Computer Methods and Programs in Biomedicine Update 2 (2022) 100070

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

Prolonged viral shedding prediction on non-hospitalized, uncomplicated SARS-CoV-2 patients using their transcriptome data

Pratheeba Jeyananthan [a,*]

[a] *Faculty of Engineering, University of Jaffna, Sri Lanka*

ARTICLE INFO

Keywords:
Prolonged viral shedding prediction
Day since onset prediction
Classification
Regression
Feature selection
Transcriptome data
COVID-19
Clinical data

ABSTRACT

Severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) is identified as a highly transmissible coronavirus which threatens the world with this deadly pandemic. WHO reported that it spreads through contact, droplet, airborne, formite, fecal-oral, bloodborne, mother-to-child and animal-to-human. Hence, viral shedding has a huge impact on this pandemic. This study uses transcriptome data of coronavirus disease 2019 (COVID-19) patients to predict the prolonged viral shedding of the corresponding patient. This prediction starts with the transcriptome features which gives the lowest root mean squared value of 16.3±3.3 using top 25 feature selected using forward feature selection algorithm and linear regression algorithm. Then to see the impact of few non-molecular features in this prediction, they were added to the model one by one along with the selected transcriptome features. However, this study shows that those features do not have any impact on prolonged viral shedding prediction. Further this study predicts the day since onset in the same way. Here also top 25 transcriptome features selected using forward feature selection algorithm gives a comparably good accuracy (accuracy value of 0.74±0.1). However, the best accuracy was obtained using the best 20 features from feature importance using SVM (0.78±0.1). Moreover, adding non-molecular features shows a great impact on mutual information selected features in this prediction.

Introduction

COVID-19 is currently alarming the world with its high transmissibility. Key transmission media of SARS-CoV-2 is infected respiratory droplets, same as other corona viruses [1]. There are many factors deciding the risk of this spread including proximity and ventilation [2]. Previous studies reported many facts related to the infection, spread and viral shedding of COVID-19. Even though prolonged viral shedding in infected people were identified, the presence of viral RNA on test does not necessarily correlate with infectivity [3]. They further reported that the relationship between quarantine after clinical recovery and transmission is uncertain. Moreover, transmission can be asymptomatic and presymptomatic, and infectivity may be highest after onset of symptoms.

As this is a communicable disease, viral shedding is very crucial. Literature shows that viral shedding and transmission of virus is very important in terms of COVID-19. Even though, it is very important to identify the viral shedding probability of an infected patient, there are very few studies in the literature related to this and it is very crucial to identify the important factors in connection with viral shedding.

In the literature, clinical data of the patients were used for viral shedding prediction [4,5]. In these studies, they either analyzed the data to predict the viral shedding [4] or cox regression is used for the same study [5], not any other prediction models. In both of these studies they identified few features such as age, sex, comorbidities and frequency of cough have high control over this viral shedding. In the literature, number of studies in this area is very few and no studies used any molecular data in this prediction. However, recent biological studies show that viral shedding is a key factor for gene expression differences and identified genes could be important for understanding the molecular mechanism variation in pathogen shedding [6]. Their finding is on low-pathogenic avian influenza (LPAIV)-infected wild-bred mallards.

Hence, this study initially utilizes transcriptome data of uncomplicated SARS-CoV-2, another infectious disease in the prediction of viral shedding of the infected patients. Feature importance, mutual information and forward feature selection are used to select the related features of the prediction. Using the selected features in the viral shedding prediction shows that 25 features selected using forward feature selection algorithm give better accuracy (lowest root mean squared error

* Corresponding author at: Computer Engineering, University of Jaffna, Faculty of Engineering, Ariviyal Nagar Kilinochchi, Kilinochchi, North, Sri Lanka.
E-mail address: pratheeba@eng.jfn.ac.lk.

https://doi.org/10.1016/j.cmpbup.2022.100070
Received 8 February 2022; Received in revised form 24 May 2022; Accepted 4 September 2022
Available online 6 September 2022

(RMSE) value) compared to other set of features. Then, with these selected features, few clinical data of those patients are added, where there are no significant improvement in the accuracy.

Further, this study predicts the day since onset (either 0 or 5 are available) of these patients. It shows that transcriptome data can classify the patients according to their day since onset with the accuracy of 0.78. For all of these studies, three different sets of features are selected using feature importance, mutual information and forward feature selection algorithms. In the regression task linear, least absolute shrinkage and selection operator (LASSO), random forest, ridge and decision tree algorithms are used with RMSE as the accuracy measure. Support Vector Machine (SVM), random forest, naïve Bayes, decision tree and K-nearest neighbors (KNN) classifiers are used in the classification task where accuracy is used for measure the accuracy of the model. In the prediction of day since onset, mutual information selected features showed a drastic accuracy improvement while adding the non-molecular features.

