iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40764

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

COVID-19 Prediction and Detection using Machine Learning and Artificial Intelligence

Aryan Mehtele[1], Aditya Modak[2], Rohit Sanhal[3], Prof. Sujit Tilak[4]
[1, 2, 3]*Department of Information Technology, Pillai College of Engineering*
[4]*Department of Computer Engineering, Pillai College of Engineering*

Abstract: ***This study focuses on the use of Artificial Intelligence and Machine Learning models to predict the possible corona positive cases and detect the spread of corona among individuals by scanning their X-Rays. Linear Regression Model is used to achieve the desired Prediction output. Along with Linear Regression we have used SVR model to attain comparative analysis among different model prediction techniques. Finally a forecasting model Holt's model was developed which seemed to be better than the above two models. Classical ML related Python libraries are extensively used in this project to train and preprocess large amounts of data. In the Detection phase, the Convolutional Neural Network model is used to detect COVID-19 infection among individuals. In the detection phase we have added multiple filtering rounds to achieve the best accuracy. The model is trained upto certain epochs to improve the accuracy and reduce the overall loss. The datasets used for all the above-mentioned models are extracted from github repositories associated with some data collection organisations. This study also includes the global analysis of COVID-19 and displays the trend shown in a graphical representation.***
Keywords: X-Rays, Linear Regression, SVR, Holt's Model, Convolutional Neural Network, epochs

I. INTRODUCTION

As the COVID-19 pandemic continues to grow, work comes to a standstill after a brisk start. The massive rise in cases has completely paralyzed the healthcare system. Detection of COVID-19 becomes the topmost priority for the government to contain this virus from further spreading. Healthcare systems are finding it difficult to keep up with the number of rising cases. Although technological integration in the healthcare field is proving effective. Artificial Intelligence based techniques have achieved the highest rates of accuracy. Readiness of the government decides a nation's fate against any further subsequent threatening waves of COVID-19. This readiness is dependent upon the preciseness of the prediction model. Machine Learning models are able to deliver the most precise results, strengthening the preparation against any further waves. In this project we aim to develop Machine assisted Prediction and Detection models which make extensive use of artificial intelligence algorithms and techniques to provide us with precise and tangible results.

II. LITERATURE SURVEY

A. Deep Learning Based COVID-19 Detection

In this paper they've taken the PA view of chest x-ray scans for covid-19 affected patients as well as healthy patients. After cleaning up the images and applying data augmentation, They've used deep learning- based CNN models and compared their performance.

B. COVID-19 Detection Using Convolutional Neural Network

In this paper they've applied three different models (InceptionV3, Xception, and ResNeXt). The analysis of this collected data is done with the help of CNN. This work mainly focuses on the use of CNN models for classifying chest X-ray images for coronavirus infected patients.

C. COVID-19 Prediction using SEIR model

This paper presents a comparative analysis of machine learning and soft computing models to predict the COVID-19 outbreak as an alternative to susceptible–infected–recovered (SIR) and susceptible-exposed-infectious-removed (SEIR) models. Among a wide range of machine learning models investigated, two models showed promising results (i.e., multi-layered perceptron, MLP; and adaptive network-based fuzzy inference system, ANFIS). Based on the results reported here, and due to the highly complex nature of the COVID-19 outbreak and variation in its behaviour across nations, this study suggests machine learning as an effective tool to model the outbreak. This paper provides an initial benchmarking to demonstrate the potential of machine learning for future research. This paper further suggests that a genuine novelty in outbreak prediction can be realised by integrating machine learning and SEIR models.

D. COVID-19 Prediction using ARIMA model

The objective of the paper is to formulate a simple average aggregated machine learning method to predict the number, size, and length of COVID-19 cases extent and wind-up period crosswise India. This study examined the datasets via the Autoregressive Integrated Moving Average Model (ARIMA). The study also built a simple mean aggregated method established on the performance of 3 regression techniques such as Support Vector Regression (SVR, NN, and LR), Neural Network, and Linear Regression. The results showed that COVID-19 disease can correctly be predicted. The result of the prediction shows that COVID-19 ailment could be conveyed through water and air ecological variables and so preventives measures such as social distancing, wearing of mask and hand gloves, staying at home can help to avert the circulation of the sickness thereby resulting in reduced active cases and even mortality.

