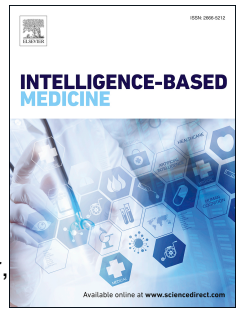

A super learner ensemble of 14 statistical learning models for predicting COVID-19 severity among patients with cardiovascular conditions

Louis Ehwerhemuepha, Sidy Danioko, Shiva Verma, Rachel Marano, William Feaster, Sharief Taraman, Tatiana Moreno, Jianwei Zheng, Ehsan Yaghmaei, Anthony Chang

PII: S2666-5212(21)00006-5

DOI: https://doi.org/10.1016/j.ibmed.2021.100030

Reference: IBMED 100030

To appear in: *Intelligence-Based Medicine*

Received Date: 1 December 2020

Revised Date: 1 March 2021

Accepted Date: 12 March 2021

Please cite this article as: Ehwerhemuepha L, Danioko S, Verma S, Marano R, Feaster W, Taraman S, Moreno T, Zheng J, Yaghmaei E, Chang A, A super learner ensemble of 14 statistical learning models for predicting COVID-19 severity among patients with cardiovascular conditions, *Intelligence-Based Medicine*, https://doi.org/10.1016/j.ibmed.2021.100030.

A super learner ensemble of 14 statistical learning models for predicting COVID-19 severity among patients with cardiovascular conditions

Louis Ehwerhemuepha[1,2], Sidy Danioko[2], Shiva Verma[3], Rachel Marano[1], William Feaster[1], Sharief Taraman[1], Tatiana Moreno[1], Jianwei Zheng[2], Ehsan Yaghmaei[1,2], Anthony Chang[1]

Affiliation:

1. Children's Hospital of Orange County, Orange, CA 92868
2. Schmid College of Science, Chapman University, Orange, CA 92866
3. Department of Computing, Data Science, and Society, University of California, Berkeley, Berkeley, CA 94720

Corresponding author: Louis Ehwerhemuepha, 1201 W La Veta Ave, Orange, CA 92868. Phone: 714-509-3146; Fax: 714-509-7653; email: lehwerhemuepha@choc.org

Financial support: There were no financial support or assistance associated with this study

Disclosure: None of the authors have financial ties to products in the study. There are no potential/perceived conflicts of interest.

Other authors information: Sidy Danioko: danio100@mail.chapman.edu; Shiva Verma: shiva.verma@berkeley.edu; Rachel Marano: rachel.marano@choc.org; William Feaster: wfeaster@choc.org; Sharief Taraman: staraman@choc.org; Tatiana Moreno: tmoreno@choc.org; Jianwei Zheng: zheng120@mail.chapman.edu; Ehsan Yaghmaei: yaghmaei@chapman.edu; Anthony Chang: achang@aol.com

Abstract

Background: Cardiovascular and other circulatory system diseases have been implicated in the severity of COVID-19 in adults. This study provides a super learner ensemble of models for predicting COVID-19 severity among these patients.

Method: The Cerner Real-World Database was used for this study. Data on adult patients (18 years or older) with cardiovascular and related circulatory diseases between 2017 and 2019 were retrieved and a total of 13 these conditions were identified. Among these patients, 33,042 admitted with positive diagnoses for COVID-19 between March 2020 and June 2020 (from 59 hospitals) were identified and selected for this study. A total of 14 statistical and machine learning models were developed and combined into a single more powerful super learning model for predicting COVID-19 severity on admission to the hospital.

Result: LASSO regression, a full extreme gradient boosting model with tree depth of 2, and a full logistic regression model were the most predictive with cross-validated AUROCs of 0.7964, 0.7961, and 0.7958 respectively. The resulting super learner ensemble model had a cross validated AUROC of 0.8006 (range: 0.7814, 0.8163). The unbiased AUROC of the super learner model on an independent test set was 0.8057 (95% CI: 0.7954, 0.8159).

Conclusion: Highly predictive models can be built to predict COVID-19 severity of patients with cardiovascular and other circulatory conditions. Super learning ensembles will improve individual and classical ensemble models significantly.

