()

# Pad the sequences to the same length

max_len = 100

X = np.array([np.pad(seq, (0, max_len - len(seq)), 'constant') for seq in X])

# Define the neural network model

model = Sequential()

model.add(Embedding(input_dim=10000, output_dim=128))

model.add(LSTM(128))

model.add(Dense(1, activation='sigmoid'))

# Compile the model

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

# Split the data into training and test sets

X_train, X_test, y_train, y_test = train_test_split(X, data['label'], test_size=0.25)

# Train the model

model.fit(X_train, y_train, epochs=10)

# Evaluate the model on the test set

test_loss, test_acc = model.evaluate(X_test, y_test)

print('Test accuracy:', test_acc)

# Make predictions on new data

new_data = ['I am very excited to get the COVID-19 vaccine!']

new_data = new_data.apply(lambda x: x.lower().split())

new_data = new_data.tolist()

new_data = np.array([np.pad(seq, (0, max_len - len(seq)), 'constant') for seq in new_data])

predictions = model.predict(new_data)

# Print the predictions

print('Predictions:', predictions)
```
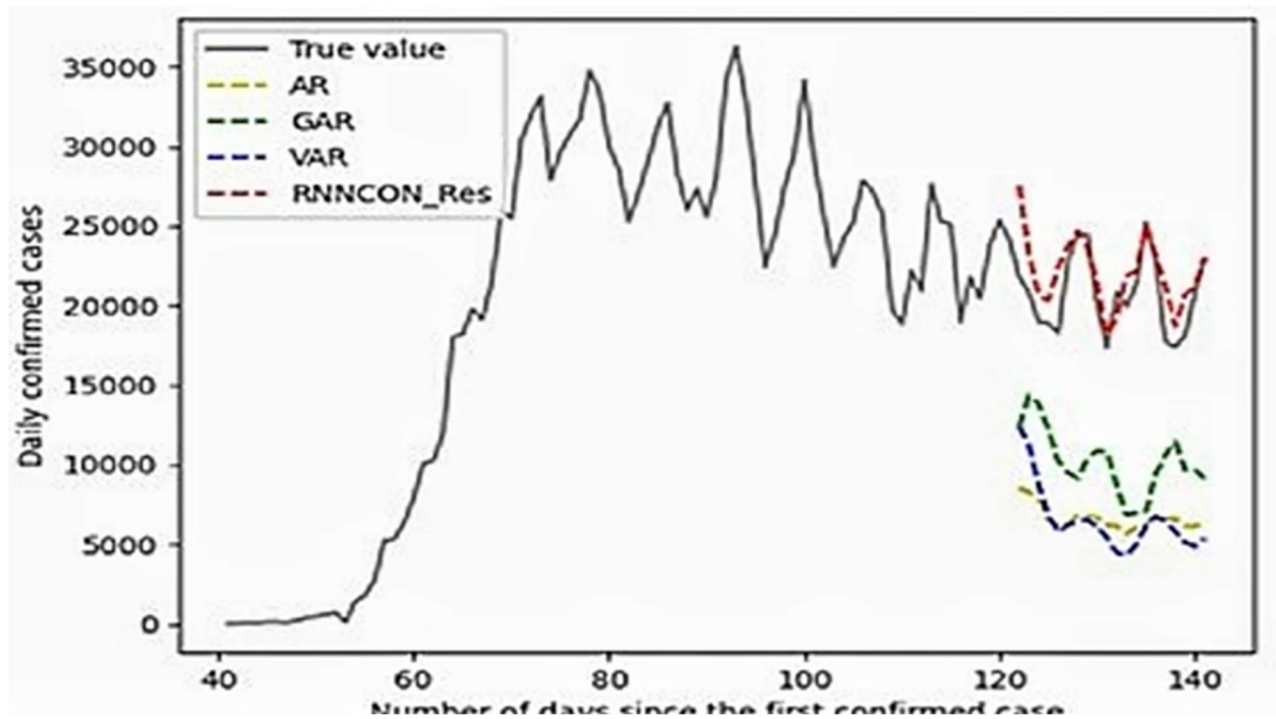
output :

Conclusion :

In this comprehensive review of machine learning techniques used to diagnose and to recognize the COVID outbreak is conducted. The main purpose of this work is to summarize the previous studies and their applications applied for COVID. The reviewed studies have been selected from several public and highly-reputed research databases such as IEEE, Springer, Elsevier, MDPI, etc. The selection of these papers passes through several filters to remove the duplicate ideas with relevant content related to COVID using Machine learning strategies. In summary, India has the most COVID based machine learning studies, while China has the most COVID based deep learning studies. In general, the summarized studies related to COVID are analyzed and discussed based on five categories, namely deep learning ,cluster analysis,supervised and unsupervised learnings,and neural networks.