Materials and Methods

Material

This study uses a publicly available data from gene expression omnibus (GEO) with the accession number of GSE178967. This is a high throughput sequencing data of the blood samples collected from108 SARS-CoV-2 positive cases at 0 days or 5 days. Majority of the patients are uncomplicated and do not need hospitalization. However these infections contribute to ongoing viral transmission, which was measured by defining early immune baseline and infection-induced signatures that predict the duration of viral shedding. The detailed experiment setup could be derived from the GEO database.

Feature selection methods

Two different studies are here with this transcriptome data, prediction of viral shedding of the COVID-19 patients and day of the patients from the onset of the infection. For both of these predictions, features are selected using mutual information and feature importance.

Feature importance

This technique calculates a score for all the input features regarding a given model called as feature importance. Feature with a higher score has a larger impact on the model, hence will be selected to the model building in the prediction of the specified target. This helps to understand the relationship between the features and the target value. Also it is useful in the identification of irrelevant features of the problem. Eventually, selecting the appropriate features of the model will help in the improvement of the prediction accuracy.

Mutual information

Mutual information measures the information carried by one random variable regarding another. It is a measure of the mutual dependence between two variable. Hence, using mutual information, one can quantify the amount of information they can obtain regarding one variable by observing another random variable.

Forward feature selection algorithm

This algorithm starts with the whole set of features in the data. In the first step, it will select the best feature (X_1) among all of those features with the capacity of predicting the target with highest accuracy. In the second step, another feature (X_2) will be selected from the rest of the features that can give the best accuracy with the already selected feature (X_1). This process will continue till the selected subset of features reach a consistent accuracy or till the specified amount of features are selected.

Machine learning algorithms

Two different works are done here including classification and regression. Five different regression and classifications algorithms are used in this study along with two feature selection algorithms.

Regression algorithms

Regression models are used in viral shedding prediction. Linear, LASSO, random forest, ridge and decision tree regressions are used here.

Linear regression

Linear regression is one of the well-known supervised machine learning algorithms shows the linear relationship between a dependent and one or more independent variables. As the target feature viral shedding is a continuous value, this is the first choice of regression algorithm. This algorithm is used under few assumptions such as there is a linear relationship between the input features and the target values, there is a small correlation (not high) between features and data are clean without noise.

Least absolute shrinkage and selection operator (LASSO) regression

This is a regression algorithm which do both variable selection and regularization during the model building to enhance the accuracy and interpretability of the model. This method introduces a bias to reduce the variance in the results and finally to a lower mean square error [7].

Random forest regression

Random forest regression is a simple model with comparably higher performance. This is an ensemble algorithm, where multiple decision trees are used in the model building. Accuracies from all the decision trees are considered in the final accuracy calculation [8].

Ridge regression

This is another regularization method for regression, as it tries to minimize the sum of squared residuals. This minimization is allowed under some penalty. Here, coefficients of the features are forced to be close to zero, not completely to zero [9]. This is the main difference between LASSO and ridge regression, where LASSO allows the zero coefficients of features.

Decision tree regression

Decision trees are working by questioning a feature in each step, generally a binary question. Initial step of this algorithm is finding out the root node (top attribute) of the tree. Generally, error value is used in this top attribute selection of regression algorithm. Based on this feature data is divided into two and algorithm will find the next attribute to split the data. This algorithm will continue either until reach the leaf node or meet the stopping criteria [10].

Classification algorithms

Five different classification algorithms are used in the classification of patients into their day since onset (either 0 or 5 are given in the data). Support vector machine, random forest, naïve Bayes, decision tree and KNN classifiers are used for this purpose.

Support vector machine (SVM)

SVM is a supervised learning algorithm, mainly used in the classification problems. This algorithm finds a hyperplane to separate the given

