Computers in Biology and Medicine 136 (2021) 104738

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Biomarkers of severe COVID-19 pneumonia on admission using data-mining powered by common laboratory blood tests-datasets

Mary Pulgar-Sánchez [a,1], Kevin Chamorro [b,c,1], Martha Fors [d], Francisco X. Mora [e], Hégira Ramírez [d], Esteban Fernandez-Moreira [f], Santiago J. Ballaz [a,f,*]

[a] *Escuela de Ciencias Biológicas e Ingeniería. Universidad Yachay Tech, Urcuquí, Ecuador*
[b] *Escuela de Matemáticas y Ciencias Computacionales. Universidad Yachay Tech, Urcuquí, Ecuador*
[c] *Universidad Técnica Del Norte, Ibarra, Ecuador*
[d] *Escuela de Medicina; Universidad de las Américas, Quito, Ecuador*
[e] *IESS Hospital Quito Sur, Quito, Ecuador*
[f] *Escuela de Medicina, Universidad Espíritu Santo, Samborondón, Ecuador*

ARTICLE INFO

Keywords:
Blood gas analyses
Clinical laboratory techniques
COVID-19
Data mining
Machine learning
Medical informatics applications

ABSTRACT

In the epidemiological COVID-19 research, artificial intelligence is a unique approach to make predictions about disease severity to manage COVID-19 patients. A limitation of artificial intelligence is, however, the high risk of bias. We investigated the skill of data mining and machine learning, two advanced forms of artificial intelligence, to predict severe COVID-19 pneumonia based on routine laboratory tests. A sample of 4009 COVID-19 patients was divided into Severe ($PaO_2 < 60$ mmHg, 489 cases) and Non-Severe ($PaO_2 \geq 60$ mmHg, 3520 cases) groups according to blood hypoxemia on admission and their laboratory datasets analyzed by the R software and WEKA workbench. After curation, data were processed for the selection of the most influential features including hemogram, pCO_2, blood acid-base balance, prothrombin time, inflammation biomarkers, and glucose. The best fit of variables was successfully confirmed by either the Multilayer Perceptron, a feedforward neural network algorithm that performed machine recognition of severe COVID-19 with 96.5% precision, or by the C4.5 software, a supervised learning algorithm based on an objective-predefined variable (severity) that generated a decision tree with 89.4% precision. Finally, a complex bivariate Pearson's correlation matrix combined with advanced hierarchical clustering (dendrograms) were conducted for knowledge discovery. The hidden structure of the datasets revealed shift patterns related to the development of COVID-19-induced pneumonia that involved the lymphocyte-to-C-reactive protein and leukocyte-to-C-protein ratios, neutrophil %, pH and pCO_2. The data mining approaches to the hematological fluctuations associated with severe COVID-19 pneumonia could not only anticipate adverse clinical outcomes, but also reveal putative therapeutic targets.

1. Introduction

The current pandemic outbreak of COVID-19 has led to millions of deaths worldwide, put a strain on healthcare systems of developed countries and collapsed the economies of low-income countries. Since the start of the pandemic, Ecuador has reported 468.414 COVID-19-infected people (26.142 per million) and 21.830 deaths (1218 per million) because of the SARS-CoV-2 infection [1]. Ecuador is one the countries with the sharpest increase of COVID-19 cases in Latin America. Although the majority are usually mild to moderate, COVID-19 disease can also progress to severe pneumonia (i.e., severe acute respiratory syndrome) and death. On hospital admission, the evaluation of the status of COVID-19 patients is critical for their clinical management, especially when dealing with limited hospital resources and personnel like in Ecuador.

Artificial intelligence may be a unique non-clinical approach to meet healthcare requirements and relieve the burden imposed by the COVID-19 pandemic [2]. One of the main criticisms is, however, the high risk of bias or optimistic predictive performance of this approach [3]. Although it may undermine trust in its efficacy, artificial intelligence is evolving all by itself. One of the most advanced techniques of artificial intelligence is data mining, which can be used for data-driven knowledge

[*] Corresponding author. School of Biological Sciences & Engineering, Yachay Tech University. Hacienda San José s/n, Proyecto Yachay, 100115, Urcuquí, Ecuador. *E-mail address:* sballaz@yachaytech.edu.ec (S.J. Ballaz).

[1] Both authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2021.104738
Received 24 April 2021; Received in revised form 14 July 2021; Accepted 2 August 2021
Available online 8 August 2021

powered by large labeled datasets [4,5]. Another promising subset of artificial intelligence is machine learning (ML), which focuses on producing systems that are able to learn from examples and improve without being explicitly programmed [6]. ML algorithms are just beginning to be investigated as prognostic tools. ML-driven predictions of COVID-19 outcomes are based on clinical data and largely radiological features extracted from X-ray computed tomography (CT) [2,7–9]. However, CT scan images are time consuming, weakly correlated with initial disease severity [10] and unable to detect small infected lung regions. CT systems are expensive and a heavy financial burden for hospitals [11].

Routinely-collected laboratory blood tests, which are readily available and inexpensive, have revealed important hematology, coagulative and biochemical fluctuations associated with COVID-19 severity [12–20]. Unfortunately, the power of data mining and ML applied to laboratory tests datasets has barely been explored [6,21–23]. Powered by a large laboratory dataset from confirmed COVID-19 patient [20], and assisted by the free R software [24] and the WEKA workbench [25] for statistical computing and graphics, this study was intended to generate blood-tests data-driven solutions that could predict the progression to severe COVID-19 pneumonia. While many studies rely on a previous selection of biomarkers to model the risk for developing severe COVID-19, this retrospective analysis used several data mining approaches [26] directly applied to a whole laboratory test dataset to find the best fit of variables that predicted disease severity [27].

2. Methods

2.1. Selection and classification of COVID-19 patients

Written informed consent was waived due to the retrospective nature of the study. We followed STROBE guidelines to report this study, which was approved by the Ethics Committee of the IESS COVID-19 Hospital Quito Sur and conducted in accordance with the ethical policies established by the Ecuadorian government [28]. The authors declare they had no access to identifying patient information when analyzing the data.

Laboratory tests was obtained from 4009 confirmed cases of COVID-19 (CDC 2019-Novel Coronavirus Real-Time RT-PCR Diagnostic Panel in upper and lower respiratory specimens) admitted at the IESS Hospital Quito Sur, a large COVID-19 hospital in Quito (Ecuador), from March 13 to June 17, 2020. Hematological analysis was performed using a Sysmex XN-550™ Hematology Analyzer (Sysmex America Inc., USA). Arterial blood gasometry was conducted on a RAPIDPoint® 500-systems blood gas system (Siemens Healthcare GmbH; Germany). All the basic information and laboratory results were downloaded from patients' electronic medical records (IBM AS-400). Consecutive adult patients with a confirmed diagnosis of COVID-19 were admitted with COVID-19-like symptoms, such as fever, cough, fatigue, shortness of breath, and headache. Drawing samples for routine laboratory tests took place upon admission [20] (a complete list is available in Supplementary material 1), whereas those patients with a pneumonia severity index (PSI) above 3 (1215 patients) in the triage evaluation were submitted to an arterial blood gas exam after a 40-min interval of average. From them, 489 patients suffered from blood hypoxemia ($PaO_2 < 60$ mmHg) and were classified as "Severe" [29]. The rest, 726 patients with a PSI >3 and $PaO_2 \geq 60$ mmHg, along with 2794 patients with no PaO_2 tests (PSI $\leq$ 3), were considered "Non-Severe" (3520 cases). This classification allowed us to predict the adverse clinical outcomes of COVID-19-induced pneumonia and the follow-up treatment plan.

