

**MSc Computing**

**Dissertation**

**CSYM023**

*Name: Ayshath safoora puthiyapurayil*

*ID:21419702*

*Word count: 2513*

***COVID-19 INFECTION DETECTION***

***USING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING***

Table of Contents

[Introduction 3](#_Toc113885804)

[Literature review /Background………………………………………………………………………………………………4](#_Toc113885805)

[Aim and objectives 5](#_Toc113885806)

[Analysis & specification of requirements 6](#_Toc113885807)

[Methodology 8](#_Toc113885808)

[Data collection techniques 10](#_Toc113885810)

Experimental design………………………………………………………………………….11

Data analysis & review………………………………………………………………………11

Presentation of the results…………………………………………………………………..14

Ethics……………………………………………………………………………………….16

Conclusion…………………………………………………………………………………..17

References…………………………………………………………………………………..18

# Introduction

After a fast start, work comes to a halt as the COVID-19 pandemic spreads further. The healthcare system is wholly paralysed as a result of the dramatic increase of cases. The government's top priority shifts to COVID-19 detection in order to stop the virus's future spread. The growing number of cases is making it harder for healthcare systems to keep up. Despite this, the use of technology in healthcare is working well. The techniques based on artificial intelligence have the best rates of accuracy. The level of government preparedness determines how well a country will fare in the face of any more COVID-19 threat waves. This preparedness depends on how accurate the forecast model is. The most accurate findings can be produced by machine learning models, boosting the defence against new waves.

Since 2019, there has been an ongoing coronavirus illness pandemic known as COVID-19. It has been linked to millions of illnesses and fatalities worldwide. To date, the COVID-19 detection mainly depends on specific testing (based, for example, on saliva or respiratory swabs). For the purpose of coronavirus infection diagnosis utilising respiratory rate, certain methods utilise smart gadgets (such as Whoop). Techniques for machine learning (ML) have emerged as a promising method for coronavirus infection identification. As a result, we present a machine learning-based COVID infection prediction in this study. Five ML models have their prediction accuracy evaluated. To choose critical features for prediction and decrease prediction time overhead without sacrificing prediction accuracy, we employ the Chi-square test and knowledge-based manual feature selection. Additionally, we investigate the accuracy of various input features (those that can be measured by medical and smart devices) and discover that the removal of some features has little to no impact on the prediction accuracy. We also suggest a Generative Adversarial Network (GAN) ML-based predictor that generates synthetic data (similar to real data) for ML training in order to address the problem of insufficient or unbalanced training data decreasing prediction accuracy. Our rigorous testing demonstrates the efficiency of our techniques in raising detection accuracy. The findings of our study can offer recommendations for creating coronavirus infection predictors based on various devices and data sources. This code is I want to put on GitHub.

**Literature Review**

**COVID-19 Detection Based on Deep Learning**

For both patients with covid-19 who are ill and healthy patients, the PA view of chest x-ray scans has been used in this study. They employed deep learning-based CNN models and compared their performance after cleaning up the photos and adding more data.

**Using a Convolutional Neural Network, COVID-19 Detection**

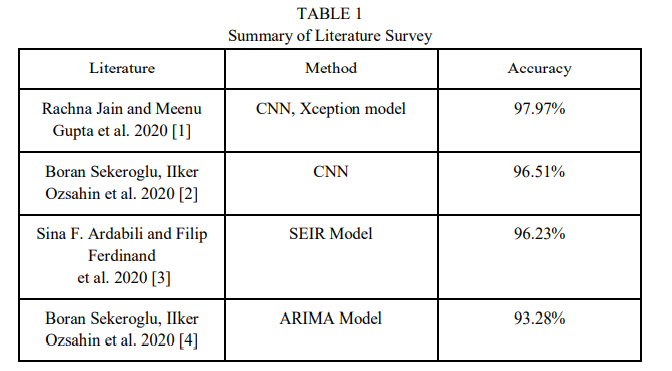
They used three different models in their paper (InceptionV3, Xception, and ResNeXt). CNN is used to assist in the analysis of this data that has been gathered. The primary focus of this research is on the classification of chest X-ray images for patients with coronavirus infections using CNN models.

