An Advanced Machine Learning Model for Disease Prediction

Ayushi Sharma, Shipra Shukla

Abstract: ***To settle on right choices and pass on about vital control measures, numerous flare-up expectation models for anticipating COVID-19 are getting utilized all round the world. Straightforward conventional models have indicated extremely less precision rate for future forecast use, because of more significant levels of vulnerability and absence of proper information. Among the different machine learning model algorithms contemplated, an ensembled model was seen as giving the best outcomes. Because of the multifaceted nature of the virus's temperament, this research paper recommends machine learning to be an extremely helpful gadget to consider in case of the ongoing pandemic. This paper gives a colossal benchmark to call attention to the probability of machine learning to be utilized as an instrument for future exploration on pandemic control and its timely prediction. Moreover, this paper delineates that the best prompts for pandemic prediction are frequently comprehended by combining machine learning, predictive analytics and visualisation tools like Tableau. The main purpose of this research is to build a perfect ML model prototype which can be later used when access to appropriate dataset (which is both large and consists of many different features) is available. Also, the secondary aim is to automate the process of reporting so as to facilitate quicker action by the concerned authorities, and help common people reach out to the correct destination for treatment or help. Furthermore, the Tableau analysis performed on the dataset is to provide more analytical depths for people with expertise in the medical domain.***

Keywords: COVID-19, machine learning, predictive analytics, Tableau.

Graphical Abstract

I. INTRODUCTION

The novel covid-19 virus still continues to pose an enormous impending threat to healthcare worldwide.

The amount of confirmed patients with the covid-19 infection has crossed the mark of 75,000 in additional than 160 nations, since the onset of the disease in Dec'19 in China. Even quite 36,000 people are dead from exposure to the present virus (till 30 March 2020) [2]. Inspite of continuous efforts from the general public health sector in controlling the spread of the virus and delaying its impact, many countries have now been subjected to an important phase of crisis, and more and more countries are sure to join this list. [3], [4], [5] All of this has led to an urgent increase within the demand for medical equipments like PPE kits, N-95masks for the security of medical staff, and also hospital beds for infected patients [6]. Urgent diagnosis and also prognosis of the virus-borne infection has become the necessity of the hour so as to reduce the burden on the healthcare sector, and also to supply for absolute best cure for the patients [10]. With having more and more accurate prediction models, the character of spread of the virus and its consequences are often referred to as it controls what the govt and other legislative bodies would decide regarding new policies and measures to contain the spread [7]. These accurate prediction models could be of great help to the healthcare workers in treating patients when only few even less than sufficient resources are available. Models starting from simple traditional ones to advanced ML models have been thought of and worked upon in order to share the related research work to help save lives around the world [6]. The research work has been split into three different modules. First one being the making of an advanced machine learning (ML) model which would help predicting the clinical testing report (positive/negative) of any person, related to the deadly Coronavirus. The second module consists of the automation of the process after prediction. Lastly, the third module consists of detailed in-depth analysis of various factors regarding the virus.The model, deployed as a webpage, would take various factors as input such as medical history, symptoms, travel history, state, etc. and then predict whether the concerned person has been infected by the COVID-19 virus or not. Later, an automation system would mail the test reports (positive/negative), along with a link to an html map tagging all COVID-19 testing hospitals in India, as soon as the model finishes training and prediction, to the concerned person who entered details on the webpage. Furthermore, a Tableau dashboard would help in getting more detailed insights about the virus. Here, the concept of predictive analytics in healthcare has been used in aiding the diagnosis. As machine learning is considered a modern-day extension of predictive analytics, both concepts have been integrated while making the model. Understanding every facet of the nature of the virus, the observations of symptoms, and what a positive outcome is, in conjunction with the present scenario, is what truly makes machine learning combined with predictive analytics one of the best approaches in aiding diagnosis [8].

