Prediction of Care for Patients in a COVID-19 Pandemic Situation Based on Hematological Parameters

Arianne Sarmento Torcate, Flávio Secco Fonseca, Antônio Ravely T. Lima, Flaviano Palmeira Santos, Tássia D. Muniz S. Oliveira, Maíra Araújo de Santana, Juliana Carneiro Gomes, Clarisse Lins de Lima, Valter Augusto de Freitas Barbosa, Ricardo Emmanuel de Souza, and Wellington P. dos Santos

1 Introduction

In December 2019, the World Health Organization (WHO) received notifications from China regarding cases of pneumonia and severe respiratory syndrome among workers at the seafood market in Wuhan, China, having unknown cause [1]. In March 2020, WHO classified the state of this disease as pandemic, known as COVID-19, which has a high rate of transmissibility, with the ability to spread fast throughout the world [2].

After a year of pandemic, the clinical manifestations of COVID-19 are not fully understood, as there is still limited information to characterize the clinical picture of the disease [3]. However, at the beginning of the pandemic, Brazilian Ministry of Health established the flu syndrome as the most common manifestation of this disease, which is defined as an acute respiratory condition, characterized by a feverish sensation or fever, accompanied by cough, sore throat, runny nose, or difficulty breathing.

The World Health Organization clarifies that the initial signs and symptoms of the disease resemble a common flu-like condition, but can vary from person to person, and may manifest through pneumonia, severe pneumonia, and severe acute

A. S. Torcate · F. S. Fonseca · M. A. de Santana · J. C. Gomes · C. L. de Lima
Graduate Program in Computer Engineering, Polytechnique School of the University of Pernambuco, Recife, Brazil

A. R. T. Lima · F. P. Santos · T. D. M. S. Oliveira · R. E. de Souza · W. P. dos Santos (✉)
Department of Biomedical Engineering, Federal University of Pernambuco, Recife, Pernambuco, Brazil
e-mail: wellington.santos@ufpe.br

V. A. de Freitas Barbosa
Academic Unit of Serra Talhada, Rural Federal University of Pernambuco, Serra Talhada, Brazil

S. K. Pani et al. (eds.), *Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis*,
https://doi.org/10.1007/978-3-030-79753-9_11

respiratory syndrome [4]. In the current context, people with comorbidities such as diabetes, cardiovascular diseases, obesity, hypertension, tuberculosis, and others are at higher risk of rapid worsening of the disease, which can lead to death.

With that in mind, the early detection and management of this disease is essential, especially in the Brazilian context, a country with continental dimensions and diverse territorial realities, with great social inequality and limitations in access to health services by the population. Thus, it is necessary to know and check information about local realities (states and cities) to make decisions in this scenario.

It is worth mentioning that, based on this problem, organizations and researchers from different areas of knowledge have sought answers to questions related to health problems caused by COVID-19. Scientific investigations aim to provide immediate actions that collaborate to control the pandemic, that is, that contribute to assist in clinical, social, or political decision-making, all based on scientific evidence, to maximize the benefits and minimize injuries and costs.

From the existing initiatives to assist the health professional in decision-making, we highlighted in this research the Artificial Intelligence (AI) and biostatistics, which together can predict, for instance, the survivor or risk of patients from their physiological parameters. These predictions allow treatment individualization, and greater chances of complete recovery. Among the subareas of AI, we highlight machine learning, whose algorithms have become one of the most used forms of classification and prediction of patterns in large data nowadays [5].

Thereby, the main goal of this research is to make predictions about the type of hospitalization and severity assessments of patients with and without COVID-19. For this, we used hematological data from patients of the units of the Unified Health System (SUS) in the city of Paudalho, Brazil. The aim is to analyze algorithms that are capable of making hospitalization predictions into one of the three possible choices: regular ward, semi-intensive care unit, and intensive care unit, corresponding to mild (non-critical), moderate, and severe cases. In order to perform this study, seven classic classifiers were applied to the data set, such as Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Tree, Random Forest, Bayesian Networks, and J48 Decision Trees, all with well-defined parameter settings tested along the experiments.

This chapter is structured as follows. After this introduction, Sect. 2 and its respective subsections highlight the theoretical background used in this project. Section 3 presents the works related to the developed research. Section 4 sets out the methodological path adopted, while Sect. 5 describes in detail the obtained results. Section 6 presents a general analysis, building discussions of the results. Final considerations and perspectives for future work are presented in Sect. 7.

2 Theoretical Foundation

In this section, we present the theoretical approaches related to the main topics of this research, with emphasis on COVID-19 and machine learning.

2.1 COVID-19

Characterized as a contagious respiratory disease, COVID-19 is associated with high mortality rates since its emergence in December 2019 [6]. According to the World Health Organization, the coronavirus pandemic is putting even the best health systems worldwide under tremendous pressure [7]. The study by Nemati et al. [6] states that on March 24, 2020, the virus had spread to more than 170 countries, with more than 422,613 confirmed cases and 18,891 deaths. In addition, mortality rates may vary between countries due to demographic differences, age distribution, and healthcare infrastructure.

The symptoms of COVID-19 are similar to a common flu condition, such as malaise, fever, fatigue, cough, mild dyspnea, anorexia, sore throat, body pain, headache, and nasal congestion [3]. However, as this disease progresses, patients feel shortness of breath, nausea culminating in pneumonia, and multiple organ failure. Because of this, the best way so far to prevent the disease is individual protection, such as hand washing, correct use of masks, and social isolation. These protective measures must be aligned with government actions, such as providing more beds in intensive care units, hiring and qualifying frontline professionals, providing basic sanitation, acquiring specialized medical equipment such as mechanical fans, and increasing and decentralizing the performance of rapid tests and vaccination of the population.

The best accepted test to diagnose COVID-19 is the molecular test, such as reverse transcription followed by polymerase chain reaction (RT-PCR), which identifies the SARS-CoV-2 viral RNA. In this type of test, secretions from the nasopharynx are collected. However, according to Iser et al. [3], citing the work developed by Woelfel et al. [8], Tolia et al. [9], and Hadaya et al. [10], these tests should be performed between the third and seventh day of infection. This period guarantees greater precision of the method and reduction of false-negative results. The problem is that in many cases, it is difficult to identify when the patient became infected. In addition, RT-PCR tests are normally performed on health professionals and on symptomatic patients who have been hospitalized, due to the high cost and the scarcity of certified laboratories for its performance. Serologic tests by immunochromatography, known as rapid tests, have become an option for: (1) people with mild to moderate symptoms, without the need for hospitalization; (2) public health system, for tracking asymptomatic cases, epidemiological survey of confirmed cases, and estimating the population's immunization rate. In this case, they should be used 7 days after the onset of symptoms. Unfortunately, rapid tests are non-specific for the detection of virus presence directly.

Briefly, the coronavirus increased the need for immediate clinical decisions and the effective use of health resources, as record pressure was imposed on health systems worldwide. The aim of developing techniques that assist in decision-making is to contribute to the control not only of the pandemic, but of the factors associated with the problem of interest. As a result, scientific investigations, ethical commitment, and the ability to perceive important clinical gaps for characterizing and defining hypotheses have become essential.

2.2 Machine Learning

Machine learning (ML) is a subarea of Artificial Intelligence. From a more technical point of view, ML stems from the difficulty in manually handling a large volume of available data, proposing intelligent systems that can learn the patterns or regularities in the data [12]. One of the objectives of ML techniques is to perform pattern detection in databases [13]. This is possible through data that provides machines with the ability to learn and, from that, recognize patterns and create relationships between variables.

The use of machine learning has become very valuable in several intelligent applications, solving most data-related problems [14]. With the algorithms present in machine learning it is possible to work with hundreds of attributes, either in the detection/use of the interactions between the attributes and thus favor the support to the diagnosis in the health area [15]. Machine learning techniques have been used in activities that mainly involve the identification of patterns [13, 15–20]. These techniques are significantly contributing to the resolution of real problems such as prediction, diagnosis, and recognition of health problems.

In this COVID-19 scenario, prediction has become a priority for public health. Thus, the use of ML algorithms has been useful to health professionals, and can help in several factors, such as for pneumonia detection by analyzing chest X-ray images [16], support for diagnosis through blood tests [13], predictive model to help doctors choose the best therapeutic strategies for patients with COVID-19 [21], prediction of mortality from blood tests [22], and others.

3 Related Works

To contribute in the context of forecasts that estimate the outbreak of COVID-19, the study by Ardabili et al. [23] presents a thorough and comparative analysis of machine learning and soft computing models, both in alternative to the Susceptible-Infected-Recovered (SIR) and Susceptible-Exposed-Infectious-Removed (SEIR) models. The methods were applied to data from five countries (Italy, Germany, Iran, the USA, and China) for the total cases obtained in 30 days. Among a range of investigated ML models, the Multilayer Perceptron (MLP) and the Adaptive Neuro-Fuzzy Inference System (ANFIS) stood out and showed promising results. The authors reinforce the importance of researchers dedicating themselves to investigating predictions to also estimate the number of infected patients, as well as the number of deaths. Thus, ML models must be analyzed and take into account the individual context of each country.

The research carried out by Nemati et al. [6] also contributes in this context by developing predictive models with the ability to predict the length of stay of patients in the hospital. For this, survival characteristics of 1182 patients were considered, where the time of discharge was chosen as a variable of interest, along with survival

analysis techniques, including statistical analysis and seven algorithms machine learning. The results obtained indicate that being male or belonging to older age groups is associated with lower probabilities of hospital discharge. In addition, the Gradient Boosting (GB) survival model surpasses other models for predicting patient survival in the researched context. Stagewise GB, on the other hand, offers the most accurate download time prediction compared to other algorithms. But, Kaplan-Meier Estimator and Cox regression methods suggest that the gender and age of hospitalized patients have a direct effect on recovery time. Finally, all findings are of great relevance to help healthcare professionals make more assertive decisions during the outbreak.

