2023 IEEE International Conference on Big Data (BigData)

DOI: 10.1109/BigData59044.2023.10386378

Combat COVID-19 at National Level using Risk Stratification with Appropriate Intervention

Xuan Jin
Crisis Strategy and Operations Group
Ministry of Health Singapore
School of Computing and Information Systems
Singapore Management University
Singapore
xuan_jin_from.csog@moh.gov.sg
xuan.jin.2021@msc.smu.edu.sg

Kar Way Tan
School of Computing and Information Systems
Singapore Management University
Singapore
kwtan@smu.edu.sg

***Abstract*—In the national battle against COVID-19, harnessing population-level big data is imperative, enabling authorities to devise effective care policies, allocate healthcare resources efficiently, and enact targeted interventions. Singapore adopted the Home Recovery Programme (HRP) in September 2021, diverting low-risk COVID-19 patients to home care to ease hospital burdens amid high vaccination rates and mild symptoms. While a patient's suitability for HRP could be assessed using broad-based criteria, integrating machine learning (ML) model becomes invaluable for identifying high-risk patients prone to severe illness, facilitating early medical assessment. Most prior studies have traditionally depended on clinical and laboratory data, necessitating initial clinic or hospital evaluations. None of these studies incorporated vaccination status, a crucial variable in a well-vaccinated population. This paper proposes a machine learning approach to nationwide risk stratification, offering intervention recommendations by harnessing nationwide datasets. Our best-performing ML model, XGBoost achieves an AUROC of 0.930 utilizing data from multiple data sources including patients' demographic information, vaccination status and medical history. For broader applicability, we also propose a parsimonious XGBoost model with an AUROC of 0.885 with a selection of five commonly collected variables, namely age, number of vaccine doses taken and number of days since the first, second and booster doses. Importantly, both of our proposed models achieve robust predictive performance without requiring the collection of clinical or laboratory data from patients. We believe that the parsimonious model, leveraging easily attainable data, has the potential for broader adoption across diverse nations, ultimately delivering paramount value to their populations.**

***Index Terms*—predictive analytics, public health, decision-support, risk stratification, COVID-19 pandemic.**

I. Introduction

Coronaviruses (CoV) are a large family of viruses causing illnesses ranging from the common cold to pneumonia. COVID-19, which is an infectious disease caused by a strain of coronavirus (SARS-CoV-2), was declared a global pandemic by World Health Organization (WHO) in March 2020. Since its identification, COVID-19 has caused immense strain on healthcare systems around the world. Governments and healthcare administrators faced a challenging task in containing the spread of SARS-CoV-2. As of June 2022, Singapore has reported a cumulative total of 1,397,000 cases and 1,400 deaths [1]. To effectively navigate this pandemic and ensure the populace's health, informed policy-making by leveraging the insights from big data to steer decision and policy-making processes could mitigate risks for high-risk patients, and deliver paramount value to citizens.

The Ministry of Health Singapore (MOH) has been adapting its policies and modus operandi towards managing COVID-19 as appropriate since the start of COVID-19 outbreak by responding to the evolving situations. In the earlier stage, all COVID-19 patients were required to be quarantined at community care facilities or admitted into hospitals to limit the spread of the disease. With stringent contact tracing and isolation of patients, Singapore managed to keep a relatively low fatality rate. However, with the number of COVID-19 cases began to surge in September 2021, there was a need to triage the patients to better make use of the limited healthcare resources. As more residents were vaccinated against COVID-19 and over 98% of the infected individuals had no or mild symptoms, the Home Recovery Programme (HRP) was introduced to allow low-risk patients to recover at home starting from September 2021 [2]. Since then, HRP had become the default recovery plan for most patients, except for those who have a higher risk of developing severe illness. A broad-based approach using standard questionnaires was used by the Telemedicine Allocation Reconciliation System (TMARS) to identify high-risk patients using a few screening criteria. The high-risk patients were given priority to be further assessed by doctors to evaluate if they were suitable for HRP or required hospital admission [3].

To further support the identification of high-risk patients, we investigate the use of machine learning (ML) models to predict if individual patient will develop severe COVID-19 illness. Severe illness related to COVID-19 is defined as the need for mechanical ventilation, ICU admission or death. Numerous studies in the literature have demonstrated that ML models can predict COVID-19 severe illness reasonably well

using clinical and laboratory data. However, our key challenge was that clinical and laboratory data was not commonly collected for most patients in Singapore. For instance, patients can self-administer an Antigen Rapid Test (ART) to test for COVID-19 infection so visiting a doctor or laboratory test was not always required. Moreover, to our best knowledge, the existing studies had not included patients' vaccination status, hence their models may not be suitable for a highly vaccinated population such as Singapore (more than 90% of the total population vaccinated as of June 2022). Therefore, we developed several ML models utilizing patients' demographic details, vaccination status, and medical history to predict severe illness. We found the best-performing model to be an XGBoost model which has good performance scores using AUROC score, accuracy, sensitivity and specificity of 0.930. 0.958, 0.542, 0.965 respectively. To improve applicability and adoption, we further train a parsimonious XGBoost model using a set of five commonly collected variables, including age, number of vaccine doses taken, and number of days since the first, second and booster doses when a patient is tested positive for COVID-19. The parsimonious XGBoost model has decent performance with AUROC score, accuracy, sensitivity and specificity of 0.885, 0.853, 0.731, 0.855 respectively. A parsimonious XGBoost model is less likely to be affected by data availability issues and can be more easily adopted by other countries.

In this paper, our contributions to the realm of big data are twofold. Firstly, we unveil the untapped potential residing in the utilization of vast national datasets to reinforce public health efforts during pandemics, focusing on both the expansive volume and intrinsic value encapsulated within big data. Our analysis sheds light on the dynamic interplay between data quantity and its intrinsic worth, accentuating the far-reaching impact of harnessing such comprehensive information at the national level. Secondly, we aspire that the insights gleaned from our study, utilizing data at the national scale, serve as a catalyst for other nations in their pursuit of establishing effective mechanisms for safeguarding public health. By sharing our learnings and findings, we hope to contribute to the advancement of global public health management, bolstered by data-driven strategies that prioritize the well-being of populations worldwide.