TABLE 1
Summary of Literature Survey

Literature	Method	Accuracy
Rachna Jain and Meenu Gupta et al. 2020 [1]	CNN, Xception model	97.97%
Boran Sekeroglu, IIker Ozsahin et al. 2020 [2]	CNN	96.51%
Sina F. Ardabili and Filip Ferdinand et al. 2020 [3]	SEIR Model	96.23%
Boran Sekeroglu, IIker Ozsahin et al. 2020 [4]	ARIMA Model	93.28%

III. SYSTEM ARCHITECTURE

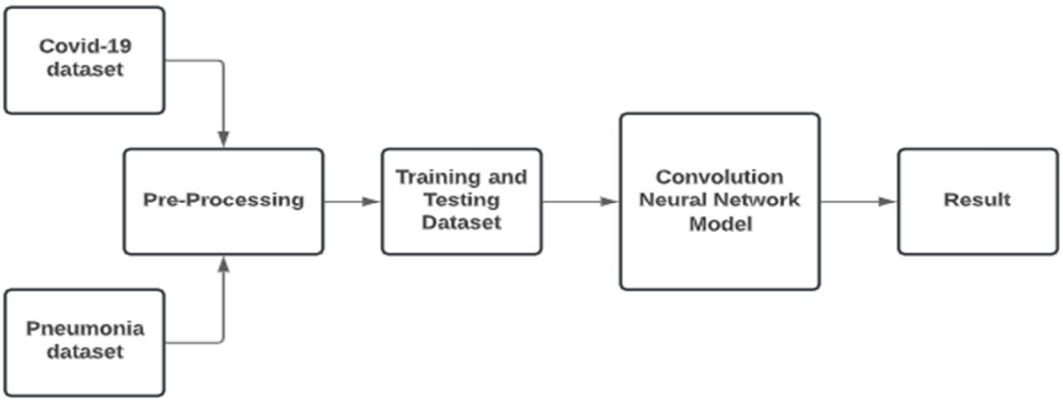


Fig. 1: System Architecture for Detection Phase

A. COVID-19 Dataset Description

First dataset involves collecting important datasets from github. COVID-19 dataset consists of X-Ray images of patients infected with COVID-19. These images are used to identify COVID-19 positive samples.

B. Pneumonia Dataset Description

The second dataset consists of Pneumonia negative patients. COVID-19 shares some similar chest conditions as Pneumonia so this dataset is used to identify COVID-19 negative samples.

C. Pre-Processing Description

Image Pre-processing here involves two phases namely image sorting, which includes merging COVID-19 dataset with pneumonia dataset and digital image processing which includes running of algorithms to perform image processing on digital images. It involves applying a wider range of algorithms to the input data to improve image data by suppressing unwanted distortions so that the CNN model may benefit from this and become more optimised and precise.

D. Training and Testing Data Description

Training dataset consists of thoroughly cleaned data, the success rate of a model depends upon the dataset. Cleaning is one of the major processes in model based projects. The training set is the material through which the computer learns how to process information. The CNN model applies its algorithms on this piece of dataset to generate it's main intelligence. It is this dataset which trains the CNN model to differentiate between COVID-19 positive and negative samples when presented with a chest X-Ray image. Testing dataset evaluates the model's intelligence gained by the training data. It lists important parameters like Accuracy and Loss rate.

E. CNN block Description

This block represents the base model of image detection which is Convolutional Neural Network Model. It makes use of iterative training on each of the images to eventually be able to recognize features, shapes. Multiple filtering techniques are used to scan images using convnets. Repetitive filtering scans are carried out to expel any errors and provide the most accurate results for this sensitive test of COVID-19.

F. Result Block Description

Final result is displayed on the screen about whether the scanned Chest X-Ray of the patient in doubt is coronavirus positive or negative.

