International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Predictive Analytics of the Type of Transmission among Covid-19 Patients in India using Machine Learning Classifiers

Jaideep Kala[1], Sonalika Bhandari[2]

[1-2]*Department of Electronics and communication Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India.*

Abstract - *The Novel Corona-virus (Covid-19) pandemic which was first reported in Wuhan, Hubei Province, China in late December 2019 has been spreading rapidly in India and across the globe with over 95,698 patients in India as of 18th May 2020. The government of India is following the approach of rapid testing and tracing contacts of infected patients and their travel history to contain the spread of the virus. Although the agencies which are tasked with tracking of patients and identifying the cause of infection in them are facing difficulties due to lack of proper methodology and due to the large number of patients being infected every day. The cause of infection among patients could be from Local Transmission, foreign travel or domestic travel. The open source database of patients provided by the government of India provides these details for only initial cases and not for majority of patients recently infected, hence tracing of people that these patients might have infected becomes a difficult task. This paper presents a prediction model using demographical and individual details for finding the type of transmission among patients which provides an insight for the spread of virus at the community level. Machine learning classifiers such as Support Vector Machine, Decision trees, Random Forests ,K-Nearest Neighbors and Naïve Bayes by training the model from the database of patients whose details of travel and acquiring of infection is known. The model was able to predict the type of transmission in the patients with 79.3% accuracy when trained on Support Vector Machine and produced detailed insights on the stage of covid-19 pandemic in India.*

Key Words: Corona-Virus, Transmission, Pandemic, India, Prediction, Machine Learning, Classifiers, Stages

1. INTRODUCTION

India witnessed its first COVID-19 case on 30th January 2020 when an Indian student returned from Wuhan, China [1]. All initial cases in the country were of people with a history of foreign travel or were close contact/family member of the infected person. For the next one month i.e. till 3rd of March India witnessed only 5 new cases which rose to just over 500 cases in the next 20 days and a lockdown was imposed country-wide to contain any further spread of the virus.[1] Although cases in India spiked from 27thof March when over 4000 suspected COVID-19 patients were recovered from a Delhi Religious congregation which had begun before the lockdown, cases began doubling every 4 ½ days as virus started spreading to every state of the country and subsequently the government began isolating individuals who came in contact of infected patients but for tracing of all the people who came in contact with them it was necessary to determine the cause of transmission in the patient at the first place. So far, efforts have been made in predicting the future cases based on the historical reported cases, analyzing the early epidemiology of COVID-19 using social media data, and modeling and investigating the effects of the case and contact-isolation in containing the COVID-19 outbreak. [2] However, it remains unclear how the disease transmits among the populations and what are the transmission patterns. This crucial information was missing from the database as some patients either were not aware of their cause of infection or weren't co-operating in sharing their details.

The database provided by the government of India till the 14th of April had over 8000 patients out of which data for around 3000 patients were well defined and complete. Features such as Age, Gender, State, Nationality, Date of infection, Status of the patient (Recovered, Deceased, Hospitalized), travel history, date of recovery/death of patients were thoroughly analyzed. Travel history/cause of infection feature plays a key role in our findings because by tracing people that the infected person met in his journey, be it of domestic/international travel or local transmission was a prime concern, in one particular case the agencies were able to isolate somewhere between 400-800 individuals that came in contact of the infected person in past one week. [3] These cases worked as primary cases that infected others. Hence, it becomes a thing of utmost importance to estimate the transmission dynamics in the initial days of infectious disease outbreak and generate predictions about the potential growth of cases. This prediction can provide insights into the epidemiology of the disease, which helps policymakers to check health system capacities. In some cases, it was found that as many as 60 people were infected by a single infected person. The methodology of contact tracing along with nation-wide lock-down has helped to keep the mortality rate low at 3.3% as compared to a global 7% mortality rate. In India, 17% of the world population resides in 2.4% area of the world, resulting in a densely populated country, which makes it a high-risk country for any infectious disease. The contribution of this paper is to explore the application of Machine learning for modeling the COVID-19 pandemic. This paper aims to investigate the generalization ability of the proposed ML models and the

accuracy of the proposed models for different lead-times. This paper contributes to the advancement of time series prediction of COVID-19. Consequently, an initial benchmarking is given to demonstrate the potential of machine learning for future research. With an in-depth characterization of age-specific social contact-based transmission, the retrospective and prospective situations of the disease outbreak, including the past and future transmission risks, the effectiveness of different interventions, and the disease transmission risks of restoring normal social activities, are computationally analyzed and reasonably explained.

2. METHOD OF STUDY

The raw data has been taken for a period of 45 days i.e. from March 1, till April 14 from the data-sharing portal covid19india.org. A processed version of this crowd-sourced database of individual patients of covid19 in India is used for studying the type of transmission. The process of data cleaning included handling of missing data, which was done by a rigorous survey of statistics given in Health ministry briefings and health bulletins of various state governments. The adopted methodology ensured that the integrity and authenticity of data is maintained for the study.

The missing values of input feature space like gender and age were handled by self-defined functions. These functions resolute to handle null or invalid data and filled them with values deduced from government statistics. The output column was prepared by extracting keywords from the textual patient details mentioned in the original data base. This was done using tools from NLTK library in python.

Prior to the imposition of lockdown in India the carriers of the Covid-19 virus could broadly be classified into three categories. Thus the problem of identifying the cause of infection in an individual reduces to a multi-class classification problem.

A classification algorithm captures pattern in training data using structured mathematical and statistical concepts. The algorithm generates a predictive model based on the training data which is used to predict a class label for an example in testing data. For the experimental analysis this paper studies five classification algorithms namely, logistic regression, random forest, Naïve Bayes, K-nearest neighbor and support vector classifier for highlighting the cause of transmission.

