communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-00987-4

Machine learning approaches to dissect hybrid and vaccine-induced immunity

Giorgio Montesi[1,5], Simone Costagli[1,5], Simone Lucchesi[1], Jacopo Polvere[1], Fabio Fiorino[1,2], Gabiria Pastore[1], Margherita Sambo[3,4], Mario Tumbarello[3,4], Massimiliano Fabbiani[3,4], Francesca Montagnani[3,4], Donata Medaglini[1], Elena Pettini[1,6] & Annalisa Ciabattini[1,6]

Abstract

Background The spread of SARS-CoV-2 Omicron variant and its subvariants, highly transmissible but responsible of milder disease, has increased unreported infection cases. Identifying unaware infected individuals is crucial for estimating the true prevalence of infection and evaluating the breadth of hybrid immunity. In this study, this challenge was addressed by applying several Machine Learning approaches.

Methods A group of 116 participants, vaccinated against SARS-CoV-2, was enrolled in the IMMUNO_COV study at Siena University Hospital, Italy. Blood samples were collected before and six months after third vaccine dose. Machine Learning analysis, involving dimensionality reduction techniques, unsupervised clustering methods and classification models, were applied to serological data including antibody responses specific for wild type SARS-CoV-2 strain as well as Delta, Omicron BA.1 and Omicron BA.2 variants. Spike- and nucleocapsid-specific B cells were also assessed in each participant.

Results Using dimensionality reduction and unsupervised clustering, participants are grouped into high- and low-responders, with infected participants mainly distributed within the high-responders. Implementation of a consensus-based approach, including k-NN, RF, and SVM models, identifies 14 participants unaware of previous infection. Their immunological profile, characterized by strong spike- and nucleocapsid-specific humoral and B cell responses, significantly differs from that of non-infected participants.

Conclusions Machine Learning approaches are applied to identify participants unaware of prior infection and to dissect their hybrid immunity profiles. Based on serological data, this cost-effective method can be a valuable tool for estimating the true prevalence of infection, improving comprehension of immune responses elicited by vaccination alone or combined with infection, and tailoring public health interventions.

Plain Language Summary

With the spread of the Omicron variants, SARS-CoV-2 infections have become more contagious but less symptomatic, resulting in many individuals becoming infected without knowing it. Using advanced computer-based techniques, applied on easily accessible blood test results, such as antibodies and immune cells levels, we developed a model capable of identifying individuals who had been previously infected but were unaware of their infection. These individuals exhibited immune responses more closely resembling those of infected individuals than non-infected ones. The application of this model not only enables the profiling of immune responses induced by mRNA vaccination alone or in combination with infection, but also provides a more accurate picture of virus spread and supports targeted prioritization of vaccine allocation.

During the period 2021–2024, several studies have closely examined the immune response triggered by the novel mRNA vaccine platform administered in the context of the SARS-CoV-2 pandemic[1–4]. Yet, the characterization of the vaccine-induced immune response persistence has been complicated by the circulation of highly transmissible and spike-mutated virus variants, that have contributed to the establishment of the so-called hybrid immunity, which arises from the combined effect of vaccine-induced stimulation and infection with the whole virus[5]. Unlike SARS-CoV-2 vaccination, natural infection stimulates the immune system with antigens from the circulating variants, broadening the spectrum of recognized antigens by the immune system and engaging the natural entry site of the virus[6,7]. Consequently, identifying individuals experiencing hybrid immunity is a pivotal task, not only to provide accurate infection rates, but also to discriminate the immune responsiveness elicited by novel mRNA-based vaccines. However, the identification of individuals with hybrid immunity has been complicated by the spread of Omicron variants, emerged since the

[1]Department of Medical Biotechnologies, Laboratory of Molecular Microbiology and Biotechnology, University of Siena, Siena, Italy. [2]Department of Medicine and Surgery, LUM University "Giuseppe Degennaro", Casamassima, Bari, Italy. [3]Department of Medical Biotechnologies, University of Siena, Siena, Italy. [4]Department of Medical Sciences, Infectious and Tropical Diseases Unit, University Hospital of Siena, Siena, Italy. [5]These authors contributed equally: Giorgio Montesi, Simone Costagli. [6]These authors jointly supervised this work: Elena Pettini, Annalisa Ciabattini. e-mail: annalisa.ciabattini@unisi.it

end of November 2021[8,9], which increased the proportion of asymptomatic and mildly symptomatic infections, frequently unreported[10].

Retrospective diagnosis of past infections is typically performed by assessing serological responses to the nucleocapsid (N) protein, a viral antigen absent from mRNA vaccines that plays a crucial role in viral genome packaging, assembly regulation, and host immune evasion[11]. Nevertheless, N-specific antibodies can rapidly decline over time, are absent in some individuals, and can also be elicited by common cold human coronaviruses[9,12–16].

Machine Learning approaches have already been successfully applied to predict vaccine immunogenicity in healthy individuals[17], people living with HIV[18], solid organ transplant recipients[19] or in individuals affected by cancer[20]. These methods have the potential to support clinical decision and guide targeted preventive measures for individuals most at risk, making them an attractive method for identifying individuals with unreported previous infection.

In this study, dimensionality reduction techniques and unsupervised clustering methods are employed to profile high and low responders by using serological data. Then, multiple Machine Learning classifiers, combined in a consensus strategy, are used to develop a model capable of distinguishing between the immunological profiles of infected and non-infected, and thus to identify unaware infected individuals. Additionally, Self-Organizing Maps are leveraged to study the memory B-cell response. This integrative approach allows accurate assessment of infection rates and comparison of the immune responsiveness elicited by vaccination alone or combined with infection, potentially offering valuable insights for the management of future epidemics/pandemics.

Methods and materials

Study design

The study included 116 healthy participants enrolled at the Infectious and Tropical Diseases Unit, Azienda Ospedaliera Universitaria Senese (Siena, Italy) in the context of the IMMUNO_COV study. Healthy volunteers were recruited by an information campaign among health care-workers and university staff. After informed consent, all immunocompetent volunteers were evaluated by clinical investigators to rule out diseases with a potential impact on immunological response. Inclusion criteria were age ≥18 years and adherence to the SARS-CoV-2 vaccination campaign. Exclusion criteria included pregnancy and immunocompromising conditions (congenital, acquired, or drug-related). The study was performed in compliance with all relevant ethical regulations and approved by the local Ethical Committee for Clinical experimentation of Regione Toscana Area Vasta Sud Est (CEAVSE, protocol code 18869, approved on the 21st December, 2020).

Eighty-two participants were vaccinated with two doses of mRNA vaccines (BNT162b2 or mRNA-1273) with intervals of 3 or 4 weeks, respectively, and 34 participants received adenovirus-based vaccines (ChAdOx1 nCoV-19), administered 12 weeks apart. A third booster dose was administered to all participants using mRNA-based vaccines, 5–7 months after the 2-dose primary vaccination cycle. Blood samples were collected before the third dose and after a 6-months follow-up. Diagnosis of SARS-CoV-2 infection among self-reported infected participants was established via antigenic or molecular testing, conducted on nasopharyngeal swabs, either self-administered or collected by healthcare professionals. Since the start of the vaccination campaign, participants also completed a survey on SARS-CoV-2 infection. Participants were asked to specify the date of the positive test. Clinical data collection and management were carried out using the software REDCap (Research Electronic Data Capture, Vanderbilt University).

Plasma and peripheral blood mononuclear cells isolation

Plasma and peripheral blood mononuclear cells (PBMCs) were isolated from venous blood samples collected in heparin-coated blood tubes (Vacutainer; BD, NJ, USA). PBMCs were isolated by density-gradient sedimentation, using Ficoll-Paque (Lymphoprep; STEMCELL Technologies, BC, Canada). Cells were gently resuspended with warm cell recovery medium [10% DMSO (Thermo Fisher Scientific) and 90% heat inactivated fetal bovine serum (Sigma Aldrich; MO, USA)] and then rapidly transferred to cryovials that were cryopreserved in liquid nitrogen after o.n. at −80 °C. Plasma samples were stored at −80 °C.

Enzyme-linked immunosorbent assay

Antigen-specific IgG were tested in plasma samples by ELISA assay[21]. IgG antibodies against full spike protein or spike RBD of wild type (wt) or Delta and Omicron variants were assessed. Briefly, micro titre plates were coated with 1 μg/ml full spike protein (S1 + S2 ECD) or RBD of wild type (wt) or other variants (lineage B.1.617, Delta; and Omicron subvariants BA.1 and BA.2; all from Sino Biological, China) blocked and added with heat-inactivated plasma samples. Anti-human horseradish peroxidase (HRP)-conjugated IgG antibody was added and plates were then developed with 3,3',5,5'-tetramethylbenzidine (TMB; Thermo Fisher Scientific) substrate. A SARS-CoV-2 Spike neutralizing antibody [SARS-CoV-2 (2019-nCoV) Spike Neutralizing Antibody, Rabbit Mab, Sino Biological] was added and titrated to each plate, to obtain a standard calibration curve. The absorbance was measured at 450 nm using a Multiskan FC Microplate Photometer (Thermo Fisher Scientific; MA, USA). Data were expressed as ng/ml. IgG specific for the wt, Delta, Omicron BA.1 and Omicron BA.2 Nucleocapsid (N) protein were assessed using the same protocol described above. N-protein of wt or other variants (Delta and Omicron subvariants BA.1 and BA.2; all from Sino Biological) was used at 1 μg/ml. Since N-specific IgG levels were similar across the wt and variants, data of Omicron BA.2 N-specific IgG, the predominant circulating variant at the time of sampling, were used. BA.2 N-specific IgG levels are expressed as the area under the curve (AUC). A pre-infection reference threshold was established by assessing IgG antibodies against the N protein of the Omicron BA.2 variant in plasma samples collected prior to vaccination (day 0) from 40 uninfected individuals.

ACE-2/RBD binding inhibition assay

ACE-2/RBD binding inhibition was tested with a SARS-CoV-2 surrogate virus neutralization test (sVNT) kit (cPass™, Genscript, USA), according to the manufacturer protocol. Plasma samples were diluted 1:20 and incubated with HRP-conjugated wt, Delta, BA.1 or BA.2 RBDs for 30 min, 37 °C. Mixtures were added to ACE2 pre-coated wells and incubated for 15 min, at Room Temperature (RT). After substrate addition and development for 15 min RT, the absorbance was measured at 450 nm on a Multiskan FC Microplate Photometer (Thermo Fisher Scientific). Results are reported as follows: percentage inhibition = (1 – sample OD value/negative control OD value) * 100. Inhibition values ≥ 30% are considered positive results, as indicated by the manufacturer.

