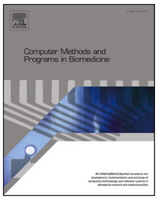
Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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



A machine learning approach to identify groups of patients with hematological malignant disorders

Pablo Rodríguez-Belenguer [a], José Luis Piñana [b,c], Manuel Sánchez-Montañés [d,*], Emilio Soria-Olivas [e], Marcelino Martínez-Sober [e], Antonio J. Serrano-López [e]

[a] *Research Programme on Biomedical Informatics (GRIB), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute, 08003 Barcelona, Spain*
[b] *Hematology Department, Hospital Clínico Universitario de Valencia, 46010 Valencia, Spain*
[c] *Fundación INCLIVA, Instituto de Investigación Sanitaria Hospital Clínico Universitario de Valencia, 46010 Valencia, Spain*
[d] *Department of Computer Science, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain*
[e] *IDAL, Intelligent Data Analysis Laboratory, ETSE, Universitat de València, 46100 Valencia, Spain*

ARTICLE INFO

Keywords:
Machine learning
High risk groups identification
Hematological disease
COVID-19
Serological response
SARS-CoV-2 mRNA vaccines
Explainable AI (XAI)

ABSTRACT

Background and Objective: Vaccination against SARS-CoV-2 in immunocompromised patients with hematologic malignancies (HM) is crucial to reduce the severity of COVID-19. Despite vaccination efforts, over a third of HM patients remain unresponsive, increasing their risk of severe breakthrough infections. This study aims to leverage machine learning's adaptability to COVID-19 dynamics, efficiently selecting patient-specific features to enhance predictions and improve healthcare strategies. Highlighting the complex COVID-hematology connection, the focus is on interpretable machine learning to provide valuable insights to clinicians and biologists.
Methods: The study evaluated a dataset with 1166 patients with hematological diseases. The output was the achievement or non-achievement of a serological response after full COVID-19 vaccination. Various machine learning methods were applied, with the best model selected based on metrics such as the Area Under the Curve (AUC), Sensitivity, Specificity, and Matthew Correlation Coefficient (MCC). Individual SHAP values were obtained for the best model, and Principal Component Analysis (PCA) was applied to these values. The patient profiles were then analyzed within identified clusters.
Results: Support vector machine (SVM) emerged as the best-performing model. PCA applied to SVM-derived SHAP values resulted in four perfectly separated clusters. These clusters are characterized by the proportion of patients that generate antibodies (PPGA). Cluster 1, with the second-highest PPGA (69.91%), included patients with aggressive diseases and factors contributing to increased immunodeficiency. Cluster 2 had the lowest PPGA (33.3%), but the small sample size limited conclusive findings. Cluster 3, representing the majority of the population, exhibited a high rate of antibody generation (84.39%) and a better prognosis compared to cluster 1. Cluster 4, with a PPGA of 66.33%, included patients with B-cell non-Hodgkin's lymphoma on corticosteroid therapy.
Conclusions: The methodology successfully identified four separate patient clusters using Machine Learning and Explainable AI (XAI). We then analyzed each cluster based on the percentage of HM patients who generated antibodies after COVID-19 vaccination. The study suggests the methodology's potential applicability to other diseases, highlighting the importance of interpretable ML in healthcare research and decision-making.

1. Introduction

Over time, groups at higher risk of suffering greater consequences from coronavirus infectious disease 2019 (COVID-19) have been identified [1]: being 60 years of age or older, suffering from severe liver disease, HIV or AIDS, or having a compromised immune system contribute to an unfavorable outcome.

The study of these risk groups is a relevant area of current scientific research [2–4]. In our case, we focus on patients with hematological malignancies (HM) who have been shown recently to be at an increased

[*] Corresponding author.
E-mail address: manuel.smontanes@uam.es (M. Sánchez-Montañés).

https://doi.org/10.1016/j.cmpb.2024.108011
Received 19 January 2023; Received in revised form 6 December 2023; Accepted 7 January 2024
Available online 9 January 2024

risk for a severe course of COVID-19, with a hospitalization rate of more than 50% and a case fatality rate of approximately 30% and reached nearly 50% in those over 70 before mass vaccination [5–8]. This contrasts with a case fatality rate of around 1% in young patients without risk factors [9]. It should be noted that, over time and during periods of high hospital demand, the workload of healthcare staff has been very tight, and this fact may have had a detrimental effect on patients' care and outcome.

It is just over three years since COVID-19 hit the world. According to data published by the WHO until 25 October 2023 [10], there have been 771.549.718 confirmed cases of COVID-19, including 6.974.473 deaths, worldwide. Patients with HM are considered at high risk of severe COVID-19 caused by the new coronavirus (SARS-CoV-2) [7]. The dramatic impact of COVID-19 in hematological patients was early observed during the first waves with mortality rates above 25% [11–16]. Although vaccination has lowered HM patients' overall mortality to under 10% [6,17] patients with cancer still experienced a disparate burden of COVID-19 mortality as compared to general population even during the Omicron variant of concern (VOC) period since the odds of developing a SARS-CoV-2 vaccine-induced antibody response are consistently lower than in the general population [18]. These facts support continuous efforts and researches in this area [19].

A strong vaccine-induced antibody response is crucial for protection against the Omicron VOC severity and mortality in these immunocompromised patients [17,20]. However, an impaired response to full SARS-CoV-2 vaccination occurs in 5% to 70% of immunocompromised patients [21]. For instance, an additional dose is now recommended in these patients [22,23]. The poor humoral response in HM patients contributes to a higher incidence of SARS-CoV-2 infection after vaccination [18]. Moreover, serological response rates in hematological patients varies significantly depending on age, disease type, the timing and type of HD treatment and adds up to the picture. Thus, it is of great interest to elucidate the conditions and/or profiles of HM patients associated with poor or null antibody response for tailored counseling regarding additional actions such as continuous preventive transmission measures, extra vaccine doses, pre-exposure neutralizing SARS-CoV-2 monoclonal antibodies prophylaxis and/or consideration of early antiviral therapy in order to mitigate the incidence and severity of breakthrough COVID-19 in poor responder patients.

Machine Learning (ML) is an area of Artificial Intelligence where algorithms are developed to automate data-driven decision-making. This process is the result of multidisciplinary work between data scientists and healthcare professionals, the main objective being to speed up the identification of groups at higher risk for decision-making. This work brings together some of the research and benefits that ML/ Deep Learning (DL) has brought to this COVID-19 pandemic in HM patients, ranging from predictive models of severe infection risk, daily analysis of the evolution of infected individuals and analysis of radiographs or medical records, among many others [15,24,25].

