ONLINE FIRST

This is a provisional PDF only. Copyedited and fully formatted version will be made available soon.

Advances in Respiratory Medicine

Formerly Pneumonologia i Alergologia Polska
Edited since 1926

ISSN: 2451-4934

e-ISSN: 2543-6031

Machine learning-based COVID-19 diagnosis by demographic characteristics and clinical data

Authors: Fatemeh Gorji, Sajad Shafiekhani, Peyman Namdar, Sina Abdollahzade, Sima Rafiei

DOI: 10.5603/ARM.a2022.0021

Article type: Research paper

Submitted: 2021-09-28

Accepted: 2021-12-05

Published online: 2022-01-31

Machine learning-based COVID-19 diagnosis by demographic characteristics and clinical data

Fatemeh Gorji et al., In-silico diagnosis of COVID-19

Fatemeh Gorji[1], Sajad Shafiekhani[2,3,4], Peyman Namdar[5], Sina Abdollahzade[6], Sima Rafiei[7]

[1]Students' Scientific Research Center, Qazvin University of Medical Sciences, Qazvin, Iran

[2]Departments of Biomedical Engineering, School of Medicine, Tehran University of Medical Sciences

[3]Research Center for Biomedical Technologies and Robotics, Tehran, Iran

[4]Students' Scientific Research Center, Tehran University of Medical Sciences, Tehran, Iran

[5]Department of Surgery, Qazvin University of Medical Sciences, Qazvin, Iran

[6]School of Medicine, Qazvin University of Medical Sciences, Qazvin, Iran

[7]Department of Healthcare Management, School of Health, Qazvin University of Medical Sciences, Qazvin, Iran

Address for correspondence: Sima Rafiei, Department of Healthcare Management, School of Health, Qazvin University of Medical Sciences, Qazvin, Iran, tel: +989123886817, e-mail: sima.rafie@gmail.com,

Abstract

Introduction: To facilitate rapid and effective diagnosis of COVID-19, effective screening can alleviate the challenges facing healthcare systems. We aimed to develop a machine learning-based prediction of COVID-19 diagnosis and design a graphical user interface (GUI) to diagnose COVID-19 cases by recording their symptoms and demographic features.

Methods: We implemented different classification models including support vector machine (SVM), Decision tree (DT), Naïve Bayes (NB) and *K*-nearest neighbor (*K*NN) to predict the result of COVID-19 test for individuals. We trained these models by data of 16973 individuals (90% of all individuals included in data gathering) and tested by 1885 individuals (10% of all individuals). Maximum relevance minimum redundancy (MRMR) algorithms used to score

features for prediction of result of COVID-19 test. A user-friendly GUI was designed to predict COVID-19 test results in individuals.

Results: Study results revealed that coughing had the highest positive correlation with the positive results of COVID-19 test followed by the duration of having COVID-19 signs and symptoms, exposure to infected individuals, age, muscle pain, recent infection by COVID-19 virus, fever, respiratory distress, loss of smell or taste, nausea, anorexia, headache, vertigo, CT symptoms in lung scans, diabetes and hypertension. The values of accuracy, precision, recall, F1-score, specificity and area under receiver operating curve (AUROC) of different classification models computed in different setting of features scored by MRMR algorithm. Finally, our designed GUI by receiving each of the 42 features and symptoms from the users and through selecting one of the SVM, KNN, Naïve Bayes and decision tree models, predict the result of COVID-19 test. The accuracy, AUROC and F1-score of SVM model as the best model for diagnosis of COVID-19 test were 0.7048 (95% CI: 0.6998, 0.7094), 0.7045 (95% CI: 0.7003, 0.7104) and 0.7157 (95% CI: 0.7043, 0.7194), respectively.

Conclusion: In this study we implemented a machine learning approach to facilitate early clinical decision making during COVID-19 outbreak and provide a predictive model of COVID-19 diagnosis capable of categorizing populations in to infected and non-infected individuals the same as an efficient screening tool.

Key words: COVID-19, diagnosis, machine learning, clinical features, demographic characteristics

Introduction

The novel coronavirus disease 2019 (COVID-19) caused a public health emergency and imposed a significant threat to global health [1]. The first case of the coronavirus disease originated from the provincial capital of Hubei, China on December 2019 and it has since spread throughout the world [2]. In March 2020 World Health Organization (WHO) mentioned COVID-19 as a pandemic that could cause millions of deaths worldwide [3]. Since then, the number of new coronavirus cases has increased as it is extremely transmittable pathogenic viral infection. In addition, the presence of asymptomatic patients in the community increases the prevalence of the disease on a larger scale [4].

This pandemic has also a deeper impact on healthcare services leading to significant challenges in many aspects, including hospital bed shortages, heavy workload, and personnel fatigue which consequently impose severe burden on health systems. Therefore, timely clinical decisions can greatly contribute to the effective and efficient use of available scarce resources. To facilitate rapid and effective diagnosis of COVID-19, effective screening can alleviate the challenges facing healthcare systems [5]. In order to realize this issue and help clinical staff in triaging patients, a number of prediction models have been developed to evaluate the risk of infection. These models use a combination of features including clinical symptoms, laboratory tests, computed tomography (CT) scans [6, 7]. However, most of the previous models have been constructed based on the data obtained from hospitalized patients, which makes it impossible to detect the disease in the community [5]. Furthermore, inadequate provision of screening services and validated tests for COVID-19 diagnosis as well as considerable costs imposed to the healthcare systems of developing countries pose some challenges to the rapid and effective disease diagnosis [8]. The length of time it takes to receive COVID-19 test is also a challenge in the way of rapid and accurate diagnosis of the disease [9].

In responding the issue, artificial intelligence (AI) and its potential use in medicine particularly during the COVID-19 pandemic has evolved to manage the disease effectively [10]. Researchers have used this method in various areas including provision of early warnings, tracking and forecasting, data analysis, diagnosis and prognosis, treatment, care and social controlling [11]. Machine learning, as one of the subsets of artificial intelligence science, has numerous applications in the diagnosis and prediction of various disease types [12].

The applicability of this method has been approved in similar studies conducted to distinguish COVID-19 from other pulmonary disorders [13]. In a study conducted by Mei et al. artificial intelligence algorithms were used to integrate CT scan findings with clinical symptoms, disease history, and laboratory results in order to detect COVID-19 positive patients in a timely manner. The AI system was also able to correctly detect 68 percent of COVID-19 patients out of those who were clinically suspected or had normal CT scan results [14]. Another research by Pourhomayoun and Shakibi applied machine learning and AI algorithms to estimate the risk of mortality among COVID-19 patients based on their clinical symptoms [15].

