Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(6): 5813–5831.
DOI: 10.3934/mbe.2022272
Received: 27 January 2022
Revised: 14 March 2022
Accepted: 28 March 2022
Published: 06 April 2022

Research article

A machine learning approach to differentiate between COVID-19 and influenza infection using synthetic infection and immune response data

Suzan Farhang-Sardroodi[1,2,*], **Mohammad Sajjad Ghaemi**[3], **Morgan Craig**[4], **Hsu Kiang Ooi**[3] **and Jane M Heffernan**[1,2,*]

[1] Modelling Infection and Immunity Lab, Mathematics Statistics, York University, Toronto, Canada
[2] Centre for Disease Modelling (CDM), Mathematics Statistics, York University, Toronto, Canada
[3] Digital Technologies Research Centre, National Research Council Canada, Toronto, ON, Canada
[4] Sainte-Justine University Hospital Research Centre and Department of Mathematics and Statistics, Université de Montréal, Montreal, Quebec, Canada

* **Correspondence:** Email: suzanfa@yorku.ca; jmheffer@yorku.ca.

Abstract: Data analysis is widely used to generate new insights into human disease mechanisms and provide better treatment methods. In this work, we used the mechanistic models of viral infection to generate synthetic data of influenza and COVID-19 patients. We then developed and validated a supervised machine learning model that can distinguish between the two infections. Influenza and COVID-19 are contagious respiratory illnesses that are caused by different pathogenic viruses but appeared with similar initial presentations. While having the same primary signs COVID-19 can produce more severe symptoms, illnesses, and higher mortality. The predictive model performance was externally evaluated by the ROC AUC metric (area under the receiver operating characteristic curve) on 100 virtual patients from each cohort and was able to achieve at least AUC=91% using our multiclass classifier. The current investigation highlighted the ability of machine learning models to accurately identify two different diseases based on major components of viral infection and immune response. The model predicted a dominant role for viral load and productively infected cells through the feature selection process.

Keywords: biological systems; mechanistic model; infectious disease; Influenza (flu); COVID-19; machine learning; classification; Logistic regression; regularization; Lasso; Ridge; PLS-DA

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and influenza viruses cause COVID-19 and influenza diseases, respectively, and mainly infect the upper and lower respiratory tract [1, 2]. Both infections present similar primary symptoms such as cough, fever, sore throat, runny or stuffy nose, tiredness, and body aches [3,4]. Early on in infection this can lead to a clinical dilemma in diagnosis [5–7]. Recently, COVID-19 has, through its worse overall decompensation due to its intensive transmission and vascular effects, caused an unrivaled global crisis [8–11]. As the globe moves to endemicity, as the striking COVID-19 outbreak continues, the concurrence of COVID-19 and influenza epidemics is impending. This motivates the current study, to design a data analysis tool that can accurately differentiate between these two infections.

One way to rapidly classify patients with influenza or COVID-19 could be through machine learning approaches. Preliminary investigation illustrated the potentials of machine-learning models for accurately distinguishing between these two viral infections, using demographics, body mass index, and vital signs in infected patients [8]. Herein, we used a simple ML-based classification to identify patients with influenza or SARS-CoV-2 based on the main features of the within-host viral dynamics and the immune response. During the past decade, in-host mathematical modelling has become an increasingly powerful tool to study inter and intracellular viral infection and the ensuing immune response [12]. Such mathematical models can deepen our understanding of virus spread within organs leading to antiviral drug inventions and optimized treatment regimens. Furthermore, using the mechanistic model to generate synthetic patient data for various infections can help us to mitigate difficulties related to clinical data analyses, such as time-inconsistent data sets that can cause biased results [13].

Recent studies have suggested that that artificial intelligence (AI) and machine learning (ML) methods can perform as well as or even better than humans at significant healthcare tasks, such as diagnosing disease [14–17]. We apply a basic mathematical model on the cellular scale (the so-called target cell-limited model [18, 19]) fit to two different sets of *in vivo* data for COVID-19 and influenza infections, to create virtual patient cohorts. Using our multi-class classifier, the patients are differentiated between the two infections. This is conducted over the entire infection and for early time-points only. Herein, we show that, with just some important in-host measurements, our method is able to discern which virus has infected a patient with a high degree of certainty. Such results can lead to the development of rapid diagnostic tests in future to aid in early patient diagnosis They can also be used in clinical trials of new therapeutics and vaccines to determine the need for new participant enrolments, the number of measurements needed from each participant, and what would be best to measure to show if a new vaccine or therapeutic is effective [20–22].

This paper is organized as follows: In section 2, through subsection 2.1, we discuss the in-host mathematical modelling of influenza and COVID-19 and parameter estimation. In subsection 2.2, we use the mechanistic model to generate synthetic patient data. In subsections 2.3 we study developing and evaluating a supervised machine learning method to discriminate the patients with different infections. The Interpretability of the developed model is discussed in subsection 2.4. The results of the prediction are presented in section 3 through subsection 3.1. Subsection 3.2 discusses the importance of the data features and determines the dominant features. The paper concludes with a discussion in Section 4.

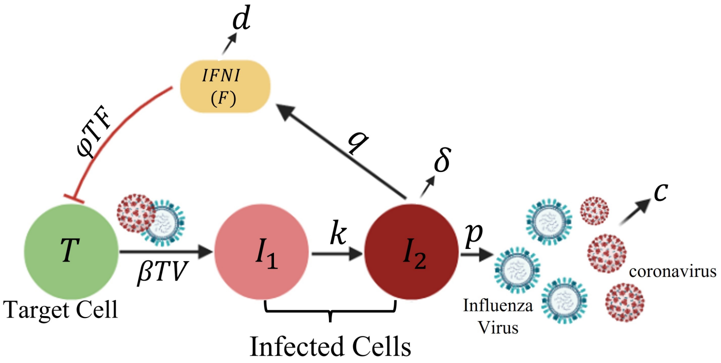

Figure 1. Schematic of viral infection. Each Target cell, T, is infected by a virus, V, with a constant rate β. During the eclipse period the productively infected cell, I_2, is being produced by the first infected cell, I_1, with a constant rate k. The Infected cell, I_2, produces virus at rate p, IFNI at rate q and dies at rate δ per cell. IFNI hinders viral infection by converting target cells to a virus-resistant state with a constant rate ϕ and decays with rate d. Free virus particles that can be influenza or coronaviruses are cleared at per-capita rate c.

2. Method

2.1. Mechanistic models

We employed a target-cell limited model of viral dynamics using five differential equations that track susceptible target cells (T), infected cells in the eclipse phase (I_1), productively infected cells (I_2), virus (V), and interferon (F) in-host. Figure 1 presents a flow diagram of the model. The system of ordinary differential equations is as follows:

$$\frac{dT}{dt} = -\beta T V - \phi T F \tag{2.1a}$$

$$\frac{dI_1}{dt} = \beta T V - k I_1 \tag{2.1b}$$

$$\frac{dI_2}{dt} = k I_1 - \delta I_2 \tag{2.1c}$$

$$\frac{dV}{dt} = p I_2 - c V \tag{2.1d}$$

$$\frac{dF}{dt} = q I_2 - d F \tag{2.1e}$$

Briefly, virus particles V can infect susceptible target cells T to produce infected cells. This is represented by the term $\beta T V$. Newly infected cells first enter the eclipse phase I_1 and become productively infected cells I_2 when within-cell processes that program the cell to make new virus particles are completed. The eclipse phase takes, on average, $1/k$ time units. Productively infected cells produce new virus particles with a rate of p, and the virus particles are cleared from the system with a rate of c. We assumed that productively infected target cells have a death rate δ. Susceptible target cells can be protected from infection by Type I interferon (IFNI), F. Type I interferons protect neighboring cells from infection and elicit an immune response [23, 24]. They are central to combating different virus infections and are regularly measured in clinical trials or infection studies in humans and animals [25]. We assumed that interferon production is proportional to the number of productively infected cells [18, 19, 26, 27], that interferon has a natural decay rate d, and that interferon protects susceptible cells by removing them from the susceptible target cells population, with a rate ϕF. This term was ignored in [18] for influenza infection. The model described by Eq. 2.1 was used in [18] and [19] to examine the kinetics of influenza A and SARS-CoV-2 viral dynamics, respectively. For the sake of simplicity, we have ignored a half-day lag in IFNI response that was considered in [18].

