Machine Learning Prediction of Hospitalization due to COVID-19 based on Self-Reported Symptoms: A Study for Brazil

Igor Miranda
Division of Computer Engineering
*UFRB**
Cruz das Almas, Brazil
igordantas@ufrb.edu.br

Gildeberto Cardoso
Division of Electrical Engineering
UFRB
Cruz das Almas, Brazil
gildeberto@ufrb.edu.br

Madhurananda Pahar
Department of Electrical Engineering
University of Stellenbosch
Stellenbosch, South Africa
mpahar@sun.ac.za

Gabriel Oliveira
Division of Computer Engineering
UFRB
Cruz das Almas, Brazil
sogabris@gmail.com

Thomas Niesler
Department of Electrical Engineering
University of Stellenbosch
Stellenbosch, South Africa
trn@sun.ac.za

***Abstract*—Predicting the need for hospitalization due to COVID-19 may help patients to seek timely treatment and assist health professionals to monitor cases and allocate resources. We investigate the use of machine learning algorithms to predict the risk of hospitalization due to COVID-19 using the patient's medical history and self-reported symptoms, regardless of the period in which they occurred. Three datasets containing information regarding 217,580 patients from three different states in Brazil have been used. Decision trees, neural networks, and support vector machines were evaluated, achieving accuracies between 79.1% to 84.7%. Our analysis shows that better performance is achieved in Brazilian states ranked more highly in terms of the official human development index (HDI), suggesting that health facilities with better infrastructure generate data that is less noisy. One of the models developed in this study has been incorporated into a mobile app that is available for public use.**

***Index Terms*—COVID-19, hospitalization prediction, self-reported symptoms, machine learning.**

I. Introduction

The rapid propagation of the SARS-CoV-2 virus and the possible need for hospitalization by patients who develop COVID-19 have overloaded healthcare systems in several countries around the world. The shortage of hospital beds, especially those in intensive care units (ICU), has been one of the main challenges to fight this disease, affecting medical and governmental decision making [1], [2].

The number of hospitalized cases has been widely adopted as a metric with which to estimate the required resources for health facilities and to define lockdown restriction levels [3]. Furthermore, machine learning tools have been used to predict the number of new as well as hospitalized cases a few weeks ahead, helping local authorities to make informed decisions [4], [5].

*UFRB - Universidade Federal do Recôncavo da Bahia

A tool to estimate the risk of hospitalization might be useful from an individual perspective by helping a patient seek treatment in time. It could also support health professionals in remote locations and in under-resourced environments to make decisions related to patient transfer and bed allocation. Point-of-care prediction methods are well suited to these cases, as they can provide practical and cost-effective strategy. In this regard, smartphone-based diagnostic and data collection tools have particular appeal [6]–[8].

Based on data collected between March and June 2020, Jehi et al. [9] used logistic regression with the least absolute shrinkage and selection operator (LASSO) to predict the risk of hospitalization for 4,536 patients with COVID-19, achieving a sensitivity of 76.9% and specificity of 72.6%. Although the study was conducted at the beginning of the pandemic and used a limited dataset, it provides a clear indication that machine learning has potential in the prediction of hospitalization due to COVID-19.

The method introduced by Sudre et al. [10] predicts hospitalization based on self-reported symptoms informed throughout 9 days. Their work uses clustering to automatically group patients into six distinct groups according to symptoms. It was noted that patients who experienced a similar level of COVID-19 severity fell into the same cluster and that the risk of hospitalization was high for patients in two of the six clusters. When using 2, 5, and 9 days of continuously reported data, a precision of 48.0%, 70.4% and 84.9% and a recall of 47.2%, 70.3% and 84.6% was achieved respectively.

Chen et al. [11] and Jimenez-Solem at al. [12] used random forests (RF) to distinguish severe and non-severe cases of COVID-19 in 362 and 3944 patients, respectively. Chen et al. considered severe cases to be patients with a respiratory rate above 30 breaths per minute or an oxygen saturation below

TABLE I: Characteristics of the study population. Datasets were compiled from open COVID-19 data provided by three state Departments of Health in Brazil, as indicated. Only laboratory-confirmed COVID-19 cases have been selected. Patients with less than two symptoms were removed. The results are presented as means and standard deviation (in parentheses) for age and as percentage values for the other features.