Introduction

The novel coronavirus disease, COVID-19, which was first reported in December 2019 in Wuhan, China, is caused by severe acute respiratory syndrome coronavirus 2, SARS-CoV-2. The virus has spread to 191 out of 195 countries with more than 63 million global cases and 1.47 million global deaths as of November 30, 2020.[1,2] The World Health Organization declared COVID-19 a global pandemic on March 11th, 2020 as the number of countries affected rose sharply from 59 on February 28th, 2020 to 122 on March 13th, 2020.[1,2]

Underlying cardiovascular and circulatory diseases have been implicated in the severity of COVID-19 in adults[3–11] since March 2020. The association between cardiovascular diseases (CVD) and COVID-19 severity is bidirectional. On the one hand, pre-existing CVD such as coronary heart disease and hypertension are known to be linked with higher COVID-19 morbidity and mortality. On the other hand, COVID-19 can induce CVD such as myocardial injury, arrhythmia, acute coronary syndrome, and venous thromboembolism among others.[7–11] In other words, while pre-existing CVD can lead to worse COVID-19 outcomes, COVID-19 can induce new CVD and potentially worsen existing disease.[7–11] Recent studies have addressed cardiovascular risk factors of COVID-19 implicating cardiovascular complications with greater COVID-19 disease burden.[12,13] This underscores the importance of studying the relationship between CVD and related circulatory conditions with respect to COVID-19 severity. Specific focus on CVD patients is therefore required given the elevated mortality rate among these patients with COVID-19. Corresponding severity prediction model for CVD patients on admission to the hospital will help with proactive care and reduce morbidity and mortality.

The application of statistical learning and artificial intelligence algorithms may provide frontline clinicians ability to provide early and targeted therapies that may help reduce morbidity.[14–19]

Furthermore, the ability to recognize, on admission, patients who will progress to severe COVID-19 would be helpful in logistics and planning in face of scarce clinical resources and has the potential to be life-saving. Consequently, the application of predictive models may help mitigate some uncertainty associated with COVID-19 disease progression.

In this study, we developed 14 statistical learning models and combined them into a super learning model that is an ensemble of ensembles and other statistical/machine learning models. The goal is to assess the extent by which these models may help predict severe COVID-19 in CVD patients who are already known to be at high risk.

Method

The Cerner Real-World Database (CRWD) was used under the Institutional Review Board of the corresponding author's institution with IRB number 2008107. The CRWD is a deidentified electronic health records database of more than 90 health systems that are either clients of Cerner® Corporation or users of a Cerner-proprietary application called HealtheIntent.[20] These health systems have agreed to share structured tabular clinical data in deidentified format and in return have access to the deidentified data of all contributing health system. A subset of the CRWD was identified by Cerner as patients who had positive labs or diagnoses for COVID-19 to help foster corresponding studies on the disease.

Patient qualification for this study was built on two inclusion/exclusion criteria. First, patients must have been admitted for COVID-19 between March 2020 and June 2020. Second, the patients must have had a CVD or related circulatory system diagnosis between 2017 and 2019. The choice of considering diagnoses between 2017 and 2019 is to ensure that only pre-existing CVD conditions not related to the emergence of the COVID-19 pandemic were considered. Consequently, qualifying patients are patients who had a history of or pre-existing CVD conditions and who were hospitalized with COVID-19 diagnosis between March 2020 and June 2020. The history or pre-existing diagnoses of CVD and other circulatory conditions considered were determined by a cardiologist and a hospitalist in the study team using the International Classification of Disease, Version 10, Clinical Modification (ICD-10-CM) codes I10 to I95. In a similar way, we considered pre-COVID-19 histories of other comorbid conditions by major body systems such as conditions affecting the digestive, nervous, and respiratory systems. A full list of all conditions and corresponding diagnosis codes are shown in the Summary Statistics in Table 1.

Demographics and health insurance payer data were retrieved for qualifying patients. The vital signs (such as body temperature, heart rate, respiratory rate, systolic blood pressure, and diastolic blood pressure) of patients on admission to the hospital with COVID-19 were captured and categorized into normal, high, and low for the age of the patient. The oxygen saturation level was also captured and categorized into the following categories: 100-95%, 94-90%, and <90%. A nuisance categorical level was created for patients with vital signs that were not measured on admission or with missing vital sign data in the database. Alternative approaches would include the use of statistical or machine learning imputation methods.

COVID-19 severity can be measured by several clinical indicators of clinical decompensation. Two of the most severe forms of decompensation are need for mechanical ventilation and in-hospital death. In this study, patients who were on mechanical ventilators or who had in-hospital death were classified as patients who progressed to severe COVID-19. All other patients were classified as having mild COVID-19. As a result, the outcome variable of this study is binary: severe COVID-19 (need for mechanical ventilators or in-hospital death) and mild COVID-19 (any other outcome with live discharge from the hospital). This binary outcome was chosen to simplify this multicenter study and to ensure that we are targeting the most severe outcomes for COVID-19.

