Machine Learning and Deep Learning Based Time Series Prediction and Forecasting of Ten Nations' COVID- 19 Pandemic

- **Country** - this is the country for which the vaccination information is provided;
- **Country ISO Code** - ISO code for the country;
- **Date**- date for the data entry; for some of the dates we have only the daily vaccinations, for others, only the (cumulative) total;
- **Total number of vaccin ations** - this is the absolute number of total immunizations in the country;
- **Total number of people vaccinated** - a person, depending on the immunization scheme, will receive one or more (typically 2) vaccines; at a certain moment, the number of vaccination might be larger than the number of people;
- **Total number of people fully vaccinated** - this is the number of people that received the entire set of immunization according to the immunization scheme (typically 2); at a certain moment in time, there might be a certain number of people that received one vaccine and another number (smaller) of people that received all vaccines in the scheme;
- **Daily vaccinations (raw)** - for a certain data entry, the number of vaccination for that date/country;
- **Daily vaccinations** - for a certain data entry, the number of vaccination for that date/country;
- **Total vaccinations per hundred** - ratio (in percent) between vaccination number and total population up to the date in the country;
- **Total number of people vaccinated per hundred** - ratio (in percent) between population immunized and total population up to the date in the country;
- **Total number of people fully vaccinated per hundred** - ratio (in percent) between population fully immunized and total population up to the date in the country;
- **Number of vaccinations per day** - number of daily vaccination for that day and country;
- **Daily vaccinations per million** - ratio (in ppm) between vaccination number and total population for the current date in the country;
- **Vaccines used in the country** - total number of vaccines used in the country (up to date);
- **Source name** - source of the information (national authority, international organization, local organization etc.);
- **Source website** - website of the source of information;

Content:

In [1]:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the /python Docker image/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all f
iles under the input directory
```

```
import matplotlib.pyplot as plt

# plotly
# import plotly.plotly as py
from plotly.offline import init_notebook_mode, iplot, plot
import plotly.express as px
import plotly as py
init_notebook_mode(connected=True)
import plotly.graph_objs as go

from pandas_profiling import ProfileReport
import scipy

# seaborn library
import seaborn as sns

# word cloud library
from wordcloud import WordCloud

import os
for dirname, _, filenames in os.walk('/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/working/) that gets preserved a
s output when you create a version using "Save & Run All"
# You can also write temporary files to /temp/, but they won't be saved outside of th
e current session
```

```
/input/covid-world-vaccination-progress/country_vaccinations_by_manufacturer.csv
/input/covid-world-vaccination-progress/country_vaccinations.csv
```

In [2]:

```
data = pd.read_csv("/input/covid-world-vaccination-progress/country_vaccinations.csv")
data.head()
```

Out[2]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred	daily_vaccinations_per_million	vaccines	source_name	source_website
0	Afghanistan	AFG	2021-02-	0.0	0.0	NaN	NaN	NaN	0.0	0.0	NaN	NaN	Johnson &Johnson, Oxford/AstraZ	World Health Orga	https://covid19.who.int/

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred	daily_vaccinations_per_million	vaccines	source_name	source_website
			22										enecа, Pfizer/Bi...	nization	
1	Afghanistan	AFG	2021-02-23	NaN	NaN	NaN	NaN	1367.0	NaN	NaN	NaN	34.0	Johnson&Johnson, Oxford/AstraZeneca, Pfizer/Bi...	World Health Organization	https://covid19.who.int/
2	Afghanistan	AFG	2021-02-24	NaN	NaN	NaN	NaN	1367.0	NaN	NaN	NaN	34.0	Johnson&Johnson, Oxford/AstraZeneca, Pfizer/Bi...	World Health Organization	https://covid19.who.int/
3	Afghanistan	AFG	2021-02-2	NaN	NaN	NaN	NaN	1367.0	NaN	NaN	NaN	34.0	Johnson&Johnson, Oxford/AstraZenec	World Health Organiz	https://covid19.who.int/

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_vaccinated_per_hundred	people_fully_vaccinated_per_hundred	daily_vaccinations_per_million	vaccines	source_name	source_website
			5										a, Pfizer/Bi...	ation	
4	Afghanistan	AFG	2021-02-26	NaN	NaN	NaN	NaN	1367.0	NaN	NaN	NaN	34.0	Johnson&Johnson, Oxford/AstraZeneca, Pfizer/Bi...	World Health Organization	https://covid19.who.int/

In [3]:

