Learning a battery of COVID-19 mortality prediction models by multi-objective optimization

Mario Martínez-García[1][0000−0002−6849−6239], Susana García-Gutierrez[2][0000−0002−8474−2479], Rubén Armañanzas[1][0000−0003−4049−0000], Adrián Díaz[1][0000−0002−2876−6177], Iñaki Inza[3][0000−0003−4674−1755], and Jose A. Lozano[1,3][0000−0002−4683−8111]

[1] Basque Center for Applied Mathematics, BCAM, Bilbao, Spain
{mmartinez, rarmananzas, adiaz, jlozano} @bcamath.org
[2] Galdakao Hospital, Osakidetza, Basque Country, Spain
susana.garciagutierrez@osakidetza.eus
[3] University of the Basque Country UPV/EHU, Computer Science Faculty, San Sebastián, Spain
inaki.inza@ehu.es

Abstract. The COVID-19 pandemic is continuously evolving with drastically changing epidemiological situations which are approached with different decisions: from the reduction of fatalities to even the selection of patients with the highest probability of survival in critical clinical situations. Motivated by this, a battery of mortality prediction models with different performances has been developed to assist physicians and hospital managers. Logistic regression, one of the most popular classifiers within the clinical field, has been chosen as the basis for the generation of our models. Whilst a standard logistic regression only learns a single model focusing on improving accuracy, we propose to extend the possibilities of logistic regression by focusing on sensitivity and specificity. Hence, the log-likelihood function, used to calculate the coefficients in the logistic model, is split into two objective functions: one representing the survivors and the other for the deceased class. A multi-objective optimization process is undertaken on both functions in order to find the Pareto set, composed of models not improved by another model in both objective functions simultaneously. The individual optimization of either sensitivity (deceased patients) or specificity (survivors) criteria may be conflicting objectives because the improvement of one can imply the worsening of the other. Nonetheless, this conflict guarantees the output of a battery of diverse prediction models. Furthermore, a specific methodology for the evaluation of the Pareto models is proposed. As a result, a battery of COVID-19 mortality prediction models is obtained to assist physicians in decision-making for specific epidemiological situations.

Keywords: COVID-19 · Mortality prediction · Multi-objective optimization · Classification evaluation

1 Introduction

The entire world has been paralyzed due to a virus, COVID-19, with unusually high levels of mortality and transmission. As of 10 January 2022, the numbers are still rising, with around 290 million infections and 5.5 million deaths since the beginning of the pandemic [1]. The fear and bewilderment experienced during the first months motivated researchers from all over the world to provide valuable information in the fight against COVID-19. The need to anticipate and correctly identify an early prognosis became an urgent challenge. Artificial intelligence through machine learning (ML) was the perfect tool to address this problem.

From the multitude of papers published, Wynants et al. [9] developed a review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19. Out of the hundreds of models collected, only the contributions of Yann et al. [10] and Knight et al. [7] were identified as clinically relevant. Yann et al. [10] proposed a mortality model trained and tested on patient data obtained just one day before discharge, using XGBoost as a classifier. Knight et al. [7] compute from a logistic regression an index between 0 and 21 that establishes a prognosis of the patient's risk mortality. Subsequently, Gupta et al. [5], from the same research group as the previous work, propose a deterioration model based on a logistic regression model.

All these models provide information of interest to healthcare professionals when making final decisions. However, the continual changes in the epidemiological situation mean that having only a single model is limited, and non-useful for physicians. At some pandemic stages physicians seek to reduce the number of deceased by improving the sensitivity of the model i.e., focusing on decreasing the number of patients predicted as surviving who subsequently decease. Nonetheless, when the resources are limited or health centres are overcrowded, focusing on improving the specificity of the model i.e., reducing the number of patients detected as deceased who subsequently survive, is a realistic option for physicians. In line with this trend, we propose not only a single model based on a single metric, but a battery of models with a diverse spectrum of performances in both areas of interest. Hence, depending on the pandemic stage, physicians will have the possibility of selecting the most suitable model.