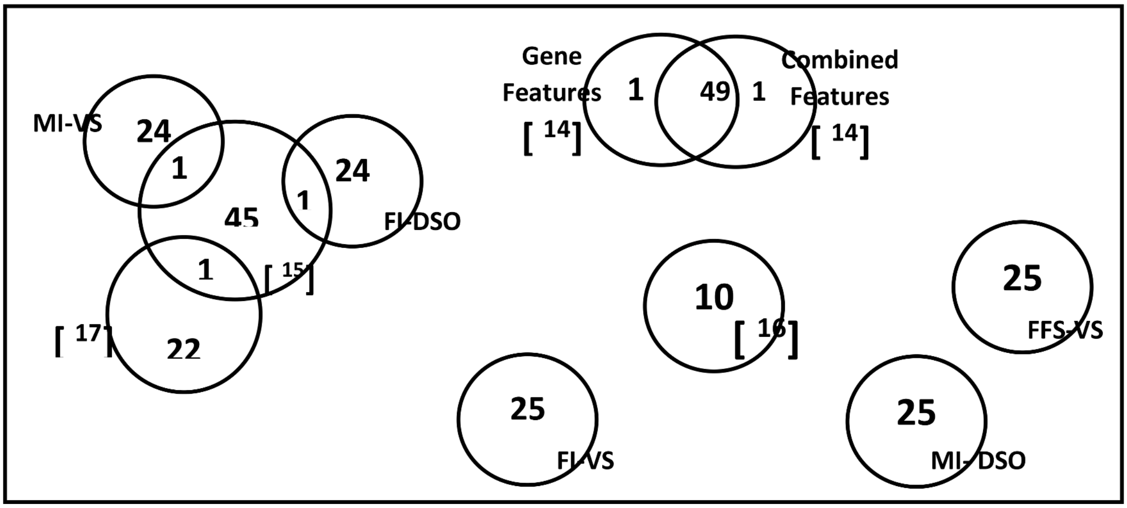

Fig. 1. Venn diagram comparing the selected features with the features from previous studies. This study has six different sets of features (2 studies, each with 3 feature selection methods) and those set are compared with the features selected from other studies in the literature. Here, references are given within the brackets. MI- Mutual Information, FI- Feature Importance, FFS- Forward feature selection, VS- Viral shedding and DSO- Day since onset.

classes, where the distance between this hyperplane and closest points (from each class) is maximized [11].

Random forest classifier

Same technique is used as described in regression, here in the classification.

Naïve Bayes classifier

This is a classification algorithm uses the concept of Bayes theorem during model building. This is a simple but comparably powerful classification algorithm which is fast with quick predictions [12]. It works under the assumption that the features are independent. Major advantage of this algorithm is, it can be trained with comparably less data.

Decision tree classifier

This is the classification version of decision trees. Same technique is used here also as described in regression.

K-nearest neighbor (KNN) classifier

This is one of the basic classification algorithm where the new point is assigned to a class with the highest number of related points, first introduced in [13]. Different distance measures including Mahalanobis distance can be used to calculate those related (neighbor) points. Number of neighbor points (K) should be defined by the user.

Cross validation

Cross validation is used to validate the performance of our model. 10-fold cross validation is used in this study. In 10-fold cross validation, whole data is split into 10 equal portions. We have 10-iterations in this method and in each iteration we will use 9 portions in the training and one portion for testing. This testing partition will change with iteration.

Accuracy measures

Measuring the performance of our model on some unseen test data is crucial in machine learning models. Three different accuracy measures are used in this study.

Root mean square error (RMSE)

Generally RMSE is used to calculate the error in a regression model. It measures how the regression fits with the original data points. It measures the root mean square value between the actual data point and the predicted data point.

Accuracy

This is one of the accuracy measures used in the classification models. This is the ratio of the correct predictions among all the predictions.

Results

Study on selected features

Twenty five features are selected for two different studies using three different feature selection methods (Supplementary Table 1 to Supplementary Table 6). To biologically validate those features, first they are compared with already identified biomarkers in few other respiratory studies [14–17]. This comparison shows that (Fig. 1) there are few common genes between these studies. This is same even among them without considering this study. This might be because the scope or target of these studies are distinct.

Then, all these selected features are tested under Gene Ontology (GO) analysis. Genes selected by feature importance for viral shedding prediction (Supplementary Table 1) shows that they are related to cytosolic transport, lumenal side of endoplasmic reticulum membrane and integral component of lumenal side of endoplasmic reticulum membrane. Same study on mutual information selected genes (Supplementary Table 2) shows that they are trans-Golgi network membrane, MHC protein complex, lumenal side of endoplasmic reticulum membrane, integral component of lumenal side of endoplasmic reticulum membrane and peptide antigen binding. These two set of features did not give any GO terms closely related to immunology or any diseases. However, forward feature selection algorithm gave a set of features those are related to citrate synthase activity, citrate (Si)-synthase activity, pyridoxal kinase activity, pyridoxal 5′-phosphate salvage and N-acetyl-beta-glucosaminyl-glycoprotein 4-beta-N-acetyl galactosaminyl transferase activity, which have already shown their relationship with some diseases. Citrate synthase activity is closely related to cancer [18, 19] and neurological diseases [20,21] and used in various infectious