A total of 14 statistical learning models (referred to as base learners from hereon) were selected for this study that encompassed LASSO regression, generalized logistic regression model (with and without forward variable selection), linear discriminant analysis (with and without LASSO variable selection), multivariate adaptive regression splines, random forest (with and without LASSO variable selection), and three extreme gradient boosting models (all with and without LASSO variable selection).[21–26] Cross-validated area under the receiver operator characteristic

curves (AUROCs) were used to estimate the performances of the base learners as well as the Super Learner model consisting of predictions from all 14 base learners. Using oracle inequalities for multi-fold cross validation,[27] the Super Learner was mathematically proven to result in better model performance than each of the base learners. Interested readers can refer to the appendix of van der Laan et al. 2007 for full exposition of the mathematical details of the proof.[28] We include a mathematical derivation of super learning in the appendix of this paper and provide a simplified graphical representation in Figure 1 here.

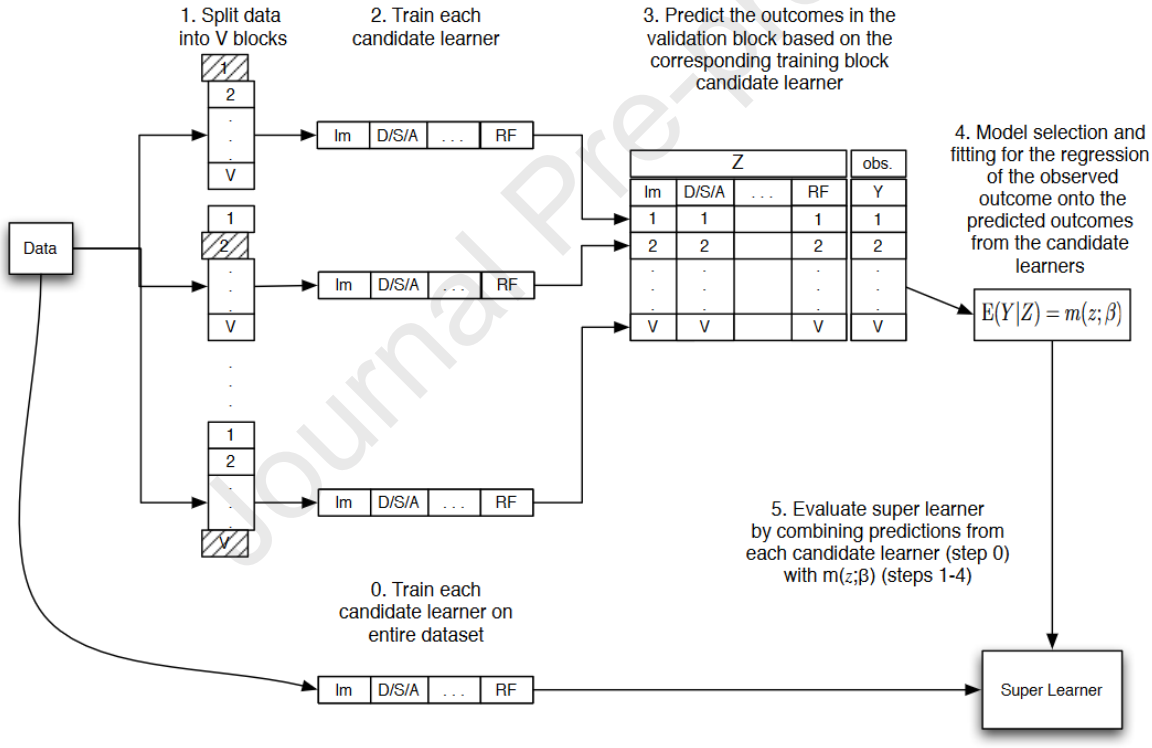

Figure 1. Visual description of super learning by van der Laan and Rose (2011, 2018).

The data for this study consist of variables capturing demographics, health insurance information, first vital signs on admission, 13 pre-existing CVD and related circulatory conditions, and pre-existing comorbid conditions. This data was split into training (70%) and test

(30%) sets. The training set was used to train all base learners and the super learner model using 10-fold cross-validation. The test set was used to provide unbiased estimates of the super learner ensemble model performance metrics. All analyses for this study were carried out in the Statistical Computing Programming Language R and the SuperLearner package.[29,30]

Results

The data used for this study consists of COVID-19 hospitalizations from 59 hospitals/health systems. There was a total of 33,042 qualifying hospitalizations of which 5,685 had mechanical ventilators or resulted in an in-hospital death. This results in a severe COVID-19 rate of 17.2%. There were 49.0% female patients, 43.9% male patients, and 7.1% of patients with unknown sex. Young adult patients (between 18 and 35 years), middle-aged adults (between 36 and 55 years), and older adults (greater than 55 years) consisted of 6.2%, 22.8%, and 71.0% of all hospitalizations indicating that these hospitalizations were skewed towards older adults. In addition, the demographics of the data indicate a skew towards White patients with 63.3% of hospitalizations. Black or African Americans, Asian or Pacific Islanders, American Indian or Alaska Natives, and patients of other racial groups consisted of 21.5, 2.5, 1.4, and 8.7% of hospitalizations. Over 50.2% of all patients were on governmental healthcare insurance plans, 30.3% on private insurance, 2.9% were self-pay, and 16.5% of other/unknown payer type.