For this purpose, a set of logistic regression models is trained in a specific way. The common log-likelihood function used to learn the logistic regression coefficients is divided into two different objective functions: one focused on deceased and the other on survivors. A multi-objective optimization is applied to both objective functions to obtain a set of models, known as Pareto or non-dominated set which can not be improved by another model in both objective functions simultaneously. Furthermore, a specific methodology for the evaluation of the Pareto models is proposed. Consequently, a battery of non-dominated COVID-19 mortality prediction models is obtained.

The paper is organized as follows. Section 2 covers data collection and preprocessing. Section 3, the design and development aspects: an in-depth explanation of the multi-objective optimization problem, the method for the evaluation

of the Pareto models and the validation of the models. Section 4 presents the final results and Section 5 a brief conclusion.

2 Clinical dataset

2.1 Data collection and characteristics

Osakidetza, the Basque Country public health service in Spain, made a prospective cohort study recruiting patients infected by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) confirmed by naso- and/or oropharyngeal swab polymerase chain reaction (PCR). Collected data contains blood tests, demographic and clinical data from the emergency department or up to 24 hours after hospital admission. The target, mortality, indicates infected deceased and hospital discharges labeled as survivors. Furthermore, all patients in the study are from Basque Country hospitals and pertain to the first (from February to April 2020) and second wave (from July to November 2020).

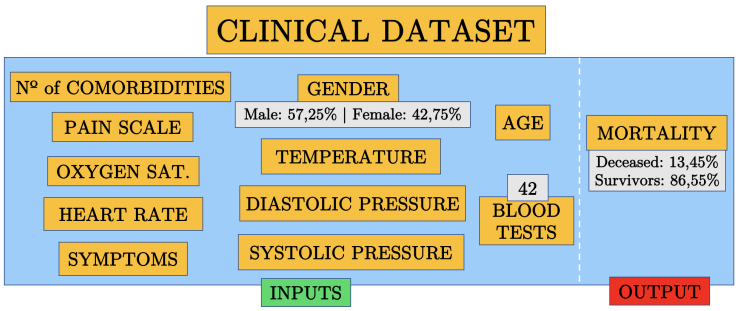


Fig. 1. Structure of the clinical dataset composed of demographic, clinical features and 42 different blood tests as inputs. Mortality represents the output of our model. (see Appendix)

2.2 Data pre-processing

In order to pre-process the data, we start by analyzing the distribution of values. Those features with unexpected distributions are studied in detail, contrasting information about their range and establishing valid ranges for collected data. All those features with a coherent distribution did not undergo any range modification. Apart from this modification, two filters are applied to treat missing values, one filter on the features and another on the patients [2].

- Feature filter. Blood tests, demographic and clinical features with more than 30% of missing values are removed from the study.
- Patient filter. Patients with three or more missing values in the features (blood tests, demographic or clinical data) are removed for further analysis.

Remaining missing values are afterwards imputed by unsupervised similarity [8]. Specifically, a five nearest-neighbours method with Euclidean distance is used to impute the data.

2.3 Final dataset

Finally, a total of 2215 patients and 53 features are retained (see Fig. 1). The final cohort is unbalanced with many more survivors than deceased (86.55% vs. 13.45%), and the sex distribution is balanced between male and female (57.25% vs. 42.75%). All features used in the modelling process are compiled in the Appendix. Moreover, features are normalized before starting with the development of the models.

3 Design and development aspects

Standard logistic regression returns a single model with specific evaluation scores that is not useful in changing epidemiological situations. However, having a battery of mortality models with different performances allows physicians and hospital managers to select the right model for a specific pandemic scenario. With this objective in mind, logistic regression coefficients are obtained by focusing on both sensitivity and specificity scores [6]. Instead of using the log-likelihood function (see Eq. 1), new functions are obtained from this one.

$$J(\theta) = -\sum_{i=1}^{N} y_i \cdot log(P_i(\theta)) + (1 - y_i) \cdot log(1 - P_i(\theta)) \tag{1}$$

The log-likelihood function is composed of the class (y_i), a summation on the N instances and the sigmoidal function (P_i) in terms of the coefficients (θ). As we address a binary mortality problem, the log likelihood function could be split into two objective functions: one for survivors (class 0) and the other for deceased patients (class 1) (see Eq. 2 and 3 respectively).