	Brazilian state of origin					
	Alagoas (AL)		**Espírito Santo (ES)**		**Santa Catarina (SC)**	
	Hosp.	Non-hosp.	Hosp.	Non-hosp.	Hosp.	Non-hosp.
Number	1846	47301	3075	74482	5842	85034
Female (%)	43.0	56.9	43.9	57.6	41.0	54.6
Age (years)	62.1 (18.0)	39.8 (16.0)	62.6 (17.7)	40.9 (16.5)	57.6 (17.2)	39.0 (15.6)
Race (%)						
Black/Indigenous/Mixed	77.1	76.2	52.1	47.8	-	-
White	11.2	17.9	32.5	37.5		
Unspecified	11.7	5.9	15.4	14.7		
Comorbidities (%)	57.0	40.6	69.0	28.4	-	-
Body pain/tiredness (%)	32.1	38.6	-	-	16.4	32.9
Breathing difficulties (%)	21.5	7.0	71.4	23.2	83.8	22.1
Coughing (%)	84.3	74.6	77.4	72.3	88.3	73.8
Diarrea (%)	-	-	15.3	21.7	20.1	13.5
Fever (%)	81.4	73.2	72.8	61.8	72.8	55.1
Headache (%)	20.3	52.8	31.2	69.2	6.5	17.4
Blocked/running nose (%)	-	-	23.2	50.1	2.3	8.4
Sore throat (%)	5.3	19.3	14.4	40.4	32.3	54.5

93% or a partial pressure of oxygen below 300mmHg. Using only the patients' comorbidity and symptoms, the system achieved an approximately 90% predictive accuracy. When this data is combined with laboratory test results (not including COVID-19 specific tests), the accuracy rose to 99%. In the study conducted by Jimenez-Solem at al., an area under the receiver operating characteristic curve (AUC) of 0.818 was reported for hospital admission.

Although the results mentioned above for hospitalization or severity prediction are promising, there is still room for improvement with regards to the number of days over which symptoms should be reported, the study population, and the labeling criteria. Furthermore, the definition of the severe cases used by [11] differs from the classification provided by the World Health Organization (WHO) [13], which may affect the system's performance in the context of the current clinical management of COVID-19 cases in most countries. Additionally, it has been recommended by the WHO that the decision for hospitalization should be made on a case-by-case basis, considering not only the clinical presentation but also the patient's demographics (age, sex, and medical history), risk factors, and even the conditions at home [13].

The current work aims to investigate the success of machine learning-based methods in estimating the risk of hospitalization due to COVID-19, using patient's medical history and self-reported symptoms, regardless of the period in which they occurred. Our final goal is to conceive a classification methodology that can be implemented as a smartphone-based solution for the self-assessment of COVID-19 severity.

We train and evaluate machine learning algorithms using official hospitalization data released by the government of Brazil so that the results can be compared with current medical practices. Brazil is one of the countries hardest hit by the Coronavirus epidemic, reaching 370,000 deaths by the beginning of April 2021.

II. Methods

A. Dataset acquisition

According to Brazilian law, it is mandatory for hospitals and other health facilities to report all disease cases to the government. During the COVID-19 pandemic, state departments of public health in Brazil have periodically released notifications of COVID-19 cases, which include anonymized patient data regarding age, gender, symptoms, previous health conditions, and other information. These data were collected during screening and eventually updated for the hospitalized patients.

To avoid social or racial bias during model training, we have considered databases from states located in Brazilian macroregions that are clearly distinguished from each other in terms of human development index. Additionally, only databases that include the diagnosis method and the case management (whether hospitalized or not) were considered.

