Health and Technology
https://doi.org/10.1007/s12553-024-00886-z

ORIGINAL PAPER

Development and validation of a symbolic regression-based machine learning method to predict COVID-19 in-hospital mortality among vaccinated patients

Filippos Sofos[1] · Erasmia Rouka[2] · Vasiliki Triantafyllia[3] · Evangelos Andreakos[3] · Konstantinos I. Gourgoulianis[4] · Efstathios Karakasidis[4] · Theodoros Karakasidis[1]

Received: 11 April 2024 / Accepted: 13 May 2024

Abstract

Purpose The continuous evolution of SARS-CoV-2 and possible future pandemics have risen concerns relevant to the effectiveness of the vaccines which are currently available. To this direction, new computational tools based on artificial intelligence (AI) and machine learning (ML) methods are incorporated, focusing on revealing hidden patterns and behaviors from, oftentimes, a great number of parameters that may affect (or not) the evolution of the pandemic.

Methods In this study, we developed and validated prediction models of COVID-19 in-hospital mortality among vaccinated patients by applying Symbolic Regression (SR)-based, ML algorithms. Considering the key role of cytokines and chemokines in the modulation of the immune response, we employed a dataset combining several of the aforementioned biomolecules with commonly used laboratory markers as well as demographic and clinical data.

Results Starting from a forty-four features dataset, we managed to restrict the total number of employed variables between 6–8 and ended up in four possible equations accurately predicting data behavior. The feature 'Days with symptoms from onset until admission' appeared in every equation, while interleukins (ILs)-17A and -6 in 3 out of 4 models. The parameters 'IL-6' and 'IL-17A', wherever combined led in a different survival effect on patients, compared to those cases where they solely appeared in an equation.

Conclusions Our method is presented for the first time and aims to be part of a broader computational and statistical framework that could aid in medical decision-making applications.

Keywords COVID-19 · Cytokines · In-hospital mortality · Machine learning · Symbolic regression · Vaccination

1 Introduction

The detection and isolation of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19 was reported on January, 2020 [1]. Globally, as of 2 August 2023, there have been 768.983.095 confirmed cases of COVID-19, including 6.953.743 deaths, reported to the World Health Organization (WHO) [2]. The emergence of SARS-CoV-2 Variants of Concern (VOCs) [3] and waning immunity after either vaccination or infection [4] have increased the risks of breakthrough infection and subsequent hospitalization especially for people with immunocompromising conditions and/or common comorbidities such as chronic kidney disease, chronic lung disease and diabetes [5]. Of note, a recent systematic review and meta-analysis highlighted the marked immune escape associated with Omicron VOCs and symptomatic disease [6]. Hence there is a challenge of predicting severe outcomes among vaccinated inpatients with breakthrough SARS-CoV-2 infection [7, 8]. Machine learning (ML) approaches can be valuable tools to this end.

Since the start of the COVID-19 pandemic, ML has been used for various applications including but not limited to

✉ Theodoros Karakasidis
thkarak@uth.gr

1 Department of Physics, Faculty of Science, University of Thessaly, Lamia, Greece

2 Department of Nursing, School of Health Sciences, University of Thessaly, Larissa, Greece

3 Biomedical Research Foundation of the Academy of Athens, Athens, Greece

4 Faculty of Medicine, School of Health Sciences, University of Thessaly, Larissa, Greece

Published online: 24 May 2024

COVID-19 detection from medical images [9] and wearables [10], outbreak predictions via wastewater surveillance [11], de novo drug design [12], assessment of vaccination hesitancy [13], identification of vaccination side effects predictors [14] and estimation of vaccines' effects on mortality [15, 16], while relevant approaches based on road networks have been also presented [17–19]. More than 20% of recent research articles in ML during the past five years refer to medical/biomedical applications. In the context of COVID-19 progression and prognosis in hospitalized patients, ML-based predictive models have shown good performance facilitating the identification of high-risk subjects thus informing proper clinical decision making [20]. More recently Baker et al. [21], included vaccination status-a variable which few prior ML studies were able to use-as a predictor of mortality among inpatients with COVID-19. Undoubtedly, all these applications have opened the road to dive deeper into COVID-19 characteristics and trends difficult to spot with usual statistical tools. However, towards transparent and trustworthy Artificial Intelligence (AI), especially in medical science applications, it is preferable for a model to be fully interpretable and understandable [22].

From a technical point of view, ML, as a subset of AI, has been mainly utilized for data science and statistical analysis tasks, through supervised and unsupervised approaches [23, 24]. It focuses on its ability to learn from data and forecast on unseen values of the implied dataset, both in classification and regression problems [25]. The ML platform is usually given in the form of a "black-box" stage, which accepts some input features and outputs one or more dependent variables. This "black-box" model is oftentimes hard to decode. It lacks in interpretability and this may pose barriers in the whole prediction procedure and decision-making, especially when survival is in question.

One of the proposed techniques exploited towards this direction is Symbolic Regression (SR) [26, 27]. SR is capable of generating analytical equations that resemble well-known theoretical and empirical mathematical equations, while driven only from data. Taking in mind genetic programming (GP) and evolutionary computing (EC) principles, the process of extracting an equation resembles the mutation and crossover mechanisms of parent/child evolution in a population [28]. Although no physical limitations are considered during evolution, the SR algorithm usually searches over an infinite pool of operators, independent variables and constants, to find the optimal expression for a given dataset, in the form of an analytical equation that combines the important input features to extract the dependent variable(s).

It is a fact that medical data from the field is oftentimes hard to obtain and leaves no chance of repeating a measurement, especially if one takes in mind that the initial period of the COVID-19 pandemic had made things even more difficult [29]. In the literature, there are cases where medical scientists had to deal with little and non-representative data from short periods to draw their results [30]. As for the number of data points that seem adequate to train an ML model, there is no clear answer. There are cases where successful ML models have been proposed, with datasets containing less than a hundred values [31, 32]. When the choice of input parameters is previously known, it is a common choice to employ transfer learning to pre-train a model with massive data and then include new data points that emerge from new measurements in a post-training phase [33]. Moreover, the current and future direction toward generative artificial intelligence would allow predictions with synthetic data [34]. Nonetheless, the complexity of the problem presented here would make it difficult to follow such approaches.

On August 9, 2023 WHO reported that the growth advantage and immune escape characteristics of the Omicron EG.5 variant, a descendent lineage of XBB.1.9.2 may increase the rate of new COVID-19 cases and become dominant in some countries or even on a global scale [35]. This continuous evolution of SARS-CoV-2 raises concerns relevant to the effectiveness of the vaccines which are currently available [36]. In this study, we aimed to develop and validate prediction models of COVID-19 in-hospital mortality among vaccinated patients by applying SR-based, ML algorithms. Considering the key role of cytokines and chemokines in the modulation of the immune response, we have employed a dataset combining several of the aforementioned biomolecules with commonly used laboratory markers as well as demographic and clinical data.