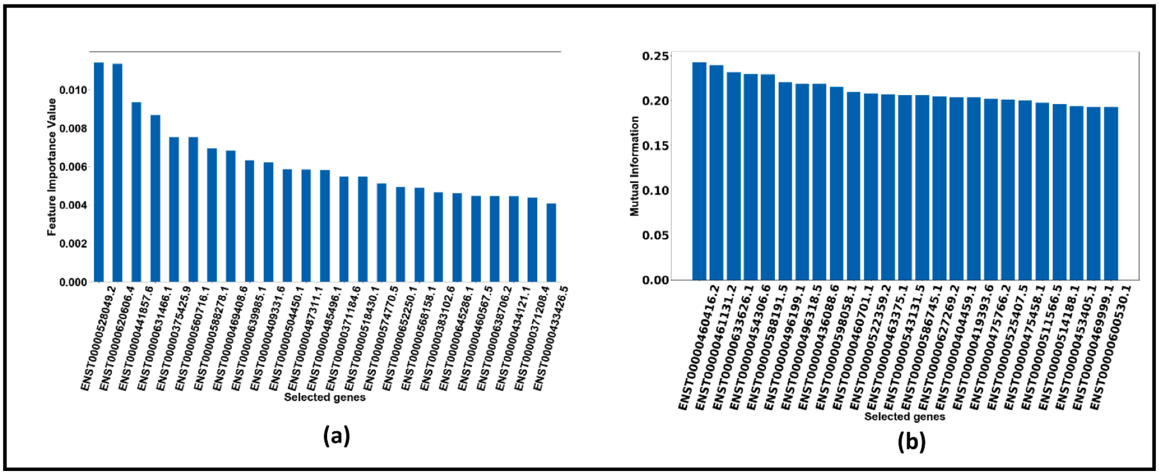

Fig. 2. Top 25 features selected in the prediction of viral shedding. (a) Features selected using feature importance (b) Mutual information selected features.

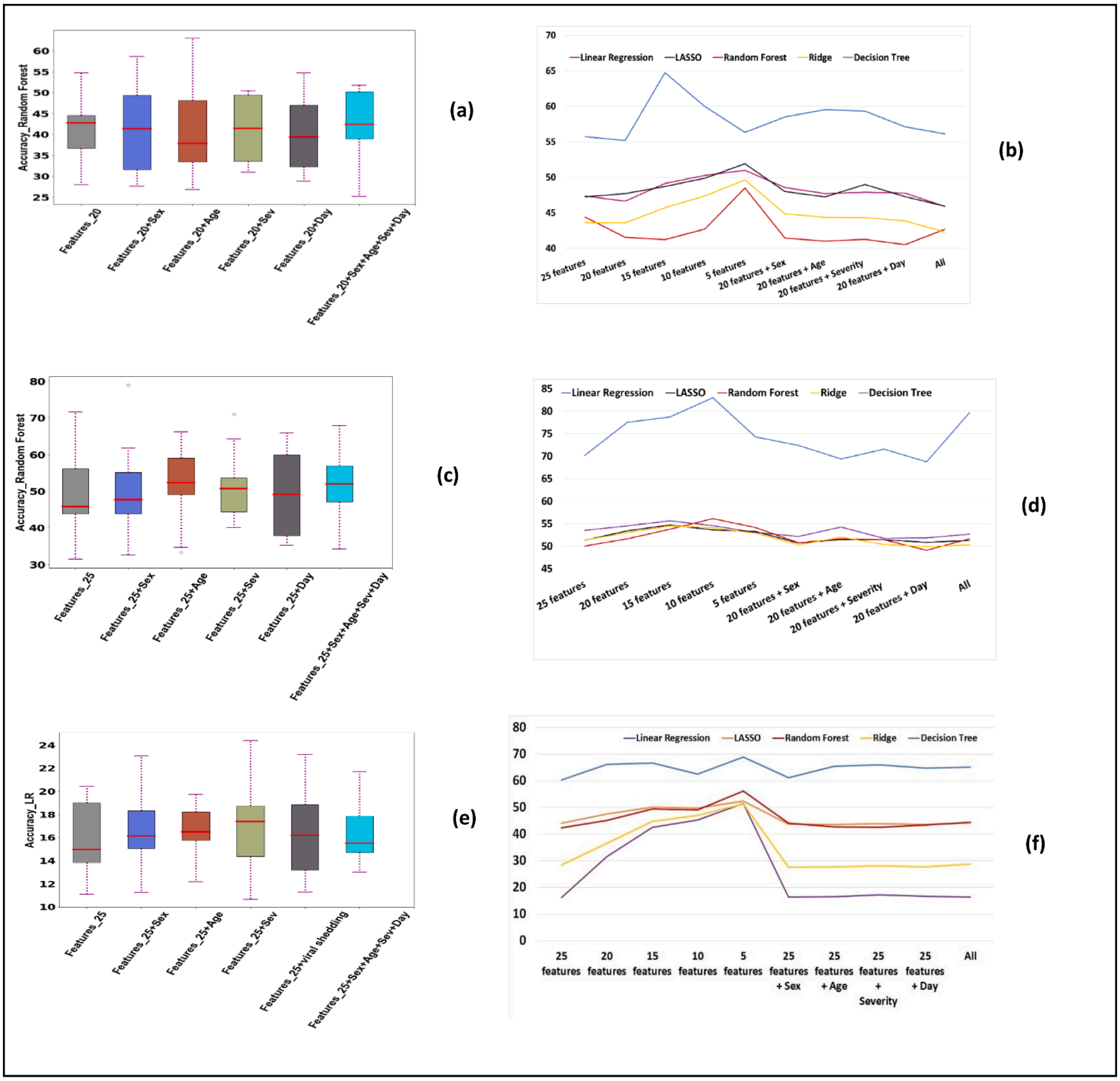

Fig. 3. Summary of accuracies by selected features in the viral shedding prediction. (a) Best performance of features (20 transcriptome selected by feature importance) using random forest regression. (b) Performance summary of all the regression algorithms using different set of features (feature importance). (c) Best performance of mutual information features (25 selected using mutual information) (d) Performance summary of mutual information selected features in prolonged viral shedding prediction (e) Twenty five features selected using forward feature selection algorithm gives best performance using linear regression (f) Different number of features are selected using forward feature selection algorithm and tested with different machine learning algorithms in the prolonged viral prediction.

