# CLM: a machine learning approach to COVID-19 screening for Indonesian health workers

**Running Head:**
A data-driven COVID-19 screening tool for the Indonesian health workers

**Author List:**
Shreyash Sonthalia[1], Muhammad Aji Muharrom[1], Levana Sani[1], Olivia Herlinda[2], Adrianna Bella[2], Dimitri Swasthika[2], Panji Hadisoemarto[3], Diah Saminarsih[4], Nurul Lutungan[2], Astrid Irwanto[1,5], Akmal Taher[6], Joseph L. Greenstein[7]

**Affiliations:**
[1] Nalagenetics Pte Ltd, Singapore, Singapore
[2] Center for Indonesia's Strategic Development Initiatives (CISDI), Jakarta, Indonesia
[3] Department of Public Health., Faculty of Medicine, Padjajaran University, Bandung, Indonesia
[4] World Health Organization, Geneva, Switzerland
[5] Department of Pharmacy, Faculty of Science, National University of Singapore, Singapore
[6] Department of Urology, Cipto Mangunkusumo Hospital, Universitas Indonesia, Jakarta, Indonesia
[7] Institute for Computational Medicine, The Johns Hopkins University, Baltimore, United States
*Dr. Greenstein's participation in this study was as an unpaid consultant for Nalagenetics. All opinions expressed and implied in this manuscript do not represent or reflect the views of the Johns Hopkins University or the Johns Hopkins Health System

**Abstract Word Count:** 384
**Text Word Count:** 3475

## ABSTRACT

The COVID-19 pandemic poses a heightened risk to health workers, especially in low- and middle-income countries such as Indonesia. Due to the limitations to implementing mass RT-PCR testing for health workers, high-performing and cost-effective methodologies must be developed to help identify COVID-19 positive health workers and protect the spearhead of the battle against the pandemic. This study aimed to investigate the application of machine learning classifiers to predict the risk of COVID-19 positivity (by RT-PCR) using data obtained from a survey specific to health workers. Machine learning tools can enhance COVID-19 screening capacity in high-risk populations such as health workers in environments where cost is a barrier to accessibility of adequate testing and screening supplies. We built two sets of COVID-19 Likelihood Meter (CLM) models: one trained on data from a broad population of health workers in Jakarta and Semarang (full model) and tested on the same, and one trained on health workers from Jakarta only (Jakarta model) and tested on an independent population of Semarang health workers. The area under the receiver-operating-characteristic curve (AUC), average precision (AP), and the Brier score (BS) were used to assess model performance. Shapley additive explanations (SHAP) were used to analyze feature importance. The final dataset for the study included 3979 health workers. For the full model, the random forest was selected as the algorithm of choice. It achieved cross-validation mean AUC of 0.818 ± 0.022 and AP of 0.449 ± 0.028 and was high performing during testing with AUC and AP of 0.831 and 0.428 respectively. The random forest model was well-calibrated with a low mean brier score of 0.122

± 0.004. A random forest classifier was the best performing model during cross-validation for the Jakarta dataset, with AUC of 0.824 ± 0.008, AP of 0.397 ± 0.019, and BS of 0.102 ± 0.007, but the extra trees classifier was selected as the model of choice due to better generalizability to the test set. The performance of the extra trees model, when tested on the independent set of Semarang health workers, was AUC of 0.672 and AP of 0.508. Our models yielded high predictive performance and may have the potential to be utilized as both a COVID-19 screening tool and a method to identify health workers at greatest risk of COVID-19 positivity, and therefore most in need of testing.

## INTRODUCTION

Since the first confirmed case of COVID-19 in Indonesia in March 2020, there have been 1.8 million confirmed cases and more than 50 thousand deaths resulting from COVID-19 infection [1]. Health workers in Indonesia are at a high risk of COVID-19 exposure and infection due to the nature of the profession, with 654 COVID-19 deaths recorded by January 2021 [2, 3]. The COVID-19 pandemic is placing an enormous burden on Indonesia's public health and economy, especially due to Indonesia having the highest fatality rate for health workers in Asia [4]. Despite implementation of large-scale social restrictions at the national and regional level, vaccination of high-risk population groups, and advocation for the usage of personal protective equipment (PPE) [5], Indonesia still had the highest daily and cumulative COVID-19 cases in The Association of Southeast Asian Nations (ASEAN) in September 2021 [6]. Furthermore, due to the often-asymptomatic nature of COVID-19 infection [7], efforts to prevent COVID-19 transmission were constrained by the ability to immediately detect and isolate the infected people [8-10].

Mass testing by reverse transcription-polymerase chain reaction (RT-PCR), the gold standard of COVID-19 diagnostic testing, remains one of the key measures to reduce transmission of COVID-19 [11]. However, the implementation of mass RT-PCR testing is limited in developing countries such as Indonesia due to financial, capital, and logistical constraints [12, 13]. Despite the rapidly increasing COVID-19 cases since the Eid al-Fitr holiday in May 2021, some regions of Indonesia have been facing limited reagent supply and inadequate laboratory capacity to provide sufficient testing. In June 2021, Indonesia had the second-lowest testing rate in Southeast Asia with only 7.5 tests per confirmed case, far below the World Health Organization (WHO) recommendation of 10-30 tests per confirmed case [14-16]. The healthcare system of Indonesia has further been weakened by surges in case numbers and patients requiring hospitalization, leading to depleted medical supplies [17]. Furthermore, in the event of a surge, one study assessed that several provinces in Indonesia would likely have suboptimal diagnostic capabilities even if using rapid diagnostic technologies in referral hospitals [18].

The limitations of mass RT-PCR implementation in Indonesia underscore the development of a COVID-19 detection method that is accurate, affordable, and accessible to users with minimum equipment and personnel required to prioritize PCR testing for health workers. Rapid diagnostic antigen testing is another common testing modality that is less accurate than RT-PCR [17]. While antigen testing provides rapid results, implementation of mass testing may still be limited by the coordination of stakeholders who may have to balance resources. A free testing or screening modality would encourage hospitals and health systems to test their workers more frequently and can be implemented in resource-strapped communities. Machine learning tools can achieve these goals and have already shown promise in several countries such as the United States, China, Israel, and Slovenia [19-22]. Several machine

learning-based models have already been developed for COVID-19 screening using data from sources such as Computed Tomography (CT) scans [19, 23], clinical symptoms [20-21], and laboratory tests [24-26]. These tools' sensitivity and specificity values are high, ranging from 0.86-0.93 and 0.56-0.98 respectively.