2.1.1. Parameter estimation

Model parameters for influenza A infection were fit to data from an experimental H1N1 influenza A/Hong Kong/123/77 infection for six patients [18] and for SARS-CoV-2 from thirteen untreated patients infected with severe acute respiratory syndrome-coronavirus [19]. The geometric average parameter values along with their 95% confidence intervals and units are summarized in Table 1. We assumed that the initial number of target cells, T_0, is equal to the total number of target cells in the upper respiratory tract and set $T_0 = 4 \times 10^8$ cells. In [19] the authors considered that the target cells distributed in a volume of 30 mL. Assuming that 1% of these cells expresses the angiotensin-converting enzyme 2 (ACE2) as a receptor for SARS-CoV-2, the target cell concentration, T_0, was expressed as 1.33×10^5 cell/ml. Model variables with initial values were estimated as in Table 2.

Table 1. Average values and confidence intervals, CI, for influenza A and SARS-CoV-2 within-host viral infection model parameters. Confidence levels of 95% display the degree of certainty around the mean for each parameter value.

Influenza Model Parameters [18]									
V_0 [95%CI] $TCID_{50}/ml^1$	R_0	β [95%CI] $(TCID_{50}/ml)^{-1}d^{-1}$	k [95%CI] d^{-1}	p [95%CI] $(TCID_{50}/ml)d^{-1}$	c [95%CI] d^{-1}	δ [95%CI] d^{-1}	q d^{-1}	d d^{-1}	ϕ $d^{-1}cell^{-1}$
$0.075[7.6E-4, 7.5]$ SD:3.5724	21.5[10.1-46.1] SD:17.15	$3.2E-5[6E-6, 1.7E-4]$ SD:7.8124	$4[3, 5.2]$ SD:1.0486	$0.046[0.012, 0.17]$ SD:0.07527	$5.2[3.1-8.7]$ SD:2.6677	$5.2[3.2-8.6]$ SD:2.5724	1	1.9 [23, 28, 29]	0
COVID-19 Model Parameters [19]									
V_0 $Copies/ml$	R_0 95%[CI]	β $(Copies/ml)^{-1}d^{-1}$	k d^{-1}	p 95%[CI] $(Copies/ml)d^{-1}$	c d^{-1}	δ 95%[CI] d^{-1}	q d^{-1}	d d^{-1}	ϕ $d^{-1}cell^{-1}$
0.1	$8.6[1.9-17.6]$ SD:12.9893	$5.68E-9$	3	$22.71[0-59.64]$ SD:49.3426	10	$0.6[0.22-0.97]$ SD:0.62051	1	0.4	1.97E-6 [30]

[1] 1 [$TCID_{50}/ml$] corresponds to 4000 [$Copies/ml$] [31].
[2] R_0 is the basic reproduction number.

Table 2. Model Variables with Initial values.

Variable	Definition	Initial Value	Unit
T	Target cell	4E+8	Cell
I_1	Infected cell (eclipse phase)	0	Cell
I_2	Productively infected cell	0	Cell
V	Viral load (flu)	7.5E-2	$TCID_{50}/ml$
	Viral load(COVID-19)	0.1	$Copies/ml$
F	type I interferon (IFNI)	0	Interferon

2.2. Generation of virtual patients

To generate a cohort of virtual patients, we followed a technique similar to the one used in [24]. Each patient is distinguished by five different in-host measurements, $\{T, I_1, I_2, V, F\}$, that are the solutions of Equation 2.1 for different sets of model parameters. Initial parameter sets representing individual virtual patients were drawn from normal distributions with means fixed to the corresponding parameter value in Table 1 and standard deviations derived from confidence interval measurements. Standard deviations were obtained from standard errors, confidence intervals, and t statistics which measure the size of the difference relative to the variation in the sample data. For each parameter value, the standard deviation was obtained by dividing the length of the confidence interval by standard errors width ($2 \times t - value$) and then multiplying by the square root of the sample size as follows

$$SD = \sqrt{N} \times SE = \sqrt{N} \times (upperlimit - lowerlimit)/(2 \times t - value) \tag{2.2}$$

Standard errors must be of means calculated from within each parameter confidence interval. The $t - value$ for a 95% confidence interval from a sample size of N was then obtained in Microsoft Excel using the $tinv$ function (i.e. $tinv(1 - 0.95, N - 1)$). From [18], the sample size for the influenza cohort is 6 patients infected by H1N1 influenza A/Hong Kong/123/77 infection. The COVID-19 cohort consisted of 13 untreated patients infected with severe acute Respiratory syndrome-coronavirus2 [19]. Therefore, the $t - value$ for influenza patients is 2.571 and for COVID-10 patients is about 2.179. From normal distributions with standard deviations, σ, and means, μ, as the original parameter values, we then generated normal distributions covering values lying around each parameter value such that $|\mu \pm \sigma - \mu| < h$. Herein, the parameter h is the user-defined value as a measure of data diversity. In the other words, the bigger the parameter h, the more diverse the synthetic data. Accordingly, the external noise can affect the data through the parameter h. The dynamics of 100 virtual patients from each cohort are shown in Figure 2. The diversity of patient data is mainly reflected in various viral load levels to agree with prior studies that different viral load is associated with the severity of diseases or different factors such as age or sex of the patients [32].

2.2.1. Consistency of the data

Generating data with time consistency for different cohorts of infections is of great importance. Data inconsistency can lead to loss of information or biased results. Since the influenza mechanistic model predicts faster clearance of influenza-infected cells than SARS-CoV-2 [19], the infection period for influenza and COVID-19 patient dynamics are not the same, see Figure 2. Therefore we limited the consistency of flu/COVID-19 cohorts to have the same number of data points during the infection time. Hereupon, as an example, we divided the main infection period (i.e., $[1 - 6]$ days for influenza patients and $[10 - 20]$ days for COVID-19 patients) into ten different sub-intervals with half-day length time steps for influenza patients and one and half-day length time steps for COVID-19 patients (see Figure 3). Hence, despite having different infection periods and time steps with different lengths to report the new virtual data point, the total number of data for the two different cohorts was the same.

In addition to the total infection period, we were also interested in studying the viral load dynamics in the early period of infection. The median incubation period for influenza A(B) virus is estimated to be 1.4(0.6) days, and for SARS-CoV-2 is around $5 - 6$ days [33]. Therefore, we assumed the time interval $[0.9, 1.3]$ days for influenza, and $[5 - 6.5]$ days for COVID-19 cohorts, corresponding to $[10^2 - 10^4] Copies/ml$ viral load. Dividing each interval into three different sub-intervals to get the time steps with length one-sixth of a day for Influenza and half a day for COVID-19 patients, we had four consistent data points for each patient.