In this study, we used machine learning method to develop a prediction algorithm to diagnose COVID-19 cases. The model was trained on data gathered from individuals in Qazvin,

northwest city of Iran tested for COVID-19 infection during the first six months of the pandemic. Through the use of machine learning algorithms, the designed model can be used as a diagnostic tool for early screening of infected population.

Materials and methods

A retrospective analysis was conducted on data obtained from 18859 individuals with any signs or symptoms of COVID-19 referred to Shahid Bolandian health center of Qazvin city, Iran to get tested by RT-PCR for the COVID-19 disease between March and August, 2020. Ethical approval of the study was obtained from the Medical Ethics Committee of Qazvin University of Medical Sciences (IR.QUMS.REC.1399.323).

Data collection

Based on the literature review and experts' opinions on the research questions, we selected 42 variables (described in Table 1) to develop machine learning models. The variables incorporated demographic characteristics including age and gender; physical symptoms such as fever, cough, myalgia, respiratory distress, loss of consciousness, anosmia, ageusia, seizure, abdominal pain, nausea, vomiting, diarrhea, anorexia, headache, vertigo, limb paresis, limb plegia, chest pain, skin lesion; clinical symptoms including CT symptoms (symptoms of coronavirus infection in CT scan of chest), SPO2, intubation (whether an intubation was carried out for suspected individual), hypertension, oxygen therapy (whether oxygen therapy was done for suspected individual); history of diseases, and a history of close contact with a COVID-19 patient or previous COVID-19 diagnosis. The outcome variable was COVID-19 test result as a binary variable (1 if the test was positive, and 0 if otherwise) which was determined by RT-PCR test as a gold standard test.

Data set preparation

First, we removed individuals with missing data or incomplete data features. Our gathered data included 42 features with categorical and numeric type variables. We performed label-encoding method to convert categorical features into numeric type (Table S1 in supplementary file). The data set after removing missing values included 18859 individuals comprised of 8983 individuals

with negative test result of COVID-19 (FR: 47.63%) and 9876 infected patients with positive test (FR: 52.37%).

Statistical analysis

We performed correlation analysis between data set features and between features and the target variable (risk of being infected by coronavirus) through the MATLAB 2019b software.

Machine learning models

In this study, we aimed to classify non-infected and infected individuals according to their features by machine learning algorithms. We implemented different classification models including support vector machine (SVM), decision tree (DT), Naïve Bayes (NB) and *K*-nearest neighbor (*K*NN) to predict the result of COVID-19 test for individuals. The SVM classification model is a supervised learning method that separates different classes in a hyperplane to find maximum marginal in multidimensional space. SVM algorithm uses different kernel functions to transform an input feature space into a separable form through adding a space dimension [16]. DT classifier is a supervised algorithm as a predictive modelling framework which can be used in many areas. Tree establishes classification systems based on multiple covariates and splits the dataset into branch-like segments [17]. NB is a probabilistic model for constructing classifiers with independent assumptions between features. NB has a good performance in training even when there is limited amount of data [18]. *K*NN is a non-parametric learning algorithm uses 'feature similarity' to predict the values of new samples by assessing similarity of new sample to the k nearest neighbors according to a specific distance metric or similarity function in the training data set. There is different distance metrics including Euclidean, Hamming or Manhattan distance [19]. We trained these models by using the data of 16973 individuals (90% of all individuals included in data gathering) and tested the models' performance by 10% of the remaining data samples. We performed hyperparameter optimization by Bayesian algorithm through training data set and optimized classifiers were tested by test data set. The values of model assessment metrics presented in next sections are assessed in the test data set. The minimum redundancy maximum relevance (MRMR) algorithm was used to assess the most influential features on the target (result of RT-PCR test for COVID-19 diagnosis). The model performance was measured by various assessment metrics including accuracy, precision, recall,

F1-score, specificity, and area under the receiver operating curve (AUROC). As shown below, these model assessment metrics except for AUROC are computed using True Positive (TP); True Negative (TN); False Positive (FP) and False Negative (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Specificity = \frac{TN}{FP + TN} \quad (5)$$

Finally, we included the trained SVM, NB, DT and KNN models into a user-friendly graphical user interface that allows users to predict the COVID-19 screening test result by recording various symptoms and features of the individuals. The GUI can be used as a primary screening tool to diagnosis COVID-19 in the shortest possible time and at the lowest possible cost.

Results

As shown in Figure 1, the correlations between different data set features (42 features) are calculated and significant correlations (p-value < 0.05) are shown as colored pixels. The black pixels in the figure indicated that there was no significant correlation between the features. The warm colors of heatmap of PRCCs represents the positive correlation between data set features and cold colors describes the negative correlations.

Figure 2 shows the significant correlations between 42 features and the output variable (result of the COVID-19 test). As it is depicted, coughing had the highest positive correlation with the output variable followed by the duration of having COVID-19 signs and symptoms. Exposure to infected individuals was found to be the third strongly correlated feature with a positive SARS-CoV-2 test. Other correlated characteristics included age, muscle pain, recent infection by COVID-19 virus, fever, respiratory distress, loss of smell or taste, nausea, anorexia, headache, vertigo, CT symptoms in lung scans, diabetes and hypertension. Characteristics that were inversely correlated with positive COVID-19 test result included lack of consciousness,

seizures, diarrhea, vomiting, paresis and limb plagia, skin inflammation, SPO_2 level, respiratory rate per minute, chronic blood and kidney disease, heart disease, asthma and lung diseases.

Ranking of the most important features of the model developed for classifying infected and non-infected patients are summarized in Figure 3. MRMR algorithm revealed that five features of great importance for predicting the result of COVID-19 test included age, chronic blood diseases, chest CT findings, exposure to COVID-19 patients, and anorexia.

In the following analyses, we assessed the predictive power of various classifiers. Figures 4 to 7 show the evaluation results of Naïve bayes, KNN, decision tree and support vector machine models, respectively. To achieve this aim, we tested the model performance in the validation data set and we evaluated different classifiers on hundred times. According to these figures, the accuracy, precision, recall, F1-score, specificity, and AUROC were computed by including the first to the 42^{nd} feature scored by MRMR analysis as predictive features into the feature space. As it is shown, by increasing the number of top scored features inserted into the feature space, the various model assessment metrics expressing the predictive power of the model increased.

Finally, Figure 8 shows the designed user-friendly graphical user interface (GUI) as a tool for in silico test of COVID-19. This GUI receives each of the 42 features and symptoms from the users and through selecting one of the SVM, KNN, Naïve Bayes and decision tree models, the prediction for the result of COVID-19 test will be conducted.