**SEIR model prediction for COVID-19. SEIR Model Prediction for COVID-19**

In contrast to susceptible-infected-recovered (SIR) and susceptible-exposed-infectious-removed (SEIR) models, this paper compares machine learning and soft computing methods to forecast the COVID-19 outbreak. Two models—the multi-layered perceptron (MLP) and the adaptive network-based fuzzy inference system (ANFIS)—out of a large range of machine learning models tested, exhibited promising outcomes. The study recommends machine learning as a useful method to simulate the COVID-19 outbreak in light of the findings presented here and the very complex nature of the outbreak as well as variations in its behaviour between countries. This study offers a preliminary benchmarking to illustrate machine learning's research potential. This research also makes the case that by combining machine learning with existing methods for epidemic prediction, a genuine novelty in outbreak. In order to illustrate the potential of machine learning for future study, this paper offers a preliminary benchmarking. This work also implies that by combining machine learning and SEIR models, a true novelty in outbreak prediction can be achieved.

**ARIMA model prediction for COVID-19**

The purpose of the paper is to develop a straightforward average aggregated machine learning method to forecast the quantity, magnitude, and duration of COVID-19 cases across India, as well as their wind-up phase. The Autoregressive Integrated Moving Average Model was used in this work to analyse the datasets (ARIMA). The performance of three regression techniques, including Support Vector Regression (SVR, NN, and LR), Neural Network, and Linear Regression, was also built upon in the study to provide a straightforward mean aggregated method. The outcomes demonstrated that COVID-19 illness can be accurately predicted. The outcome of the prediction indicates that the COVID19 illness may spread by water and air ecological variables, therefore taking preventive measures such avoiding social situations, donning masks and hand gloves, and remaining at home can assist to avert the disease.



**Aim and Objectives**

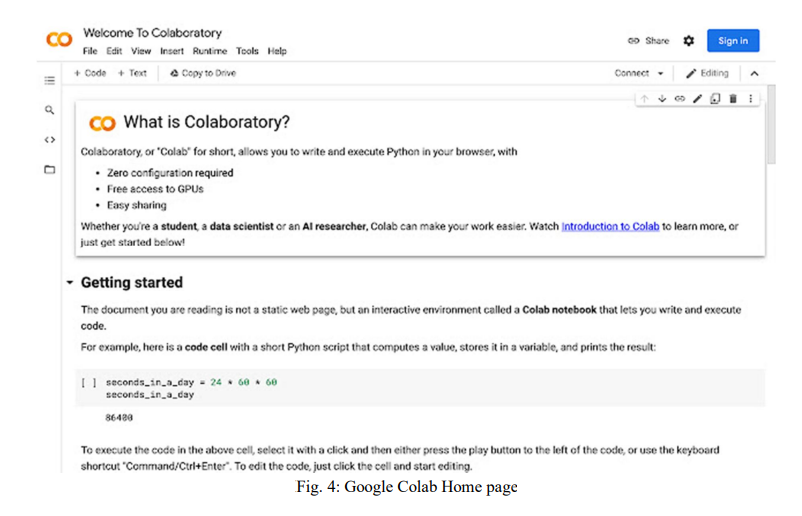
This work focuses on using AI and machine learning models to scan people's X-Rays and forecast potential corona positive instances as well as identify the spread of corona among people. To get the appropriate prediction output, a linear regression model is applied. The accuracy of several machine learning (ML) models used to predict coronavirus infection using COVID-19-related patient health data will be examined in this work.Logistic Regression (LR). In order to eliminate pointless input features and enhance predictor performance, we additionally offer two ML input feature selection techniques: the Chi-square test [9] and knowledge-based manual feature selection. We gradually remove features from the input and examine how the forecast accuracy will alter in order to account for the fact that some smart devices cannot detect specific features. Furthermore, we address the issue that the imbalanced feature in the data (with significantly fewer COVID-19 infection positive cases) and a lack of training data may have on the trained ML model's accuracy. For the purpose of training, we will specifically use Generative Adversarial Network [10] to generate synthetic data that is similar to real data. The network will be able to produce more comparable results by fitting the same dataset.

**Objective:**

Below, we provide a summary of the objective study:

* Not all of the COVID-19-related patient health data's attributes are significant or helpful for the detection of infections. To choose critical input features and exclude irrelevant ones without affecting prediction accuracy, the Chi-square test and manual feature selection based on knowledge are helpful. The accuracy of infection detection will be significantly lowered by removing some elements. As a result, the input features can be chosen in an adaptable manner.
* Although the original dataset was highly skewed, the GAN was nevertheless able to produce a balanced dataset. The accuracy of all models can be greatly increased by training them on the balanced dataset rather than the skewed dataset.