$$J_0(\theta) = -\sum_{i=1}^{G} (1 - y_i) \cdot log(1 - P_i(\theta)) \tag{2}$$

$$J_1(\theta) = -\sum_{i=1}^{K} y_i \cdot log(P_i(\theta)) \tag{3}$$

The instances ($N = G + K$) of the problem are divided into G and K survivors and deceased patients, respectively. The two new objective functions allow us to focus on key metrics for physicians when choosing a model: specificity (recall 0) and sensitivity (recall 1). Instead of performing a complex optimization of sensitivity and specificity scores, we have opted for a straightforward process: the optimization of J_0 and J_1 functions equivalent to these scores.

However, it is not feasible to compute the coefficients by gradient descent with two objective functions. Therefore, we rely on multi-objective optimization to address this issue. It is worth noting that the individual optimization of J_0, whose minimization optimizes the specificity, or J_1, whose minimization optimizes sensitivity, may be conflicting objectives. The improvement of one of them may surely imply the worsening of the other. By means of the multi-objective optimization paradigm we try to find a set of diverse models: some with balanced performances, others focused on specificity, and others on sensitivity.

3.1 Multi-objective optimization

Multi-objective optimization provides the ability to address the problem in the exposed way:

- Computation of **logistic regression coefficients**. The paradigm seeks model coefficients that optimize both objective functions in order to maximize the performance of the model.
- Obtaining a **battery of models**, known as a Pareto set, with different performance scores. The Pareto set is composed of models not improved by another model in both objective functions simultaneously.
- **Resolution of the imbalance problem**. Separation of the log-likelihood into two class-dependent functions causes both classes to have the same relevance when applying the multi-objective optimization procedure.

Multi-objective optimization development. The multi-objective optimization is undertaken by one of the most popular implementations called non-dominated sorting genetic algorithm II (NSGA-II) [3, 4].

NSGA-II procedure (see Fig. 2) is adopted taking as the individuals of the population the coefficients of a logistic regression. Note that any unspecified steps are matched to the generic one of the algorithm. The process begins with a parent population P_0 composed of M different models, where coefficients are randomly selected by means of Latin hypercube sampling. A non-domination sorting, divided into different fronts, is carried out over the pair of objective functions. After that, an individual selection, a simulated binary crossover and a polynomial mutation are used to create an offspring population Q_0, of size M. Thus, the first generation is obtained.

The procedure is different for the next generations. For the $t-th$ generation of the genetic algorithm, a combined population $R_t = P_t \cup Q_t$ of size $2M$ is initially formed. Then, population R_t is sorted according to non-domination and the best M models are selected. In order to choose exactly M models, the models of the last non-dominated front are sorted by crowding distance sorting. Consecutively, a new population P_{t+1} is obtained and used for subsequent individual selection, simulated binary crossover and polynomial mutation in order to create a new offspring population Q_{t+1}. This procedure is repeated for a number of generations established. As a result, a Pareto set of solutions with their associated objective functions values and models coefficients is obtained.

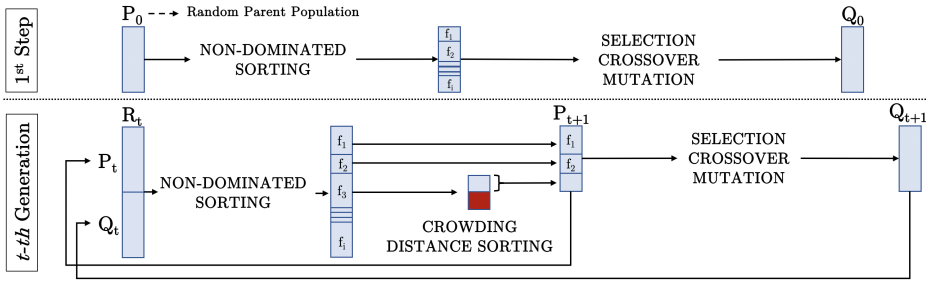


Fig. 2. NSGA-II procedure.

In our implementation the initial population is fixed to 500 random models, the number of generations is 200 and a constraint on the objective functions values ($J_0 + J_1 < 1.5$) is applied to mitigate bias.

