Original Research Article

Machine learning based prognostic model for predicting infection susceptibility of COVID-19 using health care data

Authors
R Srivatsan, Student, School of Electronics Engineering, VIT University
Email: r.srivatsan2017@vitstudent.ac.in
Prithviraj N Indi, Student, School of Electrical Engineering, VIT University
Email: prithvirajn.indi2017@vitstudent.ac.in
Swapnil Agrahari, Student, School of Mechanical Engineering, VIT University
Email: swapnil.agrahari2017@vitstudent.ac.in
Siddharth Menon, Student, School of Electronics Engineering, VIT University
Email:siddharth.menon2018@vitstudent.ac.in

Corresponding Author
Dr. S. Denis Ashok
Professor
School of Mechanical Engineering
Department of Design and Automation
VIT University,
Vellore, Tamil Nadu, 632014, India
Emai ID: denisashok@vit.ac.in
Orcid ID: 0000-0002-0276-4712

Abstract

From public health perspectives of COVID-19 pandemic, accurate estimates of infection severity of individuals are extremely valuable for the informed decision making and targeted response to an emerging pandemic. This paper presents machine learning based prognostic model for providing early warning to the individuals for COVID-19 infection using the health care data set. In the present work, a prognostic model using Random Forest classifier and support vector regression is developed for predicting the susceptibility of COVID-19 infection and it is applied on an open health care data set containing 27 field values. The typical fields of the health care data set include basic personal details such as age, gender, number of children in the household, marital status along with medical data like Coma score, Pulmonary score, Blood Glucose level, HDL cholesterol etc. An effective preprocessing method is carried out for handling the numerical, categorical values (non-numerical), missing data in the health care data set. Principal component analysis is applied for dimensionality reduction of the health care data set. From the classification results, it is noted that the random forest classifier provides a higher accuracy as compared to Support vector regression for the given health data set. Proposed machine learning approach can help the individuals to take additional precautions for protecting against COVID-19 infection. Based on the results of the proposed method, clinicians and government officials can focus on the highly susceptible people for limiting the pandemic spread.

Keywords: Machine Learning, Prognostics, COVID-19, infection susceptibility, PCA, random forests, support vector regression

Machine learning based prognostic model for predicting infection susceptibility of COVID-19 using health care data

R Srivatsan[1], Prithviraj N Indi[1], Swapnil Agrahari[1], Siddharth Menon[2], S Denis Ashok[2]

VIT University, Vellore, Tamil Nadu, 632014, India

Abstract

From public health perspectives of COVID-19 pandemic, accurate estimates of infection severity of individuals are extremely valuable for the informed decision making and targeted response to an emerging pandemic. This paper presents machine learning based prognostic model for providing early warning to the individuals for COVID-19 infection using the health care data set. In the present work, a prognostic model using Random Forest classifier and support vector regression is developed for predicting the susceptibility of COVID-19 infection and it is applied on an open health care data set containing 27 field values. The typical fields of the health care data set include basic personal details such as age, gender, number of children in the household, marital status along with medical data like Coma score, Pulmonary score, Blood Glucose level, HDL cholesterol etc. An effective preprocessing method is carried out for handling the numerical, categorical values (non-numerical), missing data in the health care data set. Principal component analysis is applied for dimensionality reduction of the health care data set. From the classification results, it is noted that the random forest classifier provides a higher accuracy as compared to Support vector regression for the given health data set. Proposed machine learning approach can help the individuals to take additional precautions for protecting against COVID-19 infection. Based on the results of the proposed method, clinicians and government officials can focus on the highly susceptible people for limiting the pandemic spread.

Methods In the present work, Random Forest classifier and support vector regression techniques are applied to a medical health care dataset containing 27 variables for predicting the susceptibility score of an individual towards COVID-19 infection and the accuracy of prediction is compared. An effective preprocessing is carried for handling the missing data in the health care data set. Principal Component Analysis is carried out on the data set for dimensionality reduction of the feature vectors.

Results From the classification results, it is noted that the Random Forest classifier provides an accuracy of 90%, sensitivity of 94% and specificity of 81% for the given medical data set.
Conclusion Proposed machine learning approach can help the individuals to take additional precautions for protecting people from the COVID-19 infection, clinicians and government officials can focus on the highly susceptible people for limiting the pandemic spread.

Keywords Machine Learning, Prognostics, COVID-19, infection susceptibility, PCA, random forests, support vector regression

Introduction

The recent outbreak of coronavirus disease 2019 (COVID-19) has created a great challenge for the healthcare system (Hui etal , 2020). Considering the lethal nature of COVID-19 outbreak and its worldwide spread, World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC) at different nations have provided provisional guidelines for protecting people from getting affected and preventing the further spread of COVID-19 virus from infected individuals. RT-PCR tests from deep nasotracheal samples and Chest CT scan are commonly used for definitive diagnosis of COVID-19. (Repici A et al. 2020). Due to the quick spread of COVID-19, physicians in the health care systems are facing extreme difficulty in the physical examination and analysis of subsequent para

clinical health care data for the accurate diagnosis of COVID-19. Hence, it is necessary develop software tools for easier way for interpreting the large scale health data set which can help the government and healthcare officials for quicker decision making during the Covid-19 pandemic situations.

