DEVELOPMENT OF PATIENTS TRIAGE ALGORITHM FROM NATIONWIDE COVID-19 REGISTRY DATA BASED ON MACHINE LEARNING

Se Young Jung
Department of Family Medicine
Seoul National University Bundang Hospital
Bundang 13620, Republic of Korea
imsyjung@gmail.com

Min Sue Park
Department of Mathematics
Pohang University of Science and Technology
Pohang 790-784, Republic of Korea
minsuepark@postech.ac.kr

Hyeontae Jo
Department of Mathematics
Pohang University of Science and Technology
Pohang 790-784, Republic of Korea
jht0116@postech.ac.kr

Hyung Ju Hwang*
Department of Mathematics
Pohang University of Science and Technology
Pohang 790-784, Republic of Korea
hjhwang@postech.ac.kr

September 21, 2021

ABSTRACT

Prompt severity assessment model of confirmed patients who were infected with infectious diseases could enable efficient diagnosis and alleviate the burden on the medical system. This paper provides the development processes of the severity assessment model using machine learning techniques and its application on SARS-CoV-2 patients. Here, we highlight that our model only requires basic patients' basic personal data, allowing for them to judge their own severity. We selected the boosting-based decision tree model as a classifier and interpreted mortality as a probability score after modeling. Specifically, hyperparameters that determine the structure of the tree model were tuned using the Bayesian optimization technique without any knowledge of medical information. As a result, we measured model performance and identified the variables affecting the severity through the model. Finally, we aim to establish a medical system that allows patients to check their own severity and informs them to visit the appropriate clinic center based on the past treatment details of other patients with similar severity.

1 Introduction

Recently, machine-learning models for various purposes have emerged, such as early detection and diagnosis of infectious diseases, and monitoring of treatment [1]. More specifically, [2, 3, 4, 5, 6, 7] provided deep learning models for detecting COVID-19 symptoms or for monitoring the patient's state using X-ray or CT scan datasets. Moreover, blood samples from the hospital can be used for evaluating the patient's severity [8, 9, 10, 11]. In order to secure the entire medical system of a nation, it is crucial to triage COVID-19 patients based on severity. In particular, most young COVID-19 patients don't need to be admitted to the hospital. For the self-quarantine population of COVID-19, accurate severity-assessment tools are necessary to appraise health status everyday.

In this paper, we propose a machine learning model that predicts the mortality of SARS-CoV-2 based on questionnaires written by patients. Indeed, many studies have been conducted to study the effect of Covid-19 based on self-reported symptoms. Multiple studies have considered self-reported symptoms as important features. For instance, [12] used linear regression to track potential SARS-CoV-2 infection based on patients' self-reported symptoms, and [13] developed an

*Corresponding Author

app which records self-reported symptoms that existing labor-intensive public health workflows may efficiently benefit from.

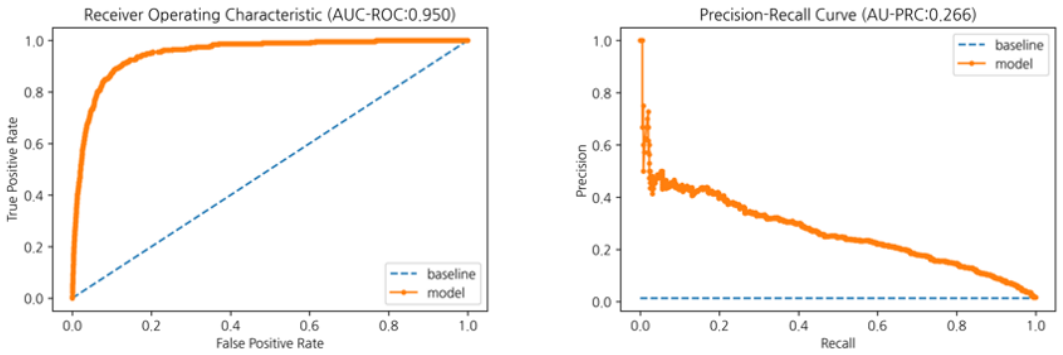
In contrast, [14] utilized Charlson comorbidity index (CCI) to predict mortality rate by using simple logistic regression, and analyzed the impact of comorbidity by measuring adjusted odds ratio. Yet, instead of simple logistic regression, we have utilized state-of-the-art machine model algorithms to predict mortality rate and derive interpretability of the prediction models.