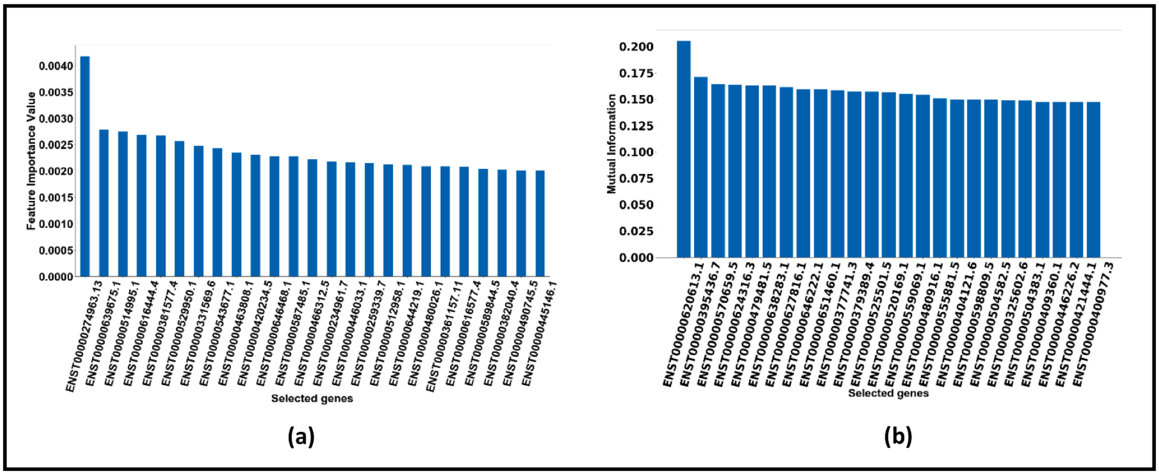

Fig. 4. Twenty five features are selected from a set of transcriptome features for the day since onset prediction using (a) Feature Importance and (b) Mutual Information.

diseases and inflammation [22]. Same as this, pyridoxal kinase was assessed as a drug target for African trypanosome 16 and showed its connection with cancer [23]. Importance of pyridoxal 5′-phosphate in many diseases including some rare diseases also studied [24,25].

Rather, GO analysis on the features selected for day since onset prediction shows some prominent functions related to immunology. Feature important selected features are closely related to the processes such as response to virus, negative regulation of viral genome replication, cell projection membrane, positive regulation of RIG-I signaling pathway, defense response to virus and regulation of viral process. Mutual information selected features are related to negative regulation of hepatocyte growth factor receptor signaling pathway, regulation of cellular response to hepatocyte growth factor stimulus, negative regulation of cellular response to hepatocyte growth factor stimulus, regulation of cellular ketone metabolic process by negative regulation of transcription from RNA polymerase II promoter and galactosylgalactosylglucosylceramide beta-D-acetylgalactosaminyltransferase activity. Finally forward feature selection algorithm selected features are negative regulation of T cell antigen processing and presentation, regulation of antigen processing and presentation of endogenous peptide antigen via MHC class I, negative regulation of antigen processing and presentation of endogenous peptide antigen via MHC class I, cellular response to iron ion starvation and EH domain binding (Maximum of five GO terms are provided here).

Viral shedding prediction

Initially this prediction starts with the selection of top 25 features using feature importance, mutual information and forward feature selection selected algorithms (Fig. 2, complete list of features are presented in the supplementary file). Those features are used as the input for five different regression algorithms, linear, LASSO, random forest, ridge and decision tree. In order to find the number of features with high accuracy, as the next step, different number of features (20, 15, 10 and 5) are used as the input to the same models. After the selection of the best set of features using RMSE, other non-molecular features (age, severity and day) are added to them one by one as the input to check their impact on this prediction.

While using top 25 features selected using forward feature selection algorithm along with linear regression algorithm, viral shedding prediction gave the best accuracy compared to others (Supplementary Table 5, 6 and 7). As the next step, non-molecular features such as sex, age, severity and number of days are individually added to those 25 features and used in the same prediction. Finally, all these features (25 selected features, sex, age, severity and number of days) are fed into these models altogether as the input of this prediction.

Fig. 3 shows that linear regression gives the best performance among all the algorithms and on the whole forward feature selection selected features perform better than other features. However, there are no significant improvements in the prediction by adding the non-molecular data Fig. 3-(a).

Day since onset prediction

This study also started with 25 top transcriptome features selected using feature importance, mutual information and forward feature selection algorithm (Fig. 4). Those 25 features are used with SVM, random forest, naïve Bayes, decision tree and KNN classifiers, followed by the utilization of 20, 15, 10 and 5 top features as the input of the model. Comparing the accuracies between these classifiers and number of feature importance selected features shows that 20 features along with SVM gives the best prediction accuracy (0.78±0.1). In this classification task, SVM outperforms other classifiers with the highest accuracy (Fig. 5). However, the second accuracy, which is almost equal to the best accuracy is with top 25 forward feature selection selected features using random forest classifier (0.74±0.1).