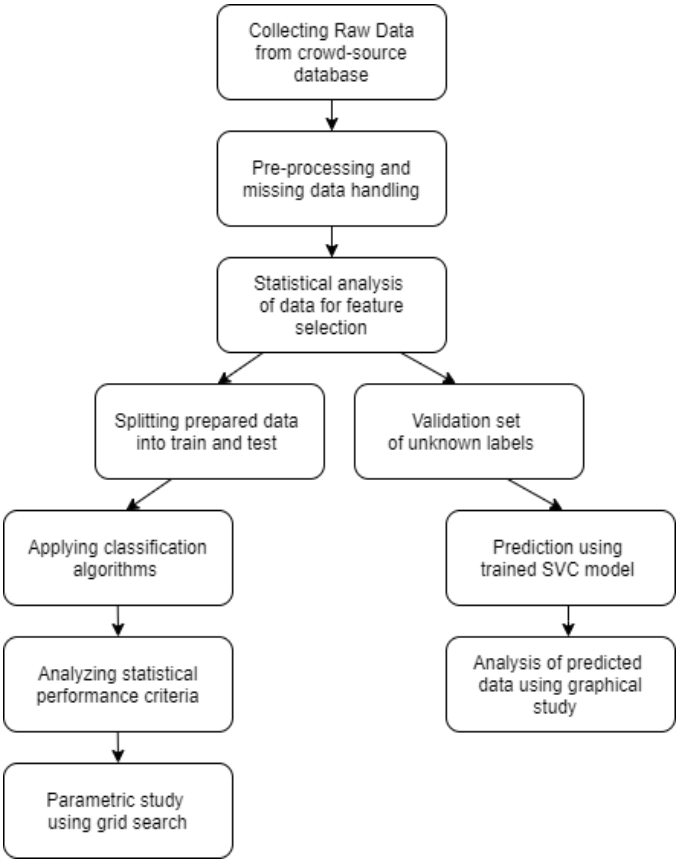


Fig 1- Proposed framework of study

3. DATASET DESCRIPTIONS

The methodology in this paper, uses classification algorithms to build an understanding of the type of transmission in Covid-19 patients of India. The correlation of features with the output data was studied to select the relevant feature space for training the classification model. The input feature set includes age, gender, nationality, state and pandemic day associated with the reported case. The feature set description is as follows:

1. AGE: On the basic demographics of India the age has been divided into four subgroups.

Age Group	Category	Label
0-20	Young age	1
20-40	Prime working age	2
40-60	Mature working age	3
60 plus	Old age	4

2. GENDER: It is a binary class representing the male and female category.

Category	Label
Male	0
Female	1

3. NATIONALITY: It is a binary class representing Indian and foreign national.

Category	Label
Indian	0
Foreign	1

4. STATE: All 28 states and 8 union territories of India are classified into four buckets based on the urban population in the state and the covid-19 trend in the area.

Category	Label
Top affected	1
Moderately affected	2
Less affected	3
Negligibly affected	4

5. PANDEMIC DAY: The column is a time series data indicating the day count of the Covid-19 pandemic.

A model is trained to predict the most probable reason out the three categories viz. International travel, Domestic travel or Local transmission. The study helps understand better about reproduction number which is defined as the average number of new infections generated by one infected individual during the entire infectious period in a fully susceptible population.

TOOLS USED

Python was selected due to its well documented resources for libraries that can be used for data analysis, visualization and classification. The main libraries imported were scikit-learn, nltk, matplotlib, pandas and numpy.

SETUP

Training data for the experiment is of the form $\{(\vec{x}_i, y_i): 1 \le i \le n\}; \vec{x}_i \in \mathbb{R}^5$, $y_i \in \{1, 2, 3\}$ and n is the number of data points in training dataset.

The experiment aims to find a function (h) such that $(h): \mathbb{R}^5 \rightarrow \{1, 2, 3\}$ which outputs a correct label.

4. ALGORITHMS

Machine learning techniques have been extensively used to study and predict outbreak infection from a long time and the results are promising as table-1 lists the major work done by authors on infectious diseases using machine learning models.

Authors	Journal	Outbreak infection	Machine learning Model
[8]	Trans-boundary and Emerging Diseases	Swine fever	Random Forest
[9]	Geospatial Health	Dengue fever	Neural Network
[10]	BMC Research Notes	Influenza	Random Forest
[11]	Journal of Public Health Medicine	Dengue/Aedes	Bayesian Network
[12]	Global Ecology and Biogeography	H1N1 flu	Neural Network
[13]	Current Science	Dengue	Adopted multi-regression and Naïve Bayes
[14]	Infectious Disease Modeling	Dengue	Classification and regression tree (CART)

Table 1 – Notable Ml methods for outbreak prediction

4.1 Logistic Regression is a binary classifier which can be extended for multiclass classification using one-versus-rest approach or one-versus-one approach. In our problem statement, one-versus-rest approach is used to classify outputs into three labels.[21]The proposed idea uses

sigmoid function to essentially form a linear decision boundary to distinguish between the output classes. Two classifiers $f1$ and $f2$ are trained such that, $f1$ classifies 1 vs {2, 3} and $f2$ classifies 2 vs {1, 3}. The points that are not classified by $f1$ and $f2$ are put into class 3. For training the classifiers hyper-parameter like solver, penalty, regularization constant (C) and multi-class model were configured as listed in table 2.[16],[17]