Still analyzing the various possibilities of developing predictive models to contribute in the context of COVID-19, we also highlight the work carried out by Batista et al. [11] at the beginning of the pandemic in Brazil. The work aimed to predict the risk of a positive diagnosis of COVID-19 with machine learning, using data resulting from admission exams in the RT-PCR tests of 235 adult patients, in the emergency room of Hospital Israelita Albert Einstein in São Paulo, from March 17 to 30, 2020. In all, five ML algorithms (neural networks, Random Forests, gradient augmentation trees, logistic regression, and SVM) were used. The results show that the best predictive performance was obtained by the SVM algorithm (AUC: 0.85; Sensitivity: 0.68; Specificity: 0.85; BrierScore: 0.16). The three most important variables for the predictive performance of the algorithm were the number of lymphocytes, leukocytes, and eosinophils, respectively.

Researchers also appear in order to propose diagnoses that are faster and cheaper. An example of this is the work developed by Kumar et al. [7]. A classifier based on ML and Deep Learning using ResNet152 on chest X-ray images of patients with COVID-19 was proposed for prediction of the new coronavirus. This work focused on the prevention of spread of the virus by asymptomatic patients. The authors used the SMOTE technique to balance the data points and then apply the algorithms. The best results were obtained using the Random Forest model, which stood out in precision (0.973), sensitivity (0.974), specificity (0.986), F1 score (0.973), and AUC (0.997). About the predictive model, the XGBoost performed better, with precision (0.977), sensitivity (0.977), specificity (0.988), F1 score (0.977), and AUC (0.998). Both models contribute to the effective clinical prediction of COVID-19.

Gomes et al. [16] proposed an intelligent system to support the diagnosis of COVID-19, investigating radiographs from different databases. Radiographies of patients with viral pneumonia, bacterial pneumonia, and healthy patients were obtained from the Kaggle website. The X-rays of patients with COVID-19 were obtained from four different databases: open source GitHub repository shared by Dr. Joseph Cohen et al. [24]; COVID-19 database, made available online by Societa Italiana di Radiologia Medica e Interventistica [25], and Peshmerga Erbil Hospital database. The authors analyzed the classification performance, using five metrics: accuracy, sensitivity, precision, specificity, and kappa index. The machine learning methods used to classify X-ray images were the Multilayer Perceptron, Support Vector Machine, Decision Trees, Bayesian Network, and Naive Bayes, and all experiments were carried out with the Weka software. The work showed that the

system can diagnose COVID-19 with an average accuracy of 89.78%. Its prototype is already developed, it was able to differentiate COVID-19 from viral and bacterial pneumonia and has low computational cost.

In the work of de Barbosa et al. [13], several experiments are also carried out with machine learning methods, such as MLP, SVM, Random Trees, Random Forest, Bayesian Networks, and Naive Bayes in order to propose an intelligent system with classic classifiers and low computational cost to support the diagnosis COVID-19 based on blood tests. Six metrics were chosen to analyze the classification performance: accuracy, precision, sensitivity, specificity, recall, and precision. The databases were made available by Hospital Israelita Albert Einstein located in São Paulo, Brazil, which are available on the Kaggle platform. Bayes Network was the best method that stood out, being able to achieve high diagnostic performance, with general precision of 95.159% $\pm$ 0.693 of general precision, kappa index of 0.903 $\pm$ 0.014, sensitivity of 0.968 $\pm$ 0.007, precision of 0.938 $\pm$ 0.010, and specificity of 0.936 $\pm$ 0.011. According to the authors, the availability of this software system combined with rapid and low-cost tests, based on blood tests, can be of great help in overcoming the testing challenges that are being faced worldwide.

With the works presented above, it is evident that the machine learning area can be applied to different types of data contributing to the pandemic scenario caused by COVID-19, considering different goals and objectives, either at the global or individual level of each country. For this, this research field has peculiarities, potentialities, and interdisciplinary alignment with other areas. For comparative effect and better understanding, each of the works cited in this section, they were summarized in Table 1, considering the main goal, methods used, and obtained results.

4 Methods

In this study, we evaluated 41 hematological data (blood tests) from patients treated at public healthcare units in the city of Paudalho, Brazil. These 41 tests (Table 2) were the features used as classification input. Patient records covered the period from December 2019 to August 2020. In order to predict hospitalization, we assessed data from three different hospitalization conditions: regular ward, semi-intensive care unit, and intensive care unit. Regular ward refers to regular service or non-critical cases, while semi-intensive care unit corresponds to moderate cases, and intensive care unit is related to severe cases. It is important to mention that all procedures involving human participants were performed in accordance with the 1964 Helsinki declaration and with the ethical standards of the institutional research committee from the Federal University of Pernambuco, registered under number 34932020.3.0000.5208.

The experiments were performed using the WEKA software, version 3.8 [26]. For a better understanding, Fig. 1 illustrates the methodological path used to carry out the experiments.

Table 1 Summary of related works

Authors/Year	Objective	Method	Results
Ardabili et al. [23]	Perform a comparative analysis of machine learning and soft computing models to predict the outbreak of COVID-19	Genetic algorithms (GA), particle swarm optimizer (PSO), gray wolf optimizer (GWO), and others	The MLP and ANFIS models stood out in the analysis due to the high generalization capacity for long-term forecasting
Nemati et al. [6]	Analyze the survival characteristics of patients with COVID-19 by computational techniques and predict the length of stay of these patients in the hospital	Statistical analysis techniques along with machine learning taking into account the survival characteristics of 1182 patients	The results obtained show that the gradient boosting model surpasses the other models for predicting patient survival, followed by the KM and cox regression methods
Batista et al. [11]	Predict the risk of a positive diagnosis of COVID-19 with machine learning, using the results of admission exams in the RT-PCR tests of 235 adult patients as predictors in the emergency room	Five machine learning algorithms were used in the experiments, namely, neural networks, random forests, gradient increase trees, logistic regression, and SVM	The best predictive performance was obtained by the SVM algorithm (with AUC: 0.85; SEN: 0.68; ESP: 0.85; BrierScore: 0.16). And, three variables were identified as most important for good predictive performance, namely: Number of lymphocytes, leukocytes, and eosinophils
Kumar et al. [7]	Propose a classifier capable of diagnosing patients with COVID-19 using chest X-ray images	ML-based classifier and deep learning using ResNet152 on chest X-ray images of patients with COVID-19 for early and non-invasive prediction of the new coronavirus	The best results were obtained using the random Forest model, which excelled in terms of accuracy (0.973), F1 score (0.973), AUC (0.997), SEN (0.974), and ESP (0.986)
Gomes et al. [16]	Development of an intelligent system to support the diagnosis of COVID-19, using radiographs from different databases	Multiclass classification, differentiating between multiple respiratory diseases, such as COVID-19, viral pneumonia, and bacterial pneumonia	The developed system can diagnose COVID-19 with up to 89.78% of average accuracy, also being able to differentiate COVID-19 from viral and bacterial pneumonia

(continued)

Table 1 (continued)

Authors/Year	Objective	Method	Results
de Barbosa et al. [13]	Propose an intelligent system capable of supporting the COVID-19 diagnosis based on blood tests	Experimenting with six classic machine learning models, namely, MLP, SVM, random trees, random Forest, Bayesian networks, and naive Bayes	The model that stood out most was the SVM, capable of achieving high diagnostic performance, with an overall accuracy of 95.159% ± 0.693 and low computational cost
This work	Analyze intelligent classifiers that are able to make hospitalization predictions considering three possible scenarios: Regular ward, semi-intensive care unit, and intensive care unit, corresponding to mild (non-critical), moderate, and serious cases	For this, we used hematological data from patients of the units of the unified health system in the city of Paudalho, Brazil. Where, seven classic classifiers were applied to the data set, such as SVM, MLP, random tree, random Forest, Bayesian networks, and J48 decision trees	The results obtained show the random Forest with 100 trees showed the best potential to perform the predictions for regular ward (ACC: 82%; KPP: 0.642; SEN: 0.730 and ESP: 0.913), as for the semi-intensive care unit (ACC: 81%; KPP: 0.633; SEN: 0.890 and ESP: 0.875), and intensive care unit (ACC: 82%; KPP: 0.640; SEN: 0.640 e ESP: 0.947)

It is valid to clarify that all steps were applied individually for each one of the three hospitalization conditions (regular, semi-intensive, and intensive). Class balancing was performed (Step 1), using SMOTE method (Synthetic Minority Oversampling Technique) [27], which aims to generate artificial instances based on existing samples to balance the classes. The settings used in this step can be seen in Table 3.

After class balancing, in Step 2 we pre-processed the data using MLPAutoencoder algorithm, which is an unsupervised learning method to select attributes and, consequently, decrease data dimensionality [28].

In Steps 3, 4, and 5 seven (7) classic classifiers were applied, both for balanced databases (Step 1) and for databases with selected attributes (Step 2). It is worth mentioning that in order to obtain individual statistical performance information for the analyses we tested each classifier 30 times, using the k-fold cross-validation method with the number of folds equal to 10. Table 4 shows which classifiers were used and their respective configuration.