Similar to our direction, [15] developed a machine-learning model that could determine whether or not a person is infected with the SARS-CoV-2 virus. The features they used mainly consist of eight types of data: sex, age, clinical symptoms (cough, fever, sore throat, shortness of breath, and headache), and whether a patient directly contacted other confirmed patients or not.

However, we found that the structural stability of individual SARS-CoV-2 could be affected by the temperature and humidity of the atmosphere [16]. In addition, hospitalization rates may vary depending on access to medical resources and the severity of previous diseases [17]. For these reasons, we utilized additional features such as the date of symptom onset (in months), the area of residence (in longitude and latitude coordinates) and the underlying symptoms of patients. A detailed description is given in the Method section (see also Figure 5, 6, 8 and Table 1).

2 Results

We split the dataset into training sets and test sets with a $80:20$ ratio, and the model was evaluated on the test set. The model achieved AUC-ROC (Area Under the Curve of Receiver Operating Characteristic) score of 0.950, F1-score of 0.861, recall as 0.807, precision as 0.923, and specificity as 0.936. Since the total number of the test set is 29,895 and there are 398 positives in the test set, the fraction of positives is 0.013, which is the baseline for AU-PRC (Area Under the Precision-Recall Curve) score. The model achieved AU-PRC score of 0.266 which has much outperformed the baseline score of 0.013. The general ROC curve and PR curve are shown in Figure 1.



(a) ROC Curve

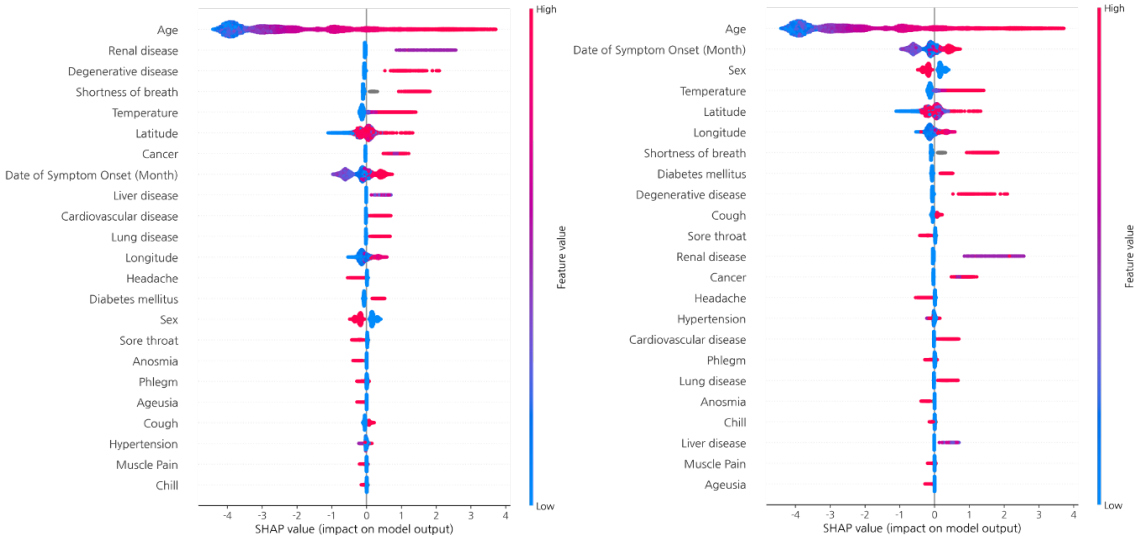
(b) Precision-Recall Curve

Figure 1: (a) ROC Curve and (b) Precision-Recall Curve

Feature importance measured by SHAP is depicted in Figure 2. In the SHAP analysis, age is proved to be the most important prognostic factor for hospitalization. Body temperature is also an important risk factor. Previous diseases before COVID-19 infection are important risk factors, such as renal disease, degenerative disease, cancer, liver, cardiovascular, and lung disease. Among initial symptoms of patients, shortness of breath is shown to be an important risk factor.