2.2. Dataset collection and feature selection

Fig. 1 summarizes the complete workflow of dataset processing. Datasets were pre-processed for cleaning non-numeric data, symbols, eliminating outliers, inconsistencies, and discriminating or replacing null values, as well as data formatting. A nonparametric outlier detection approach based on interquartile ranges (IQR) was performed to cure datasets as follows: lower outlier = Q1 - (1.5•IQR) and higher outlier = Q3 + (1.5•IQR). All possible patient information was then compiled into a single dataset integrated into a *.csv* file and subsequently converted into.*arff* extension, a format compatible with the WEKA (Waikato Environment for Knowledge Analysis) workbench. This software is a data mining toolkit to approach bioinformatics problems using simple datasets [30]. Given the high dimensionality of the data, the processing pipeline continued with a selection of the features (FS) that better predicted COVID-19 severity. The FS was performed using the WEKA software with the default configurations and using severity classification (S = Severe and N = Non-severe) as the response variable. The methods applied were as follows: Correlation Subset forward method, which retains a subset of features highly correlated with the activity and lowly correlated among them [31], this was applied with the Best first and Genetic search as search strategies. The three following methods used Ranker as search method, which was specified to retain 30 attributes: (1) the Relief-F method retained the features that best distinguish among the nearest instances [32], (2) the Correlation Attribute evaluated the Pearson correlation between the features and the severity class, and (3) the Gain Ratio Attribute Evaluation analyzed the gain ratio with respect to the severity class [33]. Finally, the Principal Component Analysis (PCA) reduced the dimensionality that collected a 95% of the variance of the chosen eigenvectors using Ranker this time with the default configuration.

2.3. Statistics

Statistical analysis of the data was performed using the SPSS software (version 24.0) for Windows (IBM, Armonk, NY, USA). Proportions for categorical variables (sex, age, and severity) were compared by the Chi-square test to analyze the nature and distribution of the sample. Normal distribution tests were computed by the Kolmogorov-Smirnov test. The independent sample Mann-Whitney *U* test was run to compare differences between the Severe and the Non-Severe groups for specific non-parametric continuous variables (denoted as Wilcoxon's *W*). Data were presented as mean ± SD. Alpha value was set at 0.05.

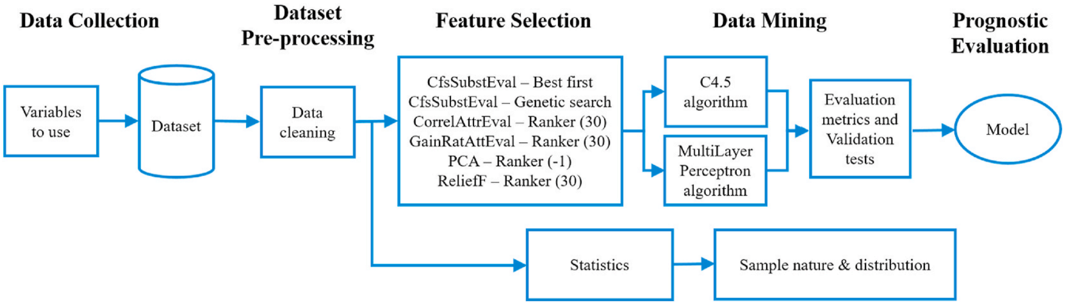
Data Collection
Dataset Pre-processing
Feature Selection
Data Mining
Prognostic Evaluation
Variables to use
Dataset
Data cleaning
CfsSubstEval – Best first
CfsSubstEval – Genetic search
CorrelAttrEval – Ranker (30)
GainRatAttEval – Ranker (30)
PCA – Ranker (-1)
ReliefF – Ranker (30)
C4.5 algorithm
MultiLayer Perceptron algorithm
Evaluation metrics and Validation tests
Model
Statistics
Sample nature & distribution

Fig. 1. Data mining and statistics workflow.

2.4. Data mining with machine learning algorithms

The C4.5 Decision Tree and the Multilayer Perceptron algorithms predicted the severity of COVID-19 cases powered by the most prevalent laboratory features. The widely used statistical classifier C4.5, also known as J48, is a supervised learning algorithm based on an objective-predefined variable (severity) used to generate a classification (decision) tree [34,35]. The Multilayer Perceptron (MLP), a class of feedforward artificial neural networks, performs machine recognition and finds patterns too complex and numerous for a human programmer to extract. Datasets were submitted to a K-fold Cross-Validation procedure (k defines the number folds in which to split a given dataset) to limit overfitting and evaluate the skill of the machine learning models and its default parameters to predict severity among COVID-19 patients. The value for k was fixed to 10, a value that has been found through experimentation to generally result in a model skill estimate with low bias and a modest variance [36].

2.5. Correlation matrix, hierarchical clustering analysis, and tanglegram

Variables of the datasets were also submitted to a bivariate Pearson's correlation matrix to measure the degree of linear relationships in either the Severe group or in the Non-Severe group. The matrices were then visualized using the *ggcorrplot* library of R to represent correlation strengths as a heatmap. Advanced hierarchical clustering analysis was applied to the color-mapped matrix to explore and summarize correlation datasets, given its visual form like a tree-shape dendrogram. The *Agnes* function with *Ward's* method showed the agglomerative hierarchical clustering of variables. Each leaf of the dendrogram corresponded to one observation (variable) and the fusion height showed the dissimilarity between two observations on the vertical axis. A cut height for cluster identification was calculated using the Average Silhouette method [37].

The tanglegram framework in R was selected to graphically compare the hierarchical grouping of variables with full linkage to *Ward*'s method and each combinatorial incongruence or misalignments identified visually. The similitude between trees was further computed by an entanglement coefficient between 0 and 1 [38]; where 1 meant a full entanglement and 0 no entanglement (the lower entanglement coefficient, the better alignment).

3. Results

3.1. Nature and distribution of the sample

There were a total of 4009 cases of COVID-19 patients, including 2007 males (50.1%) and 2002 females (49,9%). The sample was divided into "Severe" (489 cases, 12.2%; $PaO_2 = 49.4 \pm 8.7$ mmHg; $O_2Sat = 94.7 \pm 16.4\%$) and Non-Severe" (3520 cases, 87.8%; $PaO_2 = 72.1 \pm 9.5$ mmHg; $O_2Sat = 88.1 \pm 3.3\%$) groups based on a PaO_2 cut-off threshold of 60 mmHg (PaO_2 effect: $W = 121479$, $p < 0.001$; O_2Sat effect: $W = 267181$, $p < 0.001$). A total of 189 females (9.4%) and 300 males (14.9%) were classified as Severe, whereas 1813 females (90.6%) and 1707 males (85.1%) were considered Non-Severe. COVID-19 severity impacted males more than females (χ^2 (1) = 28382, $p < 0.001$) and increased with age ($W = 1.176 \times 10^6$, $p < 0.001$; Severe: 54.9 ± 20.3 years, Non-Severe: 41.4 ± 21.9 years).

