Original Research Article

MDMP&P Policy & Practice

MDM Policy & Practice
2023, Vol. 8(2) 1–21
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23814683231218716
journals.sagepub.com/home/mpp
Sage

Informed Random Forest to Model Associations of Epidemiological Priors, Government Policies, and Public Mobility

Tsaone Swaabow Thapelo, Dimane Mpoeleng, and Gregory Hillhouse

Abstract

Background. Infectious diseases constitute a significant concern worldwide due to their increasing prevalence, associated health risks, and the socioeconomic costs. Machine learning (ML) models and epidemic models formulated using deterministic differential equations are the most dominant tools for analyzing and modeling the transmission of infectious diseases. However, ML models can be inconsistent in extracting the dynamics of a disease in the presence of data drifts. Likewise, the capability of epidemic models is constrained to parameter dimensions and estimation. We aimed at creating a framework of informed ML that integrates a random forest (RF) with an adapted susceptible infectious recovered (SIR) model to account for accuracy and consistency in stochasticity within the dynamics of coronavirus disease 2019 (COVID-19). **Methods.** An adapted SIR model was used to inform a default RF on predicting new COVID-19 cases (NCCs) at given intervals. We validated the performance of the informed RF (IRF) using real data. We used Botswana's pharmaceutical interventions (PIs) and non-PIs (NPIs) adopted between February 2020 and August 2022. The discrepancy between predictions and observations is modeled using loss functions, which are minimized, interpreted, and used to assess the IRF. **Results.** The findings on the real data have revealed the effectiveness of the default RF in modeling and predicting NCCs. The use of the effective reproductive rate to inform the RF yielded an excellent predictive power (84%) compared with 75% by the default RF. **Conclusion.** This research has potential to inform policy and decision makers in developing systems to evaluate interventions for infectious diseases.

Corresponding Author
Tsaone Swaabow Thapelo, Department of Computer Science and Information Systems, Botswana International University of Science and Technology, Khurumela, 2, Palapye, Botswana; (swaabow@gmail.com).

Highlights

- This framework is initiated by incorporating model outputs from an epidemic model to a machine learning model.
- An informed random forest (RF) is instantiated to model government and public responses to the COVID-19 pandemic.
- This framework does not require data transformations, and the epidemic model is shown to boost the RF's performance.
- This is a baseline knowledge-informed learning framework for assessing public health interventions in Botswana.

Keywords
knowledge-informed learning, random forest, COVID-19, public mobility, epidemiology, policies, interventions

Date received: February 7, 2023; accepted: November 1, 2023

Infectious diseases such as coronavirus disease 2019 (COVID-19) constitute a significant concern worldwide due to their increasing prevalence, associated health risks, and socioeconomic costs. The determinants of a given infectious disease are complex, and they include but are not limited to 1) environmental factors (i.e., weather and climate) and 2) human behavioral characteristics (i.e., public mobility) and political factors such as government policies and interventions. Previous studies have shown that machine learning (ML) models[1] and epidemic compartmental models[2] formulated using deterministic differential equations are the dominant tools for examining, modeling, and analyzing the transmission of infectious diseases.

An ongoing discourse on the mentioned paradigms of modeling highlights the various viewpoints and approaches for epidemiological studies. Although ML models enable the extraction of insights from data,[3] their performance may be inconsistent[4] when exposed to stochasticity, scenarios of an unpredictable nature with data distributions that change over time. This can result in biases due to model drifts[5] that can lead to poor generalization on new cases.[6] Likewise, the capability of epidemic models is constrained to the problem of parameter dimensions and estimation.[2]

Currently, discussions persist as to whether knowledge-informed learning (KIL) is a viable avenue to bridge the gap between ML and dynamic models—to inform decision support systems for ML practitioners, policy developers, and decision makers. KIL,[7] also framed as physics-informed ML (PIML),[6] entails the synthesis of multiple viewpoints, principles, and evidence to provide informative priors for modeling. There is no standard method to incorporate prior knowledge in ML.

This research aims to construct a random forest (RF) ensemble[8] informed by the susceptible infected recovery (SIR) model,[9] then test it by conducting a retrospective study of the evolution of the COVID-19 disease. The study objective is 2-fold: 1) to predict new infections using variables that have greater associations to the state of the function dictating the growth rate of COVID-19 cases and 2) to extract the information about the constructed model using variable importance[10] and partial dependence[11] functions. The next section discusses the related literature. The third section presents the methodology to achieve the research aim and objectives. The

Department of Computer Science and Information Systems, Botswana International University of Science and Technology, Palapye, Botswana (TST); Director (Ag.) Research Innovation Technology, Research Development and Innovation, Department of Computer Science and Information Systems, Botswana International University of Science and Technology, Palapye, Botswana (DM); Head of the Department of Physics and Astronomy, Botswana International University of Science and Technology, Palapye, Botswana (GH). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors received no financial support for the research, authorship, and/or publication of this article.

fourth section presents the findings. The fifth section presents discussions, while the sixth section concludes.

Related Literature

Previous scholars have discussed various approaches of KIL[7] to improve ML models, and they revolve around the use of biases in observations, inductions, and learning.[6] Observational biases[12] can be achieved by introducing data characterizing the underlying dynamic system of interest into an ML model.[6] The inductive bias approach[13] uses mathematical expressions to 1) adapt variables (feature engineering) for data assimilation or augmentation or 2) tailor the model's structure to satisfy certain physical principles, assumptions, constraints, and boundary conditions.[6] The approach of learning biases implicitly enforces prior knowledge to impose ML constraints by properly penalizing loss functions.[6]

Common methods for training, testing, and validating ML models include percentage splitting and cross-validation using a loss function.[14] Strategies for reducing computational time include testing models using only a subset of instances[15] or via online ML using individual instances on the fly.[16] Meanwhile, ML models are known to be prune to concept drifts, phenomena in which relationships between dependent variables and independent variables change over time due to data drifts.[5] The term *data drift* defines the changes in data distributions (indicated by the model's variables) over time.[17] As new data accumulate, the model may struggle to maintain accurate predictions, leading to model drifting (reduced performance).

There is not a one-size-fits-all consensus to guide the sampling of data for ML; researchers use different splitting ratios.[14] Moreover, the topic of hyperparameter tuning to optimize loss functions continues to be a longstanding subject in predictive modeling.[18] Loss functions are used to strike a balance between fitting the training data accurately and avoiding model overfitting.[19] Chicco et al.[20] claimed that the coefficient of determination is the only informative metric in regression. To test this claim, this work examines the accuracy, consistency, and interpretability of a default RF[8] and an informed RF (IRF) when exposed to data drifts and stochasticity using standard loss functions, taking COVID-19 as a case study (see Appendix 1).

Current PIML literature is focused on deep learning models,[6] while the scope of RF is less explored despite its successful applications.[21] To support this premise, the following search string "*TS = (COVID-19 and Random Forest (RF)) and English (Languages)*" was used to extract related work from the *Web of Science* collections, focusing on the RF applications in addressing COVID-19. Only 3 of 8 publications[22–24] met our criteria. We also queried *ScienceDirect* using the following search string: "*TS = (COVID-19 and Random Forest) and TS = (prediction or forecast) and TS = (regression)*," focusing on English as a language, and retrieved 1 of 33 publications (see Table 1 and Table 4 in the Appendices).

The output from the search process supports Biau and Scornet[26] in that there is a lack of agreement regarding the optimal parameters of the RF. For instance, Biau and Scornet[26] noted that the forest's variance decreases as the number of trees grows, while accurate predictions are likely to be obtained by choosing a large number of trees. Meanwhile, Díaz-Uriarte and Alvarez de Andrés[21] argued that the value for the number of trees is irrelevant. In terms of the number of variables tried at split nodes of a tree, Díaz-Uriarte and Alvarez de Andrés[21] argued that it has a little impact on the performance of the RF and that larger values may be associated with a reduction in the RF's predictive performance. Likewise, Genuer et al.[27] claimed that the default value is either optimal or too small.

Bentéjac et al.[28] compared the performance of the XGBoost, LightGBM, CatBoost, and the RF on 28 different data sets using default and tuned settings. They found that the differences in terms of generalization accuracy of the configured versions was small. Bentéjac et al.[28] found that default versions of the RF and CatBoost produced consistent and stable results compared with their tuned counterparts. Szczepanek[29] made the same remark after comparing default settings for XGBoost, LGBoost, CatBoost, and RF in forecasting stream-flow in mountainous catchment. We chose to use the RF since it has few arguments to tune and can perform well in both default and tuned settings,[30] with the following contribution:

- This work focuses on investigating and evaluating the performance of a RF ensemble to allow a thorough analysis and understanding of its behavior, strengths, and limitations using real data. This is expected to provide insights into its accuracy and consistency under different time frames in the presence of data drifts and prior domain knowledge.
- We integrate the basic principles of observational bias and induction bias to optimize loss functions used to evaluate the default RF algorithm, showcasing its performance in predicting COVID-19 cases in the context of Botswana.

Table 1 Related Work on Applications of the Random Forest (RF) for COVID-19 Based on Web of Science and Science Direct

Reference	Purpose	Methods	Supervised Learning	Location
Lima et al.[22]	Estimation of death risk in people (>60 y old) diagnosed with COVID-19	Used variable importance to inform a RF model trained using the 10-cross-validation	Classification using area under the receiver-operating characteristic curve (AUC)	Pernambuco (Brazil)
Ramírez García and Jiménez Preciado[23]	Predicting COVID-19 dynamics and implications for economic growth	Used susceptible infected recovered and RF model	Regression using mean squared error (MSE)	United States, Brazil, United Kingdom, Mexico, China, India, Japan
Villagrana-Bañuelos et al.[24]	Predicting COVID-19 using clinical and metabolic data	Applied genetic algorithms to a RF model trained using the 5-cross-validation	Classification using AUC	Zacatecas (Mexico)
Milicevic et al.[25]	Predicting the reproduction (basic) number of COVID-19	Principal component analysis, Lasso, elastic net, RF, and gradient boosting using a splitting ratio of 0.8	Regression task: using the MSE	United States

- This analysis can inform technology transfer when incorporating known principles to enhance interpretability of ML.

Methodology

Data Collection and Understanding

The Google mobility indices used to indicate visits to public places are[31] 1) retail and recreation, 2) grocery and pharmacy stores, 3) transit stations, 4) parks and outdoor spaces, 5) workplace visitors, and 6) residential. The economic indices include 1) income support and 2) debt and contract relief. Epidemiological indices are 1) new COVID-19 cases, 2) smoothed new COVID-19 tests, 3) total COVID-19 cases, 4) effective reproduction rate ($\hat{\mathcal{R}}$), 5) positive rate, 6) new COVID-19 deaths, 7) total COVID-19 deaths, and 8) new vaccinations smoothed. The government Stringency Index (SI) is used to quantify the containment measures based on the Oxford COVID-19 Government Response Tracker (OxCGRT).