Hyper-parameter	Value
Solver	liblinear
C	5
multi_class	ovr
Penalty	l2

Table 2 – Logistic regression hyper-parameters

4.2 Gaussian Naïve Bayes is probabilistic multiclass classification algorithm that assumes value of features follow normal distribution. It is based on an strong assumption that the features in the dataset are strictly independent of each other for example it assumes pandemic day count has zero correlation with state, gender, age and nationality for any given input data. In Gaussian naïve Bayes the likelihood of features is assumed to be Gaussian, where μ and σ are the mean and variance of the continuous x computed for a given class 'c' of y.

$$P(x_i|y=c) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}}$$

In our study it was counter intuitive that despite the feature space being discrete values the Gaussian Naïve Bayes algorithm outperformed the classical Multinomial Naïve Bayes approach.[18]

4.3 Random forest is an ensemble approach which avoids over fitting by combining large number of decision trees which are formed using subset of training data and feature space.[19] For our application 75 trees along with 3 features are considered to grow a tree at each node, an individual tree in the forest forms weak or weakly correlated classifiers in itself.[20] The result from each decision tree is aggregated by the random forest classifier and the most voted class is predicted as the cause of transmission. Set of hyper-parameters used for training the model is listed in table 3.

Hyper-parameter	Value
max_depth	7
n_estimators	75
max_features	sqrt
Criterion	entropy

Table 3 – Random Forest hyper-parameters

4.4 K-NN has been studied to understand the problem using a non-parametric generalized instance based approach. For training the model the data has been scaled using standard scalar to avoid any predominant bias for any input feature. The training data samples are used as instances which are mapped as vectors in a 5 dimensional space. In the experiment optimum value of K is selected as 15, the output class for a new instance is decided by computing Manhattan distance between new data and training instances. The class with the highest frequency from the 15 most similar instances is selected as output label.[22] The analysis with KNN discarded the idea of explicit generalization of learning a prediction function f: x→y and only two hyper-parameters viz. value of K and distance function were tuned to generate the final model.

The algorithms analyzed by far relied on finding linear or probabilistic relation in data for classifying a new instance this necessarily gave good fit on training data but still left the generalization poor. Support vector machines are potential supervised learning models for analyzing nonlinear patterns because of the ability of implicitly mapping inputs into high-dimensional feature spaces.

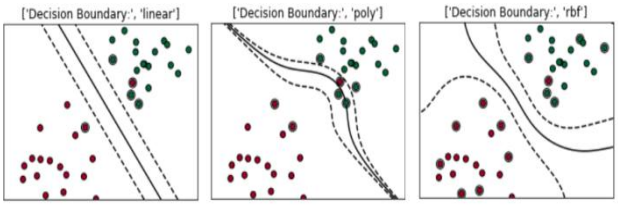
4.5 Support Vector Classifier, SVC inheritably is a binary classifier with a multiclass formulation using one vs. one or one vs. rest approach. The algorithmic approach used for SVC is a non-probabilistic idea of finding maximum margin hyper-plane for the training dataset. In a p dimensional space, a hyper-plane is a flat affine subspace of dimension p - 1. For p > 3 dimensions are hard to visualize, but are still of p -1 dimensional flat subspaces. [4] Figure 2 depicts decision boundary for a typical two dimensional input space. In our study an input data point is viewed as a 5 dimensional vector which is mapped to a possible high-dimensional feature space using kernel function $K(x_i, x_j)$ a hyper-plane in the transformed feature space may be nonlinear in the original input space.[23]

A typical applied kernel for the support vector classifier is the linear kernel, which is defined as

$$K(x_i, x_j) = (x_i^T)(x_j)$$

Hence linear kernel is an inner product between the input vectors. Another well-known kernel is the Gaussian kernel (or RBF-kernel) with γ as hyper-parameter. Using this kernel the support vector classifier is basically finding discriminating dimensions in an infinite feature space.

$$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right)$$



Hyper-parameters tuned for training the model includes selecting the right kernel function out of linear, radial bias function and polynomial, regularization parameter (C) which optimizes the model for the degree of avoidance of misclassification of training example and the gamma parameter defines how far the influence of a single training example reaches in deciding the maximum margin hyper plane. Table 4 lists the best hyper parameter set for the problem found using grid search algorithm. The algorithm is based on extended search on manually specified subset of hyper-parameter space it uses cross validation on the training dataset as performance metric. The optimal set which results in the most accurate predictions is selected for training the final model.[6]

Hyper-parameter	Value
Gamma	0.25
C	3
Kernel	RBF
Decision_function_shape	ovr
Probability	False

Table 4 – Support Vector classifier hyper-parameters

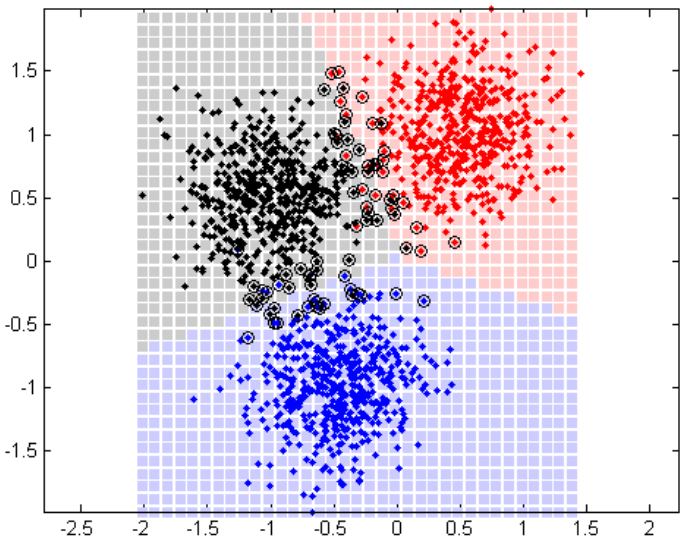


Fig 2– Support Vector classification for two dimensional input space

The performance of the trained SVC classifier is analyzed using popular ROC curve which depicts the pronounced details about the behavior of the classifier. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC curve is mainly used to study binary classifiers; figure 3 shows the ROC curve for individual classifiers trained using one versus rest approach.

