# Towards Effective and Reliable Data-driven Prognostication: An Application to COVID-19

José Solenir Lima Figuerêdo[1], Renata Freitas Araujo-Calumby[2], Rodrigo Tripodi Calumby[1]

[1] University of Feira de Santana, Brazil
jslfigueredo@ecomp.uefs.br, rtcalumby@uefs.br
[2] Feira de Santana Higher Education Unit
farm.renata@hotmail.com

**Abstract.** This study evaluates machine learning methods to predict the prognosis of patients in COVID-19 context. In addition, considering the best-performing machine learning algorithm, we applied the LIME explanation technique for machine learning models to verify how the features correlate with each decision made, in order to assist an expert regarding the groundings of the decision made by the model. The results reveal that the model developed was able to predict the patient's prognosis with an ROC-AUC = 0.8524. The prediction explanations allowed us to understand how each feature contributes to the decision made by the model, thus bringing transparency to the developed model.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: COVID-19; Machine Learning; Computer aided prognosis; Mortality prediction; Explainable AI.

## 1. INTRODUCTION

Coronavirus disease (COVID-19) is an infection caused by Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2). SARS-CoV-2 corresponds to a binuclear virus that has a broad clinical spectrum of infection [Lu et al. 2020]. When infecting a host, this agent can trigger a series of symptoms, such as fever, cough, fatigue and mild to severe respiratory complications [Yan et al. 2020]. With the large spread of SARS-CoV-2, the World Health Organization (WHO) declared a pandemic state on March 11, 2020 [WHO 2021]. According to recent statistics (April 2023), more than 764 million people have been infected and more than 6.9 million have died from COVID-19 worldwide. Considering South America, Brazil appears as the country with the most deaths. Although there is currently a general understanding that the pandemic is under control, uncertainty regarding new pandemics is still a constant concern. Scenarios like these pose great challenges for health systems, especially regarding the clinical decision-making process [White and Lo 2020]. Discussing and understanding strategies that support the decision-making process about care rationing is essential, especially in a pandemic context. Eventually, a right decision can contribute to reducing the number of deaths in situations similar to COVID-19.

Care rationing demands a complex screening process, which can influence, the quality of care to the lethality and mortality rates. For this, biomarkers of effective prognosis could be applied. The purpose of this screening process is to help determine patients who require immediate medical attention, based on the estimation of the associated mortality risk in a data-driven approach. Although this stratification process is not ideal, in many situations it becomes necessary, due to the scarcity of hospital resources, whether human or technical. Estimating the risk of death would allow early intervention and potentially reduce mortality since attention would be directed to patients similarly critical but with higher chances of death. To this end, current literature has identified different

clinical characteristics associated with the severity of COVID-19 infection, especially of citizens from Wuhan [Xie et al. 2020; Pan et al. 2020]. Therefore, these characteristics could be used to make a prognosis, especially based on Artificial Intelligence (AI) methods [Kumar et al. 2020].

AI systems have been applied to assist in the diagnosis and prognostication of many conditions [Yu et al. 2018; Soares et al. 2021]. In general, health systems are among the most promising fields for AI applications, mainly with sophisticated methods from the subarea called Machine Learning (ML). Nevertheless, the way AI systems make decisions may not be known, given the "black box" characteristic of some methods. For many algorithms, while achieving effective results, the recommendations are not easily interpretable or explainable. It is not always clear which information or specific reasoning was used by the system to reach a certain decision. Quite frequently, these constraints become barriers to the a broader adoption of ML solutions [Mittelstadt et al. 2019]. This becomes even more critical when it comes to healthcare systems, especially for critical decision-making, which commonly affects the lives of patients. In the context of COVID-19, recent studies have proposed the application AI-based methods, for example, patient mortality prediction [Yan et al. 2020]. Nevertheless, many of these works were carried out only in the Wuhan region using a restricted set of models, and in general they did not assess the interpretability of the decisions at the local/individual level.