The Containment and Health Index (CHI), adapted from the SI, was used to aggregate the effectiveness of containment measures and health regulations comprising public information campaign, face coverings, school closing, workplace closures, cancellation of public events, cancel public gatherings, restrictions on gathering size, travel restrictions, international travel controls, close public transport, restrictions on internal movement, stay-at-home requirements, contact tracing, and vaccination availability. To introduce observational biases to the RF model, we use the effective reproductive rate $\hat{R}$ from the modified SIR model[32] to argument the observational data directly in the training phase as seen in Table 2.

Table 2 shows the variables used in this work; the right-most column reports the number of missing values in the data set. Missing values indicate that the event did not occur or say the count is zero. The positive rate index has $\frac{79}{906}\approx 9\%$ of missing values (from December 31, 2021, to January 14, 2022; from March 30, 2022, to April 5, 2022; and from May 19, 2022, to August 18, 2022). These 3 batches are squeezed between constant values, so we replaced them using the last observation carry-forward (LOCF) method.[34] LOCF assumes that the most recent observation gives a reasonable estimate for missing values.[34,35] We examined the SI, $\hat{R}$, new COVID-19 cases (NCCs), and new COVID-19 deaths (NCDs) and found that no events were recorded in Botswana before their dates of observation; thus, we imputed them using zeros.[14]

Hence, we use a structured labeled data set $\mathbb{D} = (X, Y)$, with $m = 906$ instances (observations) composed of $n = 32$ variables, where X denotes predictors

Table 2 List of Variables Sourced from Multidomains, with the Time Frame of Observation and the Number of Missing Values

Variable	Domain	Source	Time Span	Type	Missing
Retail and recreation (RRV)	Mobility	Google	February 17, 2020, to August 10, 2022	Numeric	0
Grocery and pharmacy stores (GPSV)	Mobility	Google	February 17, 2020, to August 10, 2022	Numeric	0
Transit stations (TSV)	Mobility	Google	February 17, 2020, to August 10, 2022	Numeric	0
Parks and outdoor spaces (POSV)	Mobility	Google	February 17, 2020, to August 10, 2022	Numeric	0
Workplace visitors (WV)	Mobility	Google	February 17, 2020, to August 10, 2022	Numeric	0
Residential (RSV)	Mobility	Google	February 17, 2020, to July 18, 2022	Numeric	0
School closures (SC)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Workplace closures (WC)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Close public transport (CPT)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Restrictions on public gatherings (RPG)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Face coverings (FC)	Policy	OxCGRT	February 17, 2020, July 18, 2022	Factor	0
Public information campaigns (PIC)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Public transport closures (PTC)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Restrictions on internal movement (RIM)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
International travel controls (ITC)	Policy	OxCGRT	February 17, 2020, to July 18, 2022	Factor	0
Testing policy (TP)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Contact tracing (CT)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Stay-home requirements (SHR)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Factor	0
Vaccine availability (VA)	Policy	Rodés-Guirao et al.[33]	February 17, 2020, to July 18, 2022	Factor	0
Containment and Health Index (CHI)	Policy	OxCGRT	February 17, 2020, to August 10, 2022	Numeric	0
Stringency Index (SI)	Policy	OxCGRT	March 30, 2020, to August 10, 2022	Numeric	0
Income support (IS)	Economic	WHO	May 10, 2021, to August 10, 2022	Factor	0
Debt and contract relief (DCR)	Economic	WHO	May 10, 2021, to July 18, 2022	Factor	0
Effective reproduction rate ($\hat{\mathcal{R}}$)	Epidemiology	Arroyo-Marioli et al.[32]	June 30, 2020, to August 10, 2022	Numeric	0
Positive rate (PR)	Epidemiology	WHO	April 10, 2020, to May 18, 2022	Numeric	79
New COVID-19 deaths (NCDs)	Epidemiology	WHO	March 26, 2020, to August 10, 2022	Numeric	0
Total COVID-19 deaths (TCDs)	Epidemiology	WHO	March 26, 2020, to August 10, 2022	Numeric	0
Smoothed new COVID-19 tests (SNCT)	Epidemiology	WHO	April 10, 2020, to August 10, 2022	Numeric	0
Smoothed new vaccinations (SNV)	Epidemiology	WHO	March 26, 2021, to May 18, 2022	Numeric	0
Total COVID-19 cases (TCCs)	Epidemiology	WHO	March 31, 2020, to August 10, 2022	Numeric	0
New COVID-19 cases (NCCs)	Epidemiology	WHO	March 30, 2020, to August 10, 2022	Numeric	0

OXCGRT, Oxford COVID-19 Government Response Tracker; WHO, World Health Organization.

(inputs) variables while Y denotes the target (output) variable (i.e., NCCs), represented as

$$(x_{m,n}|y_m) = \left(\begin{array}{ccc|c} x_{1,1} & \cdots & x_{1,n} & y_1 \\ x_{2,1} & \cdots & x_{2,n} & y_2 \\ \vdots & \ddots & \vdots & \vdots \\ x_{m,1} & \cdots & x_{m,n} & y_m \end{array}\right) \text{such that:} \qquad (1)$$

$x \in X, \text{ and } y \in Y.$

Informed Machine Learning

Data sampling. The RF is trained and validated on data in a sequential order, with each time frame building upon the previous one. The process of training, testing, and validating the constructed RF model has 3 phases as follows[36]:

- *Training*: The informed RF is trained on a partition of labeled data set, of $\frac{2}{3}$ of the observations ($\approx$66.67%), in the training data set to learn the hidden patterns in data and adjust its internal parameters[8] based on the loss functions.
- *Testing*: The model's hyperparameters are fine-tuned based on its performance on a separate testing data set (out of bag), a $\frac{1}{3}$ partition left out during training.[8] This testing data set ($\approx$33.33%) helps to prevent overfitting. Moreover, it also enables the adjustment (tuning) of the model's hyperparameters (arguments) to optimize its performance.

- *Validation*: The model is evaluated on a completely separate (extra) data set that it has never encountered before. This validation data set is optional[14] to provide an unbiased assessment of the model's performance on new (unseen) data.

Mathematical statement. Our task is to build a regression $f: X^n \rightarrow \mathbb{R}$ and predict $y \in \mathbb{R}$, where $x \in X^n$ is a matrix of n predictor variables, while $y \in Y$ is the target variable of interest (i.e., NCCs). The following formulation was used for modeling: [y $\sim$ X]. A number of ML algorithms such as artificial neural networks (ANN), support vector machines (SVM), naive Bayes, and trees can be used for predictive modeling. Our data set (see Table 2) has portion of the variables (53.6%) classified as a signifcat, and only 43% are numeric. Tree-based algorithms are ideal for this task since they can handle both numeric and categorical (factors) variables (features) without additional data transformation.[14]

Random forest. A RF algorithm is one of the widely used ensemble models in practice.[8,14,26,37] The RF ensemble is a decision tree (DT)–based model aimed at improving model accuracy and robustness.[14] A DT uses a tree like structure,[38] in which each node denotes a decision based on the values of x_i. Basically, a DT is an interpretable model that recursively partitions the set x using a sequence of binary decisions based on "if-then-else" rules, as seen in Figure 1.

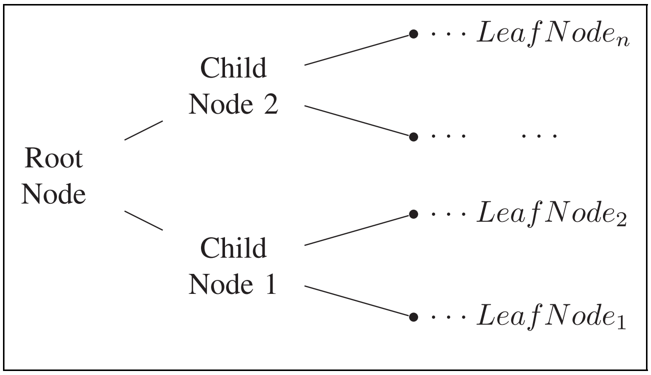

Figure 1 A tree structure with the root node on the left, branches in the middle, and leaf nodes on the right. Every instance in the training data set must end up in 1 and only 1 leaf node. Leaf nodes are collectors of instances matching the decision tree rules.

DTs are built recursively by splitting x at each test node $\mathbb{A}$ based on the chosen variable $x \in X$, its value, and the splitting criterion. The value of x is used to make a decision at each node $\mathbb{A}$, and each leaf node represents a predicted output $\bar{y} = f(x,y)$. The DT regression implements Eq. 2 to calculate numeric values of $\bar{y}$ for each instance in the set of predictors x by summing up all the numeric values $\mathbb{V}$ assigned to the leaf nodes, as determined by the DT rules (see Song and Lu[38]):

$$\bar{y} = f(x,y) = \sum_{j=1}^{J} \mathbb{V}_j 1(x \in \mathbb{A}_j), \qquad (2)$$

where $\bar{y}$ is the predicted value given x_m, x is the input instance to use in predictions, J is the total number of leaf nodes $(\mathbb{A}_j)$ in the DT, $\mathbb{V}_j$ is the value assigned to $\mathbb{A}_j$, and $\mathbb{1}(x \in \mathbb{A}_j)$ is an indicator function (returns 1 if $x \in \mathbb{A}_j$ and 0 otherwise). Building on DTs, the RF[8] uses bootstrapping (sampling with replacement) to build a DT from each bootstrapped sample $\mathbb{D}^*$. The RF introduces randomness by considering only a subset of variables at each node $\mathbb{A}_j$ when building the tree. Predictions from $\mathbb{K}$ trees are aggregated into a single composite score by averaging using the DT regression[8,26] as

$$\hat{y} = \hat{f}(x,y) = \mathbb{K}^{-1} \sum_{k=1}^{\mathbb{K}} f_k(\mathbb{K}, \mathbb{D}_k^*, \mathbb{D}_{m,n}, \omega_k, \phi_k). \qquad (3)$$

Here, $\mathbb{D}^*$ denotes bootstrap samples created from the original data set $\mathbb{D}_{m,n}$, $\hat{y} = \hat{f}(\cdot)$ is the predicted value by the ensemble, $\mathbb{K}$ is the number of trees, $f_k(\cdot)$ is the predicted value for x based on the k^{th} tree, $\omega = Mtry$ is the number of randomly chosen variables at each node, and $\phi = nodeSize$ is the minimum number of samples required to create a split at a node. Theoretically, the narration of the RF method is less conclusive, as noted in Biau and Scornet[26]; little is known about the mathematical properties of the method. For that reason, we focus only on the RF ensemble (in its default and tuned settings) to examine its hyperparameters with and without the presence of outputs from the SIR model discussed in the next section.

We chose the RF for several reasons: 1) reproducibility, it has few arguments to tune[26] and is easy to reproduce; 2) uncertainty quantification, as an ensemble, a RF uses aggregations to reduce overfitting, thus improving generalization performance[8]; 3) scalability, it also scales well with varying sample sizes and high-dimensional spaces,[8] and robustness, it is a nonparametric model and hence not sensitive to assumptions about data distributions[37,39]; and 4) model interpretability, lost in the RF compared with DT,[14] we circumvented this problem using post hoc interpretations such as variable importance[37] and partial dependence functions.[11] The SIR model is deployed to inform the RF ensemble.