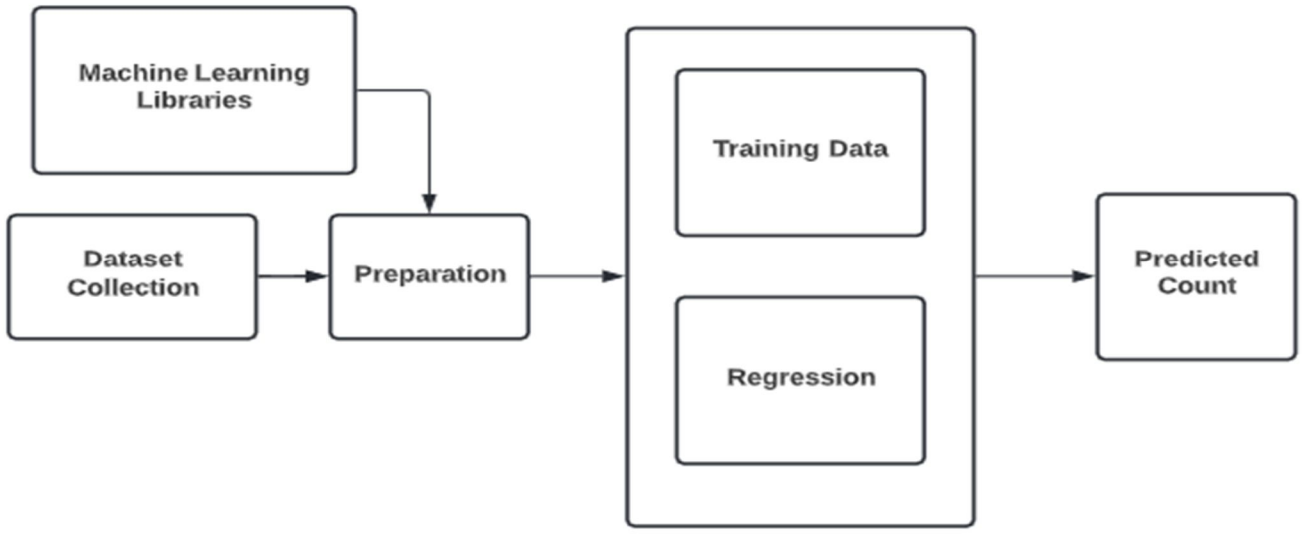


Fig. 2: System Architecture for Prediction

G. Dataset Description

First phase of this prediction model is to collect COVID-19 daily case numerical data. The data needs to be updated regularly for precise predictions.

H. Preparation Description

This phase of the project involves modifying the dataset by using python libraries. Pandas is used to transform the dataset into data frames, numpy is used in crunching large datasets, it is one of the most used libraries by a data scientist, Matplotlib is used to convert data into corresponding x and y coordinates. Using matplotlib a graphical representation of the dataset is generated.

I. Training/Regression Description

In the training phase, the Machine learning model learns using sklearn library and the data which is collected and prepared. A Regression Model is used which is useful when predicting number based problems like probability of an event.

J. Prediction Count

The results are evaluated based on predicted trend and actual trend which provides strong evidence regarding the accuracy and operability of the Model. Model is then tuned up according to the results.

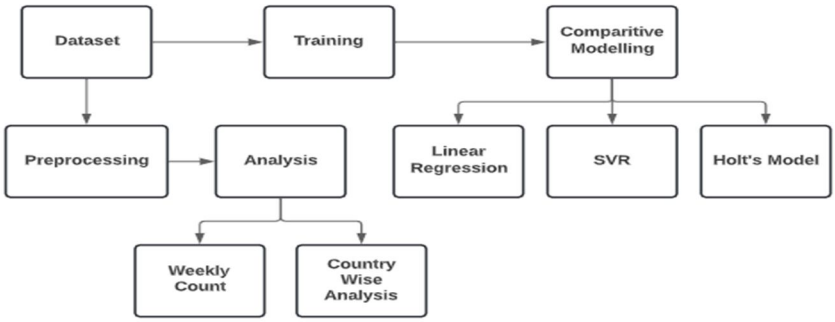


Fig. 3: System Architecture for Analysis and Comparative Study

K. Dataset Description

Dataset consists of three main components namely number of active cases, deaths and recovered cases. It contains records of all the countries making it possible to analyse the COVID-19 situation globally. Dataset is extracted from John Hopkins repository which is updated daily.

L. Preprocessing Description

In this phase the dataset is exploited to extract specific data and group up this data into similar arrays. Arrays are then processed to do a comparative analysis according to the desired output by the developer

M. Analysis Description

Arrays of data created in the preprocessing phase are used in this phase to calculate certain country specific values and time specific values and learn the pattern of outbreak of Coronavirus. Analysis provides a deep knowledge of the trend of the virus in different regions and the growth rate of different countries.

N. Training Description

In this phase python libraries such as numpy, pandas, scikit -learn are used to train the model to provide desirable outputs. Model specific algorithm is the base on which the model is trained.

O. Competitive Modelling Description

This phase focuses on developing alternate predictive models to calculate possible future values. Accuracy and loss are the parameters observed in this phase which decide the success rate of the models developed.