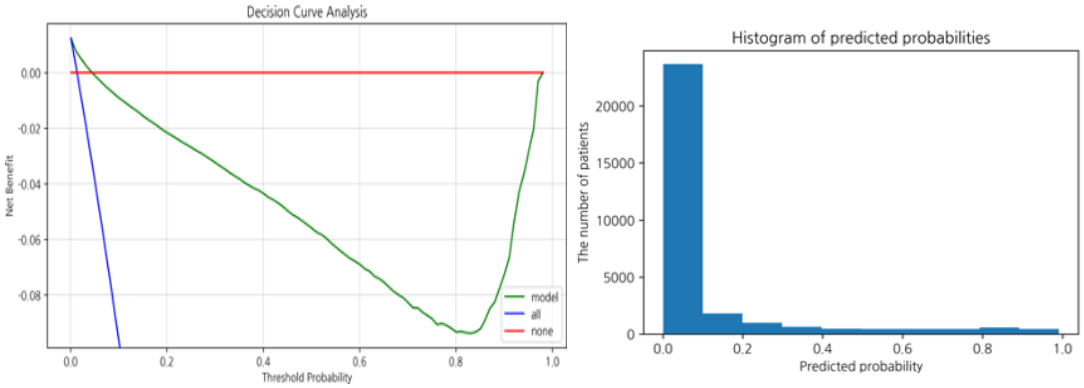
Decision curve analysis (DCA), as depicted in Figure 3, provides the range of threshold probabilities in which a prediction model shows the value and magnitude of benefit [18]. In this research, the threshold determines whether a self-quarantine patient should be hospitalized or not. The threshold should be set depending on the medical and economic environment of a country in which the model is implemented. DCA identified the optimal range of threshold in which net benefit does not fall below zero. In our model, the optimal range of threshold by the DCA is from 0 to 0.04.

We also investigated the types of medical institutions visited by patients according to their predicted mortality probabilities as shown in Figure 4. Firstly, we divided the test set into three groups: patients with predicted mortality probabilities less than 0.05, those that lie between 0.05 and 0.5, and those greater than 0.5. Then, for each group, we analyzed the types of medical institutions that the patients in-person for each group.



(a) (b)

Figure 2: **Feature importance plot.** SHapley Additive exPlanations (SHAP) beeswarm plot for the model predicting mortality rate of COVID-19. Features in the plot are sorted in descending order from top to bottom (a) by their maximum absolute and (b) by their mean absolute SHAP values. The explanation for each patient is represented by a single dot on each row, and the original feature values are represented by their colors.



(a) Decision curve analysis (b) Histogram of predicted probabilities

Figure 3: (a) Decision curve analysis and (b) the histogram of predicted probabilities.

3 Methods

3.1 Data description

Ranging from February 2020 to July 2021, the dataset is collected by the Korea Disease Control and Prevention Agency (KDCA), a government-affiliated organization, for all Koreans who tested positive for SARS-CoV-2 in Polymerase Chain Reaction (PCR). There were about 14,900 patients who tested positive, of which 2000 (1.4%) died. The dataset consists of about 149,000 patients and the features are mainly composed of three types of patient data; (i) basic personal information, (ii) types of first symptoms and (iii) underlying diseases. The detailed description of these features is given in Table. 1. As mentioned in the introduction, the area of residence is included in the data feature because it affects the virus' degree of activation, and medicalization scale. The dataset is labeled as whether the patient is dead or alive, and it is highly imbalanced because there are 2000 dead patients out of the total 149471 patients (98.7% imbalance ratio).