**Analysis and Specification Requirements**

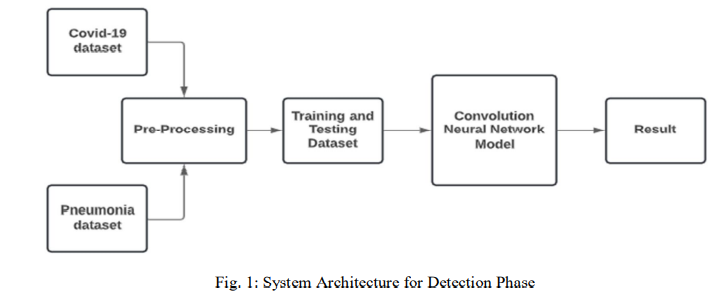


**A.Google Colab**

Google Colab is a relatively new piece of software that mimics the look and feel of Jupyter Notebook while also offering some cutting-edge features that put it ahead of its rivals. Colab assigns virtual RAM and CPU to a project, resulting in rather smooth executions. Colab is a brand-new platform for accurately and swiftly testing and designing new models. The primary benefit of Google Colab is its support for several Python libraries and the virtual RAM and CPU resources it offers, which do away with the requirement for high-performance workstations for the creation of AI models. Both novices and experts can use Colab thanks to its scalable RAM and CPU capacity. In the current fast-paced computer era, its highly scalable nature, availability at any time, and support for numerous libraries make it a powerful programme.

Collaboratively creating each model on Google Colab enables concurrent work on the models without being restricted to a single workstation. Users can upload datasets directly to Drive using the connecting to drive capability, and Colab will use the datasets directly in its execution code. Colab has thus become a new development tool in the contemporary digital computing era.

**Methodology**



A.Dataset Description for COVID-19:

The first dataset entails gathering significant data The first dataset consists of gathering significant datasets from github. X-Ray images of COVID-19-infected patients make up the COVID-19 dataset. The COVID-19 positive samples are recognised using these photos. gith's testes.

B. Pneumonia dataset description

Patients who tested negative for pneumonia make up the second dataset. Since COVID-19 and pneumonia overlap some chest symptoms, this dataset is utilised to identify COVID-19 negative samples.

C. Pre-processing description

Here, picture pre-processing consists of two stages: digital image processing, which entails executing algorithms to do image processing on digital images, and image sorting, which entails combining the COVID-19 dataset with the pneumonia dataset. To improve image data by reducing undesired distortions so that the CNN model may profit from this and become more optimised and exact, a larger range of techniques must be used to the input data.

D. Training and testing data description

The success rate of a model depends on the dataset, which is made up of meticulously cleansed data. One of the main procedures in model-based projects is cleaning. The training set is the content used to teach the computer how to process data. This dataset is used by the CNN model to run its algorithms in order to provide its primary intelligence. When presented with a chest X-ray image, the CNN model is trained using this dataset to distinguish between COVID-19 positive and negative samples.

The model's intelligence gained from the training data is assessed using the testing dataset. Important factors like Accuracy and Loss rate are listed.