SIR model. The SIR model is used to understand the dynamics of infectious diseases in a population of size N composed of 3 compartments: S, I, and R, such that $N = S(t) + I(t) + R(t)$[40]:

$$S(t + \Delta t) = S(t) - \beta \cdot \Delta t \cdot N^{-1} \cdot S(t) \cdot I(t) \mid 0 < \beta < 1, \tag{4a}$$

$$\begin{aligned} I(t + \Delta t) &= I(t) + \beta \cdot \Delta t \cdot N^{-1} \cdot S(t) \cdot \\ I(t) &- \gamma \cdot \Delta t \cdot I(t) \mid 0 < \gamma < 1, \end{aligned} \tag{4b}$$

$$R(t + \Delta t) = R(t) + \gamma \cdot \Delta t \cdot I(t). \tag{4c}$$

where S is a compartment of susceptible individuals probable of being infected at time t and I denotes *infectious* individuals. R_t are removed individuals from the I compartment (those who have gained immunity or have passed away),[32]Δ_t is a small interval of time t, $\beta \cdot \Delta_t$ is the probability that an infected can infect a susceptible, and $\gamma \cdot \Delta_t$ is the probability of an infected to be removed from the I_t group in Δt time. Taking differences of these equations leads to a system of differential equations:

$$\frac{dS(t)}{dt} = -\beta \cdot S(t) \cdot I(t), \tag{5a}$$

$$\frac{dI(t)}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t), \tag{5b}$$

$$\frac{dR(t)}{dt} = \gamma \cdot I(t). \tag{5c}$$

We extend the SIR model to include vaccinations (V) and hospitalizations (H) to get 5 compartments ($C = S, I, R, H, V$) and 10 parameters ($p = \lambda,\ \beta, \mu_1,\ \mu_2,\ \nu, h,\ \zeta, \gamma_1,\ \gamma_2, \gamma_3$), where λ are the recruits for testing, μ_1 is the natural deaths, μ_2 is the death rate of hospitalized people, ν is the vaccination rate, ζ is the death rate due to vaccine complications, and γ_1 and γ_2 are the recovery rates due to natural immunity and vaccinations, respectively, whereas γ_3 is the recovering rate of hospitalized COVID-19 patients as:

$$\frac{dS(t)}{dt} = \lambda - \beta \cdot S(t) \cdot I(t) - (\nu_1 + \mu_1) \cdot S(t), \tag{6a}$$

$$\frac{dI(t)}{dt} = \beta \cdot S(t) \cdot I(t) - (h + \gamma_1 + \nu_2 + \mu_2) \cdot I(t), \tag{6b}$$

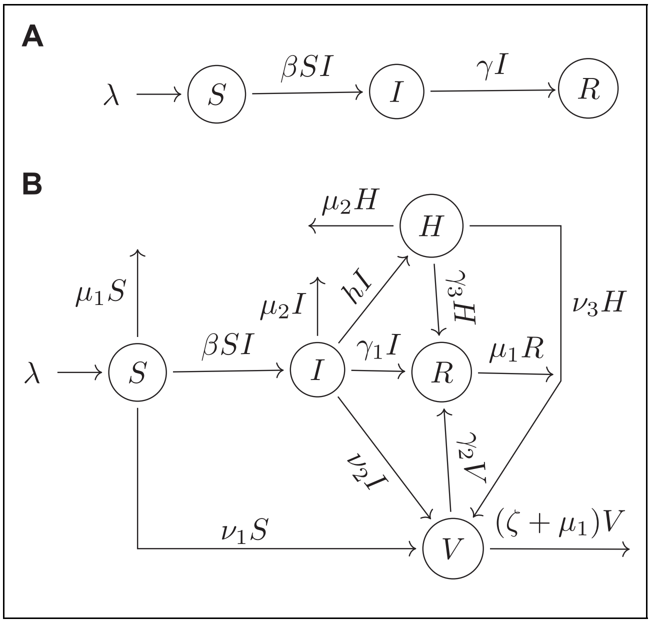

Figure 2 The basic susceptible infectious removed (SIR) model and its modification to accommodate vaccinations and hospitalizations. (a) Basic SIR model. (b) Adapted epidemic disease model (SIRVH).

$$\frac{dH(t)}{dt} = h \cdot I(t) - (\gamma_3 + \mu_2 + \nu_3) \cdot H(t), \tag{6c}$$

$$\frac{dV(t)}{dt} = \nu_1 S(t) + \nu_2 \cdot I(t) + \nu_3 \cdot H(t) - (\gamma_2 + \zeta + \mu_1) \cdot V(t), \tag{6d}$$

$$\frac{dR(t)}{dt} = \gamma_1 \cdot I(t) + \gamma_2 \cdot V(t) + \gamma_3 \cdot H(t) - \mu_1 \cdot R. \tag{6e}$$

where λ denotes tested people. Deaths of people in S, V, and R are assumed to be natural, while deaths of people in I and H are assumed to be caused by COVID-19. Data on natural deaths are scarce, so we assume that removed people are no longer infectious nor susceptible; thus, μ_1 will be ignored when modeling. Figure 2a and b show the SIR and SIRVH models.

Effective reproductive rate ($\hat{\mathcal{R}}$). This is an estimate of the reproduction rate of COVID-19[32] assuming the transmission rate β_t varies over time depending on factors such as COVID-19 interventions and public responses. The basic reproduction rate (average number of individuals infected by a single infected individual when

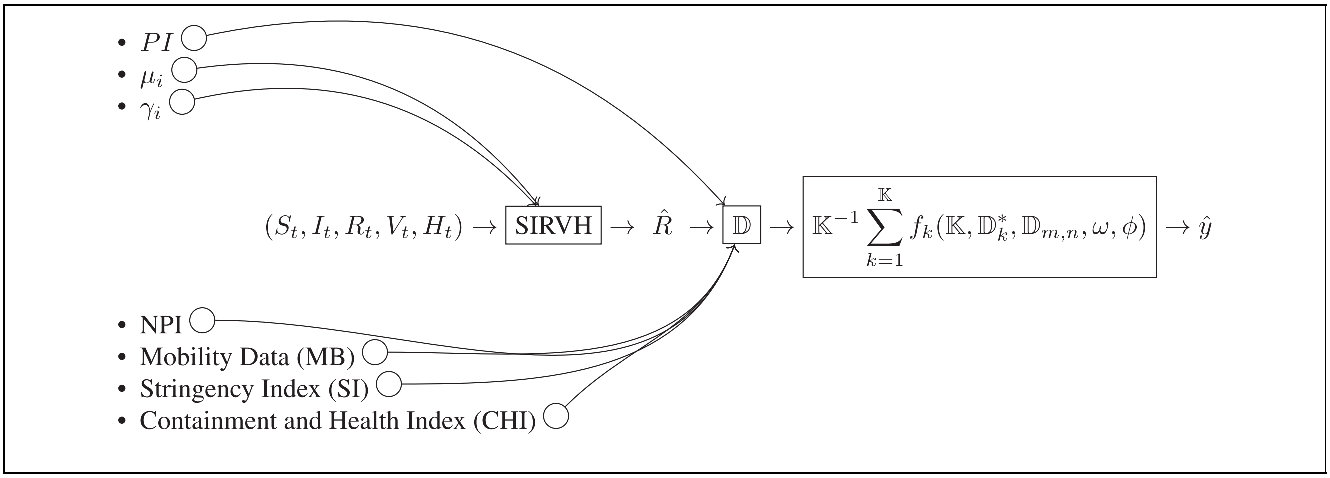
• PI
• μ_i
• γ_i
$(S_t, I_t, R_t, V_t, H_t) \rightarrow$ SIRVH $\rightarrow \hat{R} \rightarrow \mathbb{D} \rightarrow \mathbb{K}^{-1}\sum_{k=1}^{\mathbb{K}} f_k(\mathbb{K}, \mathbb{D}_k^{*}, \mathbb{D}_{m,n}, \omega, \phi) \rightarrow \hat{y}$
• NPI
• Mobility Data (MB)
• Stringency Index (SI)
• Containment and Health Index (CHI)

Figure 3 Framework for informed RF using outputs from the SIR model and observed data to predict COVID-19 cases.

everyone else is susceptible) is given as $\mathcal{R}_0 = \gamma^{-1} \cdot \beta_t$ such that $\mathcal{R}_t = \mathcal{R}_0 \cdot \beta_t N^{-1} \cdot S_{t-1}$. The plug-in estimator for $\hat{\mathcal{R}}$ at time $t + \Delta t$ is given in Arroyo-Marioli et al.[32] as

$$\hat{\mathcal{R}}(t + \Delta t) = 1 + \gamma^{-1}\hat{g}r[I(t + \Delta t)]; \tag{7}$$

where $\hat{g}r[I(t + \Delta t)]$ is an estimate of the daily growth rate $gr[I(\cdot)]$ in the number of infected individuals at time $t + \Delta t$:

$$\begin{aligned} &gr[I(t + \Delta t)] \equiv [I(t + \Delta t)]^{-1} \cdot [I(t + \Delta t) - I(t)] = \\ &\gamma \cdot [\mathcal{R}(t + \Delta t) - 1]. \end{aligned} \tag{8}$$

Using Eq. 6c, we solve $\beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t) = 0$ for β as $\beta = S(t)^{-1} \cdot \gamma$. This is achieved by following Arroyo-Marioli et al.,[32] taking NCC as the daily cases, I_t as the total number of infectious cases at time $t = t_0$, and $\gamma^{-1} = 7$ days, then rearrange Eq. 4b:

$$\begin{aligned} &I(t + \Delta t) = I(t) + \beta \cdot \Delta t \cdot N^{-1} \cdot S(t) \cdot I(t) - \gamma \cdot \\ &\Delta t \cdot I(t) = (1 - \gamma \cdot \Delta t) \cdot I(t) + NCC(t + \Delta t). \end{aligned} \tag{9}$$

Informed RF Model

Embedding physics principles in to the RF. Our final multidomain data set contains observations for the period from February 17, 2020, to August 10, 2022, with 32 variables (31 independent and 1 dependent) of which $\frac{15}{32} \approx 46.9\%$ are categorical (factors), $\frac{16}{32} \approx 50\%$ are numeric, and $\frac{1}{32} \approx 3.1\%$ for date. We set the problem under the context of concept drift. Our aim is to understand the spread of COVID-19 disease by examining changes in Y—the confirmed NCCs with respect to X (the predictors). Algebraic equations used for data augmentations are described as follows:

1. The SI is a mean score[41] computed using the adopted COVID-19 policies ϱ of dimension k on a $0 - 100$ scale:

$$SI = k^{-1} \sum_{i=1}^{k} \varrho_i. \tag{10}$$

2. The positive rate (PR) is the proportion of conducted COVID-19 tests (SNCT) that yielded a positive (NCCs):

$$\text{Positive Rate} = \frac{NCCs}{SNCT}. \tag{11}$$

The final data set $\mathbb{D}$ also includes the effective reproductive rate from the SIRVH model, Google mobility data, government policies (NI and NPI), and their averages (SI and CHI) to train the RF ensemble versions f_k from Eq. 3 as shown in Figure 3.