II. Literature Review

A wealth of existing research has delved extensively into several critical aspects surrounding the COVID-19 SARS-CoV-2 virus, contributing to our understanding and ability to safeguard public health. These studies have encompassed a range of research inquiries, including evaluating the clinical impact of the virus on human health, gauging the effectiveness of vaccines, and harnessing machine learning models to predict the likelihood of an individual developing severe illness.

A comprehensive review by Gallo Marin et al. [4] indicates that demographic features, comorbidities, clinical features and laboratory biomarkers that are associated with COVID-19 severity. For instance, older patients and male patients are associated with higher severity risk. Patients with pre-existing conditions or comorbidities, such as cardiovascular disease, chronic kidney disease, chronic lung disease, diabetes mellitus, hypertension, immunosuppression, obesity, and sickle cell diseases have increased risk of requiring mechanical ventilation and mortality. Clinical features such as hypoxia (low oxygen saturation) and specific chest radiography images are also linked to the development of severe disease due to infection of the SARS-CoV-2 virus. In addition, certain clinical and laboratory biomarkers, such as elevated D-dimer levels, C-reactive protein (CRP), lactate hydrogenase (LDH), erythrocyte sedimentation rate (SER), neutrophil-to-lymphocyte ratio (NLR), and high-sensitivity cardiac troponin are associated with worse clinical outcomes.

With the introduction of COVID-19 vaccines, a few studies had looked into the effectiveness of the vaccines and the waning of protection against severe COVID-19 illness. The first dose of BNT162b2 vaccine was found to protect against severe illness by the third week and the second dose provides significant protection within the first two months [5]. The effectiveness of two doses of BNT162b2, mRNA-1273 and ChAdOx1-S vaccines against severe illness started to decrease after 20 weeks [6] [7]. The waning effect is greater in adults above 65 years old and younger adults with underlying medical conditions.

Besides research studies that investigate the relationship between indicators and disease severity, some studies also proposed to leverage ML models to predict whether a patient will develop severe illness. Wollenstein-Betech et al. [8] built five classifiers with Logistic Regression (LR), sparse versions of LR, Support Vector Machines (SVM), Random Forest, Gradient Boosted Trees (XGBoost) with data from Brazil to predict disease severity. The authors reported area under the receiver operator characteristic curve (AUROC) in the range of 0.786–0.792 and accuracy of 0.713–0.720 for mortality. The study also reported an AUROC of 0.694–0.695 and accuracy of 0.761–0.766 for predicting whether a patient needs mechanical ventilator support. Kang et al. [9] developed a neural network model using TensorFlow to predict severe illness using data from China. Feature selection using correlation analysis was first performed to find the indicators that have a strong correlation with serious illness, to reduce the number of features from 33 to 6. A predictive model was then built with an input layer (six nodes), two hidden layers (13 units each) using rectified linear unit (ReLU) activation function, and an output layer node using Sigmoid activation function. The author reported a good prediction performance with AUROC of 0.889–0.982, sensitivity of 1.0 and specificity of 0.857. The research work by Quiroz-Juárez et al. [10] used data from Mexico to develop a neural network (NN) algorithm with two sigmoid neurons in a single hidden layer and two softmax neurons in the output layer to predict patient mortality. They have shown that NN model is able to achieve better prediction in general as compared to LR, SVM and k-nearest neighbour (kNN) at four different clinical stages (i.e., Stage 1:

initial medical assessment; Stage 2: confirmed being COVID-19 positive; Stage 3: decision point between hospitalisation or home recovery; Stage 4: intubated or in intensive care (ICU). The NN algorithm has accuracy of 0.843–0.935, sensitivity of 0.863–0.961, specificity of 0.824–0.909. Another research by Ryan et al. [11] focused on predicting the need for mechanical ventilation and patient mortality at clinically useful windows of 12, 24, 48 and 72 hours in advance using XGBoost algorithm and data from USA. The authors reported an AUROC of 0.75–0.82, accuracy of 0.595–0.668, sensitivity of 0.803–0.805, and specificity of 0.553–0.647 for predicting the need for mechanical ventilation, AUROC of 0.862–0.910, accuracy of 0.771–0.818, sensitivity of 0.818–0.826, and specificity of 0.760–0.816 for mortality prediction. Lam et al. [12] demonstrated that ML algorithm can outperform policy-based criteria in predicting severe illness using data from USA. Their XGBoost model has AUROC of 0.88, accuracy of 0.85, sensitivity of 0.80 and specificity of 0.95. Laatifi et al. [13] showed that Uniform Manifold Approximation and Projection (UMAP) is an effective dimension reduction technique for improving the prediction results of ML classifiers such as XGBoost, AdaBoost, Random Forest and ExtraTrees, using data from Morocco. They showed that UMAP can significantly improve ML predictions to an AUROC of 1.0, accuracy of 0.98–1.0, and sensitivity of 1.0 and specificity of 0.97–1.0. While the performance appeared to be positive, we noted that this study has a relatively small sample size of 340 which its validity needs to be further evaluated.

While prior literature has highlighted the potential of ML algorithms in predicting disease severity with commendable accuracy, these studies invariably relied on clinical and laboratory data for their predictive models. These approaches necessitate patient evaluation within medical facilities, which is not always feasible for nationwide implementations. In addition, a significant portion of COVID-19 cases involves mild symptoms, leading individuals to opt for home recovery. Furthermore, the existing ML models lack integrated analysis using vaccination status or vaccine waning effects in their predictive frameworks. In addressing these limitations, our pioneering ML models stand out as unique. By harnessing nationwide big data, our model derives predictions from patients' demographic profiles, medical histories, and vaccination status – all without the need for clinical or laboratory inputs. This innovative approach not only fills existing gaps but also aligns seamlessly with the context of highly vaccinated countries such as Singapore, potentially offering value to diverse nations.