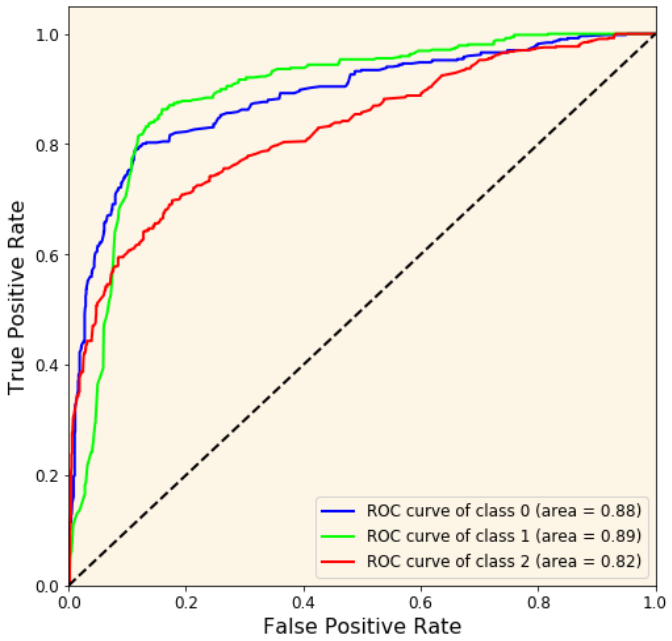


Fig 3 – ROC curve for classifiers trained using one versus rest approach

Figure 4 shows that the AUC of both micro and macro averaging is 0.87 which concluded that the trained model has good performance characteristics.

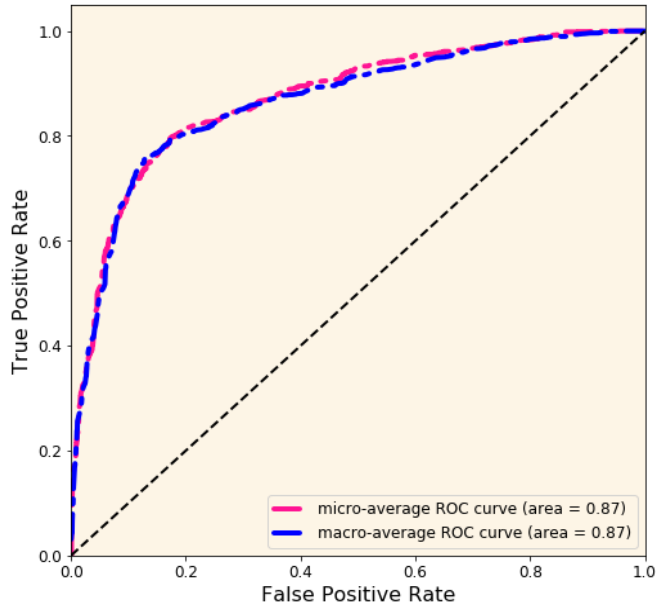


Fig 4– Micro and Macro Averaging ROC curve for SVC Classifier

For our focus of study to extend ROC curve and ROC area to multi-class classification we binarized the output in to label indicator matrix.[15] Micro-averaging technique of drawing ROC curve with each element of the label indicator matrix as a binary prediction has been used as one method while macro-averaging, technique which allots equal weight to the classification of each label is also studied.

Holdout method of using unseen testing data is also used to quantify the measure of performance for all previously discussed algorithms. Statistical parameters including accuracy, entropy loss, F1-score and precision-recall is analyzed using confusion matrix.[16] Table 5 compares the performance of all classifier. We conclude in our study that Support Vector Classifier gives the best mathematical formulation for studying the reason of transmission in Covid19 cases in India.

ALGORITHM	F1-MEASURE	TRAIN ACCURACY	TEST ACCURACY
Logistic Regression	0.60	0.62	0.59
Gaussian Naïve Bayes	0.62	0.66	0.63
Random Forest	0.70	0.78	0.72
K-NN	0.71	0.77	0.7
SVC	0.8	0.81	0.793

Table 5 – Comparison of performance metrics

5. DISCUSSION AND RESULTS

The quantitative estimates help measure the impact of isolation of individuals who came in contact with infected person with travel history or local contact this assists in reducing morbidity and peak infection rates. The mathematical predictive model of disease transmission highlights a social contact structure which can be studied using graphical representation .Figure 5shows how transmission pattern of covid-19 changed as the pandemic progressed for a period of over 70 days since the first case. The graph depicts information for about 3600 patients whose cause of infection was initially classified using records in the original database. Out of 8000 infected people till 14th of April 2020 all initial cases were linked to international travel. It can be understood by the graph that after a certain point when government banned all flights from foreign nations the green curve began to flatten but the locally transmitted cases kept growing exponentially.