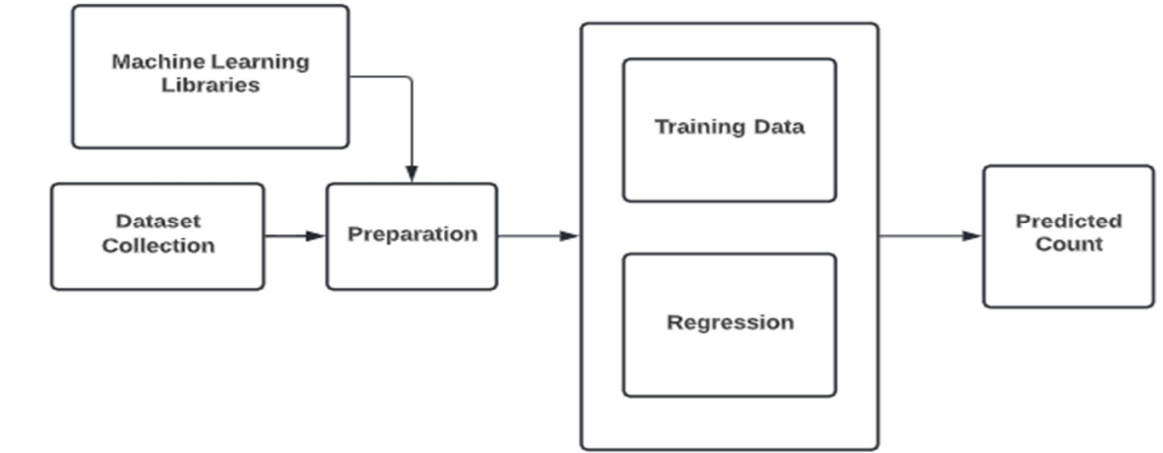
E. Block description for CNN

This block represents the Convolutional Neural Network Model, the fundamental image detection model. To finally be able to distinguish features and forms, it uses iterative training on each of the images. Convnets are used to scan images using various filtering algorithms. In order to eliminate errors and produce the most precise findings for this delicate COVID-19 test, repeated filtering scans are performed.

F. Description of Result Block

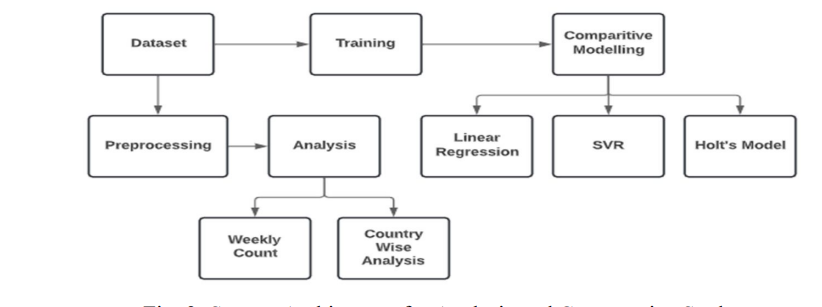
Whether the scanned Chest X-Ray of the unsure patient is coronavirus positive or negative is the final result, which is presented on the screen.

**Fig: System architecture for prediction**



**Experimental design**

**System architecture analysis**

****

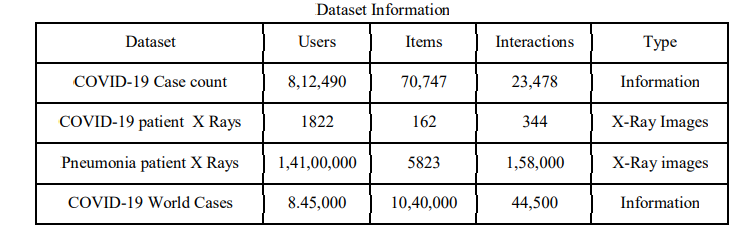
Dataset Information

The number of current cases, deaths, and recovered cases make up the dataset's three primary parts. It contains data on every nation, enabling a global analysis of the COVID-19 situation. The dataset was taken from the daily updated John Hopkins repository.

**Data collection techniques**

Datasets:

The core of the entire endeavour are the datasets for each model. These are the core of how the models function as a whole. To achieve the necessary precision, models rely heavily on reliable datasets that are regularly updated. Image of COVID-19 and Pneumonia Datasets are made up of a large number of X-Rays in two primary directions; for prediction, we need a real-time dataset that properly distinguishes between active and death counts, and for analysis, we used a static dataset to conduct analysis for a set amount of time. While Covid case numbers are extracted from Github repositories linked to reputable data collection organisations, X-ray datasets are extracted from Kaggle datasets.



**Data analysis and review**

The analysis focuses on the total number of new cases, fatalities, and recoveries in several countries as well as one particular nation. Country The countries that are most afflicted are evaluated along with the wise rise in cases, and a bar graph is used to show which countries are less affected. A country-specific analysis is also carried out to determine the weekly trend of active cases, recovered cases, and deaths.