P. Linear Regression Description

Linear Regression models a target prediction value based on independent variables. Due to limitation to linear values, non-linear dataset is forcefully adjusted by linear model giving inaccurate and highly imprecise results.

Q. Support Vector Regression Description

Support Vector regression model gives better results than linear regression model by acknowledging the non-linearity in the dataset. Although the results are slightly better than the Linear regression model, it also deviates from the desired range of values giving imprecise results which don't seem practically genuine.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

R. Holt's Forecasting Model Description

This model is a forecasting model rather than a prediction model. Holt's model uses three exponential smoothing methods to forecast values with or without any trend in the dataset. This gives this forecasting method an edge over others giving us the most precise and accurate results.

IV. REQUIREMENT ANALYSIS

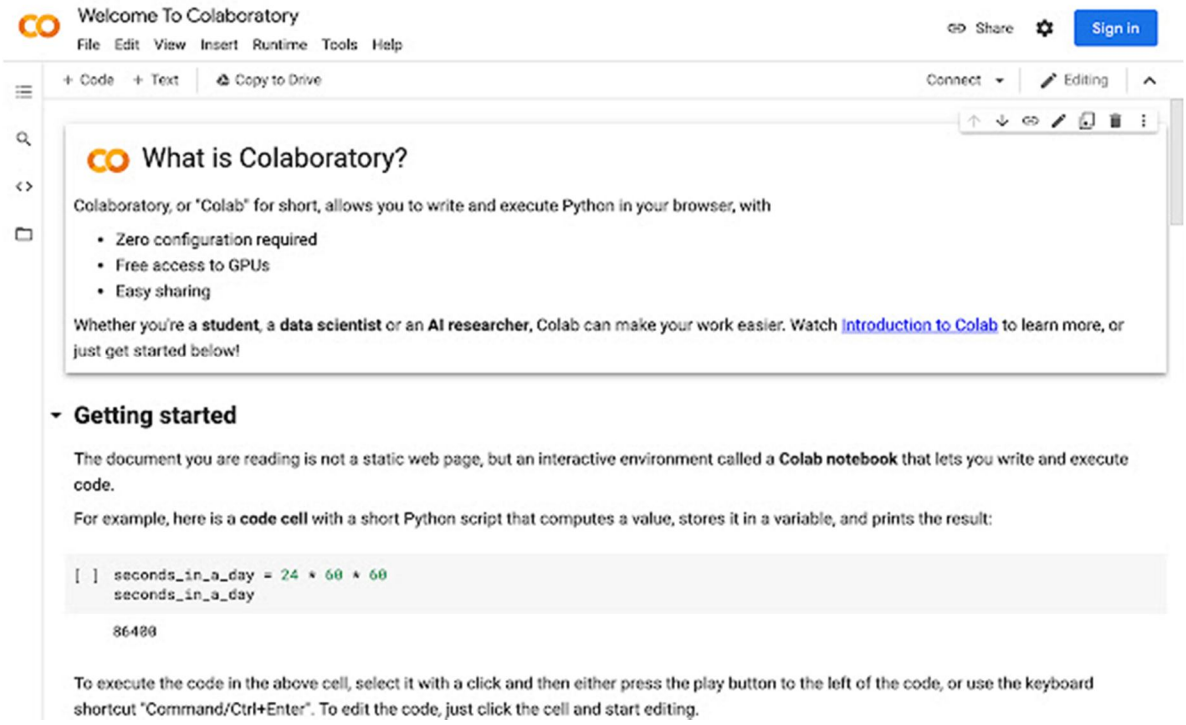


Fig. 4: Google Colab Home page

A. Google Colab

Google Colab is a relatively new software introduced by google which gives the authentic feel of Jupyter Notebook but with some advanced features putting it ahead of its competitors.

Colab allots virtual RAM and CPU to a project making the executions pretty smooth. Colab is the new platform for testing and developing new models quickly and correctly. The main feature of Google Colab is it's support for numerous python libraries and the virtual resources provided by google in the form of RAM and CPU which erases the need of a high performance workspaces for the development of Artificial Intelligence models. The expandable RAM and CPU capacity makes colab usable by everyone from beginners to professionals. It's highly scalable, anytime available along with a Support for a large number of libraries makes it a robust application in today's fast paced computer era.

Developing all the models on Google Colab collaboratively helps to simultaneously work on the models without needing to be constrained to one machine. The connecting to drive feature also enables users to upload datasets directly to drive to be used directly by colab in it's execution code. Thus colab has emerged as a developers new tool in the modern digital age of computing.