Discussion

To the best of our knowledge this is the first study which went beyond the prediction of COVID-19 cases with the aid of artificial intelligence. In fact, we developed a graphical user interface to diagnose COVID-19 by recording various symptoms and features. The developed model in this study can be used as a primary screening tool to predict an individual's health or sickness in a cost-effective manner and at the earliest possible time. The application window contains a simple and practicable interface showing the most likely outcome of an individual (1 = infected; 0 = non-infected) based on their demographic and clinical data.

The achievement of machine learning techniques in recognizing and diagnosing different types of pneumonia has been confirmed in the previous literature. For example, a study conducted by Li et al. generated a convolutional neural network (CNN) framework to

differentiate between COVID-19 and pneumonia. Findings affirmed the successfulness of AI models in accurately diagnosis of pulmonary infections when combining with lung scan results and clinical data [13]. The same technique was applied in diagnosing COVID-19 patients by several countries during the epidemic. In a study conducted to predict individual COVID-19 diagnosis using baseline demographics and laboratory data by Zhang et al. a single decision tree model was developed which used a machine learning-based framework to identify COVID-19 positive diagnosis based on patient demographics, comorbidities, and laboratory test results. Results also affirmed that the prediction model could effectively resolve the screening and testing limitations and facilitate contact tracing procedures [20]. Following the review of existing scientific literature on application of machine learning approaches in COVID-19 diagnosis we understand the significant potential of these models in helping healthcare policy makers and clinicians in accurate and timely diagnosis of the disease [21, 22].

Using correlation tests, we found significant relationships between positive COVID-19 test result and demographic-clinical data including coughing, symptom duration, exposure to infected patients, age, muscle pain, previous infection with coronavirus, fever, respiratory distress, loss of smell or taste, nausea, anorexia, headache, vertigo, lung symptoms in CT scans, diabetes and hypertension variables. Most of these features have been confirmed as identifiers of COVID-19 infection in previous studies. Zhang et al. revealed that features such as age, comorbidities, oxygen saturation, fever, and some of the laboratory test results such as lymphocyte count, and hemoglobin levels were among important predictive factors of positive COVID-19 diagnosis [20]. Literature also mentioned that decreased white blood cell count, anemia, and hypocalcemia were significantly correlated with severe progression of COVID-19 disease [23–26]. In a study using machine learning of clinical data to diagnose COVID-19, authors found that age was significantly correlated with the incidence rates of COVID-19 probably due to the greater vulnerability of elderly population which disproportionally affected them [13]. A prediction model using the support vector machine (SVM) was developed to predict the severe cases of COVID-19 patients by Sun et al. through the use of clinical data and laboratory results features. The authors found age, growth hormone secretagogues (GHSs) and total protein as the main features predicting patients in mild and severe conditions with 0.775 accuracy [25].

Another study by Yao et al. used the SVM model to classify COVID-19 patients based on the disease severity. Findings revealed that age and gender were the strongest classifying features so that male patients above 65 years old were at a higher risk of severe COVID-19 infection. To differentiate people in terms of disease severity, blood test results had the more significant role in compared with urine test result features [27]. Similarly, in a study conducted to predict the mortality of COVID-19 patients using machine learning algorithms by An et al. linear SVM got the highest performance with sensitivity of 0.92, and specificity of 0.91. They found that age, and comorbidities such as diabetes mellitus, and cancer were among important predictors of mortality and COVID-19 severity [28]. Zoabi et al. also predicted COVID-19 test results using machine learning technique based on eight features. Correspondingly the study findings affirmed the predictive role of age above 60 years old, previous contact with an infected individual, and the appearance of cough, fever, sore throat, headache and shortness of breath on positive COVID-19 test result [29]. In another study conducted by Halasz et al. to predict mortality in COVID-19 pneumonia the naïve Bayes classifier was used and six features including age, gender, mean corpuscular hemoglobin concentration, PaO_2/FiO_2 ratio, temperature, and previous stroke were identified to develop a user-friendly website to enable an easy use of the tool by physicians [30].

Differences in the predictive performance of the machine learning models were observed; the highest accuracy was achieved with Support Vector Machine (SVM) in classifying COVID-19 patients from non-infected individuals, signifying the applicability of this model in an early screening of COVID-19 cases. As literature affirmed, scientists favored the SVM due to its acceptable performance in the generalization of data.[31] Villavicencio et al. found that among different machine learning algorithms, SVM was one of the top performers with the highest value in most of the accuracy measures. According to the highest values in accuracy measures SVM was mentioned as the most reliable algorithm in terms of predicting COVID-19 cases based on the given symptoms [32].

In this study we trained different classification models according to data set gathered in Iran. The diagnostic accuracy of COVID-19 PCR kits in Iran are very low (about 75%) [33], therefore the accuracy of our classification models trained by this data were close to this value (accuracy of SVM as the best model was 0.7048 (95% CI: 0.6998, 0.7094).

Strengths and limitations of study

In our study machine learning models benefited from large sample size which improved the accuracy of models. Moreover, the dataset contained data from general population with a small number of missing data. Additionally, an added advantage of our model is its potential to be integrated with electronic health records which consequently improves the possibility of generating real-time predictions based on dynamicity of clinical data and laboratory values. There are some limitations regarding the study. First, machine learning models have not addressed how diverse comorbidities affect the diagnosis of COVID-19. Categorizing data into different subgroups and developing models would properly facilitate the identification of associations between specific comorbidities and the disease prediction. Second, further studies should be done to add laboratory test results and additional symptoms including lactate dehydrogenase (LDH), neutrophils, lymphocyte, and C-reactive protein in to the findings. Third, the diagnostic accuracy of COVID-19 PCR kits in Iran are very low and our models are trained by gathered data set in Iran, therefore the accuracy of our models are close to the accuracy of PRC kits in Iran.

Conclusion

Early detection of COVID-19 infected people can help health authorities allocate the limited resources of healthcare systems to high-risk individuals needing urgent clinical care during the pandemic. Although patient characteristics, such as lung CT scan findings, exposure to COVID-19 infected person, anorexia, chronic blood disease, age, symptoms duration, etc were proved to be associated with positive COVID-19 test, there is no available application that can reliably and easily be used by healthcare personnel in immediately diagnosis of COVID-19 patients based on their demographic and clinical characteristics. In this study we implemented a machine learning algorithm to facilitate early clinical decision making during COVID-19 outbreak and provide a predictive model of COVID-19 diagnosis capable of categorizing populations in to infected and non-infected individuals the same as an efficient screening tool.

Conflict of interest

None declared.

References

1. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020; 369: m1328, doi: 10.1136/bmj.m1328, indexed in Pubmed: 32265220.

2. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. NPJ Digit Med. 2021; 4(1): 3, doi: 10.1038/s41746-020-00372-6, indexed in Pubmed: 33398013.