In this context, the aim of this work is to propose and experimentally validate a pipeline of ML models to support computer-aided prognostication of patients in a pandemic context, like COVID-19. To achieve it, multiple predictive variables from individual information are exploited, such as sex, age, symptoms, among others. In addition, to better understand the relationship between the predictive variables and the model decisions, we enhance the machine learning decision support system with interpretability assets based on LIME technique [Ribeiro et al. 2016]. It should be noted that this study is a significant extension of a previous work [Figuerêdo et al. 2021]. We add new machine learning algorithms and added interpretability assets based on LIME in order to assist during the decision-making process.

## 2. RELATED WORKS

[Yan et al. 2020] developed predictive models to perform COVID-19 prognosis prediction according to predictive biomarkers. To support the methods, the authors used epidemiological, demographic, clinical and laboratory data. The model discovery relied on data from 375 patients from the city of Wuhan. The predictive model was built using the XGBoost algorithm. The experimental results indicated that the model managed to select three biomarkers which were enough to predict the mortality of individual patients with an accuracy of 90%. Specifically, the most predictive biomarkers were: lactic dehydrogenase (LDH), highly sensitive C-reactive protein and lymphocytes. Using also data from patients in the city of Wuhan, [Xie et al. 2020] conducted a retrospective study to assess the association between hypoxemia and mortality in patients with COVID-19 in a survival analysis. Numerous relevant results have been found, among them is the fact that hypoxemia is independently associated with in-hospital mortality. In addition, the researchers found that oxygen saturation values ($SpO_2$) greater than 90% with oxygen supplementation indicate a high probability of survival.

In [Souza et al. 2020], the authors conducted a study similar to those previously described, but using data from the state of Espírito Santo – Brazil. In addition to the geographical difference between the data used in these works, there were also differences regarding the sources of the data, such as the absence of data regarding laboratory tests, factors that are known to increase the model's effectiveness. To determine the prognosis in patients with COVID-19, the authors used numerous machine learning algorithms. For model construction purposes, data from clinical records of 13,690 patients (cases closed due to cure or death) were used. The experiments performed by the authors revealed that the outcome by COVID-19 could be predicted with an a ROC-AUC of 0.92. Likewise, using data from patients in Brazil (national scale), [Mattos et al. 2020] assessed the correlation between the manifestation of symptoms/comorbidities and the patients' survival response through Kaplan-Meier survival estimates.

The authors identified that the observed comorbidities and symptoms are in accordance with the main clinical markers of the disease already reported in the literature. In addition, the authors also identified that such clinical aspects may present different distributions of comorbidities and present symptoms differently from the results reported in patients from other countries.

Although our work has a similar objective to the works already mentioned, this work presents some significant contributions and innovations. For example, the works in [Yan et al. 2020; Xie et al. 2020] were done mostly in Wuhan, China, while ours was done with data from patients in Brazil. On the other hand, the work carried out by [Souza et al. 2020] was also conducted in Brazil, but used a limited database, containing only data from a single Brazilian state. Despite [Mattos et al. 2020] using a more comprehensive dataset than [Souza et al. 2020], the main objective was to perform an initial analysis of clinical factors related to admission in ICU or death of SARS-CoV-2, and not the development of predictive models from ML. In addition to these aforementioned remarks, the previous works included no explicit resources and analysis for explaining why the model made a particular decision. Differently, our work explicitly introduce an Explainable Artificial Intelligence (XAI) step to the predictions, with the objective of helping in the understanding of the decisions made by the model. Consequently, it helps the experts to comprehend the context and reliability for each prediction.

## 3. EXPERIMENTAL PIPELINE

The experimental process followed in this work is illustrated in Figure 1. There are four stages: data collection, preprocessing, model training (including optimization and validation) and model assessment and analysis, which includes an explainability stage of the learned models.