In this paper, we analyze a dataset that collects information on HM patients who received complete vaccination against SARS-CoV-2 [6]. Our previous work using this dataset addressed this issue by applying only probabilistic graphical models in the conditional probabilities of antibody generation according to the disorder [6]. However, the current study analyzes a more comprehensive comparison of different ML models (Decision Tree, Random Forest, Logistic Regression and Support Vector Machines) to find out which of them is more reliable in terms of predictive power and interpretability. We will obtain the SHAP values for each predictor [26] for the best ML model that will feed a Principal Component Analysis (PCA) to finally obtain clusters and analyze the profile of patients belonging to each cluster [27] and the proportion of patients that generate antibodies or not to identify patients with poor antibody generation after complete vaccination.

2. Material and methods

This section presents the data analyzed and the methods used, which are briefly described.

2.1. Data

The dataset comes from a prospective registry including 1683 patients with different HMs conducted by the Spanish Hematopoietic Transplant and Cell Therapy group (GETH-TC) in collaboration with the Spanish Society of Hematology and Hemotherapy (SEHH). Details of this multicenter registry and inclusion criteria have been reported in detail elsewhere [6]. Briefly, this registry included consecutive adult patients with a previous history of hematological disorders who were vaccinated against SARS-CoV-2 from December 30th 2020 to June 30th 2021 in 21 participating Spanish centers. All patients included in this registry gave signed informed consent according to the Helsinki declaration. The status of all included patients was updated on July 30th 2021. The local ethics committee of the Hospital Clínico Universitario de Valencia approved the registry and the study protocol (reference code 35.21). The final dataset contains 1166 evaluable HM patients with more than 100 categorical variables, which were reduced to 24 in addition to the target variable to be predicted (achieving or not an immune response against COVID-19). Details about these variables are shown in Table 1.

There were only two non-categorical variables: median time from serology to vaccination, and blood count. Both were converted to categorical according to the expert's knowledge. After this preprocessing, all variables were dummified.

2.2. Methodology

Our goal in this work is to understand the relationship between patient conditions and whether or not the patient generates antibodies to SARS-COV-2 after vaccination. Our reasoning, summarized in Fig. 1, is as follows:

1- First, we find the most accurate multivariate relationship between patient conditions and antibody generation. To do this, we train a battery of supervised machine learning algorithms and settled on the best one.

2- Once we obtain this relationship, we try to understand it. In general, Machine Learning models are not directly explainable, so we resorted to a standard XAI technique such as SHAP values. This technique yields one value for each patient and variable, so in our case we have $1166 \times 52 = 60{,}632$ values. Due to the large number of values, it does not provide a directly interpretable explanation of the relationship between the patient's conditions and whether or not the patient generates antibodies.

3- The next step is to simplify the information provided by the SHAP values so that we can extract relevant and understandable information. One strategy would be to average the SHAP values of each variable, but this would give the overall importance of each variable, offering a very limited explanation of the relationship between the patient's conditions and whether or not the patient generates antibodies. An intermediate point is to reduce the complexity of the information using a dimensionality reduction technique such as PCA and analyze whether clear clusters of patients are formed therein. Each cluster would indicate a set of patients with similar antibody generation patterns according to the model.

4- Finally, as the analyzed model is the most accurate within the battery of models, we assume that the identified clusters are the natural types of patients that exist in our dataset, considering that we want to predict whether they generate antibodies or not.

Table 1
Patient variables and number of patients (percentage) that satisfy each of them.

Variable	$n = 1166$
Prior COVID-19, n (%)	99 (8.4)
Median time from serology prior to vaccination, days (range)	0 (0–366)
Type of vaccine, n (%)	
● Moderna mRNA-1273	864 (74)
○ Response after vaccination*	700 (81)
● Pfizer-BioNTech BNT162b2	272 (23)
○ Response after vaccination*	204 (75)
● Adenoviral vector-based	30 (2.6)
○ Response after vaccination*	21 (70)
Age (years), median (range)	63 (18–97)
• 18–40 years, n (%)	128 (11)
• 41–60 years, n (%)	393 (33.7)
• 61–70 years, n (%)	314 (26.9)
• >71 years, n (%)	331 (28.4)
Male, n (%)	667 (57.2)
Baseline disease, n (%)	
• AML	149 (13.1)
• ALL	36 (3.4)
• MDS	125 (10.9)
• B-cell NHL /allo-HSCT/ASCT/CAR-T	252 (21)
• T cell NHL /allo-HSCT/ASCT	25 (2)
• Plasma cell disorders /allo-HSCT/ASCT/CAR-T	203 (17.5)
• CLL /allo-HSCT	133 (11.5)
• HD /allo-HSCT/ASCT	86 (7.8)
• cMPN /allo-HSCT	130 (11.5)
• Non-malignant disorders$^{\alpha}$ /allo-HSCT	27 (2.5)
Type of cell therapy procedure	
• Allo-HSCT	318 (27.3)
• ASCT	87 (7.5)
• CAR-T	21 (1.8)
Status disease at vaccination, n (%)	
• Complete remission	663 (56.9)
• Partial remission	139 (11.9)
• Active disease	251 (21.4)
• Non treated	114 (9.8)
Time last treatment to COVID-19 vaccine, months (range)	
• untreated	187 (16)
• Active treatment	479 (41.1)
• ≥ 6month to 1 year	90 (7.7)
• ≥ 1 year	410 (35.2)
Immunosuppressive drugs at vaccination, n (%)	220 (18.9)
Corticosteroids at vaccination, n (%)	214 (18.4)
Daratumomab, n (%)	45 (3.9)
Venetoclax, n (%)	17 (1.5)
Anti-CD-20 moAb, n (%)	217 (18.6)
• < 6months before 1st vaccine dose	77 (6.6)
• 6 to 1 year before 1st vaccine dose	24 (2.1)
• >1 year before 1st vaccine dose	115 (9.9)
BTK inhibitor therapy, n (%)	67 (5.7)
TKI therapy, n (%)	32 (2.7)
Lenalidomide maintenance, n (%)	115 (9.9)
Ruxolitinib therapy, n (%)	25 (2.1)
Blood count before vaccination (x10^{9}/mL)	
• Absolute neutrophile counts, median (range)	3.035 (0–46.7)
• Absolute neutrophile counts < 0.5 × 10^{9}/mL, n (%)	26 (2.2)
• Absolute lymphocyte counts, median (range)	1.75 (0.14–262.1)
• Absolute lymphocyte counts < 1 × 10^{9}/mL, n (%)	188 (16.1)
• Absolute lymphocyte counts < 0.5 × 10^{9}/mL, n (%)	46 (3.9)
Time from second dose to serologies, median days (range)	21 (14–61)
Median time between vaccine doses, median days (range)	28 (17–105)
response against COVID-19 detection at 3 weeks after vaccination, n (%)	925 (79.3)
Median follow-up after full vaccination, days (range)	28 (17–139)

Abbreviations, PCR, polymerase chain reaction AML, acute myeloid leukemia; ALL, acute lymphoblastic leukemia; MDS, myelodysplastic syndrome; B-cell NHL, B-cell non-Hodgkin lymphoma; T cell NHL, T cell non-hodgkin lymphoma; CLL, chronic lymphocytic leukemia; HD, Hodgkin disease; cMPN, chronic myeloproliferative neoplasm; Allo-HSCT, allogeneic stem cell transplantation; ASCT, autologous stem cell transplantation; CAR-T, T-cell chimeric antigen receptor; moAb, monoclonal antibody; BTK inhibitor, Bruton's tyrosine kinase inhibitor; TKIs, tyrosine kinase inhibitors; SCoV2-R-A, SARS-CoV-2-reactive IgG antibodies.