The data provided by the Brazilian states of Alagoas (AL) [14], Espírito Santo (ES) [15], and Santa Catarina (SC) [16] met the criteria mentioned above and were selected for this study. These states respectively occupy 27^{th} (last), 9^{th} and 3^{rd} positions in the state-wise human development index (HDI)

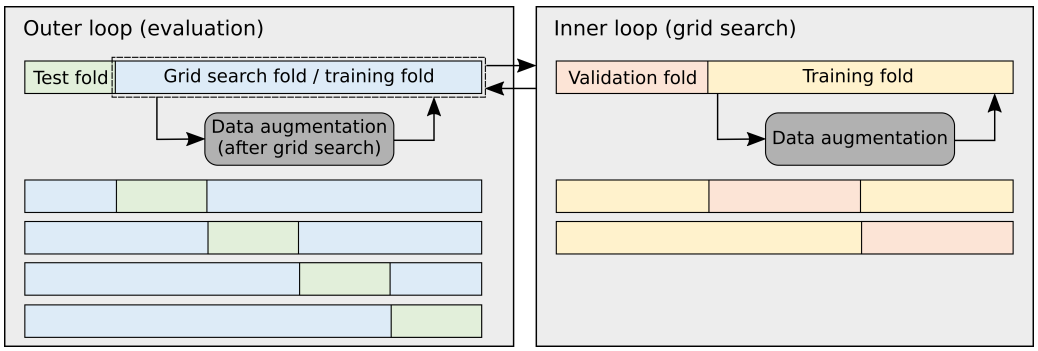


Fig. 1: Hyperparameter optimisation, training and validation scheme adopted in this study which combines nested cross-validation and data augmentation. This process has been used for each combination of machine learning algorithm and dataset.

ranking in Brazil [17]. The datasets contain data collected from March to December of 2020 in both public and private hospitals and healthcare facilities. We further removed patients with fewer than two symptoms and those without laboratory confirmation. The composition of the extracted data is shown in Table I.

Some features are not available in the AL and ES datasets because the associated symptoms were not reported. In the SC dataset, comorbidity is annotated for 51.0% of hospitalized patients and 0.1% of non-hospitalized patients. This low incidence among non-hospitalized patients is inconsistent with previous research [9], [11] and with the AL and ES datasets, which indicate the comorbidity in 40.6% and 28.4% of non-hospitalized patients respectively. For this reason, the comorbidity feature was removed from the SC dataset.

The Brazilian Institute of Geography and Statistics (IBGE) officialy classifies the Brazilian population into five racial groups: *Branco* (White), *Preto* (Black), *Amarelo* (East Asian), *Indígena* (Indigenous), or *Pardo* (mixed race) [18]. However, we considered black, indigenous, and mixed race populations as a single group since self-identification has in these cases been reported to be imprecise, with many black and indigenous persons identifying themselves as mixed race [19].

B. Experimental setup

For the proposed study, we investigated the predictive performance of three machine learning algorithms: decision trees (DT), neural networks (NN), and support vector machines (SVM). Hyperparameter optimization was performed for each of these techniques using nested cross-validation combined with data augmentation as illustrated in Fig. 1. The AUC was used as a performance measure. As datasets differ in terms of the feature they contain, this cross-validation was performed independently for each combination of machine learning algorithm and dataset.

For DTs, the maximum depth of the tree and the employed quality criterion were the hyperparameters optimised during cross-validation. For NNs, the network architecture and the activation function were the variable hyperparameters. The L2 regularization penalty was the only hyperparameter for the linear kernel SVM classifier. Table II summarizes the ranges considered for these hyperparameters.

TABLE II: Values and configurations considered for hyperparameter optimization.

DT	Maximum depth of the tree	5 to 40 with step of 5
	Quality criteria	Information gain, Gini impurity
NN	Number of hidden layers	1 to 3
	Neurons per hidden layer	32, 64, 128 neurons
	Activation function	Tanh, ReLu
SVM	L2 regularization penalty	0.1, 1, 10