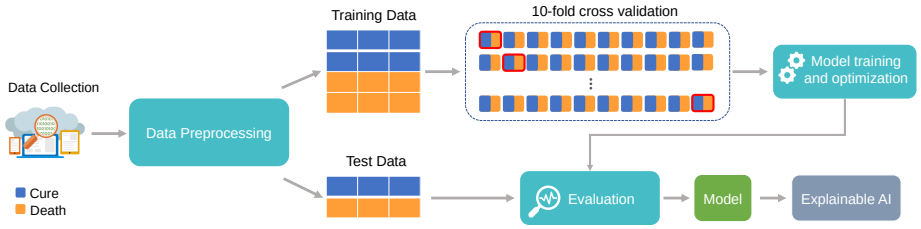


Fig. 1: Experimental process used in this study

### 3.1 Dataset and preprocessing

The experiments relied on the same database used in a previous work [Figuerêdo et al. 2021], i.e., the *Database for Severe Acute Respiratory Syndrome 2020* (SARS2020), available in the OpenDATASUS[1] portal. Such repository is maintained by the Ministry of Health of Brazil, through the Secretariat of Health Surveillance, which conducts the surveillance of Severe Acute Respiratory Syndrome in Brazil, since 2009. Previous to the publication in the portal, the database is submitted to preparation procedures that include anonymization procedures in compliance with current regulations. With the new Coronavirus pandemic, multiple data on COVID-19 cases were incorporated into the surveillance network and are updated on a weekly basis.

In this study, only completed cases (death or cure) were considered. Thus, data were discarded for the patients whose outcome was reported as unknown. After removing these cases, the database remained with 274,493 patient records, 164,535 of cure (59.94%) and 109,958 of death (40.06%). In addition to information regarding the evolution of the infection, the dataset also includes basic individual information, such as gender and age group, symptoms, comorbidities, among others. To

[1] `https://opendatasus.saude.gov.br/dataset/srag-2020`

carry out the experiments, the dataset was randomly divided into training and test sets (further details are provided in Section 3.2).

The original dataset has 156 attributes. However, a preliminary analysis showed that some of these variables do not add relevant predictive content, e.g., patient name. For this reason, some variables considered non-relevant for the task were removed, resulting in a final set of 39 variables (including the outcome attribute)[2]. Moreover, in the preprocessing phase, it was detected that the database had a large amount of missing data. Some attributes such as "Obesity" and "Kidney disease", for instance, had more than 60% absence of data. Thus, in order to avoid possible inconsistencies in the experiments, data standardization was performed considering missing data as non-occurrence of the event in particular (e.g., for the cough attribute, if the information was missing, it was indicated as the patient not having this symptom). With the exception of the "Age" attribute, all other variables were categorical. Thus, the "Age" attribute was discretized into the following categories: child (0-10), teenager (11-17), young adult (18-29), average adult (30-40), adult (41-59) and elderly (60 or more).

### 3.2 Experiments, assessment, and explanations

The database was partitioned into training and test sets. Before partitioning, a sub-sampling was applied to balance the dataset (cures and deaths). Absolute partitioning was performed through a stratified random procedure. For the test set, 30,000 records were randomly hold-out (15,000 cures and 15,000 deaths) and the remainder (202,164 samples) was retained only as the training set. The training set was used to discover the best predictive model supported by a hyperparameter optimization [3] with k-fold cross-validation and stratified sampling with $k = 10$. In turn, the test set was used to verify the effectiveness and perform the explanation analysis of the models. Five algorithms were used in this study, namely: Decision Tree, Logistic Regression, Naive Bayes, Random forest and Gradient Boosted Trees. The predictive models were assessed using the Area Under the ROC Curve (AUC). The AUC provides a numerical summary for the two-dimensional area below the ROC curve. The AUC varies between 0 and 1, with 0 representing a model that provides all predictions erroneously, while 1 represents a models that provides 100% of correct predictions. The effectiveness of the developed models was also evaluated based on classical ML measures, such as Precision, Recall and $F_1$.