2 Materials and methods

2.1 Dataset description

We retrospectively assessed data retrieved from the CoVax study which involved adult vaccinated patients with breakthrough SARS-CoV-2 infection who had been admitted to the COVID-19 Department of the University Hospital of Larissa, Greece [7]. Forty-four parameters each mapped to a mathematical representation (x_1 to x_{44}) were evaluated. The primary outcome was mortality prediction during hospitalization (y) (Table 1). The effect of vaccine type (mRNA vaccine versus viral vector vaccine) on serum cytokines and chemokines was also examined.

2.2 Symbolic regression/classification

Based on the theory of evolution, various programming techniques have emerged, mimicking the biological processes, such as mutation and cross-over, and transforming them into explainable batches of code. The EC principle, as a superset,

Table 1 Parameters investigated for mortality prediction during hospitalization

Parameter	Type	Parameter	Type
Days Post Admission (DaysPAdsm) as to specimen collection[a]	int	Ferritin	float
Interferon Alpha (IFNa)	float	Creatine kinase (CPK)	int
Interleukin 29 (IL-29)	float	Days of Hospitalization (DaysHosp)	int
Interleukin 28A (IL-28A)	float	Fractalkine	float
Interferon-Inducible T-Cell Alpha Chemoattractant (ITAC)	float	Interferon Gamma (IFNg)	float
Days since second vaccine dose (Days2Dose)	int	Interleukin 10 (IL-10)	float
Days with symptoms from onset until admission (DaysSympt)	int	C-C Motif Chemokine Ligand 20 (MIP-3a)	float
Vaccine Type (Vac_Type)	bool	Interleukin 12 (IL-12p70)	float
Sex	bool	Interleukin 13 (IL-13)	float
Body mass index (BMI)	int	Interleukin 17A (IL-17A)	float
Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF)	float	Interleukin 1 Beta (IL-1b)	float
Age	int	Interleukin 2 (IL-2)	float
Real-Time Polymerase Chain Reaction Cycle Threshold (RT-PCR CTs)	float	Interleukin 21 (IL-21)	float
White Blood Cells (WBC)	int	Interleukin 4 (IL-4)	float
Lymphocytes (Lym)	int	Interleukin 23 (IL-23)	float
Platelets (PLTs)	int	Interleukin 5 (IL-5)	float
C-reactive protein (CRP)	float	Interleukin 6 (IL-6)	float
Creatinine	float	Interleukin 7 (IL-7)	float
Urine	float	Interleukin 8 (IL-8)	float
Aspartate transaminase (AST;SGOT)	float	Macrophage Inflammatory Protein 1-Alpha (MIP-1a)	float
Alanine transaminase (ALT;SGPT)	float	Macrophage Inflammatory Protein 1-Beta (MIP-1b)	float
Lactate dehydrogenase (LDH)	int	Tumor Necrosis Factor-Alpha (TNFa)	float
Death (y)	bool		

[a]Sample collection dates were within 3 days of inpatient admission for cytokines and chemokines measurements

includes the evolutionary algorithm (EA), the GP and the SR instances [37]. These are data-based approaches focusing on extracting mathematical equations without previous knowledge of the system through an evolutionary process.

To deal with regression-based applications, SR employs the features (i.e., the input variables) of a system under investigation and creates a set of symbolic expressions, of various accuracy and complexity, aiming to describe its behavior. The process is usually represented by a tree-structure with operator nodes, leaf nodes, and branches (Fig. 1a). Each operator node contains a mathematical operator, which applies on variables and constants coming from the leaf nodes [38]. The SR algorithm initially considers one or more parent tree structures and tries to find children structures which achieve minimum loss (usually, the mean squared error - MSE) through an iterative procedure that transforms the tree shape with mutation and crossover operations. The process terminates until minimum loss is achieved. The number of nodes and branches affects the corresponding equation complexity.

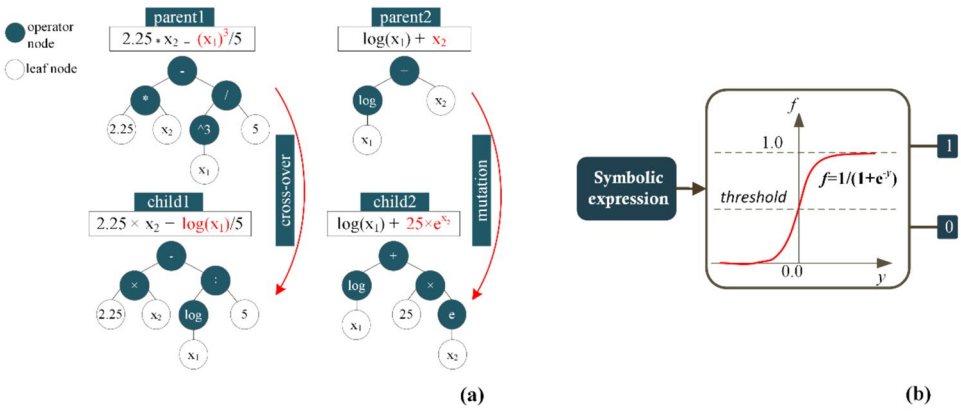

Fig. 1 The SR/SC procedures, **a** From the two, random initial parents, child$_1$ has been created with cross-over (change of a branch with a pre-existing branch) and child$_2$ with mutation (add a new branch), **b** mapping SR output to Boolean through a sigmoid function

Equations with low complexity and low MSE are selected as potential solutions to the problem. The Julia-based programming environment PySR is employed here to extract the equations [39].

The SR dataflow is usually incorporated in regression problems, where variables are integer and floating-point numbers. However, in classification problems where the output can be binary (e.g., '0' or '1'), the method must be adjusted. In this paper, we employ a symbolic classification scheme, where the SR equation enters a sigmoid function stage to transform to binary output, as shown in Fig. 1b. The output refers to the variable 'Death', with '0' (all values below the threshold line) representing the patient survival and '1' (all values above the threshold line) death.

2.3 Computational model flowchart

The computational model extracts a prediction of '1' or '0' on the output parameter 'Death', by taking into account the mathematical relations of the input features. The available data is pre-processed and statistically analyzed before entering the calculations. Next, it is randomly divided to a training and a validation set, with a percentage of 80% and 20%, respectively. Training data is used to train the SR algorithm and validation data is used for comparison at the end of the calculations. At this point, the predicted output is compared to the validated output, which is unknown to the algorithm, to check the overall accuracy.

The final expression contains the input features that affect the result, based on training data. It is worth noticing that the algorithm has no prior knowledge of the system and automatically weighs and selects the input features that seem to contribute to the extraction of the output variable. It starts with random construction of simple parent expressions and iteratively proceeds in forming child equations of various levels of complexity and error. The full suggested equation set is provided, and the final choice is made manually. The process is presented in Fig. 2.