With the capability of interpreting the hidden and complex patterns from huge, noisy or complex data, Artificial intelligence and machine learning techniques can play a major role in combating the COVID-19 pandemics. Few works have been reported by the researchers on use of machine learning techniques for the prediction and diagnosis of epidemics (Wynants L et al ,2020). An artificial intelligence based rapid diagnosis approach for COVID-19 patients developed using the analysis of Chest X ray images (Mei et al, 2020). An artificial intelligence based prediction model of the epidemics trend of COVID-19 is proposed by Yang etal, 2020. Linear Regression model is used for time series prediction of COVID-19 outbreak (Pandey et al, 2020). Mechanistic models have been reported to predict COVID-19 outbreak in real time (Liu etal, 2020). K means algorithm is applied to categorize the countries based on the number of confirmed COVID-19 cases (Carrillo-Larco et al, 2020). XGBoost machine learning model is proposed to estimate the survival ratio of severely ill Covid-19 patients (Yan et al. 2020). A classification using Fourier and Gabor methods is applied on dataset of COVID-19(Al-Karawi et al 2020). Multi Layered Perception (MLP), Adaptive Network-based Fuzzy Inference System is used for predicting (Metsky et al. 2020). Support vector machine is applied to detect severely ill COVID patients from mild symptom COVID patients (Tang et al.2020). Convolutional neural network frameworks have been proposed to detect COVID-19 from chest X-ray images. (Narin et al, 2020). A prediction model for the propagation analysis of the COVID-19 is proposed by Li etal 2020. An interpretable mortality prediction model for COVID-19 patients is developed using the health care data set (Lan etal, 2020).

It is found that many machine learning approaches has been successfully implemented for the prediction and diagnostic purposes of COVID-19 using the clinical and health care data. However, prognostic frame works for early prediction of COVID-19 infection are found to be limited which can be helpful to take proactive measures to combat the virus spread. Random forest and Support vector machine algorithms are found to be popular in achieving the satisfactory results for the different prediction applications. Hence, this paper presents Random forest and Support vector machine algorithms based prognostic approach for predicting the infection susceptibility score for each individual using the health care data. The novelty of the proposed approach is the identification of the infection susceptibility prior to infection so that the regulative and preventive rules can be made for the individuals.

Methods

In the present work, a prognostic approach is formulated using the machine learning techniques such as random forest and support vector regression for predicting the susceptibility score of COVID-19 infection as Low, medium, high. The frame work of the proposed prognostic approach is shown in Fig. 1. The major elements of the proposed method are described briefly as follows:

- A comprehensive data collection system is the base for the proposed method. Dataset includes special features like comorbidity conditions and frequency of Foreign Trips. As the medical data contains the missing values, an effective preprocessing is essential and it is carried out before it is applied to the machine learning models for the classification applications.
- Machine learning techniques consisting of Random Forests (RF) classifier and Support vector Regression for predicting infection susceptibility score of COVID-19 as discrete levels of risk factors namely - High risk (66%-100%), Medium risk (33%-66%) and Low risk (0%-33%). These classes are assigned as the targets for initiating the training process.