Exceptionally, when using the original database, without pre-processing, (Step 5), only the classifiers that obtained the best results in Steps 3 and 4 were used, namely, Random Tree, Random Forest, and J48 Decision Tree, all configured with the parameters shown in Table 3. The choice of these algorithms was due to their performance in terms of the analyzed metrics, in addition to the consumption of time and memory involved in their processing. In Step 6, some evaluative metrics were used to perform a multimodal analysis of the obtained results. The metrics

Table 2 Features extracted from patients' blood tests

List of features				
Hematocrit	Leukocytes	Serum glucose	Metamyelocytes	Segmented
Hemoglobin	Basophils	Neutrophils	Myelocytes	HB saturation arterial blood gases
Platelets	Mean corpuscular hemoglobin Mch	Urea	Myeloblasts	Total CO_2 arterial blood gas analysis
Mean platelet volume	Eosinophils	C-reactive protein	Partial thromboplastin time PTT	Promyelocytes
Red blood cells	Mean corpuscular volume MCV	Creatinine	Lactic dehydrogenase	PCO_2 arterial blood gas analysis
Lymphocytes	Monocytes	Total bilirubin	Prothrombin time pt. activity	HCO_3 arterial blood gas analysis
Mean corpuscular hemoglobin concentration Mchc	Red blood cell distribution width rdw	Direct bilirubin	Lipase dosage	Indirect bilirubin
D-dimer	Base excess arterial blood gas analysis	PH arterial blood gas analysis	PO_2 arterial blood gas analysis	Arterial FIO_2
CTO_2 arterial blood gas analysis				

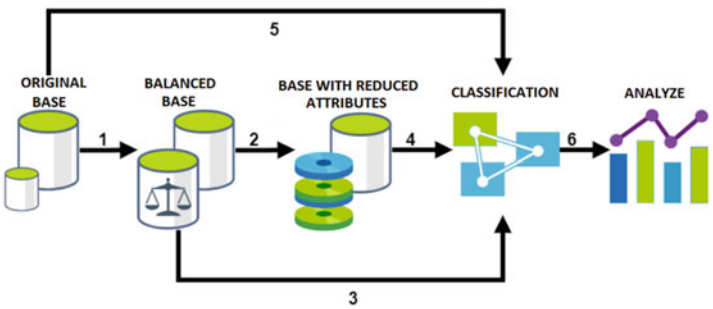

Fig. 1 Steps of the methodological path. In Step 1 the class balancing was carried out; in Step 2 we performed attribute selection; in Steps 3, 4, and 5, we trained and tested the classification models; and Step 6 consists in the metrics analyses

Table 3 Settings used for the SMOTE method

Care unit	Class value	Nearest neighbors	Percentage
Regular	0	2	95%
Semi-intensive	0	2	27.7%
Intensive	0	2	790%

Table 4 Classifiers configuration

Classifier	Parameters
Naive Bayes	Batch size: 100
Bayes net	Batch size: 100
Random tree	Batch size: 100 Seed: 1
J48 decision tree	Batch size: 100
Random Forest	Trees: 10, 20, 50 e 100 Batch size: 100
MLP	Neurons in the hidden layer: 20, 50 and 100 Learning rate: 0.3 Iterations: 500
SVM	Linear kernel ($P = 1$) Polynomial kernel ($P = 2$ and $P = 3$) RBF kernel (gamma: 0.01; 0.25 and 0.5)

were accuracy (ACC), kappa Statistics (KPP), sensitivity (SEN), specificity (ESP), ROC curve area (AUC), and training time (TT). In the results section, it is also reinforced that, for statistical purposes, the values presented refer to the averages and standard deviation of each metric, calculated from the 300 testing values found from the experiments.

5 Results

For better understanding, the results section was divided into three subsections, each referring to one of the care units. In each subsection we, respectively, assessed classifiers' performances in the prediction of regular ward hospitalization, semi-intensive care unit hospitalization, and intensive care unit hospitalization.

5.1 Regular Ward Hospitalization

In the data referring to the regular care wing, class 0 has 4110 instances and represents patients who did not need to be admitted to the regular ward. While class 1 is composed of 2105 instances and refers to patients who needed to be admitted to the regular ward. As a result, the imbalance between the two classes mentioned

is visible, with a difference of 2005 instances. To solve this problem, we applied the SMOTE method considering the real instances as examples to generate synthetic data, based on two neighbors and using the percentage of 95%. After applying SMOTE, classes 0 and 1 were balanced, both with 4110 instances.

In Step 2, MLPAutoenconder was applied to select attributes and thus, reduce the dimensionality of the data. Then, from the 41 features of the original database (listed in Table 2), only 10 were identified by the model as relevant.

The results obtained after the application of the seven classifiers (Step 3) on the balanced database (Step 1) show that the model with the best performance in relation to accuracy (82.1%), kappa statistics (0.64), sensitivity (0.73), and specificity (0.91) was the Random Forest with 100 trees. On the other hand, the worst result achieved for this case was obtained by Naive Bayes algorithm, in relation to accuracy (65%), kappa (0.30), and sensitivity (0.38); however, the average specificity value of 0.98 was slightly higher than the other methods. These results can be seen in Fig. 2, and may be further analyzed in Table 5.

In Step 4, where we applied the MLPAutocoder to select attributes in the database pre-processed with SMOTE (Step 3), the Random Forest with 100 trees stood out in terms of accuracy (75%), kappa statistics (0.651), sensitivity (0.676), and specificity (0.824). Another notable point is the good results obtained by Random Forest with 50 trees, where kappa statistics (0.651) and sensitivity (0.676) are equal to the values of Random Forest with 100 trees, differing only in accuracy (74%) and specificity (0.800).

On the other hand, the SVM with linear kernel achieved a worse performance in terms of accuracy (67%) and sensitivity (0.320), but this same model showed better kappa statistics (0.489) and specificity (0.920) than the values obtained by Naive Bayes, with kappa of 0.298, and specificity of 0.880, but still with better accuracy (69%) and sensitivity (0.390).

Figure 3 presents the results obtained in Step 4. Briefly, it is clear that the algorithms with better performances were both Random Forest, using 50 and 100 trees, as already described. Naive Bayes, followed by SVM with linear kernel, is the model that stands out negatively when compared to the others to carry out the prediction of hospitalization in the regular ward.

Regarding Step 5, among the three classifiers, the accuracy (76.9%), kappa statistics (0.523), sensitivity (0.742), and specificity (0.822) of Random Forest with 100 trees stood out in relation to the other models. However, it is worth noting that Random Forest with 10, 20, and 50 trees also obtained good results, similar to those with 100 trees. Through Fig. 4 it is possible to view this information.

Even with good results, in this scenario, Random Tree continues to occupy the position of classifier with the worst performance in relation to the other models in terms of accuracy metrics (72%), kappa index (0.453), sensitivity (0.705), and specificity (0.798).

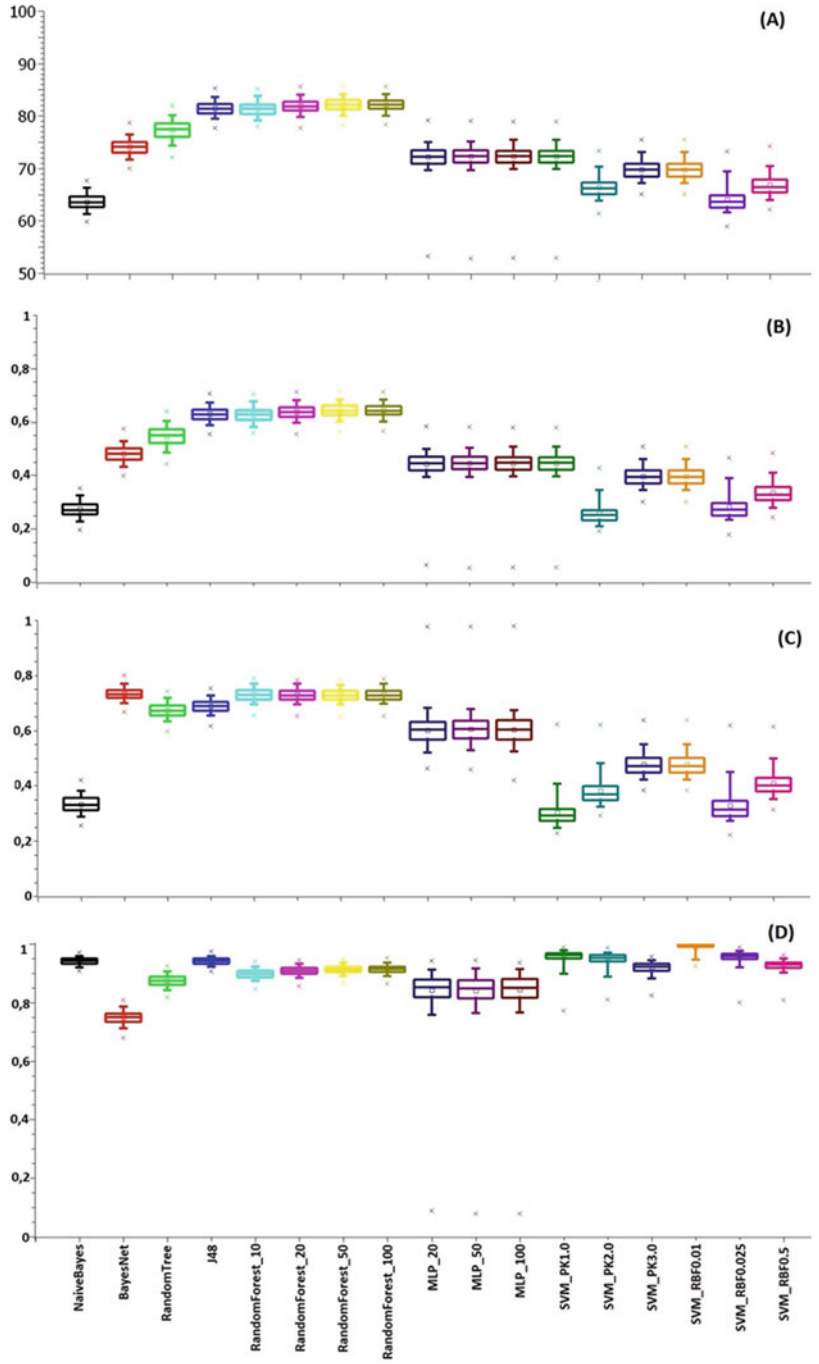

Fig. 2 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**), and specificity (**d**) to predict regular ward hospitalization using the balanced database (Step 3)

Table 5 Average and standard deviation of the evaluation metrics referring to the best models by type of care unit