In the feature importance selected features, as 20 features give the best prediction accuracy, the non- molecular data such as sex, age, severity and viral shedding are individually and collectively combined with them and used in the same prediction using five different classification algorithms. Fig. 5 shows that there are not any significant changes in the performance of these 20 features by this addition. However, we can notice minor changes in the variation, especially while adding age and with the whole features.

However, while adding these non-molecular features to the mutual information selected features, there is a drastic increment in the performance of the model from 0.59±0.03 to 0.76±0.05 (Supplementary Table 8). Adding sex of the patient to these top 25 features give this highest accuracy value along with random forest classifier, which is almost competent with the highest accuracy.

Discussion

Transcriptome data of COVID-19 patients is used in the prediction of prolonged viral shedding and day since onset prediction of the patients. Feature importance and mutual information are used in the selection of related features. This process starts with top 25 features and then decreased to 20, 15, 10 and 05 features. Number of features with highest accuracy is selected in each category and combined with few important non-molecular features to study their impact on the accuracy of those prediction.

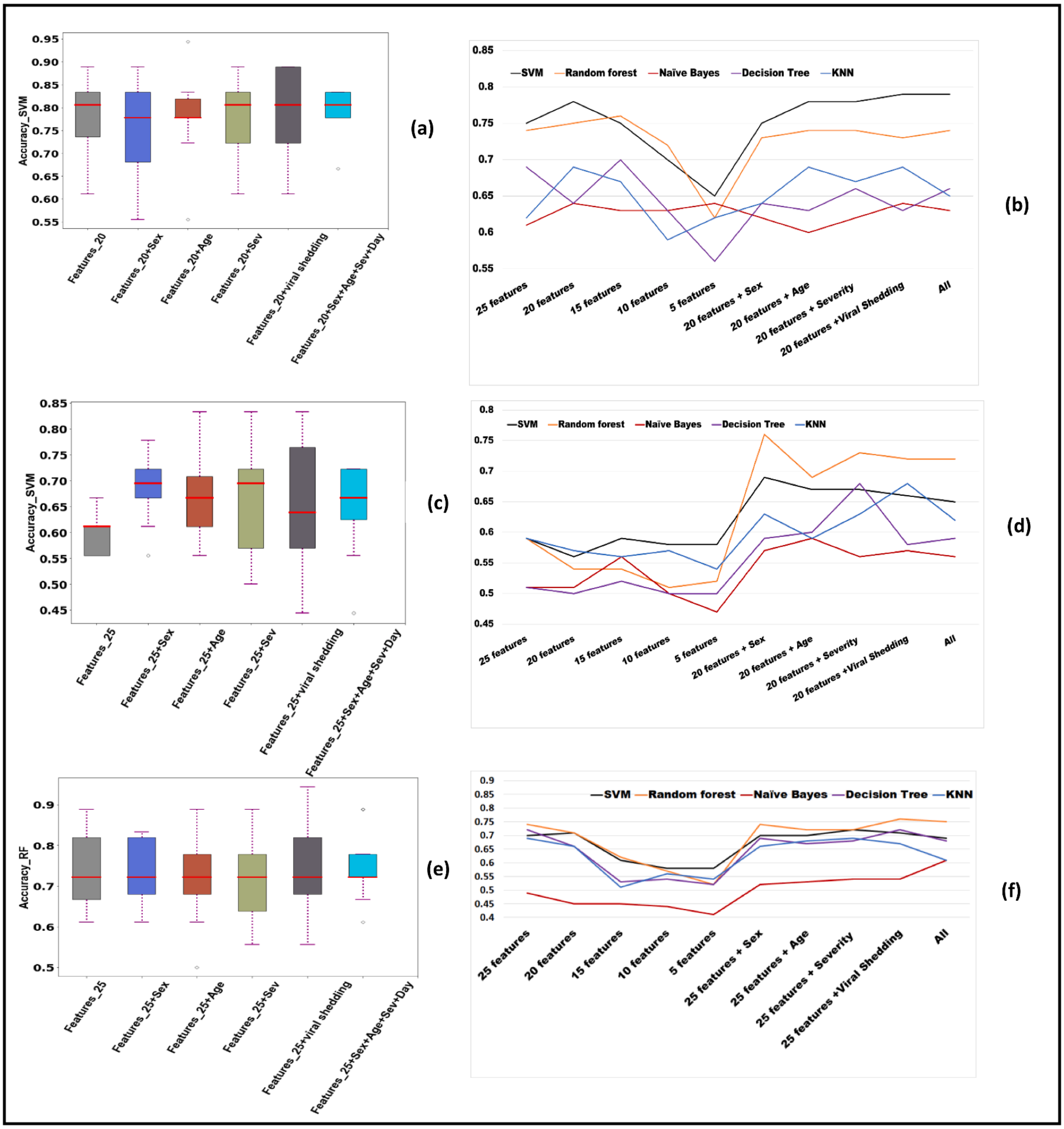

Fig. 5. Different number of transcriptome features are selected using feature importance (a and b), mutual information (c and d) and forward feature selection (e and f) and used with different machine learning algorithms to predict the day since onset of COVID-19 patients. 25, 20, 15, 10 and 5 top transcriptome features are initially selected to this prediction and the one with highest prediction accuarcy is selected. Different non-molecular features also added to that particular feature set to see their impact on this prediction. (a) Top 20 transcriptome features using SVM gives best accuracy in the feature importance selected features. Hence all the other features are added to those 20 features and the accuracies are compared after 10-fold cross validation. (b) Complete set of accuracies in the feature importance selected features (c) Top 25 features with highest accuracy in mutual information selected features using SVM and the accuracies after adding non-molecular features (d) Complete set of accuracies with mutual information selected features (e) Top 25 features selected using forward feature selection using random forest algorithm (f) Performance comparison between different classification algorithms with different number of features.

In the prediction of viral shedding, this study shows that forward feature selection aalgorithm selected features gives comparably better performance than other two sets. Also in this prediction linear regression performs better than other algorithms. Further it shows that few non-molecular data used in this data has not any significant improvement in the viral shedding prediction. Age, sex and days since onset are such features and there are previous studies regarding the viral shedding of SARS-CoV-2 patients and their association with these features [26 27]. Also it has been already stated that there is no correlation between severity and prolonged viral shedding [27].