V. DATASETS

Datasets for all the models constitute the vital part of the entire project. These are the fundamental part of the entire working of the models. Models are highly dependent on accurate and frequently updated datasets to achieve the required accuracy. COVID-19 and Pneumonia image Datasets consist of numerous X-Rays in two main orientations whereas for prediction we require a real time dataset with proper distinction between active and death count as well as for analysis phase we have used a static dataset to carry out analysis for a certain time period. X-Ray datasets are extracted from Kaggle datasets whereas Covid case counts are extracted from Github repositories associated with reputed data collecting organisations.

TABLE 2
Dataset Information

Dataset	Users	Items	Interactions	Type
COVID-19 Case count	8,12,490	70,747	23,478	Information
COVID-19 patient X Rays	1822	162	344	X-Ray Images
Pneumonia patient X Rays	1,41,00,000	5823	1,58,000	X-Ray images
COVID-19 World Cases	8.45,000	10,40,000	44,500	Information

VI. RESULTS

A. Detection Phase Results

For the detection phase we trained the model for 10 epochs with training and testing dataset and achieved high accuracy and precision. The loss seemed to be reduced with every successive epoch.

```
Epoch 1/10
8/8 [==============================] - 70s 9s/step - loss: 1.6185 - acc: 0.5391 - val_loss: 0.7066 - val_acc: 0.4062
Epoch 2/10
8/8 [==============================] - 55s 7s/step - loss: 0.6940 - acc: 0.5078 - val_loss: 0.6873 - val_acc: 0.5000
Epoch 3/10
8/8 [==============================] - 51s 6s/step - loss: 0.6767 - acc: 0.5703 - val_loss: 0.6709 - val_acc: 0.9062
Epoch 4/10
8/8 [==============================] - 50s 6s/step - loss: 0.5915 - acc: 0.6875 - val_loss: 0.4665 - val_acc: 0.9375
Epoch 5/10
8/8 [==============================] - 49s 6s/step - loss: 0.4693 - acc: 0.7578 - val_loss: 0.4883 - val_acc: 0.9688
Epoch 6/10
8/8 [==============================] - 52s 7s/step - loss: 0.3893 - acc: 0.8281 - val_loss: 0.2319 - val_acc: 1.0000
Epoch 7/10
8/8 [==============================] - 50s 6s/step - loss: 0.2416 - acc: 0.9141 - val_loss: 0.1210 - val_acc: 0.9688
Epoch 8/10
8/8 [==============================] - 50s 6s/step - loss: 0.2878 - acc: 0.8828 - val_loss: 0.2035 - val_acc: 0.9062
Epoch 9/10
8/8 [==============================] - 50s 6s/step - loss: 0.2171 - acc: 0.9297 - val_loss: 0.0817 - val_acc: 1.0000
Epoch 10/10
8/8 [==============================] - 50s 6s/step - loss: 0.1915 - acc: 0.9453 - val_loss: 0.2931 - val_acc: 0.9062
```

Fig. 5: Epoch Statistics while Training CNN model

After the training and testing phase, the model was ready for the end user. The input is a file path of the image to be scanned and Tested for possible infection. The image is converted to an array of numbers in this phase. This array is then sent to the main model. The array contains the data of the image in numerical form

```
path = "/content/drive/MyDrive/ModelCheckdataset/COVID19-Negative/IM-0170-0001.jpeg"
img = image.load_img(path, target_size=(256,256))

img = image.img_to_array(img)/255.0
img = np.array([img])
img.shape
```

Fig. 6: Path to link the image to be scanned

After examining the image the model returns a numerical value between 0 and 1. This numerical value is calculated by the model based on the matching parameters between the scanned image and its trained imageset. If the integer is less than 0.5 the image is COVID-19 positive and if it is greater than 0.5 the image is COVID-19 negative. We round off the number to the nearest integer and declare the result.

```
predict_x = model.predict(img)
predict_x = np.round(predict_x).astype(int)
print(predict_x)
```

```
[[1]]
```

```
if (predict_x == 0):
  print("Covid-19 Positive")
elif (predict_x == 1):
  print("Covid-19 Negative")
else:
  print("Check Input")
```

```
Covid-19 Negative
```

Fig. 7: Final Result of Detection

B. Prediction Phase

In this Phase the model will ask for input from the user in the form of an integer. To be specific the integer input is the number of days, it gives the number of cases after this number of days once the user fills in the number of days input and clicks enter. The model predicts the number of possible cases depending upon the training algorithm and dataset precision.

```
Enter the N.O. of days: 5
------------------------------
PREDICTION
------------------------------
Prediction - Cases after 5 days:(40.89, 'Million')
```

Fig. 8: Final Result of Prediction

Also a graph is given as an output to show the trend of the dataset and depict the fitting curve of the model.