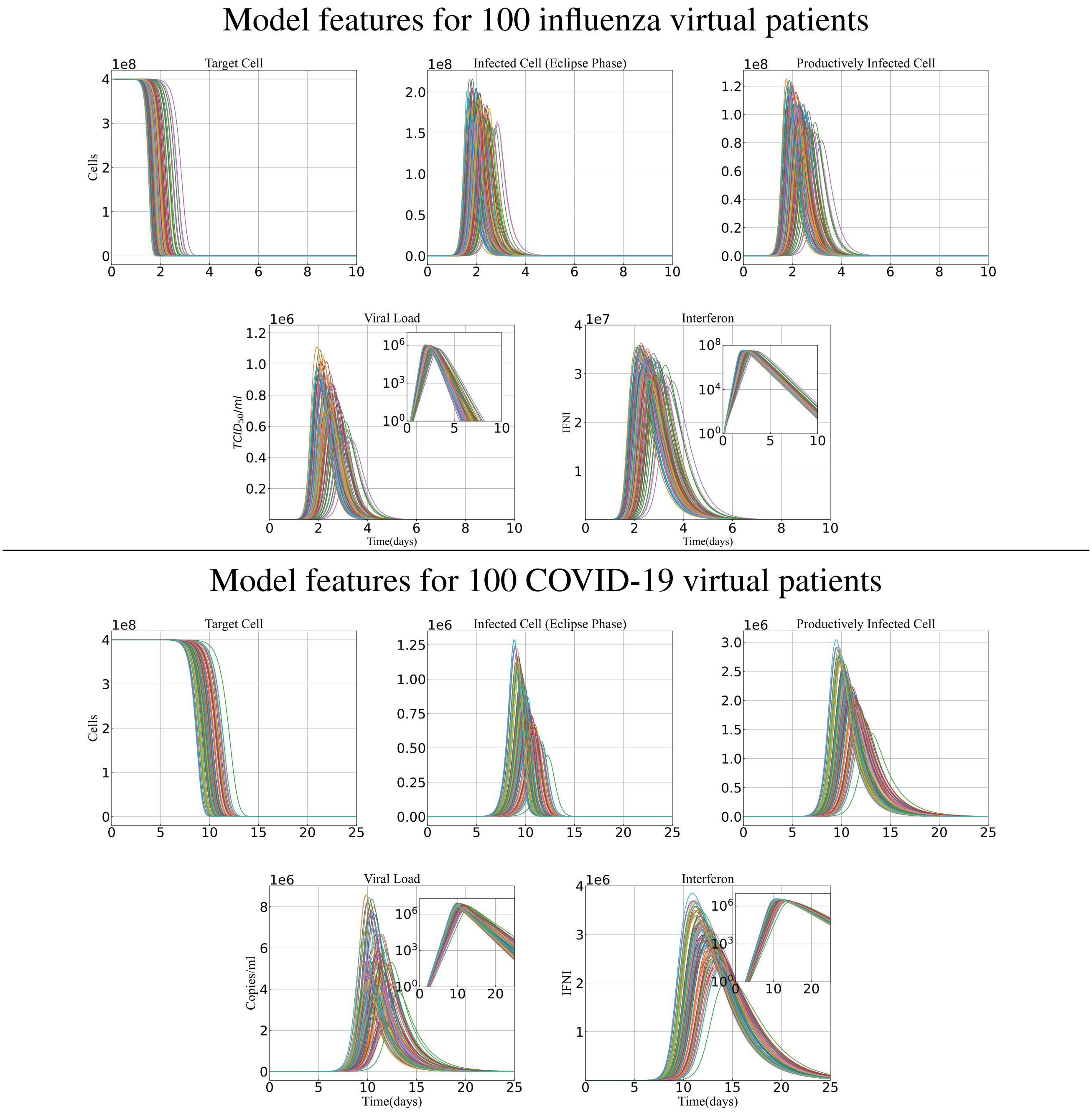

Figure 2. Cohort Dynamics. One hundred virtual patients are generated with different features of Target cells, infected/productively infected cells, viral load, and the only immune factor type I interferon for Influenza (upper two rows) and COVID-19 (lower two rows). Each solid curve with a different color represents a patient. The insets are in log scale.

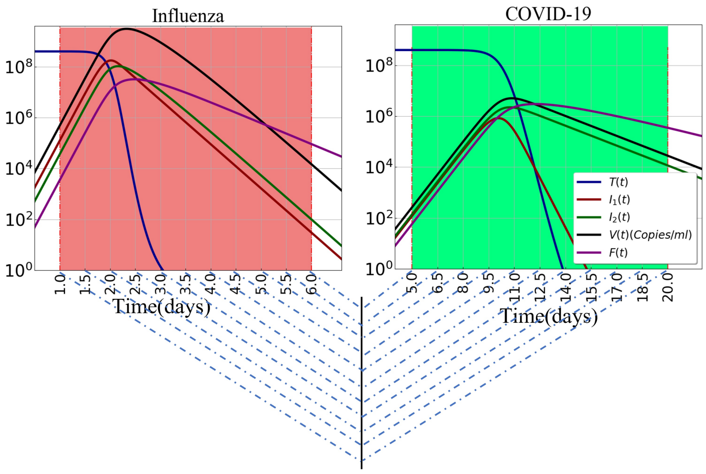

Figure 3. Consistency of the number of virtual data points during the time of infection. Dashed cross blue lines show eleven-time points of an influenza or COVID-19 patient.

2.3. Predictive model development

To distinguish between patients who encounter COVID-19 from those who are exposed to influenza, we developed a predictive model based on some biological feature selections. Accordingly, we adopted Logistic regression with ℓ_1-regularization, referred to Lasso (stands for least absolute shrinkage and selection operator) Regression, as an appropriate technical classification. Lasso regression is widely used for many supervised classification problems based on the concept of probability [34]. It can simplify the model complexity by removing irrelevant features of the data set. Recently, this algorithm was used by Han and et al. to find some additional novel immune features that accurately identified patients before the clinical diagnosis of preeclampsia [35].

Logistic regression, which is a special case of linear regression and used for binary classification, is defined by the following sigmoid function

$$h(X) = \frac{1}{1 + e^{-(\beta_0 + \beta X)}} \tag{2.3}$$

in which X is the $(n \times p)$ model feature matrix of $n = 100$ patients and $p = 5$ biological hallmarks. Defining the cost/objective (C) function of logistic regression in mean squared error format leads to a non-convexity that makes it difficult to optimally converge. Therefore, it is represented by the following equations

$$C(h(X), Y) = \begin{cases} -\log(h(X)), & \text{if } y = 1 \\ -\log(1 - h(X)), & \text{if } y = 0 \end{cases} \tag{2.4}$$

where Y is a binary response vector of outcome (CVOID-19 vs flu). Compressing the above two equations inside a single function, we have

$$J(X) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(h(x_i) + (1 - y_i) \log(1 - h(x_i)) \right] \tag{2.5}$$

Replacing the sigmoid function from equation (2.3) and applying a penalty term equal to the absolute value of the magnitude of coefficients, we can reach the following objective function (after doing some mathematical simplifications) [35]

$$J(X) = -\left[\frac{1}{n}\sum_{i=1}^{n} y_i(\beta_0 + \beta^T x_i) - \log\left(1 + e^{\beta_0 + \beta^T x_i}\right)\right] + \alpha\|\beta\|, \quad \alpha > 0 \tag{2.6}$$

The penalty term which is called the ℓ_1-regularization term is added to prevent data over-fitting. The model objective is to find a specific solution with a best-optimized cost function.

For model training and testing, we developed a $\mathcal{K}$-fold cross-validation strategy, which is a resampling method to evaluate machine learning models on a limited data sample. The procedure has a single parameter called $\mathcal{K}$ which displays the number of groups that a given data sample is to be split into. As such, the procedure is often called $\mathcal{K}$-fold cross-validation. Therefore, our regression model is not tailored to a particular data set and is exposed to all available samples of a given subject in the training set. This approach implies that the training procedure was entirely blinded to the synthetic patient data sets, and ensures the presumed independence from any intra-subject correlations that are required for Lasso classification. We fixed the number of folds of the data as $\mathcal{K} = 5$. Running the analysis on each fold, the predicted outcome will be the one with the least estimated prediction error. The regularization parameter α is estimated by a cross-validation procedure.