Database	Type of care unit	Classifier	ACC (%)	KPP	SEN	SPE	AUC	TT (ms)
Original	Regular ward	RdmForest_100	76.936 ± 1.601	0.523 ± 0.034	0.742 ± 0.027	0.822 ± 0.045	0.854 ± 0.014	12.650 ± 5.309
	Semi-intensive	RdmForest_100	79.626 ± 1.568	0.594 ± 0.031	0.858 ± 0.022	0.861 ± 0.014	0.859 ± 0.056	74.831 ± 2.243
	Intensive	RdmForest_100	91.189 ± 0.674	0.335 ± 0.060	0.984 ± 0.004	0.262 ± 0.052	0.733 ± 0.034	3.327 ± 0.135
Balanced	Regular ward	RdmForest_100	82.175 ± 1.251	0.642 ± 0.030	0.730 ± 0.022	0.913 ± 0.013	0.884 ± 0.011	15.570 ± 4.101
	Semi-intensive	RdmForest_100	81.670 ± 1.710	0.633 ± 0.028	0.890 ± 0.017	0.742 ± 2.301	0.875 ± 0.012	8.217 ± 1.179
	Intensive	RdmForest_100	82.060 ± 1.040	0.640 ± 0.020	0.694 ± 0.018	0.947 ± 0.009	0.919 ± 0.007	7.261 ± 2.324
Selected features	Regular ward	RdmForest_100	75.033 ± 1.448	0.651 ± 0.029	0.676 ± 0.022	0.824 ± 0.018	0.808 ± 0.014	11.928 ± 0.510
	Semi-intensive	RdmForest_100	71.917 ± 1.615	0.438 ± 0.032	0.764 ± 0.022	0.674 ± 0.024	0.778 ± 0.016	5.059 ± 1.121
	Intensive	RdmForest_100	77.243 ± 4.205	0.545 ± 0.085	0.664 ± 0.111	0.882 ± 0.013	0.870 ± 0.054	16.758 ± 5.753

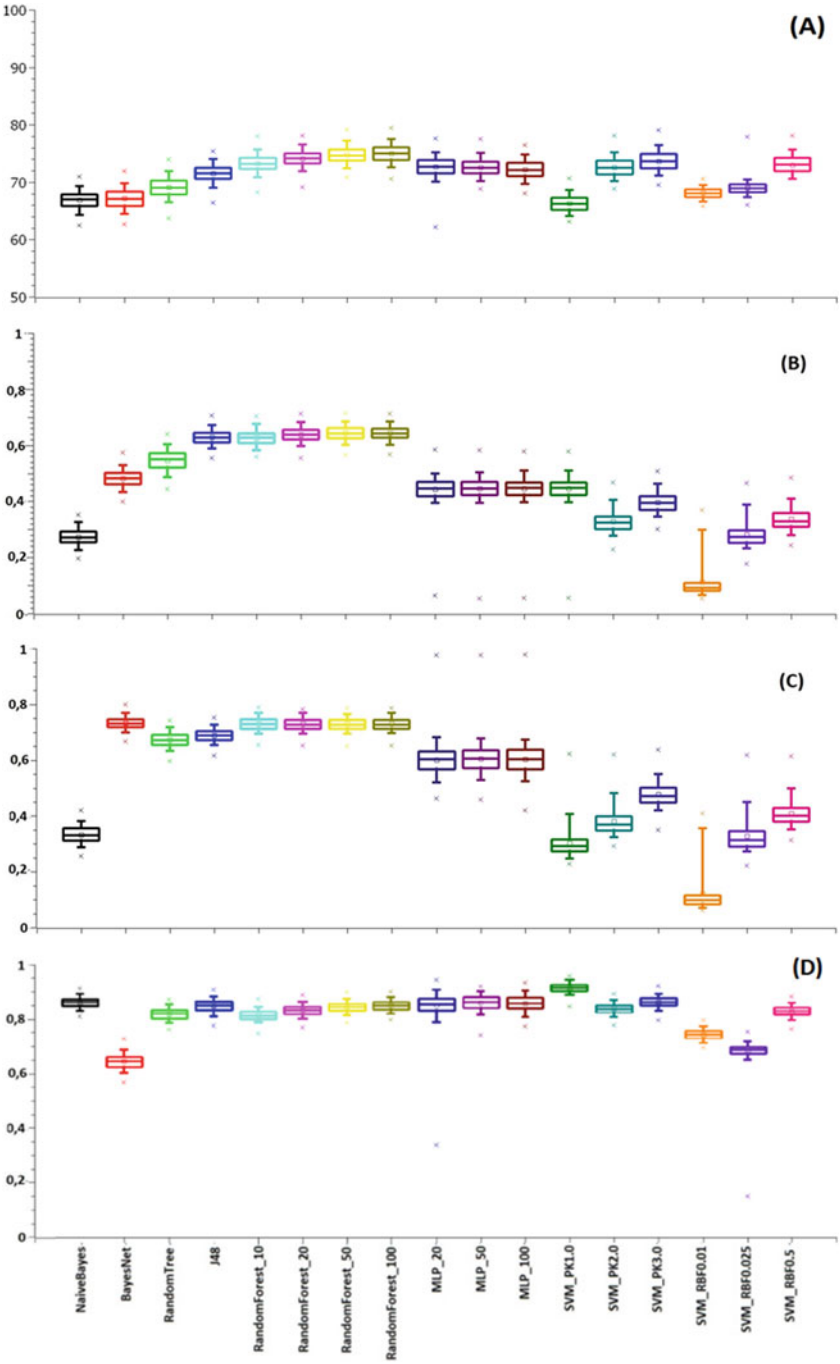

Fig. 3 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**), and specificity (**d**) to predict regular ward hospitalization using the database with features selected by MLPAutoencoder (Step 4)

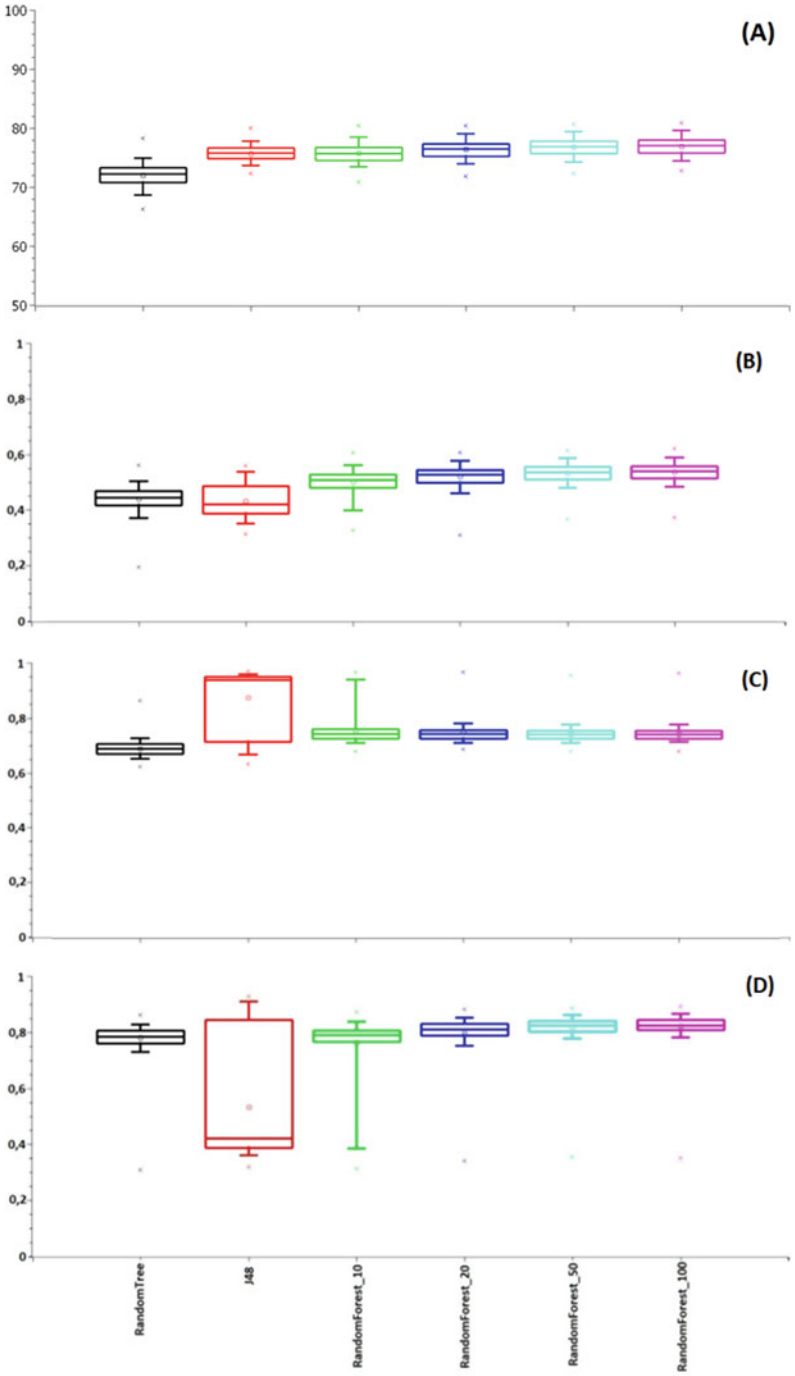

Fig. 4 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**), and specificity (**d**) to predict regular ward hospitalization using the original database, without pre-processing (Step 5)

5.2 Semi-Intensive Care Unit Hospitalization

In the data referring to semi-intensive care, class 0 is composed of 2728 instances and represents patients who did not need this type of care, while class 1 has 3487 instances and represents patients who needed care in the semi-intensive unit. There is clearly a difference of 759 instances between classes. In this unbalanced scenario, SMOTE was applied to these data in Step 1, considering two neighbors of the majority class and applying the percentage of 27.7% (as specified in Table 3). After this procedure, both classes were left with an equivalent number of instances of 3487.