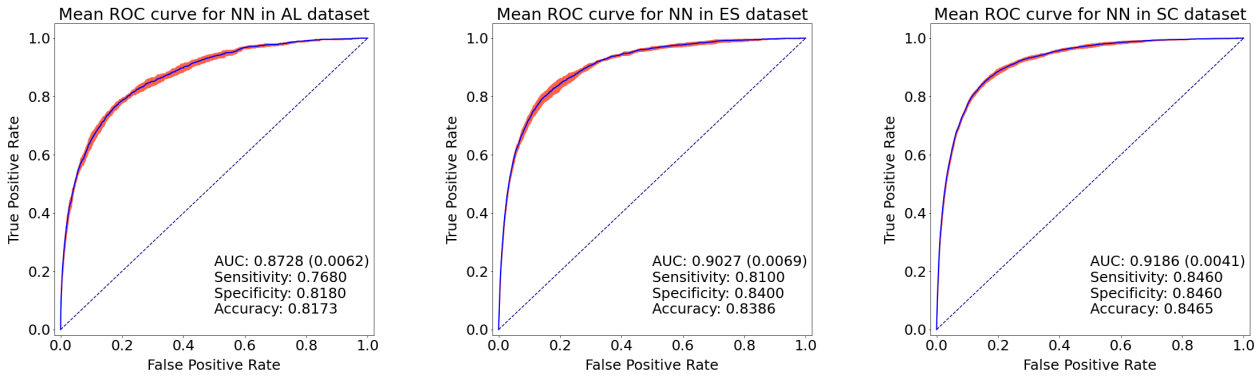
Class imbalance is an important issue for the three datasets, as pointed out in Section II-A. To mitigate this, we used stratified cross-validation and data augmentation within the scheme shown in Fig. 1. A stratified k-fold split [20] was used for both outer and inner loops. Synthetic minority oversampling (SMOTE), a data augmentation technique, was applied to the training folds in the outer loop [21]. No synthesized data was included in the outer loop test folds.

The comorbidity feature was an integer value corresponding to the number of comorbidities. For normalization purposes, the age feature was the actual age divided by 100. All the other features were considered to be dichotomous traits with 0/1 binary values.

We anticipated the potential for algorithmic bias given that previous studies have found evidence of racial health inequity in the context of COVID-19 in Brazil [22], [23]. Therefore, race was used as shown in Table I to assess bias and was performed using the AL and ES datasets. To achieve this, three 5-fold stratified cross-validation experiments were performed for: all patients; the white group; and the black/indigenous/mixed race group.

III. Results and Discussion

Table III shows the performance metrics for each cross-validation experiment. It can be seen that NN and SVM achieved the best results, without a clear advantage for either. Fig. 2 shows the receiver operating characteristic (ROC) of the experiments using NN.



(a) (b) (c)

Fig. 2: Mean ROC curves and evaluation metrics for cross-validation performed on each dataset using neural networks. The mean AUC and the standard deviation (in parenthesis) are computed across validation folds. Mean sensitivity, specificity, and accuracy are calculated using a decision threshold of 0.5. The true positive rate standard deviation is plotted in red.

Using AUC as the evaluation metric, NN performed best for the AL dataset with a mean sensitivity of 76.8% and mean specificity of 81.8%, SVM performed best for the ES dataset with a mean sensitivity of 81.2% and mean specificity of 84.0%, and NN also achieved the better results for the SC dataset with a mean sensitivity of 84.6% and mean specificity of 84.6%. The mean accuracy of the best models ranged from 79.1% to 84.7% across the three datasets. Accuracy, sensitivity, and specificity were calculated using a decision threshold of 0.5 for normalized predicted probabilities.

The hyperparameters selected during cross-validation were found to be similar across the three datasets. Parameters that were frequently found to be optimal were: entropy as quality criterion and a maximum depth of 10 for DTs; one 32-neuron hidden layer architecture using rectified linear unit (ReLU) activation for NNs; and an L2 regularization parameter of 1 for the SVM.

These results are consistent across classifiers for each dataset and across the three datasets for each classifier. This confirms the findings of the previous study [9]–[11] that machine learning can be used to predict hospitalization based on a patient's symptoms and health status with relatively high accuracy. It also shows that the machine learning algorithm adopted may affect the classification performance substantially.

In order to determine whether the classification is influenced by the racial groupings, separate evaluations were performed for the AL and the ES datasets. The results in Table IV show that the performance differences are very small, and suggest that the effectiveness of the system is not influenced by the racial grouping of the patients.

TABLE III: Evaluation metrics for cross-validation performed on each dataset/classifier pair. Accuracy, sensitivity, and specificity are mean values calculated using a threshold of 0.5. AUC standard deviation is given in parenthesis.