3.2 Validation of the models

In order to maximize the representativeness of Pareto set models, our aim is to implement NSGA-II on the entire data cohort. Nonetheless, the validation of the models from the obtained Pareto set is not trivial. Performing a cross-validation is not possible because the models obtained in the Pareto set of each fold are different and no relationship between them can be established. Therefore, we propose a novel Method for the Evaluation of PAreto Models (MEPAM). The feasibility of the method is studied by comparing MEPAM's performance internally estimated in a train partition with the performance estimated in an external test set. Once MEPAM is accepted as feasible, the process is applied to the full cohort in order to validate the final models that will be deployed. Note that both used scores, sensitivity (recall 1) and specificity (recall 0), are called recalls.

Method for the evaluation of Pareto models (MEPAM). We propose a method (see Fig. 3) for the evaluation of the models located in the Pareto set obtained with a entire data cohort. Specifically, the evaluation of the models consists of assigning to each model a recall value for each of the classes. It is needed to note that the method is described for a generic dataset. The following sections show how MEPAM is implemented on our dataset.

First of all, the multi-objective optimization framework is implemented on the **entire dataset** and a Pareto set is obtained (See (*) Fig. 3). All of the models obtained in this Pareto set are the ones we want to validate. For this purpose, and as the core of the method, the dataset is also split into **four stratified folds**: four training sets with their respective test sets. The objective of the four stratified folds is to obtain a representative set of validated models in order to be able to infer a realistic evaluation of the models from the Pareto set of the entire dataset. Thus, for each of the train subsets, the multi-objective optimization

problem is solved by NSGA-II obtaining four different Pareto sets. Moreover, each Pareto set is evaluated in its respective test fold generating a pair of recalls (R_0, R_1) for each model (See (**) Fig. 3).

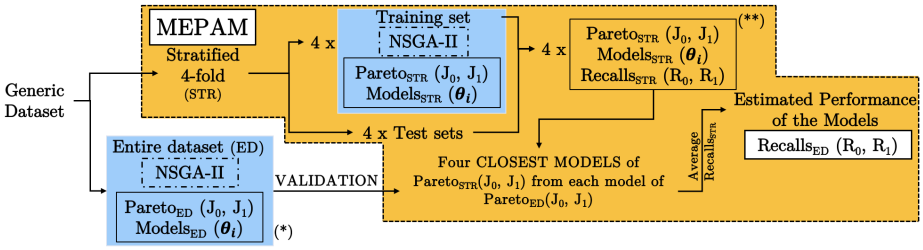


Fig. 3. Workflow for the validation of Pareto set models from a generic dataset. Method for the Evaluation of Pareto Models (MEPAM) is described within the yellow shading.

Accordingly, we proceed to validate the models from the Pareto set of the entire dataset by a recall estimation. For the validation of a single model from the Pareto set of the entire dataset, we focus on its objective functions (J_0, J_1) and collect the models with the four closest existing objective function pairs in the four Pareto sets of the stratified folds. Euclidean distance is used to collect these models for which recalls (R_0, R_1) are known. The mean of the R_0 recall of these models is considered as the estimated recall R_0 for the model to validate. The same procedure is followed for the generation of the estimated recall R_1. By repeating the process for each of the models included in the Pareto set of the entire dataset, an evaluation of the Pareto models is achieved.

MEPAM feasibility. After the explanation of the method, its feasibility (see Fig. 4) is studied on our data cohort, which is divided into train (80%) and test (20%) sets.