Memory B cell ELISpot

Spike-, RBD- and nucleocapsid-specific IgG secreting memory B cells (MBC) were evaluated using the Human IgG Single-Color ELISpot Assay (CTL Europe GmbH, Germany), according to the instructions of the manufacturer. A total of 2×10^6 PBMCs/mL were stimulated with a polyclonal B cell stimulator for 4 days, and then transferred onto multiscreen filter 96-well plates coated with the wt full spike protein (S1 + S2 ECD), RBD, nucleocapsid (all 10 μg/ml; Sino Biological) or anti-IgG capture antibody and incubated overnight at 4 °C. Plates were incubated with anti-human IgG detection solution, then with Tertiary Solution and developed by adding Blue Developer Solution. The number of spots was determined by plate scanning and analysis performed with an Immunospot S6 Ultimate Analyzer (CTL Europe GmbH). For each subject, antigen-specific IgG secreting cells are reported as percentages of total IgG secreting cells.

Multiparametric flow cytometry

Multiparametric flow cytometry was performed to identify RBD-specific B cells[22]. Approximately two million of PBMCs were stained for 30 min at 4 °C with recombinant biotinylated-RBD (BioLegend) conjugated with streptadivin (SA)-Allophycocyanin (APC), together with the following

fluorescent antibodies: CD3-BV650 (clone OKT3); CD21-FITC (clone BLY4), CD19-BUV395 (clone SJ25C1), CD10-PECF594 (clone HI10A), IgM-BV605 (clone G20-127), IgD-BV711 (clone IA6-2), CD27-BV786 (clone O323), CD20-APCH7 (clone 2H7), CD38-BUV737 (clone HB7), IgG-PECy7 (clone G18-145; all from BD Biosciences), IgA-Vio blue (clone IS11-8E10; Miltenyi Biotec; GE). All antibodies were titrated for optimal dilution. Following surface staining, cells were washed once with PBS and labeled with Zombie Aqua Fixable Viability Kit (Thermofisher) according to the manufacturer instruction. Cells were fixed in BD fixation solution (BD Biosciences) and acquired with LSRFortessa X20 SO flow cytometer (BD Biosciences). Manual data analysis was performed using *FlowJo* v10 (TreeStar; OR, USA).

Dimensionality reduction and gaussian mixture clustering

To capture complex and non-linear relationships within the high-dimensional serological feature space, Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (tSNE) were implemented as dimensionality reduction techniques. The application of UMAP was carried out using the *umap* package v0.2.10[23] in a '*umap-learn*' configuration, while tSNE was implemented through the *Rtsne* package v0.17[24] in an R environment. Both methods were applied to z-scored scaled data of serological parameters (antibody concentrations targeting spike and RBD antigens, along with ACE-2/RBD binding inhibition values, both for wt, Delta, Omicron BA.1, and Omicron BA.2 variants) resulting in 12 variables per patient. Clustering of dimensionality reduced data was performed using a Gaussian Mixture clustering, through the *mclust*[25] library v6.1.1. Clustering application was unsupervised, with the number of clusters left unspecified and automatically detected to minimize the Bayesian Information Criterion (BIC) value. Comparison between the two different dimensionality reduction techniques was performed using Within-Cluster Sum of Squares and Average Silhouette Width through the *fpc* library v2.2-13[26].

Classification models

K-Nearest Neighbor (k-NN), Support Vector Machines with a Radial Basis Function kernel (SVM-RBF) and Random Forest (RF) classifiers were employed to develop a predictive model capable of distinguishing immunological profiles of infected from non-infected individuals and thus identifying unaware infected individuals. Both the Model Construction phase, encompassing training and testing on the labeled dataset, and the Model Application to the unlabeled dataset, were performed using z-scored serological data (wt, Delta, Omicron BA.1, and BA.2 spike- and RBD-specific IgG concentrations and ACE-2/RBD binding inhibition, as well as the AUC values for BA.2 N-specific IgG) resulting in a total of 13 variables per patient. Model Construction was conducted using the *caret* library v7.0-1[27] employing a 5-fold Cross-Validation strategy to optimize hyperparameters tuning for each classifier. Performance metrics, including Accuracy, Precision and Recall were evaluated on the test set throughout the cross-validation process. For all three classifiers, variable importance was assessed to identify the most relevant features contributing to the classification task. For the RF model, feature importance was computed using the *varImp* function from the *caret* package, while for SVM-RBF and k-NN, a permutation-based approach was employed. All three pre-trained classifiers were then applied to the Unlabeled dataset during the Model Application phase. For each participant, the final classification was assigned based on a majority-voting consensus strategy among the outputs of the three models.

For Model Construction, a subset of participants was selected as a representative cohort of Infected (*mcI*) and Non-Infected (*mcNI*) participants, forming the labeled dataset. Inclusion criteria for *mcI* participants were: (i) SARS-CoV-2 positive swab; (ii) >6 N-specific MBC/10^6 cells; (iii) AUC for BA.2 N-specific IgG higher than the pre-vaccination mean plus two standard deviations (threshold=0.981 AUC value). Conversely, *mcNI* participants were defined by: (i) absence of a self-reported infection; (ii) null frequency of N-specific MBC/10^6 cells; (iii) AUC for BA.2 N-specific IgG lower than the pre-vaccination mean (threshold=0.603 AUC value). The Model Application phase was conducted on the remaining participants for whom all serological and B cellular variables were assessed, and: (i) did not meet the inclusion criteria for Model Construction; (ii) non-infection status remained uncertain (unlabeled samples).

Self-Organizing Map (SOM) analysis of flow cytometry data

The MBC population analysed using *FlowJo* v10 (TreeStar) was gated as live, singlets, $CD3^-/CD19^+$ cells, $CD20^+/CD10^-$ after exclusion of the naïve B cells ($CD27^-/IgD^+$) (Supplementary Fig. 1). MBC data were exported from *FlowJo* as FCS files and then imported into the R environment, compensated and transformed[28]. Clustering analysis was performed following the *FlowSOM* function pipeline (*FlowSOM* package v2.4.0)[29]. The expressions of CD27, IgD, CD21, CD38, IgM, IgA and IgG markers were included in the clustering analysis. Marker expression was normalized as z-score (mean = 0, standard deviation = 1) and grid size was set to 10 × 10. Similar nodes were merged in 15 metaclusters (metaclustering step). The Euclidean distance was used in both the *FlowSOM* clustering and metaclustering. Thresholds to bisect positive and negative cells for each marker expression were automatically set with the *flowDensity*[30] package v1.34.0. *FlowSOM* results were displayed as a heatmap reporting the percentage of positive cells for each marker within the metacluster. Frequency of antigen-specific B cells was calculated by importing in the R environment the RBD^+ gate defined on *FlowJo*, by using the *GetFlowJoLabels* function.

Statistical analysis

Numeric variables were reported as medians with interquartile ranges (IQR) as measure of dispersion, when not differently specified, and binary variables were reported as counts and percentages. By visual inspection (Q-Q plots and histograms) it was observed that some features, such as antibody concentrations, AUC values and antigen specific frequencies did not follow a Gaussian distribution, so only non-parametric statistical tests were employed in this study. All tests were also two-tailed, as no prior assumptions were made before the analyses. Unpaired Mann–Whitney test and Kruskall–Wallis test followed by Dunn's post-test for multiple comparative tests, were used for assessing statistical significance between 2 groups and between 3 or more groups, respectively. Fisher's exact test was used for assessing differences in number of participants positive for the SARS-CoV-2 surrogate virus neutralization test between groups, as well as for assessing differences in categorical variables in Table 1 and 2. A p value (*P) < 0.05 was considered significant. Analyses were carried out using R Statistical Software and GraphPad Prism v10 (GraphPad Software; CA, USA).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Participants

A group of 116 healthy participants, vaccinated with the 2-dose primary SARS-CoV-2 vaccination cycle, followed by a booster ~5–7 months later, was included in the study, as described in Table 1. Among them, 36 (31.03%) were male and 80 were female (68.97%), with a median age of 49 (range: 24–81). Eighty-two participants (70.69%) received mRNA vaccines during their 2-dose primary vaccination cycle (mRNA-1273 or BNT162b2), and 34 (29.31%) an adenovirus-based vaccine (ChAdOx1 nCoV-19, AZD1222). As for the booster dose, all participants received the Wuhan original monovalent mRNA vaccines. Sixty-eight participants (58.62%) never self-reported a SARS-CoV-2 infection, while 48 participants (31.38%) self-reported a previous SARS-CoV-2 infection. Among them, 9 (18.75%) self-reported infection before the booster dose, and 39 (81.25%) self-reported an infection after the booster dose.

Long-term immune response upon SARS-CoV-2 vaccine booster dose

To determine the long-term immunity generated by the vaccine and/or infection, the spike- and RBD-specific immune responses, targeting the wt

Table 1 | Characteristics of study participants and of participants classified as High and Low Responders

Clinical variables	Study participants (*n* = 116)	tSNE-GMM classification		
		High Responders (*n* = 59)	Low Responders (*n* = 57)	*P* value
Gender – *n* (%)				0.842[a]
Female	80 (68.97%)	40 (67.80%)	40 (70.18%)	
Male	36 (31.03%)	19 (32.20%)	17 (29.82%)	
Age – median (range)	49 (24–81)	47 (28-80)	51 (24–81)	0.431[b]
2-dose primary vaccination cycle – *n* (%)				0.625[a]
mRNA-1273	10 (8.62%)	4 (6.78%)	6 (10.53%)	
BNT162b2	72 (62.07%)	39 (66.10%)	33 (57.90%)	
ChAdOx1 nCoV-19, AZD1222	34 (29.31%)	16 (27.12%)	18 (31.57%)	
Vaccine booster dose – *n* (%)				0.697[a]
mRNA-1273	76 (65.52%)	40 (67.80%)	36 (63.15%)	
BNT162b2	40 (34.48%)	19 (32.20%)	21 (36.85%)	
Infected Participants – *n* (%)	48 (41.38%)	34 (57.62%)	14 (24.56%)	3.3e-04[a]
Infected before booster dose – *n* (% of total infected)	9 (18.75%)	3 (6.25%)	6 (12.50%)	0.318[a]
Infected after booster dose – *n* (% of total infected)	39 (81.25%)	31 (64.58%)	8 (16.67%)	1.19e-05[a]
Days from infection – median (IQR)[c]	106.5 (46.5–183)	94 (31–151.5)	180.5 (73–534.5)	0.037[b]

[a]Fisher test was used to assess significant differences between High Responders and Low Responders.
[b]Mann–Whitney test was used to assess significant differences between High Responders and Low Responders.
[c]Days elapsed between the date of infection and the date of 6 months post-boost blood sample collection.