Model Evaluation Using Loss Functions

To measure the average magnitude of the differences between the predicted and actual values, we used the coefficient of determination (R^2), mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE), the mean squared error (MSE)[42] and the root mean squared error (RMSE). The RMSE measures the average magnitude of the errors as:

$$RMSE = \sqrt{m^{-1} \cdot \sum_{i=1}^{m} [y_i - f(x_i, y_i)]^2}. \tag{12}$$

The MAPE[20] measures the average percentage difference between the predicted values and actual values as:

$$MAPE = m^{-1} \sum_{i=1}^{m} \left| \frac{y_i - f(x_i, y_i)}{f(x_i, y_i)} \right| \cdot 100\%. \tag{13}$$

The SMAPE[20,43] is a scale-invariant metric of the average percentage difference between predicted and actual values:

$$SMAPE = m^{-1} \sum_{i=1}^{m} \frac{|y_i - f(x_i, y_i)|}{2^{-1}(|y_i| + |f(x_i, y_i)|)} \cdot 100\%. \tag{14}$$

We use the R^2 to measure[44] the percentage of the variance of the target variable (NCCs) that is explained by the RF as

$$R^2 = 1 - \frac{\sum_{i=1}^{m} [y_i - f(x_i, y_i)]^2}{\sum_{i=1}^{m} [\tilde{y}_i - f(x_i, y_i)]^2}. \tag{15}$$

In this context, m is the number of instances, $f(x_i, y_i)$ is the predicted NCCs, y_i is the actual NCCs, while $\tilde{y}$ in Eq. 15 is the mean of the actual values y. The values of MSE, RMSE, MAPE, and MAE are bounded below by 0 and bounded upper by $+\infty$, where 0 indicates a perfect fit,[20] while $+\infty$ suggest a larger discrepancy between the predicted values and the actual values. The SMAPE metric is bounded below by 0, its upper bound is 200% to imply that the actual values and predictions are of opposite sign.[20] R^2 values range between $-\infty$ and 1, where $-\infty$ is the worst-case scenario, $R^2 = 0$ means that the model explains none of the variance of the target variable, while 1 indicates that the model perfectly explains the variability.

Post Hoc Interpretation

We also used post hoc interpretation to analyze the results generated by the proposed informed RF ensemble as follows.

Variable importance (VI) functions. We used the vip package[10] as an interface to the model-agnostic approach of quantifying how important a given variable x_i is to the dependent variable $y = f(x, \cdot)$, under the condition that the variable x_i attains a certain value $w(x_i) = \mathbb{V} \in \Re$, where w is the function operating on a given predictor variable. The variable importance of the dependent variable is the standard deviation of the function $q_w : \Re \to \Re$,[45] denoted by $Q_w \in \Re$:

$$Q_w = \sqrt{Var[q_w(w(x_i))]} = \left(\int_{\Re} (q_w(v) - \bar{q}_w)^2 \mathbb{P}r(w(x_i) = v) dv \right)^{\frac{1}{2}}, \tag{16}$$

where: $q_w = \mathbb{E}(f(X)|w(x_i) = v)$, $\bar{q}_w = \mathbb{E}[q_w(w(x_i))] = \int_{\Re} q_w(v)\mathbb{P}r(w(x_i) = v)dv$. Thus, the variable importance function determines the importance of a given variable x_i as the variability of the corresponding expected score q_w.[10,45]

Partial dependence (PD) function. The partial dependence of $\hat{f}$ on predictors $x_s \subset X$ is defined[11,46] as:

$$\hat{f}_s(x_s) = \mathbb{E}_{x_s^c}\left[f(\{x_s, x_s^c\}, y)\right] = \int f(\{x_s, x_s^c\}, y) p(x_s^c) d(x_s^c); \tag{17}$$

where $x_s^c \subset x$ of dimension $l < n$ is the complement of $x_s | x_s \cup x_s^c = x$ and $p(x_s^c)$ is the marginal probability density of the variables in x_c,[46] and $d(x_s^c)$ is a differential function that allows us to break the dependence function into small increments whose contribution are to be considered to the overall integration process. In other words, the *PD* function shows the marginal effect of 1 or more chosen predictors in the set x_s on the model's output $\hat{y}$. It is computed by taking the expectation of $\hat{y}$ with respect to x_s^c, by varying values of $x_s \in X$ while fixing other variables in x_s^c and observing the estimate $\hat{y} = f(\{x_s, x_s^c\}, y)$. According to Friedman,[46] letting $p(x_s, x_s^c) = p(x)$ to be the joint probability density over of all predictor variables in x gives:

$$p(x_s^c) = \int p(x_s, x_s^c) dx_s. \tag{18}$$

Although $p(x_s^c)$ is unknown, the PD function in Eq. 17 can be approximated using Monte Carlo[47] for a single tree[46] as

$$\bar{f}_s(x_s, y) = m^{-1}\sum_{i=1}^{m} f(\{x_s, x_s^c\}, y); \tag{19}$$

where m is the number of instances in the bootstrapped sample $\mathbb{D}_i^*$ such that $\hat{f}_s$ is the average over the K trees for the RF:

$$\hat{f}_s(x_s, y) = K^{-1}\sum_{i=1}^{K} \bar{f}_k(\{x_s, x_s^c\}, y). \tag{20}$$

Friedman[46] argued that taking the marginal distribution instead of the conditional distribution preserves the additive structure in $\hat{f}(x, y)$, such that Eq. 20 is able to recover the components of an additive function up to a certain constant[48]:

$$f(x, y) = f_s(x_s, y) + f_c(x_s^c, y). \tag{21}$$

Correlation Analysis

COVID-19 disease exhibits temporal dynamics in which interventions and their impacts change over time. Thus, we divided the data into batches of differing time frames and then used the Spearman's rank correlation coefficient[49] to identify the strength and direction of the nonlinear relationship between numerical variables based on the ranks of the data points as:

$$\rho = 1 - \frac{6\sum q_i^2}{m(m^2-1)} | -1 \le \rho \le 1; \tag{22}$$

where $q_i = rank(x_1) - rank(x_2)$ is the difference between the ranks of m paired variables (x_1, x_2) for $i = 1, \ldots, m$. In this context, $\rho = 1$ signifies a perfect monotonically increasing relationship; $\rho = -1$ depicts a perfect monotonically decreasing relationship, whereas $\rho = 0$ signifies that the variables considered are not related in a consistent monotonic manner. As a rule of thumb, redundant variables have correlation values greater or smaller than a threshold ϵ such that $\epsilon = \pm 0.95$.[14] In this work, we use the correlation matrix to gain insights on the nonlinear relationships between the predictors and target variable.

Our design method makes use of a dynamic rolling window approach, in which the RF ensemble is incrementally updated with new modeling data and parameters as the time frame progresses. A window is basically a set of labeled instances defining a context of interest. In our COVID-19 case, the rolling window approach can be considered a form of cross-validation, in which the RF ensemble is trained and validated on different subsets (i.e., by adding or dropping data) from different time frames.

Our tailored algorithm embodies the following ideas from[5] 1) modification of the concept description in response to changes in the contents of the window, 2) decision on when and how many old instances to include or delete from the window, and 3) assessment of the relative merits of concept hypothesis. We refer to our strategy as sequential learning with cross-validation (SLCV). The SLCV is useful when evaluating the model's performance on temporal (time-dependent) data over time. It allows ML practitioners to manually dissect, simulate, and examine how the model would perform when trained and adapted to new data. We use human feedback and annotations to identify the occurrences of context change in our data.

This research methodology received no external funding.

Results

The main question was whether KIL is a viable avenue for bridging the gap between pure ML (i.e., RF) models and dynamic epidemic models (i.e., SIRVH). This is vital to inform decision support systems for ML practitioners, policy developers, and decision makers. The objective was to inform the RF using outputs from the SIRVH model (i.e., the effective reproductive rate) then examine the appropriate loss function to assess the IRF model. This section showcases the performance of the IRF on 2 random data frames; later on, the process is generalized to other data frames.

A data set of m instances was divided into 2 sets[36]: a training set and a validation set, based on the predetermined splitting ratios: $[D_{Training}|D_{Validation}) = \{[75\%|25\%] \text{ or } [80\%|20\%], [85\%|15\%], [90\%|10\%]\}$. Two-thirds of the observations were used for training, and one-third of the observations in the training data were left out of the bootstrap sample to serve as a test set (out of bag).[8] Table 6 shows the results of using 0.75, that is, 75% for training and testing (i.e., $\mathbb{D}_1 = \frac{2}{3} * m$ and $\mathbb{D}_2 = \frac{1}{3} * m$) and 25% for validation ($\mathbb{D}_3$) using data from February 17, 2020, to August 10, 2020, where $\mathbb{D}_1$ is the training set, $\mathbb{D}_2$ is the testing set, and $\mathbb{D}_3$ is the validation set. The IRF attained $R^2 = 0.7$, $RMSE = 12.6$, and $MAE = 5$ in the training phase.

Figure 4 shows that the IRF attained $R^2 = 0.2$, $RMSE = 21.3$, and $MAE = 8.8$ in the validation phase. Figure 5 shows the results of using 0.75 as a splitting ratio on data from June 19, 2021, to December 31, 2021. Figure 5 shows that the IRF attained $R^2 = 0.9$, $RMSE = 552.8$, and $MAE = 298.4$ on training and

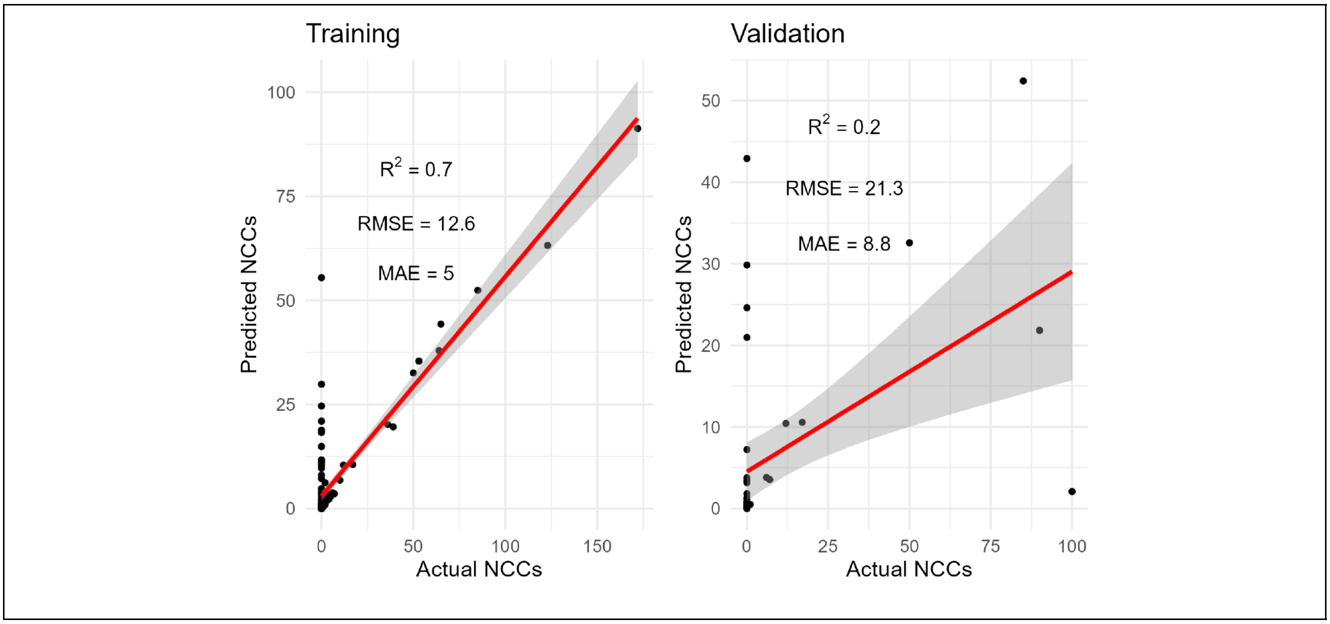

Figure 4 Outputs of the informed random forest on training and validation using the data from February 17, 2020, to August 10, 2020.