Dataset	Classifier	AUC	Sens.	Spec.	Acc.
AL	DT	0.8257 (0.0090)	0.7180	0.7940	0.7911
AL	NN	0.8728 (0.0062)	0.7680	0.8180	0.8173
AL	SVM	0.8643 (0.0087)	0.7600	0.8100	0.8082
ES	DT	0.8696 (0.0074)	0.7700	0.8180	0.8158
ES	NN	0.9027 (0.0069)	0.8100	0.8400	0.8386
ES	SVM	0.9046 (0.0068)	0.8120	0.8400	0.8378
SC	DT	0.8927 (0.0039)	0.8220	0.8300	0.8289
SC	NN	0.9186 (0.0041)	0.8460	0.8460	0.8465
SC	SVM	0.9156 (0.0043)	0.8540	0.8320	0.8339

TABLE IV: Mean AUC of 5-fold cross-validation for all patients, black/indigenous/mixed race (B/I/M) patients and white patients run for all three classifiers using AL and ES datasets.

	Brazilian state of origin					
	Alagoas			Espírito Santo		
	All	B/I/M	White	All	B/I/M	White
DT	0.8272	0.8153	0.8291	0.8651	0.8659	0.8670
NN	0.8712	0.8731	0.8764	0.9041	0.9034	0.9090
SVM	0.8644	0.8655	0.8687	0.9043	0.9044	0.9090

IV. System Availability

A neural network model (one 32-neuron hidden layer with ReLu) for hospitalization prediction using the proposed method has been incorporated in the mobile app ContraCovid, which is available for public download [a]. The app is intended for the self-monitoring of patients suffering from COVID-19 and has used the proposed predictor as one of the metrics to recommend patients to seek treatment.

[a] Available on Google Play.

V. Conclusion

We have evaluated machine learning algorithms to predict hospitalization due to COVID-19 using the patient's self-reported symptoms and previous health status. In contrast to previous studies, we do not restrict the time-frame over which the self-reporting must occur. To conduct this evaluation, we compiled our datasets based on databases officially published by three state departments of health in Brazil. In total, data from 217,580 laboratory-confirmed cases of SARS-CoV-2 patients were used to assess the performance of decision trees, neural networks, and support vector machines.

Since the information reported by the three health departments was not completely the same, independent experiments were performed. Nested cross-validation with data augmentation was applied to each dataset/algorithm pair. Achieved accuracies ranged from 79.1% to 84.7%. Neural networks and support vector machines performed best with neither offering clear advantage over the other.

The effectiveness of each algorithm was shown to be consistent across the three datasets. This suggests that machine learning can predict hospitalization due to COVID-19 using only self-reported symptoms with acceptable accuracy. Based on the official Brazilian state-wise human development index (HDI) ranking [17], the performance for the richest state was best while the performance for the poorest state was worst, with average area under the receiver operating characteristic curve (AUC) varying between 85% and 91%. This may be related to the number of hospital beds available for COVID-19 in each state, which may lead to increased noise in the data.

A comparison between data obtained for different racial groups (where such race information was available in the official data) indicate that the performance of all systems is not influenced by the racial grouping of the patients.

References

[1] J. I. Salluh, T. Lisboa, and F. A. Bozza, "Challenges for the care delivery for critically ill COVID-19 patients in developing countries: the Brazilian perspective," *Critical Care*, vol. 24, no. 1, pp. 1–3, 2020.

[2] J. I. Salluh, G. Burghi, and R. Haniffa, "Intensive care for COVID-19 in low-and middle-income countries: research opportunities and challenges," *Intensive Care Medicine*, pp. 1–4, 2020.

[3] D. O. d. C. Lino, R. Barreto, F. D. d. Souza, C. J. M. d. Lima, and G. B. d. S. Junior, "Impact of lockdown on bed occupancy rate in a referral hospital during the COVID-19 pandemic in northeast Brazil," *Brazilian Journal of Infectious Diseases*, vol. 24, no. 5, pp. 466–469, 2020.

[4] S. Debnath, D. P. Barnaby, K. Coppa, A. Makhnevich, E. J. Kim, S. Chatterjee, V. Tóth, T. J. Levy, M. d Paradis, S. L. Cohen, *et al.*, "Machine learning to assist clinical decision-making during the COVID-19 pandemic," *Bioelectronic Medicine*, vol. 6, no. 1, pp. 1–8, 2020.

[5] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," *Chaos, Solitons & Fractals*, vol. 139, p. 110058, 2020.

[6] R. Braun, C. Catalani, J. Wimbush, and D. Israelski, "Community Health Workers and Mobile Technology: A Systematic Review of the Literature," *PloS one*, vol. 8, no. 6, p. e65772, 2013.