[Table 1 here]

Univariable analyses of association between severe COVID-19 and each variable are shown in the Summary Statistics in Table 1. The analyses indicated that there were univariable associations between all variables considered and severity of COVID-19 in CVD patients except for pre-existing/history of pericarditis and digestive system comorbidities. This finding is in line with findings and studies on the impact of pre-existing comorbidities on the risk of severe COVID-19. Conclusive tests of association and causal analyses with corresponding effect sizes in multivariable statistical analyses are beyond the scope of this study.

The cross-validated model performance on the training data are shown in Table 2 in order of decreasing performance. The super learner model had a cross-validated average AUROC of 0.8006 which, as expected, is higher than those of the constituent base learners.

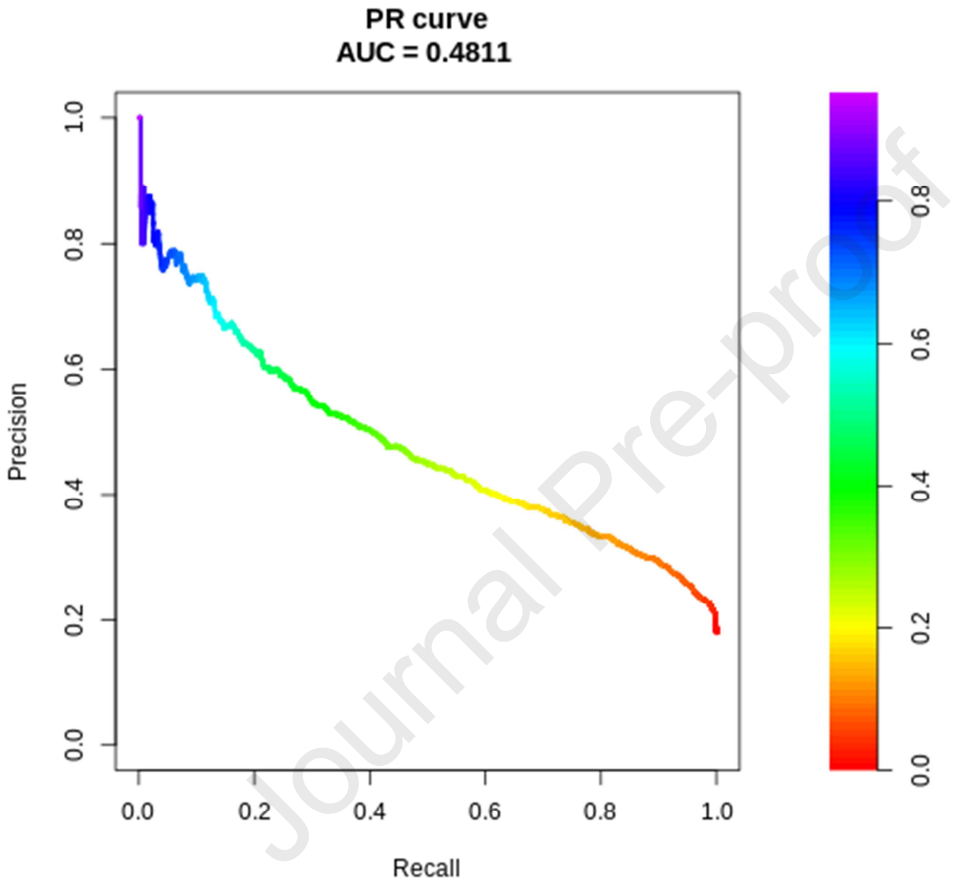

Figure 2. The precision-recall curve for the Super Learner model

The cross validated AUROCs on the training dataset is not an unbiased estimate of the super learner performance. So, the AUROC of the Super Learner Ensemble for predicting severe COVID-19 disease on patients with pre-existing CVD was estimated on the independent test set. The unbiased AUROC of the Super Learner model was 0.8057 (95% CI: 0.7954, 0.8159). At a model specificity value of 70% the sensitivity of the model was 75.2 (73.2, 77.2); the positive

predictive value was 35.4 (33.9, 36.9); negative predictive value was 92.8 (92.2, 93.5); and an F-1 score of 0.481. The area under the precision-recall curve is shown in Figure 2.