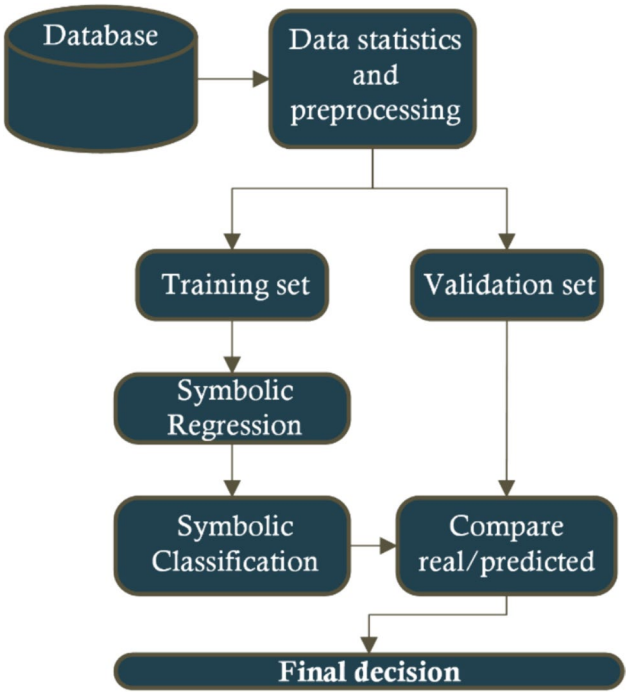

Fig. 2 Computational model flowchart

3 Results

3.1 Statistical analysis of important parameters

During pre-processing, in order to obtain better understanding of the problem, partial effects and correlations between all input parameters were calculated. The correlation diagram is shown in Fig. 3. Taking in mind the large number of input parameters and the various correlations that appear between them, it would be beneficial to apply dimensionality reduction techniques to increase interpretability and lighten the computational burden of the method. However, this is anticipated by the SR mechanism applied. The SR has performed feature selection as part of its optimization process by selecting the most relevant variables to include in the proposed expressions.

The effect of vaccine type on cytokines and chemokines levels is depicted in Fig. 4. Compared to cases who had received mRNA COVID-19 vaccine 9 (N = 54), cases who had been vaccinated with viral vector vaccine (N = 20) were found to have higher serum levels of ITAC ($p = 0.041$) and IL-10 ($p = 0.001$). No significant differences were observed for the remaining molecules.

3.2 Symbolic classification

As mentioned in Section 2.3, the choice of either a simple or a more complicated symbolic expression depends on the desired accuracy (in terms of MSE). Our computational process resulted in more than 500 possible expressions to choose. Since the most "complicated" equation is usually the most accurate, we had to weigh our decision on which equation to choose as to the specific understanding of the problem we wanted to solve. After careful investigation, we ended up in four possible expressions that achieved lower MSE and seemed to accurately follow and predict CoVax data behavior. They synergistically predict the output 'Death = 1'. Of note, the proposed equations originated without any prior hypothesis. They outputted logistic models (values '0' and '1') with a threshold of $th = 0.5556$ (Fig. 1b), mapping the output, y, as:

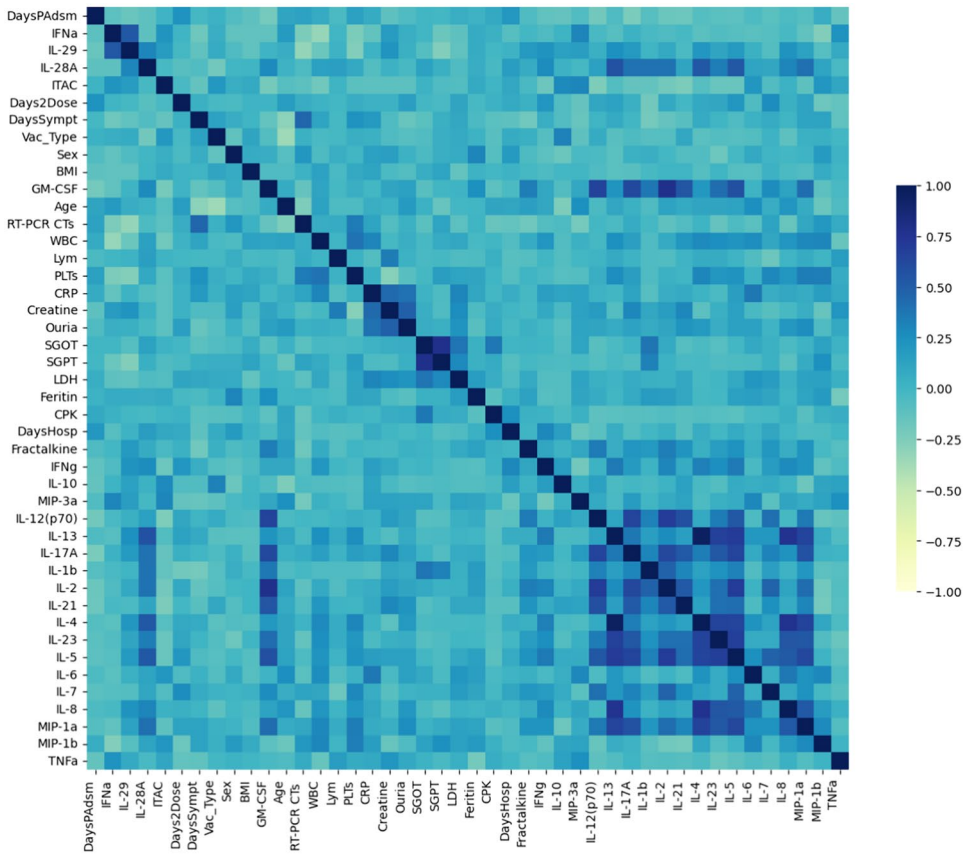

Fig. 3 Input features correlation matrix

$$f(y) = \begin{cases} 0, y < 0.5556 \\ 1, y \geq 0.5556 \end{cases} \quad (1)$$

In binary classification problems, the threshold that distinct an '1' from a '0' case is typically set to $th = 0.5$. However, this threshold is not ideal for imbalanced datasets. The optimal threshold for our binary classification problem is calculated in terms of the F1 score. The F1 score is a metric that combines both precision and recall, making it useful for imbalanced datasets [40].

Different input features were selected by the SC-derived equations. The accuracy of each equation is depicted below in terms of a confusion matrix and a ROC curve, which embeds the area under curve (AUC) metric. The confusion matrix employs the True Positive -TP (predicted:1 – actual:1), the True Negative -TN (predicted:0 – actual:0), the False Positive -FP (predicted:1 – actual:0) and the False Negative -FN (predicted:0 – actual:1) values. To compare between different models, the one with the largest AUC is the most accurate.

Next, the four predicted equations are presented. Each one employs different input features. The SC algorithm automatically selected those features from the available that affected the output at most.

3.2.1 Symbolic equations

The first equation, y_1, employed 8 input characteristics,

$$y_1 = \frac{x_{33}x_{38}}{x_{26} - 57.56} + e^{x_7\left(-x_{31}^2 - \frac{0.031}{x_{32}}\right)\left(-x_1 + x_{25} - 0.989\right)} \quad (2)$$

×1	×7	×25	×26	×31	×32	×33	×38
DaysPAdsm	DaysSympt	DaysHosp	Fractalkine	IL-13	IL-17A	IL-1b	IL-5

The second equation, y_2, employed 6 input characteristics, in an exponential form,

$$y_2 = 1.62e^{-\frac{1.752(x_7 + 0.083)\left(x_{10}x_8 + e^{0.878x_{32}} - 24.248\right)}{x_{34}x_{39}}} \quad (3)$$

×7	×8	×10	×32	×34	×39
DaysSympt	Vac_Type	BMI	IL-17A	IL-2	IL-6

The third equation, y_3, employed 6 input characteristics, common to y_2. However, it followed a different mathematical approach.