With the current situation of the world being fighting the deadly COVID-19 pandemic, there is an immediate need for quick and early diagnosis, without having to manually test every other individual. Also, there is an even greater need to know more about the deeper insights of the virus so as to know its nature and impact, which is not yet very clear. This was the sole inspiration in carrying out this research. This research could prove to be highly beneficial in the testing phase for quick actions to be taken with the people diagnosed with positive symptoms.

Revised Manuscript Received on July 30, 2020.

Ayushi Sharma, CSE, Amity University, Noida, India. E-mail: ayushi.hsharma@gmail.com

Dr. Shipra Shukla, Asst. Prof., CSE, Amity University, Noida, U.P., India. E-mail: ershiprashukla88@gmail.com

II. LITERATURE REVIEW

This global COVID-19 pandemic comes across as even weirder thanks to non-linearity which successively results in complexity of its nature[9]. Furthermore, this virus-borne pandemic is extremely different from other pandemics within the past like MERS and SARS, which forces one to consider the power of traditional ML techniques to supply accurate results[10].Aside from the varied features(known/unknown) participating within the spread, the growing complexity of differences within the behaviour across the world in various geographical areas and also the differences within the various strategies of containment adopted further increases the uncertainty of the model during a dramatic manner [11],[7].

The usage of various ML techniques to cure different outbreak medical situations of the past is shown below in Fig 1. [1].

Journal	Outbreak infection	Machine learning
Transboundary and Emerging Diseases	Swine fever	Random Forest
Geospatial Health	Dengue fever	Neural Network
BMC Research Notes	Influenza	Random Forest
Journal of Public Health Medicine	Dengue/Aedes	Bayesian Network
Informatica	Dengue	LogitBoost
Global Ecology and Biogeography	H1N1 flu	Neural Network
Current Science	Dengue	Adopted multi-regression and Naïve Bayes
Environment International	Oyster norovirus	Neural Network
Water Research	Oyster norovirus	Genetic programming
Infectious Disease Modelling	Dengue	Classification and regression tree (CART)

Fig 1.: Usage of various ML techniques to cure different outbreak medical situations of the past

III. PROBLEM STATEMENT AND SOLUTION

To build an advanced prediction model for diagnosing traces of the covid-19 virus in suspected patients, and also for predicting the people who are at risk of being affected so that they can be admitted to the hospital urgently [6]. Also, visually show different statistics related to COVID-19 so as to help with gaining deeper insights into the problem and its cure. For the solution, the first task at hand was to gather a dataset that was large enough, along with containing important features as columns. A lot of datasets were gone through, but none of them fulfilled both of the requirements. In the end, a dataset was finalised that was just a perfect blend of the two. Since the datasets related to the COVID-19 symptoms are constantly changing and also are not extensively available as of now, the dataset finalised was taken for our model creation. Here, the end goal being to atleast perfecting our advanced model to the utmost level. The data used for prediction and analysis in this paper includes the official counts as released by MoHFW, the COVID-19 India API, as well as volunteer collected de-identified open source data (from Kaggle) until the first week of May, 2020 [7].

After the acquisition of the required dataset, now the focus was towards the pre-processing of dataset. In a ML process, data pre-processing is the initial step in which the data is transformed in such a way that the machine can easily read it i.e. it can be easily picked up by the machine's algorithm. For data pre-processing, the steps performed on the dataset are data cleaning (i.e. Removing missing and duplicate values), dimensionality reduction (i.e. Feature selection – Filter method), feature encoding (i.e. converting categorical variables into numeric form).

Initially, the dataset consisted of the following features: age, region1, region2, detected state, nationality, travel history, disease history, symptoms and label. For the data cleaning step, the column region2 was dropped, as a column named region1 already had the same values. So, the duplicate column named region2 was removed.

Next for the dimensionality reduction part, correlation analysis was carried out using the Filter method. All of the features showed a positive correlation with the column showing the prediction outcome (named label), in the heatmap plotted.