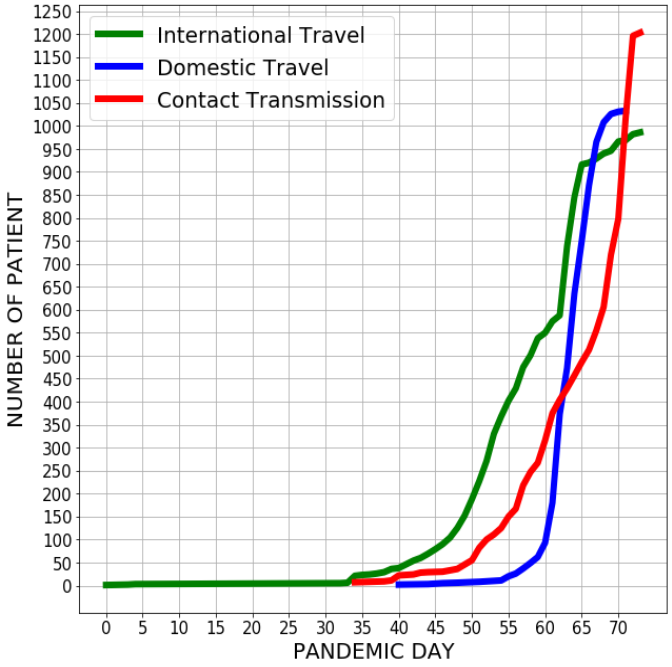


Fig 5 –Type of transmission for known data

With the identification of over 500 positive cases in the country, a nation-wide lockdown was imposed yet the spread of local infection kept growing due to unidentified patients that had travelled from abroad and weren't symptomatic at airport screening. The remaining of 4400 patients whose details of infection weren't known, were analyzed by the predictions made from the trained SVC model. Figure 6 depicts this predicted data for transmission among the patients wherein the cases of local/contact transmission showed a sudden spike while foreign travel patients gradually decreased post banning of all commercial flights.

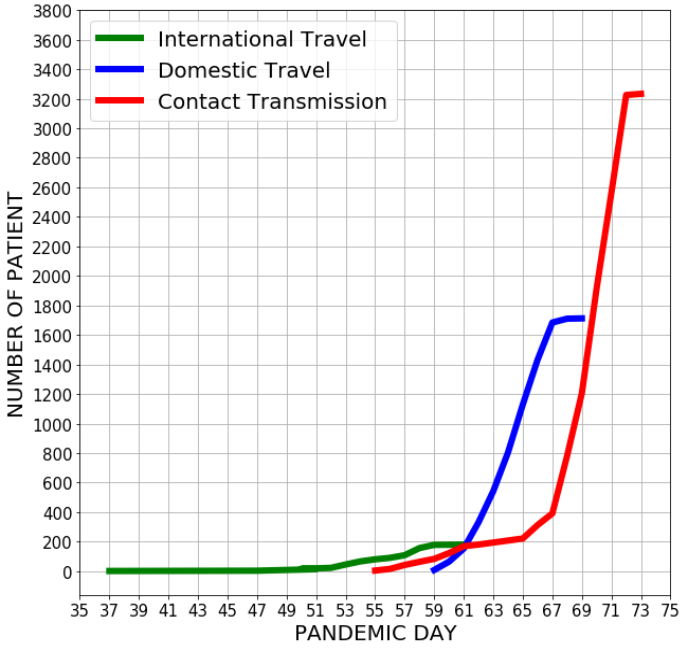


Fig 6 –Type of transmission for Predicted Data

The inter-state travel prior to the lockdown has also led to the increased COVID cases being identified in the country as seen by the blue curve.

The Bar graph in Figure 7 highlights the distribution of patients infected of the two genders for the three category of transmission. Data shows that about 3/4th of the total infected patients in India are male and rest female. The disproportional figure is due to the fact that nearly 1/3rd of the total infected people till 14th of April were the attendees of a Delhi religious congregation ,which were mostly males.

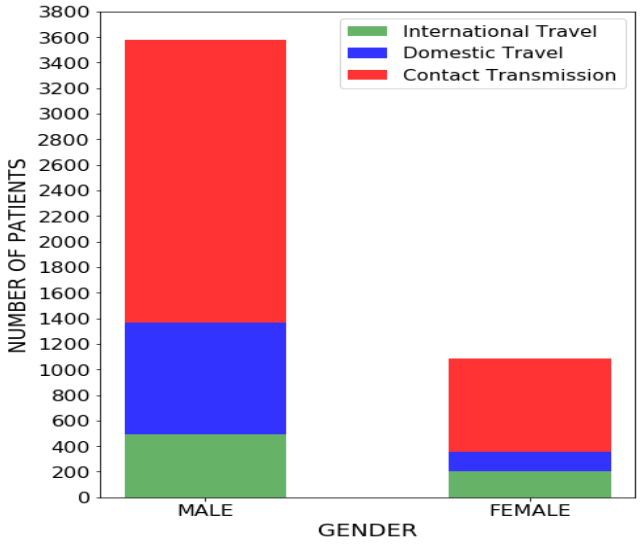


Fig 7 – Distribution of type of transmission across gender for known data

The prediction model results in Figure 8 shows an exponential increase in contact transmission among both genders while there was a slight increase in cases of domestic travel infection and a decline in patients with foreign travel.

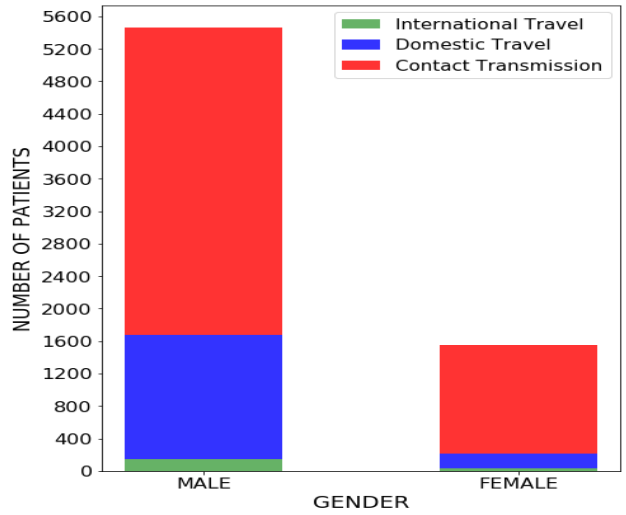


Fig 8 – Distribution of type of transmission across gender for predicted data

India has more than 50% of its population below the age of 25 and more than 65% below the age of 35. As this age group forms the major working class of the country, maximum cases of COVID-19 were from the age group of 20-40. Figure 9 and 10 shows a distribution of real data and predicted data respectively for different age groups and number of cases for each type of transmission.