III. Materials and Methods

A. Data Sources

The data used in this study was collected and provided by Ministry of Health Singapore. Patients below 18 years old were excluded in this study. It includes 316,000 patients who were infected with COVID-19 between 1 Dec 2021 and 30 Apr 2022. The list of variables used in this study is summarised in Table I. Only patients who saw doctors via telemedicine services (about 130,000 patients) had their pre-existing conditions and pregnancy status recorded by the doctors.

B. Data Exploration

The dataset exhibited a pronounced imbalance, with only 1.1% of COVID-19 patients progressing to severe illness. In terms of demographic distribution, 50.8% were female, while 49.2% were male. Among ethnic backgrounds, the representation comprised Chinese (68.3%), Malay (15.3%), Indian (12.1%), and other races (4.2%). The study encompassed COVID-19 patients aged between 18 and 114 years. Fig. 1 illustrates the trend of the COVID-19 infection and severity across age groups. The graph shows positive case numbers on the left y-axis, while the right y-axis denotes the proportion of cases evolving into severe instances. Fig. 2 shows the number of serious illness cases by age and the cumulative distribution function plot. We observed a distinct trend whereby older patients exhibited a heightened propensity for severe illness, and they made up the majority of such cases. This pattern became particularly prominent beyond the age of 60. For noting, ages beyond 100 recorded few instances of zero in Fig. 1 due to a lack of cases in that age group.

TABLE I: List of Variables

Variable type	Variable name	Description
Demographic information	Age	Patient's age: between 18 and 114 years old
	Gender	Patient's gender: 'Male' or 'Female'
	Race	Patient's race: 'Chinese', 'Malay', 'Indian' or 'Other' races
Pre-existing conditions and pregnancy status	Hospitalisation	Patient has any disease affecting their heart, lungs, kidneys, liver or brain that required hospital admission in the last 6 months
	Diabetes or hypertension	Patient has Diabetes Mellitus or Hypertension
	Weaken immune	Patient has any disease or taking medications that weaken the immune system
	Cancer	Patient has been diagnosed with cancer before
	Dialysis	Patient is on dialysis
	Organ transplant	Patient had organ transplant surgery
	Weight above 100kg	Patient weighs more than 100kg
	Pregnant	Patient is pregnant
Medical history	ICD-10 code	Patient's medical history encoded in 3 digit ICD-10 codes
Vaccination status	Number of vaccine dose	Total number of doses of vaccine taken before tested positive for COVID-19
	Number of days from first dose	Number of days between first dose of vaccine and tested positive for COVID-19
	Number of days from second dose	Number of days between second dose of vaccine and tested positive for COVID-19
	Number of days from third dose	Number of days between third dose of vaccine and tested positive for COVID-19

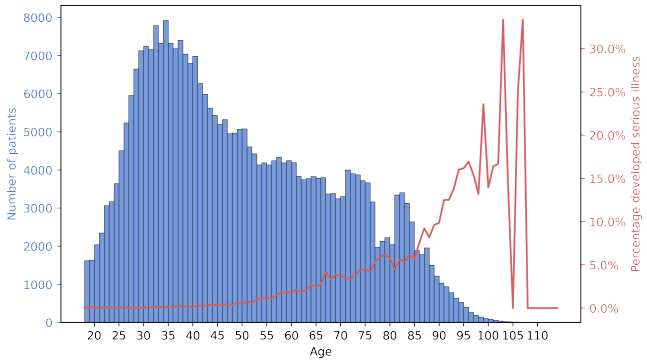


Fig. 1: Age profile and percentage of patients who developed serious illness in each age group

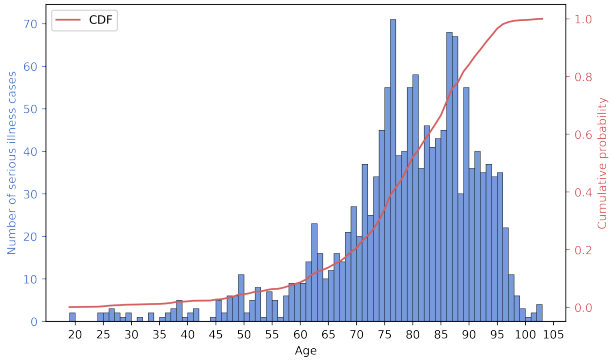


Fig. 2: Number of serious illness cases by age and cumulative distribution function

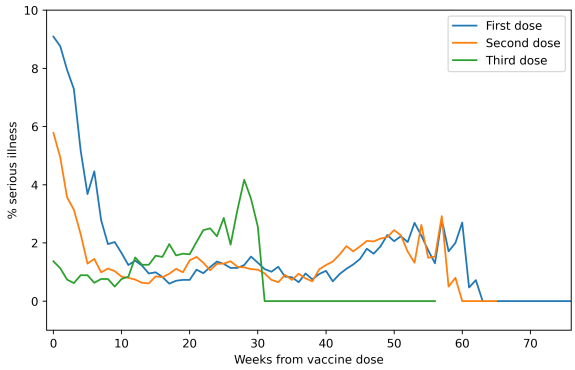


Fig. 3: Effect of vaccine over time

Among all the patients, only 2.5% were not vaccinated when tested positive for COVID-19. There are 0.6%, 27.2% and 69.7% of the patients who had received 1, 2, and 3 doses of vaccine respectively. We explored the waning effects of the vaccination doses. Similar to the findings of Chemaitelly et al. [5], Fig. 3 suggests that the protection effect of first and second doses of vaccine against severe illness improved significantly in the first 8 weeks, and reached optimum effectiveness after around 15 weeks. Contrary to observations of Fabiani et al. [6] and Andrews et al. [7], the effect of the first two doses of vaccine started to wane much later at around 40 weeks instead of after 20 weeks among the population in Singapore. This could be a result of a successful vaccination campaign by the Singapore Government, where most residents in Singapore completed their first two doses of vaccines within a short interval, thus prolonging the period of optimal protection. The protection effect of the third dose of vaccine seemed to wane faster than the first two doses, but such trend may be due to the other confounding causes. For instance, high-risk patients, predominantly the elderly, received their third vaccine dose earlier than other age groups. This demographic constitutes the majority of individuals exhibiting an extended interval between the third vaccine dose and a positive COVID-19 test result. It's important to observe that the lines corresponding to the first dose, second dose, and third dose all plateaued at zero around week 62, 60, and 31 respectively. This phenomenon arises due to the interval between dose administration and data collection for analysis.