Such as in the context of regular ward hospitalization, in Step 2, from the 41 original features, only 10 were selected by MLPAutoenconder. The prediction performance of the seven classifiers (Step 3) on the balanced database (Step 1) clearly demonstrate that the Random Forest with 100 trees stood out positively in terms of accuracy (81%), kappa statistics (0.633), sensitivity (0.890), and specificity (0.742). The Random Forest with 20 and 50 trees also performed well reaching similar results to that using 100 trees, as shown in Fig. 5.

On the other hand, Naive Bayes presents less satisfactory results regarding accuracy (61%), kappa statistics (0.230), and specificity (0.300); however, the sensitivity (0.980) of this classifier stands out in relation to the other models. The performance of this classifier is closely followed by SVM with RBF kernel and gamma of 0.01, with worse performances in the metrics of accuracy (60.8%), kappa statistics (0.240), and specificity (0.320). Similar to Naive Bayes, SVM also showed high sensitivity (0.870).

With the reduced number of attributes (Step 2), the results of Step 4 did not present major discrepancies between them. However, the best result was obtained by the Random Forest model with its respective configurations, highlighting the accuracy (71.9%) and the kappa statistics (0.438) of the model using 100 trees. In terms of sensitivity, the SVM with a 0.01 gamma and RBF kernel stood out, achieving the average value of 0.830. Regarding specificity, the Naive Bayes model performed better than the other methods, reaching 0.780. However, when comparing the training time (5 s) and the area under the ROC curve (0.778) of the 100 trees Random Forest with the other models, it presents the best results and is seen as a potential model for predicting hospitalization in semi-intensive care units in this scenario.

Analyzing from the perspective of the worst performing model, we highlight the SVM with linear kernel, which obtained the worst results in terms of accuracy (61%), kappa statistics (0.230), and specificity (0.380), improving only in sensitivity (0.850). In Fig. 6, it is possible to observe the performance of the aforementioned models.

In order to carry out a comparative analysis with the results obtained in Steps 3 and 4, in Step 5 experiments were performed to predict hospitalization in semi-intensive care units from the database without pre-processing. As shown in Fig. 7, it is clear that the Random Forest model with its respective tree configurations stands

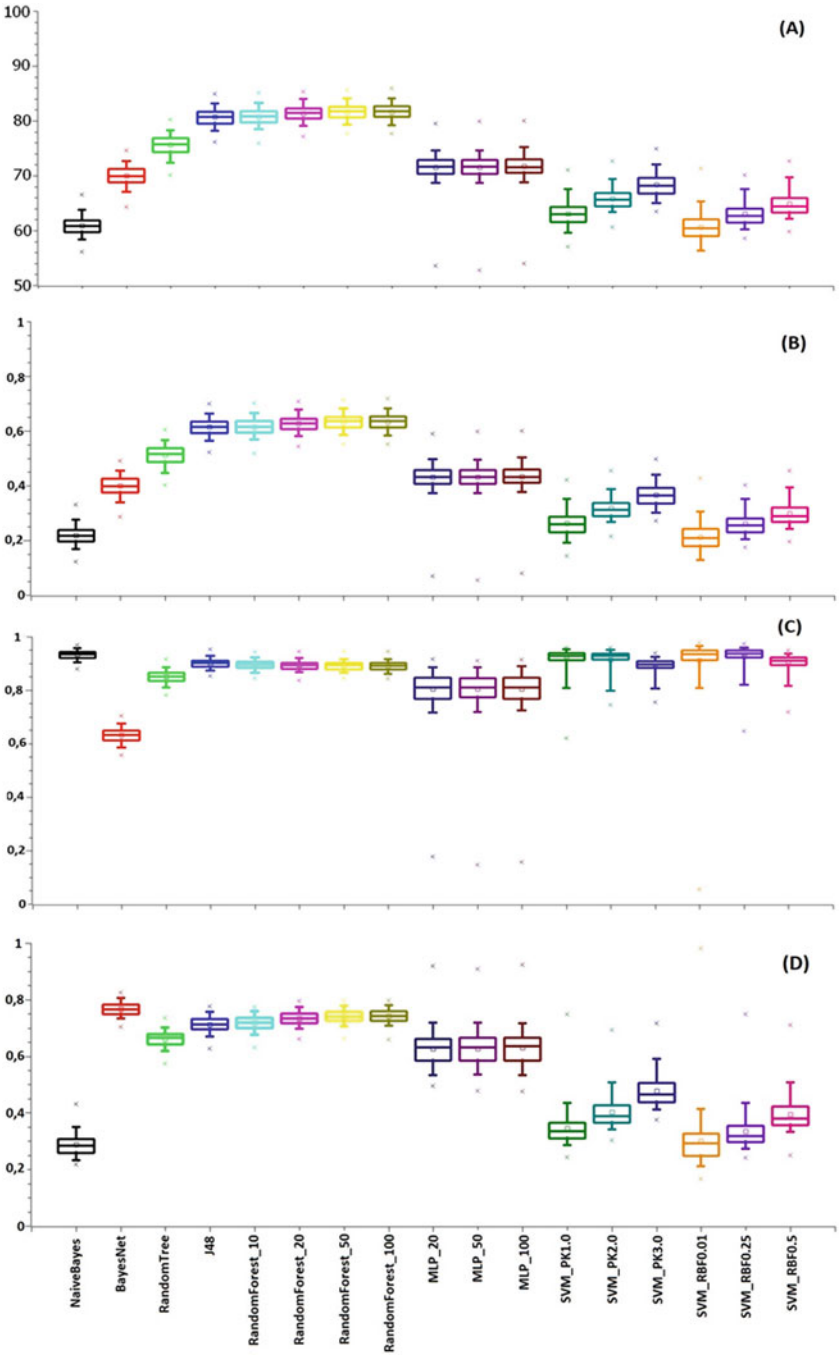

Fig. 5 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**), and specificity (**d**) to predict semi-intensive care unit hospitalization using the balanced database (Step 3)

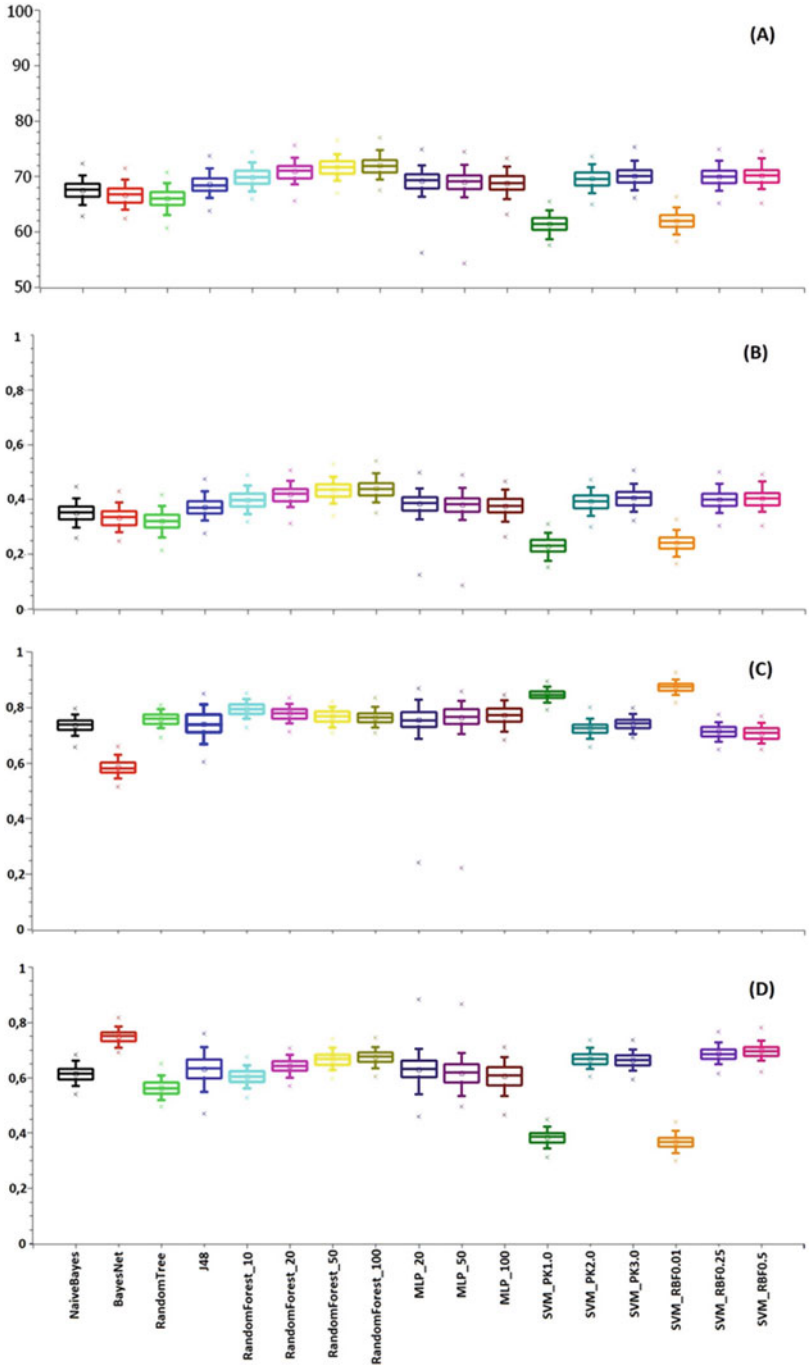

Fig. 6 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**), and specificity (**d**) to predict semi-intensive care unit hospitalization using the database with selected features (Step 4)

out, especially that of 100 trees regarding accuracy (79.6%), kappa statistics (0.594), sensitivity (0.858), and specificity (0.861).