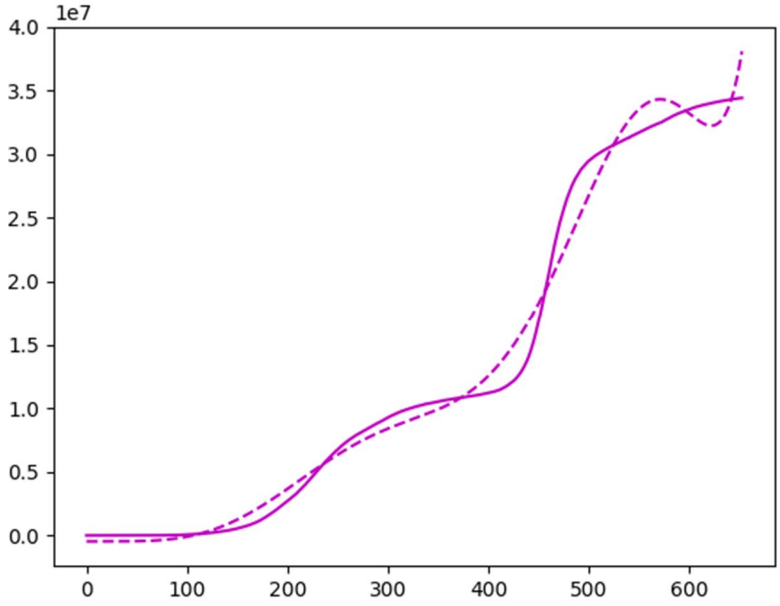


Fig. 9: Fitting curve of Model with dataset

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

C. Comparison and Analysis Phase

Analysis is focused on the number of new cases, deaths, recoveries of different countries as well as it is targeted on a specific country. Country Wise rise of cases is examined along with the countries largely affected and the countries less affected are depicted using a bar graph. Analysis based on a country is also specifically performed to discover the pattern of active cases, recovered cases and deaths on a weekly basis.