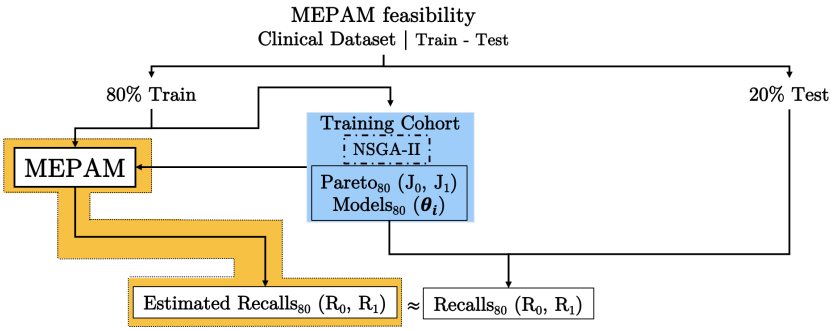


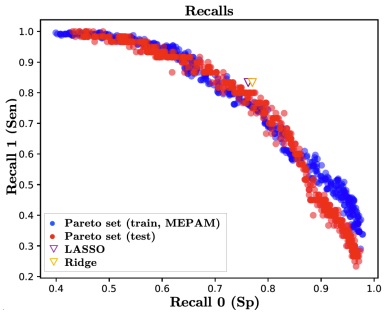
Fig. 4. Workflow to determine MEPAM feasibility.

MEPAM is applied to the training cohort for its validation by deriving an internal estimation of models recalls. Models obtained in the training cohort are externally evaluated by using the test set. Therefore, if test set recalls and those estimated internally in the train partition exhibit a low difference between their respective models, MEPAM method is assumed as feasible to be implemented in the full data cohort to evaluate the final models.

Validation of the final models. At this stage no test set is extracted and the models are computed from all available data. MEPAM is implemented to the full data cohort in order to estimate the recalls of the final models.

4 Results

MEPAM feasibility. After the execution of NSGA-II algorithm on the training cohort, a Pareto set with 500 models is obtained. On the left graph of Fig. 5, the external recall evaluation on the test set and the internally estimated recalls by MEPAM in the train partition are displayed.



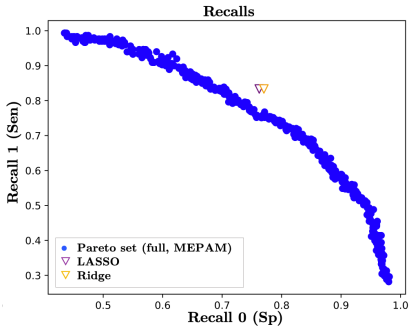
Range	No. Mod.	%	Cum. %
[0, 0.025)	89	17.8%	17.8%
[0.025, 0.05)	204	40.8%	58.6%
[0.05, 0.075)	114	22.8%	81.4%
[0.075, 0.1)	58	11.6%	93%
[0.1, 0.15)	35	7%	100%

Fig. 5. *Left graph.* R_1 (Sensitivity) vs. R_0 (Specificity). Blue dots represent the recall of the Pareto set models internally estimated by MEPAM. Red dots represent recalls externally computed on the test set. LASSO and Ridge recalls are plotted. *Right table*: Euclidean distance from estimated recalls to externally evaluated recalls on the test set.

Furthermore, LASSO and Ridge logistic regression recall values are shown as a comparative reference for the models. These models are trained with 80% of the data and tested with the remaining 20%. Although for high specificity and low sensitivity values slightly overestimated recalls are obtained, we can appreciate a solid behaviour of MEPAM.

In addition, Euclidean distance from the recalls estimated by MEPAM to those computed externally on the test set is always lower than 0.15, and 58.6% of models show a difference below 0.05 (see right Table in Fig. 5). Consequently, MEPAM is considered as a accurate performance estimation method.

Validation of the final models. NSGA-II is implemented on the full data cohort obtaining 500 different mortality prediction models in the Pareto set. The left graph of Fig. 6 highlights the difference between estimated recalls by MEPAM and those by LASSO and Ridge (trained with the 80% of the data and tested on the remaining 20%).



	Scores	
Type of Model	R_0	R_1
High Specificity	0.59	0.90
Balanced	0.77	0.77
High Sensitivity	0.92	0.62

Fig. 6. *Left graph.* R_1 (Sensitivity) vs. R_0 (Specificity). Blue dots represent estimated recalls by MEPAM on the full data cohort. LASSO and Ridge recalls are plotted. *Right table*: Examples of three models with their associated estimated performances. A model with high sensitivity, a balanced model and a model with high specificity are shown.

Although we do not obtain models as balanced as LASSO and Ridge regressions, a wide and competitive range of models is achieved. It should be noted that our objective is to provide a battery of mortality prediction models with different performances. Models with different degrees of sensitivity and specificity allow physicians to broaden the range of possibilities depending on the epidemiological situation (see right Table in Fig. 6) . In other words, depending on the availability of hospital resources, the number of patients admitted and other clinical aspects, physicians and hospital managers can choose the model that best suits a particular situation.