2.3.1. Evaluating model performance

The discriminating ability of the developed model in predicting patients with influenza from COVID-19 was evaluated using AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve analysis. AUC - ROC curve is one of the most important evaluation metrics to visualize the performance of multi-class classification problems. ROC represents a probability curve of sensitivity (true positive rate=$\frac{TP}{TP+FN}$) against 1-specificity (false positive rate=$\frac{FP}{FP+TN}$) and AUC is a performance measure of discrimination. In the other words, the AUC score is a criterion that explains how well the model is capable of discerning different cohorts. Generally, an AUC closer to 1 indicates a better overall diagnostic performance of influenza classes as influenza or COVID-19 to COVID-19.

2.4. Model interpretability

From [36, 37], "Interpretability" is the degree to which a human can understand the cause of a decision and consistently predict the model's result. The higher the interpretability of a machine learning model, the better understanding of why certain predictions have been made. Interpretable machine learning models are beneficial to extract the relevant knowledge from relationships either contained in data or learned by the model [38, 39].

Here, we looked at the regularization path which is a plot of all coefficients values against the values of α in-ℓ_1 penalization term, to see the behavior of the Lasso regression and interpret the prediction outcomes. The main purpose of Lasso regression is to classify groups of data by providing feature coefficients that can select the important features and maintain model regularization to avoid overfitting the data. Therefore, the Lasso path can give us an idea of the feature's importance.

3. Results

3.1. Prediction of Influenza versus COVID-19 infection

In this study we developed a classifier in the Lasso framework to identify patients with either influenza or COVID-19, based on four major entities of viral dynamics, $\{T(t), I_1(t), I_2(t), V(t)\}$, and one main factor of host immune response, type I interferon ($F(t)$), as the entry data features. The model was trained on data from 100 virtual patient-level data sets in each infection cohort without noise, and it was externally validated on testing sets with demographic noise (reflected in diverse viral load levels). Results in Figures 4, 5 and 6 reflect the Lasso predictions using the entire infection period (see Section 2.2.1). In Figures 4 and 5, two-dimensional scatter plots are used to compare ground truth to regression predicted values based on all model features. The hue spectrum from light to dark illustrates the probability of being in the influenza (blue) or COVID-19 (red) group. In the other words, the darker the colors, the better the prediction. Considering three attributes in the data, the predicted outcomes are improved. This is shown in three-dimensional scatter plots in Figure 6 of the ground truth and regression predicted values. ROC AUC=95% indicates a satisfactory performance of the model to distinguish between COVID-19 and influenza patients. We note that our analysis was also completed on data from 1000 virtual patients, and a similar result was obtained, ROC AUC = 93%. See Figure 7 for more details.

3.1.1. Early days of infection

We examined the model prediction for the data generated at the early days of infection after the incubation period. The results are shown in Figure 8 based on the model features. From the figure, we can see that there are some mispredictions, for small values of $I_1(t), I_2(t), V(t)$, and $F(t)$, especially when $I_2(t)$ is plotted as a function of $I_1(t)$ or $V(t)$ is plotted in terms of $I_2(t)$. In the other words, for this range of values, the influenza patients were misdiagnosed with COVID-19. In an attempt to find the reason, we compared correlations between the different variables in our model. See Figure 9. Here, we see small regions of overlap between influenza and COVID-19 models. Accordingly, the compatibility of the results between the two infections may lead to some overlaps in the model predictions. However, the ability of the model in the prediction of infections when the patients were monitored by $V(t)/F(t)$ as a function of $I_1(t)$, panels (b) and (c), or $F(t)$ in terms of $I_2(t)/V(t)$, panels (e) and (f), can be satisfactory, and thus can serve as benchmarks for clinical diagnosis. The model had a ROC AUC of 91% on the external validation data set for early infection – see Figure 7.

3.2. Significance of the features

To investigate the importance of various data features we created our ℓ_1-regularization path, which was the best way to see the behavior of the Lasso regression. The regularization path is a plot of all coefficient values in terms of the regularization parameter. Figure 10 illustrates the selection path of each feature with its corresponding coefficient in terms of the logarithm of the regularization parameter α. For each value of α, the path method on the Lasso object returns the coefficients that solve the logistic regression problem with that parameter value. The optimal value of $-\log(\alpha)$ was estimated at around 3.25 for the test set distributed over the entire infection course, and 3.04 when the early days of infection were studied. The results suggested a higher coefficient value for viral load $V(t)$ and

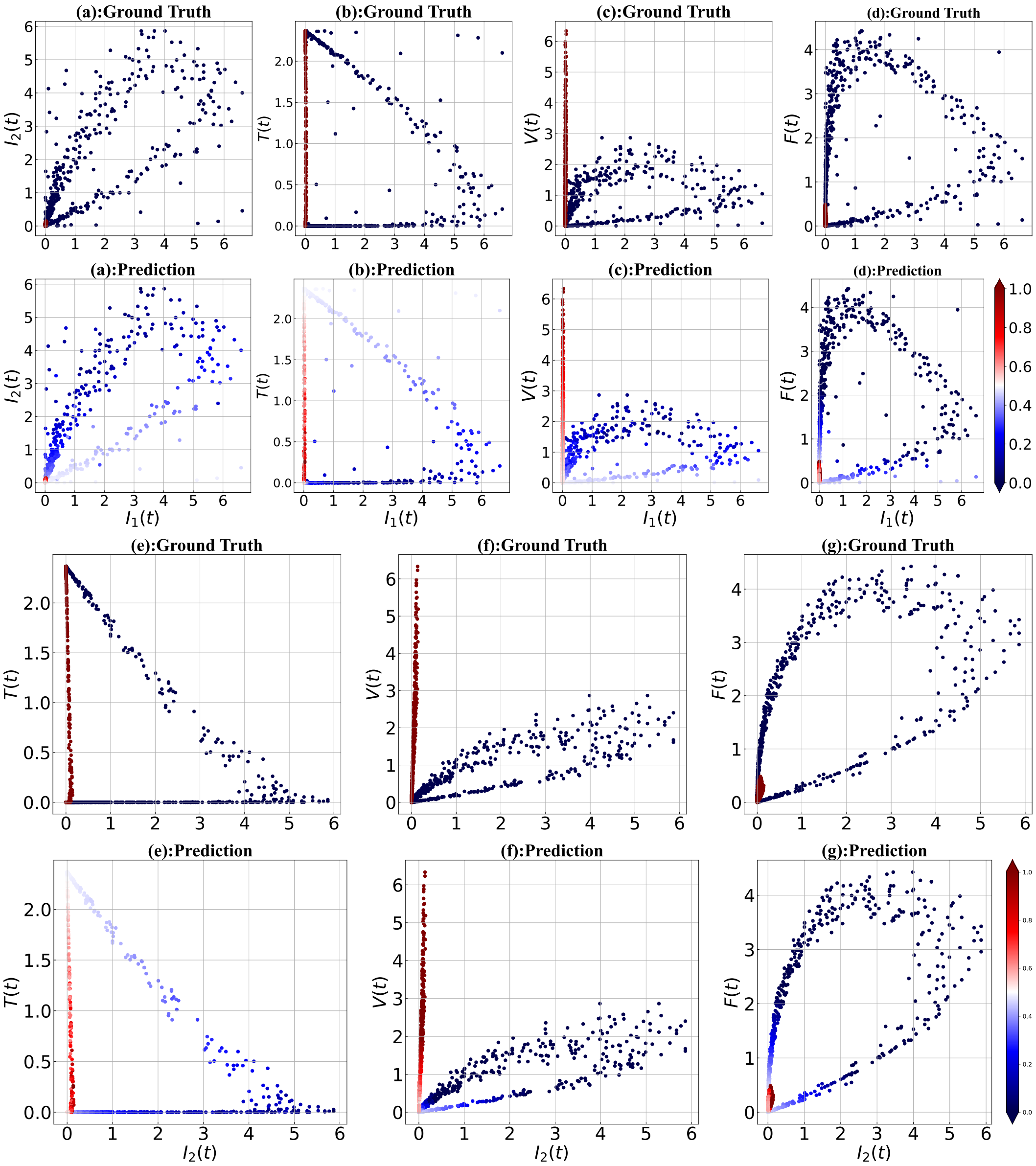

Figure 4. Two-dimensional scatter plots of ground truth and regression predicted values based on model features. Classification of the data was done for: I_2 versus I_1 in panels (a), T vs. I_1 in panels (b), V vs. I_1 in panels (c), F vs. I_1 in panels (d), T vs. I_2 in panels (e), V versus I_2 in panels (f), and F versus I_2 in panel (g). Color denotes the patient probability of being in the influenza (blue color scheme) or COVID-19 (red color scheme) cohorts. Data points, corresponding to each model feature, are rescaled by dividing by their standard deviations.