Table 2 | Characteristics of participants used for classification models

Clinical variables	Participants for Model Construction (*n* = 34)			Participants for Model Application (*n* = 57)				*P* value[b]
	All participants (*n* = 34)	*mcI* (*n* = 18)	*mcNI* (*n* = 16)	All participants (*n* = 57)[a]	*I* (*n* = 16)	*UI* (*n* = 14)	*NI* (*n* = 25)	
Gender – *n* (%)								0.009
Female	29 (85.29%)	16 (88.89%)	13 (81.25%)	33 (57.89%)	9 (56.25%)	6 (42.86%)	17 (68%)	
Male	5 (14.71%)	2 (11.11%)	3 (18.75%)	24 (42.11%)	7 (43.75%)	8 (57.14%)	8 (32%)	
Age – median (range)	50 (28–81)	49 (34–56)	50.5 (28–81)	49 (24–70)	47.5 (28–70)	42.5(24–59)	52.5 (26–69)	0.582
2-dose primary vaccination cycle - *n* (%)								0.140
mRNA-1273	1 (2.94%)	0 (0%)	1 (6.25%)	9 (15.79%)	3 (18.75%)	3 (21.43%)	3 (12%)	
BNT162b2	22 (64.71%)	10 (55.56%)	12 (75%)	35 (61.40%)	9 (56.25%)	8 (57.14%)	16 (64%)	
ChAdOx1 nCoV-19, AZD1222	11 (32.35%)	8 (44.44%)	3 (18.75%)	13 (22.81%)	4 (25%)	3 (21.43%)	6 (24%)	
Vaccine booster dose - *n* (%)								0.491
mRNA-1273	21 (61.76%)	13 (72.22%)	8 (50%)	40 (70.18%)	11 (68.75%)	10 (71.43%)	17 (68%)	
BNT162b2	13 (38.24%)	5 (27.78%)	8 (50%)	17 (29.82%)	5 (31.25%)	4 (28.57%)	8 (32%)	

[a]2 participants self-reported infected were erroneously classified as non-infected by the consensus-based model, and subsequently excluded from the analyses. [b]Statistical comparison was performed between the group of participants used for Model Construction and the group of participants used for Model Application.

strain, the Delta, Omicron BA.1 and Omicron BA.2 variants, a blood sample was collected 6 months following the booster dose. Compared to pre-boost, a significant increase in the wt spike-specific IgG levels was detected at post-boost (median values of 582.6 [210.5–1218] and 6847 [2791–14988] ng/ml, respectively, $^{***}P < 0.001$; Fig. 1a). Upon boosting, IgG levels specific for the wt spike were similar to the ones specific for the spike of the BA.2 variant (median of 7547 [2773–12886] ng/ml, respectively), and significantly higher compared to those specific for the Delta and Omicron BA.1 variants (median of 4457 [1850–8813] and 2125 [935.3–4265] ng/ml, $^{*}P = 0.028$ and $^{***}P < 0.001$, respectively; Fig. 1b). The IgG response specific for the wt RBD was also significantly higher compared to the one specific for Omicron BA.1 and BA.2 RBD (median of 10314 [4462–20192], 3548 [1241–7344] and 3170 [1206–7361] ng/ml, respectively; $^{***}P < 0.001$; Fig. 1c). The functionality of the spike-specific antibodies was assessed via their ability to block the RBD/ACE-2 interaction, employing a sVNT. Upon the booster dose, a significantly higher number of participants developed antibodies with binding inhibition capacity above the threshold value compared to the pre-boost analysis, for all viral variants ($^{***}P < 0.001$, Fig. 1d). Nevertheless, when comparing the binding inhibition capacity after the booster dose, a significant difference was observed between Omicron BA.1 and wt strain values ($^{\#\#\#}P < 0.001$; Fig. 1d). The frequency of circulating wt RBD-specific B cells, identified among non-naïve $CD19^{+}$ B cells (gating strategy in Supplementary Fig. 1) was similar before and 6 months after the booster administration (0.21 [0.11–0.34] and 0.17 [0.08–0.3] % of $CD19^{+}$ cells, respectively; Fig. 1e). Nevertheless, upon in vitro stimulation, the amount of wt spike-specific IgG-secreting MBC was significantly higher at 6 months

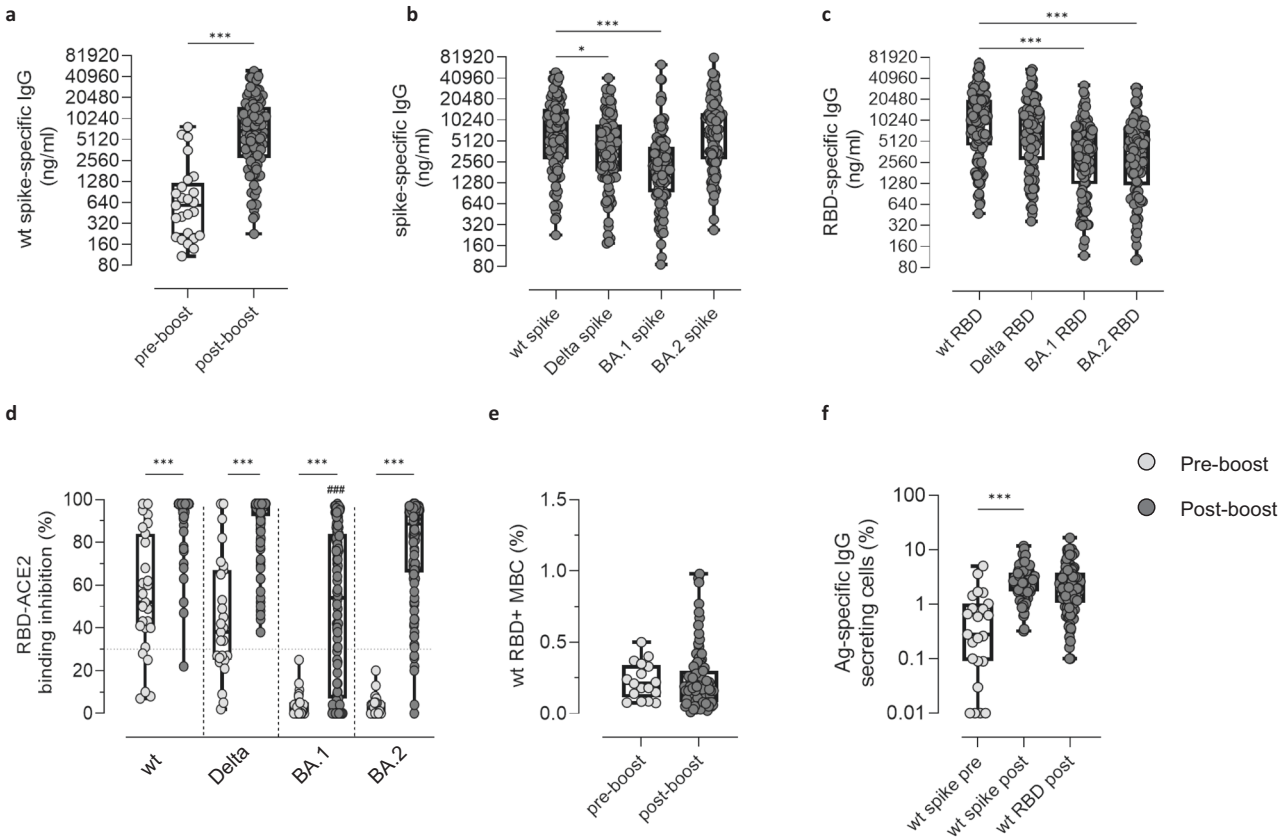

Fig. 1 | Spike- and RBD-specific immune responses. Antigen-specific humoral and cellular immune responses were evaluated in 116 participants 6 months after the booster dose (post-boost) and compared to pre-boost data. **a** Wt Spike-specific IgG assessed pre- and post-boost. Post-boost analysis of IgG specific for spike (**b**) and RBD (**c**) targeting the wt strain and the Delta, Omicron BA.1 and Omicron BA.2 variants. IgG concentrations, detected by ELISA, were expressed in ng/ml. **d** sVNT assay used to evaluate the capacity of plasma antibodies to bind the RBD of the wt, Delta, Omicron BA.1 and Omicron BA.2 strains thus blocking its interaction with ACE-2, at pre (light gray dots) and post boost (dark gray dots). Data are reported as RBD/ACE-2 binding inhibition percentage. A threshold (dotted red line), placed at 30% inhibition percentage, was used to discriminate between positive and negative samples. **e** Frequencies of wt RBD$^+$ B cells, analyzed by multiparametric flow cytometry. Data are reported as percentage of total CD19$^+$ cells of each subject. **f** Frequencies of spike- (for both pre- and post-boost time points, light gray and dark gray dots, respectively) and RBD-specific IgG secreting cells (for the post-boost time point, dark gray dots), assessed by Memory B cell ELISpot. Frequencies are reported as a percentage of total IgG secreting cells. Data are shown as box and whiskers plot showing the minimum and maximum of all the data. Statistical differences between groups were assessed using Unpaired Mann–Whitney test (**a**, **e** and **f**), Kruskal-Wallis test followed by Dunn's post-test for multiple comparisons (**b** and **c**). Fisher's exact test was used to assess differences in the number of individuals who are positive for the sVNT between pre- and post-boost data ($^*P < 0.05$; $^{***}P < 0.001$) and for the wt strain *versus* Delta, Omicron BA.1 and BA.2 viral variants at post-boost ($^{\#\#\#}P \le 0.001$, **d**). $^*P < 0.05$; $^{***}P < 0.001$. MBC, memory B cells. Sample size: pre-boost (**a**: $n = 25$; **d**: $n = 27$; **e**: $n = 17$; **f**: $n = 23$); post-boost (**a**–**f**: $n = 116$).

post-boost compared to pre-boost (2.52 [1.70–3.79] and 0.28 [0.09–1.02] % of total IgG-secreting cells respectively, $^{***}P < 0.001$; Fig. 1f).