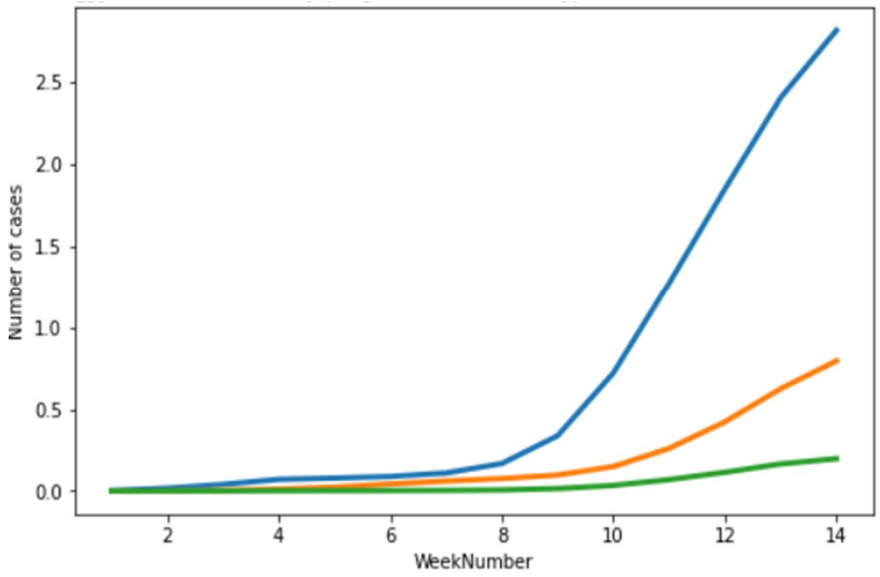


Fig. 10: Weekly Analysis of Cases in India

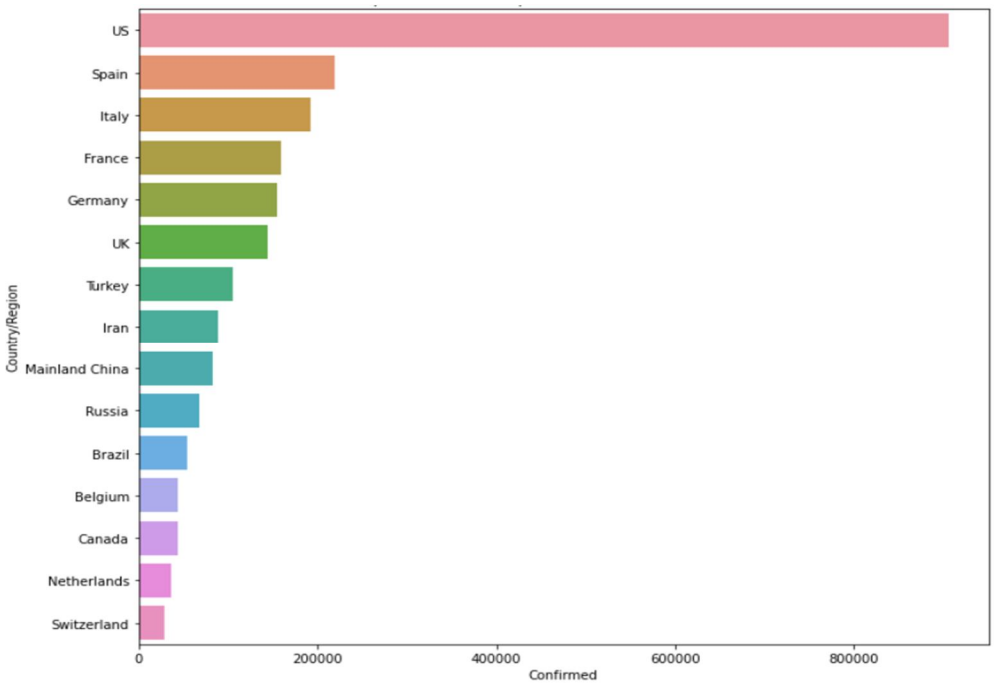


Fig. 11: Countrywise Confirmed Cases

Different models have different prediction algorithms. Comparison is done between different models which yields us which is the most precise model. Precision largely depends upon the algorithm and the behaviour of the particular model towards highly unstable datasets particularly without any trend or pattern. We developed two models namely Linear regression and Support Vector Regression for a comparative study but surprisingly both these models gave values far from the desired range of possible prediction.

	Dates	lr	svr
0	2020-04-25	1560529	3322586
1	2020-04-26	1582219	3500761
2	2020-04-27	1603909	3686599
3	2020-04-28	1625599	3880344
4	2020-04-29	1647289	4082245

Fig. 12: Comparison between Linear Regression and Support Vector Regression

Due to out of range values shown by the two models we went for Holt's Prediction model which was highly accurate and precise in predicting the possible values for the upcoming few days.

	Dates	lr	svr	Holts Linear Model Prediction
0	2020-04-25	1560529	3322586	2855246
1	2020-04-26	1582219	3500761	2933902
2	2020-04-27	1603909	3686599	3012558
3	2020-04-28	1625599	3880344	3091214
4	2020-04-29	1647289	4082245	3169870

Fig. 13 Comparison between Linear Regression, Support Vector Regression and Holt's Model