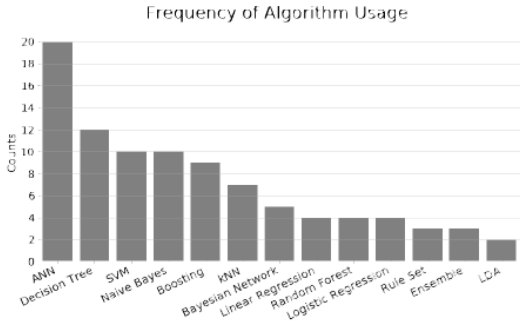

Fig 2.: Usage patterns of various ML algorithms

Lastly, for the feature encoding part, the values in columns such as gender, region1, detected state, nationality, travel history, disease history, symptoms and label, were converted from categorical form to numeric form (into a dummy variable).

To make a predictive ML model, a choice had to be made among the various ML algorithms given below, in order to select the most efficient model to train the chosen dataset:

- Logistic Regression
- Random Forest
- KNN (K-Nearest Neighbours)
- Decision Tree
- SVM (Support Vector Machine)

Given below in Fig 2., is the histogram showing usage patterns of various ML algorithms [12].

An ensemble learning method was used to enhance the predictive model's accuracy.

Ensemble method is a ML technique that combines various ML models and produces one single ensemble model with the best optimal accuracy [13].

Many people together forming a diverse group would make better decisions rather than an individual person. This is the underlying concept used for making an ensemble model [14]. For this research, the technique of max voting has been used in the final ensemble model. In this, many models are combined and used to make predictions for each data point. The output from each model is marked to be a 'vote'. The predictions collected from the most number of the models is then used as the final result of the prediction.

Eg- When 5 people were asked to rate a particular movie (out of 5), and if four of them rated it as a 4-star while two of them gave it a 5-star rating, the final rating would be taken as 4, keeping in mind the majority rating. **It can be considered as taking the mode of all the predictions [14].**

A diagrammatic representation of the ensemble model created is shown in Fig 3.

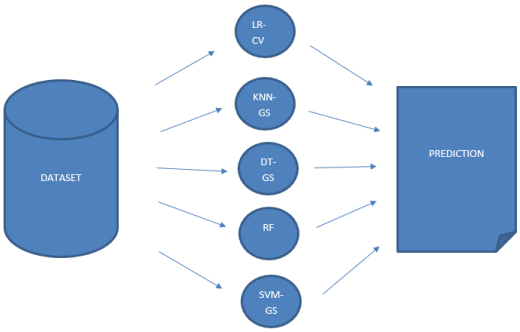

Fig 3.: Diagrammatic representation of the Ensemble Model

The ensemble model above uses multiple learning algorithms (such as Logistic Regression with GridSearch, KNN with GridSearch, Decision Tree with GridSearch, Random Forest, and Support Vector Machine with GridSearch) to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone [16].So, it was tried to combine some of the ML model techniques together as an ensemble model.

IV. RESULT AND ANALYSIS

A confusion matrix was brought to use to realize the quality of accuracy for the classifiers built using all different ML algorithms. The diagonal elements of the matrix shown below represents the amount of points that the prediction is true which is labelled as a 'true label' (TP/TN), while the off-diagonal elements of the matrix show those number of points which are mislabelled by the classifier and are a 'false label' (FP/FN).

Here, **True positives (TP)** means the cases in which we predicted yes (they have the disease), and they do have the disease.

True negatives (TN) is when the classifier predicted no, and they don't have the disease.

False positives (FP) is when the classifier predicted yes, but they don't actually have the disease.

False negatives (FN) is when the classifier predicted no, but they actually do have the disease. Greater the amount of diagonals in the confusion matrix, the higher is that the indicating power of the model to form accurate predictions [15]. **Accuracy rate** is the overall performance of the classifier i.e. how often is the classifier correct? It is calculated using the formula: (TP+TN)/total OR it can be seen in the classification report, named as weighted avg.