In conclusion, the immunological analysis performed 6 months after the booster dose highlighted the critical role of the third vaccine dose in enhancing both the humoral and antigen-specific B cell responses, not only against the spike/RBD antigens of the wild type strain, but also of the Delta and Omicron variants. However, the wide IQR values across all variables indicated a considerable dispersion of data, suggesting a heterogeneous response.

Dimensionality reduction and Gaussian mixture clustering identify high and low responders

To explore post-boost data in an unsupervised manner, the 12 serological variables previously analysed for each participant (reported in Supplementary Table 1) were computationally processed. To capture complex and non-linear relationships within this 12-dimensional feature space and obtain a meaningful two-dimensional representation, two distinct dimensionality reduction techniques, namely UMAP and tSNE, were employed. Following dimensionality reduction, the application of the unsupervised Gaussian Mixture Model (GMM) clustering algorithm identified, in both UMAP- and tSNE-derived embeddings, two distinct clusters –configuration yielding the lowest BIC value– of immune response. To quantitively compare the clustering performances of the two approaches, Within-Cluster Sum of Squares (WCSS) and Average Silhouette Width were computed. The WCSS values were 693.39 for the UMAP-GMM strategy and 494.39 for the tSNE-GMM strategy, indicating greater intra-cluster compactness in the latter. Similarly, the Average Silhouette Width was higher for the tSNE-based approach (value of 0.63) compared to the UMAP-based one (value of 0.56), reflecting better-defined clusters. Given these results, the tSNE-GMM strategy was selected for downstream analyses and its visual representation is showed in Fig. 2a.

tSNE-GMM cluster 2 consistently exhibited a significantly higher IgG response against wt, Delta, Omicron BA.1 and Omicron BA.2 spike and RBD antigens (Fig. 2b, c, $^{***}P < 0.001$), and a significantly higher proportion of participants exhibiting positive values for the RBD/ACE-2 binding inhibition against Omicron BA.1 and BA.2 variants compared to tSNE-GMM cluster 1 (Fig. 2d, $^{***}P < 0.001$ for the BA.1 variant, $^*P = 0.012$ for the BA.2 variant). Consequently, tSNE-GMM cluster 2 is hereafter referred to as High Responders (*HR*) group and tSNE-GMM cluster 1 as Low Responders (*LR*) one.

The potential impact of clinical and demographic variables including gender, age, vaccine formulations, past infections and time since infection

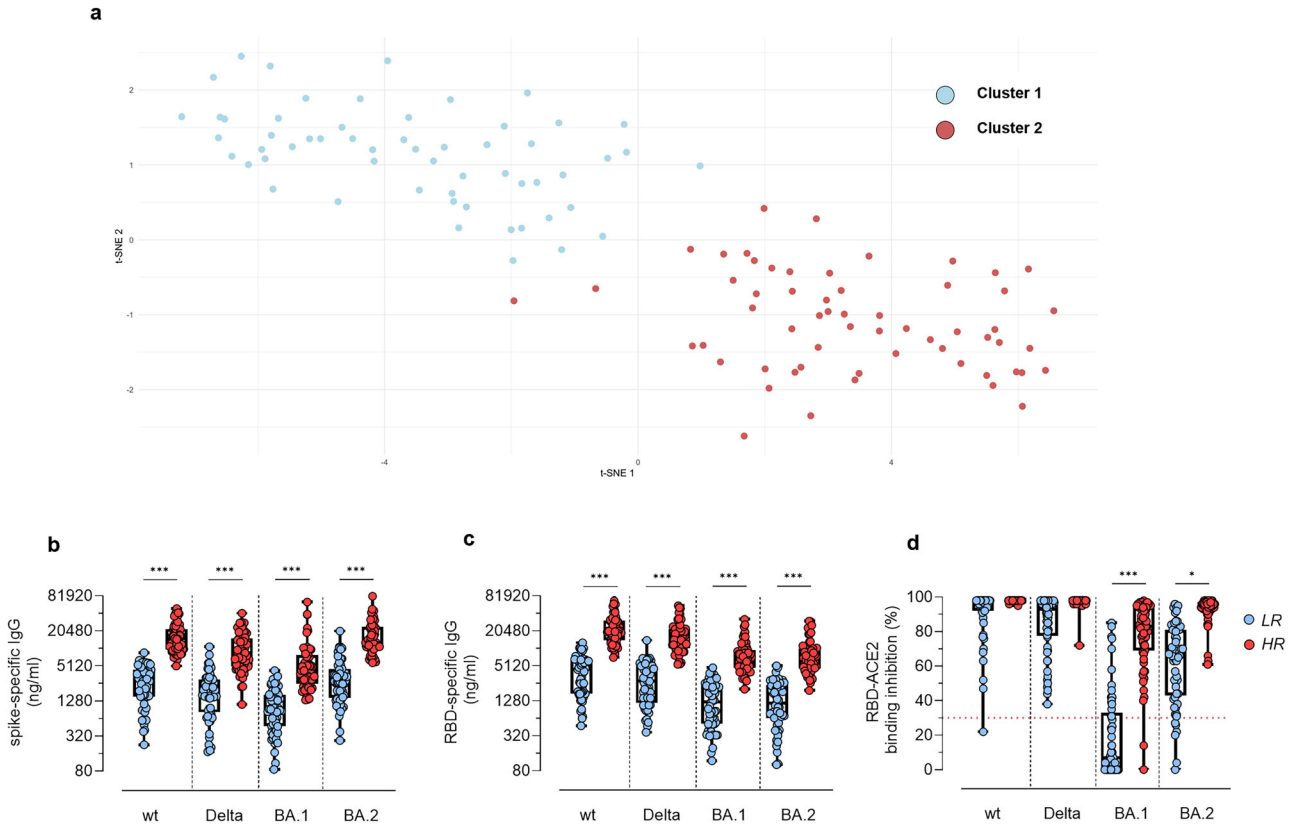

Fig. 2 | Serological data dimensionality reduction and clustering. a Participants represented in a dimensionality tSNE-reduced space derived from 12 serological variables. Clusters were identified using an unsupervised Gaussian-Mixture model. Each dot represents a single participant, and the colors represent distinct detected clusters. Spike (**b**) and RBD (**c**) specific IgG targeting the wt, Delta, Omicron BA.1 and Omicron BA.2 strains. IgG concentrations, detected by ELISA, were expressed as ng/ml. **d** sVNT assay used to evaluate the capacity of plasma antibodies to bind the RBD of the wt, Delta, Omicron BA.1 and Omicron BA.2 strains, thus blocking its interaction with ACE-2. Data are reported as RBD/ACE-2 binding inhibition percentage. A threshold (dotted red line), placed at 30% inhibition percentage, was used to discriminate between positive and negative samples. Data are shown as box and whiskers plot showing the minimum and maximum of all the data. Unpaired Mann–Whitney test was used to assess statistical differences between high (HR) and low responders (LR) in (**b, c**). Differences in the number of *HR* and *LR* participants who are positive for the sVNT against the wt strain and the other viral variants were assessed using Fisher's exact test (**d**). $^{**}P \le 0.01$; $^{***}P \le 0.001$. LR, low responders (cluster 1); HR, high responders (cluster 2). Sample size: cluster 1 (**a**: $n = 57$); cluster 2 (**a**: $n = 59$); HR (**b**–**d**: $n = 59$); LR (**b**–**d**: $n = 57$).

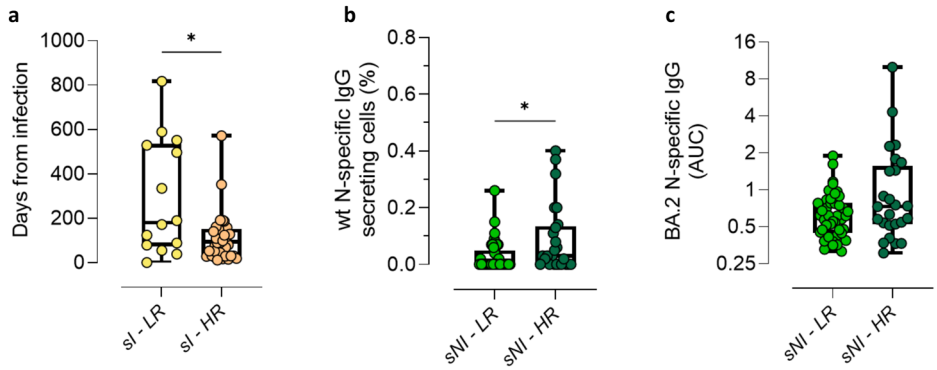

Fig. 3 | Classification of self-reported infectious status into *HR* and *LR* clusters. a Impact of the time elapsed since self-reported infection on the classification as *HR* (*sI-HR*) and *LR* (*sI-LR*). **b** Frequencies of N-specific IgG secreting MBC among *sNI* participants categorized as *HR* (*sNI-HR*) and *LR* (*sNI-LR*). Frequencies are reported as a fraction of total IgG-secreting MBC. **c** N-specific IgG targeting the Omicron BA.2 variant among sNI-HR and sNI-LR. IgG concentrations, detected by ELISA, were expressed as area under the curve (AUC). Data are shown as box and whiskers plot showing the minimum and maximum of all the data. Unpaired Mann–Whitney test was used for assessing the statistical difference between groups. $^{*}P \le 0.05$; $^{**}P \le 0.01$. sI, self-reported infected participants; sNI, self-reported non-infected participants; sI-LR and sI-HR, self-reported infected participants who were clustered as low and high responders respectively; sNI-LR and sNI-HR, self-reported non-infected participants who were clustered as low and high responders respectively; N, nucleocapsid; AUC, area under the curve; MBC, memory B cells. Sample size: sI-LR (**a**: $n = 14$); sI-HR (**a**: $n = 34$); sNI-LR (**b**: $n = 41$; **c**: $n = 43$); sNI-HR (**b**, **c**: $n = 25$).