On the contrary, in day since onset prediction, the relationship between this target and non-molecular data is noticed in the minimal level for both forward feature selection and feature importance selected features. However, in the mutual importance selected features, there is a notable influence by these non-molecular data. Adding the sex details to the model along with top 25 mutual information selected features give the accuracy value of 0.76±0.05. This is an increment from the accuracy value of 0.59±0.03 by those 25 features, which is very low compared to the accuracy value of top 20 feature importance selected features and top 25 forward feature selection selected features. However, adding sex details to mutual information selected features improve its accuracy almost equal to the accuracy of other two feature sets. Even adding age and severity also had a big impact on the prediction of day since onset. All these changes are observed only on mutual information selected features.

In the literature, the connection between gender and COVID-19 was widely studied. A close connection was reported between gender and severity [28,29]. Further, hospitalization and gender 31 and age in the prediction of disease by day-11 [31] also studied in the literature. Here, this study shows a connection between gender and day since onset. Moreover, age also has a great impact on COVID-19 [30,32,33]. This study also shows the association between age and the day from the start.

These two prediction are useful both for the clinicians and the patients. The viral shedding prediction will be useful to control the patients from further spread of the virus. Almost same like this, day since onset prediction will be helpful for the clinicians to take important treatment decisions for every patients.

Conclusion

Transcriptome data of COVID-19 patients is used in this study to predict the prolonged viral shedding and day since onset of the corresponding patient. Feature importance, mutual information and forward feature selection selected features are used in the feature selection. Best set of selected features are incorporated with some other non-molecular data to see their impact on these predictions. In the viral shedding prediction, forward feature selection selected features give the best accuracy, while feature importance selected features performs with the highest prediction accuracy in the day since onset prediction. Also this study shows the importance of sex, age, severity and day/viral shedding in the prediction of viral shedding/day since onset. In the day since onset prediction, SVM give the best accuracy value of 0.79±0.05 using 20 transcriptome features combined with all the non-molecular features. In the viral shedding prediction, linear regression gives the lowest RMSE value (16.3±3.26) using top 25 transcriptome from forward feature selection algorithm.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpbup.2022.100070.

References

[1] C Muge, K Krutika, K Jason, P Malik, Virology, transmission, and pathogenesis of SARS-CoV-2, BMJ (2020) 371.

[2] E Meyerowitz, A Richterman, R Gandhi, P Sax, Transmission of SARS-CoV-2: a review of viral, host, and environmental factors, Ann. Intern. Med. 174 (1) (2021) 69–79.

[3] A Widders, A Broom, J Broom, SARS-CoV-2: the viral shedding vs infectivity dilemma, Infect. Dis. Health 25 (3) (2020) 210–215.

[4] HC Lu, F Ling, L WeiXia, et al., A novel prediction model for long-term SARS-CoV-2 rna shedding in non-severe adult hospitalized patients with COVID-19: a retrospective cohort study, Infect. Dis. Ther. 10 (2) (2021).

[5] L Hui, Z Jing, Z Hao-Long, et al., Prolonged viral shedding of SARS-CoV-2 and related factors in symptomatic COVID-19 patients: a prospective study, BMC Infect. Dis. 21 (1) (2021).

[6] A Dolinski, J Homola, M Jankowski, J Robinson, J Owen, Differential gene expression reveals host factors for viral shedding variation in mallards (Anas platyrhynchos) infected with low-pathogenic avian influenza virus, J. Gen. Virol. 103 (3) (March 2022).