There is a concept of the number needed to treat/evaluate in clinical research involving randomized control trials. In machine learning parlance and in the case of this study, it is the average number of patients a classifier will flag to guarantee a true positive prediction. Mathematically, it is also the inverse of the positive predictive value. Consequently, the number needed to evaluate (NNE) for this model is 3 (rounded up from 2.8). That is, there will be 2 false positive prediction for every true positive prediction from the super learner.

Discussion

The base learners for predicting severe COVID-19 disease among patients with pre-existing CVD diagnoses had cross-validated training AUROCs that may result in good model performances if used individually. But the super learner model was able to take advantage of all 14 base learners and result in model performance estimates higher than those of the base learners. In other words, the performance of the resulting super learner model was in line with mathematically derived proofs indicating that the cross-validated AUROCs of any super learner will be greater than those of each individual base learners. Consequently, ranking of the base learners in comparison to the super learner is not important and, complexity does not necessary imply greater performance given the performance of the logistic regression model in relation to the more complex base learners. The goal of using super learning is not to exhaustively compare these models but to combine the strength they provide in an ensemble that is proven to result in better performance.

It is difficult to gauge the clinical value of a predictive model using the most common model performance statistics as they are usually dependent on properties of the data such as the rarity of the outcome variable without taking into consideration the true underlying cost of false positive/negative predictions. A positive predictive value of 35.3% indicate that clinicians can be certain that, on the average, 1 of 3 patients (the NNE) with pre-existing CVD that is flagged by the super learner model on admission will indeed progress to having severe COVID-19. The performance metrics quoted in the results section also indicate that only 38.1% of patients with pre-existing condition will be flagged which will consist of over 75% of patients with severe COVID-19. These numbers indicate that the super learner provides deployable levels of performance to potential afford clinical significance and improved quality of care when coupled

with appropriate clinical intervention protocols. The most likely clinical setting for the application of these type of model is in population health management where much more than age is considered in analyzing the risk COVID-19 poses to the patients within a health system. Population health initiatives aimed at improving recommended practices for reducing the risk of severe COVID-19 (such as vaccination as soon as it becomes available to high risk patients in addition to more pressing need for social distancing measures) can be targeted at the most at risk patients within the health system. These results are likely to generalize to any hospital in the US given that the data used to train the model consists of patients from 59 hospitals/health systems in the US.

Machine learning, ensemble, and super learning models suffer from the inability to provide statistically sound inference on predictors unlike advanced statistical/biostatistical models. While variable importance measures may help rank variables by how they contribute to predictive values, such measures of importance are limited in cases where little is known on underlying associations or causal factors. Shapley additive estimates[31] are a great improvement but are still limited compared to appropriate regression estimates of various risk metrics. As a result, ensemble and super learning models should be used in tandem with rigorous and advanced statistical models for discovery of significant associations and causal inference on variables that may help in the development of effective clinical intervention protocols. Additional studies addressing both associations and causality are currently being worked on as follow-up to this study.

In conclusion, the performance of base learners presented in this study show promise for the application of statistical and machine learning algorithms for predicting COVID-19 severity for CVD patients and for the general population as well. However, the AUROC of machine learning