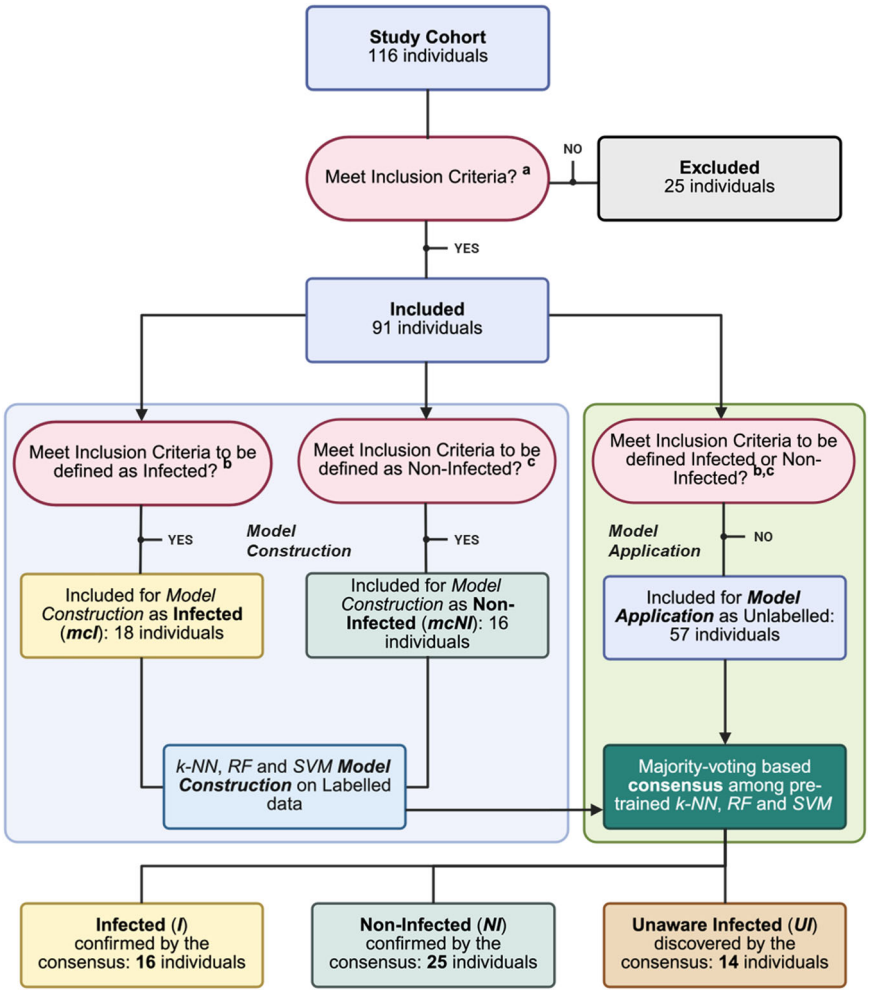

Fig. 4 | Overall strategy for the identification of Unaware Infected participants using Machine Learning Classifiers. a The study cohort (116 individuals) was initially filtered to include only those with complete serological and B-cell data. For the Model Construction (mc) phase, a subset of participants was selected as a representative cohort of infected (*mcI*) and non-infected (*mcNI*) individuals, forming the labeled dataset. **b** Inclusion criteria for *mcI* participants required a positive swab along with a frequency of N-specific MBC/10^6 cells > 6 and an AUC for IgG anti-N > the pre-vaccination mean plus two standard deviations (threshold=0.981 AUC value). **c** *mcNI* participants were defined by the absence of a self-reported infection, a null frequency of N-specific MBC/10^6 cells and an AUC for IgG anti-N ≤ than the pre-vaccination mean (threshold = 0.603 AUC value). Machine Learning classifiers employed in this phase included k-NN, SVM-RBF and RF. The Model Application phase was conducted on the remaining participants for whom all serological and B cellular variables were assessed but who did not meet the inclusion criteria for Model Construction and whose non-infection status remained uncertain (unlabeled samples). *sI*, self-reported infected participants; *sNI*, self-reported non-infected participants; *mcI*, representative group of *sI* used for Model Construction; *mcNI*, representative group of *sNI* used for Model Construction; *I*, *sI* participants classified by the model as infected; *UI*, *sNI* participants classified by the model as infected; *NI*, *sNI* participants classified by the model as non-infected; k-NN, k-Nearest Neighbors; SVM-RBF, Support Vector Machines with Radial Basis Function kernel; RF, Random Forest.

was evaluated to determine their potential influence on classification into *HR* and *LR*. Age, gender and vaccine formulations did not reveal to act as influential variables on cluster categorization ($P > 0.05$, Table 1), while a significantly higher frequency of participants who experienced a self-reported infection were classified as *HR* (71% of self-reported infected participants, $^{***}P < 0.001$).

Among self-reported Infected participants (*sI*) clustered within *HR* and *LR*, a statistically significant difference was observed when comparing the days elapsed from the last infection to the 6 months post-boost blood sample collection (Fig. 3a and Table 1). Indeed, infected participants among the *HR* group contracted the infection more recently than those falling into the *LR* one (median value = 94 days; IQR 31–151.5 days for *sI-HR*, *versus* median value = 180.5 days; IQR 73–534.5 days for *sI-LR*, $^{*}P = 0.037$).

Given the high proportion of asymptomatic and mildly symptomatic infections associated with the emergence of Omicron variants[10], it was investigated the possibility that some self-reported Non-Infected participants within the High Responders group (*sNI-HR*) might have experienced unrecognized infections. This possibility was corroborated by the observation that *sNI-HR* showed significantly higher frequencies of N-specific MBC compared to self-reported Non-Infected participants within the Low Responders group (*sNI*-LR, median value of 0.03% versus 0.00% respectively; $^{*}P = 0.013$; Fig. 3b). Moreover, *sNI-HR* tended to exhibit higher N-specific antibody response compared to *sNI-LR*, although not statistically significant (median AUC value of 0.73 and 0.61 respectively; $P > 0.05$; Fig. 3c). This suggested the potential presence of participants unaware of their infection.

Identification of unaware infected individuals via machine learning classifiers

To identify Unaware Infected participants (*UI*), a predictive model was developed leveraging three distinct Machine Learning classifiers, namely k-NN, SVM-RBF and RF. These models were trained to distinguish immunological profiles of infected and non-infected individuals based on 13 serological variables (reported in Supplementary Table 1). The analysis comprised the Model Construction phase, performed on k-NN, SVM-RBF and RF classifiers, and the Model Application phase, implemented using a majority voting-based consensus approach of the three classifiers (Fig. 4).

Model construction. From the initial cohort of 116 individuals, 25 were excluded due to incomplete serological and B cell data, as these variables were essential for the subsequent Model Construction and Application phases respectively (Fig. 4). This reduction resulted in a subset of 91 individuals. Based on predefined criteria—positive swab results, N-specific memory B cells, and anti-N IgG values (as detailed in the Methods and Materials – Classification Models section)—a subset of 34 participants was selected, comprising 18 *mcI* (model construction Infected individuals) and 16 *mcNI* (model construction Non-Infected individuals) participants. The clinical characteristics of these subgroups are reported in Table 2. These 34 participants were used to train and evaluate the three classifiers, k-NN, SVM-RBF and RF, via a 5-fold cross-validation strategy, with 70% of the data allocated for model training and the remaining 30% for testing. The classification performance of each

Table 3 | Metric performances of classifiers models during the cross-validation

Metrics	Accuracy *Mean (Min-Max)*	Recall *Mean (Min-Max)*	Precision *Mean (Min-Max)*
k-NN	0.90 (0.67-1)	0.93 (0.67-1)	0.89 (0.67-1)
RF	0.97 (0.83-1)	0.93 (0.67-1)	1 (1-1)
SVM-RBF	0.94 (0.83-1)	0.93 (0.67-1)	0.96 (0.80-1)

k-NN k-Nearest Neighbors, *SVM-RBF* Support Vector Machines with Radial Basis Function kernel, *RF* Random Forest.

Table 4 | Variable Importance analysis for the classifiers models

Variable Importance	k-NN	Random Forest	SVM-RBF
Omicron BA.2 N-specific IgG (AUC)	0.088	1	0.176
Spike-specific IgG (concentration)			
Wild type	0	0.016	0
Delta	0	0.094	0.015
Omicron BA.1	0	0.008	0.015
Omicron BA.2	0.015	0.133	0.058
RBD-specific IgG (concentration)			
Wild type	0	0.004	0
Delta	0	0.004	0
Omicron BA.1	0	0.003	0.015
Omicron BA.2	0	0.072	0.029
ACE2-RBD binding inhibition			
Wild type	0	0	0.029
Delta	0	0.003	0.029
Omicron BA.1	0.029	0.217	0.029
Omicron BA.2	0	0.278	0.044

model during this phase is reported in Table 3, demonstrating that all three models achieved optimal performances. Variable importance analysis across all three models identified Omicron BA.2 N-specific IgG AUC values, Omicron BA.2 spike-specific IgG concentrations and ACE2-BA.1 RBD binding inhibition percentages as the most important features (Table 4). However, some differences were observed in the feature importance attribution across classifiers. While the k-NN did not highlight any additional informative feature beyond those shared across models, both the SVM-RBF and the RF models assigned non-zero importance scores to a broader subset of serological variables (Table 4). In particular, the SVM-RBF assigned a zero importance score to wt spike-specific concentrations, as well as wt and Delta RBD-specific IgG concentrations, whereas the RF model excluded only the ACE2-wt RBD binding inhibition percentages. However, all three classifiers demonstrated high predictive performances and were thus retained for downstream analysis and included in the consensus-based approach during the Model Application phase.

Model application. k-NN, RF, and SVM-RBF pre-trained models were independently applied to the remaining 57 participants that did not meet the inclusion criteria of Model Construction phase and whose non-infection status was uncertain (Fig. 4). Among the 57 analysed participants, whose clinical and demographic characteristics are reported in Table 2, 18 self-reported Infection (*sI*) and 39 self-reported a Non-Infection (*sNI*). The application of the majority-voting consensus among the outputs of the three models correctly identified 16 out 18 self-reported Infected individuals, yielding a Recall of 0.89 in this Model Application phase. Recall was the only performance metric that could reliably be assessed, given the uncertainty regarding the non-infection status of the remaining participants. These 16 individuals, who self-reported a previous infection and were correctly identified by the consensus strategy, will be referred to as Infected (*I*). Among the 39 *sNI* participants, 14 were classified as infected and therefore referred to as Unaware Infected (*UI*). The remaining 25 participants were confirmed and classified as Non-Infected (*NI*) (Fig. 4).