3. World Health Organization, WHO director-general's opening remarks at the media briefing on COVID-19 March 2020.

4. Lei S, Jiang F, Su W, et al. Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of COVID-19 infection. EClinicalMedicine. 2020; 21: 100331, doi: 10.1016/j.eclinm.2020.100331, indexed in Pubmed: 32292899.

5. Gangloff C, Rafi S, Bouzillé G, et al. Machine learning is the key to diagnose COVID-19: a proof-of-concept study. Sci Rep. 2021; 11(1): 7166, doi: 10.1038/s41598-021-86735-9, indexed in Pubmed: 33785852.

6. Geelhoed GC, de Klerk NH. Emergency department overcrowding, mortality and the 4-hour rule in Western Australia. Med J Aust. 2012; 196: 122–126, doi: 10.5694/mja11.11159, indexed in Pubmed: 22304606.

7. Kim JS, Bae HJ, Sohn CH, et al. Maximum emergency department overcrowding is correlated with occurrence of unexpected cardiac arrest. Crit Care. 2020; 24(1): 305, doi: 10.1186/s13054-020-03019-w, indexed in Pubmed: 32505196.

8. Bai HX, Hsieh B, Xiong Z, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. Radiology. 2020; 296(2): E46–E54, doi: 10.1148/radiol.2020200823, indexed in Pubmed: 32155105.

9. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med. 2020; 26(8): 1224–1228, doi: 10.1038/s41591-020-0931-3, indexed in Pubmed: 32427924.

10. Ozsahin I, Sekeroglu B, Musa MS, et al. Review on diagnosis of COVID-19 from chest CT images using artificial intelligence. Comput Math Methods Med. 2020; 2020: 9756518, doi: 10.1155/2020/9756518, indexed in Pubmed: 33014121.

11. Naudé W. Artificial intelligence against Covid-19: An early review. SSRN Electronic Journal. , doi: 10.2139/ssrn.3568314.

12. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017; 15: 104–116, doi: 10.1016/j.csbj.2016.12.005, indexed in Pubmed: 28138367.

13. Li L, Qin L, Xu Z, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy. Radiology. 2020; 296(2): E65–E71, doi: 10.1148/radiol.2020200905, indexed in Pubmed: 32191588.

14. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med. 2020; 26(8): 1224–1228, doi: 10.1038/s41591-020-0931-3, indexed in Pubmed: 32427924.

15. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health (Amst). 2021; 20: 100178, doi: 10.1016/j.smhl.2020.100178, indexed in Pubmed: 33521226.

16. Suthaharan S. Support Vector Machine. Machine learning models and algorithms for big data classification. 2016: 207–235, doi: 10.1007/978-1-4899-7641-3_9.

17. Du W, Zhan Z. Building decision tree classifier on private data. 2002.

18. Keogh E. Naive bayes classifier. " Accessed: Nov. 2006.

19. Parvin, Hamid, Hoseinali Alizadeh, and Behrouz Minati. A modification on k-nearest neighbor classifier. Global Journal of Computer Science and Technology (2010). 2010.

20. Zhang J, Xie Y, Li Y, et al. COVID-19 screening on chest x-ray images using deep learning based anomaly detection. arXiv. Preprint arXiv. 2020; 200312338.

21. Yan KK, Wang X, Lam WWT, et al. Radiomics analysis using stability selection supervised component analysis for right-censored survival data. Comput Biol Med. 2020; 124: 103959, doi: 10.1016/j.compbiomed.2020.103959, indexed in Pubmed: 32905923.

22. Sevakula RK, Au-Yeung WTM, Singh JP, et al. State-of-the-art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system. J Am Heart Assoc. 2020; 9(4): e013924, doi: 10.1161/JAHA.119.013924, indexed in Pubmed: 32067584.

23. Oliveira BA, Oliveira LC, Sabino EC, et al. SARS-CoV-2 and the COVID-19 disease: a mini review on diagnostic methods. Rev Inst Med Trop Sao Paulo. 2020; 62: e44, doi: 10.1590/S1678-9946202062044, indexed in Pubmed: 32609256.

24. Liu J, Han P, Wu J, et al. Prevalence and predictive value of hypocalcemia in severe COVID-19 patients. J Infect Public Health. 2020; 13(9): 1224–1228, doi: 10.1016/j.jiph.2020.05.029, indexed in Pubmed: 32622796.

25. Sun JK, Zhang WH, Zou L, et al. Serum calcium as a biomarker of clinical severity and prognosis in patients with coronavirus disease 2019. Aging (Albany NY). 2020; 12(12): 11287–11295, doi: 10.18632/aging.103526, indexed in Pubmed: 32589164.

26. Cavezzi A, Troiani E, Corrao S. COVID-19: hemoglobin, iron, and hypoxia beyond inflammation. A narrative review. Clin Pract. 2020; 10(2): 1271, doi: 10.4081/cp.2020.1271, indexed in Pubmed: 32509258.

27. Yao H, Zhang N, Zhang R, et al. Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests.

Front Cell Dev Biol. 2020; 8: 683, doi: 10.3389/fcell.2020.00683, indexed in Pubmed: 32850809.

28. An C, Lim H, Kim DW, et al. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. Sci Rep. 2020; 10(1): 18716, doi: 10.1038/s41598-020-75767-2, indexed in Pubmed: 33127965.

29. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. NPJ Digit Med. 2021; 4(1): 3, doi: 10.1038/s41746-020-00372-6, indexed in Pubmed: 33398013.

30. Halasz G, Sperti M, Villani M, et al. A machine learning approach for mortality prediction in COVID-19 pneumonia: development and evaluation of the Piacenza Score. J Med Internet Res. 2021; 23(5): e29058, doi: 10.2196/29058, indexed in Pubmed: 33999838.

31. Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. IEEE Trans Neural Netw. 1999; 10(5): 1055–1064, doi: 10.1109/72.788646, indexed in Pubmed: 18252608.

32. Villavicencio C, Macrohon J, Inbaraj X, et al. COVID-19 prediction applying supervised machine learning algorithms with comparative analysis using WEKA. Algorithms. 2021; 14(7): 201, doi: 10.3390/a14070201.

33. Tabnak News Agency. Real numbers of deaths and vases are 2.5 times higher than the official reports. News code 1009679. 2020-10-18. http://tabnak ir/004Ef9 (19 Oct 2020).