3.2. Feature selection (FS)

Only those hematological and clinical biochemistry features with an incidence ≥ 3 in the frequency matrix (inclusion criteria) were meaningful for further analyses (a list of variables, their means and the reference values are available in the Supplementary Material file). From the original dataset, a total of 63 variables were cut down to 30 (see Table 1 for the abbreviations list) including age and sex. They were hematocrit (%), red blood cells ($\bullet 10^{12}$/L); median corpuscular volume (fL); red cell distribution width based on the standard deviation (fL); median corpuscular hemoglobin concentration or MCHC in g/dL; white blood cells ($\bullet 10^9$/L); neutrophils ($\bullet 10^9$/L); neutrophil %; monocytes ($\bullet 10^9$/L); monocytes % (MP); lymphocytes ($\bullet 10^9$/L); lymphocytes % (LP); eosinophils % (EP); neutrophil to lymphocyte ratio (NLR); leukocyte to C-reactive protein ratio (LeuCR); base excess (mEq/L); HCO_3-(mmol/L); chloride anion (mEq/L); $PaCO_2$ (mmHg); pH; lactate (mmol/L), glucose (mg/dL); prothrombin time (s); high sensitive C-reactive protein (ng/L); hemoglobin O_2 saturation (%). The following variables were added by supervised attribute selection: eosinophils ($\bullet 10^9$/L); lymphocytes-to-C-reactive protein ratio, and medium platelet volume-to-platelet ratio.

Table 1
List of abbreviations.

AEC	Absolute eosinophil cell count	LP	Lymphocyte cell %
ALC	Absolute lymphocyte cell count	LyCR	Lymphocytes-to-C-reactive protein ratio
AMC	Absolute monocyte count	MCHC	Median corpuscular hemoglobin concentration
ANC	Absolute neutrophil count	MCV	Median corpuscular volume
arff	Attribute-Relation File Format	MCV	Medium corpuscular volume
BE	Base excess	ML	Machine learning
CfsSubstEval	Correlation based Feature Selection subset evaluation function	MLP	Multilayer Perceptron algorithm
CorrelAttrEval	Correlation attribute evaluator	MP	Monocyte %
csv	Comma-separated values	MPR	Medium platelet volume-to-platelet ratio
CT	Computed tomography	NLR	Neutrophil-to-lymphocyte ratio
EP	Eosinophil %	NP	Neutrophil %
FS	Feature selection	O_2Sat	Hemoglobin O_2 saturation
GainRatAttEval	Gain ratio criterion evaluator	$PaCO_2$	Partial pressure of carbon dioxide
GLC	Glucose	PaO_2	Partial pressure of oxygen
HCT	Hematocrit	PCA	Principal Component Analysis
hs-CRP	High sensitive C-reactive protein	PSI	Pneumonia severity index
ICU	Intensive care unit	PT	Prothrombin time
IQR	Interquartile ranges	RBC	Red blood cells/ erythrocyte cell count
Lac	Lactate	RDW-SD	Red cell distribution width based on the standard deviation
IESS	Instituto Ecuatoriano de Seguridad Social	RT-PCR	Reverse transcription polymerase chain reaction
LeuCR	Leukocyte to C-reactive protein ratio	STROBE	Strengthening the reporting of observational studies in epidemiology

3.3. Classifying severe COVID-19 pneumonia by data mining and machine learning

The evaluation metrics of the algorithms when classifying patient's severity are shown in Table 2. The C4.5 and MLP algorithms showed the skill to predict severe COVID-19 pneumonia (89.6% and 96.5% precision respectively). Whereas the MLP algorithm computed the whole FS, the C4.5 algorithms only considered those variables with greater relevance for prediction (Fig. 2): $O_2Sat > pH > PaCO_2 >$ sex > lymphocytes > RDW-SD > eosinophils > lactate. Compared to C4.5, MLP had

Table 2
Algorithm evaluation metrics. Metric values for the C4.5 classification algorithm and the neural network Multilayer Perceptron (MLP) with performance parameters during cross-validation (10 folds).

Evaluation parameters	C4.5	MLP
True Positive Rate	0.894	0.965
False Positive Rate	0.136	0.036
Precision	0.896	0.965
Recall	0.894	0.965
F-Measure	0.892	0.965
Matthews correlation coefficient	0.779	0.928
ROC Area	0.977	0.988
PRC Area	0.972	0.983
Correctly Classified Instances (%)	89.3827	96.5432
Incorrectly Classified Instances (%)	10.6173	3.4568
Kappa statistic	0.7745	0.9282
Mean absolute error	0.1357	0.0396
Root mean squared error	0.2516	0.17
Relative absolute error (%)	28.2185	8.2407
Root relative squared error (%)	51.3025	34.666
Total Number of Instances	1215	1215

minimum errors in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE). In addition, the Kappa statistic value of MLP was 0.9282, which means that the prediction model was statistically significant.

3.4. *Correlation matrix, dendrograms, and tanglegram plot*

To avoid a high rate of false discoveries, only correlations whose r coefficients were either ≥ 0.8 or $\leq$ - 0.8 were considered clinically meaningful. A negative linear association between neutrophil and lymphocyte cell % was revealed by the correlation matrix in both the Non-Severe ($r =$ - 0.969, $p < 0.001$; Fig. 3b) and the Severe groups ($r =$ - 0.963, $p < 0.001$, Fig. 3a). Of note, the strong linear positive relationship between the Leukocytes-to-C-reactive protein (LeuCR) and the Lymphocytes-to-C-reactive protein (LyCR) ratios in the Non-Severe correlation matrix ($r = 0.829$, $p < 0.001$) waned in the correlation matrix of the Severe group ($r = 0.272$, $p < 0.001$).

The hierarchical clustering analysis extracted the bivariate Pearson correlation patterns in the form of dendrograms (Fig. 4). Meaningful clusters of variables were optimally compared, side by side, using a tanglegram plot (entanglement index = 0.41) for visual inspection of dissimilarities (Fig. 5). Combinatorial and topology changes included the hs-CRP, neutrophil %, $PaCO_2$, neutrophils %, and to less extent monocytes, lactate, PT, glucose, chloride anion, MCHC, and NLR variables.

4. Discussion

Using a large dataset of laboratory tests at the time of hospital admission, this study assessed the performance of data-driven knowledge algorithms for the detection of severe COVID-19 pneumonia. Besides the well-known variables with predicting ability of COVID-19 severity like sex and age, differential WBCs (e.g., NLR) and biomarkers of systemic inflammation (e.g., CRP), this study hints to features like blood acid-base balance (pH, pCO_2 and lactate), LeCR, LyCR, lymphocytes, RDW-SD, and eosinophils that may also predict poor clinical outcomes.