```
report = ProfileReport(data)
report
```

Abstract

In the paper, the authors investigated and predicted the future environmental circumstances of a COVID-19 to minimize its efects using artifcial intelligence techniques. The experimental investigation of COVID-19 instances has been performed in ten countries, including India, the United States, Russia, Argentina, Brazil, Colombia, Italy, Turkey, Germany, and France

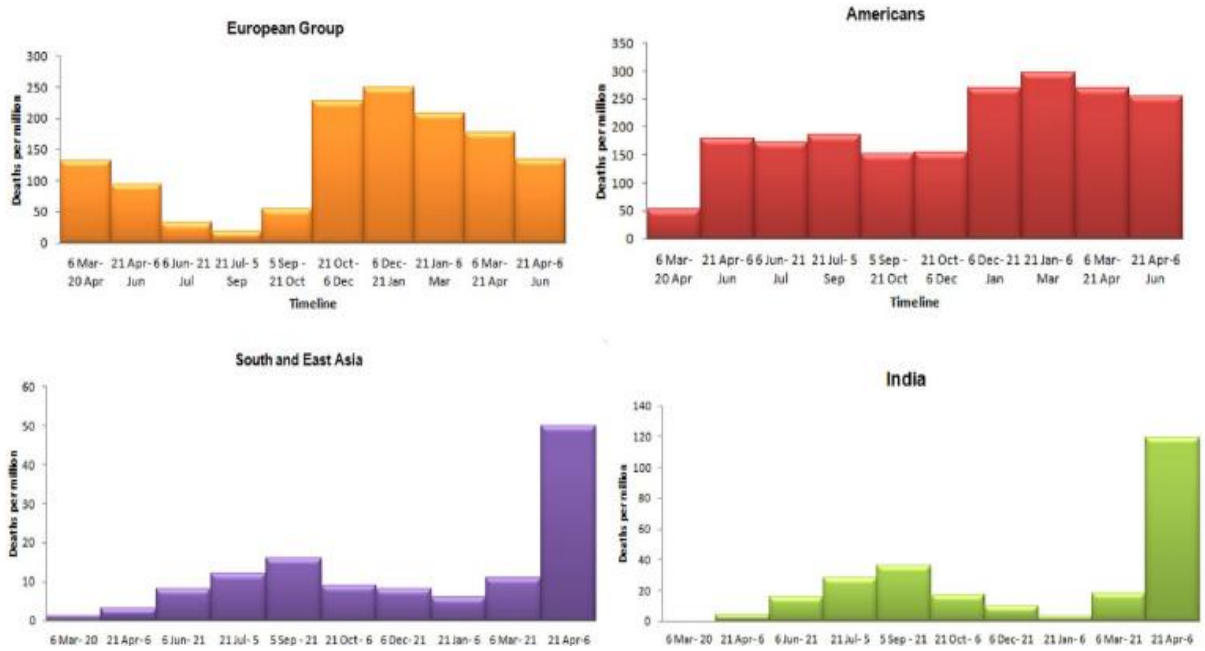
using machine learning, deep learning, and time series models. The confrmed, deceased, and recovered datasets from January 22, 2020, to May 29, 2021, of Novel COVID-19 cases were considered from the Kaggle COVID dataset repository. The country-wise Exploratory Data Analysis visually represents the active, recovered, closed, and death cases from March 2020 to May 2021. The data are pre-processed and scaled using a MinMax scaler to extract and normalize the features to obtain an accurate prediction rate. The proposed methodology employs Random Forest Regressor, Decision Tree Regressor, K Nearest Regressor, Lasso Regression, Linear Regression, Bayesian Regression, Theilsen Regression, Kernel Ridge Regressor, RANSAC Regressor, XG Boost, Elastic Net Regressor, Facebook Prophet Model, Holt Model, Stacked Long Short-Term Memory, and Stacked Gated Recurrent Units to predict active COVID-19 confrmed, death, and recovered cases. Out of diferent machine learning, deep learning, and time series models, Random Forest Regressor, Facebook Prophet, and Stacked LSTM outperformed to predict the best results for COVID-19 instances with the lowest root-mean-square and highest R_2 score values.

Keywords COVID-19 · Prediction · XG Boost · Facebook Prophet · Holt model · Stacked gated recurrent units · RANSAC regressor · Random forest regressor · Stacked long short-term memory

Introduction

Throughout history, the world has confronted several major pandemic and epidemic problems. The frst recorded pandemic occurred in Athens during the Peloponnesian War in 430 BC, followed by the Antonine Plague in 165 A.D., in 250 A.D.—the Cyprian Plague, in 541 A.D.—the Justinian Plague, in the eleventh century—leprosy, in 1350—The Black Death, in 1492—The Columbian Exchange, in 1665— The Great Plague of London, in 1817—The First Cholera Pandemic, in 1855—The Third Plague Pandemic, in 1875— Fiji Measles Pandemic, in 1889—Russian Flu, in 1918— Spanish Flu, in 1957—Asian Flu, in 1981—HIV/AIDS, in 2003-SARS, and 2019—COVID-19 [1]. While still a public health concern, Coronavirus 19 (also known as COVID-19) is an infectious sickness that occurred by the severe acute respiratory syndrome coronavirus 2. The frst recorded case of SARS (severe acute respiratory syndrome) was identifed in December of 2019 in Wuhan, China. The disease has since spread to many other nations and healthcare systems worldwide. At the same time, humans inhale contaminated air, including airborne droplets and particles that are smaller than 0.1 microns, and COVID-19 spreads [2]. Inhalation of these particles is more dangerous when people are closely packed together; nevertheless, they can be inhaled further apart, especially indoors. Infected fuids sprayed on the skin, in the eyes, nose, or mouth, or on surfaces contaminated with them may result in transmission. Someone can carry and spread the disease for up to 20 days even if they have no symptoms. During COVID19,afrstwavebeganinthe spring, which receded signifcantly throughout the summer, and a second wave appeared in the fall of 2020. The initial wave of the epidemic devastated several nations, and many patients perished. The severity of this early