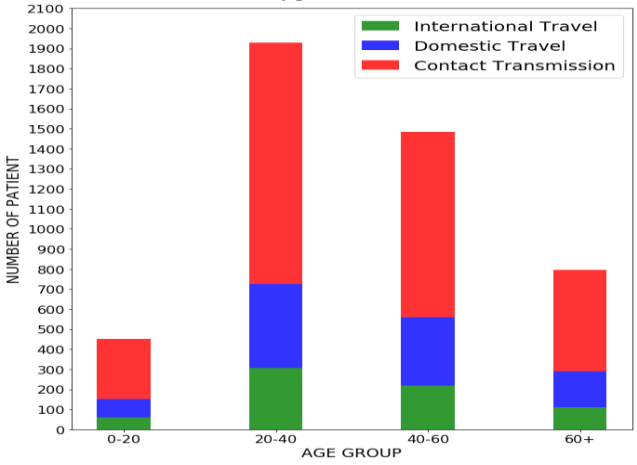


Fig 9 – Distribution of type of transmission across age groups for known data

Prediction for different age groups shows that the maximum increases in cases of local transmission was from the prime working age group (20-40) and least for the youngsters in age 0-20 as schools and universities across the nation were shut down a week prior the official nationwide lockdown.

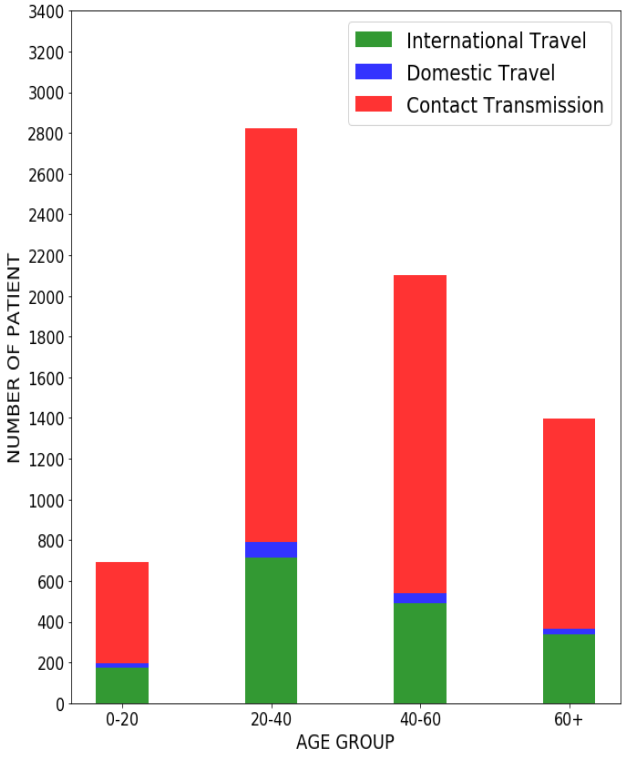


Fig 10 – Distribution of type of transmission across age groups for predicted data

The data indicates that although strict lockdown was imposed but due to the prior lifestyle of the states with large urban population density faced difficulty in combating the increase in cases. The four state groups formed for the study were based on the urban density and the prevailing Covid-19 trend of the state. The interactive work life in states like Maharashtra, specifically cities like Mumbai and the capital New Delhi are at high risk of vast community spread. Figure 11 indicates the level of community spread, the red portion in the bar graph represents all the local cases. Here Group 1 & 2 refers to the states like Maharashtra, Delhi, Tamil Nadu, Gujarat, Uttar Pradesh were daily life activities, public interaction and urban population is intense.

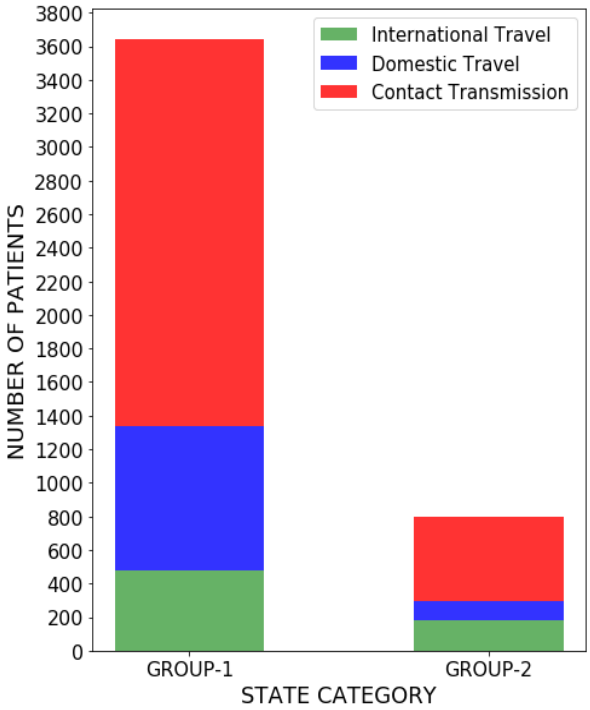


Fig 11 – Distribution of type of transmission across state group 1 and 2 for known data

Group 3 and 4 in Figure12 refers to states like Goa, Andaman and Nicobar Islands Ladakh,Uttarakhand, Jammu and Kashmir which have a lower population density and these states exhibited lesser cases of community spread.