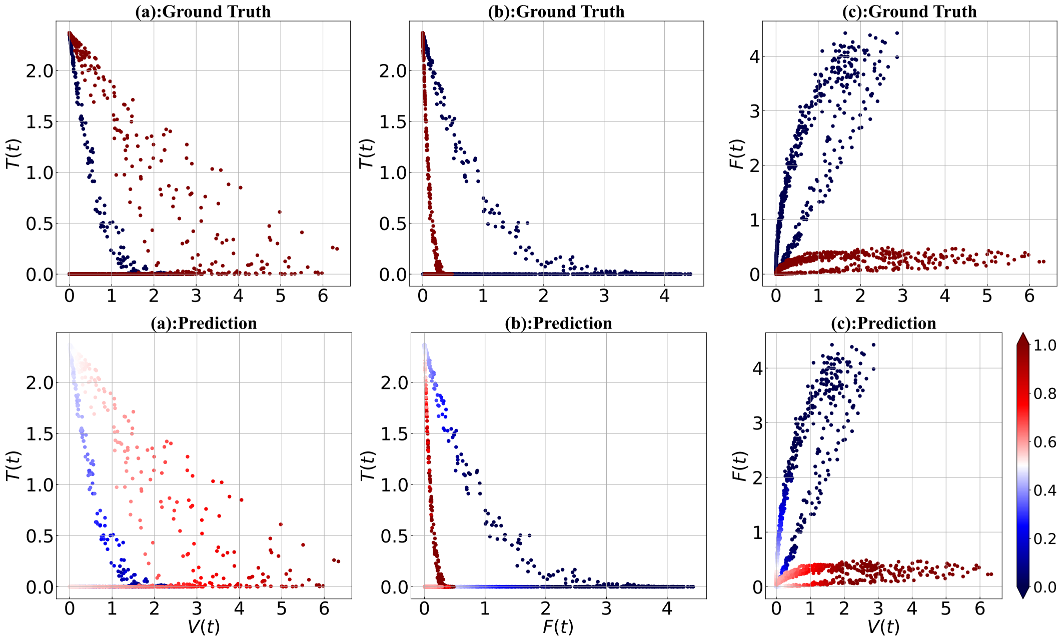

Figure 5. Two-dimensional scatter plots of the ground truth and regression predicted values for three model features T, V, F. Classification of the data was done based on: T versus V in panels (a), T versus F in panels (b), and F versus V in panels (c). Color denotes the patient probability of being in the influenza (blue color scheme) or COVID-19 (Red color scheme) cohorts.

productively infected cells $I_2(t)$ compared to the other features. In the same analysis on 1000 virtual patient data sets, the viral load had the predominant identifying role.

4. Discussion

This study presents a machine learning model to effectively classify influenza and COVID-19 virtual patients using in-host patient data. Our model employed a Lasso regression classifier trained to identify between two hundred patients, highlighted by a ROC AUC of 95%. Using within-host model structures from the literature, we generated synthetic data with five in-host measurements including target cells, eclipse phase, and productively infected cells, viral load, and type I IFN. Analyzing the feature importance revealed that the viral load and the productively infected cells are the most important components to determine if a patient is infected by influenza or SARS-CoV-2.

While our machine learning model was built on synthetic data distributed during the main infection period, it ascertained a good performance (ROC AUC = 91%) even for the early days of infection after the incubation period. However, in early infection, there were some exceptions for the small values of in-host features where the influenza patients were misdiagnosed as COVID-19. The reason was explained by the fact that during the early days of infection, influenza and COVID-19 patients have comparable in-host measurements that lead to some errors in discriminating the patients. This is interpreted as a limitation of our model even though the ROC AUC was still very high. A future extension of our work here will be in developing dynamic models which take more immune entities into account and end in a better classifier.