phase was exacerbated by a lack of specialist equipment and a lack of understanding of the disease [4]. We all learned from our mistakes during the frst wave of the pandemic, and as a result, our confdence in being able to handle the second wave much better was strong. Despite this, the second wave had considerably greater infection rates, more patients in ICUs, and, in certain countries, more fatalities [5]. Figure 1 depicts the death rates from March 6, 2020, to June 6, 2021, with Europe and the Americas having the most signifcant mortality rates compared to India and South and East Asia. Europe had 1,172,912 death cases, the Americas had 1,926,520, South and East Asia had 739,802 death cases, and India had 402,728 COVID death cases as of July



Related Work

Since 2020, researchers have made signifcant attempts to anticipate the onset of COVID illness in people or the end of the disease around the globe. Keeping this in mind, Shastri et al. [1] suggested a deep learning-based model, such as a recurrent neural network, to forecast the future circumstances of new coronaviruses by studying instances from India and the United States. Ten diferent nations with the most signifcant number of verifed cases were investigated. It was shown that the predictive accuracy of a range of six separate time series modeling approaches for coronavirus epidemic detection varied by Papastefanopoulos et al. [2]. Using an LSTM model, Chimmula et al. [3] predicted the end of the COVID-19 pandemic and worldwide epidemics due to antiviral drugs and improved access to healthcare. Indicating the date of the pandemic's demise, the writers anticipate that it will be fnished by June of 2020. Using a deep learning model, Togacar et al. [4] identifed coronavirus in datasets containing instances of pneumonia, as well as standard X-ray imaging data. The COVID-19 disease can diagnosedwith 99.27% accuracy with the model that the authors used. COVID-19 drug and vaccine research achievements were evaluated using artifcial intelligence techniques in a recent study by Arshadi et al. [5]. In addition, the scientists gave information about the compounds, peptides, and

epitopes in the CoronaDB-AI library, which were discovered both in silico and in vitro. Categorizing chest X-rays into two groups was proposed by the researchers led by Elaziz et al. [6]. The accuracy percentage for the frst and second datasets was 96.09% and 98.09%, respectively. Alimadadi et al. [7] presented a Alaska et al. [9] evaluated the efcacy of deep learning models in predicting COVID-19 illness using laboratory data from 600 patients and got 91.89% accuracy. Their approach was also utilized to help medical professionals validate test data and for clinical prediction research. The Johns Hopkins dashboard data, which were the primary source of the Punn et al.'s [10] research, were utilized with machine learning and deep learning models. The team's goal was to grasp the exponential growth of the COVID-19 and then make predictions about how widespread it may become across the country. Table 1 on the left shows the researchers who worked on the forecast and detection of COVID-19