Understanding how features affect the decision-making is considerably important for the confidence on the model. This may be decisive to select which model to be deployed or to support further actions based on model predictions [Ribeiro et al. 2016]. When using ML, especially in the healthcare, actions may not be taken based only on predictions from a black box oracle, as the consequences can be catastrophic. Thus, in addition to the traditional effectiveness assessment, we also evaluate the models using a ML explanation technique, named LIME. LIME is a technique that aims at explaining the individual predictions of a black box model by training a local surrogate model that is easier to understand (e.g., a linear model) [Ribeiro et al. 2016]. The rationale behind this approach is that a globally nonlinear model might actually be linear within a small local region of the feature space. To provide this, LIME creates a dataset of perturbed samples for a single sample of interest, predicts it with the black box model and then learns a local surrogate, which approximates the predictions of the black box model. Figure 2 (adapted from [Ribeiro et al. 2016].) illustrates the process of explaining individual predictions. This example shows a situation in which the model predicts that a patient has flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are stated as contributing to the "flu" prediction, while "no fatigue" is evidence against it. Hence, a doctor can decide whether to trust the model's prediction [Ribeiro et al. 2016]. Therefore, an explanation, as the case illustrated in Figure 4, corresponds to a small list of symptoms with relative weights that either contribute to the prediction (in orange) or are evidence against it (in blue). In this work, we considered a sample set of 8 records to evaluate the individual predictions, thus

[2]The list of the variables used can be found https://doi.org/10.6084/m9.figshare.22757123.v1
[3]The hyperparameters tested for each of the algorithms are available in https://doi.org/10.6084/m9.figshare.22756847.v1

simulating the practical application of the predictive models. It is worth noting that this explanation of the model not necessarily means causality, and such investigations are out of the scope of this work.

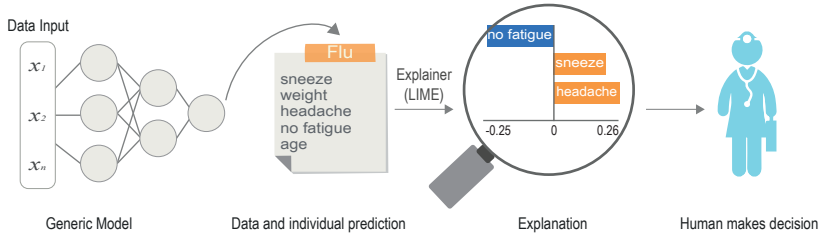


Fig. 2: Explaining individual predictions.

## 4. RESULTS AND DISCUSSIONS

In Figure 3, we present the contingency matrices for the developed models. Considering specifically True Positive (TP) and True Negative (FN) (in blue), which are strongly related to the main class of interest (i.e., death cases), it was observed that the models obtained from GB and RF were the ones that showed the best effectiveness. Considering a scenario in which a large number of patients would be evaluated by predictive models, the GB-based and RF-based would be the most suitable, as the system would make more assertive decisions regarding the identification of patients with greater chances of death, those who would therefore need intervention as soon as possible.

Still considering Figure 3, we can notice a balance between the number of False Positive (FP) and False Negative (FN) (in orange). Considering a scenario of practical use, the ideal would be the absence of errors. However, depending on the purpose of the application, some errors can have a greater negative impact than others. In our context, a FN is especially critical given it represents a patient that would not be identified as at risk and possibly not properly treated. For this scenario, the GB-based and RF-based models would be the most suitable for use, as they reached a lower FN percentage than the others. Ultimately, disregarding financial and other indirect burdens, when compared to FN, the occurrence of FP would be less harmful, considering that the patient would be directed to immediate care and submitted to further examination. However, it is important to highlight that this decision could overburden the hospital more quickly, which could deteriorate the care of patients that in fact need assistance.

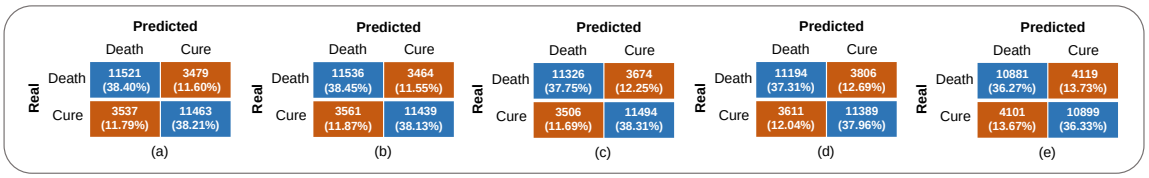


Fig. 3: Contingency matrix for the models developed. (a) Gradient Boosted Trees; (b) Random Forest; (c) Logistic Regression; (d) Decision Tree; and (e) Naive Bayes.