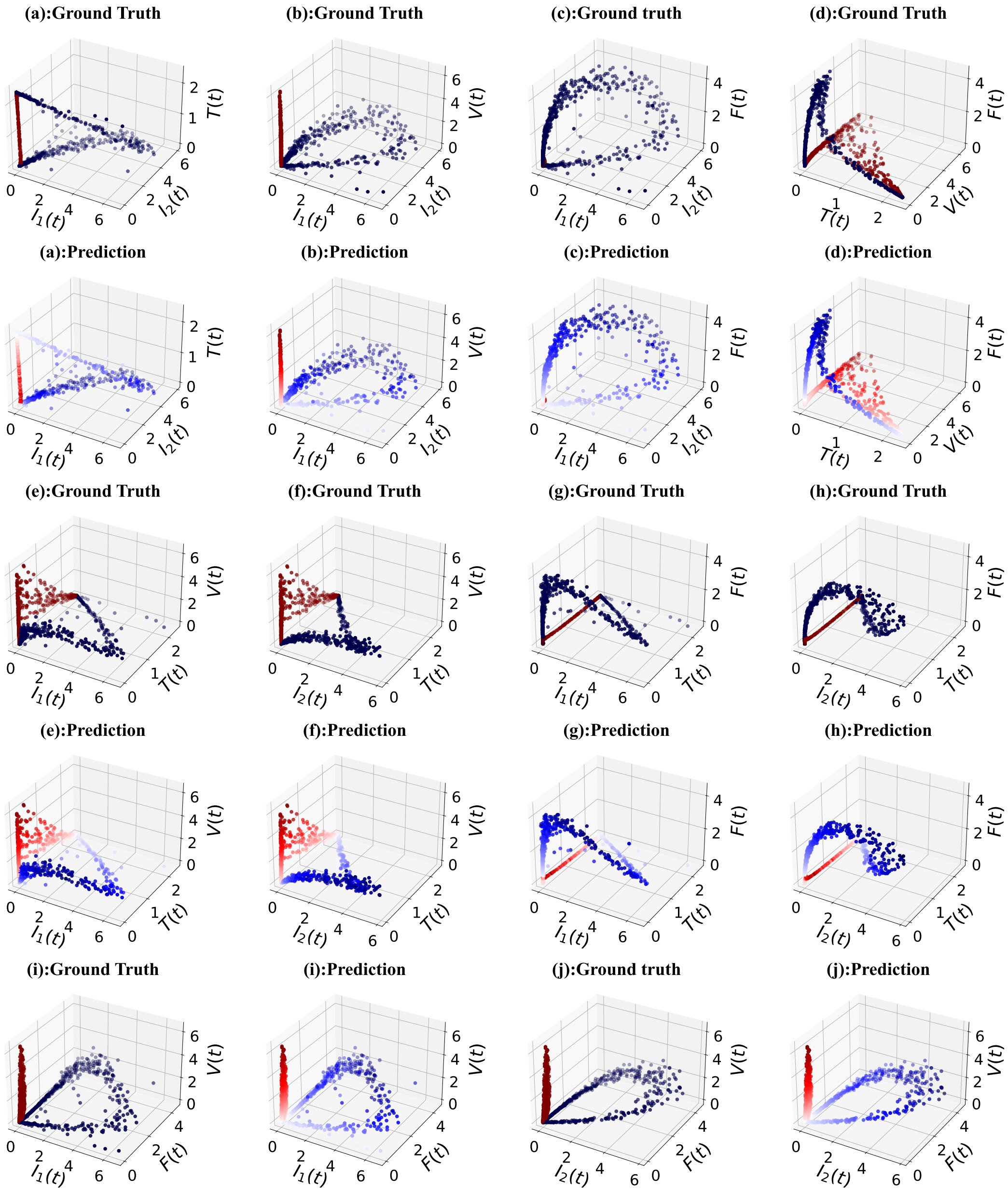

Figure 6. Three-dimensional scatter plots of the ground truth and regression predicted values based on all model features. Classification is based on I_1, I_2, T in panels (a), I_1, I_2, V in panels (b), I_1, I_2, F in panels (c), T, V, F in panels (d), I_1, T, V in panels (e), I_2, T, V in panels (f), I_1, T, F in panels (g), I_2, T, F in panels (h), I_1, F, V in panels (i), and I_2, F, V in panels (j). Shades of blue (red) indicate influenza (COVID-19) group patients. Data points are dimensionless by dividing by the corresponding standard deviations.

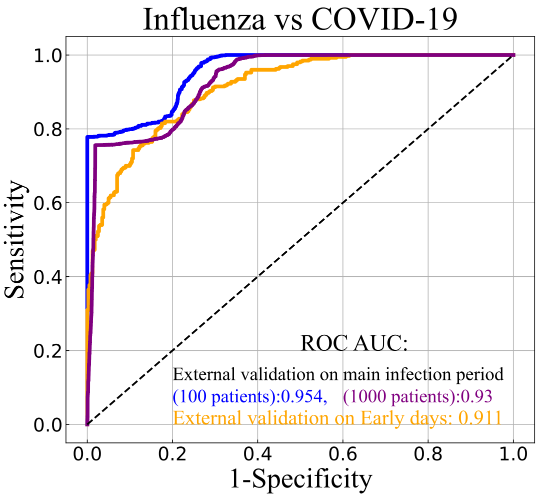

Figure 7. Receiver Operating Characteristic curve (ROC) of influenza vs COVID-19 patients. The area under the ROC curve indicates the predictive performance of the model between COVID-19 and influenza encounters on the external validation test for 100 patients from each cohort during the main (blue curve)/ early days (orange curve) of infection period, and for 1000 patients during the main infection (purple curve). The black dashed line in the diagonal has a ROC AUC of 0.5.

Our model was trained and successfully evaluated on synthetic data. The model, however, could be applied to animal or human clinical data. This could be useful, for example, if a clinical trial is complicated by the existence of an infectious disease with similar infection characteristics. The model could be applied as a low-cost classification system that would not require expensive virus typing procedures and could rely solely on viral load and interferon measurements. Additionally, the AI/ML method could be applied to determine when a clinical trial using a continuous enrollment design has accumulated sufficient data to determine whether a new pharmaceutical is effective. We note that studies like [8] that focus analysis on demographic and observational data can be cheaper to conduct than a study requiring viral load or immune system measurements, but these data can also be subject to inconsistencies and bias, affecting classification outcomes. In a future study, we will expand our analysis to a model of in-host measurements and observational data to determine if specific combinations of in-host and observational data that best classify influenza and COVID-19 infections differ.

Fourteen different AI/ML techniques in disease predictions were reviewed in [40]. Quiroz-Juárez and et al. developed an effective machine-learning algorithm for the identification of high-risk COVID-19 patients [41]. Some AI approaches that have significant contributions in the fields of health care were presented in [42] and their applications in confronting COVID-19, such as diagnosis and drug development were studied. Salehi and et al. studied machine and deep learning-based architectures performance for classification of coronavirus images such as X-ray and computed tomography [20]. Our machine learning model was developed in the Lasso framework. Ridge regression or Partial least squares discriminant analysis (PLS-DA) also can be employed, and require only small changes to our method to include this. The model demonstrated a satisfactory performance by using either Ridge or

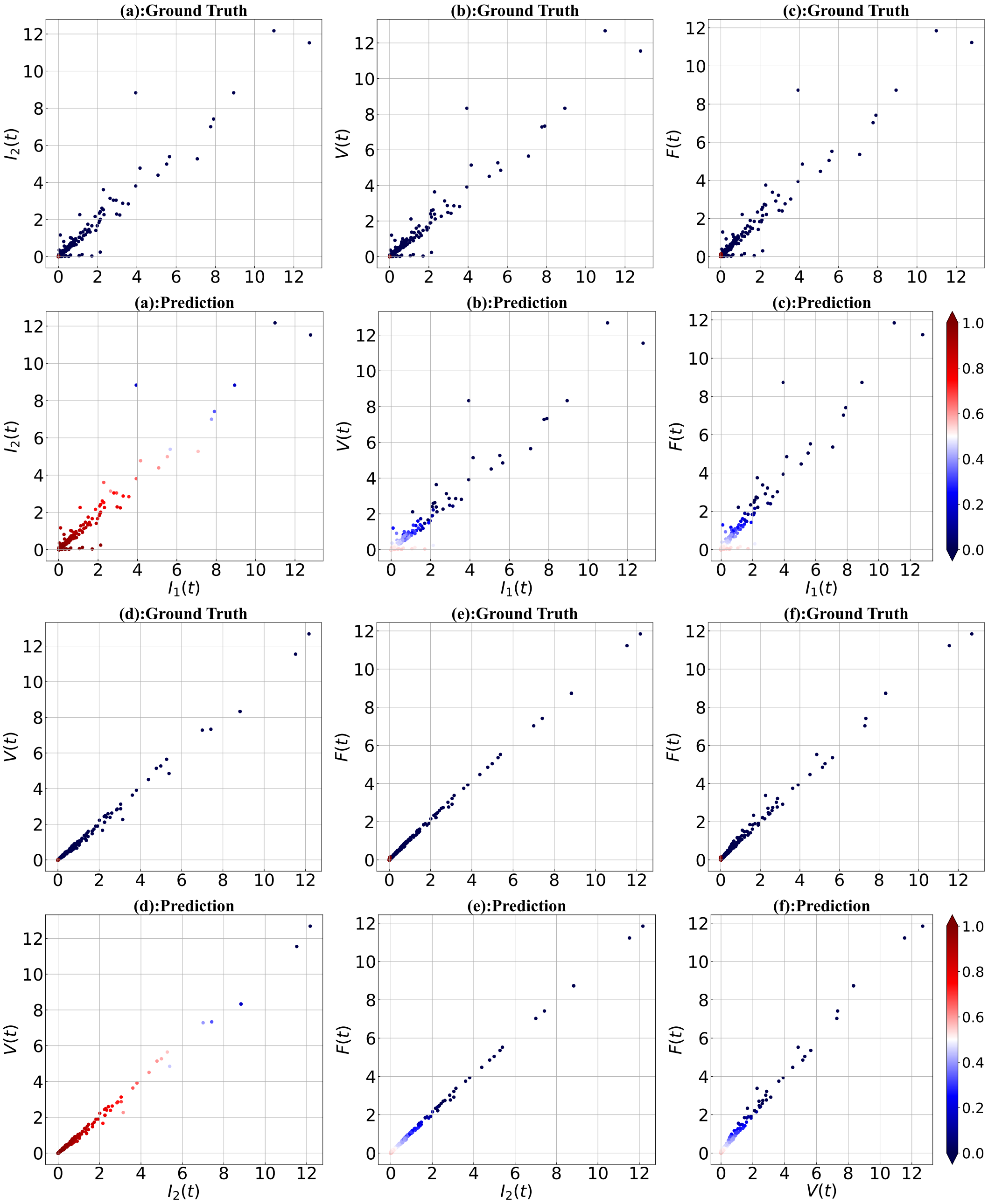

Figure 8. Early days of infection. Two-dimensional scatter plots of the ground truth and regression predicted values based on model features are shown. Classification is based on I_1, I_2 in panels (a), I_1, V in panels (b), I_1, F in panels (c), I_2, V in panels (d), I_2, F in panels (e), and V, F in panels (f). Shades of blue (red) indicate influenza (COVID-19) group patients. Data points are dimensionless by dividing by the corresponding standard deviations.