As expected, severe COVID-19 cases reached 12.2% of the sample [39–41], disease severity increased with age and males were more vulnerable than females [42–44]. Categorization of patient severity was based on the arterial blood oxygen exam, a diagnostic indicator of arterial hypoxemia in patients with community-acquired pneumonia [29]. Although COVID-19 patients experience many symptoms at many organs, our patient categorization were of predicting value for follow-up treatment plans, since COVID-19-induced respiratory failures (i.e., acute respiratory distress syndrome) causes multiple organ failures leading to death [45]. As proven by the pulse oximetry test [46], hemoglobin oxygen saturation in the Non-Severe group was almost normoxic (94.7%), whereas it was moderately hypoxic (88.1%) in the Severe group and below the recommended range (92–96%) for COVID-19 patients [47]. Importantly NLR, an indicator of adverse clinical outcomes in the progression of COVID-19 [19,23,48,49], reached 5.8 in the Severe group, whereas values above 4 do predict ICU admission [50,51]. Despite the arterial blood oxygen exam being conducted on the same day as the routine laboratory tests, this does not preclude from stating that this data mining study could predict the risk for a fatal COVID-19 evolution.

Data mining algorithms have only been used to discover symptom patterns in COVID-19 patients [52] as well as to make predictions about COVID-19 patient severity [2,7–9] and recovery [53]. The MLP

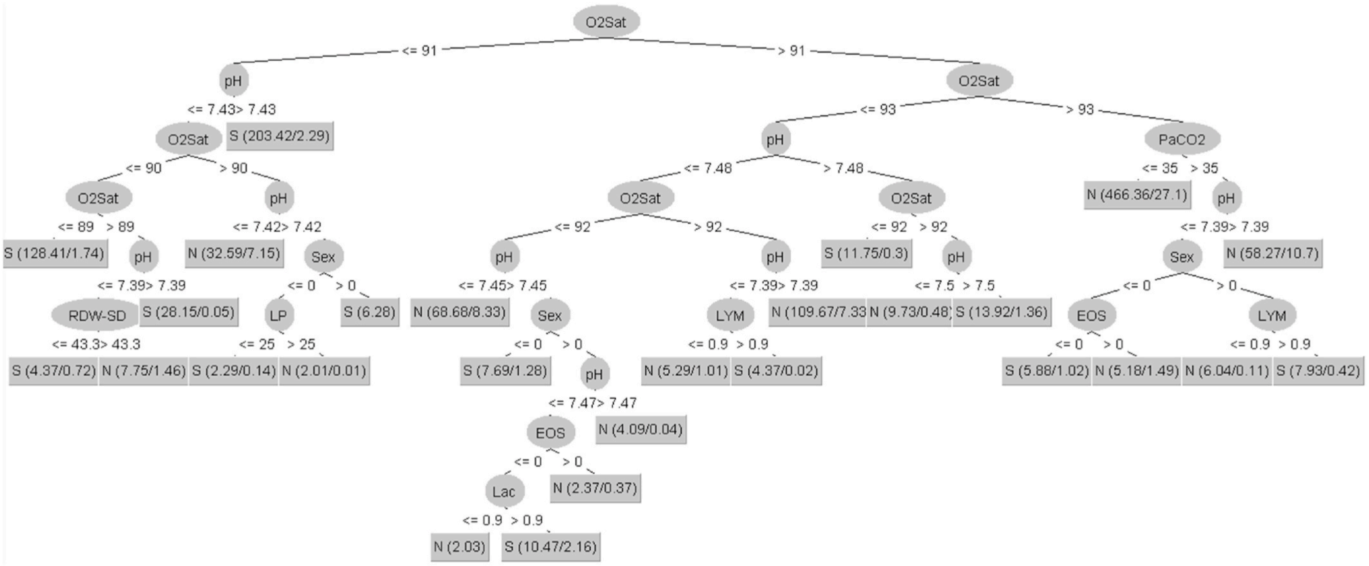

Fig. 2. C4.5 algorithm decision tree.
Notice that variables (oval) with the greatest relevance for prediction were: O_2Sat, pH, and $PaCO_2$ (critical feature values are shown in rectangles).
Abbreviations are shown in Table 1. Cl: chloride anion; N: Non-Severe; S: Severe.

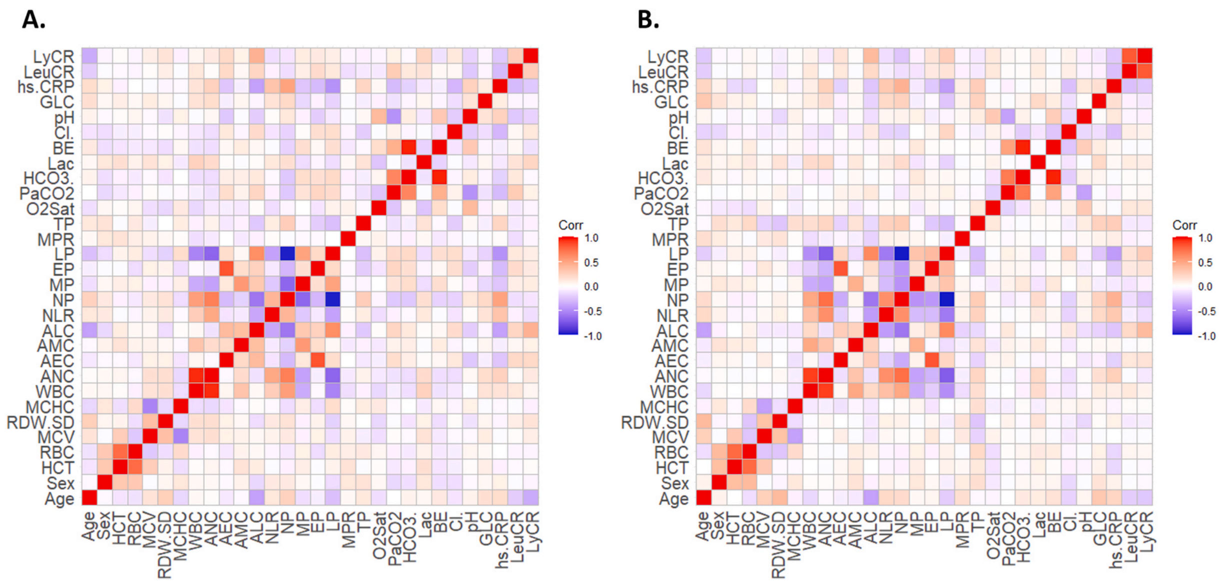

Fig. 3. Heatmap of the bivariate Pearson correlation matrix.

Correlation values are represented as colors. Red and blue are the perfect linear positive and negative relationships, respectively. White means non-linear relationship. Notice the negative neutrophils (%)-lymphocytes (%) linear correlation regardless of severity, and the positive linear LeuCR-LyCR correlation associated with Non-Severe. Left: Severe, Right: Non-Severe.

Abbreviations are shown in Table 1. Cl: chloride anion; HCO_3: bicarbonate anion.