5 Conclusion

From a cohort of first and second wave data of the Basque Country, Spain, a battery of models with different performances is obtained. The multi-objective optimization framework allows us to focus on two key metrics: sensitivity and specificity. Although the optimization of both scores may be conflicting, it can also be beneficial for the learning of models with different performances. A new procedure known as MEPAM for an honest validation of the Pareto set models is also proposed.

The strength of the battery of mortality models resides not in outstanding performances but in the provision of models with varied performances for im-

mediate use. Although having this large set of models can be overwhelming, a reduce set of different models (e.g. 3, 5) can be chosen to obtain a less aggressive and more comprehensible and explainable set of models.

A wave and its strength are not possible to predict. However, physicians have external help for any situation. From low intensity waves, where a sensitive model may be of interest to avoid fatalities, to waves of exceptional strength that collapse hospitals and deplete resources where specific models may be considered. Furthermore, the variety of models obtained in the Pareto set allows our health system to fight against any unexpected outbreak. Definitively, this battery of COVID-19 mortality prediction models is a powerful tool to support physicians and hospital managers in different epidemiological situations.

Acknowledgements. This research is supported by the Basque Government (IT1504-22, Elkartek) through the BERC 2022-2025 program and BMTF project, and by the Ministry of Science, Innovation and Universities: BCAM Severo Ochoa accreditation SEV-2017-0718 and PID2019-104966GB-I00. Furthermore, the work is also supported by the AXA Research Fund project "Early prognosis of COVID-19 infections via machine learning".

Appendix

A detailed explanation of the variables used in the model is shown in Table 1.

Table 1. Blood tests, demographic, clinical and mortality outcome information collected from medical records. Depending on the feature, mean (μ), standard deviation (σ), median or interquartile range ($Q1$-$Q3$) are displayed.

Feature	Overall
Mortality, n(%)	
Deceased	298 (13.45%)
Survivors	1917 (86.55%)
Gender, n(%)	
Male	1268 (57.25%)
Female	947 (42.75%)
Age, $\mu(\sigma)(years)$	67.12 (17.55)
Oxygen Sat., $\mu(\sigma)$(%)	95.14(3.04)
Heart rate, $\mu(\sigma)(bpm)$	79.84(14.73)
No. of comorbi., $\mu(\sigma, range)$	0.39(0.66, [0,1])
Symptoms, $\mu(\sigma, range)$	0.3(0.67, [0,3])
Pain Scale, $\mu(\sigma, range)$	0.21(0.55, [0,4])
Temperature (^{o}C), $\mu(\sigma)$	36.78(0.82)
DBP, $\mu(\sigma)(mmHg)$	74.78(11.63)
SBP, $\mu(\sigma)(mmHg)$	128.42(20.89)
Lipemia, median (Q1-Q3)	8.9 (4.0, 13.0)
Leukocytes, median (Q1-Q3)(x$10^3/\mu L$)	6.17 (4.76, 8.21)
Neutrophils, median (Q1-Q3)(%)	73.20(65.10, 80.90)
Neutrophils, median (Q1-Q3)(x$10^3/\mu L$)	4.44(3.17, 6.28)
Monocytes, median (Q1-Q3)(%)	7.60(5.50, 10.0)
Monocytes, median (Q1-Q3)(x$10^3/\mu L$)	0.46(0.33, 0.65)
Lymphocytes, median (Q1-Q3)(%)	17.40(11.40, 24.25)
Lymphocytes, median (Q1-Q3)(x$10^3/\mu L$)	1.03(0.73, 1.38)
Basophils, median (Q1-Q3)(%)	0.20(0.11, 0.40)
Basophils, median (Q1-Q3)(x$10^3/\mu L$)	0.012(0.01, 0.02)
Eosinophils, median (Q1-Q3)(%)	0.2(0, 0.55)
Eosinophils, median (Q1-Q3)(x$10^3/\mu L$)	0.01(0, 0.03)