Table 1. Data set features

categories	Feature	Feature #
—	Age	1
(Male (1), Female (0	Gender	2
(Yes (1), No (0	Being exposed by coronavirus	3

(Yes (1), No (0	Previous COVID-19 diagnosis	4
(Yes (1), No (0	Fever	5
(Yes (1), No (0	Cough	6
(Yes (1), No (0	Muscular pain	7
(Yes (1), No (0	Respiratory distress	8
(Yes (1), No (0	Loss of consciousness	9
(Yes (1), No (0	Anosmia	10
(Yes (1), No (0	Lack of sense of taste	11
(Yes (1), No (0	Convulsion	12
(Yes (1), No (0	Abdominal pain	13
(Yes (1), No (0	Nausea	14
(Yes (1), No (0	Vomiting	15
(Yes (1), No (0	Diarrhea	16
(Yes (1), No (0	Anorexia	17
(Yes (1), No (0	Headache	18
(Yes (1), No (0	Vertigo	19
(Yes (1), No (0	Limb paresis	20
(Yes (1), No (0	Limb plegia	21
(Yes (1), No (0	Chest pain	22
(Yes (1), No (0	Skin inflammation	23
—	Time duration from symptoms on	24
(Yes (1), No (0	Intubation	25
—	SPO2	26
(Less than 5 (1 (Between 5–14 (2 (Between 14–18 (3 (Between 18–22 (4 (Between 22–28 (5 (Bigger than 28 (6	Number of breaths per minutes	27
—	Body temperature	28
(Yes (1), No (0	CT finding	29
(Yes (1), No (0	Chronic liver diseases	30
(Yes (1), No (0	Diabetic diseases	31
(Yes (1), No (0	Chronic blood diseases	32
(Yes (1), No (0	HIV/AIDS	33
(Yes (1), No (0	Immunodeficiency	34
(Yes (1), No (0	Pregnancy	35
(Yes (1), No (0	Heart diseases	36
(Yes (1), No (0	Chronic kidney diseases	37
(Yes (1), No (0	Asthma patient	38

(Yes (1), No (0	Lung diseases except asthma	39
(Yes (1), No (0	Hypertension	40
(Yes (1), No (0	Oxygen therapy	41
—	Duration of hospitalization	42

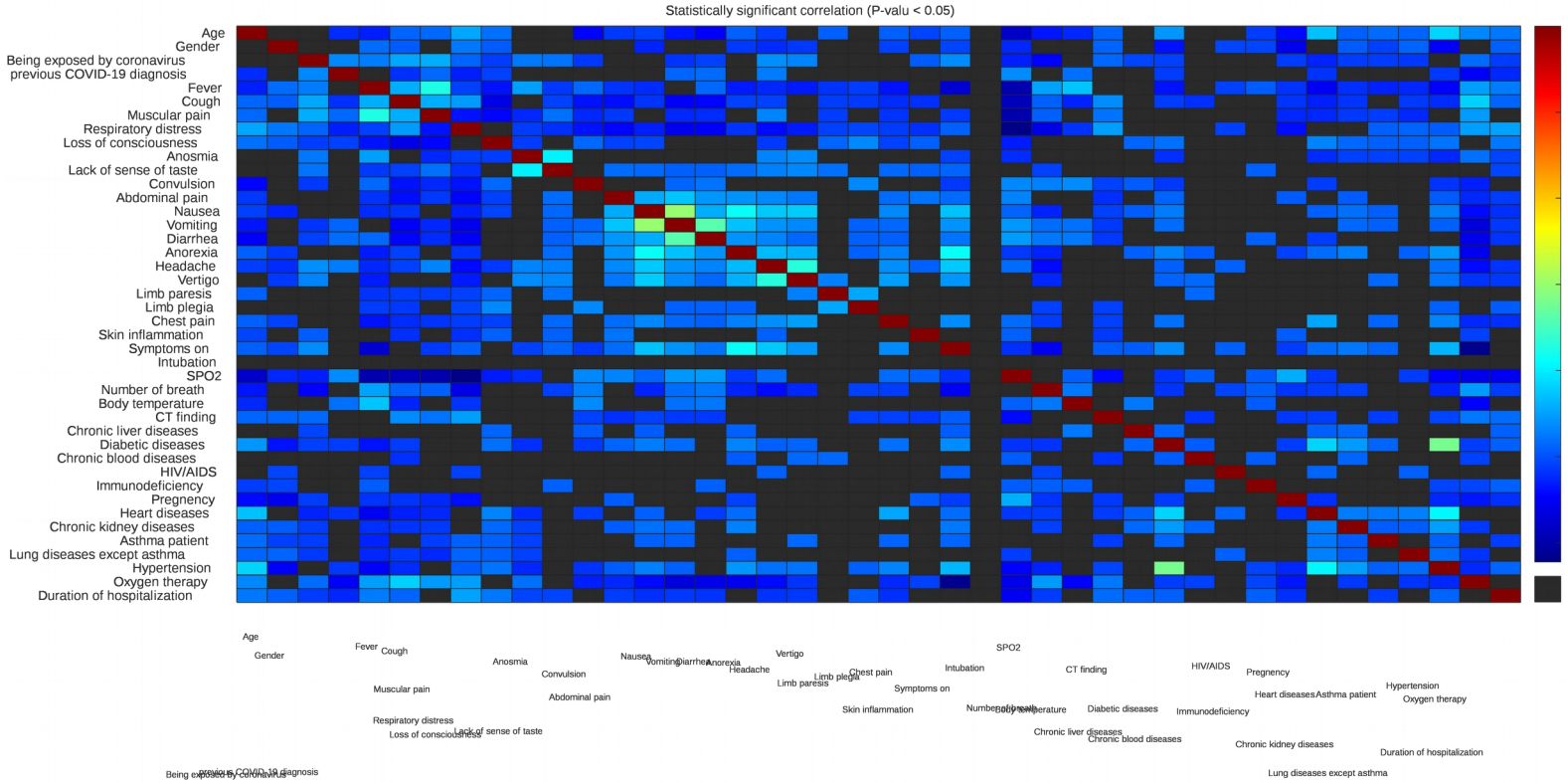

Figure 1. Statistically significant correlation between data set features

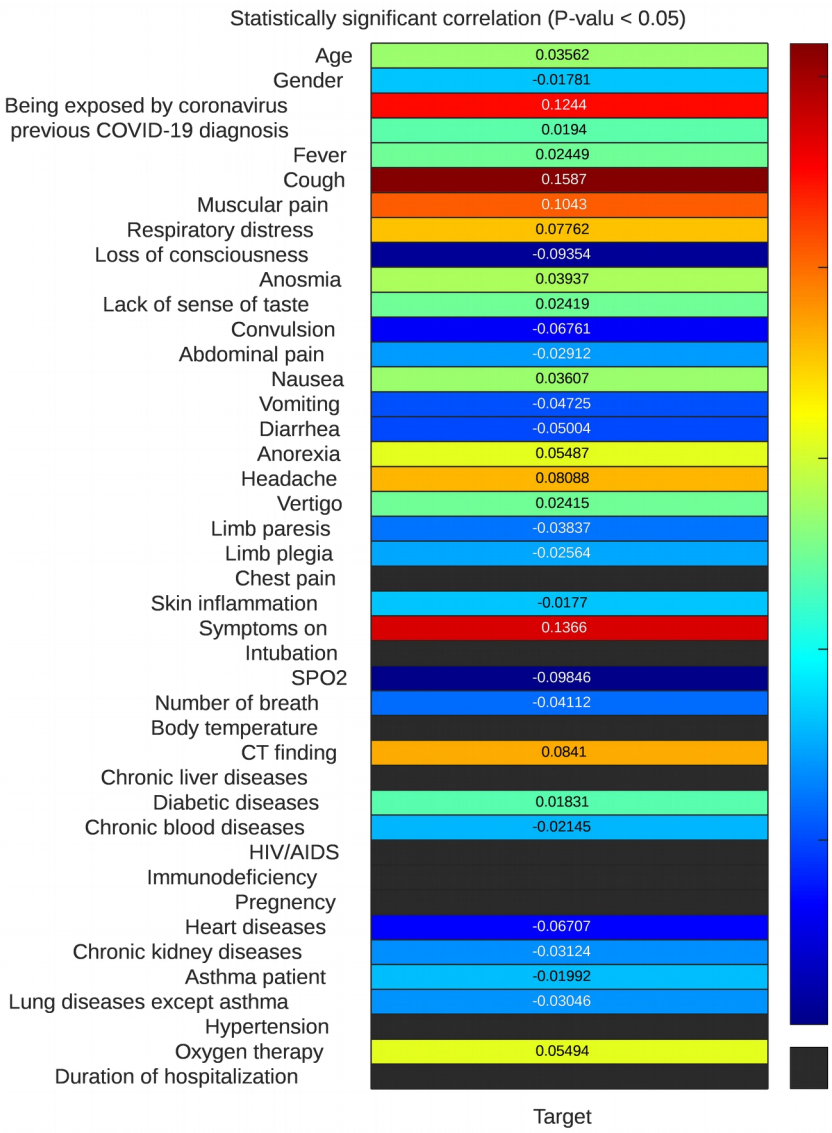

Figure 2. Statistically significant correlation between data set features and the target variable

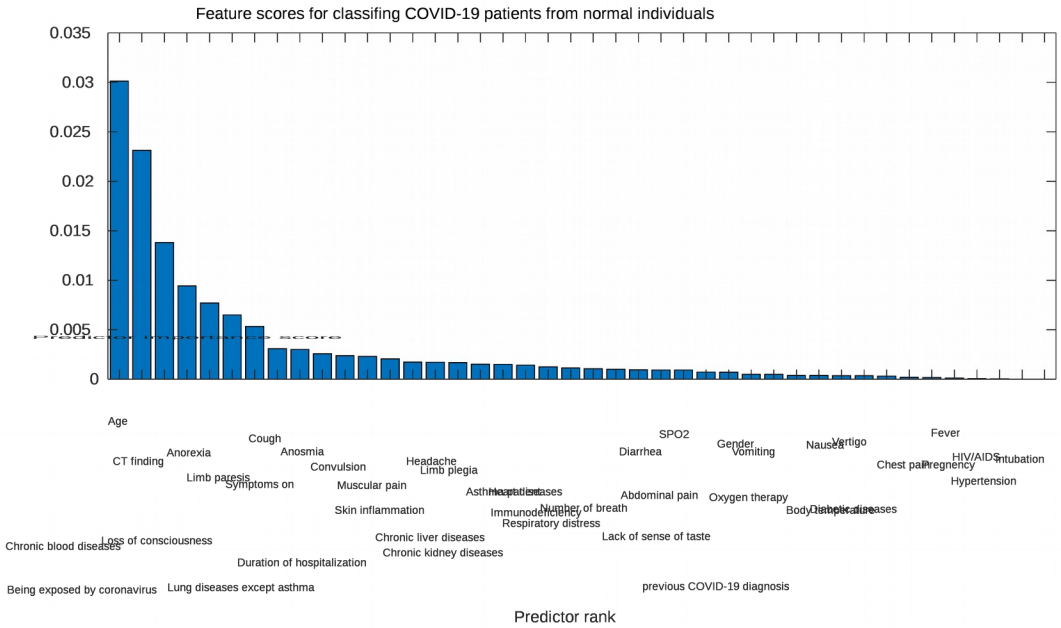

Figure 3. The Maximum relevance-minimum redundancy algorithm scored data set features according to their importance for classifying normal and infected patients