[1]https://github.com/swistakm/python-gmaps

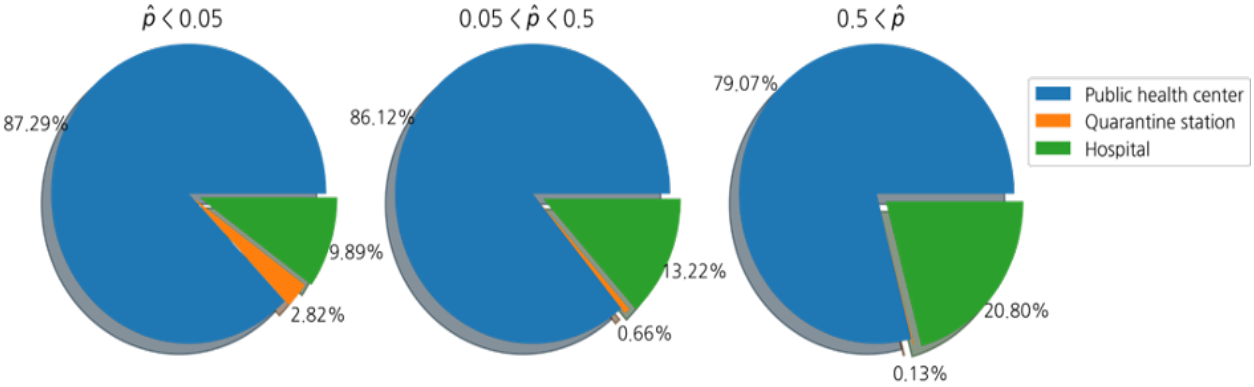


Figure 4: **Pie chart of types of medical institutions according to the patients' mortality probabilities.** Basically, since public health centers are the first places where patients receive the PCR test in general, the proportion of public health centers among the medical institutions where patients get treated is high. Nevertheless, the proportion of hospitals in the pie chart gets higher if the mortality rate of patients increases, as expected.

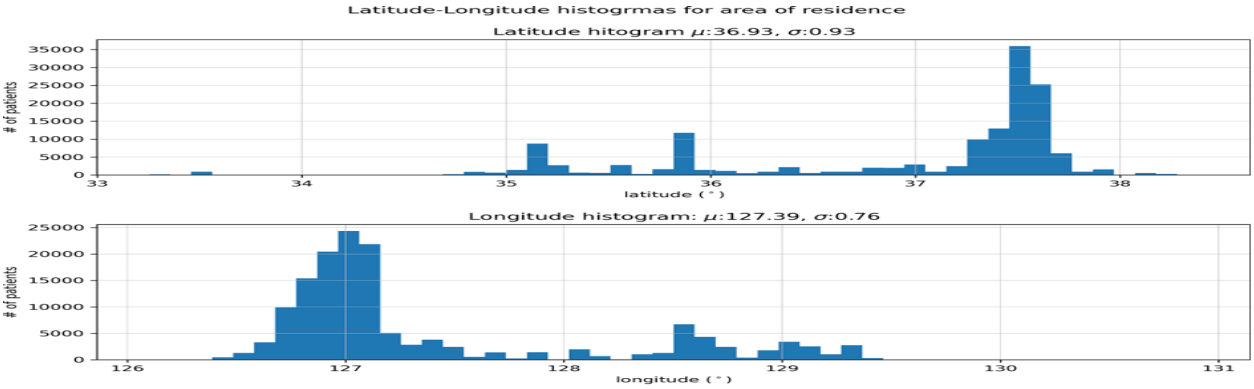


Figure 5: **Histogram of patient distribution by latitude (top) and longitude (bottom).** The x-axis represents the value of latitude (longitude), and the y-axis is the number of patients at that location. μ and σ in the title denote the mean and the standard deviation, respectively. Even though discrepancies between the actual area of residence and lat/long pair exist, they were ignored because such cases are rare.