[7] A. Bastawrous and M. J. Armstrong, "Mobile health use in low-and high-income countries: an overview of the peer-reviewed literature," *Journal of the Royal Society of Medicine*, vol. 106, no. 4, pp. 130–142, 2013.

[8] M. Tomlinson, W. Solomon, Y. Singh, T. Doherty, M. Chopra, P. Ijumba, A. C. Tsai, and D. Jackson, "The use of mobile phones as a data collection tool: a report from a household survey in South Africa," *BMC Medical Informatics and Decision Making*, vol. 9, no. 1, pp. 1–8, 2009.

[9] L. Jehi, X. Ji, A. Milinovich, S. Erzurum, A. Merlino, S. Gordon, J. B. Young, and M. W. Kattan, "Development and validation of a model for individualized prediction of hospitalization risk in 4,536 patients with COVID-19," *PloS one*, vol. 15, no. 8, p. e0237419, 2020.

[10] C. H. Sudre, K. Lee, M. N. Lochlainn, T. Varsavsky, B. Murray, M. S. Graham, C. Menni, M. Modat, R. C. Bowyer, L. H. Nguyen, *et al.*, "Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom study app," *MedRxiv*, 2020.

[11] Y. Chen, L. Ouyang, F. S. Bao, Q. Li, L. Han, B. Zhu, Y. Ge, P. Robinson, M. Xu, J. Liu, *et al.*, "An interpretable machine learning framework for accurate severe vs non-severe COVID-19 clinical type classification," *medRxiv*, 2020.

[12] E. Jimenez-Solem, T. S. Petersen, C. Hansen, C. Hansen, C. Lioma, C. Igel, W. Boomsma, O. Krause, S. Lorenzen, R. Selvan, *et al.*, "Developing and validating covid-19 adverse outcome risk prediction models from a bi-national european cohort of 5594 patients," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[13] W. H. Organization, "Clinical management of COVID-19: Interim guidance, 27 May 2020," World Health Organization, Tech. Rep., 2020.

[14] "Painel COVID-19 em Alagoas - Secretaria de Saúde do Estado de Alagoas," http://www.dados.al.gov.br/dataset/painel-covid19-alagoas, accessed: 2020-12-09.

[15] "Painel COVID-19, Secretaria de Saúde do Estado do Espírito Santo," https://coronavirus.es.gov.br/painel-covid-19-es, accessed: 2020-12-10.

[16] "COVID-19 - Casos Confirmados, Base de dados do Governo do Estado de Santa Catarina," http://dados.sc.gov.br/dataset/covid-19-dados-anonimizados-de-casos-confirmados, accessed: 2020-12-11.

[17] Instituto de Pesquisa Econômica Aplicada - IPEA, "Radar IDHM: evolução do IDHM e de seus índices componentes no período de 2012 a 2017," http://www.ipea.gov.br/portal/images/stories/PDFs/livros/livros/190416_rada_IDHM.pdf, 2019, accessed: 2020-12-11.

[18] J. Petruccelli and A. L. Saboia, "Características étnico-raciais da população: Classificação e Identidades," *Rio de Janeiro: Instituto Brasileiro de Geografia e Estatistica - IBGE*, 2013.

[19] S. R. Bailey and E. E. Telles, "Multiracial versus collective Black categories: examining census classification debates in Brazil," *Ethnicities*, vol. 6, no. 1, pp. 74–101, 2006.

[20] N. Diamantidis, D. Karlis, and E. A. Giakoumakis, "Unsupervised stratification of cross-validation for accuracy estimation," *Artificial Intelligence*, vol. 116, no. 1-2, pp. 1–16, 2000.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[22] E. M. Araujo, K. Lilly Caldwell, M. Pereira Alves dos Santos, I. Magalhães de Souza, P. Lima Ferreira Santa Rosa, A. Beatriz Silva dos Santos, and L. E. Batista, "COVID-19 morbimortality by race/skin color/ethnicity: the experience of Brazil and the United States," *Scielo Preprints - Health Sciences*, 2020.

[23] P. Baqui, I. Bica, V. Marra, A. Ercole, and M. van Der Schaar, "Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study," *The Lancet Global Health*, vol. 8, no. 8, pp. e1018–e1026, 2020.