Currently, most machine learning COVID-19 screening tools have been deployed in technologically advanced countries [19-21, 24, 26]. Since most of the available models used data from hospitalized patients, the tools may not be effective for COVID-19 screening for health workers, due to differences in available features between hospitalized patients and health workers. Hospitalized patients are highly likely to have different symptomatology, behavioral tendencies, and PPE usage requirements than health workers who are not hospitalized, the target population of this study. We aim to apply machine learning algorithms to develop software tools to augment COVID-19 screening for Indonesian health workers. This method is expected to ease the burden in Indonesia's healthcare system caused by COVID-19 through the implementation of a fast, accessible, and widespread screening methodology that can allow for accurate triage and systematic allocation of RT-PCR testing for health workers.

## METHODS

### *Study design and data*

The study was designed as a cross-sectional observational study with a total of 3979 health workers, including medical professionals and non-medical professions working at healthcare facilities. The data was collected between January 20 and September 15, 2021, at forty-four healthcare provider locations: 15 healthcare providers in Greater Jakarta Area, 1 hospital in Bandung city, and 28 healthcare providers in Semarang[1]. The hospitals and healthcare providers were selected through online recruitment, recommendations from medical associations, and a partnership agreement with the following inclusion criteria: 1) the ability to conduct swab collection for RT-PCR testing by trained health workers, 2) the presence of healthcare or non-healthcare staff with COVID-19 symptoms or close contact with COVID-19 patients, and 3) the support from the health facility management to participate in the research. The proportion of respondents from Jakarta was 3477 (87.4%) and 502 from Semarang (12.6%).

The hospitals had the authority to recommend that a member of their staff to be included in the study by following this criterion: either (1) had close contact with at least one COVID-19 patient within the last fourteen days or (2) developed COVID-19 related symptoms within the last fourteen days. For each respondent, we collected oropharyngeal or nasopharyngeal swab specimens for RT-PCR as well as data for COVID-19 symptoms, comorbidities, COVID-19 protective behaviors, working conditions, and COVID-19 vaccination status through a self-administered questionnaire. All oropharyngeal/nasopharyngeal swab specimens from respondents in participating healthcare providers in Greater Jakarta Area and Bandung were tested at Primaya Hospital, East Bekasi, while specimens from respondents in participating healthcare providers in Semarang were tested at Diponegoro National Hospital.

[1] The data from 28 healthcare providers in Semarang came from several surrounding healthcare providers in Central Java, including: 52.3% from Diponegoro National Hospital; 13.4% from Community health center (BALKESMAS); 17.2% from individual health worker from clinic/Laboratory/Pharmacy/Homecare; 5.7% from Medical and Health student around Semarang; 4.2% from Central General Hospital Dr. Kariadi; 2.3% from Health Equipment / Medical Company; 1.9% from Ken Saras Hospital; 1.9% from Kidney & Hypertension Clinic; and 1.1% from Halmahera Primary Healthcare.

### *Study Variables*

The survey questions included behavioral (protective, social, and travel) tendencies, COVID-19 vaccination status, working conditions, symptoms, comorbidities, and level of COVID-19 exposure and interaction with infected patients at health facilities. The dependent variable in this study was the result of the COVID-19 RT-PCR test taken within three days of filling the survey. Respondents with inconclusive RT-PCR results were not included in the processed dataset. Behavioral questions were chosen based on general and medical worker-specific risk factors identified in the current literature and encompassed handwashing, mask-wearing, PPE adherence, social distancing, and domestic and foreign travel tendencies. Hand-washing behaviors were assessed by the level of adherence to the six-step handwashing protocol [27, 28]. Mask-wearing and social distancing behaviors were assessed according to current WHO guidelines [29, 30].

### *Modeling and Prediction*

To predict COVID-19 diagnosis in our cohort, we trained and evaluated several machine learning classification algorithms, including random forest [31], extra trees classifier [32], and model ensembles. These were implemented using the scikit-learn Python library [33], while XGBoost [34] was implemented using scikit-learn compatible packages in Python. These models were chosen after experiments with various algorithms, such as optimization and deep learning methods, during preliminary modeling. Preprocessed respondent features were used as inputs for each model, generating an output prediction risk score with a value between 0 and 1. The output was then converted to a class label by a thresholding function. Hyperparameters were tuned and chosen using the random search optimization method in scikit learn [35]. Feature selection was implemented using the sequential feature selection method in scikit learn. Model performance was analyzed and interpreted using the area under the receiving-operating characteristic curve (AUC) and area under the precision-recall (PR) curve, average precision (AP), while model calibration was assessed using the Brier Score (BS) for each model. Feature importance and model interpretability was assessed using Shapley additive explanations (SHAP) from the SHAP package [36] in Python.

Two sets of models were developed: one using respondents from health facilities in both Jakarta and Semarang (full model), and one on respondents from Jakarta health facilities only (Jakarta model). The full model was trained with stratified 5-fold cross-validation on 80% of the dataset and tested on the remaining 20%. The Jakarta model was tested on inputs from the Semarang health workers and was developed to assess the predictive capability of this approach in an independent population of health workers within Indonesia. Due to class imbalance, adaptive synthetic oversampling (ADASYN) [37] was used during training for the positive class before validation for both modeling approaches.

## RESULTS

### *Data Summary*

Data from 3,979 health workers were included in the final dataset for model building. A summary of cohort demographics and survey responses is provided in Table 1. Approximately 74% of respondents were female and the average age was 30 years old. As health workers are currently the priority group for receiving the COVID-19 vaccine in Indonesia, 86% of

respondents had received at least the first dose of the COVID-19 vaccine. Regarding protective behaviors in the prior month, most respondents indicated they had been trained in PPE standards (98%) and six-step hand washing techniques (97%). Around 18% of respondents were currently self-isolating after having close contact with COVID-19 patients and 30% were involved in aerosol-generating procedures on COVID-19 patients. Furthermore, approximately half of positive cases were currently self-isolating. Additionally, around 73% and 59% of respondents reported having activity in a closed room and having outdoor activities at least 1-3 times a month, respectively. This study also found that 76% of respondents always used a mask outside the home, 53% always avoided shaking hands, and 53% always maintained physical distance. In terms of the use of public transportation, 83% of respondents never used mass public transportation and 76% never used door to door transportation.