Feature	Overall
PT, median (Q1-Q3)(%)	91(78, 100)
HCB, median (Q1-Q3)(x$10^6/\mu L$)	4.63(4.22, 5.04)
MCV, median (Q1-Q3)(fL)	91.20(87.6, 94.8)
PLT, median (Q1-Q3)(x$10^3/\mu L$)	181(143, 236)
CL, median (Q1-Q3)(mEq/L)	101.0(98.6, 103.6)
ALT, median (Q1-Q3)(U/L)	26(17, 41)
MCH, median (Q1-Q3)(pg)	29.9(28.6, 31.1)
INR, median (Q1-Q3)	1.06(1.00, 1.17)
CREA, median (Q1-Q3)(mg/dL)	0.92(0.76, 1.13)
CRP, median (Q1-Q3)(mg/L)	56.42(22.11, 110.64)
BR, median (Q1-Q3)(mg/dL)	0.48(0.36, 0.67)
MPV, median (Q1-Q3)(fL)	10.10(8.46, 11.10)
APTT, median (Q1-Q3)(sg)	32.54(29.60, 36.35)
NA, median (Q1-Q3)(mEq/L)	138(136, 140)
HB, median (Q1-Q3)(g/dL)	13.9(12.5, 15.0)
K, median (Q1-Q3)(mEq/L)	4.1(3.8, 4.4)
UREA, median (Q1-Q3)(mg/dL)	36(27, 50)
Haemolysis, median (Q1-Q3)	6.0(2.0, 18.0)
RDW, median (Q1-Q3)(%)	13.10(12.30, 14.05)
HCT, median (Q1-Q3)(%)	42.20(38.50, 45.60)
Jaundice, median (Q1-Q3)	1.0(0.7, 10)
D-Dimer, median (Q1-Q3)(ng/ml)	750(460, 1400)
MCHC, median (Q1-Q3)(ng/ml)	32.7(31.9, 33.4)
GLU, median (Q1-Q3)(mg/dL)	110.3(98, 132.5)
PCT, median (Q1-Q3)(ng/ml)	0.09(0.05, 0.17)
LDH, median (Q1-Q3)(U/L)	272(223, 343)

References

1. World health organization. (2022), https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
2. Armañanzas, R., Díaz, A., Martínez-García, M., Mazuelas, S.: Derivation of a cost-sensitive covid-19 mortality risk indicator using a multistart framework. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 2179–2186 (2021). https://doi.org/10.1109/BIBM52615.2021.9669288
3. Blank, J., Deb, K.: Pymoo: Multi-objective optimization in python. IEEE Access **8**, 89497–89509 (2020). https://doi.org/10.1109/ACCESS.2020.2990567
4. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. IEEE Transactions on Evolutionary Computation **6**(2), 182–197 (2002). https://doi.org/10.1109/4235.996017
5. Gupta, R.K., Harrison, E.M., Ho, A., Docherty, A.B., Knight, S.R.: Development and validation of the isaric 4c deterioration model for adults hospitalised with covid-19: a prospective cohort study. The Lancet. Respiratory Medicine. **9**(4), 349–359 (2021), https://doi.org/10.1016/S2213-2600(20)30559-2
6. Ircio, J., Lojo, A., Mori, U., Lozano, J.A.: A multivariate time series streaming classifier for predicting hard drive failures. IEEE Computational Intelligence Magazine **17**(1), 102–114 (2022). https://doi.org/10.1109/MCI.2021.3129962
7. Knight, S.R., Ho, A., Pius, R., Buchan, I.: Risk stratification of patients admitted to hospital with covid-19 using the isaric who clinical characterisation protocol: development and validation of the 4c mortality score **370** (2020). https://doi.org/10.1136/bmj.m3339
8. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P.: Missing value estimation methods for DNA microarrays . Bioinformatics **17**(6), 520–525 (06 2001), https://doi.org/10.1093/bioinformatics/17.6.520
9. Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D.: Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal **369** (2020). https://doi.org/10.1136/bmj.m1328
10. Yan, L., Zhang, H.T., Goncalves, J., Xiao, Y.: An interpretable mortality prediction model for covid-19 patients. Nature Machine Intelligence **2**(5), 283–288 (2020/05/01). https://doi.org/10.1038/s42256-020-0180-7