models are very difficult to improve on which implies that relatively small improvements are welcomed. Such improvements may translate to significant clinical impact in cases of scare resources for intervention and hospital resources have become scarcer due to the pandemic. Any improvement in a prediction task is therefore critical especially when constrained with very scarce clinical intervention resources. As a result, super learners such as the one developed in this study provide significant opportunity. The super learner model developed herein could help reduce the morbidity and mortality of CVD patients hospitalized with COVID-19 through appropriate clinical intervention and improved logistics based on predicted usage of intensive care units critical to the survival of patients with severe COVID-19. This is the first study on a model developed to predict cardiovascular predisposition to COVID-19. This multicenter study could therefore serve to frame future analyses both as a primary source and moving forward as a comparison to others.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*. 2020;20(5):533-534.
2. Parker E. Covid 2019 tracker. Published online 2020.
3. Zheng Y-Y, Ma Y-T, Zhang J-Y, Xie X. COVID-19 and the cardiovascular system. *Nat Rev Cardiol*. 2020;17(5):259-260.
4. Guzik TJ, Mohiddin SA, Dimarco A, et al. COVID-19 and the cardiovascular system: implications for risk assessment, diagnosis, and treatment options. *Cardiovasc Res*. Published online 2020.
5. Veer M, Kumar AM, Ivanova V. COVID-19 and the Cardiovascular System. *Crit Care Nurs Q*. 2020;43(4):381-389.
6. Azevedo RB, Botelho BG, de Hollanda JVG, et al. Covid-19 and the cardiovascular system: a comprehensive review. *J Hum Hypertens*. Published online 2020:1-8.
7. Clerkin KJ, Fried JA, Raikhelkar J, et al. COVID-19 and cardiovascular disease. *Circulation*. 2020;141(20):1648-1655.
8. Nishiga M, Wang DW, Han Y, Lewis DB, Wu JC. COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives. *Nat Rev Cardiol*. 2020;17(9):543-558.
9. Bansal M. Cardiovascular disease and COVID-19. *Diabetes Metab Syndr Clin Res Rev*. Published online 2020.
10. Ganatra S, Hammond SP, Nohria A. The novel coronavirus disease (COVID-19) threat for patients with cardiovascular disease and cancer. Published online 2020.
11. Guo J, Huang Z, Lin L, Lv J. Coronavirus disease 2019 (covid-19) and cardiovascular disease: a viewpoint on the potential influence of angiotensin-converting enzyme inhibitors/angiotensin receptor blockers on onset and severity of severe acute respiratory syndrome coronavirus 2 infec. *J Am Heart Assoc*. 2020;9(7):e016219.
12. Di Castelnuovo A, Bonaccio M, Costanzo S, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. *Nutr Metab Cardiovasc Dis*. 2020;30(11):1899-1913.
13. Sabatino J, De Rosa S, Di Salvo G, Indolfi C. Impact of cardiovascular risk profile on COVID-19 outcome. A meta-analysis. *PLoS One*. 2020;15(8):e0237131.
14. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*. Published online 2020:110059.
15. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Published online 2020.
16. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning

method for image-based diagnosis of COVID-19. *PLoS One*. 2020;15(6):e0235187.

17. Li L, Qin L, Xu Z, et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*. Published online 2020.

18. O'Connell R, Feaster W, Wang V, Taraman S, Ehwerhemuepha L. Predictors of pediatric readmissions among patients with neurological conditions. *BMC Neurol*. 2021;21(1):1-8.

19. Ehwerhemuepha L, Pugh K, Grant A, et al. A Statistical Learning Model for Unplanned 7-day Readmission in Pediatrics. *Hosp Pediatr*. 2020;10(1):43-51.

20. Ehwerhemuepha L, Gasperino G, Bischoff N, Taraman S, Chang A, Feaster W. HealtheDataLab - a cloud computing solution for data science and advanced analytics in healthcare with application to predicting multi-center pediatric readmissions. *BMC Med Inform Decis Mak*. 2020;20(1):1--12. doi:https://doi.org/10.1186/s12911-020-01 153-7

21. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Second Edi. Springer; 2009.

22. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning.*; 2013.

23. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. *Logistic Regression*. Springer; 2002.

24. Friedman JH. Multivariate adaptive regression splines. *Ann Stat*. Published online 1991:1-67.

25. Friedman JH, Roosen CB. An introduction to multivariate adaptive regression splines. Published online 1995.

26. Chen T, He T, Benesty M, et al. xgboost: Extreme Gradient Boosting. Published online 2019.

27. der Vaart AW, Dudoit S, van der Laan MJ. Oracle inequalities for multi-fold cross validation. *Stat Decis*. 2006;24(3):351-371.

28. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1). doi:https://doi.org/10.2202/1544-6115.1309

29. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2020. https://www.r-project.org/

30. Polley E, LeDell E, Kennedy C, van der Laan M. SuperLearner: Super Learner Prediction. Published online 2019. https://cran.r-project.org/package=SuperLearner