*Differences in SARS-CoV-2 seropositivity after complete vaccination with these compounds are statistically significant ($p < 0.0001$).

α non-malignant diseases not allografted ($n = 17$) include; aplastic anemia ($n = 5$), paroxysmal nocturnal hemoglobinuria ($n = 3$), autoimmune hemolytic anemia ($n = 3$), combined immunodeficiency disorder ($n = 2$), immune thrombocytopenia ($n = 1$) and cyclic neutropenia ($n = 3$). Among allo-HSCT recipients with non-malignant diseases there were 8 with aplastic anemia and 2 with major talassemia.

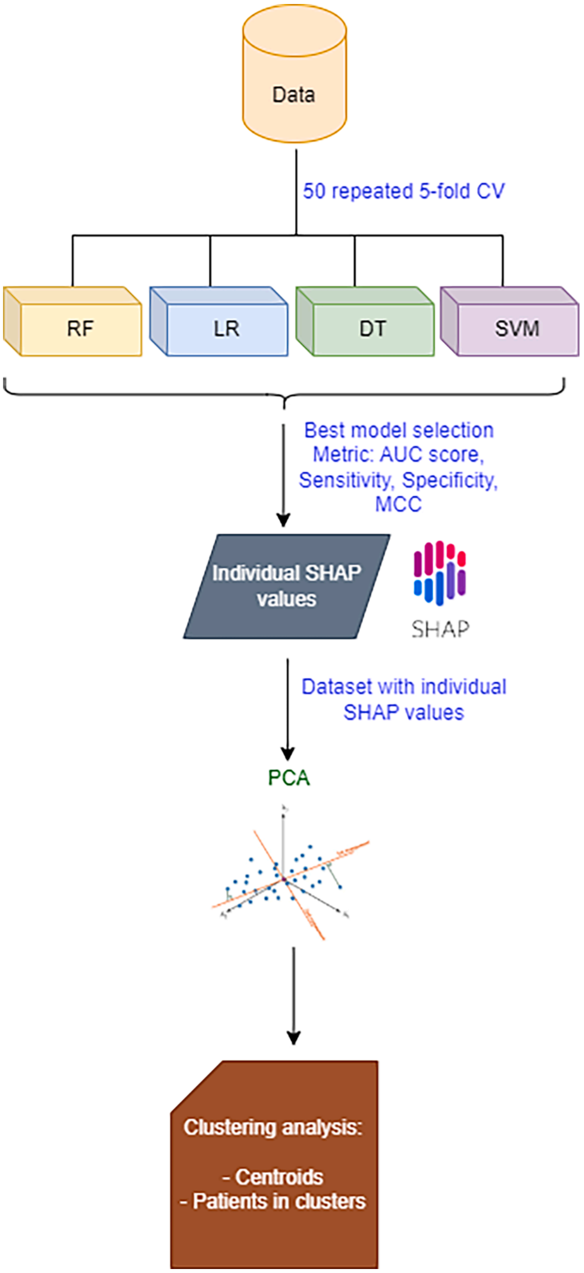


Fig. 1. Diagram of the proposed methodology.

2.3. *Predictive models*

2.3.1. *Logistic regression*

Logistic regression (LR) belongs to the group of statistical models known as generalized linear models [28], and is a classical model used in classification problems [29]. There are two operations: first, a linear combination of the independent/predictor variables is performed and a bias term is added; second, a sigmoid function is applied to estimate the probability of belonging to a given class (Eq. (1)):

$$P(\ class\ |x_1, \cdots, x_N) = \frac{1}{1 + e^{-\left(w_0 + \sum_{k=1}^{N} w_k \cdot x_k\right)}} \tag{1}$$

Here, wk (k ϵ [0, N]) are the model parameters, N is the number of predictor variables, and xk are these variables for a given patient.

2.3.2. *Decision tree*

A decision tree (DT) is a hierarchical supervised learning model that predicts the target variable by performing a sequence of queries on the predictor variables. It is composed of internal decision nodes and terminal leaves [29]. Decision trees have two main advantages that are relevant to this work:

- A decision tree is a non-parametric model; it does not assume any parametric form for class densities, and its structure is not fixed a priori. Rather, the tree grows during learning based on the complexity of the problem.
- It is a self-explanatory model, unlike other more powerful models such as neural networks, where knowledge extraction is extremely complex.

2.3.3. *Random forest*

Random forest (RF) is a substantial modification of bagging that builds a large collection of decorrelated trees and then averages them [30]. On many problems, the performance of RF is very similar to boosting, but they are easier to train and tune. Furthermore, Fernández-Delgado et al. demonstrated its superiority in a comparison of several algorithms on different problems, so RF can be considered as a reference [31]. Consequently, RFs are popular and are implemented in a wide variety of Machine Learning packages.

2.3.4. *Support vector machine*

A Support Vector Machine (SVM) is a supervised ML algorithm for solving both linear and nonlinear problems [32]. Like the previous models, they can be used for both classification and regression problems. The SVM algorithm tries to find the optimal hyperplane in the transformed space that separates the different classes in the case of a classification problem.

Finally, the quality of all models fitted to the data will be measured by the Sensitivity, Specificity, MCC, and area under the curve (AUC), by using a 50-repeated-5-fold cross-validation. The receiver operating characteristic (ROC) curve of the best model will be shown.

2.4. *SHAP values*

The importance of the variables will be extracted using SHapley Additive exPlanations (SHAP) [33]. The SHAP value is a general approach to explaining the output of any Machine Learning model. The SHAP value represents the average change in model output when considered together with the rest. It weights the effect of each variable in both directions for a positive or negative effect on the measured outcome. One of the advantages of SHAP values is that it allows to obtain the individual values of each variable for each instance. This allows to obtain a matrix of the same dimensions as the data matrix, where now each cell (i,j) reflects the importance of variable j in the prediction for instance i. This matrix will be sent to the next step.

2.5. *Dimensionality reduction of SHAP values: PCA*

Principal Component Analysis (PCA) is a linear dimensionality reduction technique [34]. The resulting variables are linearly uncorrelated, so their covariance matrix is diagonal. Applications include data visualization, reducing the number of input variables to a model, and preprocessing for clustering. There are many other popular dimensionality reduction techniques such as Multidimensional Scaling, Isomap, Local Linear Embedding (LLE), Autoencoders (AEs), and Variational Autoencoders (VAEs) [35], but PCA is one of the simplest and most widely used. In this work we will use this technique to compress the SHAP values before clustering.