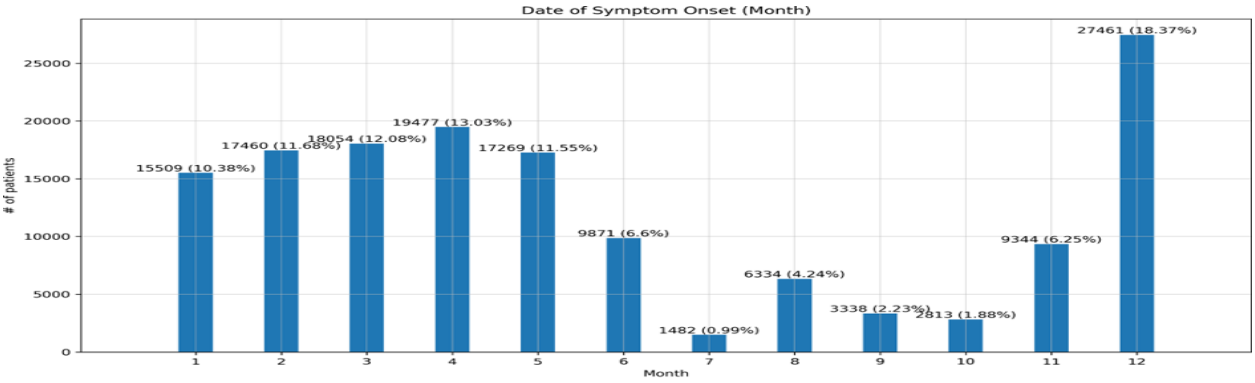


Figure 6: **Bar plot of the cumulative number of confirmed cases per month.** The size of each bar standing on the x-axis means the number of patients in that month. We marked the number of patients and its portion $(\%)$ on the top of the bar.

Type		Feature Name	Values	Value description	Missing value ratio
Basic personal information		Sex	Male: 1	$n = 75,073$ (50.25%)	0
			Female: 0	$n = 74,398$ (49.77%)	
		Age	Integer	$\mu : 44.36$ $\sigma : 20.27$	0
		Area of Residence	Latitude	Figure 5	$3,485$ (2.33%)
			Longitude		
		Date of Symptom Onset (Month)	Integer	Figure 6	$1,059$ (0.7%)
		Body Temperature	Normal: 0	Figure 7	0
			High Temperature: Quartile (Categorical)		
Symptoms (Fig. 8)	Respiratory rleated	Cough	Yes: 1	$n = 34,201$ (22.88%)	$15,273$ (10.22%)
			No: 0	$n = 99,997$ (66.90%)	
		Phlegm	Yes: 1	$n = 17,108$ (11.45%)	
			No: 0	$n = 109,120$ (78.34%)	
		Sore throat	Yes: 1	$n = 25,078$ (16.78%)	
			No: 0	$n = 109,120$ (73.00%)	
		Shortness of breath	Yes: 1	$n = 1,962$ (1.31%)	
			No: 0	$n = 132,236$ (88.47%)	
	Non-respiratory related	Muscle Pain	Yes: 1	$n = 24,017$ (16.07%)	
			No: 0	$n = 110,181$ (73.71%)	
		Headache	Yes: 1	$n = 16,337$ (10.93%)	
			No: 0	$n = 117,861$ (78.85%)	
		Chill	Yes: 1	$n = 17,227$ (11.53%)	
			No: 0	$n = 116,971$ (78.26%)	
		Ageusia	Yes: 1	$n = 4,846$ (3.24%)	
			No: 0	$n = 129,352$ (86.54%)	
		Anosmia	Yes: 1	$n = 5,498$ (3.68%)	
			No: 0	$n = 128,700$ (86.10%)	

Table 1: Characteristics of the dataset and the features used by the model in this study. Specifically, the area of residence for each confirmed patient is converted to floating point variables (latitude and longitude) using the Python Google maps[1] due to its large scale.

3.2 Evaluation

The model was evaluated on the test set using various metrics including AUC-ROC, AU-PRC, F1-score, precision, sensitivity and specificity. Moreover, we also performed a decision curve analysis on the model. ROC analysis provides