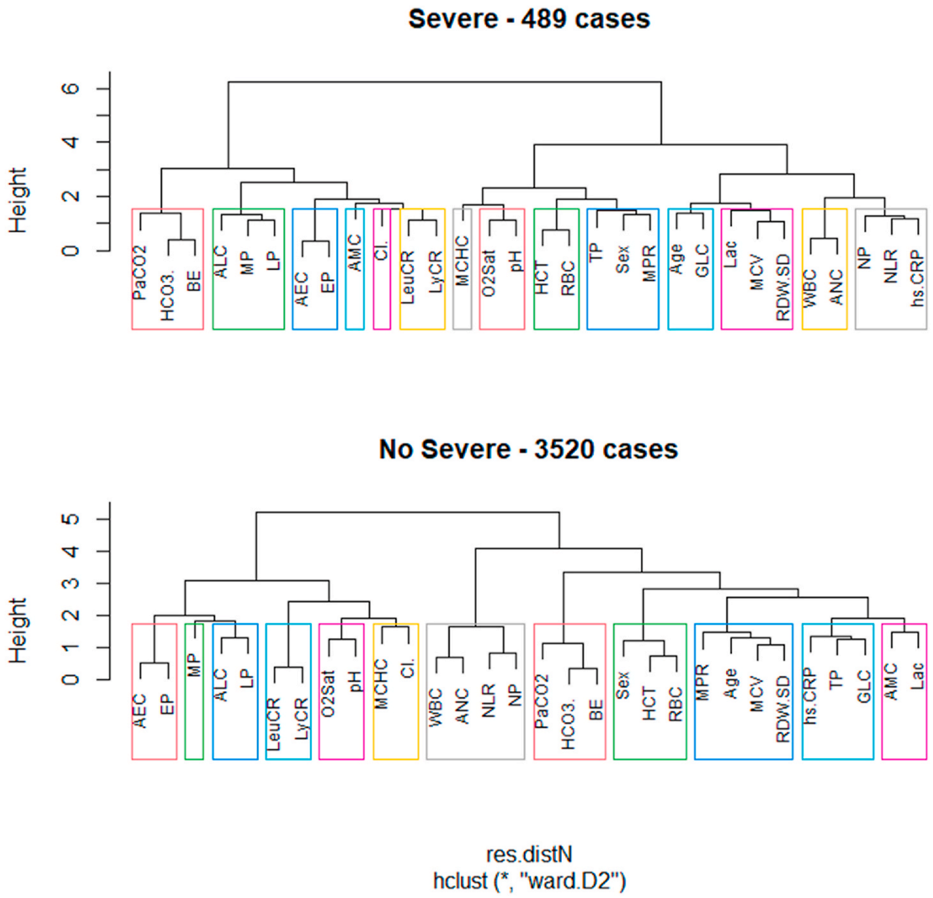

Fig. 4. Hierarchical clustering of the bivariate Pearson correlation matrix.

Dendrograms representing associations among features (variables). Branches length represents the distance between variables or clusters based on correlation patterns.

Abbreviations are shown in Table 1. Cl: chloride anion; HCO_3: bicarbonate anion.

algorithm classified the severe patients with a precision (96.5%), similar to XGBoost [9] and other multipurpose algorithms [7], but with the advantage of using laboratory tests. Its accuracy percentage was higher than the C4.5 algorithms 89.6%), which just relied on a selection of those variables with the greatest relevance, being blood pH the most striking one (Fig. 2). This finding may be interesting because, although

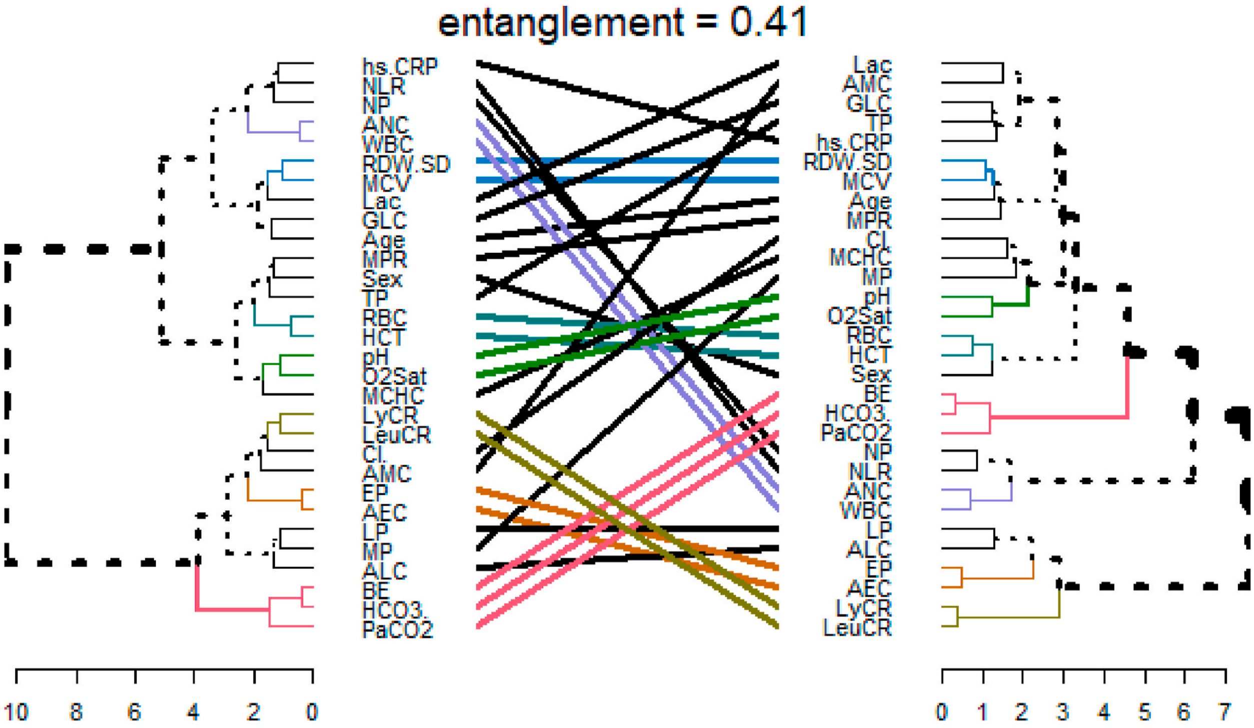

Fig. 5. Tanglegram plot.
A pair of trees (left: Severe, right: Non-Severe) of the same set of leaves with matching leaves in the two trees joined by an edge. Dashed lines represent topological changes identified by unique nodes with combinations of variables in the Non-severe group that were not present in the Severe group.
Abbreviations are shown in Table 1. Cl: chloride anion; HCO_3: bicarbonate anion.

pH differences between groups were modest, they could be enough to affect the function of the immune system [54] and to play a role in SARS-CoV-2 infectivity [55,56]. Hence, this study guarantees the utility of C4.5 and Multilayer Perceptron algorithms for patient research and stratification based on routine laboratory tests [57].

Given the data-driven knowledge approach of this analysis, the hidden structure of the datasets could give further clues about evolution to severe COVID-19. The correlation matrix (Fig. 3) revealed a the linear associations between LyCR [48,58] and LeuCR [59] in the Non-severe group (Fig. 3b) that waned in the Severe group (Fig. 3a), likely because of the possible functional exhaustion of antiviral lymphocytes [60]. Topological changes (Fig. 5) across Severe and Non-Severe dendrograms (Fig. 4) led to the discovery of subtle shift patterns of the internal environment that putatively involved the pro-inflammatory biomarker CRP [7,12,18,50,56], pCO_2 and neutrophil % in the progress to severe COVID-19 pneumonia. Neutrophils are critical components of the hyper-inflammatory process in severe COVID-19 [61] and their activity could be regulated by pCO_2 [62].