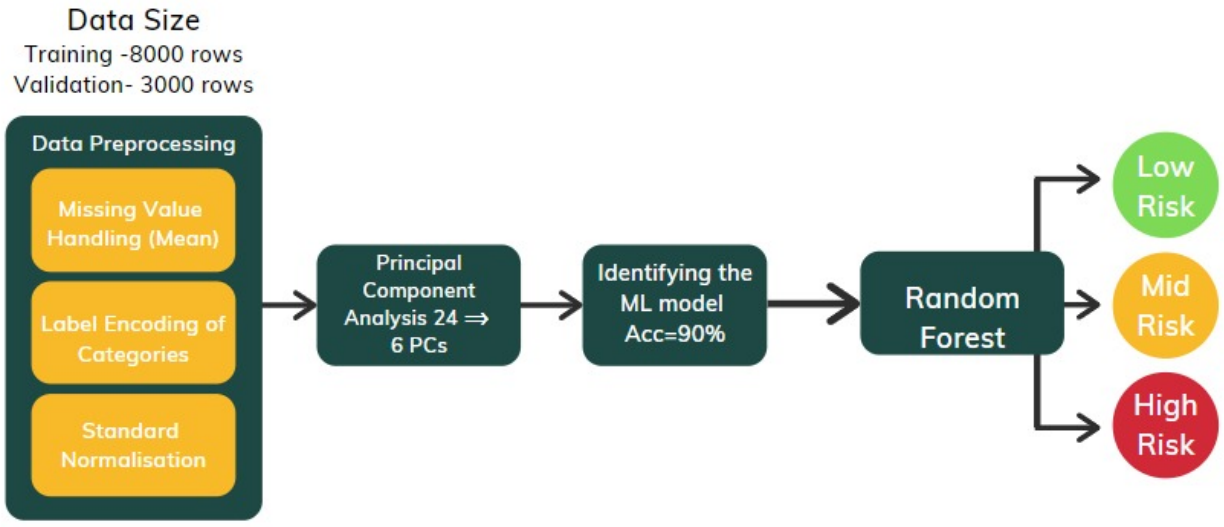


Fig 1 *Proposed prognostic approach for predicting the susceptibility score of individuals using health care data.*

Description of medical health care dataset

In the present work, open health care data set available in the online repository Kaggle is used for demonstrating the proposed prognostic approach .The data set contains 14498 rows and 27 columns. Table.1 shows the typical fields of the health care data set which include basic personal details such as age, gender, number of children in the household and marital status along with medical data like Coma score , Pulmonary score, Blood Glucose level, HDL cholesterol. Medical data chiefly includes comorbidity conditions such as Severe Acute Respiratory Infections (SARI), diabetes and heart syndromes. Vitals such as heart rate have also been considered in the modeling of the predictor infection susceptibility of COVID-19.

Data Preprocessing and Preparation

It is noted that open health care data contains many numerical and categorical values (non-numerical) and many machine learning algorithms cannot handle data in this form. Also, there can be missing values in the relevant fields of the data set and most machine learning algorithms don't support the missing values. Hence, data preprocessing and preparation is essential as the missing data would lead to inaccurate results. They are converted to numerical values using Label encoding to achieve accurate results using the machine algorithms. Further, heat map and principal component analysis is followed for preprocessing the data set.

Data Normalization:

Standardization of the dataset makes a very crucial role in the pipeline of the ML model since if the individual features do not reassemble standard normal distribution of data points, the model would become erratic in its predictions. This involves a technique of reducing mean value from each individual data point and performing a scaling operation to them in order to obtain unit variance per cell of the data.

$$x' = \frac{x-\mu}{\sigma} \quad (1)$$

Where μ denotes data's mean value and σ denotes the standard deviation obtained as square root of variance.

Table 1 *The various fields of the medical and health care data* **(**Srijan Singh 2020)

Variables	Description
people_ID	Unique ID for each person
Region	The area that the person belongs to
Gender	Gender of the person
Designation	Designation of the person
First_Name	Name of individual
Married	Marital status of individual
Children	Number of children in the family
Occupation	Sector of individual occupation
Mode_transport	Mode of transport that the individual frequently chooses to travel
cases/1M	Number of confirmed cases per 1 million population in that region
Deaths/1M	Number of Death case per 1 million population in that region
comorbidity	Co occuringmedical condition
Age	Age of the person
Coma score	Neurological coma score
Pulmonary score	Pulmonary PaO2 (mmHg)/FiO2
cardiological pressure	Cardiological Mean systolic Arterial pressure (mmHg)
Diuresis	Diuresis in mL/Day
Platelets	Hematological Platelets 10/L
HBB	Hepatic Blood bilirubin (µmol/L)
d-dimer	d-dimer concentration in the blood (ng/ml)
Heart rate	number of times a person's heart beats per minute
HDL cholesterol	High-density lipoprotein level (milligrams per decilitre)
Charlson Index	index for a patient who may have any of the listed comorbid disease conditions
Blood Glucose	strength of glucose present in the blood (millimoles per litre)
Insurance	Medical Insurance spending cover (in Rs.)
salary	Annual salary of the individual
FT/month	Average foreign trips taken by the individual per month, considering last 2 year data

Dimensionality reduction using Principal Component Analysis

As the medical and health care data set contains 27 fields, a dimensionality reduction is followed using Principal Component Analysis (PCA) which converts given features into 6 principal components (PC). Here the PCs indicate the reduced representation capturing maximum variance of the information and simultaneously reducing the dimensionality of the data.

In the first step, the mean of the values of each column or field is calculated and it is followed by finding covariance through the following equation (2).

.

$$\mathrm{cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \bar{x})(Yi - \bar{y}) \quad (2)$$

Where Xi and Yi are the individual data points and the $\bar{x}$ and $\bar{y}$ refer to the mean values of the two fields chosen at a time. 'n' refers to the total number of tuples in the dataset. Eigenvalues and the corresponding eigenvectors for the covariance matrix obtained using (3).

$$Av = \lambda v \tag{3}$$

$$Det(A - \lambda I) = 0 \tag{4}$$

Solution to (4) gives the eigenvalues and eigenvectors can be found by substituting these Eigen values in the equation (4). The eigenvectors corresponding to the maximum eigenvalues are chosen to be the principal components of the given dataset under consideration.