2.6. *Clustering: K-means*

K-means is one of the simplest and most popular clustering algorithms. This algorithm divides a dataset into k clusters by iteratively assigning data points to the centroid of the nearest cluster [29]. The goal of the algorithm is to minimize the sum of the squares of the distances from each point to its nearest centroid. The centroids are updated based on the mean of the assigned points.

2.7. *Software*

We used the Python programming languages to develop the model creation and analysis scripts. We extensively used the standard Python libraries Scikit-Learn [36], Pandas [37], Numpy [38], Matplotlib [39] and Shap [40]. Additional information is shown in Table 2.

3. Results

Fig. 1 shows a diagram of the proposed methodology. As a first step, the ML models are built using a grid search of the best hyperparameters, and evaluated by a 50 repeated – 5 fold CV. Then, the best ML model is selected according to the best AUC score, and best Sensitivity-Specificity balance along with the MCC. Subsequently, the individual SHAP values are processed by PCA and, finally, a clustering analysis of the data given by PCA is performed. Calculating SHAP values of the best classifier provides a detailed understanding of the importance of variables for each observation. PCA is used to reduce the complexity of these SHAP values while maintaining most of their variability. This dimensionality reduction facilitates interpretation and visualization by projecting the data into a lower dimensional space. Once the coordinates obtained through PCA are available, clustering is used to identify possible groupings or patterns among observations in this lower-dimensional space, which could reveal relationships between patients based on the importance of the variables captured by the SHAP values

In summary, the proposed procedure allows combining the interpretation of the importance of the variables through SHAP values, dimensionality reduction with PCA and the identification of patterns through clustering in a lower dimensional space, which could help to better understand the structure underlying patient data. By combining SHAP, PCA, and clustering, the interpretation of the model is simplified by focusing on the most significant contributions and global and group patterns.

Regarding the grid, the best fit of the hyperparameters and the AUC scores for the selected ML models are shown in Table 3.

Table 4 shows the mean and Standard Deviation for the selected models and metrics. SVM was the model with the best balance between Sensitivity and Specificity despite the fact that AUC value was exactly the same that for LR model. However, LR showed a poor performance for Sensitivity.

In Fig. 2 the ROC curve for the selected model is shown.

Table 2
Package information.

Package	Version	Applicability
scikit-learn	0.24.1	ML
numpy	1.19.5	Vector operations
matplotlib	3.3.4	Visualization
pandas	1.1.5	Dataframe operations
shap	0.40	Interpretability

Table 3
Grid search model information and AUC obtained.

Model	Grid	Best Tune	AUC
DT	criterion = [‘gini’, ’entropy’]	criterion = gini	0.65
	max_depth = [None-12]	max_depth = None	
	min_samples_split = [1–8]	min_samples_split = 2	
	min_samples_leaf = [1–5]	min_samples_leaf = 2	
RF	criterion = [‘gini’,’entropy’]	criterion = gini	0.67
	max_depth = [None-12]	max_depth = 2	
	min_samples_split = [2–5]	min_samples_split = 2	
	class_weight = [None, ’balanced’]	class_weight = balanced	
SVM	kernel = [’linear’, ‘rbf’]	kernel = rbf	0.70
	C = [0.1–10]	C = 1	
	gamma = [0.0001–0.01]	gamma = 0.01	
	class_weight = [None, ’balanced’]	class_weight = balanced	

Table 4
Mean and Standard Deviation (Std) of the proposed metrics in the models analyzed.

Model	Metric	Mean	Std
LR	AUC	0.77	0.03
LR	Sensitivity	0.32	0.06
LR	Specificity	0.95	0.02
LR	MCC	0.34	0.07
DT	AUC	0.66	0.04
DT	Sensitivity	0.43	0.07
DT	Specificity	0.84	0.03
DT	MCC	0.27	0.07
RF	AUC	0.74	0.04
RF	Sensitivity	0.55	0.08
RF	Specificity	0.79	0.04
RF	MCC	0.31	0.07
SVM	AUC	0.77	0.03
SVM	Sensitivity	0.65	0.07
SVM	Specificity	0.74	0.04
SVM	MCC	0.34	0.06

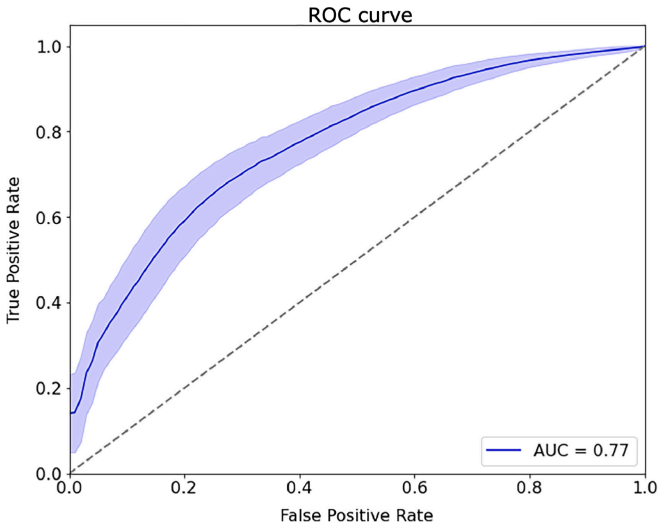


Fig. 2. ROC curve for the selected model.

Fig. 3 shows the feature importance for the SVM model, according to the SHAP analysis. Red shows a positive impact on the antibody response against SARS-CoV-2 and blue shows a negative impact. The most important predictor for the SVM model is having received corticosteroids treatment before or at the time of vaccination, and its impact on antibody generation is negative. In contrast, not having received anti-CD20 antibody therapy prior to vaccination showed a positive impact. Another negative impact was observed in patients diagnosed with chronic lymphocytic leukaemia.

Fig. 4 represents a SHAP heatmap plot in which f(x) is the probability of generating antibodies predicted by the model for each patient, and the bar on the right is the mean of the SHAP values. We observe some clear patterns. For instance, the effect of corticosteroids in reducing the probability of generating antibodies, or the effect of COVID prior vaccination in increasing this probability.

Fig. 5 shows the cumulative variance explained by the PCA of the SHAP values matrix. Ten principal components are enough to capture 80% of the variance of the SHAP values matrix.

Fig. 6 shows the dataset projected onto the first two principal components of the SHAP values matrix. Each circle represents a patient. Four clusters are obtained with the k-means algorithm. Each triangle represents a cluster centroid.

Table 5 contains the most important characteristics of the patients in each cluster according to their SHAP values and clinical expert opinion. Since the most predominant disease is B-cell non-Hodgkin lymphoma (B-cell NHL), we have also identified information on these patients along with other characteristics of interest for which there would be enough number of patients to be analyzed. Cluster 3 is the majority cluster with 71.96% of the total number of patients, followed by cluster 4 with approximately 17.32 % and clusters 1 and 2 with 9.69% and 1.03% respectively.