Approximately 80% of positive cases were symptomatic. Cough (48.52%), headache (44.43%), runny nose (39.34%), chills (38.36%), and fever (37.54%) were the five most reported symptoms among those who had COVID-19 positive test results. 90% of respondents stated that they did not have any comorbidities. Lung disease (3.11%), hypertension (2.46%), and pregnancy (2.30%) were the most reported comorbidities among those who were COVID-19 positive.

### *Model Performance and Explainability*

The Jakarta model was developed to investigate the model performance on a geographically independent test set, which was the data from the Semarang health workers. The full model contains data from both cities and is thus a model built on a wider population and would be generalizable to a more heterogeneous testing population.

#### *Jakarta Model*

The Jakarta model was trained and cross-validated on health workers from Jakarta only and was tested on respondents from Semarang. The test set was composed of 12.6% of the entire dataset. Figure 1 displays the test set receiver operating characteristic (ROC) curvd, precision-recall (PR) curve, the cross-validation calibration curve as well as both cross validation and test sensitivity, specificity, positive-predictive value (PPV), negative-predictive value (NPV), and the F1 score of several algorithms on the Semarang data. The prevalence of COVID-19 positive health workers in the validation and testing datasets was 29.5%. The random forest classifier had the best predictive performance during cross-validation, followed by the voting classifier, composed of an ensemble of a XGBoost classifier and a random forest classifier. The mean AUC for random forest on the cross-validation sets was 0.824 ± 0.008, followed by that of the voting classifier (0.822 ± 0.007) and was greater than those of XGBoost (0.786 ± 0.014) and extra trees classifier (0.815 ± 0.008) (Fig. S1A). The random forest produced the best AP (0.397 ± 0.019), followed by voting classifier (0.395 ± 0.022), extra trees (0.371 ± 0.022), and XGBoost (0.335 ± 0.030) (Fig. S1B). The calibration curves showed all models were well-calibrated, but calibration might be greatly improved if the predicted risk is scaled down (Fig. 1C). The voting classifier produced the lowest brier score of 0.096 ± 0.008.

When testing the models on the Semarang dataset, the voting classifier had the best AUC of 0.674 followed by random forest, extra trees, and XGBoost with AUCs of 0.673, 0.668, and 0.643 respectively (Fig. 1A). XGBoost had the best AP of 0.533, followed by that of the voting

classifier (0.514), random forest (0.513) and extra trees classifier (0.510) (Fig. 1B). The voting classifier model was selected as the best classifier due to producing the best test AUC and second-best F1 score.

Figure 2 displays a SHAP summary plot for the voting classifier Jakarta model. Being asymptomatic was the most important feature ranked by the SHAP analysis and were strongly associated with lower predicted risk of COVID-19 infection. Of COVID-19 symptoms, cough, fever, and chills ranked among the most important features. Of the behavioral questions, frequent handwashing in different scenarios, such as after touching or disposing trash, after shaking hands with people or after touching animals was associated with lower risk of COVID-19 infection. Health workers that wore shoe covers, hazmat suits, face shields and medical gloves regularly were also at lower risk of COVID-19 diagnosis and these behaviors ranked among the most important features. Furthermore, health workers that wore PPE such as medical gloves when handling COVID-19 specimens were also at lower risk of COVID-19. PPE such as masks and gloves have been shown to reduce risk for COVID-19 infection [36]. Furthermore, the higher average density of people in the room most frequented by the health worker was highly associated with COVID-19 positivity by the Jakarta model. Health workers that frequently work in crowded areas may be at increased risk for COVID-19.

*Full Model*

Figure 3 displays test set ROC and PR curves as well as cross validation calibration curves for all the algorithms applied to the full patient cohort training set. The prevalence of positive classes in the validation and testing datasets was 15.3%. The best performing model during 5-fold stratified cross-validation was the random forest with the highest AUC (0.818 ± 0.022), followed by voting classifier, which is an ensemble of a random forest and extra trees classifier (0.817 ± 0.024), extra trees (0.813 ± 0.026), and the XGBoost classifier (0.795 ± 0.027) (Fig. S2A). The random forest also had the best AP (0.449 ± 0.028) followed by that of the voting classifier (0.446 ± 0.034), XGBoost (0.440 ± 0.017) and extra trees (0.438 ± 0.039) (Fig. S2B). The calibration curves, derived from validation folds of 5-fold cross validation, showed all models were relatively well-calibrated, with the voting classifier and extra trees classifier producing the lowest mean brier score of 0.116 ± 0.005 (Fig. 3C).

On the held-out test set, random forest produced the highest AUC of 0.831 as compared to voting classifier (0.828), extra trees (0.822), and XGBoost (0.776) (Fig. 3A). The random forest had the highest AP (0.428) compared to voting classifier (0.426), XGBoost (0.410) and extra trees (0.402) (Fig. 3B). The random forest was overall the best performing algorithm for the full model, due to high operating point predictive performance on the held-out test set with sensitivity, specificity and F1 score of 0.787, 0.770 and 0.515 respectively. Other performance metrics such as the positive and negative predictive values are displayed in Figure 3D.

Feature importance and model explainability were assessed through SHAP values from the training set predictions. Figure 4 displays the SHAP summary plot for the top 20 most important features in the full model. Like the Jakarta model, being asymptomatic and not handling COVID-19 specimens were ranked among the most important features for reduced COVID-19 risk. Surprisingly, health workers that performed aerosol-generating procedures on COVID-19 patients, such as tracheal intubation, non-invasive ventilation, tracheostomy, and swab collection among others, were also at lower risk of infection. This observation is likely attributable to more stringent PPE requirements and behavioral measures for health workers working closely with COVID-19 patients. Fever was highly predictive of COVID-19 infection,

as well as other common COVID-19 symptoms such as chills and cough. Health workers that wore surgical hoods and medical gloves during work were at lower risk of COVID-19 infection. Inadequate access to PPE has already been reported in Indonesia [38, 39]. Employees that regularly wore a hazmat suit at the health facility were at greatly reduced risk of being COVID-19 positive. Health facilities that reported a shortage of hazmat suits in Indonesia instructed health workers to wear thin plastic raincoats while transporting COVID-19 patients [40]. These limitations highlight the importance of a sufficient supply of PPE for the health workforce during the pandemic.