Machine learning techniques for predicting the susceptibility score of COVID-19 infection

In the present work, machine learning techniques such as random forest algorithm and support vector regression are applied for predicting the susceptibility score of COVID-19 infection. Random forest algorithm (RF) is one of the most promising classifier which uses multiple decision trees (DT) to train and predict data samples. The general structure of random forest with multiple decision trees is shown in Fig.2.

The multiple ensemble DTs give rise to different classifications of infection susceptibility score. Here the value of n is chosen to be between 10 and 20 for optimum prediction. The majority scheme of vote is the terminal deciding factor of the model decision and throws the actual predicted class of the ISP of the individual as Low, medium, high.

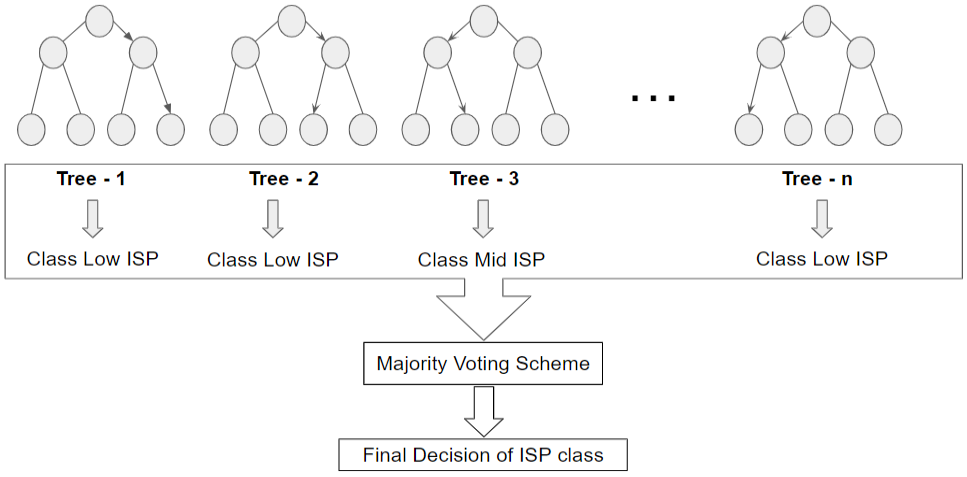


Fig 2 *Illustration of the random forest architecture*

In the random Forests classification approach, the ensemble of Decision Trees (DT) involved calculating the gini score as in equation (9).

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{5}$$

Here, the notations of the parameters are so following: ni_j refers to the significance of the node indexed j, W_j is indicating the weighted number of samples approaching the node indexed j, C_j indicates the impurity value of node indexed j, left(j) denotes the left child node from node indexed j and right(j) shows the right child node from node indexed j.

The second step is to obtain the importance given by each feature of the DT. This significance parameter can be computed using equation (10).

$$fi_i = \frac{\sum j{:}node\ j\ splits\ on\ feature\ i^{ni}j}{\sum_{k\in all\ nodes} ni_k} \quad (6)$$

where fi_i denotes the importance of feature indexed i and the ni_j refers to the importance of node indexed j.

$$norm\ fi_i = \frac{f\ i_i}{\sum_{j\in all\ features} fi_j} \quad (7)$$

These features fi_i are now normalized using the equation

$$RFfi_i = \frac{\sum_{j\in all\ trees} fi_{ij}}{T} \quad (8)$$

Then we can obtain the final feature of importance as mean of those of all the DTs (12), where RF f_i refers to the importance of feature indexed i computed through all DTs in the RF and norm fi_{ij} refers to the normalized feature significance parameter for index i in the DT indexed j and T indicates the total number of DTs.

Support Vector Regression for predicting susceptibility score of infection

Due to very high non linearity in the PCs, Support Vector Regression based approach is applied for obtaining susceptibility score of infection using the medical data set.

Assuming that the set of training medical data x_n is a multivariate set of N observations with observed response values y_n, a linear function is established as given below:

$$y = f(x) = < w, x > + b = \sum_{J=1}^{M} w_j\, x_j + b, y, b \in R, x, w \in R^M \quad (9)$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^r \begin{bmatrix} x \\ 1 \end{bmatrix} = w^r x + b\ x, w \in\ R^{M+1} \quad (10)$$

In the above equation x is a multidimensional input vector, with bias b and normal vector w. To ensure that it is as flat as possible, $f(x)$ with the minimal norm value, a convex optimization problem is formulated to minimize the following:

$$min_w \frac{1}{2} \|w\|^2 \quad (11)$$

This shows that the normal vector should be approximated during the process. Magnitude of weights is usually interpreted as flatness to the function obtained in the computation.

$$f(x, w) = \sum_{i=1}^{M} w_i\ \ x^i, x \in R, w\ \in\ R^M \quad (12)$$

To minimize the loss between the actual and predicted value which is a major constraint SVR adopts epsilon-insensitive loss function. Although asymmetrical loss functions should be used to avoid underestimation and overestimation, the functions used are usually convex in nature.