Inconsistency and conflicting data about prognostic/diagnostic biomarkers of COVID-19 severity and predicting modelling has affected the quality of studies dealing with the utility of artificial intelligence in the management of COVID patients [3]. The diversity of populations for retrospective cohorts, availability of different diagnostic tools across counties and a lack of explicit information on the predicting modeling methodology are important factors. Despite all the problems, working with retrospective laboratory data has been probed very valuable [57]. Herein, the precise patient categorization based on the COVID-19-induced pneumonia, the strict feature selection of the more influential variables, and the identification of the appropriate ML and other data mining analysis made the prediction process meet the expected objectives for patient risk stratification.

5. Summary

Data mining directly applied to routine laboratory tests discriminated with high precision COVID-19 patient's condition and the risk for ICU admission. Identified laboratory features like inflammatory biomarkers and changes in some WBC subpopulations correlated well with the literature and reaffirmed their role in COVID-19 disease progression at hematological level. Finally, our data mining supported the contention that even modest laboratory test deviations of blood pH, pCO_2, lactate and RDW-SD, typically in the "normal" range but consistently skewed in one direction, are strong enough to predict the poor clinical progression of the COVID-19 disease. The selection of data mining data-driven tools herein presented may hold promise to increase the power of clinical decision-making at the initial hospital care of COVID-19 patients.

Availability of data and materials

Data will be available on reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

Conceived and designed the study: SB, MF, MP, KC, EF, HR. Collected data: FM. Analyzed the data: MP, SB, KC, SM, MF All authors have approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all the patients whose data were used to report these findings. We are also grateful to the health professionals of the IESS COVID-19 Hospital Quito Sur in Quito (Ecuador) who treated the patients. To those who died from COVID-19, our condolences.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104738.

References

[1] Worldometer. COVID-19 CORONAVIRUS PANDEMIC., (n.d.), 2021. https://www.worldometers.info/coronavirus/. (Accessed 13 July 2021).

[2] A. Sharma, S. Rani, D. Gupta, Artificial intelligence-based classification of chest X-ray images into COVID-19 and other infectious diseases, Int. J. Biomed. Imag. 2020 (2020) 1–10, https://doi.org/10.1155/2020/8889023.

[3] L. Wynants, B. Van Calster, G.S. Collins, R.D. Riley, G. Heinze, E. Schuit, M.M. J. Bonten, D.L. Dahly, J.A.A. Damen, T.P.A. Debray, V.M.T. de Jong, M. De Vos, P. Dhiman, M.C. Haller, M.O. Harhay, L. Henckaerts, P. Heus, M. Kammer, N. Kreuzberger, A. Lohmann, K. Luijken, J. Ma, G.P. Martin, D.J. McLernon, C. L. Andaur Navarro, J.B. Reitsma, J.C. Sergeant, C. Shi, N. Skoetz, L.J.M. Smits, K.I. E. Snell, M. Sperrin, R. Spijker, E.W. Steyerberg, T. Takada, I. Tzoulaki, S.M.J. van Kuijk, B. van Bussel, I.C.C. van der Horst, F.S. van Royen, J.Y. Verbakel, C. Wallisch, J. Wilkinson, R. Wolff, L. Hooft, K.G.M. Moons, M. van Smeden, Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ 369 (2020) m1328, https://doi.org/10.1136/bmj.m1328.

[4] O.S. Albahri, J.R. Al-Obaidi, A.A. Zaidan, A.S. Albahri, B.B. Zaidan, M.M. Salih, A. Qays, K.A. Dawood, R.T. Mohammed, K.H. Abdulkareem, A.M. Aleesa, A. H. Alamoodi, M.A. Chyad, C.Z. Zulkifli, Helping doctors hasten COVID-19 treatment: towards a rescue framework for the transfusion of best convalescent plasma to the most critical patients based on biological requirements via ml and novel MCDM methods, Comput. Methods Progr. Biomed. 196 (2020) 105617, https://doi.org/10.1016/j.cmpb.2020.105617.

[5] L.J. Muhammad, M.M. Islam, S.S. Usman, S.I. Ayon, Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery, SN comput. Sci. 1 (2020) 206, https://doi.org/10.1007/s42979-020-00216-w.

[6] N. Alballa, I. Al-Turaiki, Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: a review, Informatics Med. Unlocked. 24 (2021), 100564, https://doi.org/10.1016/j.imu.2021.100564.

[7] F.T. Fernandes, T.A. de Oliveira, C.E. Teixeira, A.F. de M. Batista, G. Dalla Costa, A. D.P. Chiavegatto Filho, A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil, Sci. Rep. 11 (2021) 3343, https://doi.org/10.1038/s41598-021-82885-y.

[8] K. Gong, D. Wu, C.D. Arru, F. Homayounieh, N. Neumark, J. Guan, V. Buch, K. Kim, B.C. Bizzo, H. Ren, W.Y. Tak, S.Y. Park, Y.R. Lee, M.K. Kang, J.G. Park, A. Carriero, L. Saba, M. Masjedi, H. Talari, R. Babaei, H.K. Mobin, S. Ebrahimian, N. Guo, S. R. Digumarthy, I. Dayan, M.K. Kalra, Q. Li, A multi-center study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records, Eur. J. Radiol. 139 (2021), 109583, https://doi.org/10.1016/j.ejrad.2021.109583.

[9] I. Shiri, M. Sorouri, P. Geramifar, M. Nazari, M. Abdollahi, Y. Salimi, B. Khosravi, D. Askari, L. Aghaghazvini, G. Hajianfar, A. Kasaeian, H. Abdollahi, H. Arabi, A. Rahmim, A.R. Radmard, H. Zaidi, Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients, Comput. Biol. Med. 132 (2021), 104304, https://doi.org/10.1016/j.compbiomed.2021.104304.

[10] B. Khosravi, L. Aghaghazvini, M. Sorouri, S. Naybandi Atashi, M. Abdollahi, H. Mojtabavi, M. Khodabakhshi, F. Motamedi, F. Azizi, Z. Rajabi, A. Kasaeian, A. R. Sima, A.H. Davarpanah, A.R. Radmard, Predictive value of initial CT scan for various adverse outcomes in patients with COVID-19 pneumonia, Heart Lung 50 (2021) 13–20, https://doi.org/10.1016/j.hrtlng.2020.10.005.

[11] S. Imai, M. Akahane, T. Imamura, Computed tomography: return on investment and regional disparity factor Analysis, Front. Public Heal 6 (2019), https://doi.org/10.3389/fpubh.2018.00380.

[12] G. Bonetti, F. Manelli, A. Patroni, A. Bettinardi, G. Borrelli, G. Fiordalisi, A. Marino, A. Menolfi, S. Saggini, R. Volpi, A. Anesi, G. Lippi, Laboratory predictors of death from coronavirus disease 2019 (COVID-19) in the area of Valcamonica, Italy, Clin. Chem. Lab. Med. 58 (2020) 1100–1105, https://doi.org/10.1515/cclm-2020-0459.