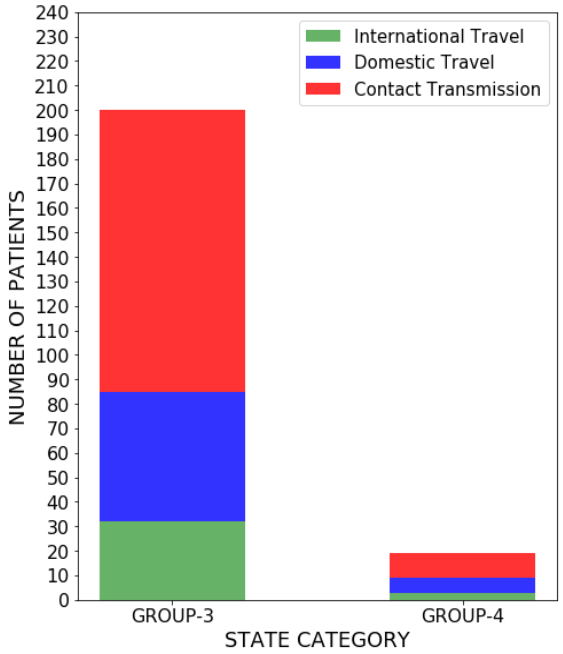


Fig 12 – Distribution of type of transmission across state group 3 and 4 for known data

Figures 9 and 10 describes the growth of cases for the three types of transmission using the SVC model for 4400 patients across the four-state groups.

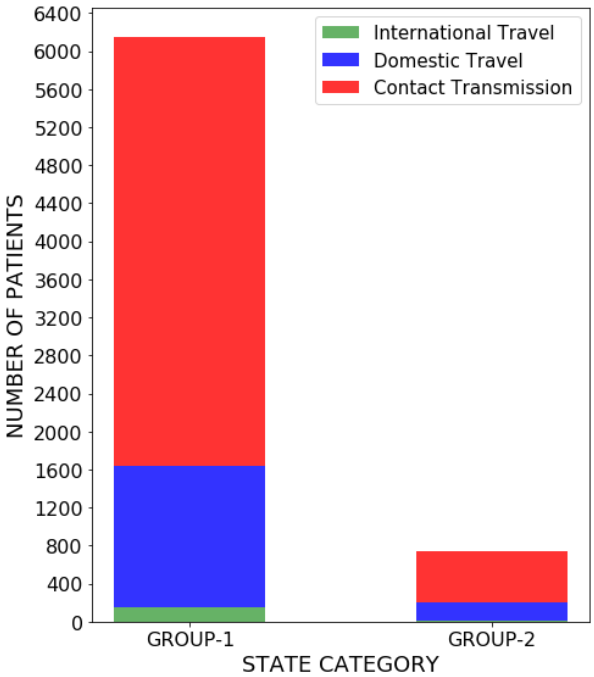


Fig 13 – Distribution of type of transmission across state group 1 and 2 for predicted data

While the cases increased in densely populated states and metropolitans, cases in low population states were under control due to strict lockdown which prohibited migration of

people from other states as well as promoted isolation of people that prevented further spread.

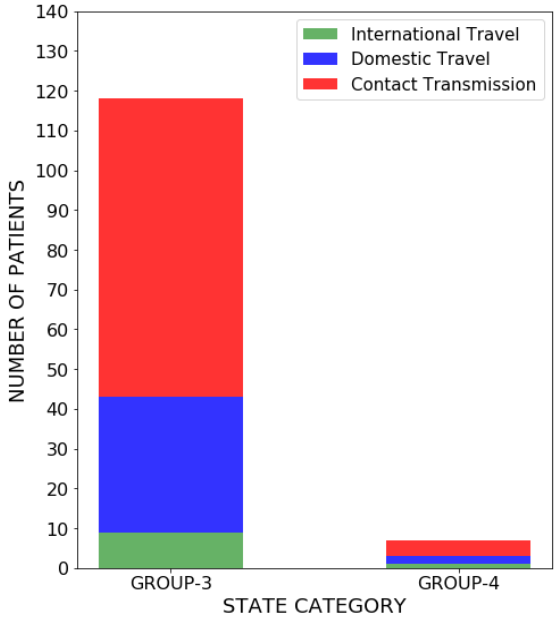


Fig 14 – *Distribution of type of transmission across state group 3 and 4 for predicted data*

6. CONCLUSIONS

The study on Covid-19 patient's dataset has prompted some definitive patterns in human to human transmission as the pandemic progressed and these results can prove helpful in planning preparedness for future pandemic or regional outbreak of any infectious disease. These patterns show that although all the initial patients were infected due to foreign travel and were isolated still individuals who came in their close proximity could carry this disease locally to other parts of the nation; hence detecting the cause of infection in any arbitrary patient could prove to be a difficult task. Although this model predicts with near 80% confidence on how a patient got infected with no background information of the patient, it can give involved agencies an upper hand in tracing down all potential factors as the search area would be reduced and would further lead to a reduction in the testing of people who are not suspected to be the carrier of the disease. The state-wise distribution of patients concludes a rapid spread of infection due to local transmission in states with a higher fraction of urban population; therefore cities like Delhi and Maharashtra should be on key focus for tracing and isolating individuals. The use of rapid antibody testing of suspected patients is beneficial as it would consume lesser time as the incubation period for the Corona-virus ranges from 1 to 12.5 days (with median estimates of 5 to 6 days), but can be as long as 14 days hence an infected person who doesn't show symptoms during this period may infect a large population. The incubation period of the virus is the time between the exposure and the display of symptoms. The data also shows that the recovery time period of a corona-virus patient ranges from 14-21 days for mild cases and can be up to 6 weeks for severe cases.[5],[7] For the classification of 3 categories of patients, SVC model outperformed other methods and showed promising results on training data as well as testing data. The model showed results in terms of predicting the time series without the assumptions that epidemiological models require. Machine learning model, as an alternative to epidemiological model, has shown potential in predicting COVID-19 transmission. The advancement of higher performance models for long-term prediction, future research should be devoted to comparative studies using various ML models for individual countries.