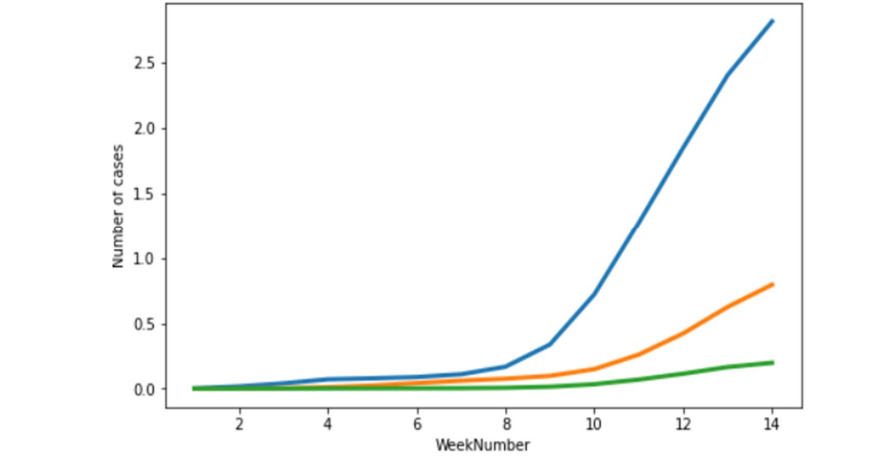


Fig: weekly analysis cases of India

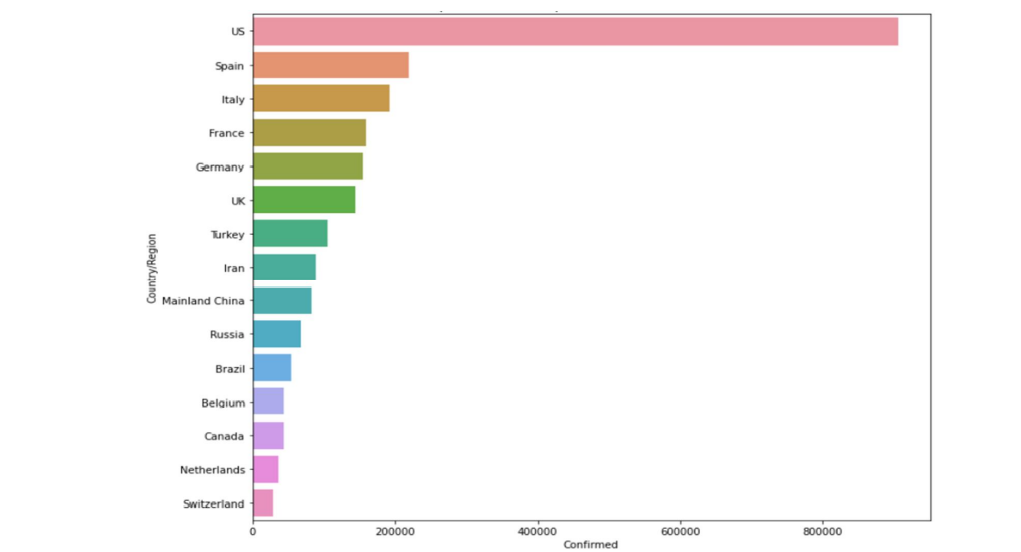


Fig: Country wise confirmed cases

Prediction algorithms differ between models. Comparing many models allows us to determine which is the most accurate. Precision heavily depends on the algorithm and how the specific model reacts to very unstable datasets, especially those lacking any kind of trend or pattern. For a comparative research, we created two models: Support Vector Regression and Linear Regression. Surprisingly, neither of these models produced results that were within the desired range of potential predictions.

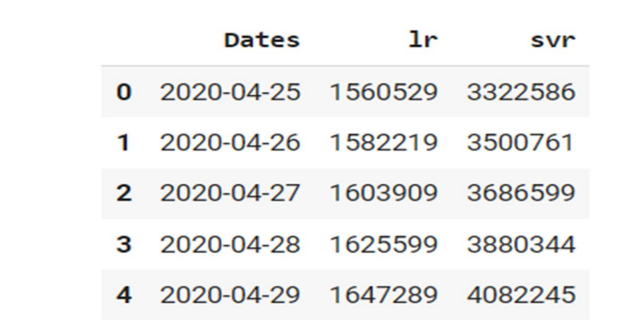


Fig : Support vector regression and linear regression are contrasted.

Due to the two models' out-of-range results, we chose Holt's Prediction, which was extremely accurate and exact in projecting potential values over the next few days.