For cluster 1, the proportion of patients that generate antibodies (PPGA) is 69.91%. Cluster 2 represents the group of patients with the lowest PPGA, 32.2%. Nevertheless, the number of patients is too small to drawn firm conclusions.

Cluster 3 presents a PPGA of 84.39%. Despite of the large number of patients with a B-Cell NHL disorder, these patients have a higher antibody production compared to those in cluster 4. This can be due to the higher proportion of patients who did not receive anti CD-20 therapy or who had received it more than 1 year ago. Finally, the PPGA in cluster 4 is 66.33%. This is the cluster with the highest ratio of patients with B-cell NHL (+ corticosteroids).

4. Discussion

The methodology presented here allows the identification of groups of patients by computing the SHAP values of the best selected ML model (SVM in our case), processing them with PCA, and performing a clustering analysis of this information to identify the natural groups. This technique provides a differential value since all the variables are dichotomous, which always adds an extra degree of complexity for ML models. The fact that the best model has an AUC of 0.77 does not mean that our analysis is incorrect, but rather that the variables in the dataset are not sufficient to predict with 100% accuracy whether the patient will generate antibodies. This is a very common occurrence in medical datasets, and it does not prevent us from performing analyses and drawing conclusions that allow us to continue advancing our knowledge of the field.

SHAP values provide a detailed explanation of how each feature contributes to the model predictions for each instance. Subsequently, PCA on the SHAP values helps to reduce the dimensionality of this detailed information, maintaining the most significant contributions. Finally, by applying clustering to the components of the PCA, common patterns in the contributions of the features can be identified. This makes it easier to identify groups of instances with similar profiles in terms of how features affect predictions which could be useful in daily clinical practice to identify HM patients at higher risk of breakthrough and severe COVID-19 that can benefit of additional preventive strategies.

The purpose of this methodology is to provide hematologists with a tool that could be applied to a given patient for decision-making. In this sense, for the implementation of this methodology, when a new patient was admitted to the hospital, the distance to the nearest centroid would be computed to know whether the proportion of patients generating

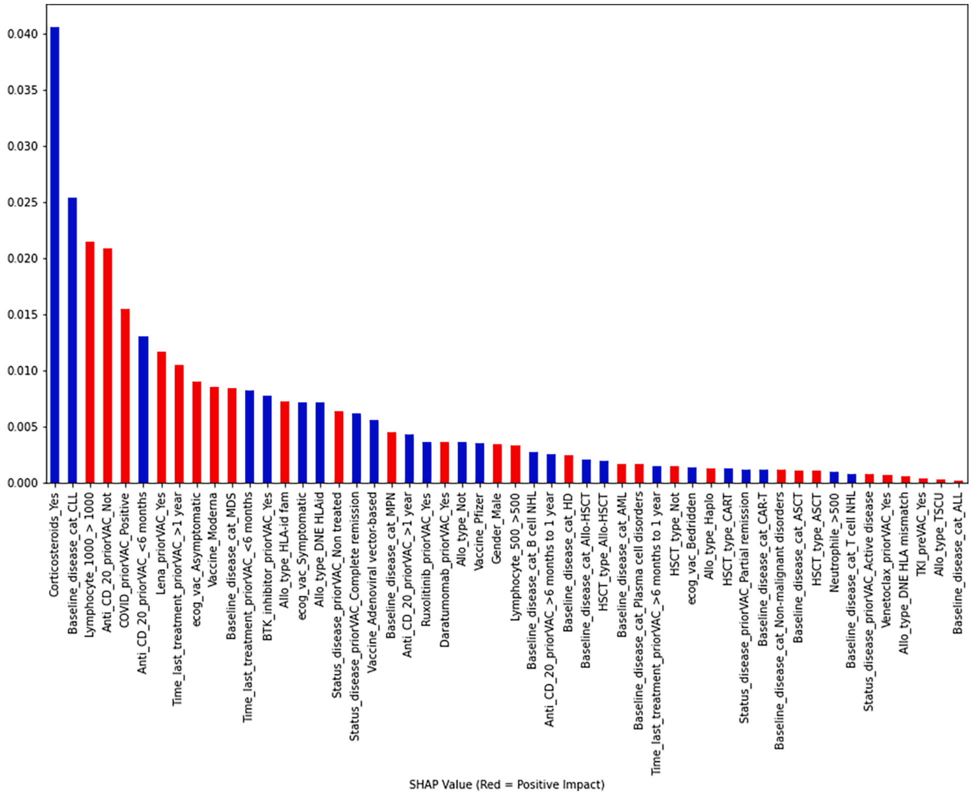


Fig. 3. Importance according to the SHAP values of each variable in the SVM model.

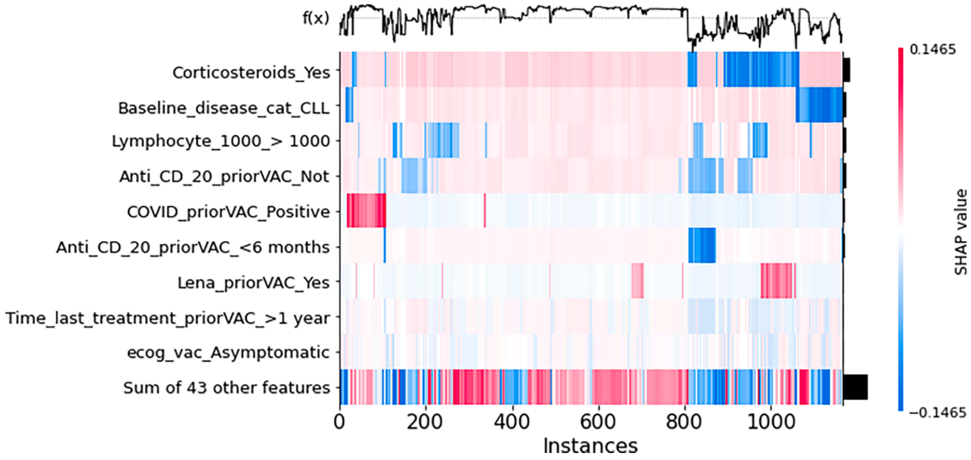


Fig. 4. SHAP heatmap plot for the 10 most important variables.

antibodies is high (cluster 3), medium (cluster 1) or low (clusters 4 and 2). In cases of medium or low probabilities, the treating physician could check antibody levels in order to indicate additional vaccine doses and/or prophylactic passive immunization with monoclonal antibodies, or in cases of breakthrough COVID-19 they could indicate early antiviral therapy which was associated with improved outcomes [20].