REFERENCES

[1] Schueller, E, Klein E , Lin G, Tseng K, BalasubramanianR, Kapoor G ,Joshi J, SriramA, Nandi, A ,LaxminarayanR"COVID-19 in India: Potential Impact of the Lockdown and Other Longer-Term Policies"The Center For Disease Dynamics, Economics & Policy April 20, 2020

[2] Lixiang L, Zihang Y, ZhongkaiD,CuiM,Jingze H, Haotian M, Deyu W, Guanhua C, Jiaxuan Z, Haipeng P, Yiming S " Propagation analysis and prediction of the COVID-19"Infectious Disease ModellingVolume 5, 2020, Pages 282-292

[3] Yang L , Zhonglei G ,Shang X ,Benyun S ,Xiao-Nong Z, Yong S "What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization", E Clinical medicine Volume 22, 100354, May 01, 2020

[4] Burges. C A tutorial on support vector machines for pattern recogni- tion. Data Mining and Knowledge Discovery, to appear, available at http ://svm. research, bell-labs, com/SVMdoc, html.

[5] Wen-Hua K, Yao L, Ming-Wei P, De-Guang K, Xiao-Bing Y, LeyiWa& Man-Qing L ."SARS-CoV-2 detection in patients with influenza-like illness" Nature micro-biology pages 675–678(2020)

[6]Chih-Wei H and Chih-Jen L "A Comparison of Methods for Multi-class Support Vector Machines" , available at https://www.csie.ntu.edu.tw/~cjlin/papers/multisvm.pdf

[7] Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet**395**, 689–697 (2020)

[8] Liang, R., Lu, Y., Qu, X., Su, Q., Li, C., Xia, S., Liu, Y., Zhang, Q., Cao, X., Chen, Q., et al. "Prediction for global African swine fever outbreaks based on a combination of random forest

algorithms and meteorological data."Transboundary Emer. Dis. 2020, 67, 935-946, doi:10.1111/tbed.13424.

[9] Anno, S., Hara, T., Kai, H., Lee, M.A., Chang, Y., Oyoshi, K.; Mizukami, Y.,Tadono, T. Spatiotemporal dengue fever hotspots associated with climatic factors in taiwan including outbreak predictions based on machine-learning. Geospatial Health 2019, 14, 183-194, doi:10.4081/gh.2019.771.

[10] Tapak, L., Hamidi, O., Fathian, M.,Karami, M. Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare centers in Iran. BMC Res. Notes 2019, 12, doi:10.1186/s13104-019-4393-y

[11] Raja, D.B.,Mallol, R., Ting, C.Y., Kamaludin, F., Ahmad, R., Ismail, S., Jayaraj, V.J.,Sundram, B.M. Artificial intelligence model as predictor for dengue outbreaks. Malays. J. Public Health Med. 2019, 19, 103-108.

[12] Koike, F.; Morimoto, N. Supervised forecasting of the range expansion of novel non-indigenous organisms: Alien pest organisms and the 2009 H1N1 flu pandemic. Global Ecol. Biogeogr.2018, 27, 991-1000, doi:10.1111/geb.12754.

[13] Agarwal, N.; Koti, S.R.; Saran, S.; Senthil Kumar, A. Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India. Curr. Sci. 2018, 114, 2281-2291, doi:10.18520/cs/v114/i11/2281-2291

[14] Titus Muurlink, O.; Stephenson, P.; Islam, M.Z.; Taylor-Robinson, A.W. Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach. Infect. Dis. Modelling 2018, 3, 322-330, doi:10.1016/j.idm.2018.11.004.

[15] Andrew P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms",Volume 30, Issue 7, July 1997, Pages 1145-115

[16] Elkan, Balakrishnan N, Zachary C. Lipton, Charles "Optimal Thresholding of Classifiers to Maximize F1 Measure",Machine Learning and Knowledge Discovery in Databases, 2014, Volume 8725

[17] Andrew Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance", ICML '04: Proceedings of the twenty-first international conference on Machine learning July 2004

[18] Perez A., Larranaga P., Inza I., "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes",International Journal of Approximate Reasoning, Volume 43, Issue 1, September 2006, Pages 1-25

[19] Liaw, Andy, and Matthew Wiener."Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[20] Strobl, Carolin, et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution." BMC bioinformatics 8.1 (2007): 25.

[21] Peng, Chao-Ying Joanne, KukLida Lee, and Gary M. Ingersoll."An introduction to logistic regression analysis and reporting." The journal of educational research 96.1 (2002): 3-14.

[22] Raikwal, J. S., and KanakSaxena. "Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set." International Journal of Computer Applications 50.14 (2012)

[23] Har-Peled, Sariel, Dan Roth, and DavZimak. "Constraint classification for multiclass classification and ranking." Advances in neural information processing systems. 2003.