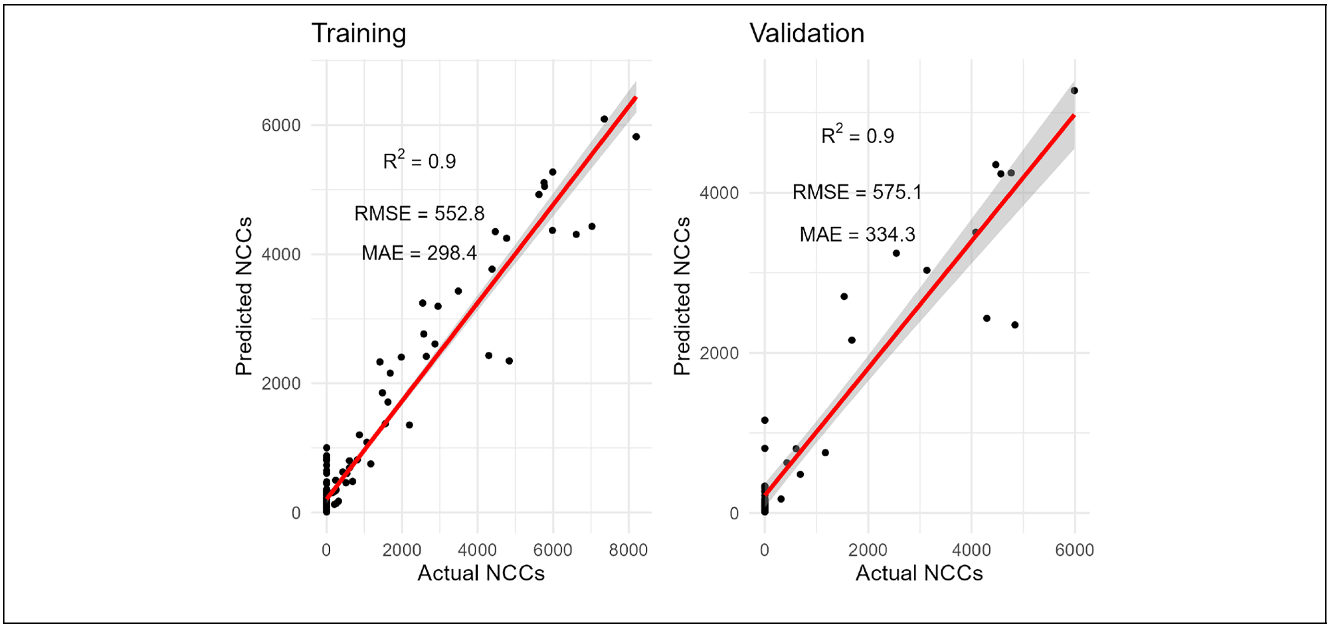

Figure 5 Outputs of the informed random forest on training and validation using data from June 19, 2021, to December 31, 2021.

R^2 = 0.9, *RMSE* = 575.1, and *MAE* = 334.3 on validation. It can be noted that the variability of the predictions is higher on the first data frame (see Figure 4) than on the second one (see Figure 5).

Using the evidence in Figures 4 and 5, it is impossible to tell whether PIML (i.e., the effective reproductive rate) can improve the performance of a default RF. To further interpret the model outputs, the partial dependence

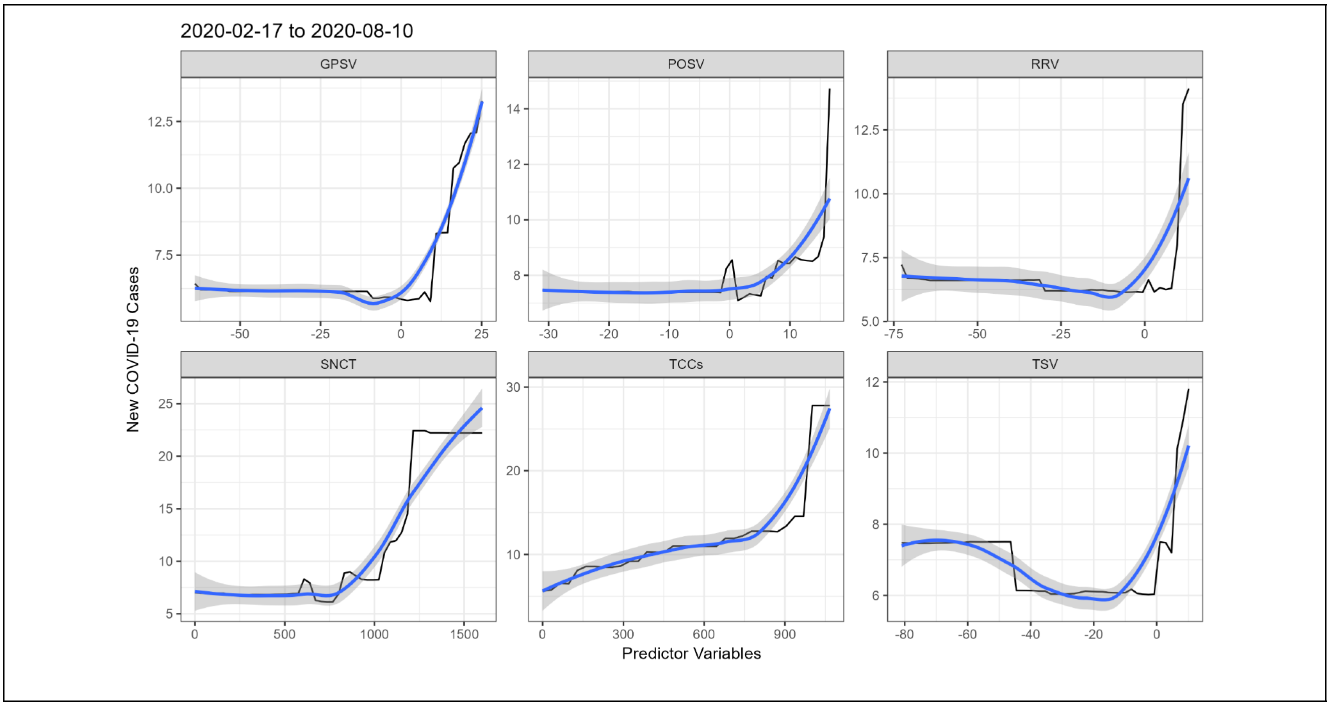

Figure 6 The dependency of predicted new COVID-19 cases (NCCs) on grocery and pharmacy (GPSV), parks and outdoor spaces (POSV), retail and recreation, COVID-19 tests (SNCT), total COVID-19 cases (TCCs), and transit stations (TSV).

functions are next used to depict relationships between predictors and the target variable, and the process is later on generalized to various data frames.

Dependence of COVID-19 Cases on Adopted Policies and Public Mobility Patterns

Figures 6 and 7 show the top 6 predictor variables of NCCs for 2 randomly selected batches (February 17, 2020, to August 10, 2020 and June 19, 2021, to December 31, 2021). The black line is the PDP trend, which illustrates a fluctuating monotonic relationship between the predictors and NCCs. The blue line shows the smoothed relationship between the predictor and the NCCs. The shaded region around the blue line represents confidence intervals showing the uncertainty in the partial dependence of the predicted NCCs for a particular predictor; a wider shaded region indicates more uncertainty.

Figure 6 reveals that the RF predicts on average low numbers of NCCs when high stringency measures such as lockdown (indicated by negative values) were adopted to restrict visits to public places like groceries and pharmacy stores (GPSV), parks and outdoor spaces (POSV), retail and recreation spaces (RRV), and transit stations (TSV). However, the number of predicted NCCs increased when restrictions on public mobility were relaxed. Likewise, the more COVID-19 tests (SNCT) were conducted, the more infectious individuals were recorded, which makes sense. Not surprisingly, the number of new cases (NCCs) predicted by the RF increased as the total number of infected individuals (TCCs) increased.

Figure 7 shows that a decrease in the values for mobility indices (i.e., visits to grocery and pharmacy store [GPSV], transit stations [TSV], retail and recreation [RRV] is associated with a decrease in the predicted NCCs, but later on, the predicted NCCs increase as these indices increase, and these agree with the reproductive rate (R), as also illustrated in Appendix 3.

Generalization of the Informed RF to Different Time Frames

Figure 8 shows a visual and informative way to intuitively assess the performance of the RF on validation using different time frames. Results show that the RF has captured a significant portion of the underlying relationship between the predictors and the target variable (NCCs). The closeness of the black points to the red line suggests that the IRF's predictions are accurate and

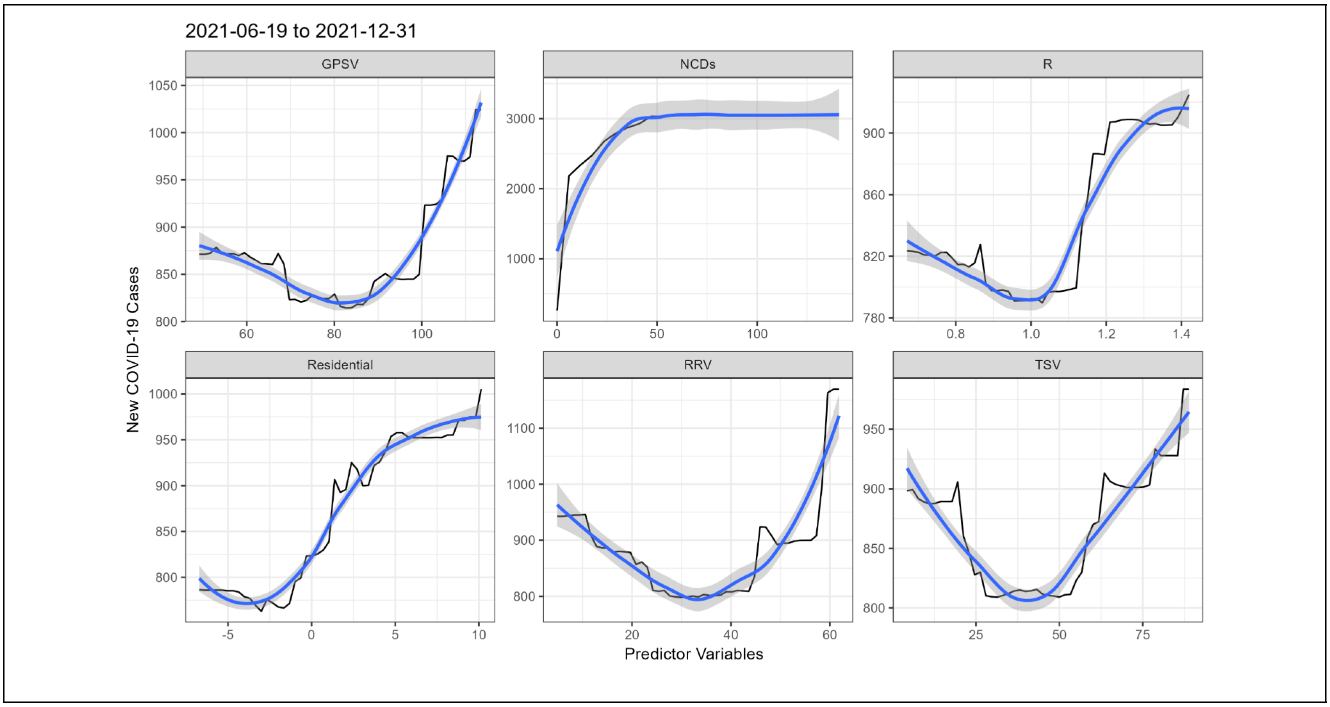

Figure 7 The dependency of the predicted new COVID-19 cases on visits to grocery and pharmacy (GPSV), new COVID-19 deaths (NCDs), effective reproductive number (R), residential, retail and recreation (RRV), and transit stations (TSV).