## DISCUSSION

This study investigated the capability of information collected via questionnaire including protective behavioral tendencies, COVID-19 vaccine status and information, symptoms, comorbidities, and working conditions among others to predict COVID-19 diagnosis for Indonesian health workers. We demonstrated that machine learning methods, specifically the random forest, XGBoost, extra trees and ensemble algorithms of these models were able to predict COVID-19 diagnosis with high performance. Models built on symptom data only have been found to be insufficient for application to clinical practice [41], perhaps due to the lack of behavioral, comorbidity, or other critical risk factors for COVID-19 infection. The models developed here incorporate many of these factors, and specifically include inputs on behavioral tendencies that are important in COVID-19 transmission and infection. The importance of behavioral tendencies is demonstrated by their high ranking in the SHAP summary plots for both models (Figs. 2 and 4), where behavioral tendencies ranked amongst the most important features for the models. Associations between features and COVID-19 positivity revealed by the models do not necessarily indicate causal relationships and must be interpreted in this context.

The full model performed well during training and testing with the random forest and was therefore chosen as the classifier of choice for the full model. The random forest performed better in terms of both AUC and AP than the extra trees, XGBoost and voting classifier models during 5-fold cross validation and on the test set. The best performing Jakarta model had poorer predictive performance on the test set as compared to that of the full model. The poorer performance of the Jakarta model on the test set of Semarang health workers is likely due to potential drifts and biases in the testing data relative to the training data. The purpose of this model was to assess performance on an independent population of health workers relative to the training data for the model. Although this model displays high test set sensitivity, reduced specificity compared to the full model may reduce adoption and applicability of models built on city-specific data to data from health workers in other cities. The full model displays generalizable performance to unseen data and may have higher potential to generalize to wider populations of health workers within Indonesia.

Improved protective behavioral tendencies, such as handwashing and sanitizing [42], were also associated with lower COVID-19 risk as reasoned by the models. The models we built are inclusive of many behavioral risk factors of COVID-19 infection among a myriad of other inputs and perform well on unseen data. Mass health worker testing, adequate PPE supply, self-isolation and quarantine, and education are the main recommendations previously issued for saving the frontline health workers during the pandemic [43]. CLM and our results may have the potential to assist in most of these guidelines, through assessing COVID-19 risk, prioritizing testing and thus isolation measures, and demonstrating the importance of utilizing PPE.

Furthermore, CLM can be used to reduce costs, as the survey can be provided to health facilities at no cost.

This study has several limitations. The first limitation is the self-reported nature of the survey, which poses risks of over or underreporting. Methods of fraud detection and error handling must be applied if using the model in real-time. The authority for hospitals to recommend their staff to be included in the study may introduce selection bias in the data collection process. Additionally, recall bias may be introduced in answers to retrospective questions in the survey, such as symptoms within the previous 14 days. Future work for the study includes collecting more data for these models, as well as investigating models for using CLM survey and other data to predict additional outcomes, such as hospitalization and mortality. Recruitment of health workers for the study also will expand to several other provinces within Indonesia.

Notwithstanding the limitations, our results demonstrate predictive capability for COVID-19 in health workers using machine learning. Our preliminary models showed high predictive performance, especially when trained and tested on similar population groups. When used in practice, CLM can be tuned by training on local populations that resemble target populations in which it will be used. The models can potentially be used to prioritize RT-PCR testing in regions where diagnostic resources are scarce. Allocating testing using the model predictions may lead to reductions in the challenges health workers in Indonesia are facing due to the pandemic. Our study may also have the potential to inform policy decisions regarding health worker PPE requirements, such as the promotion of the use of face shields and protective goggles.

# TABLES

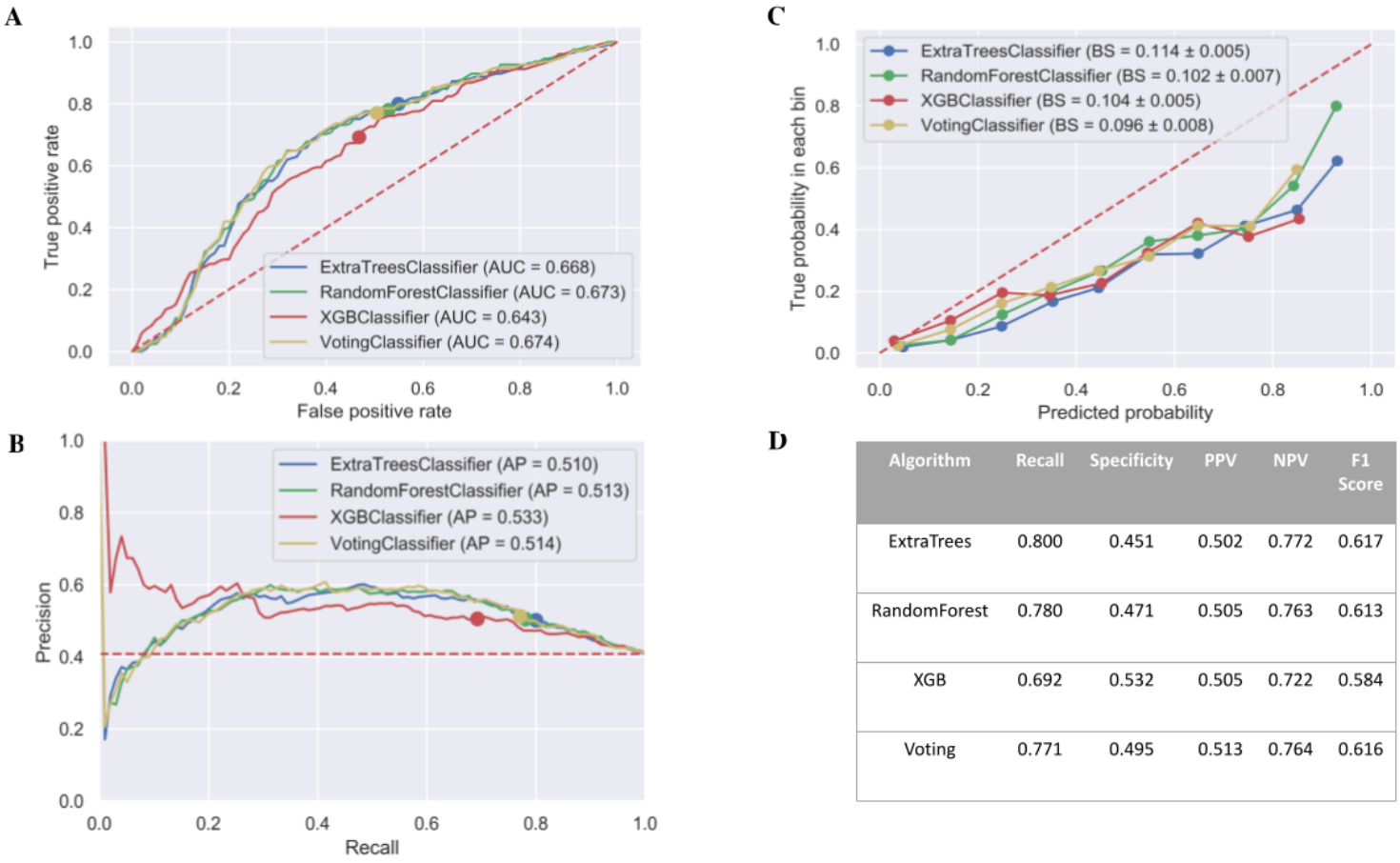
**Table 1** | Dataset descriptive statistics