Misclassification/error rate is the overall of how often is the classifier wrong? It is calculated using the formula: (FP+FN)/total OR 1-accuracy rate.

Below in Table 1., a detailed analysis of various parameters taken from confusion matrix (TP, TN, error rate) as well as the classification report (accuracy or weighted average) for the various combinations used is shown, which helps in determining which model combination should be chosen as the prediction model. It clearly shows that the parameters of the ensembled model are the best.

Table 1. : Detailed analysis of parameters of confusion matrices of all model combinations used.

ML Model	TP	TN	Accuracy rate	Error Rate
Logistic Regression + K-Fold CV	200	110	97.54%	2.46%
KNN + GridSearch	180	110	98.04%	1.96%
Decision Tree + GridSearch	180	100	98.29%	1.71%
Random Forest	180	110	98%	2%
SVM + GridSearch	180	100	98%	2%
Ensemble of the above models	180	110	98.97%	1.03%

The accuracy results, error rate, and, precision given by each of the ML model combinations were compared using the table above. Then, comparisons between the model accuracies of the models individually and along with various accuracy enhancing techniques, were carried out. Also, various different hit-and trial combinations of all these techniques mentioned above were tried. Finally, the following results (shown in Table 2.) gave the overall best accuracies, out of all the combinations.

Table 2: Comparisions of accuracies of the model combinations used.

ML Technique Combinations	Accuracy (in %)
Logistic Regression	97%
Logistic Regression+K-fold Cross Validation	97.54%
KNN	97.65%
KNN + Grid Search	98.04%
Decision Tree	98.23%
Decision Tree + Grid Search	98.29%
Random Forest	98%
SVM	97%
SVM + Grid Search	98%

The most accurate combination options for each model were taken and ensembled together using the Ensemble Learning Technique. The final result of the ensemble model came out to be the best (giving an overall accuracy of ~98.97%), shown below in Table 3.

Table 3: Final selection of models used for the ensemble model, with highest accuracy.

Model Combination	Accuracy (in %)
Logistic Regression + K-fold Cross Validation	97.54%
KNN + Grid Search	98.04%
Decision Tree + Grid Search	98.29%

Random Forest	98%
SVM + Grid Search	98%
ENSEMBLE of the above models	**98.97%**

V. AUTOMATION

The automation part was to automatically send an email containing the test reports (positive/negative), along with a link to an html map tagging all COVID-19 testing hospitals in India, as soon as the model finishes training and prediction. For building the automation system, we used the SMTPLIB module which is used to send an email to any machine available via internet connection with any IP via an STMP/ESTMP daemon, in the form of a session [17][19]. Below in Fig 5., is the snapshot of the automatic email sent to the user, containing the map file.

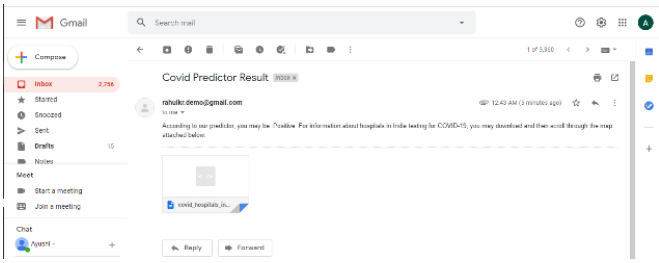

Fig 5.: Snapshot of the automated mail sent

The html map was created using the Folium library in python which uses OpenStreetMap to show the map. The map can be seen below in
Fig 6.