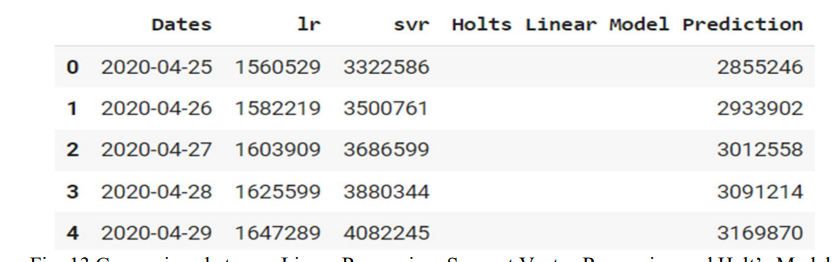
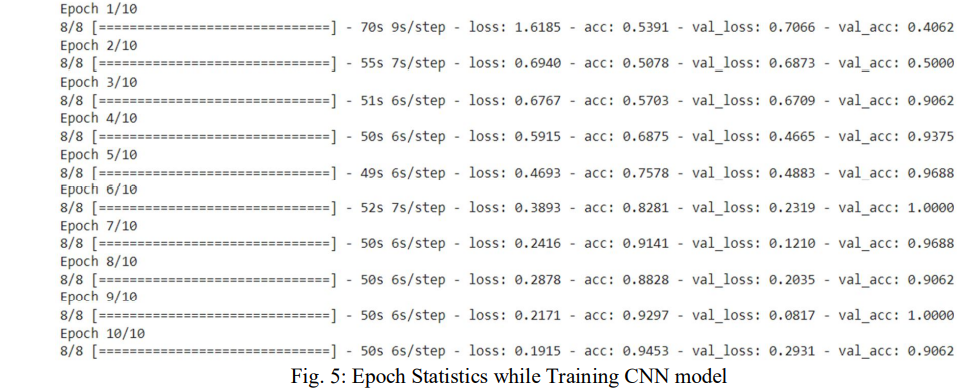


Fig: Comparison between Holt's Model, Support Vector Regression, and Linear Regression

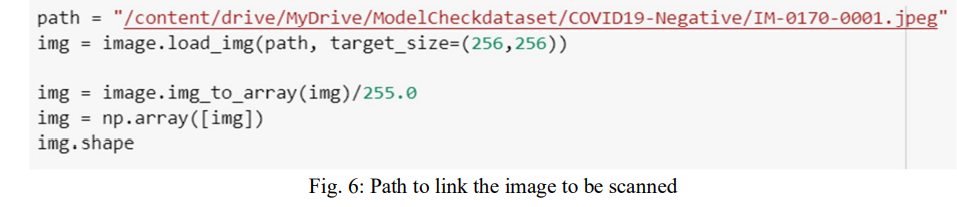
**Presentation of the results**

**Phase of Detection Results**

With training and testing datasets, we trained the model for 10 iterations during the detection phase, achieving great accuracy and precision. With each new period, the loss seemed to be lessened.

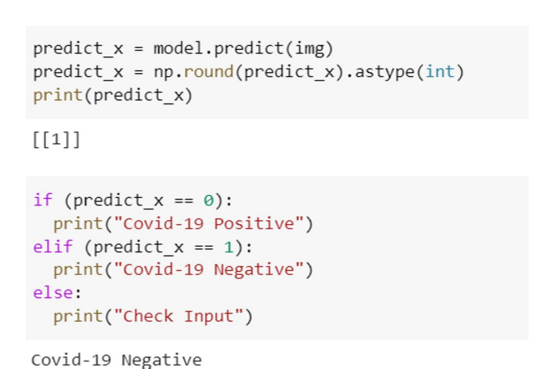


The model was prepared for use by the end user following the training and testing process. The file path of the image to be scanned and tested for potential infection serves as the input. In this stage, the image is transformed into an array of numbers. The primary model receives this array after that.The image's numerical data is contained in the array.



The model returns a numerical result between 0 and 1 after analysing the image. Based on the matching parameters between the scanned image and its training imageset, the model determines this numerical value. The picture is COVID-19 positive if the integer is less than 0.5 and COVID-19 negative if it is more than 0.5. The result is then announced after being rounded to the nearest integer.

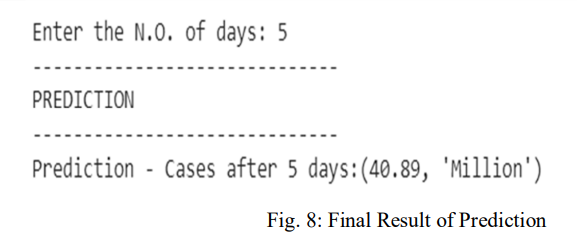
**Final result of Detection**



The result is covid negative

**Phase B: Prediction**

The model will ask the user for input in the form of an integer at this phase. When the user enters the number of days in the appropriate input field and presses Enter, the system displays the number of cases that occurred after that number of days. Depending on the precision of the dataset and the training process, the model forecasts the number of potential examples.