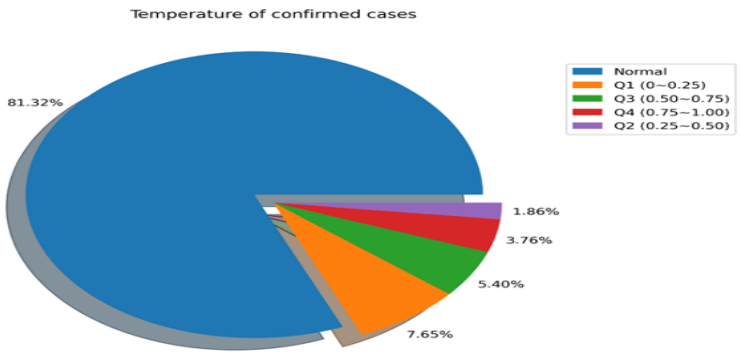


Figure 7: **Pie plot represents the ratio when divided by the body temperature of patients.** Patients who did not have fever symptoms were classified as 0. The rest of the patients who had fever symptoms were graded from 1 to 4 according to the quartile of their body temperature.

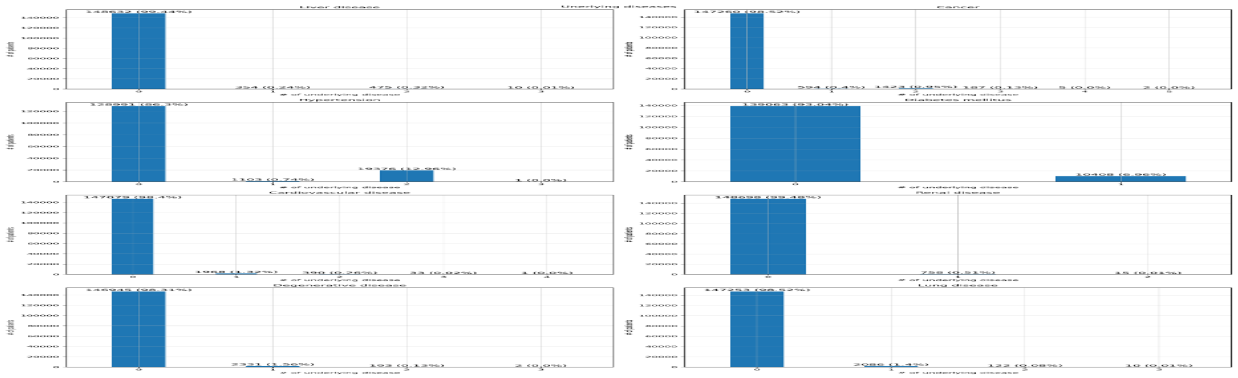


Figure 8: **Distribution of the number of underlying diseases.** 0 in the x-axis means no response to that disease, 1 or more means the number of diseases related to the title range.

information about diagnostic test performance: ROC curve consists of the True-Positive (TP) and False-Positive (FP) rate, and demonstrates the discriminatory ability of a binary classifier system by varying the discriminant thresholds. Simply speaking, the discriminatory ability of the test could be powerful when the vertex of the curve is closer to the upper left (high TP rate and low FP rate). In addition, the baseline for AUC-ROC is always going to be 0.5. On the other hand, PR curves plot the precision against the recall, and AU-PRC is especially useful for imbalanced data in a setting where we focus more on detecting the positive examples. Unlike AUC-ROC, the baseline for AU-PRC is equal to the fraction of positives. This means that obtaining an AU-PRC of 0.4 on a class with 10% positives is good but obtaining an AU-PRC of 0.6 on a class with 80% positives is unwanted [19].

3.3 Model description

We used a tree-based gradient boosting machine learning model, namely XGBoost (XGB) [20], with binary logistic objectives. This model is a decision-tree-based ensemble machine learning model that is well known for its powerful performance in classification problems in various fields. Since this is a tree-based model, it has the advantage of being able to process data with missing values well [21].

Another benefit of using gradient boosting algorithms is that it is straightforward to measure feature importance scores in prediction by calculating how useful each feature is in the construction of the weak learners within the model. Therefore, this method does not tell how positively or negatively the features affected the prediction, and does not consider the association relations among features in making predictions.

In contrast, originating from game theory, SHAP [22] algorithm was proposed to compute Shapley values [23] for each feature, where each Shapley value represents the impact that the feature to which it is associated, and predicts. When used for tree-based models, SHAP has a great advantage of being able to calculate Shapley values relatively quickly. Therefore, we have utilized it to identify the principal features in model prediction.