C. Software Used

This study employed machine learning models including Random Forest, Logistic Regression, Support Vector Machine, and Naive Bayes from Python's Scikit-learn and XGBoost libraries. For situations necessitating undersampling and/or oversampling, sampling methods from Python's Imbalanced-learn libraries were utilized.

D. Data Processing

Binary categorical variables were encoded with 1 for the positive class and 0 for the negative class. One hot encoding was applied to all the categorical variables. For patients who had received vaccinations, only those administered before their confirmed positive COVID-19 diagnosis contributed to the cumulative count of vaccine doses. The number of days between each vaccination dose and tested COVID-19 positive was calculated based on the vaccination date and the reported date of confirmation of COVID-19. In cases where a specific dose was omitted or administered after the COVID-19 confirmation, an extensive value of 10,000 days was assigned to denote the gap between vaccination and confirmed COVID-19 status. Patients who required mechanical ventilation, admitted into ICU or deceased were labeled as positive class label indicating severe illness.

E. Handling imbalance data

Given that only 1.1% of the patients had developed severe illness, we tested some of the common sampling methods on the training data set which include under-sampling of the majority negative class (Near-Miss and One-sided selection methods), over-sampling of the minority positive class (Synthetic Minority Over-sampling Technique (SMOTE)) and a combination of over- and under-sampling methods (SMOTE Tomek links (SMOTETomek)). None of the sampling methods improved the prediction performance significantly. Instead of

applying sampling methods, we noticed the model performance generally improved when the "class_weight" parameter is set to "balanced" during training. Hence, data imbalance was handled by the algorithm.

F. Data Splitting

20% of the samples were randomly chosen as the test set stratified by the sample label, i.e., severe illness. The remaining 80% of samples were split into a training set and a validation set in a ratio of 4:1 for hyperparameter tuning with five-fold cross-validation to prevent overfitting.

G. Hyperparameter Tuning

Hyperparameter tuning was conducted for each ML model and sampling method. We used Scikit-Learn's grid search function with five-fold stratified cross-validation to find the optimal hyperparameter values. The range of values of each hyperparameter is summarised in Table II. As the data set was highly imbalanced, the "class_weight" hyperparameter of Logistic Regression, GaussianNB, SVM and Random Forest models was set to "balanced". For SVM, only radial basis function kernel was used.

H. Threshold values

Delayed medical attention could have detrimental effects on patients' health. Hence, government or policy makers may wish to be more conservative in identifying at-risk patients. We tested the models using lower probability thresholds of 0.25, 0.20, 0.15 and 0.10 for the predictions, to determine if the models can achieve similar sensitivity compared to broad-based approach adopted by the Singapore Government while achieving higher specificity.

TABLE II: Range of values used in hyperparameter tuning

ML model	Hyperparameter	Range of values searched
SVM	C	0.1, 1, 10, 100
	gamma	1, 0.1, 0.01, 0.001, 0.0001
Random Forest	n_estimators	500 to 3,000 (increment of 100)
	max_depth	None, 5, 10, 15, 20, 25, 30
	min_samples_leaf	1, 10, 15, 20, 50, 70, 100, 200
	min_samples_split	2, 10, 15, 20, 50, 70, 100, 200
XGBoost	n_estimators	500 to 3,500 (increment of 100)
	learning_rate	0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001
	max_depth	1 to 35 (increment of 1)
	min_child_weight	1 to 35 (increment of 1)
	gamma	0.1 to 10 (increment of 0.1)
	subsample	0.1 to 1 (increment of 0.1)
	colsample_bytree	0.1 to 1 (increment of 0.1)
	reg_alpha	1 to 50 (increment of 1)
	scale_pos_weight	40 to 160 (increment of 10)

I. Model Evaluation

We tested a variety of common ML models, including Logistic Regression, Gaussian Naive Bayes (GaussianNB), Support Vector Machine (SVM), Random Forest and XGBoost classifiers. For the comprehensive assessment of these ML models, we employed a range of metrics encompassing AUROC, accuracy, sensitivity, specificity, F2-score, and false negative rate (FNR). Of these, AUROC emerged as the focal metric, carrying substantial importance, particularly in hyperparameter tuning. It provided a holistic gauge of model performance across varied classification thresholds, allowing future threshold adjustments as per situational needs. The trio of accuracy, sensitivity, and specificity, widely used in the literature, enabled a thorough comparison of predictive performance against existing studies. Moreover, our choice to incorporate the F2-score stemmed from our emphasis on recall over precision, accentuating the significance of reducing false negatives rather than false positives. This emphasis holds paramount importance in identifying patients prone to severe illness, thereby safeguarding their well-being. Similarly, the inclusion of FNR furnished a direct criterion for model selection, underscoring the imperative to minimize false negatives. This approach ensures timely medical attention for patients predisposed to severe illness, further cementing the efficacy of the chosen model.

IV. Results and Discussion

A. Comparison of ML Models

Two distinct datasets were curated from various data sources. Dataset 1 encompasses patients' demographic particulars, pre-existing conditions, pregnancy status, and vaccination records, while Dataset 2 comprises patients' demographic details, medical history, and vaccination records. The rationale behind this diverse data selection lies in the recognition that each model may exhibit varying degrees of applicability. Dataset 1 leverages telemedicine data, where pre-existing conditions are self-reported. This configuration proves beneficial for scenarios in which comprehensive medical histories, such as electronic medical records, are not ubiquitously implemented. In contrast, Dataset 2 offers a more comprehensive approach, incorporating individual patient's record of medical history. However, we acknowledge that in certain contexts, such as in countries where medical records are not centralized or electronically managed, this approach may not be applicable, making self-reported measures a more suitable choice. By investigating the performance of both model types, we aim to provide the relevant model that can be employed as needed. Table III displays the performance scores of the ML models, while Fig. 5 illustrates the graphical representations of the AUROC scores. The feature set labels in the AUROC plots correspond to the usage of each of the three datasets.