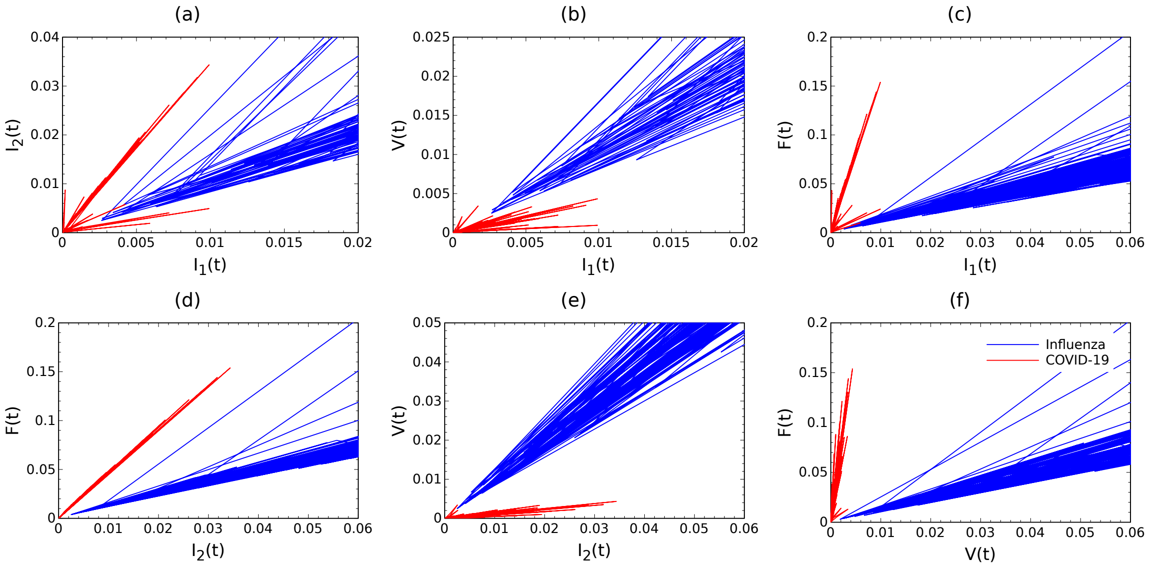

Figure 9. Comparison of in-host measurements, $\{T, I_1, I_2, V, F\}$, between influenza and COVID-19 virtual patients where plotted as a function of each other. Blue(red) solid lines represent the ratio of the features for one hundred influenza (COVID-19) patients. Data points are divided by the corresponding standard deviations for each feature.

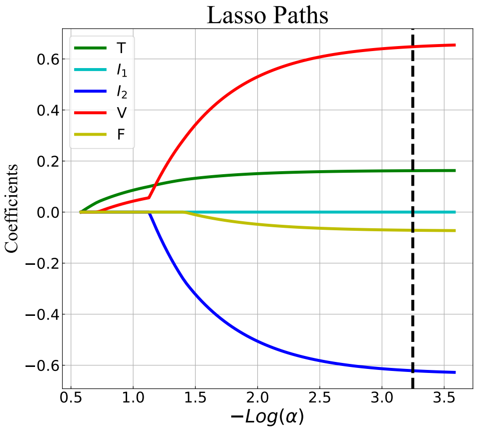

Figure 10. Lasso coefficients of five sample features, $\{T, I_1, I_2, V, F\}$, as a function of the logarithm of regularization parameter, $-\log\alpha$. Each colored line represents the value taken by a different coefficient in the optimization objective for Lasso. The black dashed line indicates the selected regularization parameter with the value of $-\log(\alpha) \approx 3.25$. This number was $\approx$ 3.04 with the same Lasso Paths when the early days of the infection period were considered.

PLS regression – (ROC AUC= 95%) for the main infection period and –(ROC AUC= 89%) for the early days of infection.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the NRC Pandemic Response Challenge Program Grant No. PR016-1.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. L. D. Manzanares-Meza, O. Medina-Contreras, SARS-CoV-2 and influenza: A comparative overview and treatment implications, *Bol. Med. Hosp. Infant. Mex.*, **77** (2020), 262–273. https://doi.org/10.24875/bmhim.20000183

2. K. Subbarao, S. Mahanty, Respiratory virus infections: Understanding COVID-19, *Immunity*, **52** (2020), 905–909. https://doi.org/10.1016/j.immuni.2020.05.004

3. H. Faury, C. Courboulès, M. Payen, A. Jary, P. Hausfater, C. E. Luyt, et al., Medical features of COVID-19 and influenza infection: A comparative study in Paris, France, *J. Infect.*, **82** (2021), e36–e39. https://doi.org/10.1016/j.jinf.2020.08.017

4. X. Zheng, H. Wang, Z. Su, W. Li, D. Yang, F. Deng, et al., Co-infection of SARS-CoV-2 and influenza virus in early stage of the COVID-19 epidemic in Wuhan, China, *J. Infect.*, **81** (2020), e128–e129. https://doi.org/10.1016/j.jinf.2020.05.041

5. S. Azekawa, H. Namkoong, K. Mitamura, Y. Kawaoka, F. Saito, Co-infection with SARS-CoV-2 and influenza A virus, *IDCases*, **20** (2020), e00775. https://doi.org/10.1016/j.idcr.2020.e00775

6. H. Khorramdelazad, M. H. Kazemi, A. Najafi, M. Keykhaee, R. Z. Emameh, R. Falak, Immunopathological similarities between COVID-19 and influenza: Investigating the consequences of Co-infection, *Microb. Pathog.*, **152** (2021), 104554. https://doi.org/10.1016/j.micpath.2020.104554

7. P. K. Bhatraju, B. J. Ghassemieh, M. Nichols, R. Kim, K. R. Jerome, A. K. Nalla, et al., Covid-19 in critically ill patients in the Seattle region—case series, *NEJM.*, **382** (2020), 2012–2022. https://doi.org/10.1056/NEJMoa2004500

8. N. Yanamala, N. H. Krishna, Q. A. Hathaway, A. Radhakrishnan, S. Sunkara, H. Patel, et al., A vital sign-based prediction algorithm for differentiating COVID-19 versus seasonal influenza in hospitalized patients, *NPJ Digit. Med.*, **4** (2021), 1–10. https://doi.org/10.1038/s41746-021-00467-8

9. M. Ackermann, S. E. Verleden, M. Kuehnel, A. Haverich, T. Welte, F. Laenger, et al., Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in Covid-19, *NEJM.*, **383** (2020), 120–128. https://doi.org/10.1056/NEJMoa2015432

10. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, et al., Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia, *NEJM.*, **382** (2020), 1199–1207. https://doi.org/10.1056/10.1056/NEJMoa2001316

11. N. Zhu, D. Zhang, W.Wang, X. Li, B. Yang, J. Song, et al., A novel coronavirus from patients with pneumonia in China, 2019, *NEJM.*, **382** (2020), 727–733. https://doi.org/10.1056/NEJMoa2001017