consistent across the entire time span of the data, with the IRF attaining R^2 values greater than or equal to 0.7 (except for batch 2). This indicates that the IRF's predictions are very close to the true values; thus, the IRF is a good fit for our data.

Thus, to assess the overall fit of the RF model, we focused on the R^2 and MAE metrics using 4 different splitting ratios (i.e., 0.75, 0.80, 0.85, and 0.95). To select the optimal splitting ratio, we used the percentage hits gained by the default RF per data batch. A hit is defined as a prediction with a score that is greater than or equal to a given threshold. We set our threshold as 0.7[14] on a scale from 0 to 1, thus penalizing $R^2 \leq 0.7$. The model producing more hits is considered better.

Table 3 shows that the default RF performed better when using 0.75 as a splitting ratio by yielding 93%; however, using 0.90 as a splitting ratio yielded the lowest performance (61%). In general, the default RF performs better (84%) when including the effective reproductive rate $(\hat{\mathcal{R}})$ from the SIRVH model, compared with (75%) when this variable $(\hat{\mathcal{R}})$ is not included in the data set. This information indicates that the PIML strategy applied to the RF ensemble helped to improve its predictive performance. To further interpret outputs generated by the IRF, we provide a visual assessment of the impacts of policies and public mobility responses on the predictability of new COVID-19 cases using variable importance functions (see the next section).

Impact of Adopted Policies and Public Mobility Patterns on COVID-19 Cases

Figure 9 shows that the stay-home requirements (SHR) variable, having a variable importance value of 8.75, is the most important when predicting new cases (NCCs) between February 17, 2020, and June 19, 2020. Thus, changes in SHR have the most significant impact on the RF's predictions. The second variable is total COVID-19 cases (TCCs) with 6.28, followed by positive rate (PR) with 5.31. Meanwhile, closing public transport (CPT) with 3.76 and parks and outdoor spaces (POSV) with 3.13 were ranked in the 9th and 10th positions, respectively. Likewise, POSV was the most important variable in the second batch (February 17, 2020, and June 19, 2020) with 5.72, followed by smoothed new COVID-19 tests (SNCT) with 5.23 and transit station visits (TSV) with 4.99, while PR with 3.73 and total COVID-19 deaths (TCDs) with 2.25 were ranked 9th and 10th, resepectively.

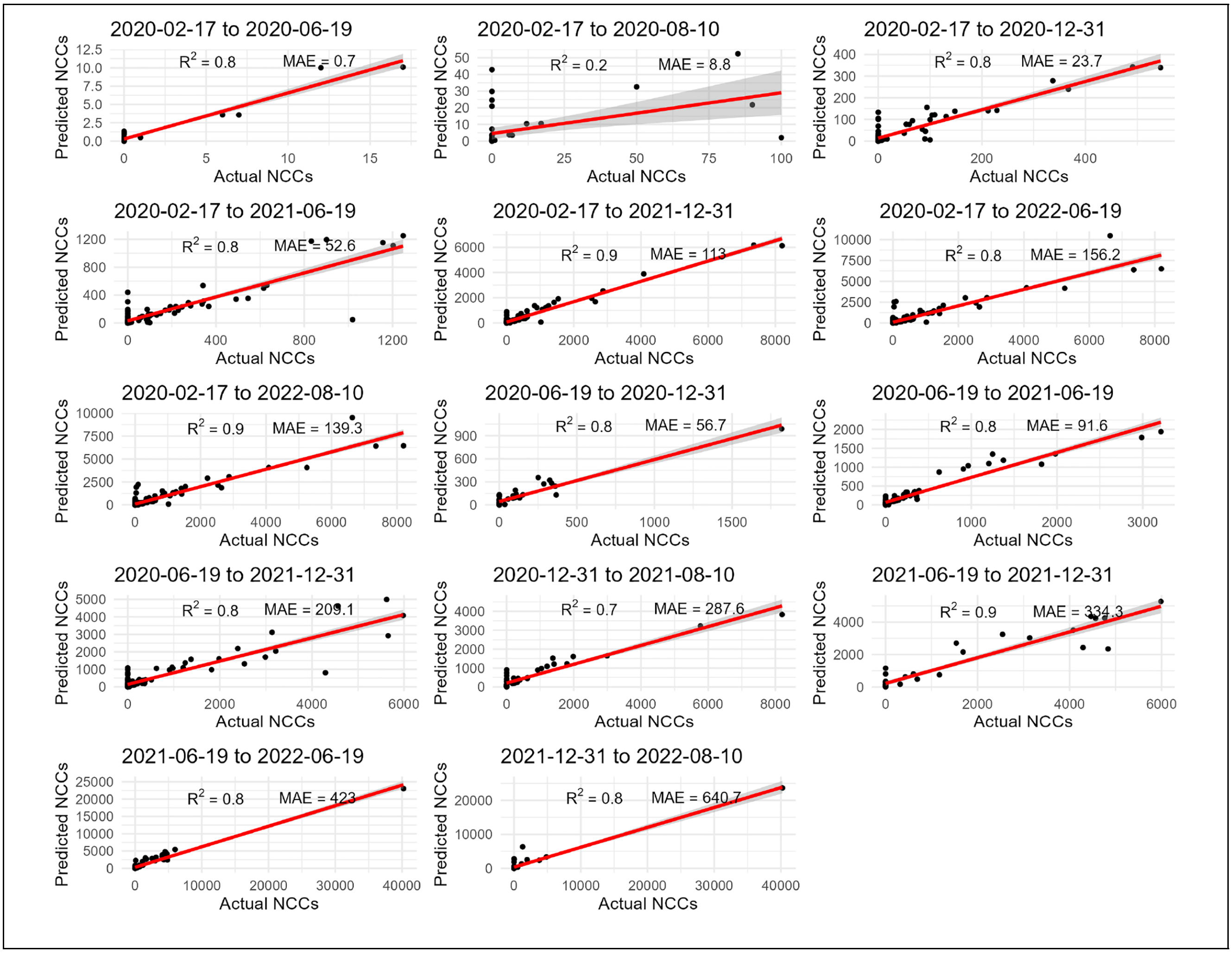

Figure 8 Report on the outputs of the informed random forest model on the validation data using the selected 14 data batches.

Table 3 Aggregated Results of the Random Forest ($\mathbb{K} = 500$, $\omega = 10$, $\phi = 5$) with and without the Effective Reproductive Rate ($\hat{\mathcal{R}}$)

Splitting Ratio	Without $\hat{\mathcal{R}}$	With $\hat{\mathcal{R}}$	Overall Performance	Ranking
0.75	93%	93%	93%	1
0.80	79%	93%	86%	2
0.85	71%	86%	79%	3
0.90	57%	64%	61%	4
Overall performance	75%	84%		

The COVID-19 deaths (NCDs) variable is a strong predictor for NCCs from batch 3 to 14. The prominence of NCDs in these batches suggests that it may contain valuable information for predicting NCCs during these time frames. However, looking at Table 10 and 11 are in the Appendices, we see that a number of COVID-19 waves occurred in these time frames, which were also accompanied by many COVID-19 deaths, and this

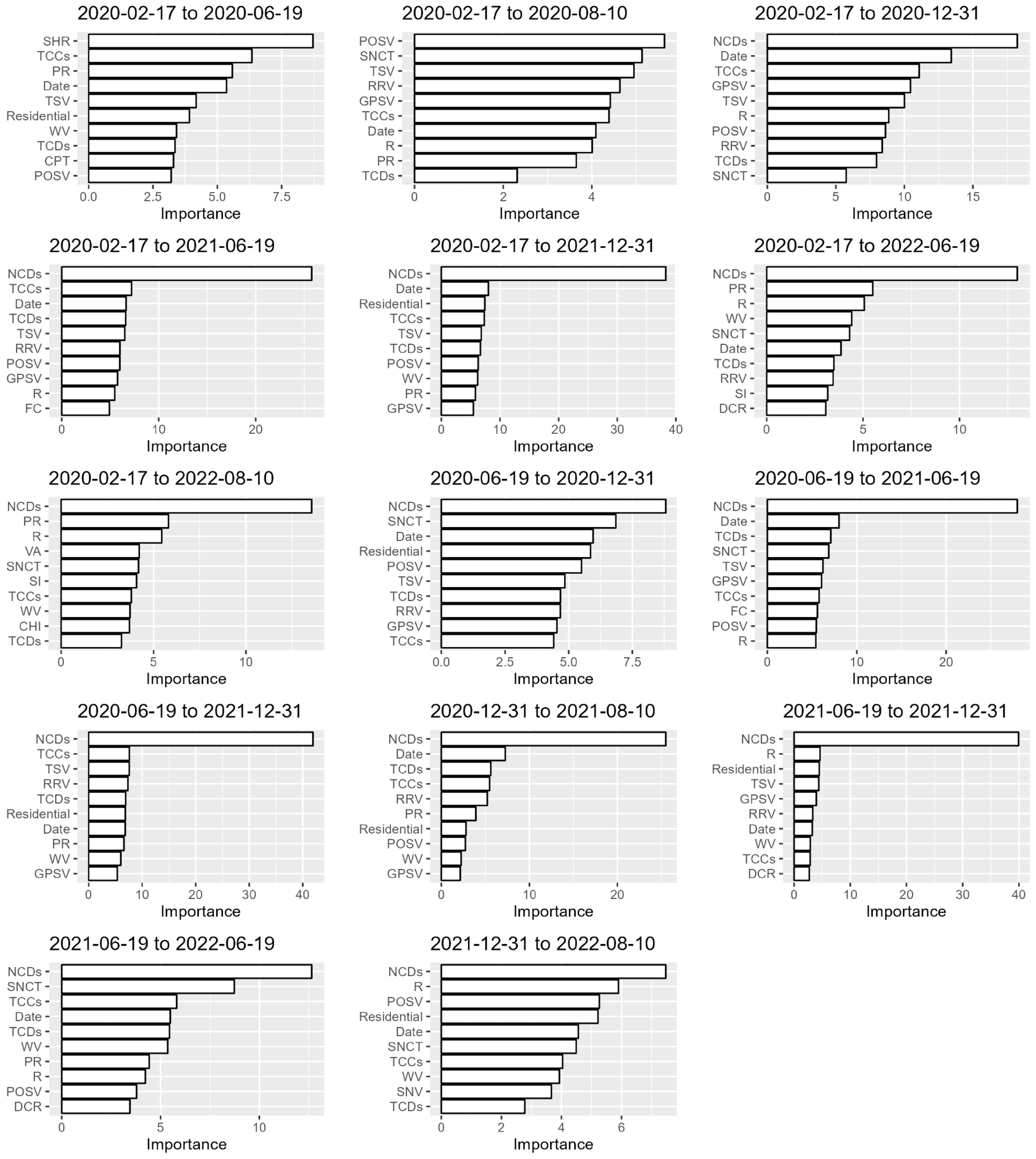

Figure 9 Importance of variables on predicting new COVID-19 cases (NCCs) with a random forest.