[7] R Tibshirani, Regression shrinkage and selection via the lasso, J. Royal Statist. Soc. Series B (Methodological) 58 (1) (1996).

[8] HT Kam, Random decision forests, Proceedings of the Third International Conference on Document Analysis and Recognition 1 (1995) 278–282.

[9] EH Donald, WS Donald, Ridge, a computer program for calculating ridge regression estimates, Upper Darby, Pa, Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station 236 (1977) 10.

[10] J Quinlan, Simplifying decision trees, Int. J. Man Mach. Stud. 27 (3) (1987) 221–234.

[11] C Cortes, V Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995).

[12] F Nir, G Dan, G Moises, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.

[13] F Evelyn, LH Joseph, Discriminatory analysis. nonparametric discrimination: consistency properties, USAF School Aviat. Med. (1951).

[14] Y Zhang, Z Li, T Zeng, et al., Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles, Front. Genet. 13 (2021). January.

[15] Y Zhang, H Li, T Zeng, et al., Identifying transcriptomic signatures and rules for SARS-CoV-2 infection, Front. Cell Dev. Biol. (2021) 11. Jan.

[16] Y Zhang, T Zeng, L Chen, S Ding, T Huang, Y Cai, Identification of COVID-19 infection-related human genes based on a random walk model in a virus-human protein interaction network, Biomed. Res. Int. 8 (2020). July.

[17] P Ranjan, N Singh, A Kumarea, NLRC5 interacts with rig-i to induce a robust antiviral response against influenza virus infection, Eur. J. Immunol. 45 (3) (2015). March.

[18] L Chen, T Liu, J Zhou, Y Wang, X Wang, W Di ea, Citrate synthase expression affects tumor phenotype and drug resistance in human ovarian carcinoma, PLoS One 9 (12) (2014).

[19] Z Cai, Y Deng, J Ye, et al., Aberrant expression of citrate synthase is linked to disease progression and clinical outcome in prostate cancer, Canc. Manage. Res. (2020) 12.

[20] M Ullian, B Gantt, A Ford, B Tholanikunnel, E Spicer, W Fitzgibbon, Potential importance of glomerular citrate synthase activity in remnant nephropathy, Kidney Int. 63 (1) (2003). Jan.

[21] XX Cui, X Li, SY Dong, YJ Guo, T Liu, YC Wu, SIRT3 deacetylated and increased citrate synthase activity in PD model, Biochem. Biophys. Res. Commun. 484 (4) (2017) 767–773.

[22] X Tingting, C Yue, Research progress of [68Ga]Citrate PET’s utility in infection and inflammation imaging: a review, Mol. Imaging Biol. (2019) 22.

[23] A Joseph, J Pan, J Michels, G Kroemer, M Castedo, Pyridoxal kinase and poly(ADP-ribose) affect the immune microenvironment of locally advanced cancers, Oncoimmunology 10 (1) (2021).

[24] C Barbara, M Riccardo, O Elisa, A Alessandra, V Carla, The chaperone role of the pyridoxal 5 ’-phosphate and its implications for rare diseases involving B6-dependent enzymes, Clin. Biochem. (2013) 47.

[25] Salvo Md, M Safo, R Contestabile, Biomedical aspects of pyridoxal 5′-phosphate availability, Front. Biosci. (Elite Ed) 4 (3) (Jan 2012) 897–913.

[26] L Hui, Z Jing, Z Hao-Long, et al., Prolonged viral shedding of SARS-CoV-2 and related factors in symptomatic COVID-19 patients: a prospective study, BMC Infect. Dis. 21 (1) (2021) 1282.

[27] A Widders, A Broom, J Broom, SARS-CoV-2: the viral shedding vs infectivity dilemma, Infect. Dis. Health 25 (3) (2020) 210–215.

[28] R Federico, N Luca, G Arianna, et al., Covid-19 and gender: lower rate but same mortality of severe disease in women—an observational study, BMC Pulmonary Med. 21 (1) (2021).

[29] G Catherine, RZ Vera, K NH, M Rosemary, L KS, Impact of sex and gender on COVID-19 outcomes in Europe, Biol. Sex Differ. 11 (1) (2020).

[30] Q Virginia, S Cristina, S Alessandra, et al., Sex differences in a cohort of COVID-19 Italian patients hospitalized during the first and second pandemic waves, Biol. Sex Differ. 12 (1) (2021).

[31] Elisa G, Alessia S, Monica C, et al. Assessment of COVID-19 progression on day 5 from symptoms onset. *BMC Infect. Dis.*;21(1).

[32] C Cheng, Z DongDong, D Dejian, et al., The incubation period of COVID-19: a global meta-analysis of 53 studies and a Chinese observation study of 11 545 patients, Infect. Dis. Poverty 10 (1) (2021).

[33] A Ghazal, L Katya, Y Lanbo, et al., Predictors of COVID-19 severity: a literature review, Rev. Med. Virol. 31 (1) (2021).