| Variable | All sample [N=3,984] | | PCR test result | | | |
|---|---|---|---|---|---|---|
| | | | Negative [N=3,374] | | Positive [N=610] | |
| | Freq (Mean) | % (SD) | Freq (Mean) | % (SD) | Freq (Mean) | % (SD) |
| **(1) Protective Behaviors** | | | | | | |
| **Knowledge of standard PPE** | | | | | | |
| No | 81 | 2.03 | 74 | 2.19 | 7 | 1.15 |
| Yes | 3,903 | 97.97 | 3,300 | 97.81 | 603 | 98.85 |
| **Self-isolating after contact with COVID-19 patient** | | | | | | |
| No | 3,258 | 81.77 | 2,954 | 87.55 | 304 | 49.84 |
| Yes | 726 | 18.23 | 420 | 12.45 | 306 | 50.16 |
| **Performs aerosol-generating procedure on COVID-19 patient** | | | | | | |
| No | 2,784 | 69.87 | 2,316 | 68.63 | 468 | 76.72 |
| Yes | 1,200 | 30.13 | 1,058 | 31.37 | 142 | 23.28 |
| **Knowledge of 6 step hand washing techniques** | | | | | | |
| No | 108 | 2.71 | 94 | 2.79 | 14 | 2.30 |
| Yes | 3,876 | 97.29 | 3,280 | 97.21 | 596 | 97.70 |
| **Used a mask outside of the home** | | | | | | |
| Never | 64 | 1.61 | 58 | 1.72 | 6 | 0.96 |
| Sometimes | 361 | 9.06 | 307 | 9.10 | 54 | 8.85 |
| Often | 512 | 12.85 | 439 | 13.02 | 73 | 11.97 |
| Always | 3,047 | 76.48 | 2,570 | 76.16 | 477 | 78.2 |
| **Removed part or all of mask when met someone outside the home** | | | | | | |
| Never | 2,341 | 59.71 | 1,995 | 60.15 | 346 | 57.28 |
| Sometimes | 1,202 | 30.67 | 1,001 | 20.20 | 201 | 33.28 |
| Often | 184 | 4.70 | 151 | 4.56 | 33 | 5.46 |
| Always | 193 | 4.92 | 169 | 5.10 | 24 | 3.97 |
| **Visited closed, ventilated, or air-conditioned room** | | | | | | |
| Never | 1,065 | 26.74 | 940 | 27.87 | 125 | 20.49 |
| 1-3 times a month | 390 | 9.79 | 332 | 9.84 | 58 | 9.51 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1-3 times a week | 377 | 9.47 | 321 | 9.52 | 56 | 9.18 |
| 4-6 times a week | 1,385 | 34.75 | 1,139 | 33.74 | 246 | 40.33 |
| More than 6 times a week | 767 | 19.26 | 642 | 19.03 | 125 | 20.49 |
| **Had in-person meetings** | | | | | | |
| Never | 2,387 | 59.93 | 1,999 | 59.26 | 388 | 63.61 |
| 1-3 times a month | 978 | 24.53 | 857 | 25.38 | 121 | 19.84 |
| 1-3 times a week | 278 | 6.98 | 225 | 6.67 | 53 | 8.69 |
| 4-6 times a week | 201 | 5.05 | 168 | 4.98 | 33 | 5.41 |
| More than 6 times a week | 140 | 3.51 | 125 | 3.71 | 15 | 2.46 |
| **Had outdoor activities** | | | | | | |
| Never | 1,621 | 40.69 | 1,356 | 40.19 | 265 | 43.44 |
| 1-3 times a month | 1,215 | 30.50 | 1,039 | 30.79 | 176 | 28.85 |
| 1-3 times a week | 544 | 13.65 | 458 | 13.57 | 86 | 14.1 |
| 4-6 times a week | 323 | 8.11 | 286 | 8.48 | 37 | 6.07 |
| More than 6 times a week | 281 | 7.05 | 235 | 6.97 | 46 | 7.54 |
| **Used mass public transportation** | | | | | | |
| Never | 3,264 | 81.93 | 2,756 | 81.68 | 508 | 83.28 |
| 1-3 times a month | 375 | 9.41 | 325 | 9.63 | 50 | 8.2 |
| 1-3 times a week | 119 | 2.99 | 98 | 2.90 | 21 | 3.44 |
| 4-6 times a week | 114 | 2.86 | 102 | 3.02 | 12 | 1.97 |
| More than 6 times a week | 112 | 2.81 | 93 | 2.76 | 19 | 3.11 |
| **Used door to door transportation** | | | | | | |
| Never | 3,028 | 76.00 | 2,547 | 75.49 | 481 | 78.85 |
| 1-3 times a month | 545 | 13.68 | 468 | 13.87 | 77 | 12.62 |
| 1-3 times a week | 160 | 4.02 | 134 | 3.97 | 26 | 4.26 |
| 4-6 times a week | 125 | 3.14 | 113 | 3.35 | 12 | 1.97 |