31. Shapley LS. A value for n-person games. *Contrib to Theory Games*. 1953;2(28):307-317.

Table 1. Summary statistics on all variables

Variables	Levels	COVID-19 Infection, n (%) Mild	Severe	Unadjusted p values
Sex	Female	9668 (50.30)	1625 (41.59)	< 0.001
	Male	8342 (43.40)	1865 (47.73)	
	Unknown	1212 (6.31)	417 (10.67)	
Age	Young Adults (18 to 35yrs)	1344 (6.99)	95 (2.43)	
	Middle-Aged Adults (36 to 55yrs)	4776 (24.85)	529 (13.54)	
	Older Adults (> 55yrs)	13102 (68.16)	3283 (84.03)	
Race	White	12341 (64.20)	2254 (57.69)	
	Black or African American	4052 (21.08)	956 (24.47)	
	Asian or Pacific islander	446 (2.32)	111 (2.84)	
	American Indian or Alaska Native	264 (1.37)	51 (1.31)	
	Other racial group	1621 (8.43)	420 (10.75)	
	Unknown racial group	498 (2.59)	115 (2.94)	
Payer	Governmental Insurance	9445 (49.14)	2289 (58.59)	
	Private Insurance	6218 (32.35)	716 (18.33)	
	Self-pay	626 (3.26)	41 (1.05)	
	Unknown	2933 (15.26)	861 (22.04)	
Vital signs on admission				
Temperature	Normal	11898 (61.90)	2392 (61.22)	< 0.001
	High	2440 (12.69)	886 (22.68)	
	Low	117 (0.61)	126 (3.22)	
	Unknown	4767 (24.80)	503 (12.87)	
Heart rate	Normal	8396 (43.68)	2000 (51.19)	
	High	2743 (14.27)	1246 (31.89)	
	Low	470 (2.45)	121 (3.10)	
	Unknown	7613 (39.61)	540 (13.82)	
Respiratory rate	Normal	12864 (66.92)	1556 (39.83)	
	High	3332 (17.33)	1903 (48.71)	
	Low	27 (0.14)	37 (0.95)	
	Unknown	2999 (15.60)	411 (10.52)	
Systolic blood pressure	Normal	4456 (23.18)	954 (24.42)	
	High	9926 (51.64)	1753 (44.87)	
	Low	1910 (9.94)	768 (19.66)	
	Unknown	2930 (15.24)	432 (11.06)	
Diastolic blood pressure	Normal	4642 (24.15)	770 (19.71)	
	High	4927 (25.63)	808 (20.68)	

	Low	6722 (34.97)	1897 (48.55)	
	Unknown	2931 (15.25)	432 (11.06)	
Oxygen saturation	100 - 95%	13219 (68.77)	1819 (46.56)	
	94 - 90%	2516 (13.09)	864 (22.11)	
	< 90%	699 (3.64)	820 (20.99)	
	Unknown	2788 (14.50)	404 (10.34)	
Pre-existing cardiovascular and related circulatory conditions				
Hypertensive heart diseases (I10-I16)	No	2774 (14.43)	426 (10.90)	
	Yes	16448 (85.57)	3481 (89.10)	
Ischemic heart diseases (I20-I25)	No	13742 (71.49)	2438 (62.40)	< 0.001
	Yes	5480 (28.51)	1469 (37.60)	
Pulmonary heart diseases (I26-I27)	No	17676 (91.96)	3448 (88.25)	
	Yes	1546 (8.04)	459 (11.75)	
Pericarditis (I30-I32)	No	18851 (98.07)	3833 (98.11)	0.932
	Yes	371 (1.93)	74 (1.89)	
Endocarditis and heart valves disorders (I33-I39)	No	17343 (90.22)	3412 (87.33)	
	Yes	1879 (9.78)	495 (12.67)	
Cardiomyopathy (I42-I43)	No	18043 (93.87)	3536 (90.50)	< 0.001
	Yes	1179 (6.13)	371 (9.50)	
Atrioventricular and other conduction disorders (I44-I45)	No	17593 (91.53)	3478 (89.02)	
	Yes	1629 (8.47)	429 (10.98)	
Cardiac arrest (I46)	No	19143 (99.59)	3880 (99.31)	0.026
	Yes	79 (0.41)	27 (0.69)	
Arrythmias (I47-I49)	No	15167 (78.90)	2776 (71.05)	
	Yes	4055 (21.10)	1131 (28.95)	
Heart failure (I50)	No	15599 (81.15)	2624 (67.16)	
	Yes	3623 (18.85)	1283 (32.84)	< 0.001
Cerebrovascular disorders (I60-I69)	No	16782 (87.31)	3234 (82.77)	
	Yes	2440 (12.69)	673 (17.23)	
Disorders of the arteries, arterioles, and capillaries (I70)	No	16244 (84.51)	3093 (79.17)	
	Yes	2978 (15.49)	814 (20.83)	
Disorders of the veins and lymphatic vessels/nodes (I80)	No	16675 (86.75)	3316 (84.87)	0.002
	Yes	2547 (13.25)	591 (15.13)	
Hypotension (I95)	No	17085 (88.88)	3359 (85.97)	< 0.001
	Yes	2137 (11.12)	548 (14.03)	
Pre-existing comorbid conditions				
Infectious and parasitic diseases (A00-B99)	No	12836 (66.78)	2428 (62.14)	
	Yes	6386 (33.22)	1479 (37.86)	
Malignant neoplasms (C00-C96)	No	17086 (88.89)	3390 (86.77)	
	Yes	2136 (11.11)	517 (13.23)	< 0.001
Endocrine, nutritional, and metabolic diseases (E00-E89)	No	3308 (17.21)	440 (11.26)	
	Yes	15914 (82.79)	3467 (88.74)	
Mental, behavioral, and	No	9645 (50.18)	1833 (46.92)	