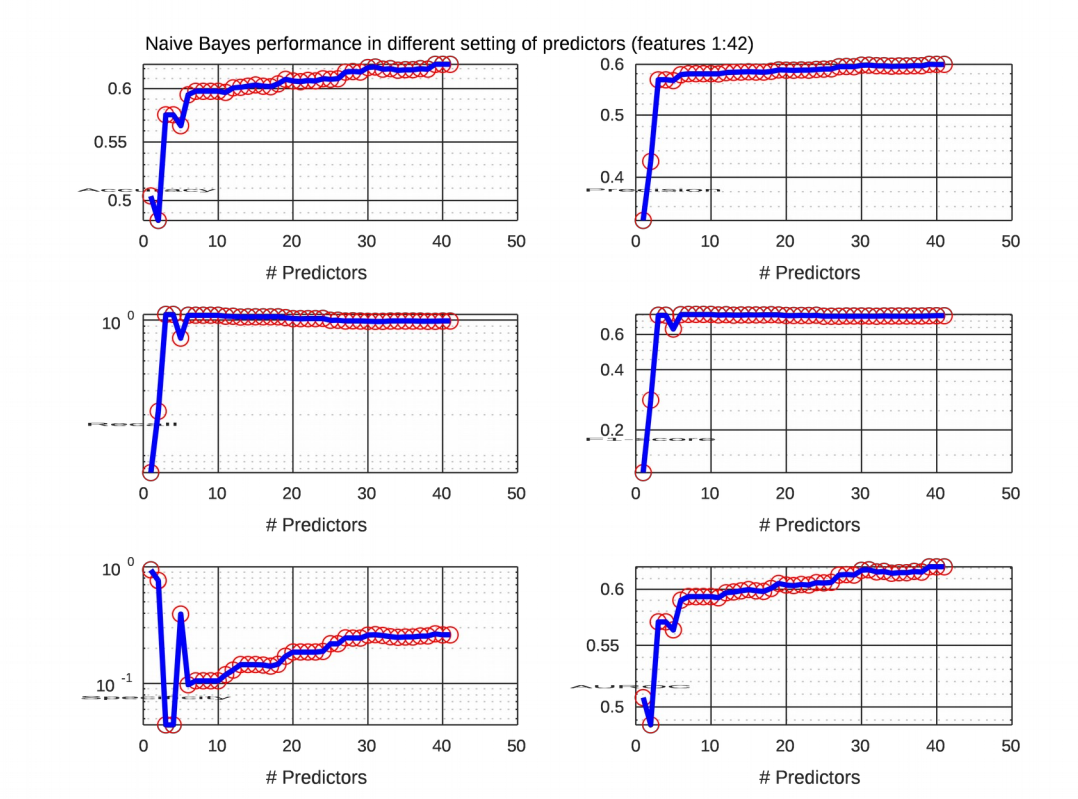

Figure 4. Naïve Bayes assessment metrics for classifying normal and infected individuals in different settings of top scored features by MRMR algorithm

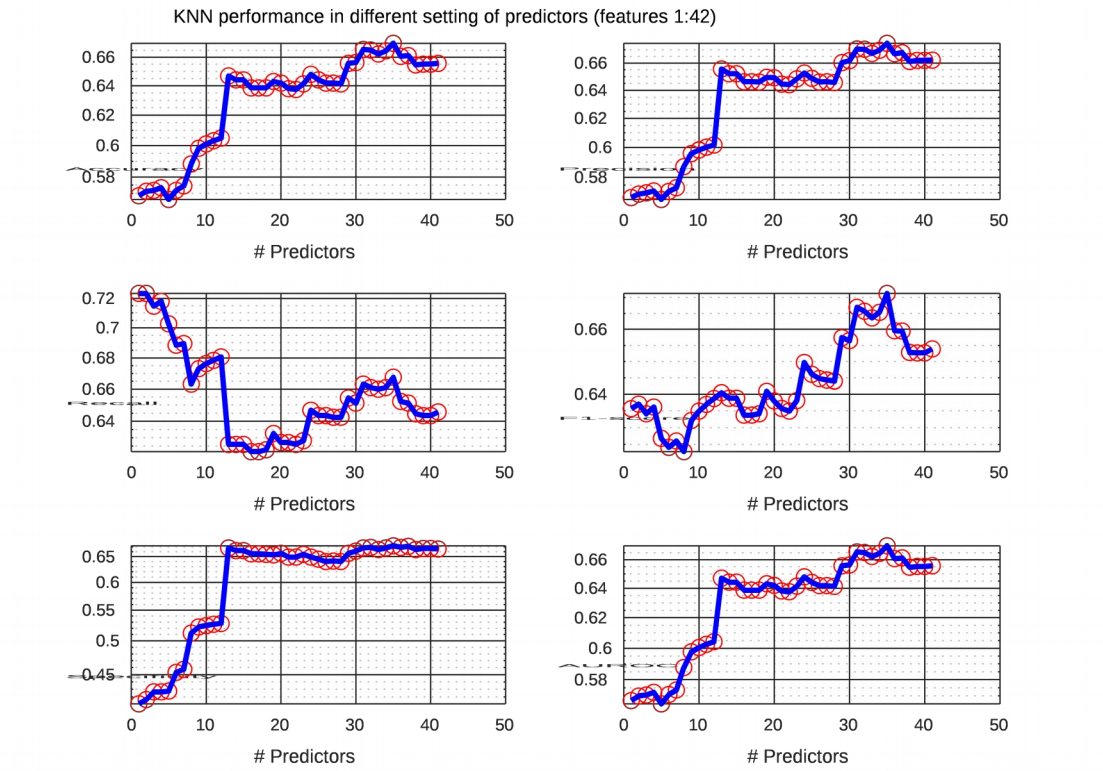

Figure 5. KNN assessment metrics for classifying normal and infected individuals in different settings of top scored features by MRMR algorithm

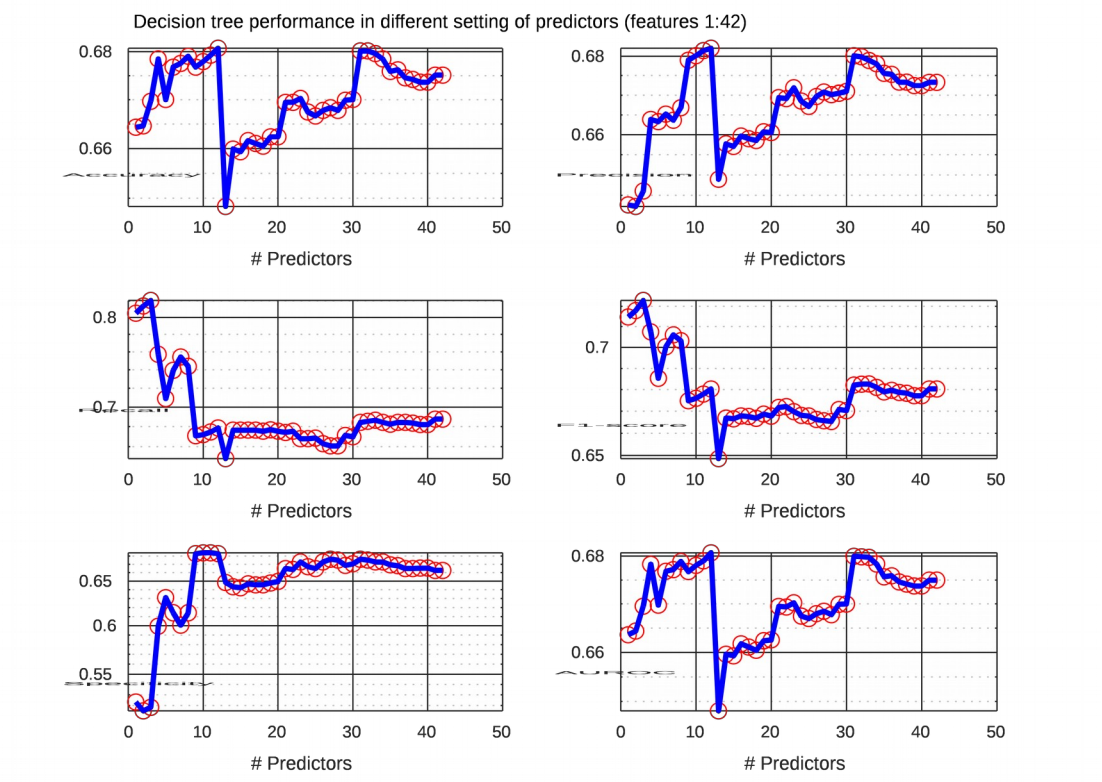

Figure 6. Decision tree assessment metrics for classifying normal and infected individuals in different settings of top scored features by MRMR algorithm