By applying the SHAP methodology on the SVM model we obtained individual importance values for each variable and patient (SHAP values), and after applying PCA to these values and performing clustering analysis, four distinct clusters emerged. Therefore, as can be seen in Table 4, it is consistent that in patients who did not receive anti-CD-20 monoclonal antibody therapy prior to vaccination the antibody response rate was higher (patients of cluster 3). This is physiologically consistent since anti-CD-20 monoclonal antibody (mAbs) therapy targets B-cell lymphocytes which are responsible of generating antibodies in response to infections or vaccines in human beings. Prior treatment with anti-CD20 mAbs consistently impairs SARS-CoV-2 antibody production [41–44] and it is an independent factor significantly associated with COVID-19 mortality and prolonged viral shedding, especially in those infected with the SARS-CoV-2 Omicron VOC, even in fully vaccinated hematologic patients [41,45]. Conversely, patients that did not receive it before vaccination had higher probabilities of developing antibodies and lower risk of severe breakthrough COVID-19. In this sense, diseases such as B-cell NHL or chronic lymphocytic leukemia (CLL) are negative contributors in mounting a proper antibody response and merit special attention to improve preventive strategies [17].

Besides, Fig. 6 shows how PCA analysis on the SHAP values allows to obtain different groups of patients. Even so, it has been possible to obtain separate clusters with their centroids. The interesting part of such an approach is that for new patients the same transformations and

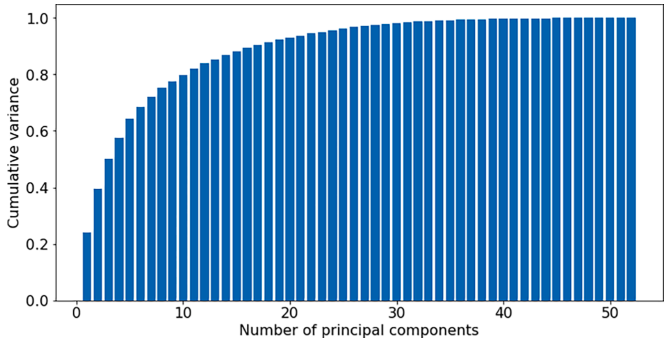


Fig. 5. Cumulative variance of the principal components calculated from the SHAP value matrix.

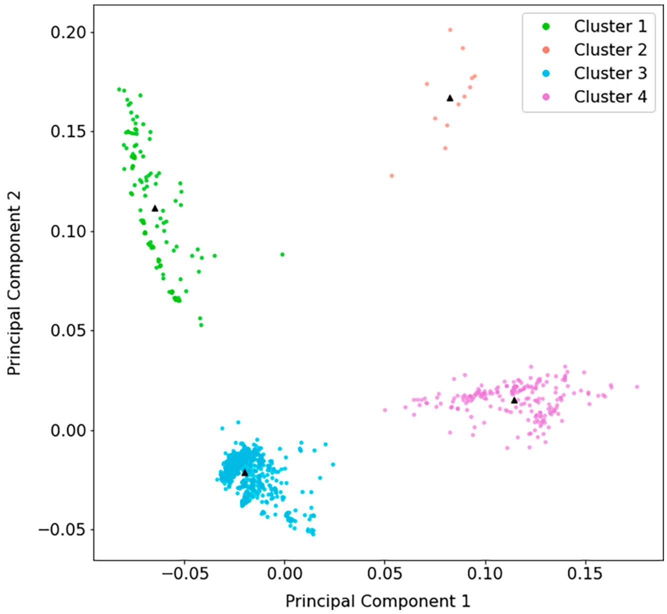


Fig. 6. Clusters obtained after applying k-means on the first two principal components obtained by PCA on the SHAP values matrix. The centroids of each cluster are marked with triangles.

calculations could be applied. Then, by computing the distance to any of the clusters it would be possible to estimate a priori the probabilities of generating antibodies against SARS-CoV-2 after complete vaccination. At the light of our result, the cluster of patients who most likely will generate antibodies would be cluster 3, followed by 1, 4 and 2.

Although the sample size could be regarded as a limitation, it should be considered that HM are categorized as rare diseases, given that their prevalence falls below the arbitrary threshold of < 6 per 100,000 inhabitants, as reported by The Portal for Rare Diseases and Orphan Drugs (accessible at: https://www.orpha.net/consor/cgi-bin/Clinics_Networks_Disease.php?lng=EN). In fact, our dataset comprised 1600 patients within a single study, representing 11.6% and 22.5% of the cases included in larger systematic review and meta-analyses reported to date in this scenario [46,47]. Our findings align consistently and robustly with the outcomes of both meta-analyses. This concordance underscores the reliability of our results and the valuable insights they offer into the diverse nature of immune responses among patients with hematological diseases.

Regarding the limitations of our work, an important aspect is the computational cost of the proposed methodology. Of all the steps performed, the most expensive by far is the computation of the SHAP values (1166×52 values), which took 30 h on a computer with 32 GB of DDR4 RAM and a 3.6 GHz Intel i9 CPU. Each of the other computational steps took on the order of seconds.

Table 5
Number of patients in each cluster.

	Cluster 1 (Y:79, N:34)	Cluster 2 (Y:4, N:8)	Cluster 3 (Y:708, N:131)	Cluster 4 (Y:134, N:68)
No Anti-CD-20	Y: 78, N: 29	Y: 4, N: 8	Y: 618, N: 72	Y: 100, N: 41
Anti-CD-20 < 6 months	Y: 0, N: 3	–	Y: 16, N: 34	Y: 9, N: 15
Anti-CD-20 > 1 year	Y: 1, N: 2	–	Y: 66, N: 18	Y: 21, N: 7
B-cell NHL	–	–	Y: 82, N: 48	Y: 30, N: 27
B-cell NHL + Corticosteroids	–	–	Y: 82, N: 48	Y: 30, N: 27
B-cell NHL + Allo- HSCT	–	–	Y: 246, N: 51	Y: 14, N: 7
B-cell NHL + Active disease	–	–	Y: 15, N: 8	Y: 5, N: 10
B-cell NHL + Complete remission	–	–	Y: 44, N: 31	Y: 22, N: 14
Last treatment < 6 months	–	–	Y: 28, N: 13	Y: 7, N: 15
CLL	Y: 79, N: 34	Y: 4, N: 8	–	–
Corticosteroids	–	Y: 4, N: 8	–	Y: 134, N: 68

Patients who generate Y:Yes and N:Not; B-cell NHL, B cell non-Hodgkin lymphoma; CLL, chronic lymphocytic leukemia; Allo-HSCT, allogeneic hematopoietic stem cell transplantation.– Patients who do not generate anti- SARS-CoV-2 antibodies. In each Cluster the number of patients generating anti-SARS-CoV-2 antibodies for different conditions is shown.

Another aspect that may appear to be a limitation is the use of PCA as a dimensionality reduction algorithm for SHAP values. Apart from this algorithm, there are other dimensionality reduction techniques that can be used [35]. Although not shown in our paper, we have tested other techniques such as Autoencoders or Variational Autoencoders, without obtaining improvements over PCA. However, PCA may not be the optimal dimensionality reduction technique on other datasets. In that case nonlinear dimensionality reduction techniques, such as those mentioned above, could be used. Something similar happens with the clustering algorithm we have used: in our case a well-known algorithm such as k-means has worked well. However, in other datasets this may be a limitation, so that more complex clustering algorithms would have to be considered in that case.