[13] B.H. Foy, J.C.T. Carlson, E. Reinertsen, R. Padros I. Valls, R. Pallares Lopez, E. Palanques-Tost, C. Mow, M.B. Westover, A.D. Aguirre, J.M. Higgins, Association of red blood cell distribution width with mortality risk in hospitalized adults with SARS-CoV-2 infection, JAMA Netw. Open. 3 (2020), e2022058, https://doi.org/10.1001/jamanetworkopen.2020.22058.

[14] X. Liu, R. Zhang, G. He, Hematological findings in coronavirus disease 2019: indications of progression of disease, Ann. Hematol. 99 (2020) 1421–1428, https://doi.org/10.1007/s00277-020-04103-5.

[15] G. Ponti, M. Maccaferri, C. Ruini, A. Tomasi, T. Ozben, Biomarkers associated with COVID-19 disease progression, Crit. Rev. Clin. Lab Sci. (2020) 1–11, https://doi.org/10.1080/10408363.2020.1770685.

[16] S. Sun, X. Cai, H. Wang, G. He, Y. Lin, B. Lu, C. Chen, Y. Pan, X. Hu, Abnormalities of peripheral blood system in patients with COVID-19 in Wenzhou, China, Clin. Chim. Acta 507 (2020) 174–180, https://doi.org/10.1016/j.cca.2020.04.024.

[17] C. Wang, R. Deng, L. Gou, Z. Fu, X. Zhang, F. Shao, G. Wang, W. Fu, J. Xiao, X. Ding, T. Li, X. Xiao, C. Li, Preliminary study to identify severe from moderate cases of COVID-19 using combined hematology parameters, Ann. Transl. Med. 8 (2020), https://doi.org/10.21037/atm-20-3391, 593–593.

[18] X. Yuan, W. Huang, B. Ye, C. Chen, R. Huang, F. Wu, Q. Wei, W. Zhang, J. Hu, Changes of hematological and immunological parameters in COVID-19 patients, Int. J. Hematol. 112 (2020) 553–559, https://doi.org/10.1007/s12185-020-02930-w.

[19] X. Li, C. Liu, Z. Mao, M. Xiao, L. Wang, S. Qi, F. Zhou, Predictive values of neutrophil-to-lymphocyte ratio on disease severity and mortality in COVID-19 patients: a systematic review and meta-analysis, Crit. Care 24 (2020) 647, https://doi.org/10.1186/s13054-020-03374-8.

[20] S.J. Ballaz, M. Pulgar-Sánchez, K. Chamorro, E. Fernández-Moreira, H. Ramírez, F. X. Mora, M. Fors, Common laboratory tests as indicators of COVID-19 severity on admission at high altitude: a single-center retrospective study in Quito (Ecuador), Clin. Chem. Lab. Med. 59 (2021), https://doi.org/10.1515/cclm-2021-0156 e326–e329.

[21] R. Sardar, A. Sharma, D. Gupta, Machine learning assisted prediction of prognostic biomarkers associated with COVID-19, using clinical and proteomics data, Front. Genet. 12 (2021), https://doi.org/10.3389/fgene.2021.636441.

[22] F. Paiva Proença Lobo Lopes, F.C. Kitamura, G.F. Prado, P.E. de A. Kuriki, M.R. T. Garcia, Machine learning model for predicting severity prognosis in patients infected with COVID-19: study protocol from COVID-AI Brasil, PloS One 16 (2021), e0245384, https://doi.org/10.1371/journal.pone.0245384.

[23] X. Wang, X. Li, Y. Shang, J. Wang, X. Zhang, D. Su, S. Zhao, Q. Wang, L. Liu, Y. Li, H. Chen, Ratios of neutrophil-to-lymphocyte and platelet-to-lymphocyte predict all-cause mortality in inpatients with coronavirus disease 2019 (COVID-19): a retrospective cohort study in a single medical centre, Epidemiol. Infect. 148 (2020) e211, https://doi.org/10.1017/S0950268820002071.

[24] R Foundation The, The R Project for Statistical Computing, 2020.

[25] University of Waikato, WEKA Software, 2021. http://www.cs.waikato.ac.nz/ml/weka/.

[26] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, B. Goethals, Mining Association Rules in Graphs Based on Frequent Cohesive Itemsets, 2015, pp. 637–648, https://doi.org/10.1007/978-3-319-18032-8_50.

[27] B.M. Henry, M.H.S. de Oliveira, S. Benoit, M. Plebani, G. Lippi, Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis, Clin. Chem. Lab. Med. 58 (2020) 1021–1028, https://doi.org/10.1515/cclm-2020-0369.

[28] Ministerio de Salud Pública, Ecuador, Reglamento de los comités de ética de investigación en seres humanos, 2014. https://www.gob.ec/sites/default/files/regulations/2018-10/Documento_Reglamento Comités Ética Investigación Seres Humanos.pdf. (Accessed 5 January 2021).

[29] K.P. Levin, B.H. Hanusa, A. Rotondi, D.E. Singer, C.M. Coley, T.J. Marrie, W. N. Kapoor, M.J. Fine, Arterial blood gas and pulse oximetry in initial management of patients with community-acquired pneumonia, J. Gen. Intern. Med. 16 (2001) 590–598, https://doi.org/10.1046/j.1525-1497.2001.016009590.x.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (2009) 10–18, https://doi.org/10.1145/1656274.1656278.

[31] M.A. Hall, Correlation-based Feature Selection for Machine Learning, 1999.

[32] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Mach. Learn. 53 (2003) 23–69, https://doi.org/10.1023/A:1025667309714.

[33] M. Trabelsi, N. Meddouri, M. Maddouri, A new feature selection method for nominal classifier based on formal concept analysis, Procedia Comput. Sci. 112 (2017) 186–194, https://doi.org/10.1016/j.procs.2017.08.227.

[34] S.L. Salzberg, Programs for machine learning by J. Ross quinlan. Morgan kaufmann publishers, Inc, Mach. Learn. 16 (1994) 235–240, https://doi.org/10.1007/BF00993309, 1993.

[35] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2011, https://doi.org/10.1016/C2009-0-19715-5.

[36] G. James, D. Witten, T. Hastie, R. Tibshirani, Cross-validation, in: An Introduction to Statistic Learning. With Application in R, 2013, pp. 175–187, https://doi.org/10.1007/978-1-4614-7138-7.

[37] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley \& Sons, 2009.

[38] T. Galili, dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering, Bioinformatics 31 (2015) 3718–3720, https://doi.org/10.1093/bioinformatics/btv428.

[39] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, L. Zhang, Epidemiological and clinical characteristics of 99

cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study, Lancet 395 (2020) 507–513, https://doi.org/10.1016/S0140-6736(20)30211-7.

[40] W.A. Pongpirul, S. Wiboonchutikul, L. Charoenpong, N. Panitantum, A. Vachiraphan, S. Uttayamakul, K. Pongpirul, W. Manosuthi, W. Prasithsirikul, Clinical course and potential predictive factors for pneumonia of adult patients with Coronavirus Disease 2019 (COVID-19): a retrospective observational analysis of 193 confirmed cases in Thailand, PLoS Neglected Trop. Dis. 14 (2020), e0008806, https://doi.org/10.1371/journal.pntd.0008806.