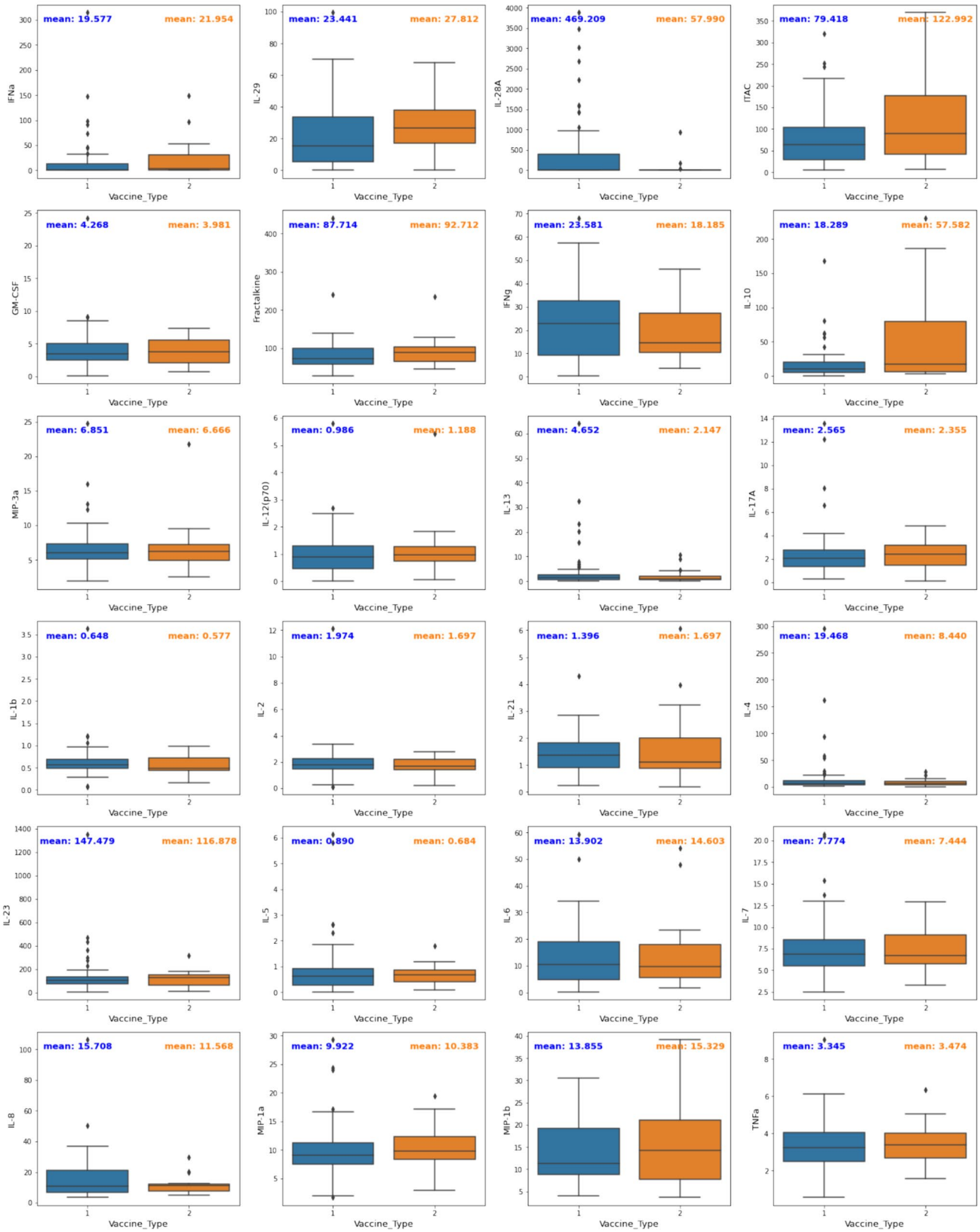

Fig. 4 Comparison of serum cytokines and chemokines levels in mRNA vaccine (Group 1) and viral vector (Group 2) vaccine breakthrough cases. Mean value for each group is shown. T-Test p-value was > 0.05 for all molecules except for ITAC ($p = 0.041$) and IL-10 ($p = 0.001$)

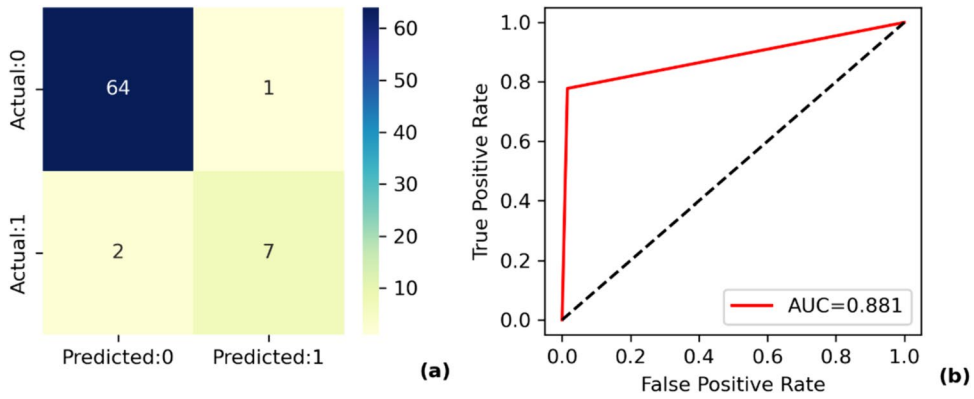

Fig. 5 Metrics for y_1, **a** Confusion matrix and **b** ROC curve

$$y_3 = \frac{0.202 x_{34} x_{39}}{x_8 \sqrt{x_{10} + x_{34} - x_{39}} \left(x_{32} x_7 + \log\left(x_{32}\right) - 1.007\right)} - 0.072518 \tag{4}$$

×7	×8	×10	×32	×34	×39
DaysSympt	Vac_Type	BMI	IL-17A	IL-2	IL-6

Finally, the fourth equation, y_4, employed 7 input characteristics, in an exponential form, as

$$y_4 = e^{\frac{x_{25} x_7 \left(x_{31} + \left(0.015 x_{12} - 1.61\right)\left(x_{40} + e^{\frac{x_{31}}{x_{38}}}\right) + 1.36\right)}{x_{39}}} \tag{5}$$

×7	×12	×25	×31	×38	×39	×40
DaysSympt	Age	DaysHosp	IL-13	IL-5	IL-6	IL-7

3.2.2 Accuracy metrics

Accuracy metrics for the classification output of each equation is presented in Figs. 5, 6, 7 and 8. The equation that achieved better metrics was y_1. The confusion matrix (Fig. 5a) revealed that the model was capable of spotting 7 out of 9 '1s' ('Death') from the dataset. Moreover, there was one FP prediction. The ROC curve (Fig. 5b) gave the AUC = 0.881, which was the higher obtained from this investigation.