VII. CONCLUSION

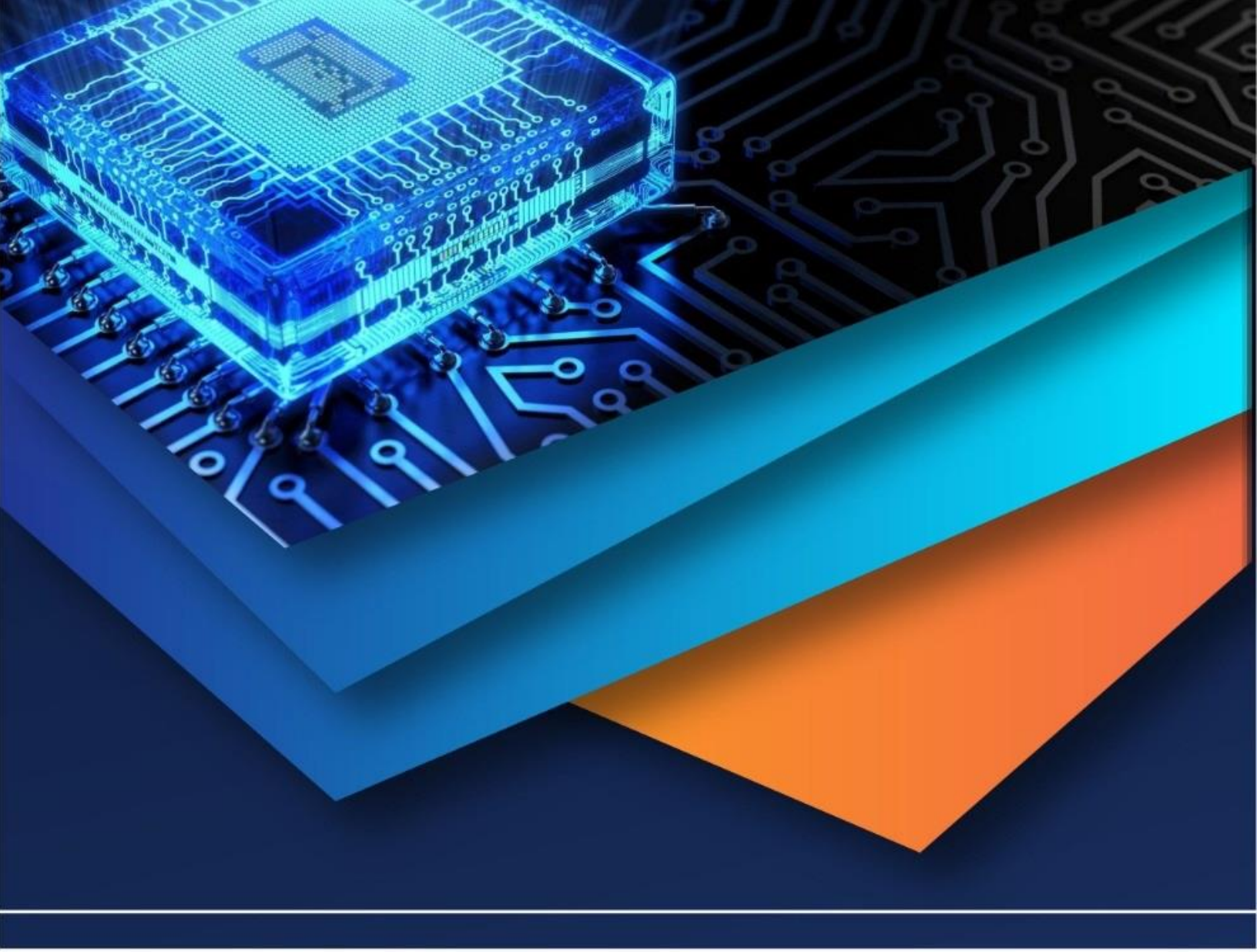
Thus from our above study we can say that the success of an Artificial intelligence model depends upon the dataset used and the base algorithm used by the model. For detection phase training the model to more number of epochs improves the accuracy by many folds. The imageset needs to be processed to remove any incorrectly oriented image. Additional filters help to extract minor details from training images which tends to give an edge to the model while in production, whereas in case of prediction we need to choose the correct model depending upon the dataset and algorithm. Linear Regression and SVR models seem to be too rigid to handle unsupervised data which can be handled well by Holt's model or Polynomial Regression. Overall for perfect analysis the dataset needs to be updated frequently and empty spaces can be filled up with garbage values.

VIII. ACKNOWLEDGEMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Sujit Tilak for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. SatishKumar Verma and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

[1] Sina F. Ardabili, Amir Mosavi, Pedram Ghamis,Filip Ferdinand. Annamaria R,Varkonyi-Koczy,Uwe Reuter ,Timon Rabczuk,Peter M, Atkinson, "COVID-19 Outbreak prediction with Machine Learning", Lancaster Environment Centre, Germany, 2020.
[2] Rachna Jain, Meenu Gupta, Soham Taneja,Jude Hemanth, "Deep Learning based detection and analysis of COVID-19 on chest X-Ray images", Springer Nature , October 2020.
[3] Boran Sekeroglu, IIker Ozsahin, "Detection of COVID-19 from Chest X-ray Images using Convolutional Neural Networks",SLAS Technology,18 September 2020.
[4] Roseline Oluwaseun and Joseph Bamidele, "MACHINE LEARNING PREDICTION FOR COVID 19 PANDEMIC IN INDIA", Landmark University, Nigeria, April 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY