Chapter 23

Artificial Intelligence Applied to the Analysis of Relevance of Laboratory Biomarkers for Diagnosis of COVID-19

https://doi.org/10.56238/colleinternhealthscienv1-023

R. B. de Alvarenga Neto
Graduate Program in Control and Automation Engineering - PROPECAUT - Ifes, Vitoria, Brazil

D. C. de A. Pinto Gomes
Laboratory of Neurochemistry and Behavior, Graduate Program in Biochemistry, Center for Health Sciences, UFES, Vito'ria, Brazil

L. A. Pinto
Graduate Program in Control and Automation Engineering - PROPECAUT - Ifes, Vitória, Brazil

ABSTRACT

This paper presents a study on the diagnosis of COVID-19 through the analysis of biomarkers from clinical tests using Machine Learning techniques. The research was conducted in two stages. In the first, the influence of clinical markers on the diagnosis of the disease was investigated, and in the second, the performance of several Machine Learning algorithms was investigated in the classification of patients with symptoms similar to COVID-19.

The experiments used a database provided by Hospital Israelita Albert Einstein with laboratory test results from 5,644 patients submitted to the RT-PCR test. Among the results found in the variable selection stage, indicators from the leukocyte group were more relevant for COVID-19 detection. In the classification step, the best results were obtained for the

The results were obtained with Stacking using 20 descriptors selected by Decision Tree (Acuracy = 0.9778; Sensitivity = 0.9527; Specificity = 1.000). The results indicate

it is feasible to use Machine Learning techniques together with variable selection to obtain models with good predictive power for the diagnosis of COVID-19.

Keywords: diagnostic de COVID-19, selecion de varia'veils, machine learning, biomarkers from cl'ınicos exams.

1 INTRODUCTION

In December 2019, an epidemic of acute pneumonia began in China that has been termed COVID-19, caused by the new coronavirus (SARS-CoV-2), which has spread rapidly, infecting people around the world. According to the Coronavirus Resource Center (CRC), a repository of information about the pandemic maintained by the Johns Hopkins University School of Medicine, from the beginning of the pandemic to the current date (May 28, 2022), 528,527,600 cases have been reported worldwide, with a total number of deaths of 6,286,811. In Brazil, until that date there were 30,921,145 registered cases with 666,319 deaths, [2].

Research has proposed the use of laboratory parameters, obtained from the patient's admission exams in the hospital as indicators of infection and disease progression. The research of [3] and [4] concluded that the percentage of lymphocytes in the blood (LYM%) is the most important and consistent among the hematological parameters for diagnosis of COVID-19. In the work of [5], a comparative analysis of hematologic parameter values was performed between patients with COVID-19 with moderate-severity disease compared to those with high-severity disease. The values of interleukin-6 (IL-6), D-dimer (D-D), glucose, thrombin time, fibrinogen, and C-reactive protein showed significant differences between the two groups. Other studies have also proposed the association between COVID-19 and the values of biomarkers obtained through conventional clinical exams.

The use of Machine Learning techniques for the elaboration of disease diagnosis based on biochemical markers was proposed in [6], [7] and [8]. Specifically in relation to COVID-19, the works of [9], [10] and [11], investigated the use of Machine Learning to diagnose and predict disease progression from laboratory biomarker values. In a study by [12], the authors used Machine Learning techniques and neural networks to predict the need for hospitalization. According to the authors in [12], patients with positive SARS-CoV-2 diagnosis showed an increase in monocytes and a reduction in platelets, leukocytes, eosinophils, basocytes, and lymphocytes.

In this context, this paper presents a study on the application of Machine Learning algorithms to assist in the diagnosis and prediction of the evolution of COVID-19, through clinical parameters obtained from hospitalized patients. For this purpose a dataset consisting of biomarkers from the former groups of CBC, biochemical, viral panel, venous blood gas, arterial blood gas, and urine was used. The original dataset contains 111 variables, corresponding to the biochemical markers obtained through laboratory tests of 5,644 patients, made available publicly and free of charge by Hospital Israelita Albert Einstein1 , located in the city of Sa˜o Paulo - Brazil.

The ensembles of individual classifiers built with Bagging, Boosting and Stacking algorithms, consisting of the individual k-Nearest-Neighbor, A' Decision Tree and Support Vector Machine (SVM) classifiers, were used to build the classification models.

2 THEORETICAL FRAMEWORK

In this section we briefly describe the Relief-F and A' Decision Tree variable selection algorithms, as well as the Bagging, Boosting and Stacking model combination algorithms used for classification.

A. Variable selection

Relief-F [13] is based on the logic of the k-nearest-neighbors algorithm, and has as its central idea the estimation of the quality of variables based on their ability to distinguish classes of samples that are close to each other in the sample space, being able to correctly estimate the quality of the attributes in problems with strong dependency among them. Furthermore, according to [14], the estimates of the importance of the variables are easily interpretable and contextualized in the problem domain. According to [15], Relief-F can be applied well to remove irrelevant features, but it is not necessarily the most appropriate to select the best among the most relevant features.

Decision trees are methods that use a tree-like structured graphical representation, whose goal is to identify groups of objects with common characteristics. To determine the importance of the variables, the decision tree is based on the information gain at each stage, and thus determines the relative information gain among different variables [16]. In this work, the selection of variables by decision tree was implemented using CART, which is an algorithm that implements multivariate trees.

B.	Combination Algorithms of Classifiers

In supervised learning, the technique of combining classifiers, ensemble, also known as multiple classifier system, is used to improve the performance of unstable classifiers or those with low predictive power [17]. This paper discusses the three main ensemble models, namely Boosting, Bag- ging and Stacking.

Boosting acts to reduce the variational error that arises when models are not able to identify relevant trends in the data. This happens by evaluating the difference between the predicted value and the actual value. The algorithm trains the classifiers in sequential order and for each classification cycle the weights for the classification errors are updated. They are increased as a form of punishment for incorrectly classified samples and reduced for correctly classified samples. The training set used for each member is chosen based on the performance of the previous classifiers in the session. In Boosting, the examples incorrectly predicted by the previous classifiers of the session are chosen more often than the correctly predicted examples [18].

Bagging (Bootstrap Aggregating) is a classifier combination technique designed to increase the accuracy of the prediction that uses resampling with replacement to make the selection procedure completely random. When a sample is selected without replacement, subsequent variable selections are always dependent on the previous selections, making the criteria non-random. In Bagging, the resampling of the training set does not depend on the performance of previous classifiers [17].

Stacking is an ensemble learning method that combines several machine learning algorithms by means of meta-learning. The algorithm trains a second-level meta-classifier considering the prediction result of the individual classifiers. In this ensemble approach, the individual classifiers are trained on a complete training data set, and

the meta-classifier is trained with the final results of the basic-level model. In general, Stacking can improve the accuracy of the prediction of individual models [19].

3 ME'ALL

This section presents the data set used, the description of the preprocessing step, and the configurations of the variable selection algorithms and classifiers.

A. Data Set

The research was developed on a set of data made available by Hospital Israelita Albert Einstein. O The set consists of 111 variables, most of which are biomarker values from laboratory tests, in addition to the level of unbalance. To mitigate the effect of the unbalance, the Adasyn algorithm (Adapage quartile and the result of RT-PCR tests of 5,644 pative Synthetic) [21]. Adasyn is based on the algorithm patients with symptoms similar to those of COVID-19. In all cases, laboratory tests (complete blood count, viral panel, biochemicals, venous blood gas, arterial blood gas and urine) were performed when the patients were admitted to the hospital. After the admission exams, the patients were placed in hospital, and each one was referred to the initial treatments, according to the results of the exams performed at admission. Patients with mild symptoms were referred to a ward, patients diagnosed with intermediate symptoms were referred to semi-intensive care, and patients with severe symptoms were admitted to the ICU.

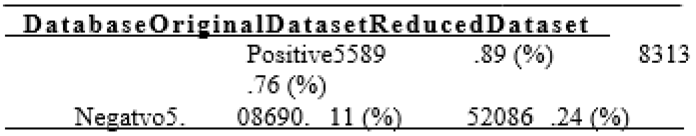
The data set presents problems, such as a large amount of missing data and unbalance of the number of samples between classes (about 90% concentrated in the negative class), which causes bias in the classification. To reduce the amount of unknown values, the samples with a high number of attributes with unknown values were manually removed. After this removal step, of the 5,644 patients that were included in the original set, 603 samples (patients) and 37 variables (exams retained for modeling) were left for the subsequent steps, which is equivalent to 11% of the total. Of the 603 patients retained, 83 (14%) were diagnosed positive for SARS-CoV-2. Since a significant majority of the samples with a large number of missing test values belonged to tests from the viral panel, venous blood gas, arterial blood gas, and urine groups, it was decided to eliminate all tests from these groups. Thus, in the variable selection and classifier construction phases, only the biochemical markers of the complete blood count and biochemistry tests groups were used.

B. Data preprocessing

In the original set, before the manual removal of variables, 88% of the attribute values were unknown. After manual removal of the variables with large amounts of missing values, the total number of unknown values among the retained variables was reduced to approximately 37.65%. To fill in the missing attributes, we used the Alternating Least Square method [20], which estimates the values of these attributes based on the values of the known attributes. Table 1 shows the number of samples per class. As can be seen, the data set has a very high ellipsis.

Smote, its goal being to reduce the bias introduced by the unbalanced distribution, and especially to adaptively shift the decision frontier to focus on the most difficult-to-learn examples. The principle behind this algorithm is to use the density distribution as a criterion to automatically decide the number of synthetic data that need to be generated for each example of the monetary class [21].

Table 1: Database unbalance



DatabaseOriginalDatasetReducedDataset			
	Positive5589 .76 (%)	.89 (%)	8313
Negatvo5.	08690. 11 (%)	52086 .24 (%)	

C. Variable Selection and Classification

The research was conducted in two stages. In the first stage, the Relief-F and Decision Tree-based techniques of variable selection were used to determine the order of influence of biomarkers in the construction of the classification models. Tests with the Relief-F algorithm were performed for k equal to 10, 20, 30, 40, 50 and 60, and for the Decision Tree, the maximum number of division numbers was n = 1, 5, 10, 15 and 20.

In the second step we built ensembles of clas- sifiers based on Bagging, Boosting, and Stacking. For the construction of the Bagging ensemble, Random Flo- rests with 10, 20, 50, 100 and 200 trees were created. For the

Boosting was used with Adaboost with Decision Trees with depth equal to 1 and number of iterations equal to 50, 100, 150, 200, 250 and 300. For the construction of the ensemble by Stacking, the algorithms selected were Decision Tree with 1, 10, 20 and 30 division no., k-NN with 1, 3, 5 and 7 neighbors, and SVM with kernels, Polynomial, Gaussian and Linear. For the experimental stage, the data set was partitioned, with 70% of the samples for the training/validation set and 30% for the test set, having

The k-Fold algorithm with k = 5 was used.

All tests were performed in the Mat- environment.

lab®R2018b on an Intel(R) Core(TM) i5-2410M CPU@2.30 GHz machine and 8GB RAM.

4 RESULTS

The results of the variable selection methods and the classification ensembles are presented below.

A. Variable selection

Tables 2 and 3 show, in descending order, the list of the 10 most relevant variables obtained by the selection algorithms Relief-F and Decision Tree, respectively. The columns "Group" in the tables refer to the groups CBC (H) and Biochemistry (B), to which the descriptors selected for the test phase belong. The letter I in the group columns corresponds to the age quartile of the patients. For ranking, in each case, the score of each variable was calculated as the average of the scores obtained in all algorithm configurations.

Table 2: Variable selection - Relief-F

Rank	Group	Descriptor
1	H	Leucocitos
2	H	Eosinofilos
3	H	Red Gloves
4	H	Platelets
5	I	Age quartile
6	B	Sodio
7	H	Monocytes
8	B	Alkaline phosphatase
9	H	Hematocrite
10	H	Neutrofilos

When comparing the results of the two varietal selection techniques with the reference works presented in Section I, we can observe the importance of blood count biomarkers, especially those belonging to the Leukocyte family, which had 4 markers (leukocytes, eosinophils, monocytes, and neutrophils) present among the 10 most relevant for both varietal selection methods. For the Biochemical marker group, sodium obtained the best ranking position (sixth most important variable by Relief-F and seventh by the A' decision tree), This is in agreement with the studies of [22], which suggests that the rich serum can be used to identify the risk of disease progression to more severe stages in patients with COVID-19. Another highlight among the descriptors in the biochemical group was C-reactive protein, a fact that is in agreement with the research carried out by [23], [5], [4].