neurodevelopmental disorders (F01-F99)	Yes	9577 (49.82)	2074 (53.08)	
Diseases of the nervous system (G00-G99)	No	10163 (52.87)	1821 (46.61)	
	Yes	9059 (47.13)	2086 (53.39)	
Diseases of the respiratory system (J00-J99)	No	8405 (43.73)	1607 (41.13)	0.003
	Yes	10817 (56.27)	2300 (58.87)	
Diseases of the digestive system (K00-K95)	No	8202 (42.67)	1669 (42.72)	0.970
	Yes	11020 (57.33)	2238 (57.28)	
Diseases of the skin and subcutaneous tissue (L00-L99)	No	13573 (70.61)	2683 (68.67)	0.016
	Yes	5649 (29.39)	1224 (31.33)	
Diseases of the musculoskeletal system and connective tissue (M00-M99)	No	6390 (33.24)	1408 (36.04)	< 0.001
	Yes	12832 (66.76)	2499 (63.96)	
Diseases of the genitourinary system (N00-N99)	No	8091 (42.09)	1350 (34.55)	
	Yes	11131 (57.91)	2557 (65.45)	

Table 2. Cross-validated (training) AUROC

Algorithm	Cross-validated (training) AUROC		
	Average	Minimum	Maximum
Super Learner	0.8006	0.7814	0.8163
Lasso regression	0.7964	0.7759	0.8143
Extreme gradient boosting, max. tree depth of 2 (all variables)	0.7961	0.7774	0.8136
Logistic regression (all variables)	0.7958	0.7755	0.8137
Logistic regression (forward variable selection)	0.7957	0.7764	0.8147
Extreme gradient boosting, max. tree depth of 2 (LASSO variable selection)	0.7956	0.7746	0.8131
Linear discriminant analysis (LASSO variable selection)	0.7948	0.7718	0.8110
Linear discriminant analysis (all variables)	0.7947	0.7713	0.8107
Multivariate adaptive regression splines	0.7906	0.7733	0.8105
Random forest (all variables)	0.7869	0.7761	0.7981
Random forest (LASSO variable selection)	0.7845	0.7709	0.7974
Extreme gradient boosting, max. tree depth of 4 (LASSO variable selection)	0.7817	0.7680	0.7963
Extreme gradient boosting, max. tree depth of 4 (all variables)	0.7804	0.7708	0.7963
Extreme gradient boosting, max. tree depth of 6 (all variables)	0.7668	0.7488	0.7820
Extreme gradient boosting, max. tree depth of 6 (LASSO variable selection)	0.7663	0.7581	0.7787

Table 3. Super learner weights (on base learners)

	Super Learner Weight	
Base learners	**Mean**	**SD**
Multivariate adaptive regression splines	0.203	0.048
Extreme gradient boosting, max. tree depth of 2 (all variables)	0.145	0.036
Extreme gradient boosting, max. tree depth of 2 (LASSO variable selection)	0.131	0.030
Linear discriminant analysis (LASSO variable selection)	0.110	0.020
Linear discriminant analysis (all variables)	0.106	0.018
Random forest (LASSO variable selection)	0.070	0.056
Random forest (all variables)	0.065	0.050
Logistic regression (forward variable selection)	0.054	0.035
LASSO regression	0.031	0.022
Extreme gradient boosting, max. tree depth of 6 (all variables)	0.024	0.018
Extreme gradient boosting, max. tree depth of 4 (LASSO variable selection)	0.023	0.030
Logistic regression (all variables)	0.019	0.017
Extreme gradient boosting, max. tree depth of 6 (LASSO variable selection)	0.012	0.020
Extreme gradient boosting, max. tree depth of 4 (LASSO variable selection)	0.007	0.015

Highlights

1. Patients with cardiovascular diseases are at high risk of severe COVID-19
2. Machine learning and artificial intelligence may help in identification of the highest risk patients
3. Individual base learners (statistical learning models) have good performance in predict the severity of COVID-19 in patients with pre-existing cardiovascular diseases
4. Super learning, a mathematically proven approach to ensemble learning, will further improve the performance of base learners
5. A super learner model for patients with pre-existing cardiovascular conditions resulted in improved model performance that may have meaningful clinical impact

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: