

Effectiveness, Explainability and Reliability of Machine Meta-Learning Methods for Predicting Mortality in Patients with COVID-19: Results of the Brazilian COVID-19 Registry

Polianna Delfino-Pereira

Instituto de Avaliação de Tecnologia em Saúde

Cláudio Moisés Valiense De Andrade

Universidade Federal de Minas Gerais

Virginia Mara Reis Gomes

Centro Universitário de Belo Horizonte

Maria Clara Pontello Barbosa Lima

Universidade Federal de Ouro Preto

Maira Viana Rego Souza-Silva

Universidade Federal de Minas Gerais

Marcelo Carneiro

Hospital Santa Cruz

Karina Paula Medeiros Prado Martins

Universidade Federal de Minas Gerais

Thaís Lorena Souza Sales

Federal University of São João del-Rei

Rafael Lima Rodrigues De Carvalho

Instituto de Avaliação de Tecnologia em Saúde

Magda Carvalho Pires

Universidade Federal de Minas Gerais

Lucas Emanuel Ferreira Ramos

Universidade Federal de Minas Gerais

Rafael T Silva

Universidade Federal de Minas Gerais

Adriana Falangola Benjamin Bezerra

Hospital das Clínicas da Universidade Federal de Pernambuco

Alexandre Vargas Schwarzbald

Hospital Universitário de Santa Maria

Aline Gabrielle Sousa Nunes

Confederação Nacional das Cooperativas Médicas

Amanda de Oliveira Maurilio

Federal University of São João del-Rei

Ana Luiza Bahia Alves Scotton

Hospital Regional Antônio Dias

André Soares de Moura Costa

Mater Dei Hospital

Andriele Abreu Castro

Hospital Moinhos de Vento

Bárbara Lopes Farace

Hospital Risoleta Tolentino Neves

Christiane Corrêa Rodrigues Cimini

Universidade Federal dos Vales do Jequitinhonha e Mucuri

Cíntia Alcântara De Carvalho

Hospital João XXIII

Daniel Vitorio Silveira

Confederação Nacional das Cooperativas Médicas

Daniela Ponce

São Paulo State University

Elayne Crestani Pereira

Universidade do Sul de Santa Catarina

Euler Roberto Fernandes Manenti

Hospital Mãe de Deus

Evelin Paola de Almeida Cenci

Hospital Universitário Canoas

Fernanda Barbosa Lucas

Hospital Santo Antônio

Fernanda d'Athayde Rodrigues

Hospital de Clínicas de Porto Alegre

Fernando Anschau

Hospital Nossa Senhora da Conceição

Fernando Antônio Botoni

Universidade Federal de Minas Gerais

Fernando Graça Aranha

Hospital SOS Cardio

Frederico Bartolazzi

Hospital Santo Antônio

Gisele Alsina Nader Bastos

Hospital Moinhos de Vento

Giovanna Grunewald Vietta

Universidade do Sul de Santa Catarina

Guilherme Fagundes Nascimento

Confederação Nacional das Cooperativas Médicas

Helena Carolina Noal

Hospital Universitário de Santa Maria

Helena Duani

Universidade Federal de Minas Gerais

Helóisa Reniers Vianna

Hospital Universitário Ciências Médicas

Henrique Cerqueira Guimarães

Hospital Risoleta Tolentino Neves

Isabela Moraes Gomes

Universidade Federal de Minas Gerais

Jamille Hemerito Salles Martins Costa

Hospital Márcio Cunha

Jessica Rayane Corrêa Silva Da Fonseca

Hospital Semper

Júlia Di Sabatino Santos Guimarães

Pontifícia Universidade Católica de Minas Gerais

Júlia Drumond Parreiras De Moraes

Hospital Universitário Ciências Médicas

Juliana Machado Rugolo

São Paulo State University

Joanna d'Arc Lyra Batista

Federal University of Fronteira Sul

Joice Coutinho De Alvarenga

Hospital João XXIII

José Miguel Chatkin

Pontifical Catholic University of Rio Grande do Sul

Karen Brasil Ruschel

Hospital Mãe de Deus

Leila Beltrami Moreira

Hospital de Clínicas de Porto Alegre

Leonardo Seixas De Oliveira

Hospital Santa Rosália

Liege Barella Zandona

Hospital Bruno Born

Lilian Santos Pinheiro

Universidade Federal dos Vales do Jequitinhonha e Mucuri

Luanna da Silva Monteiro

Hospital Metropolitano Odilon Behrens

Lucas de Deus Sousa

Hospital Regional Antônio Dias

Luciane Kopittke

Hospital Nossa Senhora da Conceição

Luciano de Souza Viana

Hospital Márcio Cunha

Luís César De Castro

Research Center of Vale do Taquari

Luísa Argolo Assis

Pontifícia Universidade Católica de Minas Gerais

Luísa Elem Almeida Santos

Centro Universitário de Patos de Minas

Maderson Álvares de Souza Cabral

Universidade Federal de Minas Gerais

Magda Cesar Raposo

Federal University of São João del-Rei

Maiara Anschau Floriani

Moinhos Research Institute

Maria Angélica Pires Ferreira

Hospital de Clínicas de Porto Alegre

Maria Aparecida Camargos Bicalho

Fundação Hospitalar do Estado de Minas Gerais

Mariana Frizzo De Godoy

Hospital São Lucas da PUCRS

Matheus Carvalho Alves Nogueira

Mater Dei Hospital

Meire Pereira De Figueiredo

Hospital Santo Antônio

Milton Henriques Guimarães Júnior

Hospital Márcio Cunha

Monica Aparecida de Paula De Sordi

São Paulo State University

Natália da Cunha Severino Sampaio

Hospital Eduardo de Menezes

Neimy Ramos De Oliveira

Hospital Eduardo de Menezes

Pedro Ledic Assaf

Hospital Metropolitano Doutor Célio de Castro

Raquel Lutkmeier

Hospital Nossa Senhora da Conceição

Reginaldo Aparecido Valacio

Hospital Metropolitano Odilon Behrens

Renan Goulart Finger

Hospital Regional do Oeste

Rochele Mosmann Menezes

Hospital Santa Cruz

Rufino de Freitas Silva

Hospital São João de Deus

Saionara Cristina Francisco

Hospital Metropolitano Doutor Célio de Castro

Silvana Mangeon Meireles Guimaraes

Hospital Semper

Silvia Ferreira Araujo

Hospital Semper

Talita Fischer Oliveira

Universidade Federal de Minas Gerais

Tatiana Kurtz

Hospital Santa Cruz

Tatiana Oliveira Fereguetti

Hospital Eduardo de Menezes

Thainara Conceição De Oliveira

Hospital Universitário Canoas

Túlio Henrique Oliveira Diniz

Hospital São João de Deus

Yara Cristina Neves Marques Barbosa Ribeiro

Hospital Metropolitano Doutor Célio de Castro

Yuri Carlotto Ramires

Hospital Bruno Born

Marcos André Gonçalves

Universidade Federal de Minas Gerais

Milena Soriano Marcolino (✉ milenamarc@gmail.com)

Universidade Federal de Minas Gerais

Bruno Barbosa Miranda de Paiva



Universidade Federal de Minas Gerais

Research Article

Keywords: COVID-19, prognosis, prediction model, machine learning

Posted Date: January 11th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1164411/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on March 1st, 2023. See the published version at <https://doi.org/10.1038/s41598-023-28579-z>.

Abstract

The majority prognostic scores proposed for early assessment of coronavirus disease 19 (COVID-19) patients are bounded by methodological flaws. Our group recently developed a new risk score - ABC₂SPH - using traditional statistical methods (least absolute shrinkage and selection operator logistic regression - LASSO). In this article, we provide a thorough comparative study between modern machine learning (ML) methods and state-of-the-art statistical methods, represented by ABC₂SPH, in the task of predicting in-hospital mortality in COVID-19 patients using data upon hospital admission. We overcome methodological and technological issues found in previous similar studies, while exploring a large sample (5,032 patients). Additionally, we take advantage of a large and diverse set of methods and investigate the effectiveness of applying meta-learning, more specifically Stacking, in order to combine the methods' strengths and overcome their limitations. In our experiments, our Stacking solutions improved over previous state-of-the-art by more than 26% in predicting death, achieving 87.1% of AUROC and MacroF1 of 73.9%. We also investigated issues related to the interpretability and reliability of the predictions produced by the most effective ML methods. Finally, we discuss the adequacy of AUROC as an evaluation metric for highly imbalanced and skewed datasets commonly found in health-related problems.

Introduction

The number of patients with coronavirus disease 2019 (COVID-19), as well as the related deaths, have increased exponentially during the COVID-19 pandemic. Although over 7.6 billion doses of COVID-19 vaccines have been administered, variants are still emerging, and COVID-19 seems to be an issue governments worldwide will need to keep grappling with (1,2).

Given the current scenario, there is an urgent need for an early disease stratification tool upon hospital admission, to allow the early identification of risk of death in patients with COVID-19, assisting in the management of disease and optimizing resource allocation, hopefully assisting to save lives during the pandemic. Although several scores have been proposed for the early assessment of COVID-19 patients at hospital admission, the majority of them are bounded by methodological flaws and technological limitations, meaning that reliable prognostic prediction models are scarce (3–5).

A state-of-the-art method for this prediction task has recently been proposed by our group with the development of a new risk score - ABC₂SPH - using traditional statistical methods (least absolute shrinkage and selection operator - LASSO regression), which exploits a rich set of information, including patient's demographics, comorbidities, vital signs and laboratory parameters at the time of presentation, for assessing prognosis in COVID-19 patients. The model has shown high discriminatory value (AUROC 0.844, 95% CI 0.829 to 0.859), confirmed in the Brazilian (0.859 [95% CI 0.833 to 0.885]) and Spanish (0.899 [95% CI 0.864 to 0.934]) validation cohorts, and with better discrimination ability than other existing scores (4).

In this context, artificial intelligence (AI), and more specifically machine learning (ML), techniques have been explored in various fields for dealing with the pandemic, such as detecting outbreaks, diagnosis,

interpretation of imaging exams, vaccines development and prognosis prediction (6,7), but comprehensive comparative studies to investigate whether ML techniques have superior performance to statistical methods are still scarce.

Indeed, in several other contexts (8) ML techniques have demonstrated superior effectiveness (i.e., accuracy) when compared to traditional statistical methods (e.g., logistic regression), due for instance, their capability of dealing with collinearity and redundancy, as well the ability to find non-linear correlations among the variables. However, current studies in the mortality prediction for COVID-19 using ML techniques are limited, regarding either methodological or technological aspects.

In this scenario, the contributions of this article are fivefold. First, we provide a **thorough comparative study** among state-of-the-art ML methods, including modern techniques, such as transformer and convolutional neural networks, boosting algorithms, support vector machines (SVM), k-nearest neighbors, as well as state-of-the-art statistical methods, represented by ABC₂-SPH, in the task of determining **in-hospital mortality** in COVID-19 patients, using data **upon hospital admission**.

Second, given the profusion and diversity of the compared methods, we investigate the effectiveness of meta-learning ensemble strategies, most notably Stacking (9), that combine the methods' outputs (probabilities), in order to exploit the ML methods' strengths and overcome their limitations.

Third, we study the **reliability** of the predictions of the most effective methods by correlating the probability of the outcome and the effectiveness (accuracy) of the methods. Few studies have investigated this important aspect of the predictions, which has practical impact in the applicability of the methods. Fourth, we investigated how **interpretable** (or explainable) are the predictions produced by the most effective methods, using modern interpretability tools.

Related Work

This study also included a narrative review on existing prediction models for COVID-19 mortality. A literature search of Medline and MedRxiv was conducted in August 2021, with no language or date restrictions. The search terms "COVID-19", "SARS-CoV-2" were used, combined with "mortality", "prognosis", "risk factors", "hospitalizations" or "score". Text screening retained 76 studies included in the S1 Table.

The existing literature largely focuses on American and Chinese in-hospital patients (53.94%). In fact, models validated in one country cannot be extrapolated to the population as a whole, since there is heterogeneity among countries in different characteristics such as populations features (including genetics, race, ethnicity, prevalence of comorbidities), socioeconomic factors, and the healthcare systems themselves (access, hospitals patient load, practice and available resources) (10).

Another remarkable point is the sample size. Larger population studies are needed to allow certain metrics of model performance to be estimated with more accurate and reliable results. In contrast, smaller samples reduce the ability to identify risk factors and increase the likelihood of overfitting (11). Among the analyzed

models, 17.10% were developed and validated in a sample of 500-1000 patients, and 35.52% had less than 500 patients. Only 47.36% of the studies used a sample with more than 1000 patients.

Most of the studies (60.52%) used traditional statistical methods, including multivariate logistic regression, LASSO and Cox regression analysis. Artificial intelligence techniques were used in 39.47% of studies, among them stands out machine learning, including random forest (RF), XGBoost and SVM. Only 11.8% of works exploit modern neural network methods.

Overall, the majority of developed models are limited by methodological bias, such as absence of external validation (testing) in 51.31%, so the assessment of accuracy in those studies may be overestimated, and only 23.68% reported having followed the methodological recommendations from Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) (11).

The model performance was evaluated in most studies by area under the receiver operating curve (AUROC). The mean AUROC for training ranged from 0.64-0.96 for traditional statistical methods, and from 0.74-0.96 for models using AI techniques. While more common in healthcare-related literature, the AUROC values can be misleading, especially when there is significant class imbalance (12). To properly assess the performance of different models, it is of utmost importance to use other metrics that consider imbalance issues, such as macro-average F1-score (macro-F1), used in only 13.33% studies.

Finally, few studies deeply analyzed the impact of the variables in the final model or on the final model outcome. Most studies did not investigate how reliable the made predictions are in terms of the correlation between the probability of the prediction and the accuracy. This analysis has implications on the practical use of this technology. An accurate but unreliable method has its practical applicability diminished. We explicitly tackle these issues in our study.

Materials And Methods

This is a substudy of the Brazilian COVID-19 Registry, a multi-hospital cohort study previously described (13). Adult patients with laboratory-confirmed COVID-19, admitted consecutively in any of the 36 participating hospitals, from March 1 to September 30, 2020 were enrolled. Individuals were not included if they were transferred between hospitals and data from the first or last hospitals was not available, as well those who were admitted for other reasons and developed COVID-19 symptoms during their stay (4).

Trained hospital staff or interns collected clinical and laboratory data from the medical records (S2 Table) (4,11). Laboratory exams were performed at the discretion of the treating physician. A prespecified case report form was used, applying Research Electronic Data Capture (REDCap) tools (14). Error checking code was developed in R, to identify data entry errors, as previously described (4), and the results were sent to each center for checking and correction, before further data analysis.

Data analysis

The development, validation and reporting of the models followed guidance from the TRIPOD guideline and the Prediction model Risk of Bias Assessment Tool (PROBAST) (11,15).

At that time, 36 Brazilian hospitals participated in the cohort, located in 17 cities, from five Brazilian states (4). A total of 5032 patients were admitted between March 1, 2020 and September 31, 2020, and the full group was used to perform a 10-fold cross validation procedure, which was repeated 3 times (at a total of 30 performance measurements for each of the classifiers presented in our study). The overall study population included 45.9% women, with a mean age of 60 ± 17 years, 27.17% needed mechanical ventilation and 20.15% died.

In order to properly assess the performance of different models, we chose to use three different metrics, each assessed through the aforementioned 10-fold cross validation procedure, for each classifier. Our evaluation metrics include both micro-average and macro-average F1-score (micro-F1 and macro-F1), and the AUROC. The F1 score is the harmonic mean between precision and recall scores, for each class (i.e. one score to estimate how well the model can predict which patients will die, and one to estimate the same regarding which patients will not die). The "average" part, described as either "micro" or "macro", refers to how these results are aggregated. In "macro" averaging, all classes are taken as equally important, while in "micro" averaging, class imbalance is not accounted for in the final result and all individual predictions are considered equally important (16).

As for the specific models compared in our study, we trained two modern neural network benchmarks – the FNet transformer, with and without virtual adversarial training, which is a regularization technique – and a deep convolutional Resnet. We also experimented with a support vector machine classifier, a boosting model (microsoft research's Light Gradient Boosting Machine), and the K-nearest neighbors algorithm, as well as a stacking of these methods.

We compare these ML alternatives to traditional statistical methods, including a Generalized Additive Model (GAM), which has rarely been used in this scenario before, and LASSO regression, the current state-of-the-art. GAM was used before in ABC₂-SPH, but only to select variables for the lasso regression, which yielded an inferior result when compared to LASSO regression, whereas in our work, we directly tune GAM to the classification task, thus obtaining better results, as we shall see.

The choice of neural networks to include in our study was motivated by current state-of-the-art methods, even though, in general, neural networks tend to perform better in situations where massive amounts of data are available (17,18). Usually, the ability to compare distant input positions in the query vectors is related to the neural network's depth. Transformer architectures, as introduced by Vaswani et al (2017), gained rapid success due to their capacity of doing so in a constant number of operations, achieving state-of-the-art results in many tasks (19). That is the reason we chose a FNet Transformer classifier. For comparison purposes, we also included a Resnet model, which held similar success for image classification, due to the capacity of building very deep networks. Due to the relative drop in performance of neural networks when fewer data samples are present in training, we also included a training variant where we perform virtual adversarial training, as introduced in Miyato et al (2017), in which the model's

decision boundary is smoothed in the most anisotropic direction through a gradient-based approximation (20).

Additionally, we included a standard support vector machine classifier, which learns a separation hyperplane between classes, while maximising the separation margin, and a K-nearest neighbors classifier, which yields predictions based on spatial similarities between training samples and new query points. Motivated by the results shown in Shwartz et al (2021) (21), we included a boosting algorithm (LightGBM), which is usually an effective model in tabular data, as concluded in Ke et al (2017) (22). As the final classifier, we included a meta-learning ensemble-based Stacking model, which learns to combine the prediction outputs of all previous classifiers in order to improve classification effectiveness. We compare these methods to Generalised Additive Models (GAM) and LASSO regression, the latter being the current state-of-the-art model for this task, as demonstrated in our previous work.

We ran all classification tests using a 10 fold cross validation procedure, after which we calculated confidence intervals for each result, and confirmed statistical significance by applying a Wilcoxon signed-rank test with 95% confidence.

For the parametrization of our models, we used the values presented in Table 1, where the values in brackets are evaluated in the validation set of the cross validation process. For deep network models we use the early stop to optimize the model, which optimizes the weights until the model has no significant improvement in the validation set.

Table 1
Parameterization of methods.

Method	Parametrization
SVM	C: [10^{-3} , 10^{-2} , 10^{-1} , 10^0 , 10^1 , 10^2] Kernel: [linear, rbf, poly, sigmoid] class_weight: [None, 'balanced']
RF	N-estimators: [10, 50, 100, 200, 500, 1000, 2000]
KNN	Neighbors: [2, 4, 8, 16, 32]
LASSO	Alpha: [10^{-3} , 10^{-2} , 10^{-1} , 10^0 , 10^1 , 10^2]
LIGHT_GBM	N-estimators: [10, 50, 100, 200, 500, 1000, 2000] learning_rate: [10^{-3} , 10^{-2} , 10^{-1} , 30^{-1}] colsample_by_tree: [0.5, 1.0]
CNN	Early Stop
FNet	Early Stop
FNet + VAT	Early Stop
GAM	No tuning
Stacking	Meta-Classifer: Logistic Regression, Alpha: [10^2]
List of model names: CNN = convolutional neural network, FNet = fourier transformation neural network, FNet + VAT = fourier transformation neural network with virtual adversarial training, GAM = generalized additive models, KNN = K-nearest neighbors, LASSO = lasso regression, LIGHT_GBM = light gradient boosting machines, RF = random forest, SVM = support vector machines, STACKING = a stacking classifier, which combines all others.	

Ethics approval and consent to participate

The study protocol was approved by the Brazilian National Commission for Research Ethics (CAAE 30350820.5.1001.0008). Individual informed consent was waived due to the severity of the situation and the use of deidentified data, based on medical chart review only. All methods were carried out in accordance with relevant guidelines and regulations.

Results And Discussion

Classification results for the prediction of death are shown in Table 1. Neural network models (CNN - convolutional neural networks - and FNet - Fourier transform neural network - / FNet + VAT - Fourier transform neural network + virtual adversarial training) produced the worst results, while boosting ('LightGBM' - Light Gradient Boosting Machine), Stacking and one traditional statistical model ('Generalized Additive Models - GAM') produced the best overall results, when considering micro-F1, macro-F1 and AUROC.

The less effective results of the Neural network are somewhat expected as the size of the dataset is not that huge, with fewer than 10,000 samples. Typically, we expect neural networks of large capacity (millions to billions of parameters) to excel in tasks where very large datasets are available (millions to billions of training instances), which is very rare in health-related problems. In such large-scale datasets, neural networks can capture very complex relationships. However, in smaller sample sizes, they show a remarkable tendency to overfit, hence obtaining poor results in terms of validation error (17,18).

Thus, tree-based ensemble models such as random and boosting forests tend to be more robust to small sample sizes and to overfitting, which is exactly the behavior we observed in our experiments (23). SVM and K-nearest neighbors (KNN), which are simpler models, with fewer parameters, also tend to perform reasonably well on smaller datasets being better than the neural network models.

We should stress that the statistical models LASSO regression and GAM showed very competitive results. Unexpectedly, GAM was the runner up method considering all metrics, being even better than LASSO and some traditional ML methods such as SVM and KNN. In our work, we directly tuned GAM to the classification task, using the cross-validation procedure, which yielded superior performance. GAM and LightGBM are statistically tied regarding all evaluation metrics considering a Wilcoxon signed-rank test with 95% confidence.

In any case, the best single overall model, with statistical significance, under all considered metrics, was the Stacking model, which is a combination of the output of all other individual models, which, in turn, allows us to better discriminate between patients with higher clinical risk at admission time. When considering micro and macro-F1, F1 for death and AUROC at the task of predicting death, Stacking was significantly (statistically) better than all other models. The largest gains were in F1 to predict death with gains of up more than 26% over LASSO, the previous state-of-the-art. The Stacking technique improves the F1-score results for the class of interest (death) by 7% over RF, by 5% for LightGBM and by 6% for GAM, which were the three individual best models in this metric. The combination of models based on different classification premises, potentially made stacking more robust. If a single classifier makes a wrong prediction, the others can still make corrections, increasing the robustness of the final stacking model.

Table 2. Micro-F1, macro-F1 and AUROC results for the prediction of COVID-19 in-hospital death.										
	MICRO-F1		MACRO-F1		F1 (DEATH)		F1 (NO DEATH)		AUROC	
	mean	CI	mean	CI	mean	CI	mean	CI	mean	CI
KNN	0.807	0.002	0.492	0.007	0.091	0.014	0.892	0.001	0.781	0.010
FNet + VAT	0.810	0.013	0.677	0.020	0.470	0.038	0.884	0.009	0.772	0.019
FNet	0.814	0.008	0.686	0.017	0.486	0.030	0.887	0.005	0.789	0.015
CNN	0.815	0.013	0.693	0.016	0.500	0.026	0.886	0.009	0.796	0.016
SVM	0.839	0.010	0.691	0.031	0.478	0.058	0.904	0.005	0.833	0.012
LASSO	0.842	0.009	0.677	0.024	0.446	0.044	0.908	0.005	0.859	0.006
LIGHT_GBM	0.846	0.008	0.723	0.016	0.538	0.028	0.908	0.005	0.865	0.008
GAM	0.847	0.006	0.720	0.014	0.532	0.026	0.908	0.003	0.855	0.012
RF	0.850	0.005	0.717	0.013	0.524	0.024	0.911	0.003	0.863	0.007
STACKING	0.855	0.007	0.739	0.018	0.564	0.032	0.913	0.004	0.871	0.007

List of model names, from top to bottom (ordered by MicF1): CNN = convolutional neural network, FNet = fourier transformation neural network, FNet + VAT = fourier transformation neural network with virtual adversarial training, GAM = generalized additive models, KNN = K-nearest neighbors, LASSO = lasso regression, LIGHT_GBM = light gradient boosting machines, RF = random forest, SVM = support vector machines, STACKING = a stacking classifier, which combines all others.

The ROC curves for all evaluated models are shown in Fig 1. From this Figure, we can see the separation of two distinct groups: one group of models with inferior results, composed of neural network models and K-nearest neighbors, and another group of models with superior (indistinguishable) results, consisting of SVM, RF, LightGBM, GAM and the Stacking.

Despite similarities in the curves and at AUROC values, these classifiers can yield quite different results when compared with micro-F1 and macro-F1, or class-specific F1 scores, which shows that (1) AUROC score is not an adequate metric for evaluating and comparing models, especially in face of high imbalance/skewness and that (2) even though some models, like Stacking and GAM have very similar AUROC scores, their capacity to discriminate relevant outcomes like death is quite different (0.532 F1 score for GAM and 0.564 for Stacking, a significant difference of 6%).

Interestingly, using such curves, we can sensibly calibrate the trade-off between sensitivity and specificity, further customizing the way such models can be used. In particular, when applying Stacking, our model can be tailored to the early identification of high-risk patients, with good discrimination capacity.

Explainability

Explainability is an essential aspect of the task if ML methods are to be trusted and actually used by practitioners. Various prognostic factors have been proposed in the stratification of COVID-19 patients, based on their risk of death, that includes clinical, laboratory and radiological variables. Among these risk factors, stand out advanced age, multiple comorbidities on admission (such as hypertension, diabetes mellitus, cardiovascular diseases and others), abnormal levels of C-reactive protein (CRP), lymphocytes, neutrophils, D-dimer, blood urea nitrogen (BUN) and lactate dehydrogenase (LDH).

A very interesting feature of some ML models, in particular decision trees, RF and boosting forests, is the explainability of these models. This is still a very active research area, but modern advances in tools and visualization alternatives allow us to represent which features are most important to the model and at which polarities and intervals. In this context and as previously cited, the best model in our tests was the Stacking, that is a meta-model, in which inputs are the outputs of other classifiers. Since we aim to explain a classifier that works on the level of the features themselves (instead of a meta-level of other classifier outputs), we will provide explanations for the second best model, LightGBM. Furthermore, tree-based boosting and bagging algorithms rank as some of the most explainable machine learning models, and also lead many benchmarks, particularly for tabular data where data samples are not that large. Their unique combination of explainability, reliability and performance, added to the fact that stacking is a meta-classifier are why we will exploit the boosting model (which, in our case, outperformed the bagging model - random forests/RF), to analyse the correlations among variables.

In a sense, some traditional models, such as regression models, also have a good explainability, as it is possible to assess the coefficients of each attribute, to measure how important a feature is. These models however do not measure up to modern tree-based algorithms in many scenarios, especially in cases with larger datasets (24). Another key difference between these models is that, in the case of regression models, we have to explicitly remove collinear variables, but these variables, even though they might not improve classification performance, still yield valid model explanations.

In decision tree based algorithms, each node represents a feature. The closer to the root (i.e. the 'first' node of each tree), the more the feature is able to differentiate the data classes. For example, in Fig 2, feature 'SF ratio' with the value less than 233 and the feature 'lactate' with a value less than 1.68 mmol/L results in a subset with 5.9% of the dataset where the 'death' outcome is more common.

These algorithms look for the values of the features that further separate the classes, while trying to decrease the coefficient or entropy values of the class label (which are measures of purity and information) in each partition in the decision tree – this coefficient is called the GINI Index. Such index and the entropy score tend to isolate records that represent the most frequent class in a branch.

In Fig 3, we present SHapley Additive exPlanation (SHAP) values for our boosting model. This is a special type of explainability technique, which allows us to not only probe which features were important to the model, but also which polarities or intervals push predictions to each of the training classes, and additionally, allows us to evaluate why the model predicted any single instance (25).

For any simple model, such as regression, the model itself is a reasonable explanation of what was learned. However, for more complex models, which in turn are capable of learning more complex solutions (provided enough data is present at training time), we cannot use the model to explain itself, since it is a complex solution. In these situations, shapley values build upon the idea that the explanation of a model can itself be a model. This technique has been recently introduced, and further expands on the explainability of machine learning models, making them even more useful, as they become more interpretable (25).

With the help of SHAP values in Figs 3 and 4, we can extract interesting knowledge from our boosting model, the best individual ML model that works with the base features. We can see for instance that the most important feature in the prediction of death COVID-19 is age. This is coherent with previous medical literature, and serves as an additional validation to the model. Other scores and a recent meta-analysis have shown age as a key prognostic determinant in COVID-19 (26–29). The meta-analysis included more than half million of COVID-19 patients from different countries, and observed that the risk increased exponentially after the fifth decade of life. It is important to highlight that this fact could be influenced by both the physiological aging process and, especially by the individuals functional status and reserve, what may hinder the intrinsic capacity to fight against infections, increasing susceptibility to the infection and severe clinical manifestations (30).

The second most important feature is the supplemental oxygen requirement, which, as per Fig 3, lower values (blue tones) indicate higher risk. Although COVID-19 is a multisystem disease, it is well known that lung involvement is the mainstay for assessing disease severity, and oxygen requirement upon hospital admission has been shown to be an independent predictor for severe COVID-19 in several studies (31,32).

We also observed that lower values of platelets and higher levels of urea and C-reactive protein also increase risk of mortality, which is in line with what was previously observed using statistical models (33). Other studies suggest that C-reactive protein was a marker of a cytokine storm developing in patients with COVID-19 and was associated with the disease mortality (34–36).

Interestingly, ML models can return explanations in the form of intervals, such as the behavior seen in Fig 3 for sodium and bicarbonate levels, which imply there is a "safe interval" for which risk is lower, but values either too high or too low yield higher risk of death. This is an intrinsic limitation of regression models, and the variable may be seen as non-significant due to the fact that it is a non-linear association.

From a clinical perspective, those results are in line with a recent study, which demonstrated that hyponatremia and hypernatremia during COVID-19 hospitalization are associated with a higher risk of death and respiratory failure, respectively (37). With regards to bicarbonate, low levels are related to acidosis, and high levels are usually related to advanced chronic obstructive pulmonary disease (COPD) with retention of carbon dioxide, both of them conditions well-known to be associated with worse prognosis in clinical practice (38–40). This sort of non-linearity cannot be captured by simple regression models, since we can only measure how large coefficient values are, and correlate that to the importance of each feature.

In exploiting LASSO regression in our previous work (4), we had to exclude some features which had shown to be important in the boosting model due to high collinearity. This may explain the difference in the features included in both models, despite the fact that all features included in both had previous evidence of association with COVID-19 prognosis.

Another interesting remark is shown in Fig 4, in which we can see the relative importance of each feature. Here, again, age is the most important single feature (due to higher mean SHAP value), which is in line with previous studies (3,26,27). However, the remaining features when combined yield higher predictive value in this task than just age.

Reliability

Finally, we investigate issues related to the reliability of the models. Neural network models are, for instance, known for having irregular error rates, regardless of prediction confidence. At the other end of the spectrum, boosting and bagging models tend to have a very interesting reliability profile, with a tendency to have lower error rates at high confidence scores, and higher error rates at lower confidence scores. This enables tuning the trade-off between accuracy and sensitivity for some specific classifiers.

Accordingly, we show in Fig 5 the reliability profile for our best model (Stacking). In this Figure, the x-axis shows prediction ranges for the model's confidence score, while the y-axis shows the percentage of hits or misses for the model. Note that the model makes more correct predictions (hits, in green) when it is more certain of the prediction (range 0.87-0.96). Thus, this classifier yields a useful reliability profile, in relation to its confidence score. This kind of characteristic means we can tune how many patients the model will indicate, as well as how sensitive or specific that indication can be. Such tuning can be tailored to any healthcare service, accounting for intensive care unit beds, available professionals and so on.

Based on S1 Table, there were few prediction studies that had extensive analysis utilizing AI techniques. In this study, AI techniques were compared to traditional statistical methods to develop a model to predict COVID-19 mortality, considering demographic, comorbidity, clinical presentation and laboratory analysis data. We observed that regarding the prediction of the class of interest (death), the best individual methods was a ML one (LightGBM) closely followed by a statistical model (GAM), both being better than neural network models, and both being surpassed by a meta-learning ensemble model – Stacking – which was the best overall solution, considering all criteria for the posed prediction problem.

We would like to emphasize that, despite the fact that in medical research the AUROC is widely used as the sole measure of models' discriminatory ability, our data reassures that it is an insufficient metric for evaluating and comparing models. F1 Score is a more robust metric, especially in larger, more complex and imbalanced datasets, which are common in health-related scenarios.

Conclusion

In this study, modern AI techniques performed better than traditional statistical methods to predict COVID-19 mortality. The meta-learning strategy based on Stacking, surpassed the state-of-the-art LASSO

regression method by more than 26% for prediction death. We also demonstrated that AUROC score was an insufficient metric for evaluating and comparing models. Even though some models, like Stacking and GAM have very similar AUROC scores, their capacity to discriminate relevant outcomes like death is quite different (0.53 F1 score for GAM and 0.56 for Stacking, which yields a 5.6% difference). Finally, we investigated issues related to the explainability and prediction reliability of the best ML models, concluding that they are potentially very useful for practical purposes in real settings. Age was the main mortality risk predictor, but urea, C-reactive protein, lactate, respiratory rate, heart rate, NRL, neutrophils, sodium and pCO₂ may also significantly influence the disease outcome.

Declarations

Acknowledgments

We would like to thank the hospitals which are part of this collaboration, for supporting this project: Hospital Bruno Born; Hospital Cristo Redentor; Hospital das Clínicas da Faculdade de Medicina de Botucatu; Hospital das Clínicas da UFMG; Hospital das Clínicas da Universidade Federal de Pernambuco; Hospital de Clínicas de Porto Alegre; Hospital Santo Antônio; Hospital Eduardo de Menezes; Hospital João XXIII; Hospital Julia Kubitschek; Hospital Mãe de Deus; Hospital Márcio Cunha; Hospital Mater Dei Betim-Contagem; Hospital Mater Dei Contorno; Hospital Mater Dei Santo Agostinho; Hospital Metropolitano Dr. Célio de Castro; Hospital Metropolitano Odilon Behrens; Hospital Moinhos de Vento; Hospital Nossa Senhora da Conceição; Hospital Regional Antônio Dias; Hospital Regional de Barbacena Dr. José Américo; Hospital Regional do Oeste; Hospital Risoleta Tolentino Neves; Hospital Santa Cruz; Hospital Santa Rosália; Hospital São João de Deus; Hospital São Lucas da PUCRS; Hospital Semper; Hospital SOS Córdio; Hospital Tacchini; Hospital Unimed-BH; Hospital Universitário Canoas; Hospital Universitário Ciências Médicas; Hospital Universitário de Santa Maria.

We also thank all the clinical staff at those hospitals, who cared for the patients, and all undergraduate students who helped with data collection.

Author contributions statement

Substantial contributions to the conception or design of the work: BBMP, CMVA, MAG and MSM.

Substantial contributions to the acquisition, analysis, or interpretation of data for the work: all authors.

Drafted the work: BBMP, PDP, CMVA, VMRG, MC, MAG and MSM.

Revised the manuscript critically for important intellectual content: all authors.

Final approval of the version to be published: all authors.

Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: BBMP, MAG and MSM.

Competing interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study was supported in part by Minas Gerais State Agency for Research and Development (*Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG*) [grant number APQ-00208-20], National Institute of Science and Technology for Health Technology Assessment (*Instituto de Avaliação de Tecnologias em Saúde – IATS*)/ National Council for Scientific and Technological Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq*) [grant number 465518/2014-1], and CAPES Foundation (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) [grant number 88887.507149/2020-00].

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*. 2020 May;20(5):533–4.
2. Callaway E. Could new COVID variants undermine vaccines? Labs scramble to find out. *Nature*. 2021 Jan 14;589(7841):177–8.
3. Fumagalli C, Rozzini R, Vannini M, Coccia F, Cesaroni G, Mazzeo F, et al. Clinical risk score to predict in-hospital mortality in COVID-19 patients: a retrospective cohort study. *BMJ Open* [Internet]. 2020 Sep 25;10(9):e040729. Available from: <https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2020-040729>
4. Marcolino MS, Pires MC, Ramos LEF, Silva RT, Oliveira LM, Carvalho RLR, et al. ABC2-SPH risk score for in-hospital mortality in COVID-19 patients: development, external validation and comparison with other available scores. *International Journal of Infectious Diseases* [Internet]. 2021 Sep;110:281–308. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1201971221006056>
5. Gupta RK, Marks M, Samuels THA, Luintel A, Rampling T, Chowdhury H, et al. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. *European Respiratory Journal* [Internet]. 2020 Dec;56(6):2003498. Available from: <http://erj.ersjournals.com/lookup/doi/10.1183/13993003.03498-2020>
6. Borges do Nascimento IJ, Marcolino MS, Abdulazeem HM, Weerasekara I, Azzopardi-Muscat N, Gonçalves MA, et al. Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future Studies. *Journal of Medical Internet Research*. 2021 Apr 13;23(4):e27275.
7. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. *Radiology*. 2020 Sep;296(3):E156–65.

8. Mohnen SM, Rotteveel AH, Doornbos G, Polder JJ. Healthcare Expenditure Prediction with Neighbourhood Variables – A Random Forest Model. *Statistics, Politics and Policy*. 2020 Dec 16;11(2):111–38.
9. Gomes C, Goncalves M, Rocha L, Canuto S. On the Cost-Effectiveness of Stacking of Neural and Non-Neural Methods for Text Classification: Scenarios and Performance Prediction. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021;4003–14.
10. Núñez-Gil IJ, Fernández-Pérez C, Estrada V, Becerra-Muñoz VM, El-Battrawy I, Uribarri A, et al. Mortality risk assessment in Spain and Italy, insights of the HOPE COVID-19 registry. *Internal and Emergency Medicine*. 2021 Jun 9;16(4):957–66.
11. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*. 2015 Jan 6;162(1):W1–73.
12. Brabec J, Machlica L. Bad practices in evaluation methodology relevant to class-imbalanced problems. 2018 Dec 4;
13. Marcolino MS, Ziegelmann PK, Souza-Silva MVR, Nascimento IJB, Oliveira LM, Monteiro LS, et al. Clinical characteristics and outcomes of patients hospitalized with COVID-19 in Brazil: Results from the Brazilian COVID-19 registry. *International Journal of Infectious Diseases*. 2021 Jun;107:300–10.
14. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O’Neal L, et al. The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*. 2019 Jul;95:103208.
15. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*. 2019 Jan 1;170(1):51.
16. Cuadros-Rodríguez L, Pérez-Castaño E, Ruiz-Samblás C. Quality performance metrics in multivariate classification methods for qualitative analysis. *TrAC Trends in Analytical Chemistry*. 2016 Jun;80:612–24.
17. Cunha W, Mangaravite V, Gomes C, Canuto S, Resende E, Nascimento C, et al. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*. 2021 May;58(3):102481.
18. Cunha W, Canuto S, Viegas F, Salles T, Gomes C, Mangaravite V, et al. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management*. 2020 Jul;57(4):102263.
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L PI. Attention is all you need. *Conference on Neural Information Processing System*. 2017;
20. Miyato T, Maeda S, Koyama M, Ishii S. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. 2017 Apr 12;
21. Shwartz-Ziv R, Armon A. Tabular Data: Deep Learning is Not All You Need. 2021 Jun 6;

22. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*. 2017;2017-Decem:3147–55.
23. Salles T, Rocha L, Gonçalves M. A bias-variance analysis of state-of-the-art random forest text classifiers. *Advances in Data Analysis and Classification*. 2021 Jun 19;15(2):379–405.
24. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology*. 2007 Nov;88(11):2783–92.
25. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. 2020 Jan 17;2(1):56–67.
26. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ*. 2020 Sep 9;m3339.
27. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. *JAMA Internal Medicine*. 2020 Aug 1;180(8):1081.
28. Bonanad C, García-Blas S, Tarazona-Santabalbina F, Sanchis J, Bertomeu-González V, Fácila L, et al. The Effect of Age on Mortality in Patients With COVID-19: A Meta-Analysis With 611,583 Subjects. *Journal of the American Medical Directors Association*. 2020 Jul;21(7):915–8.
29. Chowdhury MEH, Rahman T, Khandakar A, Al-Madeed S, Zughaier SM, Doi SAR, et al. An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. 2020 Jul 29;
30. Aliberti MJR, Szlejf C, Avelino-Silva VI, Suemoto CK, Apolinario D, Dias MB, et al. COVID-19 is not over and age is not enough: Using frailty for prognostication in hospitalized patients. *Journal of the American Geriatrics Society*. 2021 May 5;69(5):1116–27.
31. Bhargava A, Fukushima EA, Levine M, Zhao W, Tanveer F, Szpunar SM, et al. Predictors for Severe COVID-19 Infection. *Clinical Infectious Diseases*. 2020 Nov 5;71(8):1962–8.
32. Daher A, Balfanz P, Aetou M, Hartmann B, Müller-Wieland D, Müller T, et al. Clinical course of COVID-19 patients needing supplemental oxygen outside the intensive care unit. *Scientific Reports*. 2021 Dec 26;11(1):2256.
33. Bashash D, Hosseini-Baharanchi FS, Rezaie-Tavirani M, Safa M, Akbari Dilmaghani N, Faranoush M, et al. The Prognostic Value of Thrombocytopenia in COVID-19 Patients; a Systematic Review and Meta-Analysis. *Archives of academic emergency medicine*. 2020;8(1):e75.
34. Zhang J, Cao Y, Tan G, Dong X, Wang B, Lin J, et al. Clinical, radiological, and laboratory characteristics and risk factors for severity and mortality of 289 hospitalized COVID-19 patients. *Allergy*. 2021 Feb 24;76(2):533–50.
35. Ouyang S-M, Zhu H-Q, Xie Y-N, Zou Z-S, Zuo H-M, Rao Y-W, et al. Temporal changes in laboratory markers of survivors and non-survivors of adult inpatients with COVID-19. *BMC Infectious Diseases*. 2020 Dec 11;20(1):952.
36. Izcovich A, Ragusa MA, Tortosa F, Lavena Marzio MA, Agnoletti C, Bengolea A, et al. Prognostic factors for severity and mortality in patients infected with COVID-19: A systematic review. Lazzeri C, editor.

37. Tzoulis P, Waung JA, Bagkeris E, Hussein Z, Biddanda A, Cousins J, et al. Dysnatremia is a Predictor for Morbidity and Mortality in Hospitalized Patients with COVID-19. *The Journal of Clinical Endocrinology & Metabolism*. 2021 May 13;106(6):1637–48.
38. Gunnerson KJ. Clinical review: the meaning of acid-base abnormalities in the intensive care unit part I - epidemiology. *Critical care (London, England)*. 2005 Oct 5;9(5):508–16.
39. Raphael KL, Murphy RA, Shlipak MG, Satterfield S, Huston HK, Sebastian A, et al. Bicarbonate Concentration, Acid-Base Status, and Mortality in the Health, Aging, and Body Composition Study. *Clinical Journal of the American Society of Nephrology*. 2016 Feb 5;11(2):308–16.
40. Pahal P, Hashmi MF, Sharma S. Chronic Obstructive Pulmonary Disease Compensatory Measures. *StatPearls*. 2021.

Figures

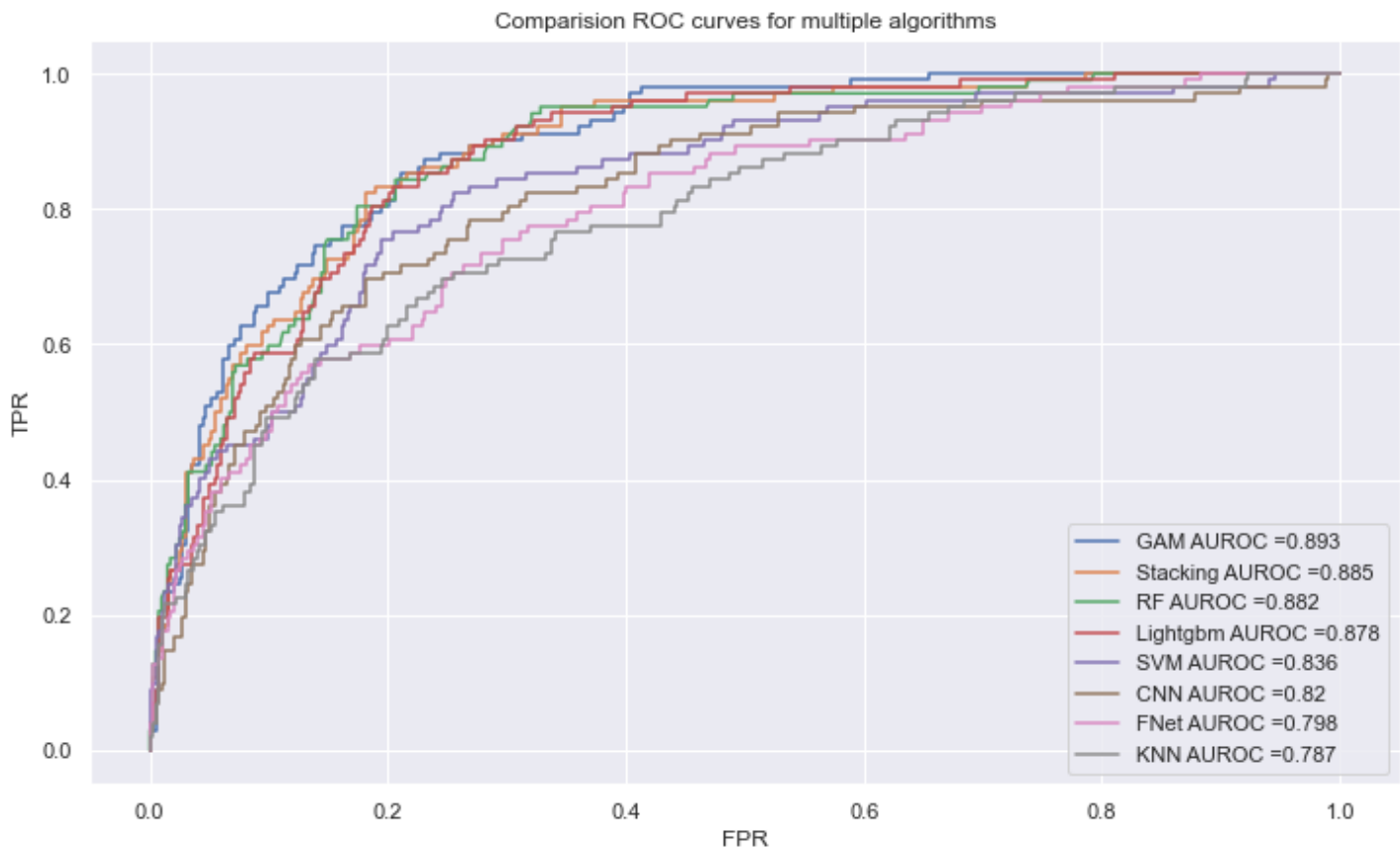


Figure 1

Receiver Operating Characteristic (ROC) curve comparing multiple models, trained on the prediction of the death outcome.

Sample decision tree with max_depth=2 from our dataset

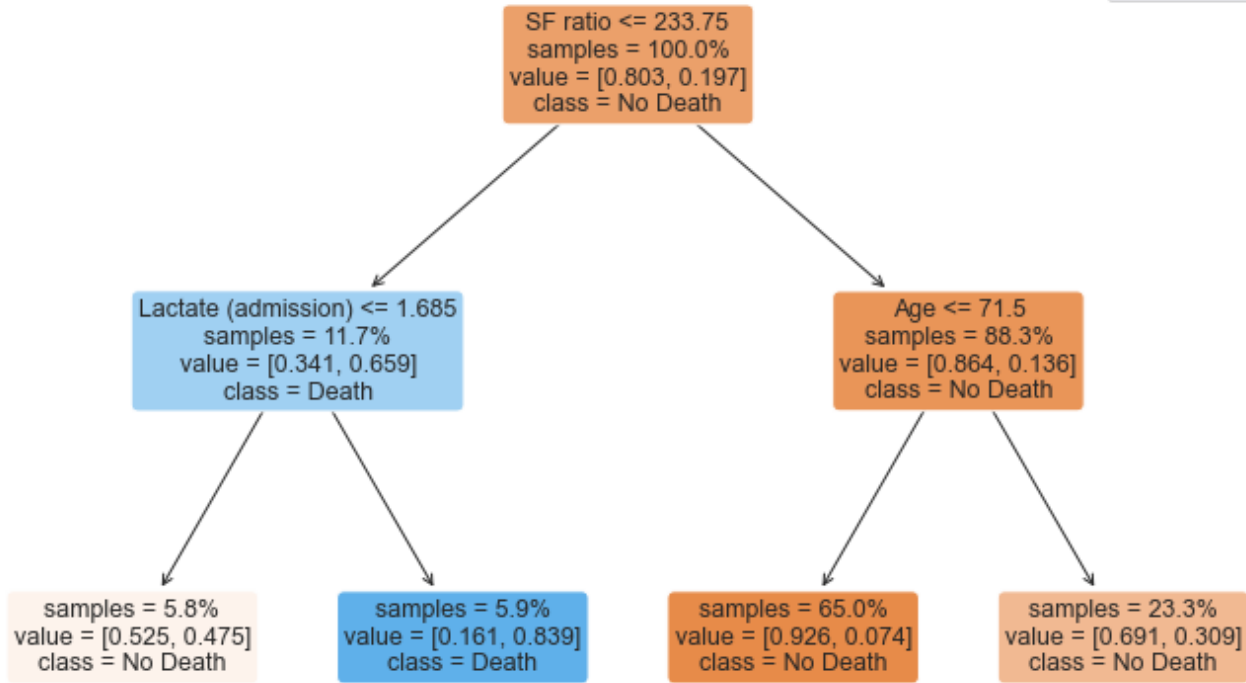


Figure 2

A sample decision tree with depth 2, trained on our dataset. At each level but the last, the first line of text in each box shows the variable and its cut before the split.

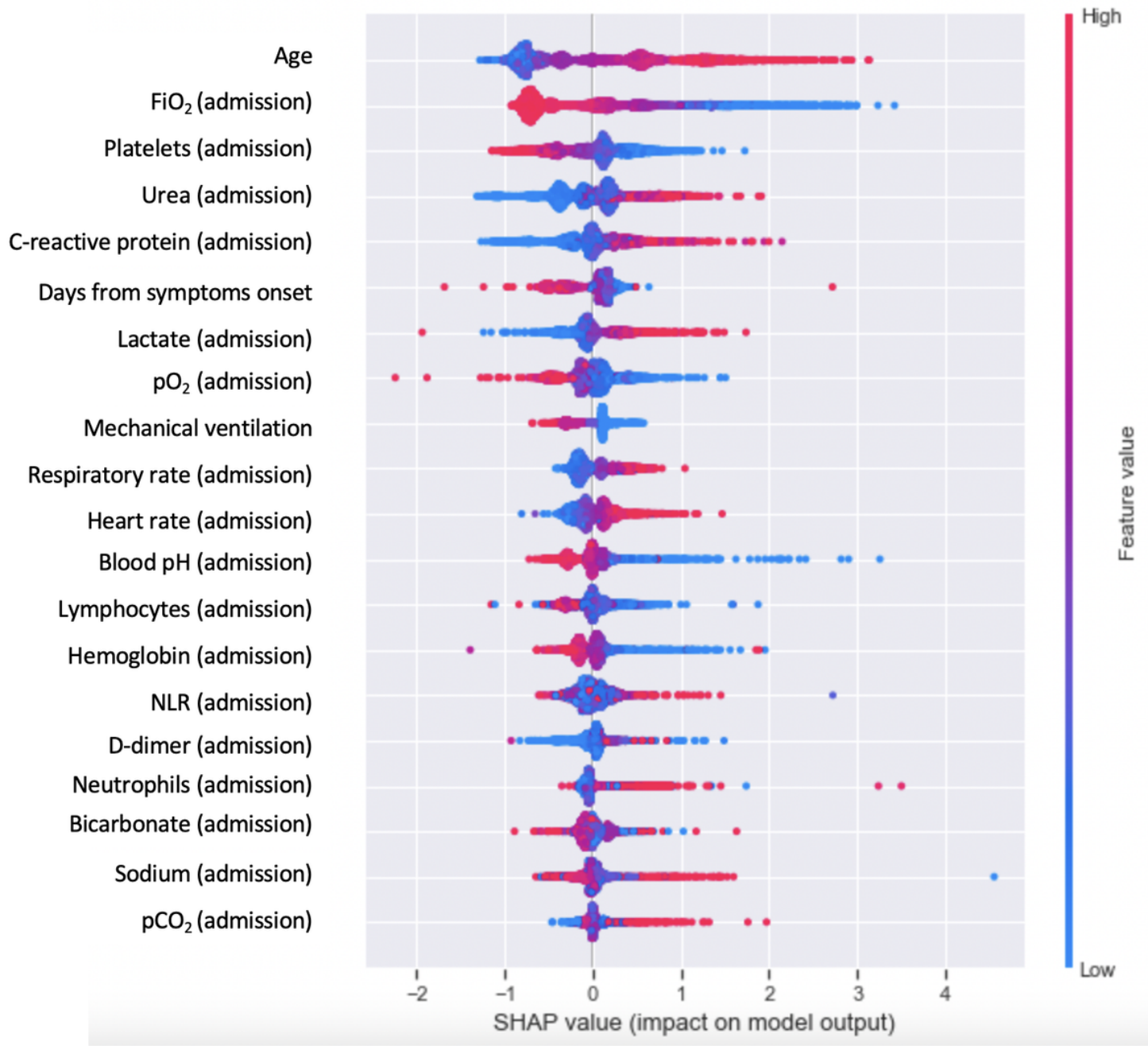


Figure 3

SHAP values for the LightGBM model trained on the prediction of the death outcome.