B. Classification

In the second stage, the classification stage, ensembles (Bagging, Boosting, and Stacking) of classifiers were built to improve the predictive accuracy and reduce the error components due to bias and variance. The tables with the best accuracy results are presented below

Table 3: Variable selection - Decision tree

Rank	Group	Descriptor
1	H	Eosinofilos
2	H	Leucocitos
3	H	Platelets
4	H	Red Gloves
5	H	Promyelocytes
6	B	C-reactive protein
7	B	Sodio
8	B	Hematocrite
9	H	Monocytes
10	I	Age quartile

(Acc.), Sensitivity (Sens.), and Specificity (Spec.), as well as the number of descriptors (Desc.) used to build the models and the Variable Selection (SV) method. The labels DT and RFF in the tables refer to the variable selection methods A' Decision and Relief-F trees, respectively, applied for the classification with Bagging (Table 4), Boosting (Table 5), and Stacking (Table 6) ensembles.

Table 4 shows the best results presented by Bagging, where the first column contains the number of trees for the construction of Random Forests. The best result obtained for this ensemble was with 200 trees, with 37 described.

Table 4: Bagging classification

N Árv	Acc.	Sens.	Spec.	SV	Desc.
	100,	94000,	93920,	9397DT	
	25				
	200,	94980,	94870,	9492DT30	
	500,	95650,	95510,	9556DT30	
	1000,	95870,	95950,	9595DT30	
	2000,	**96290,**	**96150,9619**	**-**	**37**

The results of the classification using Boosting (Ad- aboost) are shown in Table 5. The first column indicates the number of iterations used for testing the Boosting en- sembles. The best result was obtained with 200 iterations, and the variables were selected by the random forest.

Stacking ensembles were constituted with the individual classifiers, A Decision Tree, with

different numbers of partition numbers, k-NN, with different values of k, and SVM, with different types of kernel. Table 6 presents the results obtained for the individual classifiers and for the Stacking meta-classifier.

The models were obtained for different configurations of the classifiers and different amounts of selected variables. The best results were achieved with

Table 5: Boosting classification

N iter.	Acc.	Sens.	Spec.	SVDesc	
			.		
50	0.94290	.	94460	RFF	30
			.9446		
100	0.94920.	95060	.9506	DT	20
150	0.95240.	95390	.9539	RFF	30
200	**0.96190**	**.**	**96330**	**RFF**	**30**
			.9633		
250	0.95870	.	96030	RFF	30
			.9603		
		3000.9587		0.96030 .	
		9603RFF30			

Table 6: Stacking classification for variable selection by decision tree with 20 descriptors

StackingAcc	.	Sens.	Spec.
Tree-1	0.8413	0.8986	0.7904
Tree-10	0.8508	0.8311	0.8683
Tree-20	0.8190	0.8446	0.7964
Tree-30	0.8381	0.8176	0.8563
kNN-1	0.8698	0.9797	0.7725
kNN-3	0.8381	0.9865	0.7066
kNN-5	0.8095	0.9797	0.6587
kNN-7	0.8032	0.9662	0.6587
SVM-Polinomial	0.9048	0.9730	0.8443
SVM-Gaussian	0.9175	0.8243	1.0000
SVM-Linear	0.8603	0.8986	0.8263
	STK0.	**97780.**	**95271.0000**

Stacking for the top 20 descriptors selected by the Decision Tree, with the following results: Accuracy = 0.9778; Sensitivity = 0.9527; Specificity = 1.000, demonstrating the feasibility of using Machine Learning techniques to diagnose COVID-19 from a reduced set of conventional laboratory tests, which can reduce the time and cost of diagnosing the disease for patients requiring hospital admission or who have been treated in an outpatient clinic.

C. Comparison with related work

The authors in [24] diagnosed Covid-19 using convolutional neural networks and lung X-rays from different data sets. The results obtained in terms of accuracy ranged from 96% to 99%. In [25], a 95.33% accuracy was obtained in the Covid-19 diagnosis using a combination of convolutional neural networks and Support Vector Machine (SVM). In [26], the authors used several classification and variable selection algorithms to diagnose Covid-19 from biochemical markers of blood and urine tests. The best result based on model accuracy was 95.169%.

Comparing the best result obtained in the present work, where the accuracy was 97.78%, with the results of the research presented in this section, it can be seen that the results of the proposed method are comparable to the diagnostic approaches used both in traditional machine learning methods and in Deep Learning-based approaches.

5 CONCLUSION

This research investigated the development of techniques to diagnose COVID-19 in patients who were admitted to hospitals with symptoms consistent with the disease. To this end, Machine Learning techniques were used to build classification models. The experiments were performed on a dataset provided by Hospital Israelita Albert Einstein, (Sa˜o Paulo/Brazil). The dataset, with a high level of sample imbalance between classes and a large number of missing values, was composed of a collection of biomarkers obtained through laboratory tests pertaining to six groups: complete blood count, viral panel, biochemicals, venous blood gas, arterial blood gas and urine test.

The investigations focused on two aspects relevant to the study. In the first step, we investigated the importance of individual examinations for the diagnosis of

COVID-19. In this step, the selection techniques Relief-F and Decision Tree were used to select the variables with the highest capacity to separate the classes

"negative" and "positive". The results obtained in both methods confirm the findings of research conducted over the years 2020 and 2021, which investigated the effects of COVID-19 on laboratory biomarker values. The results pointed out that the main biomarkers for predicting COVID-19 evaluated belonged mostly to the CBC and Biochemical groups.

When comparing the results of the two selection techniques with the reference works, we note the importance of the biomarkers of blood count, especially those belonging to the Leukocyte family, which had 4 markers (leukocytes, eosinophils, monocytes, and neutrophils) present among the 10 most relevant for both methods of variable selection. For the Biochemical markers group, sodium obtained the best ranking position (sixth most important variable by Relief-F and seventh by the Decision Tree), which is in agreement with the studies of [22], which indicates that sodium can be used for risk identification in patients with COVID-19. Another highlight among the descriptors in the biochemical group was C-reactive protein, which is in agreement with research by [23], [5], [4].

In the second step ensembles (Bagging, Boosting and Stacking) of individual classifiers were built. The models were obtained for different configurations of the classifiers and different amounts of selected variables. The best results were achieved with Stacking with the top 20 descriptors selected by the Decision tree, and the following results were obtained: Accuracy = 0.9778; Sensitivity=0.9527 and Specificity=1.000. The results demonstrate the feasibility of using Machine Learning techniques to diagnose COVID-19 from a reduced set of conventional laboratory tests, which can reduce the cost of diagnosing the disease for patients who need hospital admission, or who have required outpatient care.

THANKS

The authors thank the Cooperac¸a˜o CAPES/FAPES - PDPG, TIC+TAC project, for the financial support (TO 133/2021, Process No. 2021-CFT5C).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Guo Y. R., Cao Q. D., Hong Z. S. et al. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (covid-19) outbreak.

- an update on the status. Military Medical Research, v.7(1):11.

2. CRC (2022). Coronavirus Resource Center: Covid-19 dashboard by the center for systemsscience and engineering csse at johns hopkins university jhu. Accessed: 2022-28-05. Available at: https://coronavirus.jhu.edu/map.html.

3. Li T., Qi W., Duanyang Z., et al (2020). Lymphopenia predicts disease severity of covid-19: a descriptive andpredictive study.Signal Trans- duction and Targeted Therapy, v.5(33).

4. Zhao,K., Li, R., Wu, X. et al. (2020) Clinical features in 52 patients with covid-19 who have increased leukocyte count: a retrospective analysis. European Journal of Clinical Microbiology Infectious Dis- eases, v.39(12).

5. Gao, Y., Li T., Han M. et al (2020). Diagnostic utility of clinical labo- ratory data determinations for patients with the severe covid-19. J Med Virol, v.92(7):791-796.

6. Gunc˘ar, G., Kukar, M., Notar, M. et al. (2018). An application of ma- chine learning to haematological diagnosis. Scientific reports, 8(1), 411.

7. Vogado, L. H.S., Veras, R. M.S., Arau'jo, F. H.D. et al. (2018) Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification. Eng. Appl. Artif. Intell., Pergamon Press,Inc., USA, v. 72, n. C, p. 415-422.

8. Luo, Y., Szolovits, P., Dighe, A. S. et al. Using Machine Learning to Predict Laboratory Test Results. American Journal of Clinical Pathol- ogy, v. 145, n. 6, p. 778-788, 06 2016. ISSN 0002-9173.

9. Batista, A. F. de M., Miraglia, J. L., Donato, T. H. R. et al. (2020). Covid-19 diagnosis prediction in emergency care patients: a machine-learning approach. medRxiv, Press.

10. Wynants, L., Calster, B. V., Collins, G. et al. (2021) Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. BMJ, v. 369, p. m1328.

11. Sun, L., Song, F., Shi, N. et al. (2020) Combination of four clinical in- dicators predicts the severe/critical symptom of patients infected covid-

19. Journal of Clinical Virology, v. 128, p. 104431, ISSN 1386-6532.

12. Banerjee, A., Ray, S., Vorselaars, B. et al. (2020). Use of machine learn- ing and artificial intelligence to predict sars-cov-2 infection from full blood counts in a population.International Immunopharmacology, El- sevier, v.86:106705.

13. Kira, K.; Rendell, L. A. (1992). The feature selection problem: Tra- ditional methods and a new algorithm. In: AAAI-92 Proceedings, pp. 129-134.

14. Sikonja, R. M.; Kononenko, I. (2003). Theoretical and empirical anal- ysis of ReliefF and RReliefF. Machine Learning Journal (2003) v.53, pp.23-69.

15. Jovic, A.; Brkic, K.; Bogunovic, N., (2015). A review of feature selec- tion methods with applications. In:2015 38th International Convention on Information and Communication Technology, Electronics and Mi- croelectronics (MIPRO). [S.l.: s.n.], pp. 1200-1205.

16. Hartshorn, S., (2016). Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners. [S.l.: s.n.].

17. Breiman, L. (1996). Bagging predictors. Mach Learn, v.24, pp.123-140.

18. Asmita, S. and Shukla, K. k. (2014). Review on the Architecture, Algo- rithm and Fusion Strategies in Ensemble Learning. International Jour- nal of Computer Applications, v.108(8), pp:21-28.

19. Patel, A. Stacking -Ensemble meta Algorithms for improve predic- tions -¡https://medium.com/ml-research-lab/stacking-ensemble-meta- algorithms-for-improve-predictions-f4b4cf3b9237¿ Accessed May 28, 2022.

20. Severson, K.; Molaro, M.; Braatz, R. (2017) Principal component anal- ysis of process datasets with missing values. Processes, v.5, pp:38.

21. Lara, D. F.; Pozo, A. T. R.; Garcia, L. M. L. d. S. (2013). Empirical studies of balancing methods for classification. In: ASAI. ASAI, Simposio Argentino de Inteligencia Artificial. [S.l.], 2013. pp:955-960.

22. Tzoulis, P.; Waung, J. A.; Bagkeris, E. et al. (2021). Dysnatremia is a predictor for morbidity and mortality in hospitalized patients with covid-19. The Journal of Clinical Endocrinology Metabolism, Oxford, v.106, pp: 1637-1648.

23. Seyit, M.; Avci E.; Nar, R. et al. (2021). Neutrophil to lymphocyte ratio, lymphocyte to monocyte ratio and platelet to lymphocyte ratio to predict the severity of covid-19. American Journal of Emergency Medicine, Elsevier, v.40, pp: 110-114.

24. Narin, A; Kaya, C.; Pamuk, Z. (2020). Automatic detection of coron- avirus disease (COVID-19) using X-ray images and deep convolutional neural networks. Pattern Analysis and Applications, Springer Science and Business Media LLC, v.24, 3, pp: 1207-1220.

25. Sethy, P. K.; Behera; S. K.; Ratha; P. K.; Biswas; P. (2020). Detection of Coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine. International Journal of Mathematical, Engineering and Management Sciences, v. 5, 4, pp: 643-651.

26. de Freitas Barbosa, V.A., Gomes, J.C., de Santana, M.A. et al., (2022). Heg.IA: an intelligent system to support diagnosis of Covid-

19 based on blood tests. Res. Biomed. Eng. 38, 99-116 (2022). https://doi.org/10.1007/s42600-020-00112-5