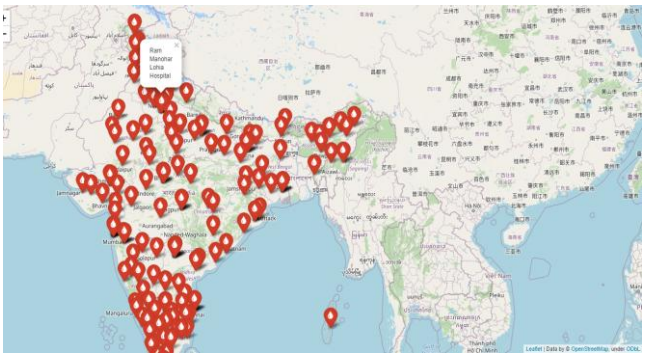

Fig 6.: Folium map showing tags of hospitals testing and treating COVID-19

VI. DEPLOYMENT

Deployment is a very crucial step to make any ML model of some use to the public (end users). Here, Flask (a web application framework written in Python) has been used to deploy the final ensembled ML model as a web page. Below, in Fig 7., the layout of the basic "COVID-19 Predictor" webpage can be seen.

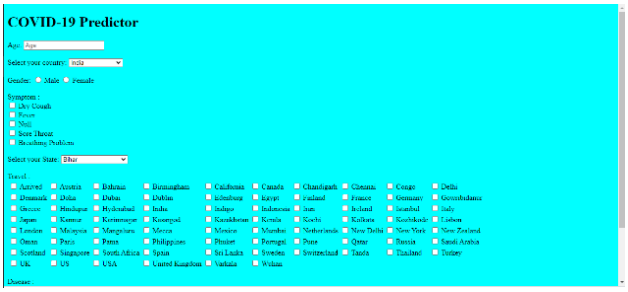

Fig 7.: Snapshot of the basic layout of the COVID-19 Predictor webpage

VII. VISUALISATION

For the aspect of visualisation, Tableau worksheets, which would help in gaining deeper insights into the nature and impact of the virus, were created. It would also come in handy in knowing about the current preparation level and ways to enhance it. Here, a live data extract has been used as a data source which constantly gets updated on a daily basis. Snapshots of some of the sheets (both static and interactive) are shown below from Fig 8. to Fig 13.
NOTE: The Tableau worksheets below show visualisations related to COVID-19 upto atmost the first week of May.

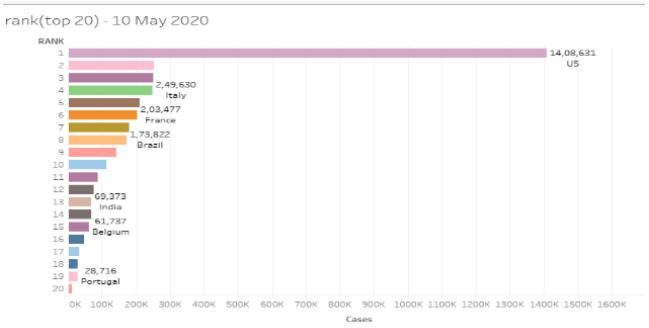

Fig 8.: Top-20 Rank-wise (as per cases)

Fig 8.: An interactive sheet showing the details of total cases of top-20 countries (rank-wise), updated daily. Here, the top 20 countries in the view keep changing as the total cases increase or stay same, as the date changes.

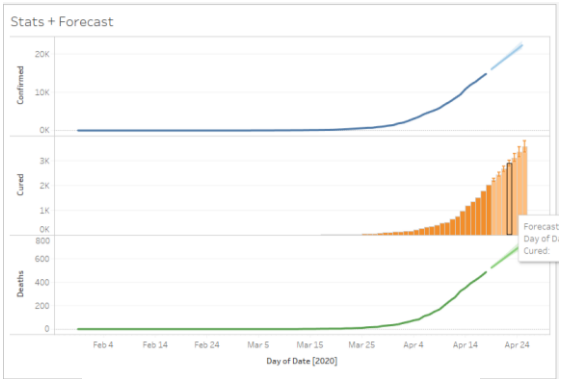

Fig 9.: Case Statistics and Forecast

Fig 9.: A static sheet showing statistics and forecast of confirmed, cured and death cases in India, along the entire timeline. Here, it can be noticed that the graph is not too steep i.e. there's no exponential growth even in the predicted phase, which is a good sign.