Finally, analyzing the model with the worst performance in this scenario, we highlight the Random Tree model in terms of accuracy (73%), kappa statistics (0.512), sensitivity (0.710), and specificity (0.690). However, it is worth noting that in the general context, all three models (Random Tree, J48, and Random Forest) showed positive results and better performance than SVMs and MLPs. But among these three classifiers, Random Forest showed the best overall performance.

5.3 Intensive Care Unit Hospitalization

In the data regarding prediction of hospitalization in intensive care units, class 0 has 5592 instances and represents patients who did not need to be admitted to the intensive care unit. While class 1 is composed of 623 instances and refers to patients who needed intensive care. Clearly, it is possible to identify an imbalance between classes, with a difference of 4969 instances. In order to perform the class balancing, in Step 1 (SMOTE method), it was necessary to carry out an expansion of 790% of the smaller class based on two neighbors of the majority class, resulting in two balanced classes, each one with 5592 instances.

During the analysis, from the predictions of hospitalization in intensive care units using the database pre-processed with the SMOTE method (Step 3), the algorithm that obtained the best accuracy (82%) was Random Forest (as shown in Fig. 8a), configured with 50 and 100 trees, respectively. Bayes Net also stood out for accuracy, with 78.4%. For this same metric, the models that had the most outlier values were Naive Bayes (60%) and SVMs (ranging from 63% to 68%). This information can be seen in Fig. 8.

In Fig. 8, we also see the kappa statistics (Fig. 8b) obtained by the models in Step 3, which ended up following a pattern similar to the previous metric: with Random Forest of 100 trees achieving the best average kappa (0.65). Naive Bayes model with kappa of 0.22, followed by the SVMs with kappa statistics ranging from 0.23 to 0.43, showed less satisfactory results.

Regarding sensitivity, the SVM models performed better, reaching 0.98 in the identification of true positives. While Naive Bayes obtained only 0.26, it stands out, therefore, in three parameter analyses, as the worst model to be used in this context. The specificities of the SVMs models, on the other hand, were the lowest, thus disqualifying these models for the identification of true negatives.

In Step 4, after applying MLPAutocode, there was a reduction from 41 to 10 attributes in the database pre-processed with SMOTE (Step 3). By analyzing the accuracy of the models in detail, as shown in Fig. 9, it is clear that the Random Forest models of 100, 50, 20, and 10 trees stood out in comparison to the other algorithms, with accuracy between 76% and 78%. On the other hand, it was also found that SVM models did not obtain good results in this metric, specially SVM

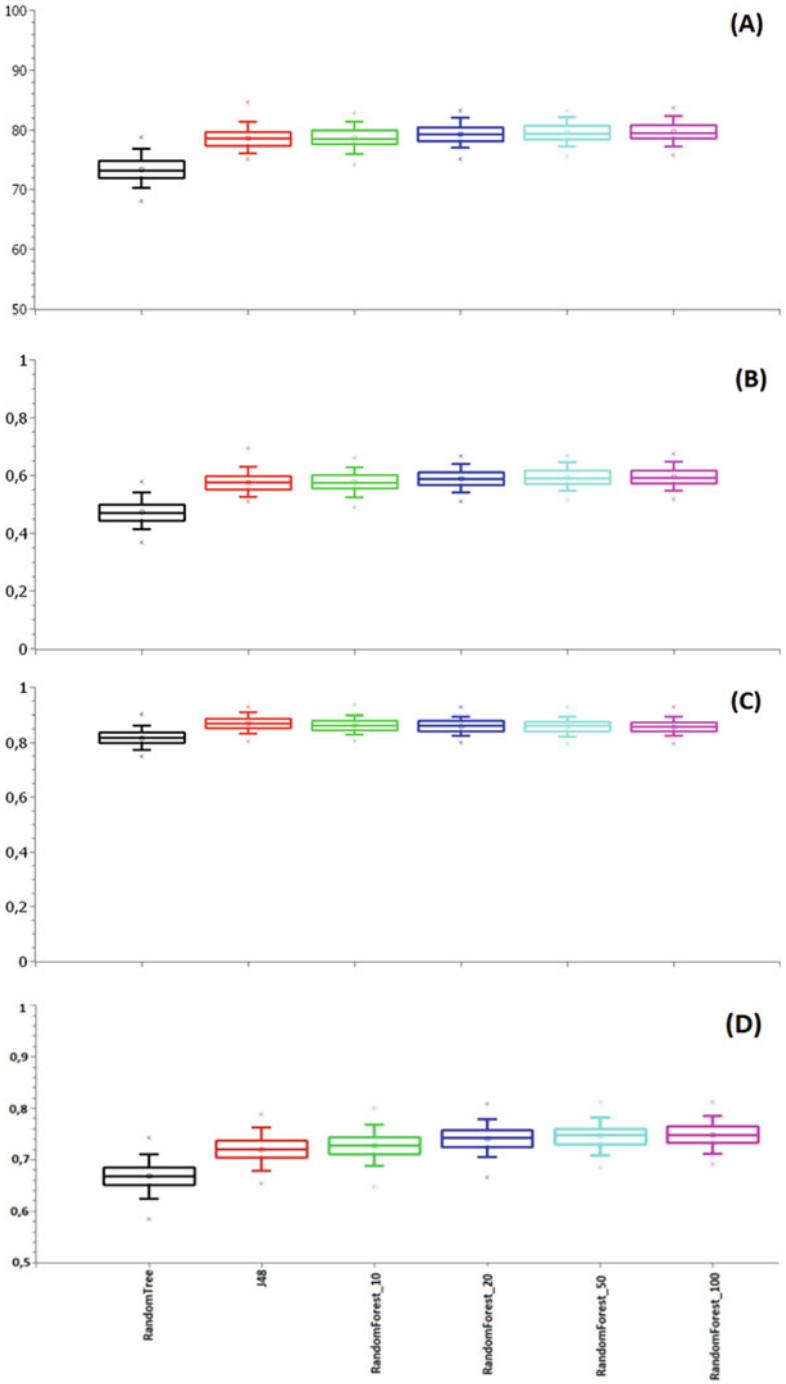

Fig. 7 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**), and specificity (**d**) to predict semi-intensive care unit hospitalization using the original database, without pre-processing (Step 5)

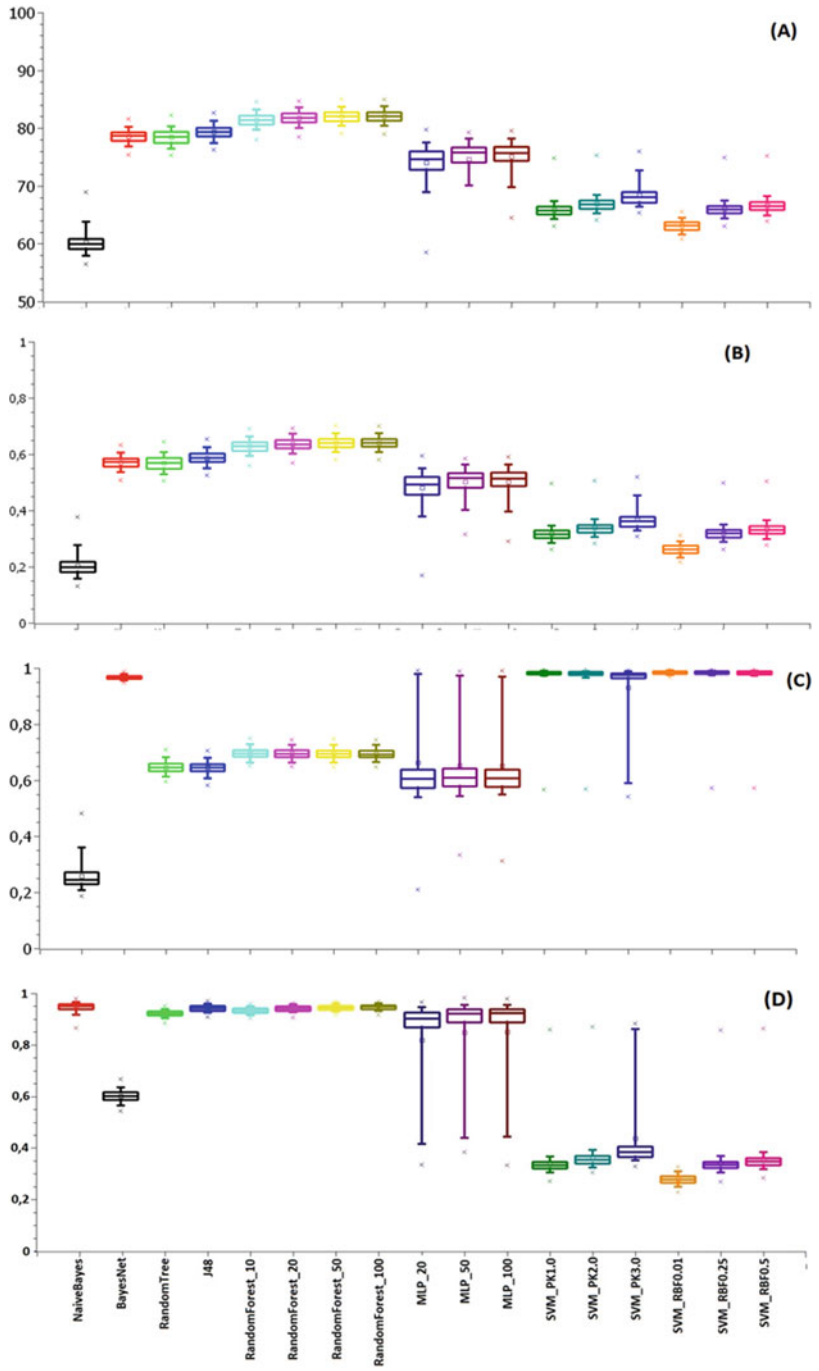

Fig. 8 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**) and specificity (**d**) to predict intensive care unit hospitalization using the class-balanced database (Step 3)

with RBF kernel and gamma of 0.01 (63%), followed by SVM with linear kernel (64%).