In the first three equations, y_1, y_2, and y_3, the prediction gave two FN results (Figs. 5a, 6a and 7a), meaning that two patients were predicted to live but they, finally, died. The accuracy result was further deteriorated in y_4, where three FN results were obtained (Fig. 8a). Nevertheless, y_4, achieved an FP = 0, i.e., it spotted all 'Death' = 0 instances.

3.2.3 Symbolic equation comparison

The parameters defined as important and exploited to construct the mathematical equations by the SC algorithmic procedure were derived without any prior hypothesis during the calculations. Some of them appeared more than once, and this repetition may be further evidence of significance. Table 2 presents the number of times specific features appeared in the equations. The variable 'DaysSympt' appeared in all models, while 'IL-17A' and 'IL6' in 3 out of 4 models.

Diving deeper into the proposed equations, arguing on their partial effect on the respective models, it was observed that Equation y_1, the most accurate of four, written as

$$y_1 = \frac{(IL1\_b)(IL\_5)}{\text{Fractalkine} - 57.56} e^{-DaysSympt\left((\text{IL\_13})^2 + \frac{0.031}{IL\_17A}\right)(DaysHosp - DaysPAdsm - 0.989)} \tag{6}$$

translates as follows:

$$y_1 \uparrow when \begin{cases} IL1 \\ IL5 \\ IL17A \\ DaysPAdsm \end{cases} \uparrow$$

$$y_1 \downarrow when \begin{cases} \text{Fractalkine} \\ \text{DaysSympt} \\ \text{IL13} \\ \text{DaysHosp} \end{cases} \uparrow$$

Although equations y_2 and y_3 had different mathematical operators, they presented similar qualitative trends (increase or decrease behavior) when their input variables changed.

$$y_2 = 1.62 e^{-\frac{1.752(\text{DaysSympt} + 0.083)\left(BMI * Vac\_Type + e^{0.878 IL\_17A} - 24.248\right)}{IL_2 IL6}} \tag{7}$$

$$y_3 = \frac{0.202 * \text{IL}_2 * \text{IL}_6}{\text{VacType} \sqrt{\text{BMI} + \text{IL}_2 - \text{IL}_6} \left(\text{IL}_{17\text{A}} * \text{DaysSympt} + \log(\text{IL\_17A}) - 1.007\right)} - 0.072518 \tag{8}$$

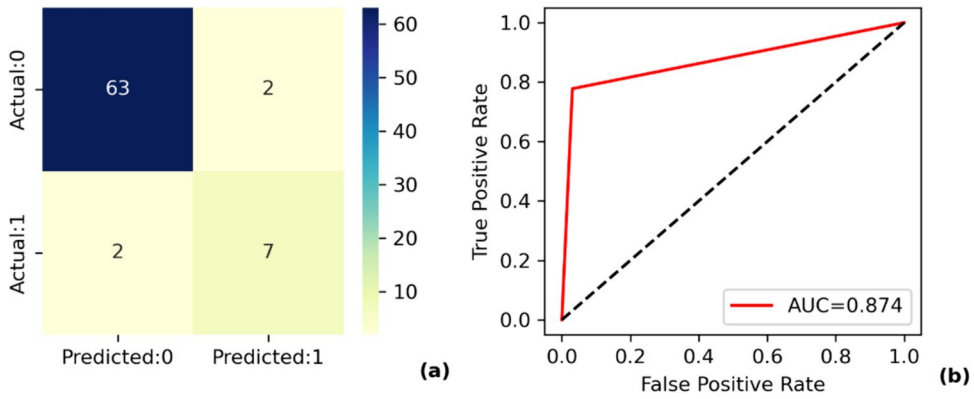

Fig. 6 Metrics for y_2, **a** Confusion matrix and **b** ROC curve

The increase behavior:

$$y_2, y_3 \uparrow when \begin{cases} IL6 \\ IL2 \end{cases} \uparrow$$

while,

$$y_2, y_3 \downarrow when \begin{cases} \text{BMI} \\ \text{DaysSympt} \end{cases} \uparrow$$

$$y_2, y_3 \downarrow when\ IL17A \uparrow$$

As far as equation y_4 is concerned (the one with the lowest AUC), it was written as:

$$y_4 = e^{\frac{DaysHosp * DaysSympt\left(IL\_13 + (0.015 * Age - 1.61)\left(IL\_7 + e^{\frac{IL\_13}{IL\_5}}\right) + 1.36\right)}{IL\_6}} \quad (9)$$

Here, most features increased the equation output as:

$$y_4 \uparrow when \begin{cases} IL13 \\ Age \\ DaysHosp \\ IL5 \\ IL7 \\ DaysSympt \end{cases} \uparrow$$

while a decreased output resulted only from:

$$y_4 \downarrow when\ IL6 \uparrow$$

This is the only model where an increase in 'DaysSympt' led to an increase in y_4.

Table 3 illustrates the patients' labels with outcome 'Death' = 1 discovered by each equation. None of the four proposed equations alone finds this outcome for all patients. To achieve maximum prediction ability, one could employ synergistically Eqs. 1 and 4 to ensure indices 55 and 54, and one of the Eqs. 2 or 3 that reveal all the remaining values.

All equations spotted the right outcome for labels 7, 22, 30, 42 and 53. The outcome for patient no. 54 was only predicted by y_4. We observed that equations y_1, y_2, and y_3 that included the 'IL17A' term failed to predict it. As far as 'IL6' is concerned, this might have had an impact on patient no. 55 prediction since, wherever 'IL6' was present (e.g., in y_2, y_3, and y_4), the prediction failed. Another observation is that label 47 was found only by y_2 and y_3, which combined levels of 'IL6' and 'IL17', while y_1 and y_4, where these two features did not apply concurrently, failed to predict it.