| More than 6 times a week | 126 | 3.16 | 112 | 3.32 | 14 | 2.30 |
| **Boarded an airplane** | | | | | | |
| Never | 3,897 | 97.82 | 3,299 | 97.78 | 598 | 98.03 |
| 1 time | 44 | 1.1 | 37 | 1.10 | 7 | 1.15 |
| 2-4 times | 32 | 0.8 | 27 | 0.80 | 5 | 0.82 |
| More than 4 times | 11 | 0.28 | 11 | 0.33 | 0 | 0.00 |
| **Avoiding shaking hand with someone outside the home** | | | | | | |
| Never | 228 | 5.72 | 196 | 5.81 | 32 | 5.25 |
| Sometimes | 803 | 20.16 | 679 | 20.12 | 124 | 20.33 |
| Often | 831 | 20.86 | 701 | 20.78 | 130 | 21.31 |
| Always | 2,122 | 53.26 | 1,798 | 53.29 | 324 | 53.11 |
| **Maintaining physical distancing** | | | | | | |
| Never | 88 | 2.21 | 78 | 2.31 | 10 | 1.64 |
| Sometimes | 716 | 17.97 | 614 | 18.20 | 102 | 16.72 |
| Often | 1,082 | 27.16 | 902 | 26.73 | 180 | 29.51 |
| Always | 2,098 | 52.66 | 1,780 | 52.76 | 318 | 52.13 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Visiting other city or country** | | | | | | |
| No | 3,738 | 93.93 | 3,181 | 94.28 | 557 | 91.31 |
| Yes | 246 | 6.17 | 193 | 5.72 | 53 | 8.69 |
| **(2) Symptoms** | | | | | | |
| Asymptomatic | 2,450 | 62.74 | 2,375 | 70.38 | 125 | 20.49 |
| Anosmia or ageusia | 322 | 8.08 | 170 | 5.04 | 152 | 24.92 |
| Diarrhea | 142 | 3.57 | 94 | 2.79 | 48 | 7.87 |
| Chills | 467 | 11.72 | 233 | 6.91 | 234 | 38.36 |
| Runny nose | 591 | 14 | 351 | 10.41 | 240 | 39.34 |
| Headache | 674 | 16.92 | 403 | 11.95 | 271 | 44.43 |
| Cough | 706 | 17.73 | 410 | 12.16 | 296 | 48.52 |
| Fever | 419 | 10.52 | 190 | 5.63 | 229 | 37.54 |
| Sore throat | 463 | 11.62 | 278 | 8.24 | 185 | 30.33 |
| Myalgia | 538 | 13.51 | 337 | 9.99 | 201 | 32.95 |
| Malaise | 366 | 9.19 | 232 | 6.88 | 134 | 21.97 |
| Fatigue | 237 | 5.95 | 138 | 4.09 | 99 | 16.23 |
| Phlegm production | 231 | 5.80 | 132 | 3.91 | 99 | 16.23 |
| Dizziness | 161 | 4.04 | 96 | 2.85 | 65 | 10.56 |
| Stomach pain | 128 | 3.21 | 85 | 2.52 | 43 | 7.05 |
| Nausea or vomiting | 261 | 6.55 | 169 | 5.01 | 92 | 15.08 |
| Delirium | 15 | 0.38 | 11 | 0.33 | 4 | 0.66 |
| Breathing difficulties | 72 | 1.81 | 44 | 1.30 | 28 | 4.59 |
| Tingling or numbness | 35 | 0.88 | 22 | 0.65 | 13 | 2.13 |
| Watery eyes | 69 | 1.73 | 24 | 0.71 | 45 | 7.38 |
| Skin rash | 29 | 0.73 | 20 | 0.59 | 9 | 1.48 |
| Loss of appetite | 201 | 5.05 | 108 | 3.20 | 93 | 15.25 |
| Night sweats | 128 | 3.21 | 60 | 1.78 | 68 | 11.15 |
| Chest pain | 60 | 1.51 | 35 | 1.04 | 25 | 4.10 |
| Sudden hearing loss | 22 | 0.55 | 11 | 0.33 | 11 | 1.80 |
| Heartbeat irregularities | 56 | 1.41 | 40 | 1.19 | 16 | 2.62 |
| Acute seizure | 1 | 0.03 | 1 | 0.03 | 0 | 0.00 |
| **(3) Comorbidities** | | | | | | |
| Not comorbidities | 3,639 | 91.34 | 3,091 | 91.61 | 548 | 89.84 |
| Lung disease | 81 | 2.03 | 62 | 1.84 | 19 | 3.11 |
| Immune system disease | 17 | 0.43 | 11 | 0.33 | 6 | 0.98 |
| Hypertension | 109 | 2.74 | 94 | 2.79 | 15 | 2.46 |
| Other endocrine disease | 8 | 0.20 | 8 | 0.24 | 0 | 0.00 |
| Diabetes | 19 | 0.48 | 15 | 0.44 | 4 | 0.66 |
| Pregnancy | 105 | 2.64 | 91 | 2.7 | 14 | 2.30 |
| Coronary heart disease | 24 | 0.60 | 22 | 0.65 | 2 | 0.33 |
| Cancer | 6 | 0.15 | 6 | 0.18 | 0 | 0.00 |
| Stroke | 3 | 0.08 | 3 | 0.09 | 0 | 0.00 |
| Liver disease | 9 | 0.23 | 5 | 0.15 | 4 | 0.66 |