5. Conclusions

In this work we showed that it is possible to obtain a clustering of HM patients that characterizes the serological response by using an approach based on SHAP values according to the best ML model (SVM). With this methodology we obtained four clusters. Cluster 1contains the highest proportion of patients with CLL. The size of cluster 2 (only 12 patients) limits its interpretation. Cluster 3 contains the majority of patients (71.96%), with a PPGA of 84.39%. However, patients within this cluster treated with anti-CD 20 therapy before 6 months had a lower PPGA (47.06%). Finally, cluster 4 presents a PPGA of 66.33%. Additionally, patients belonging to this cluster with a B-cell NHL disorder had the highest risk of poor humoral response. The results of this cluster are probably influenced by the great proportion of patients receiving corticosteroids, which is the most important variable with a negative contribution in antibody generation. This methodology could be used in clinical practice to identify patients at risk of poor serological response by computing the distance to the different centroids who may benefit from additional preventive measures.

CRediT authorship contribution statement

Pablo Rodríguez-Belenguer: Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing, Formal analysis. **José Luis Piñana:** Conceptualization, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Manuel Sánchez-Montañés:** Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Emilio Soria-Olivas:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Marcelino Martínez-Sober:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Antonio J. Serrano-López:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

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.

Acknowledgements

We thank the Spanish Hematopoietic Transplant and Cell Therapy group (GETH-TC) for sharing the data. The co-author Manuel Sánchez-Montañés has been supported by grants PID2021-122347NB-I00 and PID2021-127946OB-I00 (funded by MCIN/AEI//10.13039/501100011033 and ERDF "A way of making Europe"), and from project FACINGLCOVID-CM under grant PD2022-004-REACT-EU.

References

[1] COVID-19 symptoms and what to do. nhs.uk. Published March 20, 2023. Accessed November 27, 2023. https://www.nhs.uk/conditions/covid-19/covid-19-symptoms-and-what-to-do/.

[2] C. Jung, B. Mamandipoor, J. Fjølner, et al., Disease-Course Adapting Machine Learning Prognostication Models in Elderly Patients Critically Ill With COVID-19: multicenter Cohort Study With External Validation, JMIR Med. Inform. 10 (3) (2022) e32949, https://doi.org/10.2196/32949.

[3] M. Pourhomayoun, M. Shakibi, Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making, Smart Health (Amst) 20 (2021) 100178, https://doi.org/10.1016/j.smhl.2020.100178.

[4] S. Subudhi, A. Verma, A.B. Patel, Prognostic machine learning models for COVID-19 to facilitate decision making, Int. J. Clin. Pract. 74 (12) (2020) e13685, https://doi.org/10.1111/ijcp.13685.

[5] J.L. Piñana, P. Rodríguez-Belenguer, D. Caballero, et al., Applicability of probabilistic graphical models for early detection of SARS-CoV-2 reactive antibodies after SARS-CoV-2 vaccination in hematological patients, Ann. Hematol. 101 (9) (2022) 2053–2067, https://doi.org/10.1007/s00277-022-04906-8.

[6] J.L. Piñana, L. López-Corral, R. Martino, et al., SARS-CoV-2 vaccine response and rate of breakthrough infection in patients with hematological disorders, J. Hematol. Oncol. 15 (1) (2022) 54, https://doi.org/10.1186/s13045-022-01275-7.

[7] P. Langerbeins, M. Hallek, COVID-19 in patients with hematologic malignancy, Blood 140 (3) (2022) 236–252, https://doi.org/10.1182/blood.2021012251.

[8] M. Sánchez-Montañés, P. Rodríguez-Belenguer, A.J. Serrano-López, E. Soria-Olivas, Y Alakhdar-Mohmara, Machine Learning for Mortality Analysis in Patients with COVID-19, Int. J. Environ. Res. Public Health 17 (22) (2020) 8386, https://doi.org/10.3390/ijerph17228386.

[9] Older adolescent (15 to 19 years) and young adult (20 to 24 years) mortality. Accessed November 27, 2023. https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-older-adolescent-(15-to-19-years)-and-young-adult-(20-to-24-years)-mortality.

[10] WHO Coronavirus (COVID-19) Dashboard. Accessed November 27, 2023. https://covid19.who.int.

[11] J.L. Piñana, R. Martino, I. García-García, et al., Risk factors and outcome of COVID-19 in patients with hematological malignancies, Exp. Hematol. Oncol. 9 (2020) 21, https://doi.org/10.1186/s40164-020-00177-z.

[12] J. García-Suárez, J. de la Cruz, Á. Cedillo, et al., Impact of hematologic malignancy and type of cancer therapy on COVID-19 severity and mortality: lessons from a large population-based registry study, J. Hematol. Oncol. 13 (1) (2020) 133, https://doi.org/10.1186/s13045-020-00970-7.

[13] L. Pagano, J. Salmanton-García, F. Marchesi, et al., COVID-19 infection in adult patients with hematological malignancies: a European Hematology Association Survey (EPICOVIDEHA), J. Hematol. Oncol. 14 (1) (2021) 168, https://doi.org/10.1186/s13045-021-01177-0.

[14] A. Sharma, N.S. Bhatt, A. St Martin, et al., Clinical characteristics and outcomes of COVID-19 in haematopoietic stem-cell transplantation recipients: an observational cohort study, Lancet Haematol. 8 (3) (2021) e185–e193, https://doi.org/10.1016/S2352-3026(20)30429-4.

[15] P. Ljungman, R. de la Camara, M. Mikulska, et al., COVID-19 and stem cell transplantation; results from an EBMT and GETH multicenter prospective survey, Leukemia 35 (10) (2021) 2885–2894, https://doi.org/10.1038/s41375-021-01302-5.

[16] J.M. Ribera, M. Morgades, R. Coll, et al., Frequency, clinical characteristics and outcome of adults with acute lymphoblastic leukemia and COVID 19 infection in the first vs. second pandemic wave in Spain, Clin. Lymphoma Myeloma Leuk. 21 (10) (2021) e801–e809, https://doi.org/10.1016/j.clml.2021.06.024.

[17] J.L. Piñana, L. Vazquez, M. Calabuig, et al., One-year breakthrough SARS-CoV-2 infection and correlates of protection in fully vaccinated hematological patients, Blood Cancer J. 13 (1) (2023) 8, https://doi.org/10.1038/s41408-022-00778-3.

[18] J. La, J.T.Y. Wu, W. Branch-Elliman, et al., Increased COVID-19 breakthrough infection risk in patients with plasma cell disorders, Blood 140 (7) (2022) 782–785, https://doi.org/10.1182/blood.2022016317.