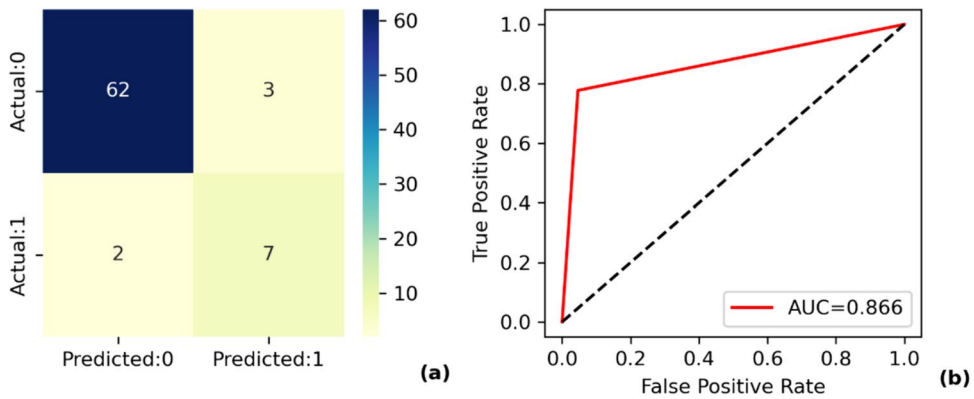

Fig. 7 Metrics for y_3, **a** Confusion matrix and **b** ROC curve

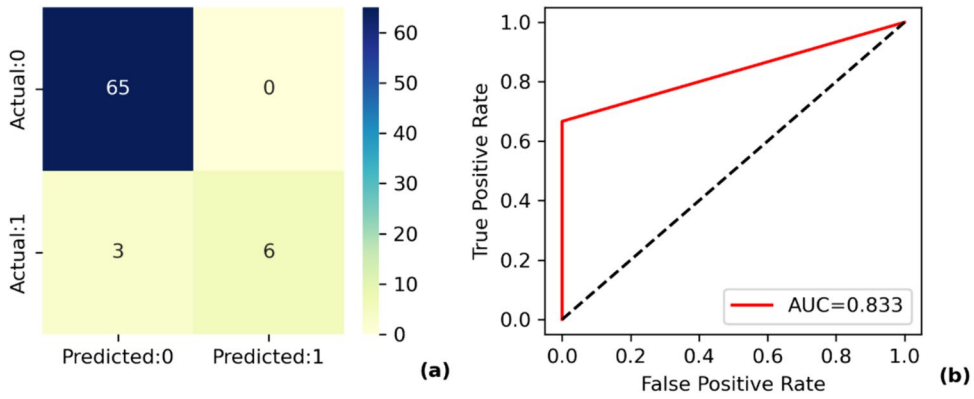

Fig. 8 Metrics for y_4, **a** Confusion matrix and **b** ROC curve

Table 2 Number of times the input features appeared in every equation

Input Property	Times of appearance/ total equations	Input Property	Times of appearance/ total equations
DaysSymp[a]	4/4	IL-13	2/4
IL-17A	3/4	IL-2	2/4
IL-6	3/4	IL-5	2/4
BMI	2/4	DaysHosp	2/4

[a]Days with symptoms from onset until admission

Table 3 True Positive values found by each proposed equation

'Death'=1	Patient index								
y_1	7	22	30	42	43	47	53	54	55
y_2	7	22	30	42	43	47	53	54	55
y_3	7	22	30	42	43	47	53	54	55
y_4	7	22	30	42	43	47	53	54	55

Numbers are the patient labels in the dataset

Colored cells denote successful prediction

4 Discussion

The continuous evolution of SARS-CoV-2 [36] but also of other RNA viruses such as influenza virus and the emerging tick-borne severe fever with thrombocytopenia syndrome virus imposes continuous global public health challenges which emphasizes the need for future pandemic preparation and response [41]. ML-based methods represent valuable tools towards this end [42]. In this study, we applied SR-based ML algorithms towards the development and validation of mortality prediction models for hospitalized patients with breakthrough COVID-19 infection. Starting from a forty-four features dataset, we managed to restrict the total number of employed variables between 6–8 and ended up in four possible equations accurately predicting CoVax data behavior.

The feature 'Days with symptoms from onset until admission' appeared in every equation, while 'IL-17A' and 'IL-6' in 3 out of 4 models. Of note, the parameters 'IL-6' and 'IL-17A', wherever combined (e.g., in y_2 and y_3), led in a different survival effect on patients, compared to those cases where they solely appeared in an equation (e.g., in y_1 and y_4). It was recently shown that increased IL-6 and IL-17A levels are associated with severe and long COVID-19, respectively [43]. Moreover, high serum IL-6 at the time of hospitalization has been associated with disease severity and patient survival [44]. Excessive levels of IL-6 promote generation of IL-17A and vice versa, resulting in amplified production of both cytokines. IL-6 and IL-17A can also result in viral persistence either independently or synergistically [45]. IL-13, -2, -5 appeared in two out of four proposed equations. These three interleukins as well as the IL-17 biomolecule have been reported as important factors in the determination of risk for mechanical ventilation and/or death in COVID-19 inpatients [46].

Since cytokines function in the regulation of innate and adaptive immunity, in this study we also explored the effect of vaccine type on serum cytokines and chemokines. Our analysis showed that serum levels of ITAC and IL-10 are significantly lower in hospitalized breakthrough cases who had received mRNA-based COVID-19 vaccines compared to those who had received viral-vector vaccines. IL-10 is considered a pleiotropic cytokine which may play a double role acting either as an anti-inflammatory molecule or as an immune stimulating factor in COVID-19, depending on the timing of its secretion [47]. As to ITAC (CXCL11), it

has been reported that it is significantly upregulated following SARS-CoV-2 infection [48] and that its expression in early-disease plasma samples may differentiate between patients developing critical versus non-critical disease [49]. More recently, a randomized controlled trial assessing the longitudinal association of COVID-19 vaccination with cytokine and chemokine concentrations among adult outpatients with symptomatic SARS-CoV-2 infection found that days since full vaccination and type of vaccine received are not correlated with cytokine and chemokine concentrations [50].

In regards to the interval between COVID-19 disease onset and admission, the available evidence has suggested that this is a variable having different prognostic values in different countries or regions with disparate health systems adopting different anti-epidemic strategies [51]. Similarly, the hospital length of stay for COVID-19 patients is dependent on a number of factors including but not limited to patient's age, accessibility to health services and availability of resources [52]. In the CoVax study, both the absence of anti-S SARS-CoV-2 antibodies and poor clinical outcomes of COVID-19 disease were associated with a shorter period between symptom onset and hospital admission which is in line with the outcomes generated by Eqs. 1–3 [7].

With respect to the 'BMI' variable which, in this study appeared in two out of four equations, a meta-analysis of 208 studies with 3 550 997 participants from over 32 countries found that the risk of COVID-19-related hospitalizations and death steadily increases with increasing levels of obesity noting however that the most recent studies show a weaker association between obesity and COVID-19 outcomes compared with the earlier ones [53]. More recently, it was reported that individuals with obesity show a reduction in the maintenance of humoral vaccine responses which has implications for vaccine prioritization policies [54]. Our models have predicted a negative association between BMI and mortality. However, it should be noted that BMI was included as a continuous variable in the proposed equations and was not subdivided into categories. The "obesity paradox" has been previously commented in the CoVax study [7]. Overall, the parameters that were defined in this study as important and thus exploited to construct the mathematical equations by the SC algorithmic procedure have been associated in the literature with COVID-19 clinical outcomes which enhances the validity of our method.

5 Conclusions

In this study we have provided fully interpretable analytical equations that capture mortality by selecting only the patient metrics regarded as most important (6–8 features only), significantly reducing the overload imposed by examining all features (forty-five features). Notwithstanding the fact that SR has been mainly employed in regression-based problems, our method applies successfully in classification problems, too, where the equation output is a binary decision point ('0' or '1'). To our knowledge, this method is presented for the first time and aims to be part of a broader computational and statistical framework that could aid in medical decision-making applications.

Author contributions Conceptualization, T.E.K. and E.R.; methodology, F.S., T.E.K.; software, F.S.; investigation, E.R., V.T., E.A., and E.K; data curation, E.R. and F.S.; writing—original draft preparation, E.R., V.T., E.A., E.K and F.S.; writing—review and editing, K.G. and T.E.K.; supervision, E.R., T.E.K., and K.G.; All authors have read and agreed to the published version of the manuscript.