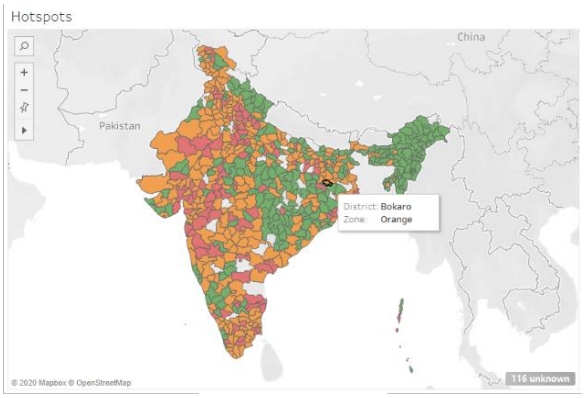

Fig 10: Hotspots

Fig 10.: A static sheet showing the three hotspot zones (district wise) – red, orange and green in different states of India. A district is decided to be designated as a red zone if it satisfies *any one* of the following conditions:

- it has a total of 4 infections while the state is fortunate to have a total of only 5,
- or if district disease-load increases from 1 to 2 in four days? [18]

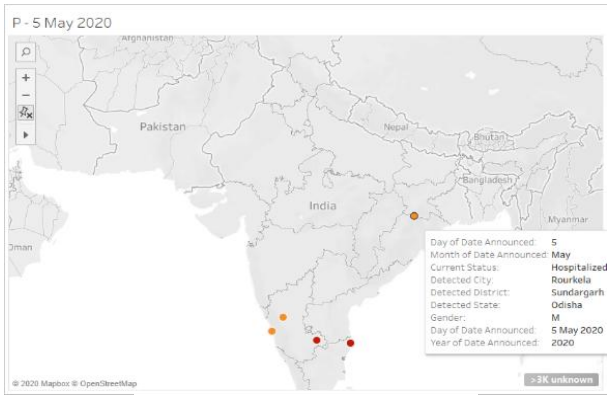

Fig 11.: Patient Health Details

Fig 11.: An interactive sheet showing all the patient details of all affected people, along the entire timeline. The status of patients shown is of three categories – Deceased, hospitalized, and recovered and is shown with the colours red, orange and green respectively. With each date of the month, it can be seen how the status of cases have changed.

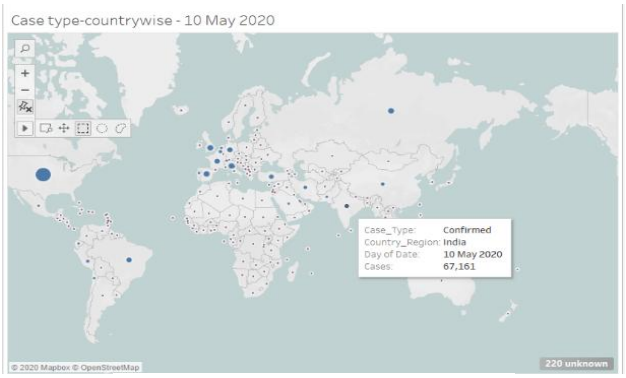

Fig 12.: Case Types (Country-wise)

Fig 12.: An interactive Tableau sheet showing the country wise-case types (confirmed & death) around the world, along a timeline starting from the beginning of the pandemic to the latest date. Here, the size of the dot is proportional to the number of cases in that country. Just with the click of a button, the dot sizes increase with change in dates along the entire timeline.