We highlight that the kappa statistics presented a behavior very similar to the accuracy. The Random Forest model stood out from the other methods, with kappa reaching 0.55. The SVM models with RBF kernel with 0.01 gamma and SVM with linear kernel performed worse again, with kappa of 0.25 and 0.29, respectively.

Unlike what has been analyzed and reported so far, for sensitivity (Fig. 9c) in Step 4, we see that the models with the best sensitivity were Naive Bayes (0.98), followed by SVM with linear kernel (0.97) and the SVM with RBF kernel and gamma of 0.01 (0.89), while Random Tree and J48 showed sensitivity of 0.60 and 0.61, respectively. We also noticed that the same models that performed better in sensitivity achieved worse values for specificity (Fig. 9d), indicating a discrepancy between true-positive and negative predictions.

For comparative purposes, as shown in Fig. 10, only the Random Forest, J48, and Random Tree models were executed on the database without pre-processing (Step 5), since these algorithms showed better overall performances in the previous steps. As a result, the accuracies of J48 (94%) and Random Forest models (from 90% to 93%) were outstanding. Random Tree, on the other hand, achieved 87% of accuracy, decreasing in relation to the other models.

Still at this step, during the analysis of the kappa statistics, the Random Forest models of 50 and 100 trees continued to stand out, with average values of 0.38, followed by J48 with kappa of 0.37. The Random Tree classifier did not show good performance, with an average kappa equal to 0.27.

Figure 10 also presents the results related to the sensitivity in Step 5. In this metric, the findings demonstrate that J48 stands out with an average value of 0.99, closely followed by Random Forest, with 0.98. Random Tree obtained 0.95 for the same indicator. As for specificity, the discrepancy with the pattern presented so far is visible. This time, the J48 model performed worse among the models, with a true-negative rate of 0.22. The Random Forest model with 100 trees, on the other hand, reached the highest average specificity value (0.26).

After a comparative and multimodal analysis in each of the stages, to predict the three types of hospitalization proposed, the best models were selected. In order to further show these results, the indicators for the best models in each scenario were compiled in Table 5, together with the averages and standard deviations of accuracy (ACC), kappa statistics (KPP), sensitivity (SEN), specificity (SPE), area under the ROC curve (AUC), and the training time (TT).

Finally, it is of great value to mention that at all steps for the three different types of care unit (regular, semi-intensive, and intensive), the model that stood out positively regarding the evaluation metrics was the Random Forest, mainly with the configuration of 50 and 100 trees. On the other hand, the models that had less satisfactory results, in all three conditions, were Naive Bayes, SVMs, and Random Tree (Stage 5). However, in order to obtain a more accurate and general analysis, the discussion of these results will be further explored in the following section.

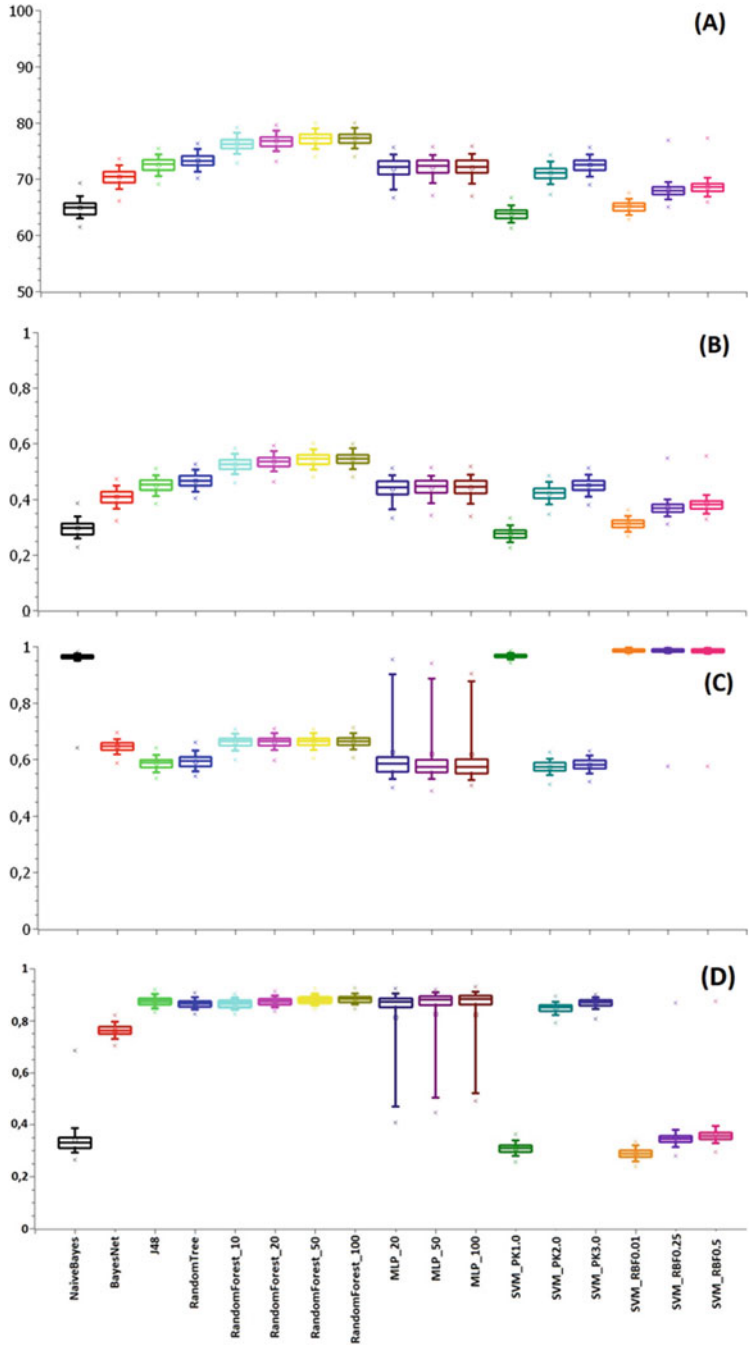

Fig. 9 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**) and specificity (**d**) to predict intensive care unit hospitalization using the database with selected features (Step 4)

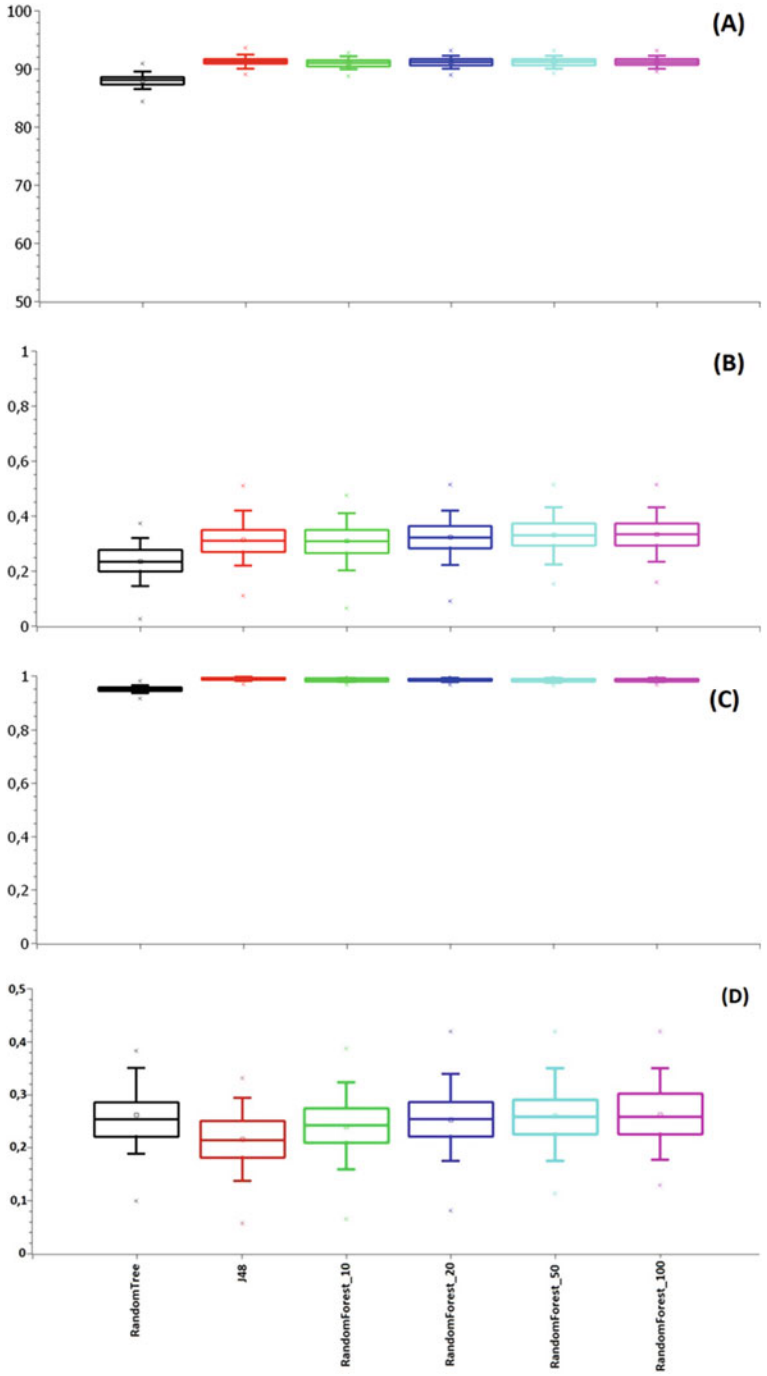

Fig. 10 Results of accuracy (**a**), kappa statistics (**b**), sensitivity (**c**), and specificity (**d**) to predict intensive care unit hospitalization using the original database, without pre-processing (Step 5)

6 Discussion

In general, the Random Forest model with its four configurations (10, 20, 50, and 100 trees), had better performance. In all the generated databases (original, balanced, and with attributes selection), its performance was satisfactory, not only considering the average values of its metrics, but also the standard deviations and the constancy of these models. About the training time, it was identified that very high degree parameters (such as SVM and MLP) showed an exponential growth for an insignificant performance gain. Although the re-training is done on average once a week, the metrics were not positive enough to encourage the use of these models.