inflated other variables. The effective reproductive rate $(\hat{\mathcal{R}})$ was among the top 10 predictors except in batches 1, 5, 8, 10, and 11. It is important to note that variable importance scores are based on the correlation between the predictors and the target variable (NCCs) and that correlation does not imply causation. Moreover, these scores do not tell us about the impact of each predictor on NCCs; for that, we used the partial dependence functions.

Discussion

We have discussed a basic yet powerful approach of incorporating prior knowledge into a data-driven model. This approach 1) can be instantiated using any supervised ML algorithm. 2) has a post hoc phase that takes into account the effect of each predictor variable on the dependent variable (NCCs) in the model, 3) is consistent and uses the same interpretation regardless of the time frame it is exposed to, and 4) has the potential to help interpret nonlinear interaction effects in the presence of data drifts and outliers. In the end, we built an informed RF to predict the number of new COVID-19 cases for a given location (i.e., Botswana). We chose the RF model due to its ability to handle outliers[8,14] that may arise due to novel data drifts.

The approach used so far involves a dynamic and iterative process of training, testing, and validating the RF ensemble using the data sets $\mathbb{D}_1$, $\mathbb{D}_2$, and $\mathbb{D}_3$, respectively, over time. We fix the starting date, increasing time horizons, and moving the starting or final point into the future. Moving these points into the future implies that we are periodically updating the RF ensemble with unseen data streams, which can be of a different distribution. This technique helps the model in learning relationships within the data, allowing it to make predictions based on relevant data. This approach has potential to aid in assessing how well the RF generalizes to unseen data streams and whether it provides accurate predictions in the presence of data drifts.

Research Contribution

We proposed a reproducible framework for KIL that integrates ML and epidemic model outputs to examine government responses and public responses. This work extends the existing evidence on epidemiology by addressing both prediction and inferences using observational biases (i.e., multiple data sources) and induction biases (feature engineering to generate variables such as SI, CHI, and PR) for data aggregation. Following the evidence in YeŞİlkanat,[36] we note that the default RF performs well in predicting COVID-19 cases with an average of 75%. Using outputs from the SIRVH model (i.e., $\hat{\mathcal{R}}$) significantly improved the performance of the default RF to obtain an average performance of 84%, clearly showing the benefits of KIL.

Besides confirming the power of KIL[7] in boosting a default RF ensemble, the steps taken in this work can be packaged into a reproducible framework to help current and future users of ML in identifying appropriate ways to use prior knowledge in mitigating ML challenges (i.e., handling noisy data, missing data, and insufficient data) for policy analysis and assessment.

Figure 7 shows that between February 17, 2020, and August 10, 2020, a decrease in mobility indices (i.e., visits to grocery and pharmacy store [GPSV], transit stations [TSV], and retail and recreation [RRV]) was associated with a decrease in the predicted NCCs, but later on, the predicted NCCs increased as these indices increased. The same applies to time period between June 19, 2021, and December 31, 2021, during which an increase in the values for mobility indices (i.e., visits to transit stations and residential, retail, and recreation, see Figure 11; parks and outdoor spaces, see Figure 10) was associated with an increase in the predicted NCCs. This can be attributed to unpredictable public health measures and public responses.

In terms of previous published work on the examination of adopted government policies and their impacts on public mobility and COVID-19 cases, our findings corroborate studies from China,[50] South Korea,[51] and the United States,[52] which reported that stringent COVID-19 policies and mass vaccination campaigns reduced the number of daily COVID-19 new cases and deaths. Such hypotheses are also made for populated countries such as Brazil, China, India, and the United States[53] as well as for low- and middle-income countries (Botswana, India, Jamaica, Mozambique, Namibia, and Ukraine).[54] This work extends the study by Lane et al.[54] by exploring analyses of up to 906 (they used 100) days from the onset of COVID-19.

This study contributes to ongoing debates on the effect of parameter tuning or estimation in the context of RF. Bentéjac et al.[28] compared the performance of the XGBoost, LightGBM, CatBoost, and the RF on 28 different data sets using default and tuned settings and found that default versions of the RF and CatBoost generated consistent and stable results compared with their tuned counterparts. Szczepanek[29] made the same claim after comparing default settings for XGBoost, LGBoost, CatBoost, and RF in forecasting daily stream flow in mountainous catchment. This study confirms the case of a default RF.

The use of MSE, RMSE, and MAPE seems to be less informative when the model is exposed to data drifts (COVID-19 waves). In our case, the RMSE can range from 0.3 to as large as 3075.1 (see Table 12 in the

Appendices), indicating that data drifts and outliers have a significant impact on RMSE.[14,20] Compared with RMSE, the MAE showed to be more robust to outliers, with values ranging between 0.2 and 834.5 (see Table 12 in the Appendices). We believe that MAE and R^2 are more informative and robust to data drifts.

Chicco et al.[20] argued that the MSE, RMSE, MAE, and MAPE are unsuitable for regression analysis.[55] Prior studies suggested the use of R^2 because "it considers the distribution of all the ground truth values, and produces a high score only if the model correctly predicts most of the actual values."[20(P. 17)] To corroborate such claims, we suggest that R^2 is ideal for evaluating the performance of the RF since both R^2 and the RF do not require any data transformations.[20] This results in a model that is quick to deploy, guided by prior knowledge, and has evidence to inform policy and decision making in emergency responses.

Lessons Learnt

There is no universal approach to best develop an informed model that is interpretable and efficacious. Different tasks require different approaches. The best strategy might vary across problems, and at times one needs to integrate different actors, tools, methods, and techniques. When fitting KIML models to data containing data drifts, it is essential to use specific expertise (domain knowledge) to inform the choice of appropriate data (observational bias), model structures (induction bias), and the choice and even modification of loss functions (learning bias) to enable model assessment and interpretation (see the "Related Literature" section).

Research Implications

- This work shows an efficacious way to combine epidemic model outputs with RF to accurately forecast the complex spatiotemporal dynamics of COVID-19 while capturing key model properties such as consistency, accuracy, and interpretability.
- The study contributes to health care by offering insights into relationships between the indices derived based on policies, public mobility, epidemiology, and economic interventions during a pandemic. It provides evidence-based information on the effectiveness of different interventions in reducing the spread of COVID-19 at a national level.
- The integration of ML (i.e., RF) and epidemic (i.e., SIRVH) models to examine COVID-19 interventions can contribute to knowledge and technology transfer from research into practice. This can enhance collaboration between ML researchers, health care providers, and policy makers to develop decision support systems for infectious diseases.[56]

Research Limitations

- The data used is an aggregation of indices from multiple sources, making it difficult to pinpoint data for specific locations (i.e., towns, villages, or districts). Moreover, we used only Google mobility data because they are openly available, making other potential sources to be ignored due to data access implications beyond the researchers' control. This is a challenge that may lead to selection bias, a scenario in which certain locations are overrepresented or underrepresented.
- We used only the RF model for predictions; other ML models of deep learning could be used to enable comparisons.
- This work is limited to the batch approach of building RF models in an offline fashion (i.e., using percentage splitting). The proposed IRF uses raw data as they are, which might not accommodate direct integration and comparison with ML models such as ANN and SVM, which require additional data transformation. Likewise, the consolidated information is from public documents, such as government gazettes reporting on the adopted policies; links to the sources could be modified by the time of publication. Finally, the domain knowledge was fuzzy, hence limiting the analyses for inferences. To compensate for this limitation, we collected the data from various sources to complement the extracted evidence.

Generalization, Reliability, and Validity

We used historical data comprising government stringency policies, public mobility indices, economic indices, and epidemiological priors. The distributions of the chosen data sets differ due to evolving factors (i.e., stringency policies and human behavioral changes). To assess external validity, we examined whether a default RF and an IRF can effectively predict the number of COVID-19 cases of different time frames. Specifically, we examined whether the predictions generated by the RF model remain accurate and consistent when applied to shorter and longer time series with different data distributions.

This study showed that both the default RF and the IRF can accurately and consistently capture variations and trends specific to selected data batches of time series with data drifts and outliers (see Figure 8). These findings can be generalized and applied to other domains such as weather forecasting and financial markets, in which the data distributions are characterised by stochasticity.

Conclusions

This study proposed an informed learning approach that integrates ML with an epidemic model to accurately depict, analyze, and infer the dynamics of infectious diseases. For that, we used a case study of COVID-19 disease, incorporating prior knowledge on epidemic modeling into a default RF using known assumptions, mathematical expressions, and equations. We used multidisciplinary indicators that include epidemiological indices, government policies, public mobility, and economic indicators to predict the number of COVID-19 cases. We used the effective reproductive rate to inform the RF and compare it with the default RF. We assessed the models to test and validate their feasibility, predictability, and consistency in the presence of data drifts. Experimental results revealed that both the default and IRF generate accurate and effective results that are consistent in capturing the hidden nonlinear relationships in the presence of stochasticity (i.e., COVID-19 waves). Interestingly, our approach can be easily transferred from research into practice without background knowledge on numerical analysis (i.e., stability conditions) and mathematical optimization. However, the proposed approach needs some technical knowledge of programming and detailed awareness of the functions to process fuzzy data for implementation of the appropriate models using available packages or libraries. This approach can be extended to perform parameter estimation of compartmental models for related infectious diseases. The proposed IRF to analyze and forecast time series provides extracts with potential to improve preparedness and response strategies to health outbreaks. Further investigations are needed (i.e., using transformed data and other ML models, using data from different countries) to strengthen the reliability and validity of this study. Most importantly, this work seeks to motivate multidomain collaboration to inform continuous development of ML-based decision support systems for tracking, monitoring, and assessing health outbreaks and their interventions.

Acknowledgments

We acknowledge Google, for mobility data; the World Health Organization, for health data; and the R community, for R packages. The data used in this work are available online. Any other material will be made available upon request to the first author. Moreover, the code and data will be made available on GitHub upon acceptance.