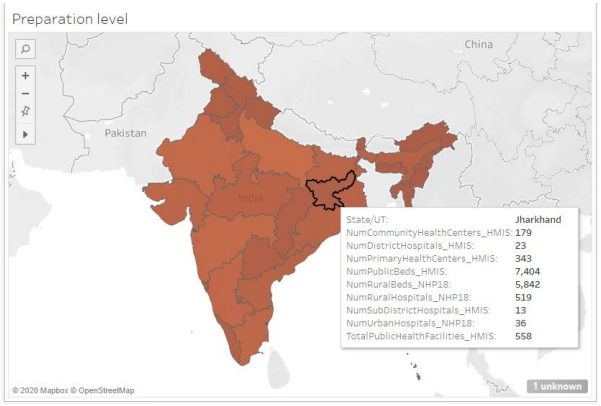

Fig 13.: Preparation Level

Fig 13.: A static sheet showing the preparation level of states in India, in terms of the total number of public healthcare facilities. It is shown using a colour range from brown to blue, where more brownish shade colour means more number of total public healthcare facilities in that state. Here, it can be concluded that some states such as Rajasthan, Maharashtra need to become more prepared with the availability of their healthcare facilities.

VIII. LIMITATIONS AND CONSTRAINTS

The absolute motto of prediction model (one dedicated to the job of predicting whether a person is infected with a disease or not) is to supply help to the healthcare unit in decision-making. Thus, it's essential to identify a target population on which the future predictions made would be of some use to medical field, and a dataset containing patient symptoms data on which the prediction model could be developed then tested. This targeted population must be known accurately such the performance of the developed model are often tested, and thus the users could know which is that the audience while making predictions. Obviously, because of time constraints and lack of data from targeted audience, the sort and amount of knowledge entries isn't enough as of now. With new publications and research work on covid-19 related ML models being subjected to the medical work field with such rapidity, the predictor model built can't be viewed as an up so far predictor with all the new symptoms of COVID-19. supported the predictors included in various other models on COVID-19, more and more factors must be considered incorporating several candidate predictors: like temperature of patient's body, respiratory signs, loss of hearing/taste, etc [6]. An important point to be considered is that the model in itself gives very accurate results. Since there are limited datasets available, with accurate and enormous amount of patient entries, the dataset chosen gave the simplest accuracy results. Nevertheless, a minimum of the simplest prototype for a COVID-19 Predictor has been built as of now. within the future, as more and more patient data regarding symptoms is formed available, the dataset are often always updated and trained for an ML model with even better accuracy, to supply with more and more accurate information

for the utilization of healthcare sector because the research findings increase on this subject worldwide[6].

IX. CONCLUSION

The ensemble advanced ML model was created that successfully predicts the COVID-19 status (positive/negative) of any person who enters his/her details in the webpage. The automation part would prove to be useful for quick action to be taken if a particular person is predicted to be positive. The advanced ML model prototype can be considered as the best as of now. This model can be used as a way to predict COVID-19 positivity (however mild, moderate, or severe), without the need to perform extensive testing for every other person in the country, as it's a very tedious task to test everyone in a country like India, with more than 1.3 billion population. Thus, as soon as someone tests COVID-19 positive by this model, he/she could go to their nearby hospitals immediately and isolate themselves. Furthermore, the dashboard made can be used for future analysis and forecasting for the COVID-19 curve. It can be very beneficial for COVID-19 researchers to get more detailed insights about the spread and nature of the virus. This would help in combating the virus in a better, more efficient way.