For Dataset 1, Random Forest has the highest AUROC of 0.912, good accuracy, specificity and F2-score of 0.970, 0.975 and 0.430 respectively, but has the lowest sensitivity of 0.588. XGBoost has the second highest AUROC of 0.906 and accuracy of 0.921, while its sensitivity (0.701) and specificity (0.923) were more balanced than Random Forest. Both XGBoost and Random Forest models performed better by using Dataset 2 compared to using Dataset 1. AUROC of Random Forest increased slightly from 0.912 to 0.916, and yielded a more balanced sensitivity and specificity of 0.858 and 0.813

TABLE III: Comparing the performances of baseline ML models

Dataset	ML Model	AUROC	Accuracy	Sensitivity	Specificity	F2-score	FNR
Dataset 1	Logistic Regression	0.832	0.749	0.787	0.748	0.145	0.213
	GaussianNB	0.788	0.811	0.608	0.813	0.144	0.392
	SVM	0.837	0.814	0.701	0.815	0.166	0.299
	Random Forest	0.912	0.970	0.588	0.975	0.430	0.412
	XGBoost	0.906	0.921	0.701	0.923	0.305	0.299
Dataset 2	Logistic Regression	0.786	0.589	0.643	0.588	0.115	0.357
	GaussianNB	0.885	0.948	0.341	0.959	0.256	0.659
	SVM	0.860	0.942	0.330	0.953	0.236	0.670
	Random Forest	0.916	0.814	0.858	0.813	0.278	0.142
	XGBoost	0.930	0.958	0.542	0.965	0.426	0.458
Dataset 3	Random Forest	0.884	0.801	0.806	0.801	0.245	0.194
	XGBoost	0.885	0.853	0.731	0.855	0.277	0.269
Lam et al.	XGBoost	0.880	0.850	0.800	0.950	-	-
Keng et al.	NN	0.953	-	1.000	0.857	-	-

respectively. AUROC of XGBoost increased significantly from 0.906 to 0.930, which is the highest among all models using Datasets 1 and 2. The variations in outcomes can be ascribed to data reliability factors. The data pertaining to patients' pre-existing conditions and pregnancy status may be less robust due to their reliance on self-reporting by patients and their coverage of only eight major pre-existing conditions. In contrast, the medical history data encompasses patients' actual past clinical visits, with each clinical episode professionally encoded in ICD-10 by medical providers. This medical history data offers a much more refined dataset where each condition is represented in a more detailed health category than high-level classification of just eight major conditions.

B. Parsimonious Models

While using all variables in Datasets 1 and 2 produced good performances, a parsimonious model that uses fewer and more commonly collected variables may be more easily adopted in real life. For instance, healthcare administrators or doctors may need to rely on a simple checklist to make decisions when ML prediction using a computer is not available. If any countries want to implement our ML model, they may not have the same set of variables collected. They are less likely to be affected by data availability if fewer variables are required and the variables are commonly collected data.

To explore the feasibility of a parsimonious model, we explored variable reduction by selecting five variables among those with highest feature importance scores using a Random Forest model as shown in Fig. 4. In addition, these top 5 variables can be readily sourced from a single data source, eliminating the need to combine data from multiple sources. This streamlined dataset, denoted as Dataset 3, comprises of age, number of days since each of the three doses, and the total number of vaccine doses taken. Age, being a universally available and easily obtainable data, and vaccination information, commonly collected by countries with vaccination programmes in place, make up these essential variables. When employing this set of five variables, a marginal decrease in performance was observed compared to using Datasets 1 and 2. The AUROC score of Random Forest and XGBoost decreased by 0.028 (from 0.912 to 0.884) and 0.021 (from 0.906 to 0.885) respectively, compared to Dataset 1. These scores decreased by 0.032 (from 0.916 to 0.884) and 0.046 (from 0.930 to 0.885), respectively, compared to Dataset 2. The detailed results are shown in Table III. Using logistic regression, we tested two additional models using only the age variable or only the vaccination-related features. The models yielded significantly lower AUROC of 0.731 and 0.615 respectively. Given the slight decline in performance for the parsimonious models using the top 5 variables, it is reasonable to conclude that both ML models are suitable for most COVID-19 patients, given that the relevant information is available and could achieve greater applicability.

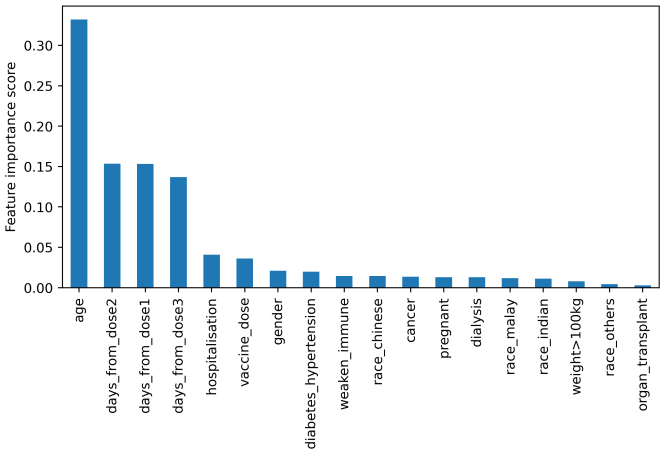
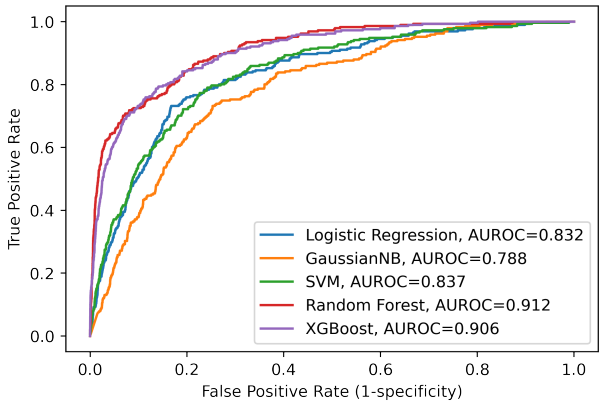
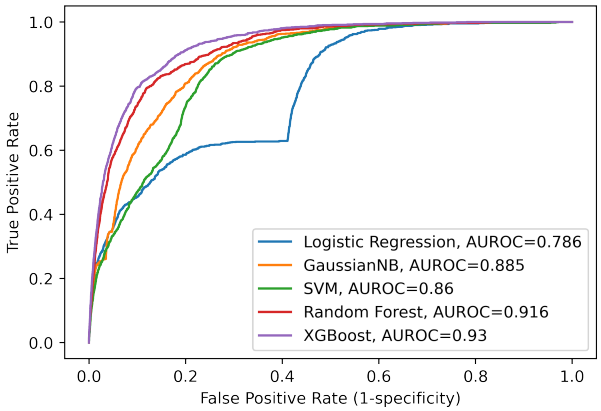