Additionally, a graph that displays the trend of the dataset and the model's fitting curve is provided as an output.

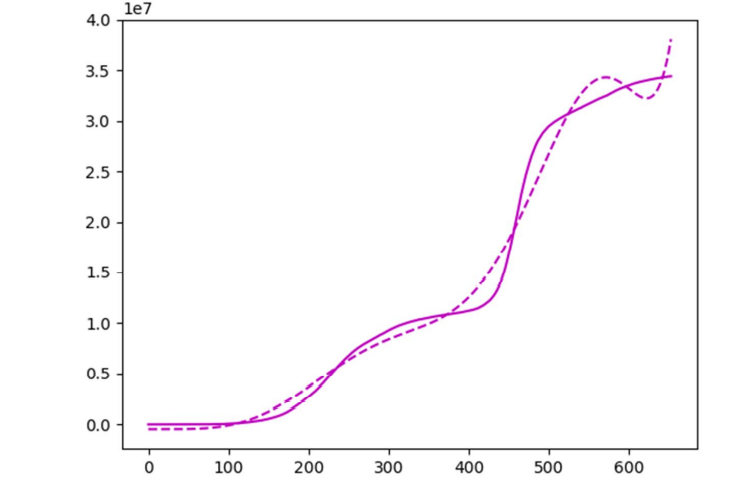


Fig: fitting of curve of model with dataset

**Ethics**

This work focuses on using AI and machine learning models to scan people's X-Rays and forecast potential corona positive instances as well as identify the spread of corona among people. Using a linear regression model,to produce the necessary Prediction results. SVR model has been utilised in conjunction with linear regression to achieve comparative analysis between various model prediction methods. Finally, the Holt's model for predicting was created, and it appeared to be superior than the other two models. This project uses a lot of traditional ML-related Python tools to train and preprocess a lot of data. The Convolutional Neural Network model is utilised to identify COVID-19 infection in people during the detection phase. To get the best accuracy, we implemented numerous filtering rounds during the detection phase. To increase accuracy and lower overall loss, the model is trained up to a specified number of epochs. All of the aforementioned models' datasets were taken from github repositories connected to certain data collection organisations.

**Conclusion**

Therefore, based on the results of the study mentioned above, we can conclude that the dataset and base algorithm utilised by an artificial intelligence model determine its success. Training the model for more epochs during the detection phase greatly increases accuracy. Any improperly oriented photos must be removed through processing of the imageset. While choosing the right model for prediction depends on the dataset and algorithm, additional filters help to extract minute information from training photos that tend to give the model an advantage while in production. SVR and linear regression models appear to be too inflexible to handle unsupervised data, whereas Holt's model or polynomial regression can handle effectively.

**References**

[1] J. J. Amon, “Covid-19 and detention: Respecting human rights.” Health and Human Rights,, vol. 22, no. 1, pp. 367–370, 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] Centers for Disease Control and Prevention, “Symptoms of coronavirus,” December 2020. [Online]. Available: https://www.cdc.gov/coronaviru s/2019-ncov/symptoms-testing/symptoms.html

[4] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, Applied logistic regression. John Wiley & Sons, 2013, vol. 398.

[5] W. Fan and K. Zhang, Bagging: Encyclopedia of Database Systems. Springer US, 2009. [6] Imbalanced-learn API, “Balanced bagging classifier.” [Online]. Available: https://imbalanced-learn.org/stable/generated/imblearn.ens emble.BalancedBaggingClassifier.html

[7] J. H. Friedman, “Stochastic gradient boosting,” Elsevier Science Publishers B. V., vol. 38, no. 4, 2002.

[8] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” CoRR, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754

[9] Scikit-learn Chi2, “sklearn.feature selection.chi2,” 2011. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feat ure\ selection.chi2.html\#sklearn.feature\ selection.chi2

[10] C. Nash, “Create data from random noise with generative adversarial networks,” 2017. [Online]. Available: https://www.toptal.com/machine -learning/generative-adversarial-networks

[11] Boran Sekeroglu, IIker Ozsahin, ”Detection of COVID-19 from Chest X-ray Images using Convolutional Neural Networks”,SLAS Technology,18 September 2020.

[12] 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.

[13] Roseline Oluwaseun and Joseph Bamidele, ”MACHINE LEARNING PREDICTION FOR COVID 19 PANDEMIC IN INDIA”, Landmark University, Nigeria, April 2020