## FIGURES

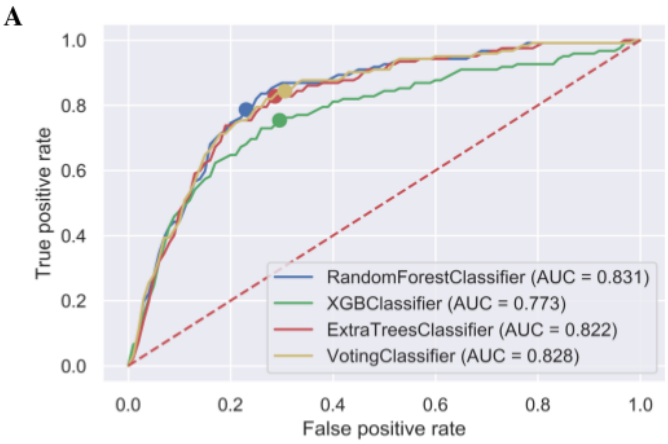


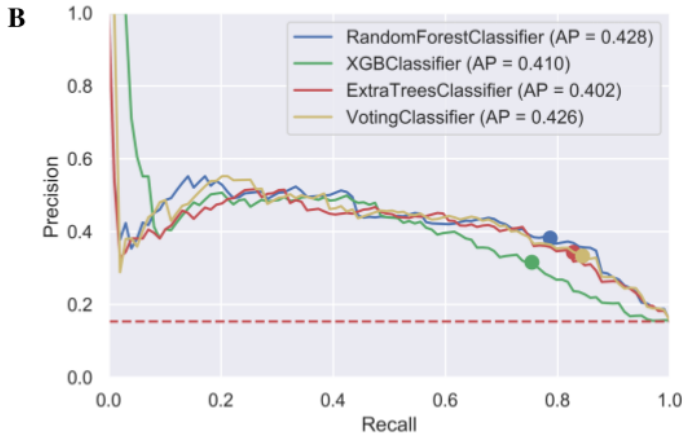

D

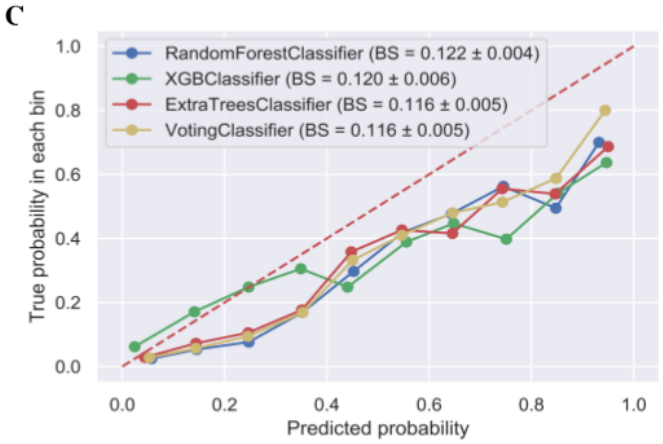
| Algorithm | Recall | Specificity | PPV | NPV | F1 Score |
|---|---|---|---|---|---|
| ExtraTrees | 0.800 | 0.451 | 0.502 | 0.772 | 0.617 |
| RandomForest | 0.780 | 0.471 | 0.505 | 0.763 | 0.613 |
| XGB | 0.692 | 0.532 | 0.505 | 0.722 | 0.584 |
| Voting | 0.771 | 0.495 | 0.513 | 0.764 | 0.616 |

**Fig.1** | **Jakarta model performance. A**, ROC curve on the test set (Semarang data) for all classifiers. **B**, PR curve on test set for all classifiers. **C**, Calibration curve for all classifiers generated from validation sets during 5-fold cross validation. **D**, Test set performance statistics using operating threshold derived from validation sets during 5-fold cross validation.


**Fig.2** | **Jakarta model explainability.** SHAP analysis on training set predictions.




D

| Algorithm | Recall | Specificity | PPV | NPV | F1 Score |
|---|---|---|---|---|---|
| ExtraTrees | 0.828 | 0.714 | 0.344 | 0.861 | 0.486 |
| RandomForest | 0.787 | 0.770 | 0.382 | 0.844 | 0.515 |
| XGB | 0.754 | 0.705 | 0.316 | 0.811 | 0.446 |
| Voting | 0.844 | 0.694 | 0.333 | 0.867 | 0.478 |

**Fig.3** | **Full model performance. A**, ROC curve on the test set for all classifiers. **B**, PR curve on test set for all classifiers. **C**, Calibration curve for all classifiers generated from validation sets during 5-fold cross validation. **D**, Test set performance using operating threshold derived from validation sets during 5-fold cross validation.


**Fig.4** | **Full model explainability.** SHAP analysis on training set predictions.

## REFERENCES

[1] KawalCovid19. Informasi Terkini COVID-19 di Indonesia [Internet]. [cited 2021 Jun 2]. Available from: https://kawalcovid19.id/

[2] World Health Organization. Prevention, Identification and Management of Health Worker Infection in the Context of COVID-19 [Internet]. Available from: https://www.who.int/publications/i/item/10665-336265

[3] Lidwina A. 654 Tenaga Kesehatan Gugur Lawan Pandemi Covid-19 di Indonesia [Internet]. [cited 2021 Jun 2]. Available from: https://databoks.katadata.co.id/datapublish/2021/01/28/654-tenaga-kesehatan-gugur-lawan-pandemi-covid-19-di-indonesia

[4] Widadio NA. Coronavirus kills 647 health workers in Indonesia [Internet]. [cited 2021 Jun 1]. Available from: https://www.aa.com.tr/en/asia-pacific/coronavirus-kills-647-health-workers-in-indonesia/2125642

[5] Fitria Chusna Farisa. Setahun Covid-19: Upaya Indonesia Akhiri Pandemi, dari PSBB hingga Vaksinasi [Internet]. 2021 [cited 2021 Jun 17]. Available from: https://nasional.kompas.com/read/2021/03/02/10213641/setahun-covid-19-upaya-indonesia-akhiri-pandemi-dari-psbb-hingga-vaksinasi?page=all

[6] CSIS. Southeast Asia Covid-19 Tracker [Internet]. [cited 2021 Jun 17]. Available from: https://www.csis.org/programs/southeast-asia-program/projects/southeast-asia-covid-19-tracker

[7] Day M. Covid-19: identifying and isolating asymptomatic people helped eliminate virus in Italian village. BMJ [Internet]. 2020 Mar 23;m1165. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj.m1165

[8] Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. J Med Syst [Internet]. 2020 Aug 1;44(8):135. Available from: http://link.springer.com/10.1007/s10916-020-01597-4