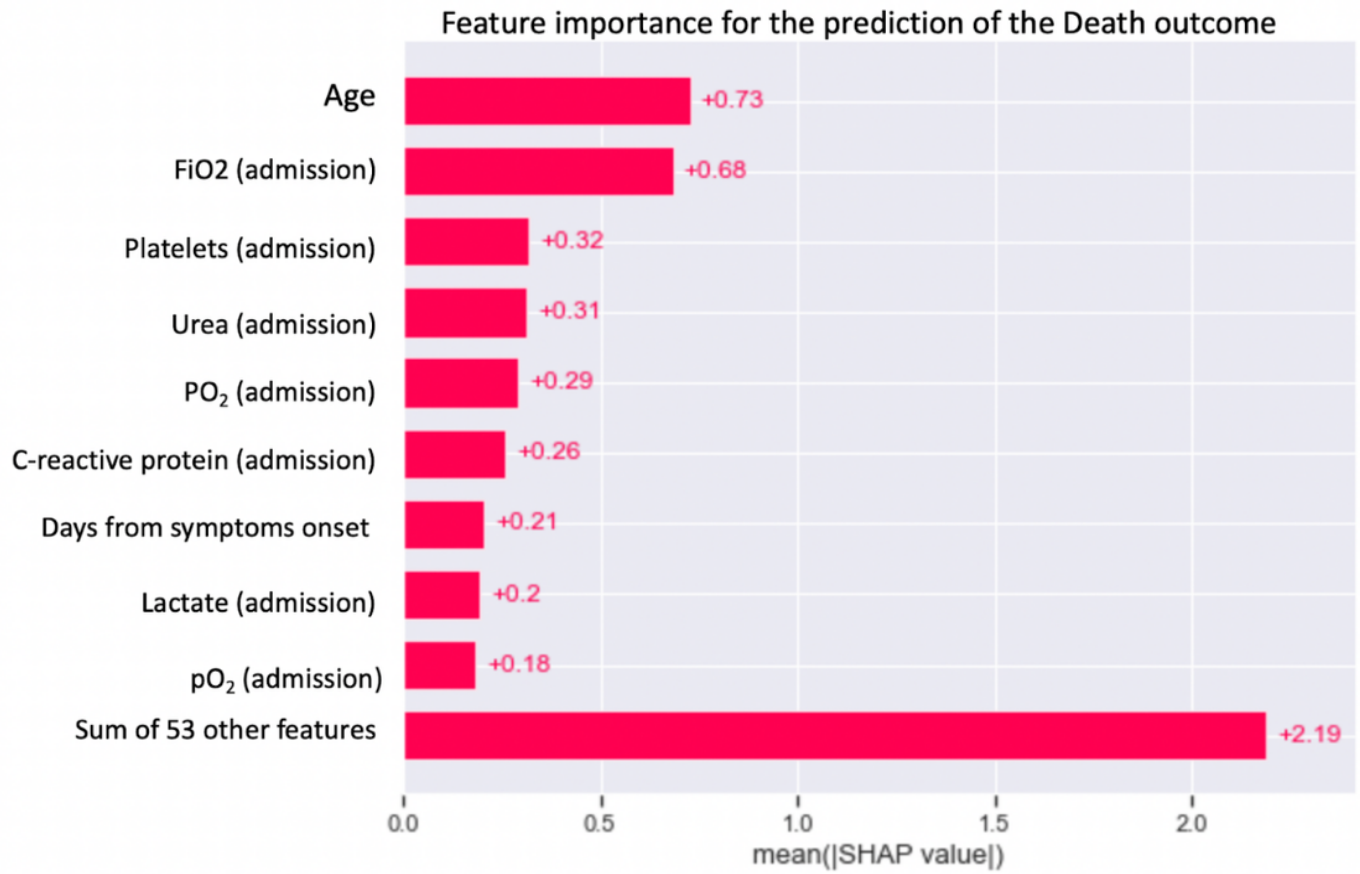


Figure 4

Mean SHAP values for each feature in the prediction of either death.

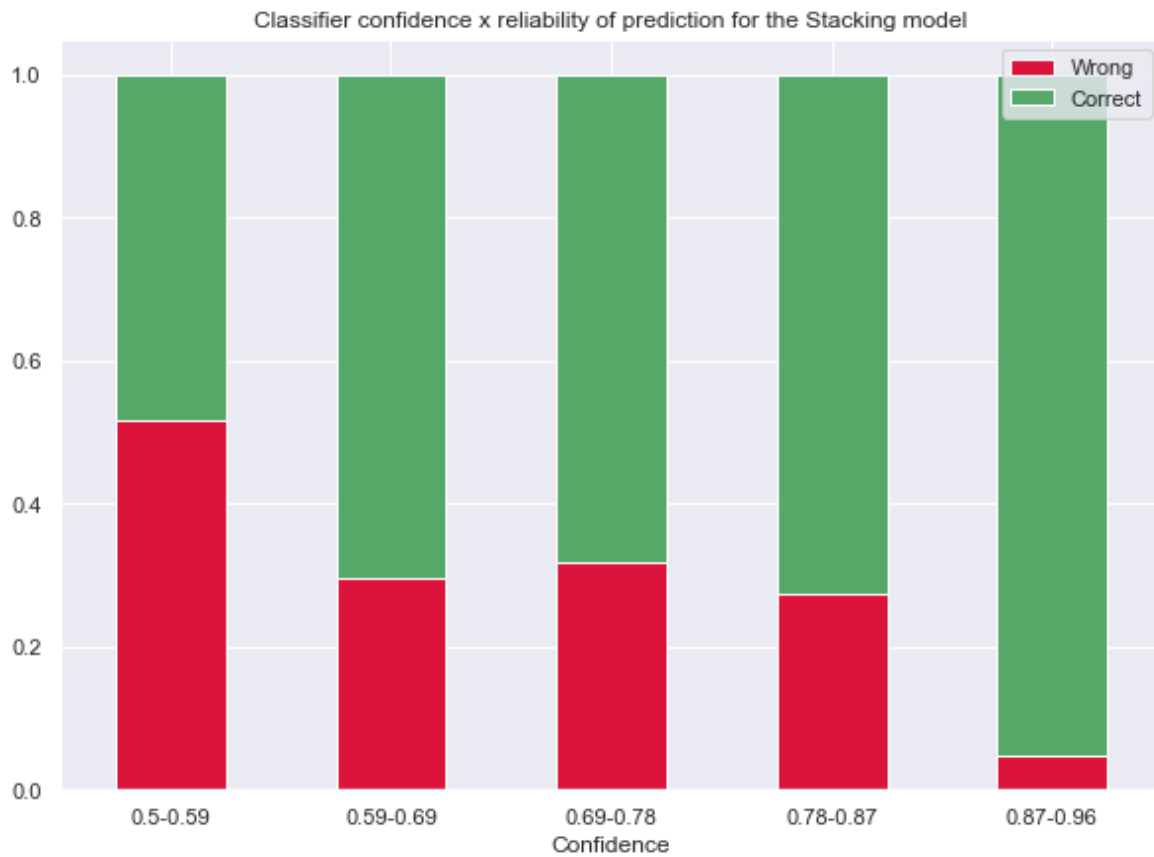


Figure 5

Error rates for each confidence threshold in the Stacking model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS1andS2.pdf](#)