Since most of COVID data is asymmetrical, linear methods wouldn't provide accurate results. Nonlinear methods in SVR are handled by mapping the features to higher dimensional space called kernels.

$$min_w \frac{1}{2} \|w\|^2 + C\ \sum_{i=1}^{N} \varepsilon_i + \ \varepsilon_i^* \quad (13)$$

To achieve higher accuracy we replace all instances of x with K(xi,xj) from the earlier linear formula which leads to primal formulation shown in the above equation. The transformation of features to kernel space is shown in the above equation.

Performance of classification

In order to evaluate the classification performance of proposed machine learning algorithms, performance metrics such as accuracy, precision, sensitivity and specificity are calculated using the following formula.

$$\text{Accuracy} \quad \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

$$\text{Precision} \quad \frac{TP}{TP+FP} \tag{15}$$

$$\text{Sensitivity} \quad \frac{TP}{TP+FN} \tag{16}$$

$$\text{Specificity} \quad \frac{TN}{TN+FP} \tag{17}$$

Here TP= True Positive; TN= True Negative; FP= False Positive; FN= False Negative

Results

The proposed machine learning approaches are accomplished in Python computing and programming environment using the major computing libraries and mathematical functions used are in Numpy, Pandas, Scikit learn in Jupyter Notebook environment. Sample values of the data entities are described in Fig.3.

	1	2	3	4	5	6	7	8	9	10	11	12	13
	people_ID	Region	Gender	Designation	Name	Married	Children	Occupation	Mode_transp	cases/1M	Deaths/1M	comorbidity	Age	Coma score
1	1	Bhuban	Female	Mrs	mansi	YES	1	Farmer	Public	2	0	Hypertension	68	8
2	2	Bhuban	Female	Mrs	riya masi	YES	2	Farmer	Walk	2	0	Diabetes	64	15
3	3	Bhuban	Female	Mrs	sunita	NO	1	Cleaner	Public	2	0	None	19	13
4	4	Bhuban	Female	Mrs	anjali @ b	YES	1	Driver	Car	2	0	Coronary Hear	33	9
5	5	Bhuban	Female	Mrs	champa k	NO	2	Manufac	Car	2	0	Diabetes	23	7

	14	15	16	17	18	19	20	21	22	23	24	25	26
	Pulmonary score	cardiologica	Diuresis	Platelets	HBB	d-dimer	Heart rate	HDL choles	Charlson	Blood Gluco	Insurance	salary	FT/month
1	<400	Normal	441	154	93	233	82	58	27	7	3600000	1E+06	2
2	<100	Stage-02		121	56	328	89	68	5	6	1600000	400000	1
3	<300	Elevatec	416	124	137	213	77	43	40	6	3400000	900000	1
4	<200	Stage-0	410	98	167	275	64	60	27	7	700000	2E+06	1
5	<400	Normal	390	21	153	331	71	64	32	7	3200000	1E+06	1

Fig.3 *Sample health care data set (Srijan Singh 2020)*

The given medical health care data set is preprocessed, and principal component analysis is carried out for dimensionality reduction of data set. The hardware specifications of training, testing and inference include a processor of Intel Xeon CPU at 2.20GHz and a memory of 0.88GB. The inference timings using the best performing and most accurate RF model for 10000 users' details was 2.85704 seconds.

Heat Map of medical health care data

As the medical and health care data set contains many fields, a heat map is developed to quickly check correlations and visualizing the correlation matrix. Fig. 4 shows a heat map of correlation between the features of the dataset.

This heat map gives an understanding of the vital and detrimental correlated factors of medical data. It will be helpful to decide components of concern as against as the highly correlated ones. Unnecessary fields such as Name, Insurance, Salary, People_ID are found to be non-contributing to our analysis and prediction, thus they are removed during data preparation.

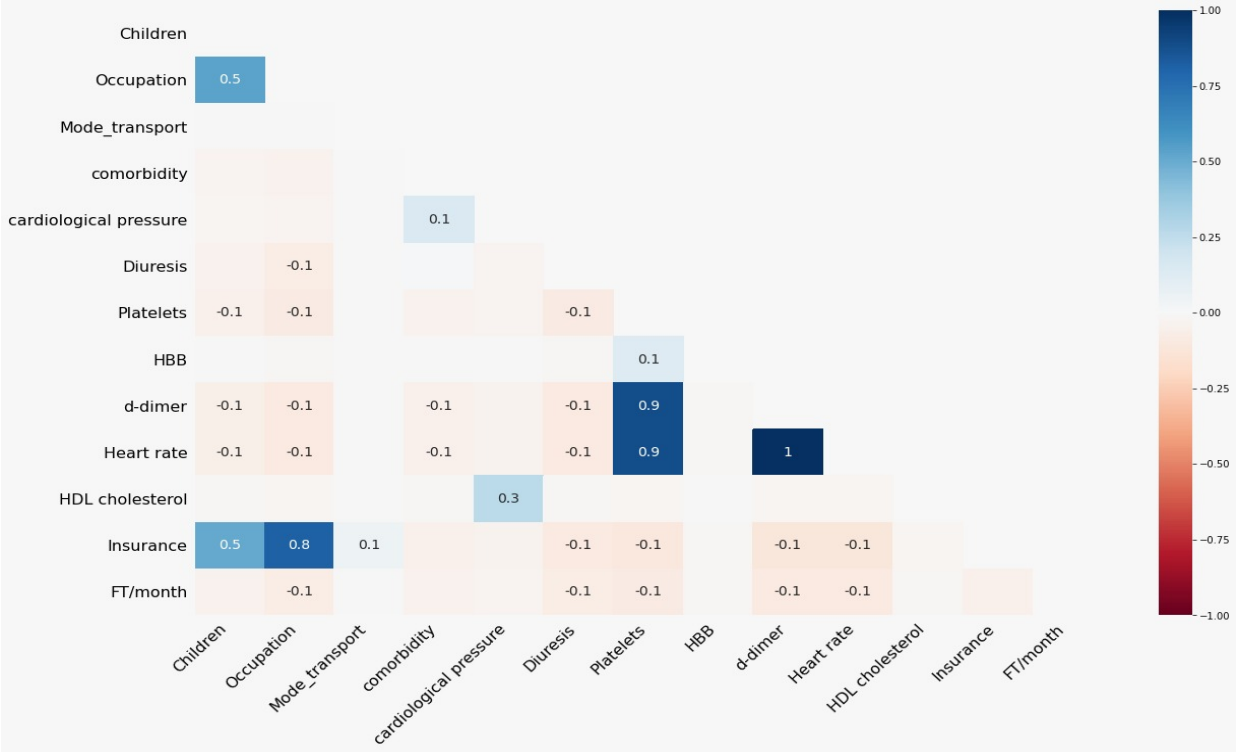


Fig.4 *Heat Map of medical and health care data.*

Estimated Principal Components of health care data

In the present work, PCA is applied on the health care data set to extract 6 principal components from 27 field values and the estimated explained variance values are given in Table.2.

Table 2 *Estimated principal component and the variance*

Principal Component	Explained Variance
PC1	0.55962616
PC2	0.18990593
PC3	0.1070075
PC4	0.09964971
PC5	0.01466063
PC6	0.01237666

Developed random forest tree structure for classification of COVID Inspection Susceptibility

Fig 5 shows the structure of random forest decision tree that classifies using the principal components of the dataset into 3 classes infection susceptibility of COVID-19 as Low, medium and high which are based on the estimated gini score. The color codes Green, Blue, Orange indicates the classes of infection susceptibility of COVID-19 as Low, medium and high respectively.