Funding This research received no external funding.

Data availability The data that support the findings of this study are available on request from the corresponding author.

Declarations

Informed consent statement Informed consent was obtained from all subjects involved in the study.

Author agreement All authors have read and approved the final version of the manuscript.

Institutional review board statement The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the University Hospital of Larissa, Greece (46943/29.11.2021).

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382:727–33. https://doi.org/10.1056/NEJMoa2001017.
2. WHO Coronavirus (COVID-19) Dashboard. n.d. https://covid19.who.int . Accessed 13 Aug 2023.
3. Scovino AM, Dahab EC, Vieira GF, Freire-de-Lima L, Freire-de-Lima CG, Morrot A. SARS-CoV-2's variants of concern: a brief characterization. Front Immunol. 2022;13:834098. https://doi.org/10.3389/fimmu.2022.834098.
4. Pooley N, Abdool Karim SS, Combadière B, Ooi EE, Harris RC, El Guerche SC, et al. Durability of vaccine-induced and natural immunity against COVID-19: a narrative review. Infect Dis Ther. 2023;12:367–87. https://doi.org/10.1007/s40121-022-00753-2.
5. Smits PD, Gratzl S, Simonov M, Nachimuthu SK, Goodwin Cartwright BM, Wang MD, et al. Risk of COVID-19 breakthrough infection and hospitalization in individuals with comorbidities. Vaccine. 2023;41:2447–55. https://doi.org/10.1016/j.vaccine.2023.02.038.
6. Menegale F, Manica M, Zardini A, Guzzetta G, Marziano V, d'Andrea V, et al. Evaluation of waning of SARS-CoV-2 vaccine-induced immunity: a systematic review and meta-analysis. JAMA Netw Open. 2023;6:e2310650. https://doi.org/10.1001/jamanetworkopen.2023.10650.
7. Livanou E, Rouka E, Sinis S, Dimeas I, Pantazopoulos I, Papagiannis D, et al. Predictors of SARS-CoV-2 IgG spike

antibody responses on admission and clinical outcomes of COVID-19 disease in fully vaccinated inpatients: The CoVax study. World J Pers Med. 2022;12:640. https://doi.org/10.3390/jpm12040640.
8. Kandeel A, Fahim M, Deghedy O, Alim W, Fattah MA, Afifi S, et al. Clinical features and severe outcome predictors of COVID-19 vaccine breakthrough infection among hospitalized patients: results from Egypt severe acute respiratory infections sentinel surveillance, 2021–2022. BMC Infect Dis. 2023;23:130. https://doi.org/10.1186/s12879-023-08097-z.
9. Yang D, Martinez C, Visuña L, Khandhar H, Bhatt C, Carretero J. Detection and analysis of COVID-19 in medical images using deep learning techniques. Sci Rep. 2021;11:19638. https://doi.org/10.1038/s41598-021-99015-3.
10. Nestor B, Hunter J, Kainkaryam R, Drysdale E, Inglis JB, Shapiro A, et al. Machine learning COVID-19 detection from wearables. Lancet Digital Health. 2023;5:e182–4. https://doi.org/10.1016/S2589-7500(23)00045-6.
11. Ai Y, He F, Lancaster E, Lee J. Application of machine learning for multi-community COVID-19 outbreak predictions with wastewater surveillance. PLoS ONE. 2022;17:e0277154. https://doi.org/10.1371/journal.pone.0277154.
12. Floresta G, Zagni C, Gentile D, Patamia V, Rescifina A. Artificial intelligence technologies for COVID-19 de novo drug design. Int J Mol Sci. 2022;23:3261. https://doi.org/10.3390/ijms23063261.
13. Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Syst Appl. 2023;212:118715. https://doi.org/10.1016/j.eswa.2022.118715.
14. Abbaspour S, Robbins GK, Blumenthal KG, Hashimoto D, Hopcia K, Mukerji SS, et al. Identifying modifiable predictors of covid-19 vaccine side effects: a machine learning approach. Vaccines. 2022;10:1747. https://doi.org/10.3390/vaccines10101747.
15. Magazzino C, Mele M, Coccia M. A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality. Epidemiol Infect. 2022;150:e168. https://doi.org/10.1017/S0950268822001418.
16. Drikakis D, Sofos F. Can artificial intelligence accelerate fluid mechanics research? Fluids. 2023;8:212. https://doi.org/10.3390/fluids8070212.
17. Uddin S, Khan A, Lu H, Zhou F, Karim S. Suburban road networks to explore COVID-19 vulnerability and severity. Int J Environ Res Public Health. 2022;19:2039. https://doi.org/10.3390/ijerph19042039.
18. Uddin S, Lu H, Khan A, Karim S, Zhou F. Comparing the impact of road networks on COVID-19 severity between delta and omicron variants: a study based on greater Sydney (Australia) suburbs. Int J Environ Res Public Health. 2022;19:6551. https://doi.org/10.3390/ijerph19116551.
19. Uddin S, Khan A, Lu H, Zhou F, Karim S, Hajati F, et al. Road networks and socio-demographic factors to explore COVID-19 infection during its different waves. Sci Rep. 2024;14:1551. https://doi.org/10.1038/s41598-024-51610-w.
20. Moulaei K, Shanbehzadeh M, Mohammadi-Taghiabad Z, Kazemi-Arpanahi H. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med Inform Decis Mak. 2022;22:2. https://doi.org/10.1186/s12911-021-01742-0.
21. Baker TB, Loh W-Y, Piasecki TM, Bolt DM, Smith SS, Slutske WS, et al. A machine learning analysis of correlates of mortality among patients hospitalized with COVID-19. Sci Rep. 2023;13:4080. https://doi.org/10.1038/s41598-023-31251-1.
22. Bruckert S, Finzel B, Schmid U. The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell. 2020;3. https://doi.org/10.3389/frai.2020.507973.
23. Sofos F, Stavrogiannis C, Exarchou-Kouveli KK, Akabua D, Charilas G, Karakasidis TE. Current trends in fluid research in the era of artificial intelligence: a review. Fluids. 2022;7:116. https://doi.org/10.3390/fluids7030116.
24. Frank M, Drikakis D, Charissis V. Machine-learning methods for computational science and engineering. Computation. 2020;8:15. https://doi.org/10.3390/computation8010015.
25. Chowdhury MA, Hossain N, Ahmed Shuvho MB, Fotouhi M, Islam MS, Ali MR, et al. Recent machine learning guided material research - A review. Comput Conden Matter. 2021;29. https://doi.org/10.1016/j.cocom.2021.e00597.
26. Derner E, Kubalík J, Ancona N, Babuška R. Constructing parsimonious analytic models for dynamic systems via symbolic regression. Appl Soft Comput. 2020;94:106432. https://doi.org/10.1016/j.asoc.2020.106432.
27. Casadei F, Pappa GL. Multi-region symbolic regression: combining functions under a multi-objective approach. Nat Comput. 2021;20:753–73. https://doi.org/10.1007/s11047-021-09851-5.
28. Sofos F, Charakopoulos A, Papastamatiou K, Karakasidis TE. A combined clustering/symbolic regression framework for fluid property prediction. Phys Fluids. 2022;34:062004. https://doi.org/10.1063/5.0096669.
29. Oliveira D, Miranda R, Leuschner P, Abreu N, Santos MF, Abelha A, et al. OpenEHR modeling: improving clinical records during the COVID-19 pandemic. Health Technol. 2021;11:1109–18. https://doi.org/10.1007/s12553-021-00556-4.
30. Kamalov F, Thabtah F. Forecasting Covid-19: SARMA-ARCH approach. Health Technol. 2021;11:1139–48. https://doi.org/10.1007/s12553-021-00587-x.
31. Elton DC, Boukouvalas Z, Butrico MS, Fuge MD, Chung PW. Applying machine learning techniques to predict the properties of energetic materials. Sci Rep. 2018;8:9059. https://doi.org/10.1038/s41598-018-27344-x.
32. Sofos F, Karakasidis TE. Machine learning techniques for fluid flows at the nanoscale. Fluids 2021;6:96. https://doi.org/10.3390/fluids6030096.
33. Chang R, Wang Y-X, Ertekin E. Towards overcoming data scarcity in materials science: unifying models and datasets with a mixture of experts framework. Npj Comput Mater. 2022;8:1–9. https://doi.org/10.1038/s41524-022-00929-x.
34. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5:493–7. https://doi.org/10.1038/s41551-021-00751-8.
35. COVID-19: WHO tracking EG.5 variant of interest | UN News 2023. https://news.un.org/en/story/2023/08/1139617 . Accessed 13 Aug 2023.
36. Pagani I, Ghezzi S, Alberti S, Poli G, Vicenzi E. Origin and evolution of SARS-CoV-2. Eur Phys J Plus. 2023;138:157. https://doi.org/10.1140/epjp/s13360-023-03719-6.
37. Angelis D, Sofos F, Karakasidis TE. Artificial intelligence in physical sciences: symbolic regression trends and perspectives. Arch Computat Methods Eng. 2023;30:3845–65. https://doi.org/10.1007/s11831-023-09922-z.
38. Wang Y, Wagner N, Rondinelli JM. Symbolic regression in materials science. MRS Commun. 2019;9:793–805. https://doi.org/10.1557/mrc.2019.85.
39. Cranmer M, Sanchez Gonzalez A, Battaglia P, Xu R, Cranmer K, Spergel D, et al. Discovering symbolic models from deep learning with inductive biases. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors, et al. Advances in neural information processing systems, vol. 33. Curran Associates: Inc; 2020. p. 17429–42.
40. Zou Q, Xie S, Lin Z, Wu M, Ju Y. Finding the best classification threshold in imbalanced classification. Big Data Res. 2016;5:2–8. https://doi.org/10.1016/j.bdr.2015.12.001.