Contribution Outline

The overall goal of this research is to build models that can calculate two necessary evaluative measures: RMSE and R_2 Score for confrmed, death, and recovered cases from ten diferent nations to help future forecasts. The steps are as follows:

Step 1: Initially, data are pre-processed to capture char

acteristics utilizing various variables, such as active cases,

recovered cases, and COVID-19 fatality cases.

Step 2: Exploratory Data Analysis of COVID-19's active

cases, closed cases, confrmed cases, recovered cases, and

death cases are calculated to summarize or interpret the

information that is hidden in rows or columns, and scaling

techniques such as Min–Max have been applied to normalize

each feature that is obtained from these attributes.

Step 3: Later, utilizing confrmed cases, recovered cases,

and death cases from 10 diferent nations, the gathered data

were used to anticipate the future conditions of a new Coro

naVirus. To get the fndings, several machine learning mod

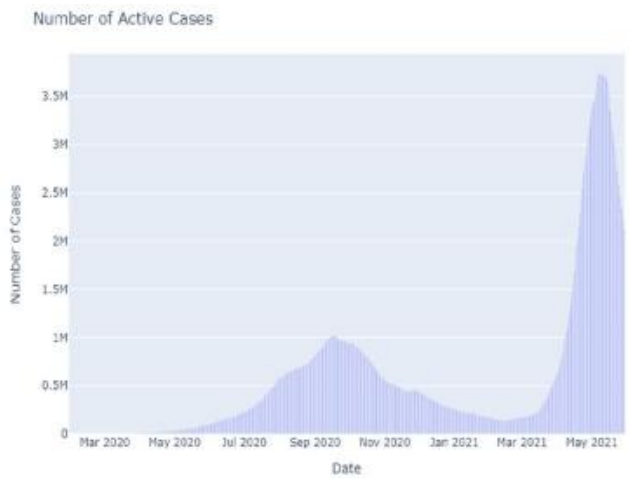
els, time series models, and deep learning models were used,

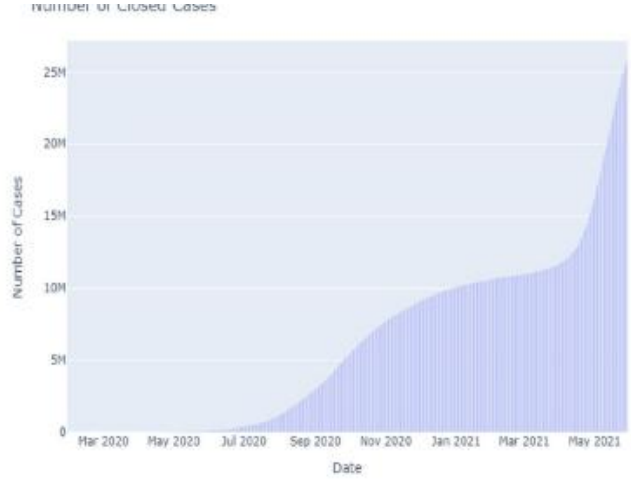
alysis of the existing work

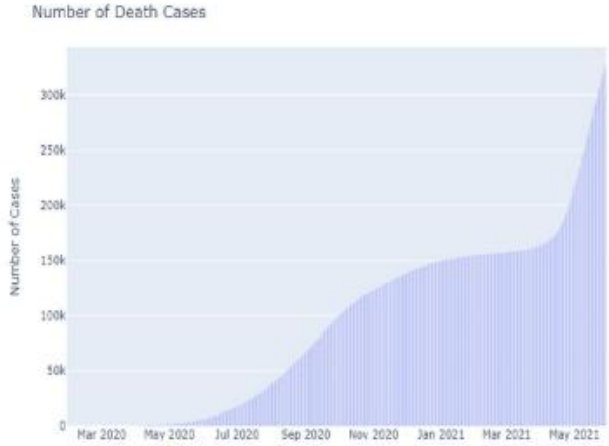
me	Dataset	Technique	Results	Limitations
[11]	1065 CT pathogenic images	Transfer Learning Model, CNN, Graph-Net	Accuracy: 89.5% Specificity: 0.88 Sensitivity: 0.87	Factors such as low signal-to-noise ratio and complex data integration led to reducing the efficacy of deep learning models
hyay et al. [12]	Data collected from WHO (Jan. 16–20,2020)	Long Short-Term Memory, Gated Recurrent Unit	Accuracy: 87%	The model failed to represent the spatio temporal components of the LSTM network
al. [4]	Data collected from Qatar University	Stacking Technique, Fuzzy Color, Deep Learning Model	Classification accuracy: 99.27%	Publications of COVID-19 images were limited. The system did not work with the low resolution and different size input images
. [1]	Dataset was sourced from the Ministry of Health and Family Welfare	Deep Neural Network, Long Short-Term Memory, Recurrent Neural Network, Polynomial Regression	Accuracy ConvLSTM: 98%	The comparative analysis had been performed only for two countries
al. [13]	COVID-19 chest X-ray dataset	Bayesian Deep Learning	Accuracy: 80%	After reviewing the data, it was impossible to conclude anything regarding markers for imaging, discoveries concerning improved diagnosis and therapy for COVID-19
[10]	Data collected from Jan 22, 2020 to Apr 1 2020 at Johns Hopkins University	Support Vector Machine, Deep Neural Network, Long Short-Term Memory, Polynomial Regression	RMSE confirmed: 455.92 Death: 117.94 Recovered: 809.71	The study needed to work on more algorithms to enhance the RMSE score