[9] Xie J, Tong Z, Guan X, Du B, Qiu H, Slutsky AS. Critical care crisis and some recommendations during the COVID-19 epidemic in China. Intensive Care Med [Internet]. 2020 May 2;46(5):837–40. Available from: http://link.springer.com/10.1007/s00134-020-05979-7

[10] Grasselli G, Pesenti A, Cecconi M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy. JAMA [Internet]. 2020 Apr 28;323(16):1545. Available from: https://jamanetwork.com/journals/jama/fullarticle/2763188

[11] WHO. WHO provides one million antigen-detecting rapid diagnostic test kits to accelerate COVID-19 testing in Indonesia [Internet]. 2021 [cited 2021 Jun 17]. Available from: https://www.who.int/indonesia/news/detail/17-03-2021-who-provides-one-million-antigen-detecting-

rapid-diagnostic-test-kits-to-accelerate-covid-19-testing-in-indonesia

[12] WHO. Global partnership to make available 120 million affordable, quality COVID-19 rapid tests for low- and middle-income countries [Internet]. 2020 [cited 2021 Jun 18]. Available from: https://www.who.int/news/item/28-09-2020-global-partnership-to-make-available-120-million-affordable-quality-covid-19-rapid-tests-for-low--and-middle-income-countries

[13] Syambudi I. Pasokan Reagen PCR Menipis, Testing COVID-19 Terbengkalai [Internet]. 2021 [cited 2021 Jun 17]. Available from: https://tirto.id/pasokan-reagen-pcr-menipis-testing-covid-19-terbengkalai-ga6z

[14] BBC. Lonjakan Covid-19 di Indonesia diprediksi sampai awal Juli, daerah lain bisa menyusul Kudus [Internet]. 2021 [cited 2021 Jun 17]. Available from: https://www.bbc.com/indonesia/indonesia-57492990

[15] Our World in Data. Coronavirus (COVID-19) Testing - Statistics and Research [Internet]. [cited 2021 Jun 17]. Available from: https://ourworldindata.org/coronavirus-testing

[16] Tirto. Ridwan Kamil Kritik Pelacakan COVID-19 RI Jauh dari Standar WHO [Internet]. 2020 [cited 2021 Jun 17]. Available from: https://tirto.id/ridwan-kamil-kritik-pelacakan-covid-19-ri-jauh-dari-standar-who-f9RL

[17] Mahendradhata, Yodi et al. “The Capacity of the Indonesian Healthcare System to Respond to COVID-19.” Frontiers in public health vol. 9 649819. 7 Jul. 2021, doi:10.3389/fpubh.2021.649819

[18] Hendarwan, Harimat et al. “Assessing the COVID-19 diagnostic laboratory capacity in Indonesia in the early phase of the pandemic.” WHO South-East Asia journal of public healthvol. 9,2 (2020): 134-140. doi:10.4103/2224-3151.294307

[19] Scohy, Anaïs et al. Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* **129,** (2020). doi:10.1016/j.jcv.2020.104455

[20] Wang, S. *et al.* A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *Eur. Radiol.* (2021) doi:10.1007/s00330-021-07715-1.

[21] Shoer, S. *et al.* A Prediction Model to Prioritize Individuals for a SARS-CoV-2 Test Built from National Symptom Surveys. *Med* **2**, 196-208.e4 (2021).

[22] Tostmann, A. *et al.* Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among health workers, the Netherlands, March 2020. *Eurosurveillance* **25**, (2020).

[23] Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digit. Med.* **4**, 3 (2021).

[24] Li, W. T. *et al.* Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. *BMC Med. Inform. Decis. Mak.* **20**, 247 (2020).

[25] Bayat, V. *et al.* A Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Prediction Model From Standard Laboratory Tests. *Clin. Infect. Dis.* (2020) doi:10.1093/cid/ciaa1175.

[26] Kukar, M. *et al.* COVID-19 diagnosis by routine blood tests using machine learning. (2020).

[27] Feng C, Huang Z, Wang L, Chen X, Zhai Y, Zhu F, Chen H, Wang Y, Su X, H. S. & Al., E. A Novel Triage Tool of Artificial Intelligence-Assisted Diagnosis Aid System for Suspected COVID-19 Pneumonia in Fever Clinics. *medRxiv* (2020).

[28] Ran, Li et al. Risk Factors of health workers With Coronavirus Disease 2019: A Retrospective Cohort Study in a Designated Hospital of Wuhan in China. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **71,** 16 (2020).

[29] Arias, Ariadna V et al. Assessment of hand hygiene techniques using the World Health Organization's six steps. *Journal of infection and public health* **9,** 3 (2016).

[30] World Health Organization. Coronavirus disease (COVID-19) advice for the public: When and how to use masks [Internet]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks/

[31] Breiman L. Random Forests. *Machine Learning* **45**, 5-32 (2001).

[32] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach Learn **63**, 3–42 (2006).

[33] Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project. 2013.

[34] Chen T and Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

[35] Bergstra J & Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* **13**, 281–305 (2012).

[36] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017; Long Beach, CA.

[37] He H, Bai Y, Garcia E. A. and Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[38] Centers for Disease Control and Prevention. How to Protect Yourself & Others [Internet]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html

[39] Yunus F, Andarini S. Letter from Indonesia. *Respirology* **25,** 1328-9 (2020). doi: 10.1111/resp.13953

[40] Legido-Quigley, Helena & Asgari-Jirhandeh, Nima. (2018). Resilient and people-centred health systems: Progress, challenges and future directions in Asia. World Health Organization. Regional Office for South-East Asia.

[41] Callahan, A., Steinberg, E., Fries, J.A. et al. Estimating the efficacy of symptom-based screening for COVID-19. *npj Digit. Med.* **3,** 95 (2020). https://doi.org/10.1038/s41746-020-0300-0

[42] Huang, F et al. COVID-19 outbreak and health worker behavioural change toward hand hygiene practices. *The Journal of hospital infection* **111,** 27-34 (2021). doi:10.1016/j.jhin.2021.03.004

[43] Nagesh, Shubha, and Stuti Chakraborty. "Saving the frontline health workforce amidst the COVID-19 crisis: Challenges and recommendations." *Journal of global health vol*. 10,1 (2020): 010345. doi:10.7189/jogh-10-010345

[44] Dyer O. COVID-19: Indonesia becomes Asia's new pandemic epicentre as delta variant spreads. *BMJ* (2021).