To further confirm the *UI* classification assigned by the consensus strategy, the frequency of N-specific MBC, assessed by ELISPOT, was compared between the *UI* and *NI* groups. Participants classified by the consensus approach in the *UI* group showed a statistically significant higher frequency of N-specific MBC compared to *NI* participants (median 0.09% and 0% respectively; $^{**}P = 0.003$; Supplementary Fig. 2), confirming an unaware infected profile. In summary, the application of this strategy allowed for the reliable identification of 14 participants with an unreported infection history based on their immunological profiles, demonstrating its potential to uncover hidden infection status.

Characterization of the immunological profile of participants stratified in Infected, Unaware Infected and Non-Infected participants

The immunological response was analysed based on the stratification of participants into the *I*, *UI* and *NI* groups as determined by the consensus strategy. The 2 participants self-reported Infected but erroneously classified by the model as *NI*, along with the 34 used for the Model Construction phase (*mcI* and *mcNI*) were excluded from this analysis. Participants classified as *UI* exhibited levels of IgG specific for wt and BA.2 RBD (median of 27,311 and 9487 ng/ml, respectively) comparable to participants classified as *I* (median of 21,914 and ng/ml and 8876 ng/ml, respectively). Moreover, their IgG levels were statistically higher compared to *NI* participants (median of 5710 and 2188 ng/ml for wt and BA.2 RBD-specific IgG; $^{***}P < 0.001$; Fig. 5a, b), while no significant differences in the proportion of participants above the binding inhibition threshold value were observed between groups (all $P > 0.05$; Fig. 5c, d). Similar results were observed when the analysis was performed for the serological response specific for Delta and BA.1 variants (Supplementary Fig. 3). A statistically significant higher frequency of wt RBD^+ B cells was observed in *I* and *UI* (median of 0.27% and 0.28%) compared to the *NI* group (median of 0.16%; $^{*}P = 0.018$ and $^{**}P = 0.005$ respectively; Fig. 5e). Participants classified as *I* and *UI* also presented statistically higher frequencies of circulating IgG secreting RBD-specific MBC capable of reactivating upon in vitro stimulation compared to *NI* participants (median frequency of 4.55% in *I*, 3.71% in *UI* and 1.54% in NI; $^{***}P < 0.001$; Fig. 5f).

To compare the phenotypes of the RBD^+ B cells developed among the *I*, *UI* and *NI* groups, the SOM clustering algorithm was applied to the multidimensional flow cytometry data (Fig. 6). According to the combination of the expression of 7 markers (IgD, CD27, CD21, CD38, IgM, IgA, IgG), 12 MBC clusters were identified among the total $CD19^+$ no naïve B cells, and grouped in Ig-switched MBC (IgD^- $CD27^+$), plasmablast/plasma cells (PB/PC; IgD^- $CD38^+$), double negative (DN; IgD^- $CD27^-$) and unswitched MBC (IgD^+ $CD27^+$) (Fig. 6a). Most of the RBD^+ B cells fell into IgG^+ resting MBC (cluster 3), DN $CD21^+$ MBC (cluster 4), DN $CD21^-$ MBC (cluster 12) and IgG^+ activated MBC (cluster 13) (Fig. 6b). When comparing the phenotypes of RBD^+ B cells among *I*, *UI* and *NI*, statistically higher levels of RBD^+ IgG+ resting B cells (cluster 3) were detected in participants belonging to the *I* and *UI* groups compared to *NI* (median of 24.53%, 28.1% and 14.69%, respectively, $^{*}P = 0.041$ and 0.016, respectively, Fig. 6c). Conversely, *NI* showed statistically higher levels of RBD^+ DN1 $CD21^+$ B cells (cluster 4) compared to *I* (median of 37.17% and 20% respectively, $^{*}P = 0.02$, Fig. 6d).

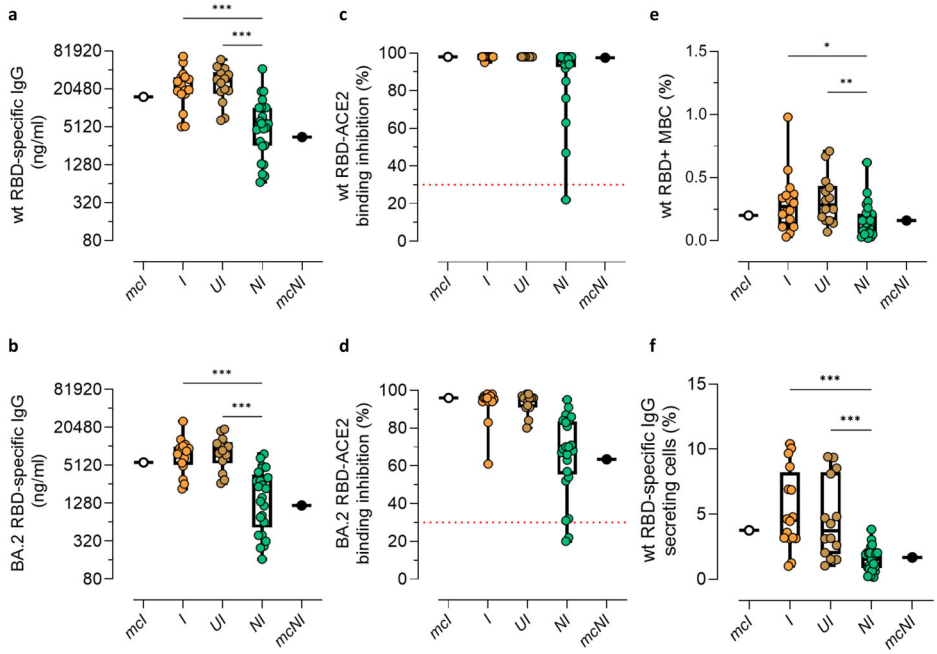

Fig. 5 | RBD-specific immune responses in groups with different immunological profile, as classified by the consensus-based model. Humoral and B cellular RBD-specific immune responses were analysed in self-reported infected participants classified as infected (*I*), and in self-reported non-infected participants classified as Unaware Infected (*UI*) and as non-infected (*NI*). **a** RBD-specific IgG targeting the wt strain and **b** the Omicron BA.2 variant are reported as ng/ml. **c** sVNT used to assess the capacity of plasma antibodies to bind the RBD of wt, and **d** Omicron BA.2 strains. Data are reported as RBD/ACE-2 binding inhibition percentage. A threshold (dotted red line), placed at 30% inhibition percentage, was used to discriminate between positive and negative samples. **e** Frequencies of wt RBD$^+$ B cells, analyzed by multiparametric flow cytometry. Data are reported as fraction of total CD19$^+$ B cells of each subject. **f** Frequencies of wt RBD-specific IgG secreting cells, evaluated by using the ELISpot assay upon in vitro stimulation. Frequencies are reported as a fraction of total IgG-secreting cells. Data are shown as box and whiskers plot showing the minimum and maximum of all the data. Kruskal-Wallis test, followed by Dunn's post-test for multiple comparisons, was used for assessing statistical differences between *I*, *UI* and *NI* groups (**a**, **b**, **e**, **f**). Differences in the number of *I*, *UI* and *NI* participants who are positive for the sVNT against the wt strain and the Omicron BA.2 variant were assessed using Fisher's exact test (**c**, **d**). The individual data points in white (*mcI*) and black (*mcNI*) represent the median value of the representative group of self-reported infected and self-reported non-infected participants used in the Model Construction phase respectively, and were not included in the statistical analyses. $^{**}P \le 0.01$; $^{***}P \le 0.001$. *sI,* self-reported infected participants; *sNI*, self-reported non-infected participants; *mcI,* representative group of *sI* used for Model Construction; *mcNI,* representative group of *sNI* used for Model Construction; *I, sI* participants classified by the model as infected; *UI* , *sNI* participants classified by the model as infected; *NI, sNI* participants classified by the model as non-infected; MBC, memory B cells; wt RBD$^+$, wt RBD-specific B cells. Sample size: *mcI* (**a**–**f**: $n = 18$); *I* (**a**–**f**: $n = 16$); *UI* (**a**–**f**: $n = 14$); *NI* (**a**–**f**: $n = 25$); *mcNI* (**a**–**f**: $n = 16$).

Discussion

In the present study, different Machine Learning approaches were applied to analyse a group of healthy individuals upon the third dose with mRNA SARS-CoV-2 vaccines, to discover participants who were unaware of a previous infection, and to dissect hybrid and vaccine-induced immunity. Indeed, the application of these integrative approaches allowed to identify a subset of 14 individuals unaware of a previous SARS-CoV-2 infection. Following the emergence of the Omicron variant, associated with lower pathogenicity but higher transmissibility[31], tracking infections, especially asymptomatic ones, has become significantly more challenging. Nevertheless, recognizing individuals with unaware infection and correctly distinguishing the immunological profile induced solely by vaccination from that induced by a combination of vaccination and infection booster (hybrid immunity) is fundamental for understanding the durability and persistence of immune responses induced by mRNA-based vaccines. Defining a reliable method for the correct identification of unaware infected individuals could also prove highly valuable in refining vaccination policies during epidemics/pandemics, ensuring the proper administration of vaccine doses.

Immunological data employed in this study included serological and B cellular data assessed against wt, Delta, Omicron BA.1 and Omicron BA.2 spike and RBD antigens, 6 months after the booster dose. Data detected in the group clearly confirmed that the third vaccine dose strongly boosted the immune response towards the vaccine wt spike antigen and elicited an evident Omicron-specific immunity, as previously reported[8,32–34]. However, the cohort displayed substantial variability in the distribution of both humoral and cellular data, possibly due to unaware SARS-CoV-2 infection.

To analyse this variability in an unsupervised manner, dimensionality reduction techniques, UMAP and tSNE, followed by an unsupervised Gaussian Mixture clustering (GMM) were employed. A comparison of the two strategies revealed GMM clustering on tSNE-reduced data as the optimal method yielding the best performances in terms of WCSS and Average Silhouette Score. This approach allowed to objectively stratify the population without relying on predefined labels, thus minimizing biases. Additionally, it enabled to classify the population by simultaneously analysing 12 serological variables, thus providing a comprehensive representation of the immunological profile of each subject, resulting in a classification with clear biological relevance. The algorithms classified the population into high responders, characterized by elevated humoral responses, and low responders, displaying significantly lower humoral responses. The high responder cluster predominantly comprised individuals who self-reported a prior infection. Notably, in individuals who declared a prior infection the impact of the infection in promoting higher humoral responses decreased over time from the infection. This suggests that the effect of prior infection on antibody concentration and RBD/ACE-2 binding inhibition capacity wanes within a few months from the infection.

For what it concerns individuals self-reported as non-infected, the ones falling in the high responder group exhibited higher responses towards the N-protein compared to their low responder counterparts, suggesting the potential presence of unaware infected individuals within the self-reported

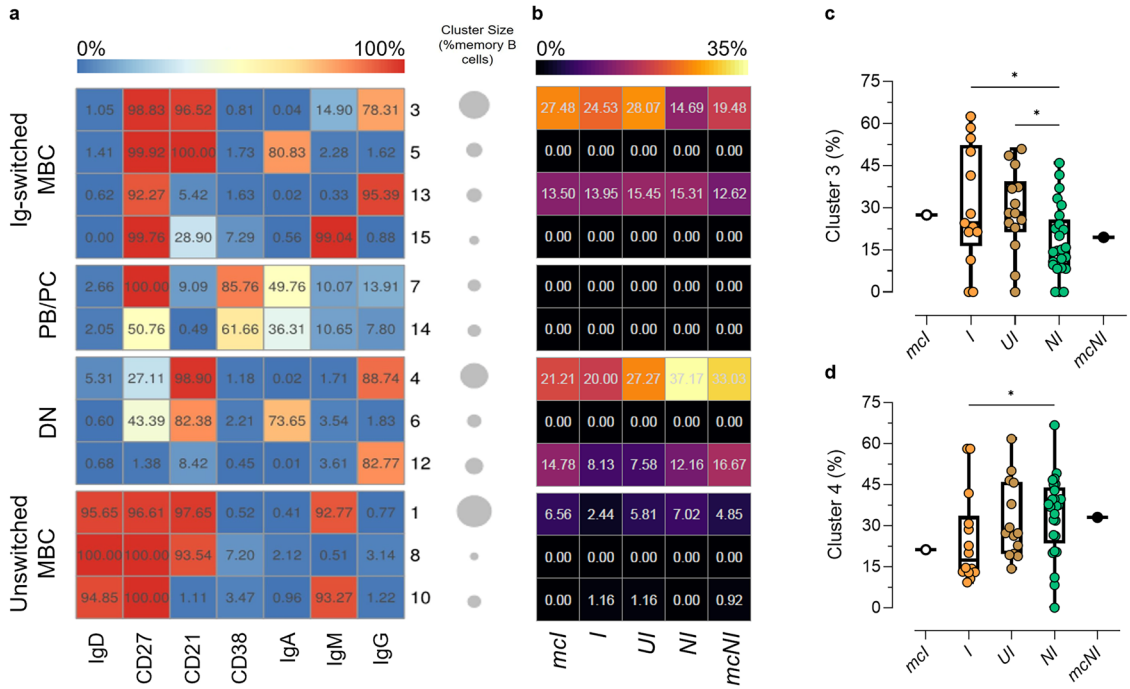

Fig. 6 | Clustering analysis of wt RBD + B cell subsets in participants classified as infected, Unaware Infected and non-infected by the consensus-based model. *FlowSOM* algorithm was used to characterize different phenotypes of memory B-cells among self-reported infected participants classified as infected (*I*), and self-reported non-infected participants classified as Unaware Infected (*UI*) and as non-infected (*NI*). **a** Heatmap of clusters from the *FlowSOM* analysis of total memory B-cells, with markers (IgD, CD27, CD21, CD38, IgA, IgM, IgG) reported in column, and the identified clusters in row. Clusters are grouped in Ig-switched MBC, plasmablasts/plasmacells (PB/PC), double negative (DN) and unswitched MBC. The percentage of cells positive for each marker is visualized with a color scale from blue (0%) to red (100%). A bubble plot reported the size of each cluster, with circle areas proportional to the number of cells included in each cluster. **b** Median frequencies of wt RBD$^+$ B cells in each clusters reported as a fraction of total wt RBD$^+$ B cells in a color scale from black (lowest value) to light yellow (highest value). **c** Frequencies of wt RBD$^+$ B cells within cluster 3 (IgG$^+$ resting) and **d** cluster 4 (DN CD21$^+$IgG$^+$), expressed as a fraction of wt RBD$^+$ B cells in each sample. Data are shown as box and whiskers plot showing the minimum and maximum of all the data. Kruskal-Wallis test, followed by Dunn's post-test for multiple comparisons, was used for assessing statistical differences between *I*, *UI* and *NI* groups. The individual data points in white (*mcI*) and black (*mcNI*) represent the median value of the representative group of self-reported infected and self-reported non-infected used to train the consensus model respectively, and were not included in the statistical analyses. *$P \le 0.05$. *sI,* self-reported infected participants; *sNI*, self-reported non-infected participants; *mcI,* representative group of *sI* used for Model Construction; *mcNI,* representative group of *sNI* used for Model Construction; *I*, *sI* participants classified by the model as infected; *UI*, *sNI* participants classified by the model as infected; *NI*, *sNI* participants classified by the model as non-infected; wt RBD$^+$, RBD wt-specific B cells. Sample size: *mcI* (**c**, **d**: $n = 18$); *I* (**c**, **d**: $n = 16$); *UI* (**c**, **d**: n = 14); *NI* (**c**, **d**: $n = 25$); *mcNI* (**c**, **d**: $n = 16$).

non-infected group. However, the N-specific immune response as biomarker of infection has shown important limitations such as the demonstrated limited persistence of N-specific antibodies over time, the existence of individuals who remain seronegative shortly after infection[9,12–14,16], and the reported cross-reactivity of antibodies developed against common cold human coronaviruses[15].

To overcome such limitations, several Machine Learning classifiers, including k-NN, SVM with Radial Basis Function kernel and Random Forest, were employed to develop a model capable of identifying individuals unaware of a potential prior infection using easily obtainable and cost-effective serological data. The analysis was structured in two main phases: Model Construction and Model Application. In the Model Construction phase, conducted on a subset of labeled individuals, all three models, k-NN, RF and SVM-RBF, achieved optimal performances in terms of accuracy, precision and recall. A variable importance analysis further revealed that Omicron BA.2 N-specific IgG, Omicron BA.2 spike-specific IgG and Omicron BA.1 RBD/ACE-2 binding inhibition were the most important features across all three methods. During the Model Application phase, each of the three pre-trained classifiers was independently applied to the unlabeled dataset, and final class assignments were determined through a majority-voting consensus across their individual predictions. Among participants self-reported as non-infected, the model identified approximately 25% individuals (14 out of 57) with an immunological profile similar to self-reported infected participants, who were therefore classified as unaware infected, a percentage in line with the proportion of asymptomatic infections observed since the emergence of the Omicron variant in several studies[10,35,36].

The immunological profile of the unaware infected subgroup, showed (i) a higher frequency of N-specific MBC compared to model-confirmed non-infected individuals and, (ii) values of wt and BA.2 RBD-specific antibodies, frequency of wild type RBD$^+$ MBC and circulating IgG-secreting cells comparable to those observed in model-confirmed infected individuals and significantly higher compared to the non-infected ones. The identification of previously unrecognized infections can improve the comprehension of immune responses elicited by vaccination alone or combined with infection. By determining the true prevalence of past infections, including asymptomatic cases, the Machine Learning strategy here presented provides a more accurate understanding of virus spread. Moreover, it can guide vaccine prioritization by identifying individuals with robust hybrid immunity and distinguishing those who require additional booster doses, thereby optimizing vaccine allocation, particularly in situations of limited vaccine supply.

The multiparametric analysis of wild type RBD$^+$ MBC, performed employing the SOM clustering algorithm, showed a prevalence of RBD-specific IgG$^+$ resting memory B cells in model-confirmed infected and unaware infected groups, while DN1 (CD27$^-$ IgD$^-$ CD21$^+$) subset was higher

in the model-confirmed non-infected group. DN are a B cell subset identified both in healthy individuals[37] and in chronic infected patients[22], individuals affected by autoimmune diseases[38] and elderly[39]. In healthy individuals, DN B cells express switched BCR, developmental markers, and somatically hypermutated Ig genes[37,38,40,41], while in aged and virally infected individuals show characteristics of exhausted cells[22,42]. In our dataset, the prevalence of IgG^+ resting MBC in vaccinated and infected individuals could be due to a multifaceted immune activation, possibly triggered by the broader spectrum of antigens encountered during natural infection. On the other hand, the increased presence of DN1 B cells in those who received only vaccination may suggest the vaccine's unique ability to activate alternative differentiation pathways[28].

The study has some known limitations. First, the relatively small number of participants compared to the high dimensionality of the immunological data, that may limit the robustness and generalizability of the findings regarding the application of Machine Learning algorithms. Additionally, while a positive swab test was available for self-reported infected participants, the absence of periodic sampling prevented the identification of non-reported infections in self-reported non-infected individuals via PCR testing. This led to the adoption of restrictive selection criteria for the inclusion in the Machine Learning classifiers in the Model Construction phase. While this improved the reliability of the chosen non-infected group, it also reduced the number of participants available for constructing the classifiers, potentially impacting their performances.

Future developments could involve the application of the proposed Machine Learning strategies to larger datasets where infection status is confirmed by PCR validation and information on disease severity is available. Such applications may offer deeper insights into the robustness and performance of the proposed approaches, and help determine whether infection severity, in addition to time since infection, influences the stratification of individuals into responder clusters.

In conclusion, the Machine Learning strategies described allowed to identify individuals unaware of a prior infection and to dissect hybrid and vaccine-induced immunity, revealing a unique immunological signature associated with hybrid immunity. The Machine Learning strategies here implemented for identifying individuals unaware of a previous infection could be easily adapted to other pathogens, and could represent a valuable tool for monitoring infections in clinical trials. In the context of an epidemic/pandemic virus, this approach could be applied for comparing the capacity of various vaccine formulations in preventing infection *versus* disease. Therefore, this Machine Learning methodology can enhance epidemiological surveillance by providing accurate infection rates, and guide vaccine prioritization for individuals who need them most.

Data availability

The dataset supporting the findings of this study has been deposited in the Zenodo data repository and is publicly available at https://doi.org/10.5281/zenodo.15518709[43]. The repository includes anonymized immunological and infection-related data used for the Machine Learning and statistical analyses described in the manuscript.

Code availability

The code underlying the Machine Learning analyses reported in this study is publicly available at https://github.com/Giomu/ML_ADHVII.git. A snapshot of the repository at the time of publication has been deposited on Zenodo at https://doi.org/10.5281/zenodo.15496792[44].

Abbreviations

HR	High Responders
LR	Low Responders
sI	participants self-reported Infected
sNI	participants self-reported Non Infected
sI-HR	participants self-reported Infected in the High Responders group
sI-LR	participants self-reported Infected in the Low Responders group
sNI-HR	participants self-reported Non Infected in the High Responders group
sNI-LR	participants self-reported Non Infected in the Low Responders group
UI	self-reported Non Infected participants unaware of a previous infection
mcI	group of self-reported Infected used for Model Construction
mcNI	group of self-reported Non Infected used for Model Construction
I	participants self-reported Infected classified as Infected by the consensus strategy
NI	participants self-reported Non Infected classified as Non Infected by the consensus strategy

Received: 12 December 2024; Accepted: 20 June 2025;
Published online: 08 July 2025

References

1. Painter, M. M. et al. Rapid induction of antigen-specific CD4+ T cells is associated with coordinated humoral and cellular immunity to SARS-CoV-2 mRNA vaccination. *Immunity* **54**, 2133–2142.e3 (2021).
2. Turner, J. S. et al. SARS-CoV-2 mRNA vaccines induce persistent human germinal centre responses. *Nature* **596**, 109–113 (2021).
3. Baden, L. R. et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N. Engl. J. Med.* **384**, 403–416 (2021).
4. Wang, Z. et al. mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* **592**, 616–622 (2021).
5. Interim statement on hybrid immunity and increasing population seroprevalence rates. https://www.who.int/news/item/01-06-2022-interim-statement-on-hybrid-immunity-and-increasing-population-seroprevalence-rates#.
6. Lasrado, N. & Barouch, D. H. SARS-CoV-2 hybrid immunity: the best of both Worlds. *J. Infect. Dis.* **228**, 1311–1313 (2023).
7. The Lancet Infectious Diseases, null. Why hybrid immunity is so triggering. *Lancet Infect. Dis.* **22**, 1649 (2022).
8. Muik, A. et al. Neutralization of SARS-CoV-2 Omicron by BNT162b2 mRNA vaccine-elicited human sera. *Science* **375**, 678–680 (2022).
9. Van Elslande, J. et al. Longitudinal follow-up of IgG anti-nucleocapsid antibodies in SARS-CoV-2 infected patients up to eight months after infection. *J. Clin. Virol.* **136**, 104765 (2021).
10. Yu, W. et al. Proportion of asymptomatic infection and nonsevere disease caused by SARS-CoV-2 Omicron variant: A systematic review and analysis. *J. Med. Virol.* **94**, 5790–5801 (2022).
11. Bai, Z., Cao, Y., Liu, W. & Li, J. The SARS-CoV-2 nucleocapsid protein and its role in viral structure, biological functions, and a potential target for drug or vaccine mitigation. *Viruses* **13**, 1115 (2021).
12. Van Elslande, J. et al. Lower persistence of anti-nucleocapsid compared to anti-spike antibodies up to one year after SARS-CoV-2 infection. *Diagn. Microbiol. Infect. Dis.* **103**, 115659 (2022).
13. Haveri, A. et al. Persistence of neutralizing antibodies a year after SARS-CoV-2 infection in humans. *Eur. J. Immunol.* **51**, 3202–3213 (2021).
14. Gallais, F. et al. Evolution of antibody responses up to 13 months after SARS-CoV-2 infection and risk of reinfection. *EBioMedicine* **71**, 103561 (2021).
15. Dobaño, C. et al. Immunogenicity and crossreactivity of antibodies to the nucleocapsid protein of SARS-CoV-2: utility and limitations in

seroprevalence and immunity studies. *Transl. Res.* **232**, 60–74 (2021).
16. Tomic, A. et al. Divergent trajectories of antiviral memory after SARS-CoV-2 infection. *Nat. Commun.* **13**, 1251 (2022).
17. Marcinkevics, R. et al. Machine learning analysis of humoral and cellular responses to SARS-CoV-2 infection in young adults. *Front. Immunol.* **14**, 1158905 (2023).
18. Montesi, G. et al. Predicting humoral responses to primary and booster SARS-CoV-2 mRNA vaccination in people living with HIV: a machine learning approach. *J. Transl. Med.* **22**, 432 (2024).
19. Azarfar, G. et al. Using machine learning for personalized prediction of longitudinal coronavirus disease 2019 vaccine responses in transplant recipients. *Am. J. Transplant.* **25**, 1107–1116 (2025).
20. Konnova, A. et al. Predictive model for BNT162b2 vaccine response in cancer patients based on blood cytokines and growth factors. *Front. Immunol.* **13**, 1062136 (2022).
21. Ciabattini, A. et al. Evidence of SARS-CoV-2-Specific Memory B cells six months after vaccination with the BNT162b2 mRNA Vaccine. *Front. Immunol.* **12**, 740708 (2021).
22. Polvere, J. et al. B cell response after SARS-CoV-2 mRNA vaccination in people living with HIV. *Commun. Med.* **3**, 13 (2023).
23. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw*. **3**, 861 (2018).
24. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res*. **9**, 2579–2605 (2008).
25. Scrucca, L., Fraley, C., Murphy, T. B. & Raftery, A. E. *Model-Based Clustering, Classification, and Density Estimation Using Mclust in R* (CRC Press, 2023).
26. Hennig, C. fpc: Flexible Procedures for Clustering. R package version 2.2-13, https://CRAN.R-project.org/package=fpc (2024).
27. Kuhn, M. Building predictive models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).
28. Ciabattini, A. et al. Trajectory of spike-specific B cells elicited by two doses of BNT162b2 mRNA vaccine. *Cells* **12**, 1706 (2023).
29. Van Gassen, S. et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015).
30. Malek, M. et al. flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinforma. Oxf. Engl.* **31**, 606–607 (2015).
31. Suzuki, R. et al. Attenuated fusogenicity and pathogenicity of SARS-CoV-2 Omicron variant. *Nature* **603**, 700–705 (2022).
32. Pastore, G. et al. Homologous or heterologous administration of mRNA or adenovirus-vectored vaccines show comparable immunogenicity and effectiveness against the SARS-CoV-2 Omicron variant. *Expert Rev. Vaccines* **23**, 432–444 (2024).
33. Pajon, R. et al. SARS-CoV-2 omicron variant neutralization after mRNA-1273 booster vaccination. *N. Engl. J. Med.* **386**, 1088–1091 (2022).
34. Gruell, H. et al. mRNA booster immunization elicits potent neutralizing serum activity against the SARS-CoV-2 Omicron variant. *Nat. Med.* **28**, 477–480 (2022).
35. Shang, W. et al. Percentage of asymptomatic Infections among SARS-CoV-2 omicron variant-positive individuals: a systematic review and meta-analysis. *Vaccines* **10**, 1049 (2022).
36. Shi, N. D. J. et al. The asymptomatic proportion of SARS-CoV-2 omicron variant infections in households: a systematic review. *Influenza Other Respir. Viruses* **18**, e13348 (2024).
37. Fraussen, J. et al. Phenotypic and Ig repertoire analyses indicate a common origin of IgD-CD27- Double Negative B cells in healthy individuals and multiple sclerosis patients. *J. Immunol.* **203**, 1650–1664 (2019).
38. Wei, C. et al. A new population of cells lacking expression of CD27 represents a notable component of the B cell memory compartment in systemic lupus erythematosus. *J. Immunol.* **178**, 6624–6633 (2007).
39. Colonna-Romano, G. et al. A double-negative (IgD-CD27-) B cell population is increased in the peripheral blood of elderly people. *Mech. Ageing Dev.* **130**, 681–690 (2009).
40. Wu, Y.-C. B., Kipling, D. & Dunn-Walters, D. K. The relationship between CD27 negative and positive B cell populations in human peripheral blood. *Front. Immunol.* **2**, 81 (2011).
41. Beckers, L., Somers, V. & Fraussen, J. IgD−CD27− double negative (DN) B cells: origins and functions in health and disease. *Immunol. Lett.* **255**, 67–76 (2023).
42. Frasca, D., Diaz, A., Romero, M. & Blomberg, B. B. Human peripheral late/exhausted memory B cells express a senescent-associated secretory phenotype and preferentially utilize metabolic signaling pathways. *Exp. Gerontol.* **87**, 113–120 (2017).
43. Montesi, G. et al. Machine learning approaches to dissect hybrid and vaccine-induced immunity dataset. Zenodo, https://doi.org/10.5281/zenodo.15518709 (2025).
44. Giomu. Giomu/ML_ADHVII. Zenodo, https://doi.org/10.5281/zenodo.15496792 (2025).

Acknowledgements

This study was supported by the Department of Medical Biotechnologies of University of Siena (D.M.), by NextGenerationEU projects PNRR MUR Extended Partnership Initiative on Emerging Infectious Diseases project (PNRR PE13 INF_ACT—CUP B63C22001400007) and PNRR MUR M4 C2 Inv. 1.5, CUP Master B63C22000680007, CUP Project E93C24001570007 "Characterization of antigen-specific antibody and memory B-cell responses following vaccination". We would like to thank all the volunteers who participated to the study, the Infectious Disease Unit nursing staff who choose to cooperate for blood withdrawal.

Author contribution

Conceptualization: A.C., D.M., E.P., G.M., S.C. Volunteers enrollment: F.M., M.T., M.F., M.S.; Data Curation: S.C., G.P., F.F., J.P., E.P., M.S., F.M.; Immunological Analyses: S.C., E.P., J.P., F.F.; Computational Analyses: G.M., S.L.; Supervision of the study and Project Administration: A.C., E.P., D.M; Data Visualization: S.C., G.M, S.L.; Writing the Manuscript: S.C., G.M., A.C, E.P, D.M. All the authors edited and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-00987-4.

Correspondence and requests for materials should be addressed to Annalisa Ciabattini.

Peer review information *Communications Medicine* thanks Ghazal Azarfar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.