[41] S. Tian, N. Hu, J. Lou, K. Chen, X. Kang, Z. Xiang, H. Chen, D. Wang, N. Liu, D. Liu, G. Chen, Y. Zhang, D. Li, J. Li, H. Lian, S. Niu, L. Zhang, J. Zhang, Characteristics of COVID-19 infection in beijing, J. Infect. 80 (2020) 401–406, https://doi.org/10.1016/j.jinf.2020.02.018.

[42] A. Borghesi, A. Zigliani, R. Masciullo, S. Golemi, P. Maculotti, D. Farina, R. Maroldi, Radiographic severity index in COVID-19 pneumonia: relationship to age and sex in 783 Italian patients, Radiol. Med. 125 (2020) 461–464, https://doi.org/10.1007/s11547-020-01202-1.

[43] A. Ortolan, M. Lorenzin, M. Felicetti, A. Doria, R. Ramonda, Does gender influence clinical expression and disease outcomes in COVID-19? A systematic review and meta-analysis, Int. J. Infect. Dis. 99 (2020) 496–504, https://doi.org/10.1016/j.ijid.2020.07.076.

[44] A. Pradhan, P.-E. Olsson, Sex differences in severity and mortality from COVID-19: are males more vulnerable? Biol. Sex Differ. 11 (2020) 53, https://doi.org/10.1186/s13293-020-00330-7.

[45] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen, B. Cao, Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study, Lancet 395 (2020) 1054–1062, https://doi.org/10.1016/S0140-6736(20)30566-3.

[46] A.T. Society/, A.C. Of C. Physicians, ATS/ACCP statement on cardiopulmonary exercise testing, Am. J. Respir. Crit. Care Med. 167 (2003) 211–277, https://doi.org/10.1164/rccm.167.2.211.

[47] N. Shenoy, R. Luchtel, P. Gulani, Considerations for target oxygen saturation in COVID-19 patients: are we under-shooting? BMC Med. 18 (2020) 260, https://doi.org/10.1186/s12916-020-01735-2.

[48] W. Ullah, B. Basyal, S. Tariq, T. Almas, R. Saeed, S. Roomi, S. Haq, J. Madara, M. Boigon, D.C. Haas, D.L. Fischman, Lymphocyte-to-C-Reactive protein ratio: a novel predictor of adverse outcomes in COVID-19, J. Clin. Med. Res. 12 (2020) 415–422, https://doi.org/10.14740/jocmr4227.

[49] Y. Yufei, L. Mingli, L. Xuejiao, D. Xuemei, J. Yiming, Q. Qin, S. Hui, G. Jie, Utility of the neutrophil-to-lymphocyte ratio and C-reactive protein level for coronavirus disease 2019 (COVID-19), Scand. J. Clin. Lab. Invest. 80 (2020) 536–540, https://doi.org/10.1080/00365513.2020.1803587.

[50] A. Ciccullo, A. Borghetti, L. Zileri Dal Verme, A. Tosoni, F. Lombardi, M. Garcovich, F. Biscetti, M. Montalto, R. Cauda, S. Di Giambenedetto, Neutrophil-to-lymphocyte ratio and clinical outcome in COVID-19: a report from the Italian front line, Int. J. Antimicrob. Agents 56 (2020), 106017, https://doi.org/10.1016/j.ijantimicag.2020.106017.

[51] C. Qin, L. Zhou, Z. Hu, S. Zhang, S. Yang, Y. Tao, C. Xie, K. Ma, K. Shang, W. Wang, D.-S. Tian, Dysregulation of immune response in patients with coronavirus 2019 (COVID-19) in wuhan, China,, Clin. Infect. Dis. 71 (2020) 762–768, https://doi.org/10.1093/cid/ciaa248.

[52] M. Tandan, Y. Acharya, S. Pokharel, M. Timilsina, Discovering symptom patterns of COVID-19 patients using association rule mining, Comput. Biol. Med. 131 (2021) 104249, https://doi.org/10.1016/j.compbiomed.2021.104249.

[53] L.J. Muhammad, M.M. Islam, S.S. Usman, S.I. Ayon, Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery, SN Comput. Sci. 1 (2020) 206, https://doi.org/10.1007/s42979-020-00216-w.

[54] F. Erra Díaz, E. Dantas, J. Geffner, Unravelling the interplay between extracellular acidosis and immune cells, Mediat. Inflamm. 2018 (2018) 1–11, https://doi.org/10.1155/2018/1218297.

[55] Z.-Y. Yang, Y. Huang, L. Ganesh, K. Leung, W.-P. Kong, O. Schwartz, K. Subbarao, G.J. Nabel, pH-dependent entry of severe acute respiratory syndrome coronavirus is mediated by the spike glycoprotein and enhanced by dendritic cell transfer through DC-SIGN, J. Virol. 78 (2004) 5642–5650, https://doi.org/10.1128/JVI.78.11.5642-5650.2004.

[56] B. Wang, Adjusting extracellular pH to prevent entry of SARS-CoV-2 into human cells, Genome 64 (2021) 595–598, https://doi.org/10.1139/gen-2020-0167.

[57] A. Alnor, M.B. Sandberg, C. Gils, P.J. Vinholt, Laboratory tests and outcome for patients with coronavirus disease 2019: a systematic review and meta-analysis, J. Appl. Lab. Med. 5 (2020) 1038–1049, https://doi.org/10.1093/jalm/jfaa098.

[58] S. Ghahramani, R. Tabrizi, K.B. Lankarani, S.M.A. Kashani, S. Rezaei, N. Zeidi, M. Akbari, S.T. Heydari, H. Akbari, P. Nowrouzi-Sohrabi, F. Ahmadizar, Laboratory features of severe vs. non-severe COVID-19 patients in Asian populations: a systematic review and meta-analysis, Eur. J. Med. Res. 25 (2020) 1–10, https://doi.org/10.1186/s40001-020-00432-3.

[59] T. Yamada, M. Wakabayashi, T. Yamaji, N. Chopra, T. Mikami, H. Miyashita, S. Miyashita, Value of leukocytosis and elevated C-reactive protein in predicting severe coronavirus 2019 (COVID-19): a systematic review and meta-analysis, Clin. Chim. Acta 509 (2020) 235–243, https://doi.org/10.1016/j.cca.2020.06.008.

[60] M. Zheng, Y. Gao, G. Wang, G. Song, S. Liu, D. Sun, Y. Xu, Z. Tian, Functional exhaustion of antiviral lymphocytes in COVID-19 patients, Cell. Mol. Immunol. 17 (2020) 533–535, https://doi.org/10.1038/s41423-020-0402-2.

[61] M.Z. Tay, C.M. Poh, L. Rénia, P.A. MacAry, L.F.P. Ng, The trinity of COVID-19: immunity, inflammation and intervention, Nat. Rev. Immunol. 20 (2020) 363–374, https://doi.org/10.1038/s41577-020-0311-8.

[62] R.J. Coakley, C. Taggart, C. Greene, N.G. McElvaney, S.J. O'Neill, Ambient pCO2 modulates intracellular pH, intracellular oxidant generation, and interleukin-8 secretion in human neutrophils, J. Leukoc. Biol. 71 (2002) 603–610. http://www.ncbi.nlm.nih.gov/pubmed/11927646.