Authors' Contributions

Conceptualization, TST, DM, and GH; methodology, TST; validation, TST, DM, and GH; data analysis, TST; formal analysis, TST, DM, and GH; data curation, TST; writing—original draft preparation, TST; review and editing, TST, DM, and GH; visualization, TST; supervision, DM and GH. All authors have read and agreed to the final version of this article to be published.

ORCID iD

Tsaone Swaabow Thapelo https://orcid.org/0000-0002-1658-0498

Supplemental Material

Supplementary material for this article is available online at https://doi.org/10.1177/23814683231218716.

References

1. Stevenson F, Hayasi K, Bragazzi NL, et al. Development of an early alert system for an additional wave of covid-19 cases using a recurrent neural network with long short-term memory. *SSRN* 3838420, 2021. Available from: https://www.mdpi.com/1660-4601/18/14/7376
2. Ning X, Jia L, Wei Y, Li X-A, Chen F. Epi-DNNs: epidemiological priors informed deep neural networks for modeling COVID-19 dynamics. *Comput Biol Med.* 2023;158: 106693. DOI: 10.1016/j.compbiomed.2023.106693
3. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–60. DOI: 10.1126/science.aaa84
4. Schultz MG, Betancourt C, Gong B, et al. Can deep learning beat numerical weather prediction? *Philos Trans R Soc A*. 2021;379(2194):20200097. DOI: 10.1098/rsta.2020.0097
5. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. *Mach Learn.* 1996;23:69–101. Available from: https://link.springer.com/article/10.1023/A:1018046501280
6. Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. *Nat Rev Phys*. 2021;3(6):422–40. Available from: https://www.nature.com/articles/s42254-021-00314-5
7. von Rueden L, Mayer S, Beckh K, et al. Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans Knowl Data Eng*. 2021;35(1):614–33. DOI: 10.1109/TKDE.2021.3079836
8. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. Available from: https://link.springer.com/article/10.1023/a:1010933404324

9. Magal P, Ruan S. Susceptible-infectious-recovered models revisited: from the individual level to the population level. *Math Biosci*. 2014;250:26–40. DOI: 10.1016/j.mbs.2014.02.001
10. Greenwell BM, Boehmke BC. Variable importance plots: an introduction to vip. 2019. [cited December 14, 2023]. Available from: https://koalaverse.github.io/vip/articles/vip.html
11. Greenwell BM. pdp: an R package for constructing partial dependence plots. *R J*. 2017;9(1):421. Available from: https://journal.r-project.org/articles/RJ-2017-016/RJ-2017-016.pdf
12. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhyā: Indian J Stat*. 1973;35(4): 417–46. Available from: https://www.jstor.org/stable/25049893
13. Amit R, Meir R. Lifelong learning and inductive bias. *Curr Opin Behav Sci*. 2019;29:51–4. DOI: 10.1016/j.cobeha.2019.04.003
14. Dean A. *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. New York: John Wiley and Sons; 2014. Available from: https://dl.acm.org/doi/abs/10.5555/2670086
15. Sun Q, Pfahringer B, Mayo M. Towards a framework for designing full model selection and optimization systems. In: *International Workshop on Multiple Classifier Systems*. New York: Springer; 2013. p 259–70. Available from: https://link.springer.com/chapter/10.1007/978-3-642-38067-9_23
16. Savitha R, Ambikapathi AM, Rajaraman K. Online RBM: growing restricted Boltzmann machine on the fly for unsupervised representation. *Appl Soft Comput*. 2020;92: 106278. DOI: 10.1016/j.asoc.2020.106278
17. Wang H, Fan W, Yu PS, Han J.Mining concept-drifting data streams using ensemble classifiers. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM; 2003. p 226–35. DOI: 10.1145/956750.956778
18. Probst P, Wright MN, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019;9(3):e1301. DOI: 10.1002/widm.1301
19. Wang Q, Ma Y, Zhao K, Tian Y. A comprehensive survey of loss functions in machine learning. *Ann Data Sci*. 2022;9: 187–212. Available from: https://link.springer.com/article/10.1007/s40745-020-00253-5
20. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci*. 2021;7:e623. DOI: 10.7717/peerj-cs.623
21. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 2006;7:1–13. Available from: https://link.springer.com/article/10.1186/1471-2105-7-3
22. Lima TPF, Sena GR, Neves CS, et al. Death risk and the importance of clinical features in elderly people with covid-19 using the random forest algorithm. *Rev Bras Saúde Mater Infant*. 2021;21:445–51. DOI: 10.1590/1806-9304202100S200007
23. Ramírez García A, Jiménez Preciado AL. Covid-19 and economics forecasting on advanced and emerging countries. *EconoQuantum*. 2021;18(1):21–43. DOI: 10.18381/eq.v18i1.7222
24. Villagrana-Bañuelos KE, Maeda-Gutiérrez V, Alcalá-Rmz V, et al. Covid-19 outcome prediction by integrating clinical and metabolic data using machine learning algorithms. *Rev Invest Clin*. 2022;74(6):314–27. DOI: 10.24875/ric.22000182
25. Milicevic O, Salom I, Rodic A, et al. $PM_{2.5}$ as a major predictor of covid-19 basic reproduction number in the USA. *Environ Res*. 2021;201:111526. DOI: 10.1016/j.envres.2021.111526
26. Biau G, Scornet E. A random forest guided tour. *Test*. 2016;25(2):197–227. Available from: https://link.springer.com/article/10.1007/s11749-016-0481-7
27. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett*. 2010;31(14): 2225–36. DOI: 10.1016/j.patrec.2010.03.014
28. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev*. 2021;54:1937–67. Available from: https://link.springer.com/article/10.1007/s10462-020-09896-5
29. Szczepanek R. Daily streamflow forecasting in mountainous catchment using XGBoost, lightGBM and catboost. *Hydrology*. 2022;9(12):226. DOI: 10.3390/hydrology9120226
30. Lei C, Deng J, Cao K, et al. A comparison of random forest and support vector machine approaches to predict coal spontaneous combustion in gob. *Fuel*. 2019;239:297–311. DOI: 10.1016/j.fuel.2018.11.006
31. Aktay A, Bavadekar S, Cossoul G, et al. Google COVID-19 community mobility reports: anonymization process description (version 1.1). arXiv preprint arXiv:2004.04145, 2020. Available from: https://arxiv.org/abs/2004.04145
32. Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C. Tracking R of COVID-19: a new real-time estimation using the kalman filter. *PLoS One*. 2021;16(1): e0244474. DOI: 10.1371/journal.pone.0244474
33. Rodés-Guirao L, Appel C, Giattino C, et al. Coronavirus pandemic (COVID-19). *Our World in Data*. 2020. Available from: https://ourworldindata.org/coronavirus
34. Shao J, Zhong B. Last observation carry-forward and last observation analysis. *Stat Med*. 2003;22(15):2429–41. DOI: 10.1002/sim.1519
35. Blankers M, Koeter MWJ, Schippers GM. Missing data approaches in ehealth research: simulation study and a tutorial for nonmathematically inclined researchers. *J Med Internet Res*. 2010;12(5):e1448. DOI: 10.2196/jmir.1448
36. YeŞİlkanat CM. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest

machine learning algorithm. *Chaos Solitons Fractals*. 2020;140:110210. DOI: 10.1016/j.chaos.2020.110210

37. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning—a brief history, state-of-the-art and challenges. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. New York: Springer; 2020. p 417–31. Available from: https://link.springer.com/chapter/10.1007/978-3-030-65965-3_28
38. Song Y-Y, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*. 2015;27(2):130. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/
39. Jebb AT, Parrigon S, Woo SE. Exploratory data analysis as a foundation of inductive research. *Hum Resour Manag Rev*. 2017;27(2):265–76. Available from: https://www.sciencedirect.com/science/article/pii/S1053482216300353
40. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol*. 2004;160(6):509–16. DOI: 10.1093/aje/kwh255
41. Hale T, Angrist N, Goldszmidt R, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat Hum Behav*. 2021;5(4): 529–38. Available from: https://www.nature.com/articles/s41562-021-01079-8
42. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. Berlin: Spinger; 2013. DOI 10.1007/978-1-4614-7138-7
43. Meade N. Long range forecasting: from crystal ball to computer. *J Oper Res Soc*. 1986;37(5):533–5. DOI: 10.1057/jors.1986.91
44. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. New York: Springer; 2009. Available from: https://link.springer.com/book/10.1007/978-0-387-21606-5
45. Zien A, Krämer N, Sonnenburg S, Rätsch G. The feature importance ranking measure. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009,* Bled, Slovenia, Proceedings, Part II 20. New York: Springer; 2009. p 694–709. Available from: https://link.springer.com/chapter/10.1007/978-3-642-04174-7_45
46. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–232. Available from: https://www.jstor.org/stable/2699986
47. Raychaudhuri S. Introduction to Monte Carlo simulation. In: *Winter Simulation Conference*. New York, NY: IEEE; 2008. p 91–100. DOI: 10.1109/WSC.2008.4736059
48. Hooker G. Discovering additive structure in black box functions. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'04. New York, NY: Association for Computing Machinery; 2004. p 575–80. Available from: https://dl.acm.org/doi/abs/10.1145/1014052.1014122
49. Xiao C, Ye J, Esteves RM, Rong C. Using spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurr Comput*. 2016;28(14):3866–78. DOI: 10.1002/cpe.3745
50. Chan H-Y, Chen A, Ma W, Sze N-N, Liu X. COVID-19, community response, public policy, and travel patterns: a tale of Hong Kong. *Transp Policy*. 2021;106:173–84. Available from: https://www.sciencedirect.com/science/article/pii/S0967070X21000895
51. Arora AS, Rajput H, Changotra R. Current perspective of COVID-19 spread across South Korea: exploratory data analysis and containment of the pandemic. *Environ Dev Sustain*. 2021;23(5):6553–63. Available from: https://link.springer.com/article/10.1007/s10668-020-00883-y
52. Pan Y, Darzi A, Kabiri A, et al. Quantifying human mobility behaviour changes during the covid-19 outbreak in the united states. *Sci Rep*. 2020;10(1):1–9. Available from: https://www.nature.com/articles/s41598-020-77751-2#citeas
53. Dey SK, Rahman MM, Siddiqi UR, Howlader A, Tushar MA, Qazi A. Global landscape of COVID-19 vaccination progress: insight from an exploratory data analysis. *Hum Vaccin Immunother*. 2022;18(1):2025009. Available from: https://www.tandfonline.com/doi/full/10.1080/21645515.2021.2025009
54. Lane J, Means AR, Bardosh K, et al. Comparing COVID-19 physical distancing policies: results from a physical distancing intensity coding framework for Botswana, India, Jamaica, Mozambique, Namibia, Ukraine, and the United States. *Global Health*. 2021;17(1):1–12. Available from: https://link.springer.com/article/10.1186/s12992-021-00770-9
55. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: empirical comparisons. *Int J Forecast*. 1992;8(1):69–80. DOI: 10.1016/0169-2070(92)90008-W
56. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med*. 2020;26(7):1037–40.