41. Choi YK. Emerging and re-emerging fatal viral diseases. Exp Mol Med. 2021;53:711–2. https://doi.org/10.1038/s12276-021-00608-9.
42. Syrowatka A, Kuznetsova M, Alsubai A, Beckman AL, Bain PA, Craig KJT, et al. Leveraging artificial intelligence for pandemic preparedness and response: a scoping review to identify key use cases. Npj Digit Med. 2021;4:1–14. https://doi.org/10.1038/s41746-021-00459-8.
43. Queiroz MAF, Neves PFM das, Lima SS, Lopes J da C, Torres MK da S, Vallinoto IMVC, et al. Cytokine profiles associated with acute COVID-19 and long COVID-19 syndrome. Front Cell Infect Microbiol. 2022;12:922422. https://doi.org/10.3389/fcimb.2022.922422.
44. Del Valle DM, Kim-Schulze S, Huang H-H, Beckmann ND, Nirenberg S, Wang B, et al. An inflammatory cytokine signature predicts COVID-19 severity and survival. Nat Med. 2020;26:1636–43. https://doi.org/10.1038/s41591-020-1051-9.
45. Hou W, Jin Y-H, Kang HS, Kim BS. Interleukin-6 (IL-6) and IL-17 synergistically promote viral persistence by inhibiting cellular apoptosis and cytotoxic T cell function. J Virol. 2014;88:8479–89. https://doi.org/10.1128/jvi.00724-14.
46. Gibellini L, De Biasi S, Meschiari M, Gozzi L, Paolini A, Borella R, et al. Plasma cytokine atlas reveals the importance of TH2 polarization and interferons in predicting COVID-19 severity and survival. Front Immunol. 2022;13:842150. https://doi.org/10.3389/fimmu.2022.842150.
47. Carlini V, Noonan DM, Abdalalem E, Goletti D, Sansone C, Calabrone L, et al. The multifaceted nature of IL-10: regulation, role in immunological homeostasis and its relevance to cancer, COVID-19 and post-COVID conditions. Front Immunol. 2023;14:1161067. https://doi.org/10.3389/fimmu.2023.1161067.
48. Callahan V, Hawks S, Crawford MA, Lehman CW, Morrison HA, Ivester HM, et al. The pro-inflammatory chemokines CXCL9, CXCL10 and CXCL11 are upregulated following SARS-CoV-2 infection in an AKT-dependent manner. Viruses. 2021;13:1062. https://doi.org/10.3390/v13061062.
49. Zhang Y, Xu C, Higuita NIA, Bhattacharya R, Chakrabarty JH, Mukherjee P. Evaluation of I-TAC as a potential early plasma marker to differentiate between critical and non-critical COVID-19. Cell Stress. 2021;6:6–16. https://doi.org/10.15698/cst2022.01.262
50. Zhu X, Gebo KA, Abraham AG, Habtehyimer F, Patel EU, Laeyendecker O, et al. Dynamics of inflammatory responses after SARS-CoV-2 infection by vaccination status in the USA: a prospective cohort study. Lancet Microbe. 2023;4:e692-703. https://doi.org/10.1016/S2666-5247(23)00171-4.
51. Peng L, Lv Q-Q, Yang F, Wu X-M, Zhang C-C, Wang Y-Q, et al. The interval between onset and admission predicts disease progression in COVID-19 patients. Ann Transl Med. 2021;9:213. https://doi.org/10.21037/atm-20-5320
52. Alimohamadi Y, Yekta EM, Sepandi M, Sharafoddin M, Arshadi M, Hesari E. Hospital length of stay for COVID-19 patients: A systematic review and meta-analysis. Multidiscip Respir Med. 2022;17:856. https://doi.org/10.4081/mrm.2022.856.
53. Sawadogo W, Tsegaye M, Gizaw A, Adera T. Overweight and obesity as risk factors for COVID-19-associated hospitalisations and death: systematic review and meta-analysis. BMJ Nutr Prev Health. 2022;5:10–18. https://doi.org/10.1136/bmjnph-2021-000375.
54. van der Klaauw AA, Horner EC, Pereyra-Gerber P, Agrawal U, Foster WS, Spencer S, et al. Accelerated waning of the humoral response to COVID-19 vaccines in obesity. Nat Med. 2023;29:1146–54. https://doi.org/10.1038/s41591-023-02343-2.