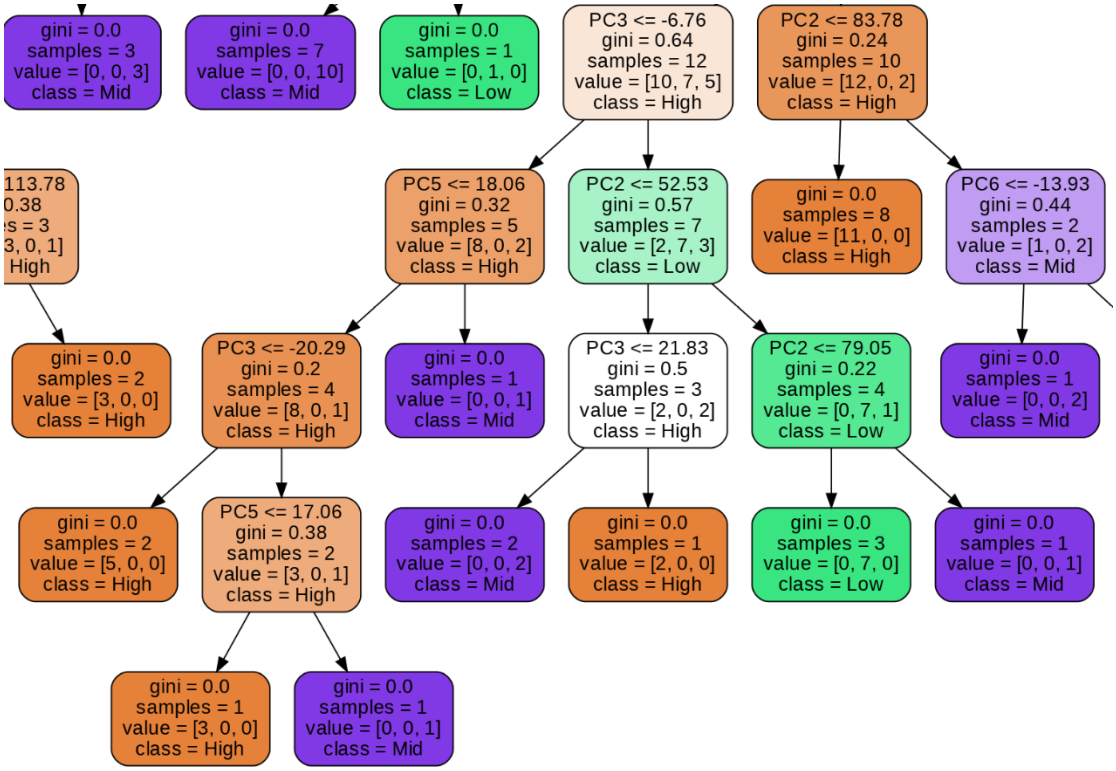


Fig 5 *A glimpse of an individual tree structure of the RF model.*

The trees developed contain 100- 200 nodes indicating the complex relationships mapped between the estimated principal components of dataset and the target classes. The computed gini score is also displayed at each level.

Discussions

Dimensionality reduction using Principal components of medical data set

Fig..6 shows that the most contributing principal component PC 1 which has an explained variance of over 50%. The next significant component explains less than half of the variance and the consequent component is more or less similar in the percentage of variance captured. The last two components show minute percentages of variance captured that are below 2%.

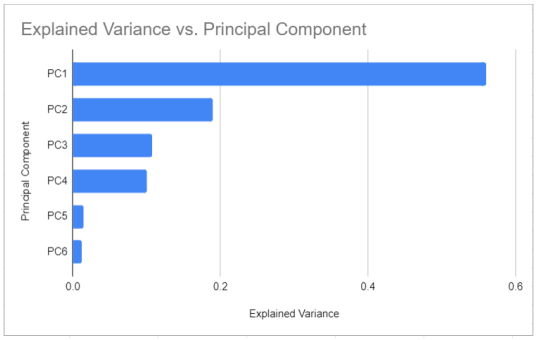


Fig.6 *PCA and the most contributing features towards the total variation in the dataset.*

From Table 2 and Fig 5, total variance captured by all PCs is found to be 0.983226 which indicates that the variance retention of over 98% of the variance in the data. These results validate the selection of features of the health care data set for the classification of

Classification performance of random forest

In order to validate the accuracy of the prediction of susceptibility score of infection, the proposed machine learning approach is applied to test data and compared with the actual results to those predicted by the algorithm. From the confusion matrix as shown in Fig.7, it is found that Sensitivity = 94%, Specificity = 81% and Precision = 44%. as shown in Fig. It is found that the random forest approach gave an overall classification accuracy of 90% for the validation data set of medical data.

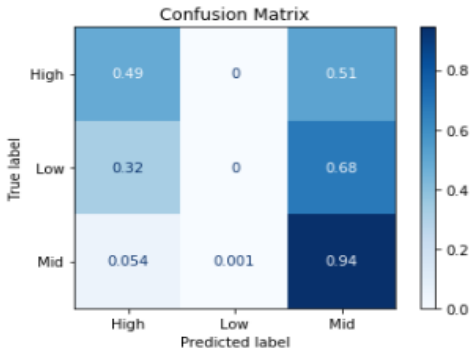
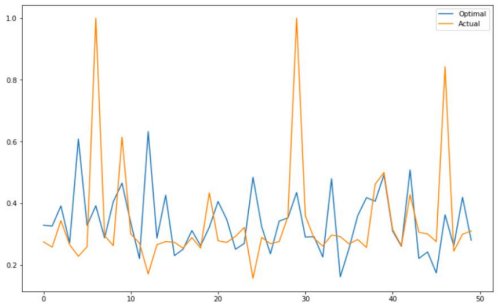


Fig.7 *Confusion Matrix for random forest*

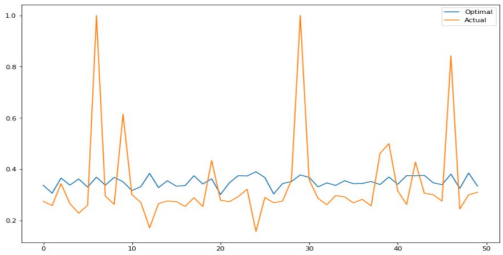
The high sensitivity shows that a high proportion of actual positives are classified correctly. The high specificity rate also shows that a good percentage of the 'safe' population is identified as not susceptible to the infection.