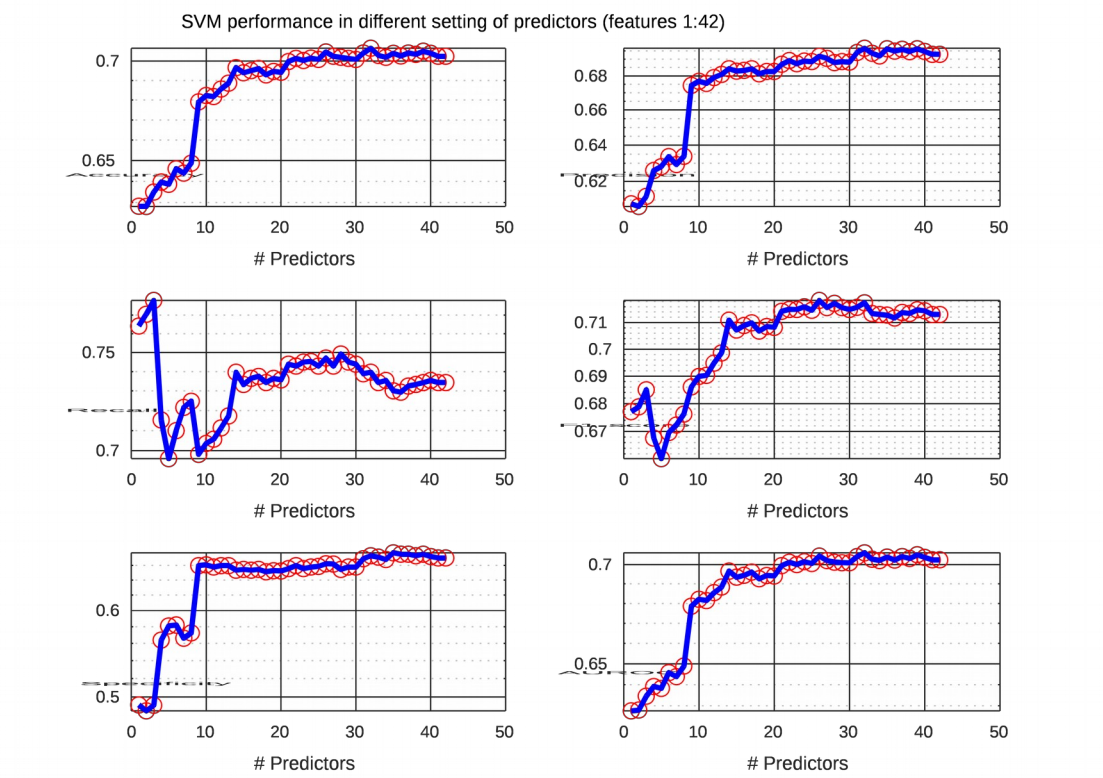

Figure 7. SVM assessment metrics for classifying normal and infected individuals in different settings of top scored features by MRMR algorithm

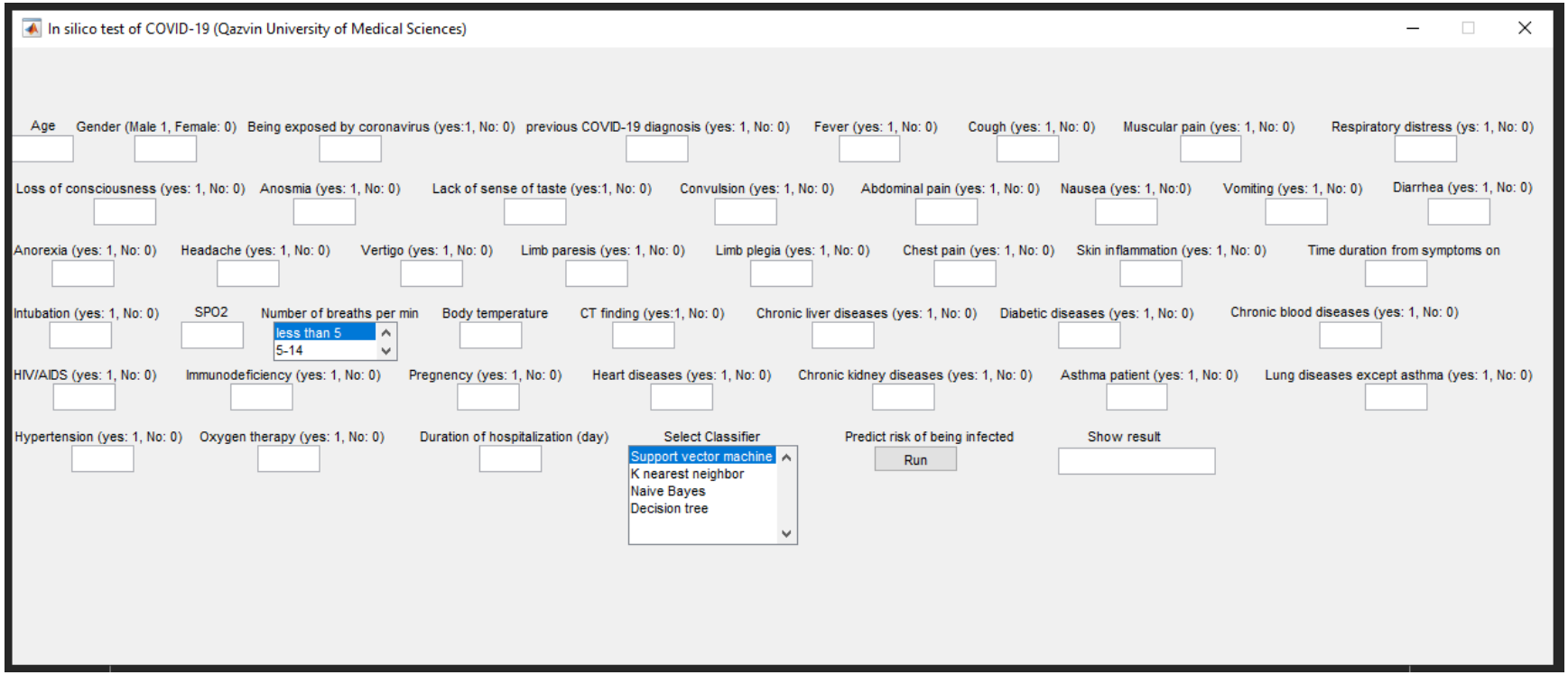

Figure 8. Graphical user interface unit for silico test of COVID-19