Table I shows the effectiveness of the models developed, considering the Recall, Precision, $F_1$ and AUC measures. In general, the models achieved promising effectiveness, especially the GB and RF. Among the models developed, the NB achieved the worst effectiveness. It is worth mentioning that the classification process used a standard probability threshold of 0.5. The AUC values achieved (above 80% for all cases), show that the models developed were able to obtain high and promising predictive effectiveness. In summary, these results point to the possible effective use of machine learning models to face current and similar problems as those imposed by the COVID-19 pandemic.

Table I: Prediction effectiveness of the developed models in terms of Recall, Precision and $F_1$.

| Algorithm/Classifier | Recall | Precision | $F_1$ | AUC |
|---|---|---|---|---|
| Gradient Boosted Trees | 0.7661 | 0.7661 | 0.7661 | 0.8524 |
| Random Forest | 0.7691 | 0.7641 | 0.7666 | 0.8471 |
| Logistic Regression | 0.7607 | 0.7607 | 0.7607 | 0.8438 |
| Decision Tree | 0.7528 | 0.7528 | 0.7528 | 0.8382 |
| Naíve Bayes | 0.7260 | 0.7260 | 0.7260 | 0.8114 |

These results described in this paper become even more relevant, considering a high percentage of missing data had to be handled. In addition, it is worth mentioning that among the attributes used to characterize users, there was no data from clinical tests and laboratory tests. Such kind o data would possibly contribute to the process of class separation, thus improving the process of identifying cases with greater chances of death, consequently increasing the effectiveness of the models.

For model explanation, LIME was applied over the GB models, since it achieved the greatest global effectiveness in terms of AUC. In addition, among the algorithms used, the GB is also the one that most represent a "black box" algorithm. In the application of LIME, we limit the number of features to 10 in the process of explaining predictions. Although this value represents less than 50% of the total variables of the dataset, it is highly recommended to use a reduced number of features, otherwise, the explanations could be difficult to understand. Explanations of the individual predictions are illustrated in Figure 4. The explanation illustration regards 8 records selected at random, including 2 cases from each prediction assessment category, i.e., TP, TN, FN, and FP. For these cases, the explanations correspond to a list of 10 features with relative weights - features that contribute to the prediction (in orange) or are evidence against it (in blue). These explanations can be used to help an expert in the decision-making process. The specialist, with knowledge of the domain, can use the provided explanations to accept (trust) or reject a prediction by more clearly understanding the reasoning behind it.

With LIME, an individualized analysis is performed for each patient. For instance, considering TP-Case 1 in Figure 4 (a), we observe that the use of invasive ventilatory support contributed positively to the prediction of death. Likewise, the age of 61 was also positively correlated. On the other hand, the fact that the patient is hospitalized and is white is presented as not contributing to the prediction of death. It also suggests that a greater number of positive correlation features led the model to predict the case as in risk of death. Thus, when observing these data, a specialist in the field of application could perceive, for example, that although this patient is hospitalized, a number of features commonly related to the patient's death are shown as contributing to the prediction of death. With these decision-support resources, the specialist may take the decision made by the model as reliable and more probably correct.

Figure 4 (b) illustrates the cases of true negatives. Analyzing TN-Case 2, we found that the fact the patient feels respiratory discomfort and is brown contributes positively to the prediction of death. On the other hand, the facts that the patient is 10 years old, is not in an ICU, and uses non-invasive ventilatory support are pointed out as not contributing to the prediction of death. The contributions against the prediction of death outweigh the correlated contributions. Therefore, it suggests why the system indicates the patient's cure. When confronted with these explanations, a professional with domain knowledge could verify that the decision made by the system is consistent.