Comparison of Kernel functions for Support vector regression

In the present work, two kernel functions such as RBF and linear function is used for predicting the COVID Infection Susceptibility of individuals for the given health care data. Fig 8 shows the results obtained from the SVR model that was trained with a Radial Basis function (RBF) and linear kernel.



(a) RBF kernel



(b) linear kernel

Fig.8 *Comparison of Kernel functions used in Support vector regression*

It can be seen that there is a lesser deviation between the predicted values and actual values using the RBF kernel function in support vector regression.

.

Conclusions

This paper proposed Random forest and Support vector regression based prognostic model for predicting susceptibility of COVID-19 infection as low, medium and high using health care data of an individual. A medical data set available in the online repository is used for demonstrating the proposed approach. Heat map and principal component analysis (PCA) is applied to identify the dominant features of the medical data set. Using PCA, 6 principal components are extracted from 27 fields of medical data set and explained variance values are estimated for the given data set. The first principal component is found to provide the explained variance of 0.5596. Total variance captured by all PCs is found to be 0.983226 which highlights the effectiveness of the dimensionality reduction of medical data set. From the confusion matrix and the performance metrics of random forest approach, it is found that the random forest approach gave an overall classification accuracy of 97% for the validation data set of medical data which is found to be better than the support vector regression. It is found that RBF kernel function for support vector regression is superior in prediction of infection susceptibility of individual for COVID-19.

These results highlighted the application of machine learning approaches for interpreting health care data in understanding the infection severity of individual for COVID-19. From the larger public health care perspective, proposed approach will be helpful in identification of individuals who are highly susceptible for the COVID-19 infection in a containment zone which can give a decisive role to physicians and government officials for planning the more aggressive treatment and a better chance of survival. Also the early detection can also help hospitals prioritize intensive-care resources.

Acknowledgements

The authors thank the management of Vellore Institute of Technology, Vellore for providing the necessary facilities to carry out this research work.

Declarations

Conflicts of Interest The authors declare that they do not have any conflicts of interests.

Funding There was no funding done for this research work

Availability of Data and Material

The datasets presented in this study can be found online in open source repositories.

Code Availability For the reproducible code, please check out the GitHub repository at: https://github.com/srivatsanrr/autonom_covid

Following open source repositories are used for implementation of our work: Keras: https://keras.io; Sklearn: https://scikit-learn.org/stable/. statsmodels: https://www.statsmodels.org/stable/index.html.

References

Repici A, Maselli R, Colombo M, Gabbiadini R, Spadaccini M, Anderloni A, et al. Coronavirus (COVID-19) outbreak: what the department of endoscopy should know. Gastrointest Endosc 2020:1–6. doi:10.1016/j.gie.2020.03.019.

Mei, X., Lee, H., Diao, K. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat Med* (2020). https://doi.org/10.1038/s41591-020-0931-3

Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M. Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. BMJ 2020;369. doi:10.1136/bmj.m1328.

Pandey, Gaurav,. "SEIR and Regression Model based COVID-19 outbreak predictions in India." *arXiv preprint arXiv:2004.00958* (2020).

Liu, Dianbo,.A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019* (2020).
Carrillo-Larco Rodrigo M., Castillo-Cara Manuel. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Wellcome Open Research. 2020;5(56):56.
Yan, Li. Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan. *medRxiv* (2020).
Al-Karawi, Dhurgham. Machine Learning Analysis of Chest CT Scan Images as a Complementary Digital Test of Coronavirus (COVID-19) Patients *medRxiv* (2020).
Metsky, Hayden C. CRISPR-based COVID-19 surveillance using a genomically-comprehensive machine learning approach *bioRxiv* (2020).
Tang, Zhenyu Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images *arXiv preprint arXiv:2003.11988* (2020).

Narin, Ali, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks arXiv *preprint arXiv: 2003.10849* (2020).

Li L, Yang Z, Dang Z, Meng C, Huang J, Meng H, et al. Propagation analysis and prediction of the COVID-19. Infect Dis Model 2020;5:282–92. doi:10.1016/j.idm.2020.03.002.

Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y . An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2020;2:283–8. doi:10.1038/s42256-020-0180-7.

Hui DS, Azhar EI, Memish ZA, Zumla A. Human Coronavirus Infections—Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome (MERS), and SARS-CoV-2. vol. 2. 2nd ed. Elsevier Inc.; 2020. doi:10.1016/b978-0-12-801238-3.11634-4.

Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020;12:165–74. doi:10.21037/jtd.2020.02.64.

Srijan Singh, 2020- April, Flipr Hiring Challenge, 1, Retrieved May 2020 from https://www.kaggle.com/srijansingh53/flipr-hiring-challenge/version/1