Fig. 4: Feature importance scores from Random Forest Model

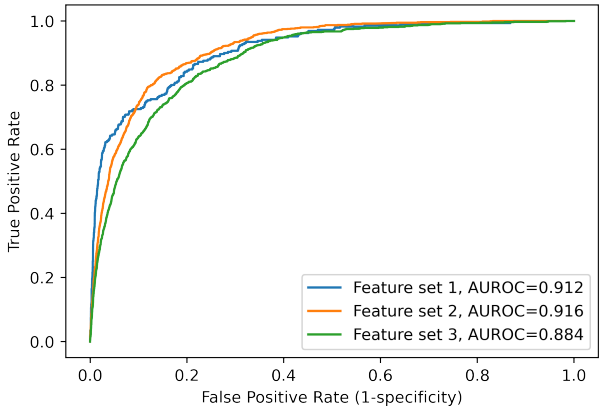
Compared to the results of Lam et al. [12], which also used XGBoost for predicting severe illness, all of our XGBoost models in Table III have higher AUROC scores (+0.005 to +0.05). Compared to the neural network model of Keng et al. [9], all of our XGBoost models have lower AUROC scores (-0.023 to -0.073). Other studies broke down severe illness into multiple sub-categories so they are not directly comparable to our model. All of the aforementioned studies, including Lam



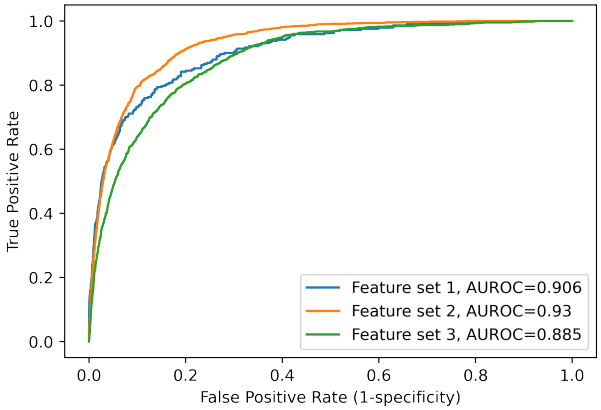
(a) AUROC plot of all ML models using Dataset 1



(b) AUROC plot of all ML models using Dataset 2



(c) AUROC plot of Random Forest using different datasets



(d) AUROC plot of XGBoost using different datasets

Fig. 5: AUROC plots of various ML models

et al. and Keng et al., relied on clinical data collected from patients in clinical or laboratory settings, (e.g., hemoglobin, albumin, globulin, nitrogen level and oxygen saturation in the blood etc) and none of them included vaccination status or vaccine waning effect into their prediction model. While our models exhibit slightly reduced performance compared to Keng et al., our innovation lies in utilizing solely patients' demographic information, vaccination status, and pre-existing medical conditions or medical history for prediction, without necessitating clinical or laboratory data. This approach proves particularly relevant for nations with vaccinated populations opting for home-based recoveries.

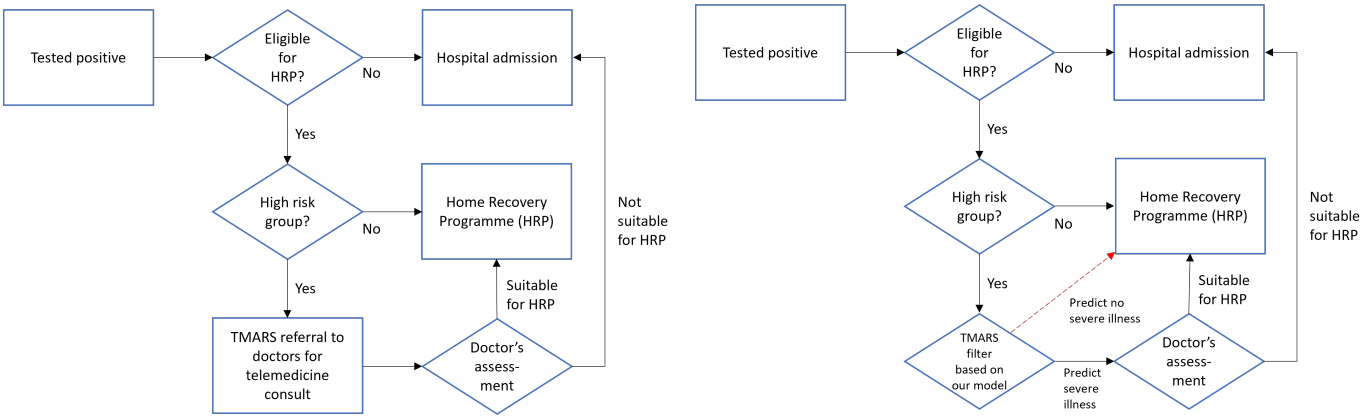
Furthermore, our research has shown that substantial prediction accuracy can be achieved using a streamlined set of five commonly captured variables. This not only augments the adaptability of our model but also resonates significantly for countries lacking access to patients' comprehensive medical histories. We hope that our contribution stands as a notable effort, spotlighting the value that nationwide big data can bring to the forefront.

V. Application in Singapore's Context

In 2021, Singapore's Ministry of Health (MOH) implemented the National Sorting Logic to risk stratify COVID-19 patients. It was then integrated with Telemedicine Allocation Reconciliation System (TMARS) to manage patients. The management workflow is illustrated in Fig. 6a. After a patient is tested positive for COVID-19, he or she will be assessed for suitability of HRP either by a doctor or self-assessment based on the following list of criteria. Patients not eligible for HRP will be admitted into hospitals, while patients eligible for HRP will be further assessed if they belong to the high-risk group based on the following list of broad-based sorting criteria.

The following individuals (adult) are not eligible for HRP:

- Partially vaccinated or unvaccinated persons aged 80 years and older
- Pregnant women with gestation age 36 weeks and above
- Partially vaccinated or unvaccinated pregnant women



(a) Current workflow

(b) Proposed improvised workflow

Fig. 6: Workflow for managing COVID-19 patients

The following individuals (adult) belong to the high-risk group:

- Persons aged 70 years and older
- Persons who are/have
  - pregnant
  - on dialysis
  - had organ transplant surgery
  - been diagnosed with cancer previously
  - any disease or taking medications that weaken the immune system
  - any disease affecting their heart, lungs, kidneys, liver or brain that required a hospital admission in the last 6 months

The default recovery plan for low-risk patients will be the HRP. In the case of high-risk patients who haven't received prior medical attention, TMARS will facilitate teleconsultation with doctors. Subsequently, these doctors will evaluate the necessity of hospital admission versus home recovery under the HRP. Our ML prediction model can seamlessly integrate into the existing workflow by becoming an additional layer of filter within TMARS's operations, as depicted in Fig. 6b.

The XGBoost models have FNR between 0.269 and 0.458 as shown in Table III, in which the predictions missed a significant proportion of patients who potentially developed serious illness. To be more conservative and identify as many high-risk patients, we reduced the probability threshold for predicting severe illness from 0.5 to 0.25, 0.20, 0.15, 0.10. Compared to the broad-based sorting logic described above, when we set the probability threshold for severe illness to 0.1 to our XGBoost models, they achieved equal or lower FNR of $0.033 - 0.089$ and improved specificity from 0.310 to $0.509 - 0.783$ (see Table IV).

Based on the preceding analysis and experimentation involving various thresholds, our ML models exhibit promising performance in the context of COVID-19 patient risk stratification, particularly in identifying potential patients at risk of developing severe illnesses. The model utilizing Dataset 1, which incorporates self-reported medical conditions, proves advantageous for scenarios where minimizing system integration or data integration during deployment is a priority. Conversely, the model employing Dataset 2 emerges as the most comprehensive and top-performing model, demonstrating superior overall prediction capabilities. Dataset 3, known as the parsimonious model, presents a noteworthy advantage in its reduced data dependency, reliance on easily attainable variables, and potential for widespread adoption. Given that patients retain the option to seek medical attention at hospital emergency departments in the event of exacerbated symptoms, we maintain the perspective that a minimal number of false negative cases would have negligible repercussions on public health at large. Additionally, this model exhibits the lowest False Negative Rate (FNR), indicative of sufficient recall in identifying patients who may progress to severe illness, thus ensuring comprehensive patient care.

At the zenith of the infectious period, Singapore had a daily surge of over 20,000 COVID-19 cases, placing immense strain on local healthcare resources. Should a subsequent wave of COVID-19 infections emerge, implementing one of our ML models can effectively curtail the total number of patients necessitating physician evaluation, all the while ensuring that high-risk patients are not overlooked. This highlights the invaluable role of using big data in safeguarding public health, particularly when faced with a significant influx of patients.

VI. Limitations and Future Work

In this study, the maximum duration between vaccination and tested positive for COVID-19 was 76 weeks (536 days). As the protection effect of vaccine is expected to continue to wane, the validity of our model should be tested again over a longer time horizon. Lastly, SARS-CoV-2 is continuously mutating as it circulates through the world population. The effect of different strands of SARS-CoV-2 was not taken into account during this study due to data limitations. As new

TABLE IV: Comparing the performances of broad-based approach and XGBoost models

Broad-based approach / ML model	Probability threshold for severe illness	Accuracy	Sensitivity	Specificity	F2-score	FNR
Broad-based	-	0.317	0.921	0.310	0.070	0.089
XGBoost with Dataset 1	0.25	0.827	0.814	0.827	0.202	0.186
	0.20	0.791	0.845	0.791	0.181	0.155
	0.15	0.741	0.880	0.739	0.157	0.120
	0.10	0.663	0.921	0.660	0.131	0.089
XGBoost with Dataset 2	0.25	0.841	0.773	0.909	0.389	0.227
	0.20	0.851	0.818	0.884	0.359	0.182
	0.15	0.846	0.859	0.856	0.316	0.141
	0.10	0.853	0.924	0.783	0.269	0.076
XGBoost with Dataset 3	0.25	0.663	0.923	0.658	0.186	0.077
	0.20	0.624	0.943	0.618	0.173	0.057
	0.15	0.578	0.960	0.571	0.160	0.040
	0.10	0.517	0.967	0.509	0.144	0.033

variants continually emerge, the necessity of re-evaluating our models in subsequent timeframes becomes apparent.

VII. Conclusion

This study introduces the utilization of comprehensive national big data to effectively stratify risks for the preservation of public health in Singapore. We proposed the use of an XGBoost machine learning model aimed at predicting the likelihood of severe illness development among COVID-19 patients. This model harnessed patients' demographic details, vaccination status, and medical history. The best performing XGBoost model exhibited commendable performance metrics: an AUROC score of 0.930, along with accuracy, sensitivity, and specificity values of 0.958, 0.542, and 0.965 respectively. In addition, we further trained a parsimonious model, comprising only five variables (age, number of vaccine doses administered, and days since first, second and third dose), which yielded an AUROC score of 0.885, alongside accuracy, sensitivity, and specificity of 0.853, 0.731, and 0.855. Our approach accomplished comparable performance without incorporating clinical or laboratory data, while taking patients' vaccination status into consideration. This is particularly pivotal for a populace characterized by high vaccination coverage and a preference for home-based recovery. Moreover, the parsimonious model, reliant on commonly collected variables, stands less vulnerable to data availability concerns. With our experience and model tailored for Singapore, we aspire for its applicability to extend to other nations, ultimately contributing to the broader big data community by showcasing the substantial value of its integration within nationwide healthcare applications.

References

[1] "Overview of covid-19 situation in singapore," https://www.gov.sg/features/covid-19, last accessed: 2023-09-02.

[2] "Updating our healthcare protocols for a more covid-19 resilient nation," https://www.moh.gov.sg/news-highlights/details/updating-our-healthcare-protocols-for-a-more-covid-19-resilient-nation, last accessed: 2023-09-02.

[3] T. C. Koh, J. Y. Goh, W. K. Yau, T. W. Kok, M. Y. Lee, K. C. Goh, M. W. Sng, and S. J. Chong, "Enhanced monitoring system to better monitor high-risk covid-19 patients," *Journal of Medical Systems*, vol. 47, no. 1, 2023.

[4] B. G. Marin, G. Aghagoli, K. Lavine, L. Yang, E. J. Siff, S. S. Chiang, T. P. Salazar-Mather, L. Dumenco, M. C. Savaria, S. N. Aung, T. Flanigan, and I. C. Michelow, "Predictors of COVID-19 severity: A literature review," *Reviews in Medical Virology*, vol. 31, no. 1, pp. 1–10, Jul. 2020.

[5] H. Chemaitelly, P. Tang, M. R. Hasan, S. AlMukdad, H. M. Yassine, F. M. Benslimane, H. A. A. Khatib, P. Coyle, H. H. Ayoub, Z. A. Kanaani, E. A. Kuwari, A. Jeremijenko, A. H. Kaleeckal, A. N. Latif, R. M. Shaik, H. F. A. Rahim, G. K. Nasrallah, M. G. A. Kuwari, H. E. A. Romaihi, A. A. Butt, M. H. Al-Thani, A. A. Khal, R. Bertollini, and L. J. Abu-Raddad, "Waning of BNT162b2 vaccine protection against SARS-CoV-2 infection in qatar," *New England Journal of Medicine*, vol. 385, no. 24, p. e83, Dec. 2021.

[6] M. Fabiani, M. Puopolo, C. Morciano, M. Spuri, S. S. Alegiani, A. Filia, F. D'Ancona, M. D. Manso, F. Riccardo, M. Tallon, V. Proietti, C. Sacco, M. Massari, R. D. Cas, A. Mateo-Urdiales, A. Siddu, S. Battilomo, A. Bella, A. T. Palamara, P. Popoli, S. Brusaferro, G. Rezza, F. M. Ippolito, and P. Pezzotti, "Effectiveness of mRNA vaccines and waning of protection against SARS-CoV-2 infection and severe covid-19 during predominant circulation of the delta variant in italy: retrospective cohort study," *BMJ*, p. e069052, Feb. 2022.

[7] N. Andrews, E. Tessier, J. Stowe, C. Gower, F. Kirsebom, R. Simmons, E. Gallagher, S. Thelwall, N. Groves, G. Dabrera, R. Myers, C. N. Campbell, G. Amirthalingam, M. Edmunds, M. Zambon, K. Brown, S. Hopkins, M. Chand, S. N. Ladhani, M. Ramsay, and J. L. Bernal, "Duration of protection against mild and severe disease by covid-19 vaccines," *New England Journal of Medicine*, vol. 386, no. 4, pp. 340–350, Jan. 2022.

[8] S. Wollenstein-Betech, A. A. B. Silva, J. L. Fleck, C. G. Cassandras, and I. C. Paschalidis, "Physiological and socioeconomic characteristics predict COVID-19 mortality and resource utilization in brazil," *PLOS ONE*, vol. 15, no. 10, p. e0240346, Oct. 2020.

[9] J. Kang, T. Chen, H. Luo, Y. Luo, G. Du, and M. Jiming-Yang, "Machine learning predictive model for severe COVID-19," *Infection, Genetics and Evolution*, vol. 90, p. 104737, Jun. 2021.

[10] M. A. Quiroz-Juárez, A. Torres-Gómez, I. Hoyo-Ulloa, R. de J. León-Montiel, and A. B. U'Ren, "Identification of high-risk COVID-19 patients using machine learning," *PLOS ONE*, vol. 16, no. 9, p. e0257234, Sep. 2021.

[11] L. Ryan, C. Lam, S. Mataraso, A. Allen, A. Green-Saxena, E. Pellegrini, J. Hoffman, C. Barton, A. McCoy, and R. Das, "Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: A retrospective study," *Annals of Medicine and Surgery*, vol. 59, pp. 207–216, Nov. 2020.

[12] C. Lam, J. Calvert, A. Siefkas, G. Barnes, E. Pellegrini, A. Green-Saxena, J. Hoffman, Q. Mao, and R. Das, "Personalized stratification of hospitalization risk amidst COVID-19: A machine learning approach," *Health Policy and Technology*, vol. 10, no. 3, p. 100554, Sep. 2021.

[13] M. Laatifi, S. Douzi, A. Bouklouz, H. Ezzine, J. Jaafari, Y. Zaid, B. E. Ouahidi, and M. Naciri, "Machine learning approaches in covid-19 severity risk prediction in morocco," *Journal of Big Data*, vol. 9, no. 1, Jan. 2022.