. [9]	Samples collected from the Albert Einstein Israelite Hospital in Sao Paulo, Brazil	Artificial Neural Network, Convolution Neural Network, Long Short-Term Memory	Accuracy: 86.66% F1 Score: 91.89% Recall: 99.42% AUC: 62.50% Precision: 86.75%	The primary disadvantage of the study was the sheer amount of data. To increase the number of patients for whom the lab findings could not be assessed, the procedure was applied on 600 patients
. [14]	180 COVID-19 and 200 chest X-ray images	CNN model, SVM, ResNet50	Accuracy: 91.6%	The study needed to incorporate work on different imagistic patterns of COVID-19
. [15]	337 patient images from real-world data	Deep learning, nCOVnet	Accuracy: 97.62%	The system worked on a small dataset
[6]	Dataset collected from Joseph Paul Cohen and Paul Morrison Lan Dao	Manta Ray Foraging Optimization, Fractional Multichannel Exponent Moments	Accuracy: 96.09% Accuracy: 98.09%	The system dealt with resource limitations and high CPU time

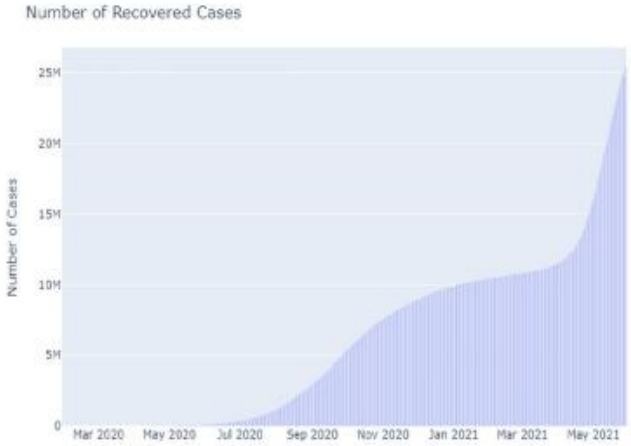
Table 2 Analysis of COVID-19 cases among the top ten countries

Countries	Confirmed cases	Death cases	Recovered cases
India	27,894,800	325,972	25,454,320
USA	33,251,939	594,306	–
Russia	4,995,613	118,781	46,16,422
Argentina	3,732,263	77,108	3,288,467
Brazil	16,471,600	461,057	14,496,224
Colombia	3,363,061	87,747	3,141,549
Italy	4,213,055	126,002	3,845,087
Turkey	33,251,939	47,271	5,094,279
Russia	4,995,613	118,781	4,616,422
Germany	3,684,672	88,413	3,479,700









This work employed an exploratory analysis of ten diferent countries after pre-processing to assess its features via statistical graphs. Figures shown below depicts the graphical analysis of active cases, death cases, closed points, and recovered cases that have been recorded from Jan 2020 to May 2021. It wasdetermined in Fig. 3 that 27,894,800 cases had been confrmed, 2,114,508 were still active, 325,972 had died, 25,780,292 had been closed, and 25,454,320 people had been recovered from Jan 2020 to 29 May 2021. Additionally, the numbers of confrmed cases, deaths, and recovered cases each day were, respectively, 57,397, 671, and 52,375. According to Fig. 4, it has been discovered that US has 3,325,189,940 instances with high certainty, 3,266,576,333 cases with moderate certainty, 594,306 cases with low certainty, and 0 cases with a medium certainty which were seen from January 1st, 2020 to May 29th, 2021. Additionally, the daily average of confrmed cases was reported as 673,128, while the daily average of deaths was recorded as 12,030. Finally, the daily average of recovered cases was recorded as 0. As demonstrated in Fig. 5, the numbers of confrmed, active, and death cases have been as follows: 49,956,313.0, 260,410.0, 118,781.0, 47,352,203.0, and 46,164,322.0 from January 1, 2020 to May 29, 2021. Finally, the total number of confrmed cases was 10,300. The number of death cases was 245, and the total number of recovered cases was 9518. In Fig. 6, it was discovered that Argentina has reported 373,263.0 total cases, with 366,688.0 currently active

cases,77,108currentlyknowndeathcases,336,575 previously known to be closed cases, and 328,467 previously known recovered cases from January 1st, 2020 to May 31st, 2021.

In addition to this, there were around 8239.0 confrmed

cases of the disease each day, approximately 170.0 deaths

per day, and approximately 7259.0 recovered cases per day