The cases described above represented two situations in which the system correctly predicted the evolution of the patient's health status. Although this situation is ideal, the system is susceptible to errors, generating false negatives or false positives. Figure 4 (c) shows cases of false negatives. Many features are presented as not contributing to death, which outweighs the features that contribute to death, leading the system to mistakenly predict the patient's cure. However, it is important to carefully analyze this decision. Although in both cases the contributing features against death outweigh the

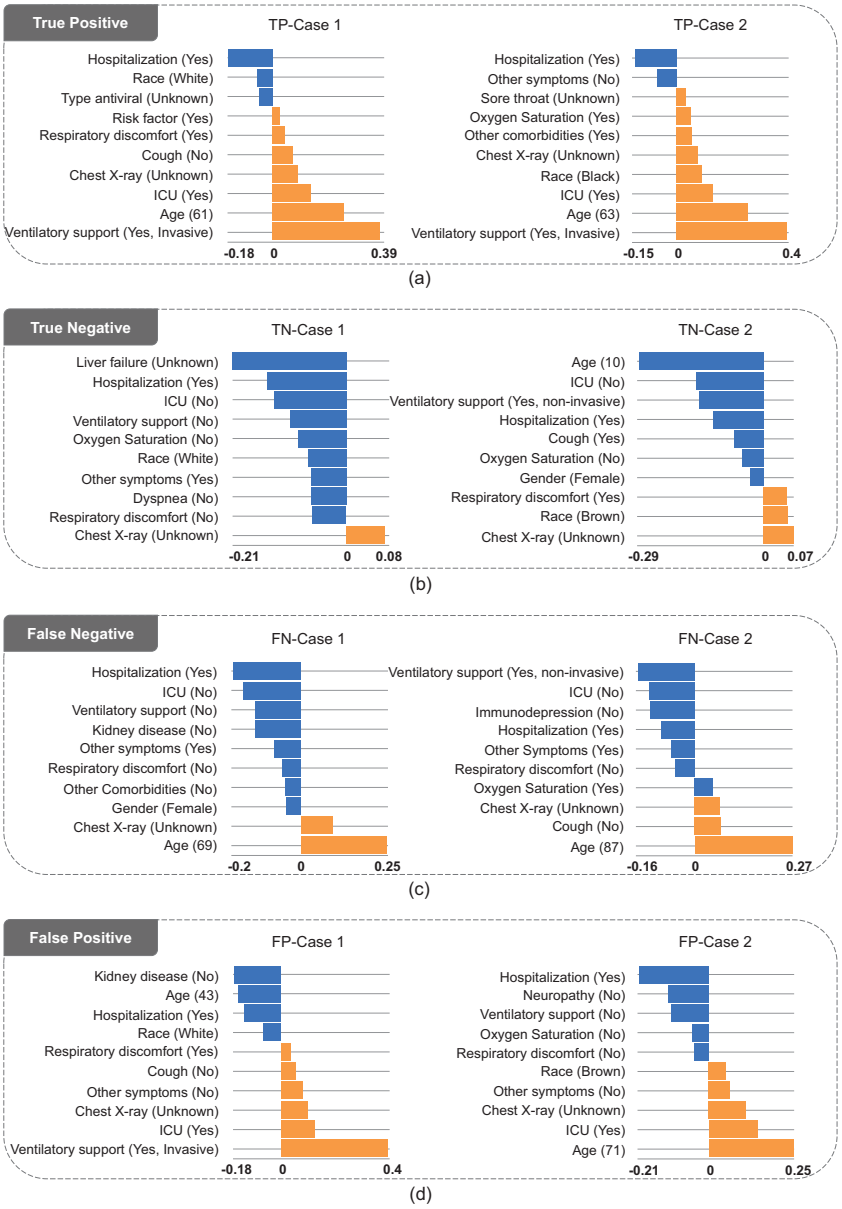


Fig. 4: Explaining individual predictions. (a) TP cases; (b) TN cases; (c) FN cases; (d) FP cases.

features that contribute in favor, these are elderly patients. That is, these are patients belonging to the known risk group. In addition, this feature presented a reasonably high death contribution weight. Hence, the professional's knowledge is still crucial in the decision-making process. The health professional can link their knowledge to the explanations provided, to verify that the prediction is possibly inaccurate.