12. M. S. Ciupe, J. M. Heffernan, In-host modeling, *Infect. Dis. Model.*, **2** (2017), 188–202. https://doi.org/10.1016/j.idm.2017.04.002

13. D. Kyte, J. Ives, H. Draper, T. Keeley, M. Calvert, Inconsistencies in quality of life data collection in clinical trials: A potential source of bias? Interviews with research nurses and trialists, *PLoS One*, **8** (2013), e76625. https://doi.org/10.1371/journal.pone.0076625

14. F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, et al., Artificial intelligence in healthcare: Past, present and future, *Stroke Vasc. Neurol.*, **2** (2017). http://dx.doi.org/10.1136/svn-2017-000101

15. T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, *Future Healthc. J.*, **6** (2019), 94. https://doi.org/10.7861/futurehosp.6-2-94

16. A. Bohr, K. Memarzadeh, The rise of artificial intelligence in healthcare applications, *Artif. Intell. Med.*, (2020), 25–60. https://doi.org/10.1016/B978-0-12-818438-7.00002-2

17. M. Mirbabaie, S. Stieglitz, N. Nicholas RJ. Frick, Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction, *Health Technol.*, **11** (2021), 693–731. https://doi.org/10.1007/s12553-021-00555-5

18. P. Baccam, C. Beauchemin, C. A. Macken, F. G. Hayden, A. S. Perelson, Kinetics of influenza A virus infection in humans, *Virol. J.*, **80** (2006), 7590–7599. https://doi.org/10.1128/JVI.01623-05

19. A. Gonçalves, J. Bertrand, R. Ke, E. Comets, X. De Lamballerie, D. Malvy, et al., Timing of antiviral treatment initiation is critical to reduce SARS-CoV-2 viral load, *CPT Pharmacometrics Syst. Pharmacol.*, **9** (2020), 509–514. https://doi.org/10.1002/psp4.12543

20. A. W. Salehi, P. Baglat, G. Gupta, Review on machine and deep learning models for the detection and prediction of Coronavirus, *PloS one*, **33** (2020), 3896–3901. https://doi.org/10.1016/j.matpr.2020.06.245

21. A. Alimadadi, S. Aryal, I. Manandhar, B. P. Munroe, B. Joe, Xi. Cheng, Artificial intelligence and machine learning to fight COVID-19, *Physiol. Genomics*, **52** (2020), 200–202. https://doi.org/10.1152/physiolgenomics.00029.2020

22. A. W. Salehi, P. Baglat, G. Gupta, Alzheimer's disease diagnosis using deep learning techniques, *Int. J. Eng. Adv. Technol.*, **9** (2020), 874–880. https://doi.org/10.35940/ijeat.C5345.02

23. P. Cao, A. W. Yan, J. M. Heffernan, S. Petrie, R. G. Moss, L. A. Carolan, et al., Innate immunity and the inter-exposure interval determine the dynamics of secondary influenza virus infection and explain observed viral hierarchies, *PLoS Comput. Biol.*, **11** (2015), e1004334. https://doi.org/10.1371/journal.pcbi.1004334

24. A. L. Jenner, R. A. Aogo, S. Alfonso, V. Crowe, X. Deng, A. P. Smith, et al., COVID-19 virtual patient cohort suggests immune mechanisms driving disease outcomes, *PLoS Pathog.*, **17** (2021), e1009753. https://doi.org/10.1371/journal.ppat.1009753

25. F. McNab, K. Mayer-Barber, A. Sher, A. Wack, A. O'garra, Type I interferons in infectious disease, *Nat. Rev. Immunol.*, **15** (2015), 87–103. https://doi.org/10.1038/nri3787

26. N. Néant, G. Lingas, Q. Le Hingrat, J. Ghosn, I. Engelmann, Q. Lepiller, et al., Modeling SARS-CoV-2 viral kinetics and association with mortality in hospitalized patients from the French COVID cohort, *Proc. Natl. Acad. Sci.*, **118** (2021), e2017962118. https://doi.org/10.1073/pnas.2017962118

27. L. B. Ivashkiv, L. T. Donlin, Regulation of type I interferon responses, *Nat. Rev. Immunol.*, **14** (2014), 36–49. https://doi.org/10.1038/nri3581

28. K. A. Pawelek, G. T. Huynh, M. Quinlivan, A. Cullinane, L. Rong, A. S. Perelson, Modeling within-host dynamics of influenza virus infection including immune responses, *PLoS Comput. Biol.*, **8** (2012), e1002588. https://doi.org/10.1371/journal.pcbi.1002588

29. F. G. Hayden, R. Fritz, M. C. Lobo, W. Alvord, W. Strober, S. E. Straus, Local and systemic cytokine responses during experimental human influenza A virus infection. Relation to symptom formation and host defense, *J. Clin. Investig.*, **101** (1998), 643–649. https://doi.org/10.1172/JCI1355

30. N. K. Vaidya, A. Bloomquist, A. S. Perelson, Modeling Within-Host Dynamics of SARS-CoV-2 Infection: A Case Study in Ferrets, *Viruses*, **13** (2021), 1635. https://doi.org/10.3390/v13081635

31. L. Bordi, G. Sberna, E. Lalle, P. Piselli, F. Colavita, E. Nicastri, et al., Frequency and duration of SARS-CoV-2 shedding in oral fluid samples assessed by a modified commercial rapid molecular assay, *Viruses*, **12** (2020), 1184. https://doi.org/10.3390/v12101184

32. W. H. Mahallawi, A. D. Alsamiri, A. F. Dabbour, H. Alsaeedi, A. H. Al-Zalabani, Association of viral load in SARS-CoV-2 patients with age and gender, *Front. Med.*, **8** (2021), 39. https://doi.org/10.3389/fmed.2021.608215

33. K. Ejima, K. S. Kim, C. Ludema, A. I. Bento, S. Iwanami, Y. Fujita, et al., Estimation of the incubation period of COVID-19 using viral load data, *Epidemics*, **35** (2021), 100454. https://doi.org/10.1016/j.epidem.2021.100454

34. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. B Stat. Methodol.*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

35. X. Han, M. S. Ghaemi, K. Ando, L. S. Peterson, E. A. Ganio, A. S. Tsai, et al., Differential dynamics of the maternal immune system in healthy pregnancy and preeclampsia, *Front. Immunol.*, **10** (2019), 1305. https://doi.org/10.3389/fimmu.2019.01305

36. T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artif. Intell.*, **267** (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

37. B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, *Adv. Neural Inf. Process. Syst.*, **29** (2016), 2288—2296.

38. C. Molnar, Interpretable machine learning, *Lulu. Com.*, (2020).

39. W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, *Proceedings of the National Academy of Sciences*, **116** (2019), 22071–22080. https://doi.org/10.1073/pnas.1900654116

40. O. Dogan, S. Tiwari, M. A. Jabbar, S. Guggari, A systematic review on AI/ML approaches against COVID-19 outbreak, *Complex Intell. Syst.*, **7** (2021), 2655–2678. https://doi.org/10.1007/s40747-021-00424-8

41. M. A. Quiroz-Juárez, A. Torres-Gómez, I. Hoyo-Ulloa, R. d. J. León-Montiel, A. B. U'Ren, Identification of high-risk COVID-19 patients using machine learning, *PLoS One*, **16** (2021), e0257234. https://doi.org/10.1371/journal.pone.0257234

42. M. M. Rahman, F. Khatun, A. Uzzaman, S. I. Sami, M. A. Bhuiyan, T. S. Kiong, A comprehensive study of artificial intelligence and machine learning approaches in confronting the coronavirus (COVID-19) pandemic, *PLoS One*, **51** (2021), 446–461. https://doi.org/10.1177/00207314211017469