In the context of regular ward care, the Random Forest model was the most successful. Both with the configuration of 100 trees, as well as those of 50, 20, and 10 trees. Another classifier that also showed positive results was the J48. It is worth mentioning that in this context, as shown in the results section, the SVMs and MLPs have not shown satisfactory results.

The prediction of hospitalization in semi-intensive care units shares the same findings as the regular ward. In both, Random Forest with its four configurations performed better. On the other hand, SVMs and Bayes Net declined in the results of the evaluation indicators. It is important to note that most of the classifiers were more promising with the balanced database.

Considering the intensive care unit hospitalization, the Random Forest models with 100 and 50 trees showed better accuracy and sensitivity with the original database. Kappa index, specificity, and area under the ROC curve were best evaluated on the balanced database. Still for this type of care, the Naive Bayes model was the worst classifier, taking into account that its metrics are the lowest.

One curiosity found, which is worth reporting, refers to the calculated sensitivity for recommendation in care in semi-intensive and intensive care units. In this indicator, SVM models were the best evaluated, SVM with RBF kernel and 0.01 gamma, followed by SVM with linear kernel. Despite their good performance for the identification of true positives, these classifiers showed discrepancies in terms of accuracy, kappa index, specificity, and area of the ROC curve.

Finally, in view of the evidence, the results weighed against the original bases. This leads to belief that the balance of bases, more than the reduction in the number of attributes, result in better classifications of this problem. This leaves the possibility that, perhaps, the MLPAutocode is not the most appropriate technique in this context of prediction. In testing this hypothesis, future works may use other techniques and, consequently, different configurations to carry out a comparative analysis of the results.

7 Conclusion

Considering this atypical pandemic moment caused by the new coronavirus, the use of predictive models has been helping health professionals in order to control the spread of the virus and the burden on the health system. In this context, our work contributes to this theme by proposing a comparative analysis of seven classic classifiers capable of making predictions regarding care and assessments of the severity of patients with and without COVID-19, seen at the Unified Health System (SUS) units of the municipality of Paudalho, Brazil, based on the evaluation of hematological data (blood tests).

The results obtained through the evaluative metrics show that among the seven classifiers used, both for predicting regular appointments and for attending a semi-intensive care unit, the Random Forest algorithm showed better performance with all configurations, in relation to the other models. In both cases, the SVM also stood out, but negatively, thus being the least suitable model to be used in this scenario.

Besides the findings, we highlight that analyzing the results from different quantitative and qualitative perspectives is important both for a better understanding of the problem, as well as for choosing the best solution in the researched context. For this reason, several aspects, ranging from the robustness of the model to the time of its execution, must be taken into account.

It is worth noting that the predictive models should assist the health professional in making decisions. The model is only for support and streamline the process. In the case of this work, the prediction can help by assisting in the screening of regular patients, so that those with moderate and severe cases receive care as soon as possible.

Finally, as perspectives for future work, it is intended to apply to the three databases (referring to regular, intensive, and semi-intensive care) other pre-processing techniques, as well as other methods for reducing and selecting the number of attributes with different configurations, in order to verify comparatively the classifiers' performance in relation to the original knowledge base and the base with the pre-processed data.

References

1. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2019). A novel coronavirus from patients with pneumonia in China. *The New England Journal of Medicine, 382*, 727–733.
2. Croda, J. H. R., & Garcia, L. P. (2020). Resposta imediata da Vigilância em Saúde à epidemia da COVID-19. *Epidemiologia e Serviços de Saúde,* 29(1), Brasília.
3. Iser, B. P. M., Sliva, I., Raymundo, V. T., Poleto, M. B., Scuelter-Trevisol, F., & Bobinski, F. (2020). Definição de caso suspeito da COVID-19: uma revisão narrativa dos sinais e sintomas mais frequentes entre os casos confirmados. *Epidemiologia e Serviços de Saúde, 29*(3), e2020233.
4. World Health Organization (WHO). (2019). *Coronavirus disease (COVID-19) pandemic*. Geneva: World Health Organization. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

5. Teodoro, L. A., & Kappel, M. A. A. (2020). Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. *Revista Brasileira de Informática na Educação, 28*, 838–863.
6. Nemati, M., Ansary, J., & Nemati, N. (2020). Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. *ScienceDirect, Patterns, 1*(5), 100074.
7. Kumar, M., Patel, A. K., Shah, A. V., Raval, J., Rajpara, N., Joshi, M., & Joshi, C. G. (2020). First proof of the capability of wastewater surveillance for COVID-19 in India through detection of genetic material of SARS-CoV-2. *Science of the Total Environment, 746*, 141326.
8. Woelfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Mueller, M. A., Niemeyer, D., Kelly, T. C. J., Vollmar, P., Rothe, C., Hoelscher, M., Bleicker, T., Brunink, S., Schneider, J., Ehmann, R., Zwirglmaier, K., Drosten, C., & Wendtner, C. (2020). Clinical presentation and virological assessment of hospitalized cases of coronavirus disease 2019 in a travel-associated transmission cluster. *medRxiv*. Available from: https://doi.org/10.1101/2020.03.05.20030502.
9. Tolia, V. M., Chan, T. C., & Castillo, E. M. (2020). Preliminary results of initial testing for coronavirus (COVID-19) in the emergency department. *Western Journal of Emergency Medicine, 21*(3), 503–506. Available from: https://doi.org/10.5811/westjem.2020.3.47348.
10. Hadaya, J., Schumm, M., & Livingston, E. H. (2020). Testing individuals for coronavirus disease 2019 (COVID-19). *JAMA, 323*(19), 1981. Available from: https://doi.org/10.1001/jama.2020.5388.
11. Batista, A. F. M.; Miraglia, J. L.; Donato, T. H. R., & Chiavegatto Filho, A. D. P. (2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *medRxiv*.
12. Torcate, A. S.; Barbosa, J. C. F., & de Oliveira Rodrigues, C. M. (2020). Utilizando o Learning Analytics com o K-Means para Análise de Dificuldades de Aprendizagem na Educação Básica. In *Anais do XXVI Workshop de Informática na Escola* (pp. 31–40). SBC, November.
13. de Barbosa, V. A. F., Gomes, J. C., Santana, M. A., Albuquerque, J. E. A., Souza, R. G., Souza, R. E., & Santos, W. P. (2020). Heg.IA: Um sistema inteligente para apoiar o diagnóstico de Covid-19 com base em exames de sangue. *medRxiv preprint*. https://doi.org/10.1101/2020.05.14.20102533. this version posted May 18, 2020.
14. Jordan, M. I., & Mitchell, T. M. (2015). Aprendizagem de máquina: tendências, perspectivas e perspectivas. *Science, 349*, 255–260. Available from: https://doi.org/10.1126/science.aaa8415.
15. Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., & Notar, M. (2018). An application of machine learning to haematological diagnosis. *Scientific Reports, 8*, 1–12.
16. Gomes, J. C., Barbosa, V. A. D. F., Santana, M. A., Bandeira, J., Valença, M. J. S., Souza, R. E., Ismael, A. M., & Santos, W. P. (2020). IKONOS: uma ferramenta inteligente para apoiar o diagnóstico de COVID-19 por análise de textura de imagens de raios-X [publicado online antes da impressão, em 3 de setembro de 2020]. *Pesquisa em Engenharia Biomédica*; 1–14. https://doi.org/10.1007/s42600-020-00091-7.
17. Tanner, L., Schreiber, M., Low, J. G., Ong, A., Tolfvenstam, T., Lai, Y. L., Ng, L. C., Leo, Y. S., Thi Puong, L., Vasudevan, S. G., Simmons, C. P., Hibberd, M. L., & Ooi, E. E. (2008). Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS Neglected Tropical Diseases, 2*(3), e196. https://doi.org/10.1371/journal.pntd.0000196. PMID: 18335069; PMCID: PMC2263124.
18. Luo, Y., Szolovits, P., Dighe, A. S., & Baron, J. M. (2016). Using machine learning to predict laboratory test results. *American Journal of Clinical Pathology, 145*, 778–788.
19. Cordeiro, F. R., Santos, W. P. S., & Silva-Filho, A. G. (2017). Analysis of supervised and semi-supervised growcut applied to segmentation of masses in 635 mammography images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 5*, 297–315.
20. Silva, W. W. A., Santana, M. A., Silva Filho, A. G., Lima, S. M. L., & Santos, W. P. (2020). Morphological extreme learning machines applied to the detection and classification of mammary lesions. In T. K. Gandhi, S. Bhattacharyya, S. De, D. Konar, & S. Dey (Eds.), *Advanced machine vision paradigms for medical image analysis*. London: Elsevier.

21. Ji, Y., Ma, Z., Peppelenbosch, M. P., & Pan, Q. (2020). Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health, 8*(4), e480.
22. Yan, L., Zhang, H. T., Goncalves, J. et al. (2020, May 14). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, *ed 2*, 283–288. Available from: https://doi.org/10.1038/s42256-020-0180-7.
23. Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. *Algorithms, MedRxiv, 13*(10), 249.
24. Cohen, J. P., Morrison, P, & Dao, L. (2020). COVID-19 image data collection. arXiv:2003.11597. Available from: https://arxiv.org/abs/2003.11597.
25. Di Radiologia Medica and Intervencionista. (2020). *Covid-19 Database*. Available from: https://www.sirm.org/category/senza-categoria/covid-19/.
26. Witten, I. H., & Frank, E. (2020). Data mining: Practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record, 31*(1), 76–77.
27. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2020). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
28. Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016). Rainfall prediction: A deep learning approach. In *International Conference on Hybrid Artificial Intelligence Systems* (pp. 151–162). Springer, Cham; April.