Figure 4 (d) presents explanations for predictions that generated false positives. Taking the prediction on FP-Case 1, notice that the contributions towards a death prediction are higher than the opposite, especially the ventilatory support attribute. However, it is worth noting that this particular patient has some features that do not contribute to the death and must be better analyzed. For example, the patient has no kidney disease, is hospitalized, and is not an elderly patient. Hence,

a professional with domain knowledge would more carefully consider their decision from these data. After all, even though there are features indicating a strong relationship with death, some factors, such as age, could lead the professional to identify that prediction is possibly misleading. Ultimately, for this particular case, with hospital resources available, healthcare professionals could rely on the system's response and continue with the care for this patient.

## 5. CONCLUSIONS

The experiments carried out indicate that the model developed is capable of predicting patients' prognosis, with the model obtained with GB as the most effective. The GB model reached $ROC - AUC = 0.8524$. Using the LIME ML model explainability technique, we illustrate for a sample of patients, how each feature influences decision-making, showing whether the feature correlated negatively or positively with the prediction provided by the model. In summary, the results showed the potential of using this technique as a strategy to increase users' confidence in the models, refine the decision-making process, and increase its transparency, and, therefore, enable wider adoption. Although this work was developed considering the context of COVID-19, the procedures performed could be replicated in similar health-related contexts. In future work we intend to evaluate deep learning methods, also with a larger amount of data, as it is continuously updated. With this, it would be possible to verify, among other aspects, whether the risk factors have remained the same over time and if more accurate predictions emerge from more data or more complex models. Additionally, for a better understanding of the predictions, complementary explainability techniques may be integrated.

## ACKNOWLEDGMENTS

This work was partially supported by UEFS AUXPPG 2023 and CAPES PROAP 2023 grants.

## REFERENCES

Figuerêdo, J. et al. Machine learning for prognosis of patients with covid-19: An early days analysis. In *Anais do XVIII ENIAC*. SBC, Porto Alegre, RS, Brasil, pp. 59–70, 2021.

Kumar, A. et al. A review of modern technologies for tackling covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14 (4): 569 – 573, 2020.

Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 395 (10224): 565–574, 2020.

Mattos, J. et al. Clinical risk factors of icu & fatal covid-19 cases in brazil. In *Anais do VIII KDMiLe*. SBC, Porto Alegre, RS, Brasil, pp. 33–40, 2020.

Mittelstadt, B. et al. Explaining explanations in ai. In *Proceedings of the Conference on FAT*. ACM, New York, NY, USA, pp. 279–288, 2019.

Pan, D. et al. A predicting nomogram for mortality in patients with covid-19. *Frontiers in Public Health* vol. 8, pp. 461, 2020.

Ribeiro, M. T. et al. "why should i trust you": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*. ACM, New York, USA, pp. 1135–1144, 2016.

Soares, F. et al. Analysis and prediction of childhood pneumonia deaths using machine learning algorithms. In *Anais do IX KDMiLe*. SBC, Porto Alegre, RS, Brasil, pp. 16–23, 2021.

Souza, F. S. H. et al. Predicting the disease outcome in covid-19 positive patients through machine learning: a retrospective cohort study with brazilian data. *medRxiv*, 2020.

White, D. B. and Lo, B. A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic. *JAMA* 323 (18): 1773–1774, 05, 2020.

WHO. Coronavirus disease 2019 Situation Report. `https://covid19.who.int/`, 2021. Accessed 01 April 2021.

Xie, J. et al. Association Between Hypoxemia and Mortality in Patients With COVID-19. *Mayo Clinic Proceedings* 95 (6): 1138–1147, jun, 2020.

Yan, L. et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* 2 (5): 283–288, may, 2020.

Yu, K.-H. et al. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2 (10): 719–731, Oct, 2018.