[19] A.L. Potter, V. Vaddaraju, S. Venkateswaran, et al., Deaths Due to COVID-19 in Patients With Cancer During Different Waves of the Pandemic in the US, JAMA Oncol. 9 (10) (2023) 1417–1422, https://doi.org/10.1001/jamaoncol.2023.3066.

[20] J.L. Piñana, I. Heras, T.F. Aiello, et al., Remdesivir or Nirmatrelvir/Ritonavir Therapy for Omicron SARS-CoV-2 Infection in Hematological Patients and Cell Therapy Recipients, Viruses 15 (10) (2023) 2066, https://doi.org/10.3390/v15102066.

[21] J.L. Piñana, M. Guerreiro, C. Solano, SARS-CoV-2 Immunity in Hematopoietic Stem Cell Transplant and Cell Therapy Recipients: what Do We Know, and What Remains to Be Determined? Hemato 4 (2) (2023) 170–183, https://doi.org/10.3390/hemato4020014.

[22] S. Cesaro, P. Ljungman, M. Mikulska, et al., Recommendations for the management of COVID-19 in patients with haematological malignancies or haematopoietic cell transplantation, from the 2021 European Conference on Infections in Leukaemia (ECIL 9), Leukemia 36 (6) (2022) 1467–1480, https://doi.org/10.1038/s41375-022-01578-1.

[23] S. Cesaro, M. Mikulska, H.H. Hirsch, et al., Update of recommendations for the management of COVID-19 in patients with haematological malignancies, haematopoietic cell transplantation and CAR T therapy, from the 2022 European Conference on Infections in Leukaemia (ECIL 9), Leukemia 37 (9) (2023) 1933–1938, https://doi.org/10.1038/s41375-023-01938-5.

[24] J. Kang, T. Chen, H. Luo, Y. Luo, G. Du, M. Jiming-Yang, Machine learning predictive model for severe COVID-19, Infect. Genet. Evol. 90 (2021) 104737, https://doi.org/10.1016/j.meegid.2021.104737.

[25] R. Rehouma, M. Buchert, Y.P.P. Chen, Machine learning for medical imaging-based COVID-19 detection and diagnosis, Int. J. Intell. Syst. 36 (9) (2021) 5085–5115, https://doi.org/10.1002/int.22504.

[26] M. Smith, F. Alvarez, Identifying mortality factors from Machine Learning using Shapley values – a case of COVID19, Expert Syst. Appl. 176 (2021) 114832, https://doi.org/10.1016/j.eswa.2021.114832.

[27] S. Lu, R. Chen, W. Wei, M. Belovsky, X. Lu, Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions, AMIA Annu. Symp. Proc. 2021 (2022) 813–822.

[28] D.R. Cox, The regression analysis of binary sequences, J. R. Stat. Soc. Series B: Stat. Methodol. 20 (2) (1958) 215–232.

[29] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2, Springer, 2009.

[30] L. Breiman, Bagging predictors, Mach Learn 24 (2) (1996) 123–140, https://doi.org/10.1007/BF00058655.

[31] M. Fernández-Delgado, E. Cernadas, S. Barro, D Amorim, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn Res. 15 (1) (2014) 3133–3181.

[32] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support Vector Regression Machines, in: MC Mozer, M Jordan, T Petsche (Eds.), Advances in Neural Information Processing Systems, MIT Press, 1996. Vol 9, https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

[33] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst. 41 (3) (2014) 647–665, https://doi.org/10.1007/s10115-013-0679-x.

[34] K.LIII Pearson, On lines and planes of closest fit to systems of points in space, London, Edinburgh, Dublin Philos. Mag. J. Sci. 2 (11) (1901) 559–572, https://doi.org/10.1080/14786440109462720.

[35] B. Ghojogh, M. Crowley, F. Karray, A. Ghodsi, Elements of Dimensionality Reduction and Manifold Learning, Springer Nature, 2023.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-Learn: machine Learning in Python, J. Mach. Learn Res. 12 (null) (2011) 2825–2830.

[37] W. McKinney, Data Structures for Statistical Computing in Python, in: Walt S van der, J Millman (Eds.), Proceedings of the 9th Python in Science Conference, 2010, pp. 56–61, https://doi.org/10.25080/Majora-92bf1922-00a.

[38] C.R. Harris, K.J. Millman, S.J. van der Walt, et al., Array programming with NumPy, Nature 585 (7825) (2020) 357–362, https://doi.org/10.1038/s41586-020-2649-2.

[39] Hunter JD. Matplotlib, A 2D Graphics Environment, Comput. Sci. Eng. 9 (3) (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.

[40] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17, Curran Associates Inc., 2017, pp. 4768–4777.

[41] S.A. Apostolidis, M. Kakara, M.M. Painter, et al., Cellular and humoral immune responses following SARS-CoV-2 mRNA vaccination in patients with multiple sclerosis on anti-CD20 therapy, Nat. Med. 27 (11) (2021) 1990–2001, https://doi.org/10.1038/s41591-021-01507-2.

[42] B. Kornek, F. Leutmezer, P.S. Rommer, et al., B Cell Depletion and SARS-CoV-2 Vaccine Responses in Neuroimmunologic Patients, Ann. Neurol. 91 (3) (2022) 342–352, https://doi.org/10.1002/ana.26309.

[43] Z.L.E. van Kempen, E.M.M. Strijbis, M.M.C.T. Al, et al., SARS-CoV-2 Antibodies in Adult Patients With Multiple Sclerosis in the Amsterdam MS Cohort, JAMA Neurol. 78 (7) (2021) 880–882, https://doi.org/10.1001/jamaneurol.2021.1364.

[44] A. Werner, S. Schäfer, O. Zaytseva, et al., Targeting B cells in the pre-phase of systemic autoimmunity globally interferes with autoimmune pathology, iScience 24 (9) (2021) 103076, https://doi.org/10.1016/j.isci.2021.103076.

[45] C. Cattaneo, L. Masina, C. Pagani, et al., High mortality in fully vaccinated hematologic patients treated with anti-CD20 antibodies during the "Omicron wave" of COVID-19 pandemic, Hematol. Oncol. 41 (1) (2023) 205–207, https://doi.org/10.1002/hon.3064.

[46] M. Noori, S. Azizi, F. Abbasi Varaki, S.A. Nejadghaderi, D Bashash, A systematic review and meta-analysis of immune response against first and second doses of SARS-CoV-2 vaccines in adult patients with hematological malignancies, Int. Immunopharmacol. 110 (2022) 109046, https://doi.org/10.1016/j.intimp.2022.109046.

[47] J.S.K. Teh, J. Coussement, Z.C.F. Neoh, et al., Immunogenicity of COVID-19 vaccines in patients with hematologic malignancies: a systematic review and meta-analysis, Blood Adv. 6 (7) (2022) 2014–2034, https://doi.org/10.1182/bloodadvances.2021006333.