REFERENCES

1. COVID-19 Outbreak Prediction with Machine Learning Sina F. Ardabili 1, Amir Mosavi 2,3,*, Pedram Ghamisi 4, Filip Ferdinand 2 , Annamaria R. Varkonyi-Koczy 2, Uwe Reuter 3, Timon Rabczuk 5, Peter M. Atkinson 6
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis2020:S1473-3099(20)30120-1. doi:10.1016/S1473-3099(20)30120-1. pmid:32087114CrossRefPubMedGoogle Scholar
3. Arabi YM, Murthy S, Webb S. COVID-19: a novel coronavirus and a novel challenge for critical care. Intensive Care Med2020. doi:10.1007/s00134-020-05955-1. pmid:32125458CrossRefPubMedGoogle Scholar
4. Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. JAMA2020. doi:10.1001/jama.2020.4031. pmid:32167538CrossRefPubMedGoogle Scholar
5. Xie J, Tong Z, Guan X, Du B, Qiu H, Slutsky AS. Critical care crisis and some recommendations during the COVID-19 epidemic in China. Intensive Care Med2020. doi:10.1007/s00134-020-05979-7. pmid:32123994
6. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020; 369 doi: https://doi.org/10.1136/bmj.m1328 (Published 07 April 2020) Cite this as: BMJ 2020;369:m1328
7. Predictive Model with Analysis of the Initial Spread of COVID-19 in India Shinjini Ghosh1 Massachusetts Institute of Technology 77 Massachusetts Avenue, MA 02139, USA
8. The APP Solutions. (August 27, 2019). How predictive analytics is changing healthcare industry. Medium. https://medium.com/@TheAPPSolutions/how-predictive-analytics-is-changing-healthcare-industry-999646a97d59
9. Ivanov, D. Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transp. Res. Part E Logist. Transp. Rev. 2020, 136, doi:10.1016/j.tre.2020.101922.
10. Koolhof, I.S.; Gibney, K.B.; Bettiol, S.; Charleston, M.; Wiethoelter, A.; Arnold, A.L.; Campbell, P.T.; Neville, P.J.; Aung, P.; Shiga, T., et al. The forecasting of dynamical Ross River virus outbreaks: Victoria, Australia. Epidemics 2020, 30, doi:10.1016/j.epidem.2019.100377.
11. Darwish, A.; Rahhal, Y.; Jafar, A. A comparative study on predicting influenza outbreaks using different feature spaces: application of influenza-like illness data from Early Warning Alert and Response System in Syria. BMC Res. Notes 2020, 13, 33, doi:10.1186/s13104-020-4889-5.
12. The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review Rory Bunkera,_, Teo Susnjakb aNagoya Institute of Technology. Gokisocho, Showa Ward, Nagoya, Aichi, 466-8555, Japan bMassey University. Massey University East Precinct, Dairy Flat Highway (SH17), 0632, New Zealand
13. Lutins, Evan. (Aug 2, 2017). Ensemble Methods in Machine Learning: What are They and Why Use Them?. Towards Data Science. https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f
14. Singh, Aishwarya. (JUNE 18, 2018). A Comprehensive Guide to Ensemble Learning (with Python codes). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble- models/
15. Scikit-learn.org. https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html
16. Patel, Ashish. (May 15, 2019). Ensemble Learning- The heart of Machine learning. Medium. https://medium.com/ml-research-lab/ensemble-learning-the-heart-of-machine-learning-b4f59a5f9777
17. Python.org. https://docs.python.org/3/library/smtplib.html
18. Sen, Arijit. (April 18, 2020). Rules to Define India's COVID-19 Hotspots Are Omitting More Stressed Districts. The Wire. https://science.thewire.in/health/covid-19-hotspot-districts-red-zone/
19. Shukla, S., Kumar, M., An Improved Energy Efficient Quality of Service Routing for Border Gateway Protocol, Computer and Electrical Engineering, Elsevier, Vol. 67, 2018

AUTHORS PROFILE

Ayushi Sharma is currently pursuing her Bachelor's degree in Computer Science and Engineering from Amity University, Noida, Uttar Pradesh. She is a highly motivated and hardworking student, with an in-depth knowledge in the area of data science and machine learning through various electives at college as well as through personal self-learning expeditions. Always willing to learn new things and become an expert in her area of work, with a passion for challenges, innovation, and working with people and communities. Her research interests include machine learning, deep learning, data science in general.

Shipra Shukla received B. Tech, M. Tech and Ph. D. degrees in Computer Science and Engineering. Presently, she is working as Assistant Professor in the department of computer science and engineering at Amity University, Noida, India. She has obtained more than 6 years of research and teaching experience. Her current research interests are computer networking, Machine Learning techniques, IoT and Routing in the internet.