3.4 Hyperparameter tuning

We used a Bayesian optimization algorithm [24] to tune the hyperparameters of models. Unlike grid search that tries every pre-selected hyperparameter, Bayesian optimization is an approach that uses Bayes theorem to effectively direct the search for the optimal hyperparameter. Since hyperparameter spaces of tree-based models are huge, the time required to tune the models can be greatly reduced by using Bayesian optimization. For each step in the process of Bayesian optimization, we used 5-fold cross validation to measure validation accuracy.

4 Data availability

The datasets generated and analyzed during the study are provided by the KDCA and are not currently publicly available in compliance with the Personal Information Protection Act.

5 Code availability

The code for the statistical analyses will be made available upon request to the corresponding author. All code was written in Python 3.7.9 using the following open source software packages: scikit-learn v0.24.1, xgboost v1.4.2, bayesian-optimization v1.1.0.

6 Acknowledgments

Hyung Ju Hwang is funded by the National Research Foundation of Korea (NRF) grant funded by the South Korean government (MSIT) (No. 2017R1E1A1A03070105) and by the Institute for the Information and Communications Technology Promotion (IITP) grant funded by the South Korean government (MSIP) (No.2019-0-01906, Artificial Intelligence Graduate School Program (POSTECH)) and by the ITRC(Information Technology Research Center) support program(IITP-2018-0-01441).

References

[1] Raju Vaishya, Mohd Javaid, Ibrahim Haleem Khan, and Abid Haleem. Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):337–339, 2020.

[2] Ophir Gozes, Maayan Frid-Adar, Hayit Greenspan, Patrick D Browning, Huangqi Zhang, Wenbin Ji, Adam Bernheim, and Eliot Siegel. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *arXiv preprint arXiv:2003.05037*, 2020.

[3] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Ruixuan Wang, Huiying Zhao, Yunfei Zha, et al. Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[4] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *European radiology*, pages 1–9, 2021.

[5] Cheng Jin, Weixiang Chen, Yukun Cao, Zhanwei Xu, Zimeng Tan, Xin Zhang, Lei Deng, Chuansheng Zheng, Jie Zhou, Heshui Shi, et al. Development and evaluation of an artificial intelligence system for covid-19 diagnosis. *Nature communications*, 11(1):1–14, 2020.

[6] Narinder Singh Punn and Sonali Agarwal. Automated diagnosis of covid-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks. *Applied Intelligence*, 51(5):2689–2702, 2021.

[7] Farah E Shamout, Yiqiu Shen, Nan Wu, Aakash Kaku, Jungkyu Park, Taro Makino, Stanisław Jastrzębski, Jan Witowski, Duo Wang, Ben Zhang, et al. An artificial intelligence system for predicting the deterioration of covid-19 patients in the emergency department. *NPJ digital medicine*, 4(1):1–11, 2021.

[8] Cong Feng, Lili Wang, Xin Chen, Yongzhi Zhai, Feng Zhu, Hua Chen, Yingchan Wang, Xiangzheng Su, Sai Huang, Lin Tian, et al. A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected covid-19 pneumonia in fever clinics. *MedRxiv*, pages 2020–03, 2021.

[9] Kenji Ikemura, Eran Bellin, Yukako Yagi, Henny Billett, Mahmoud Saada, Katelyn Simone, Lindsay Stahl, James Szymanski, DY Goldstein, and Morayma Reyes Gil. Using automated machine learning to predict the mortality of

patients with covid-19: Prediction model development study. *Journal of medical Internet research*, 23(2):e23458, 2021.

[10] Ae-Young Her, Youngjune Bhak, Eun Jung Jun, Song Lin Yuan, Scot Garg, Semin Lee, Jong Bhak, and Eun-Seok Shin. A clinical risk score to predict in-hospital mortality from covid-19 in south korea. *Journal of Korean medical science*, 36(15), 2021.

[11] Miguel Marcos, Moncef Belhassen-García, Antonio Sánchez-Puente, Jesús Sampedro-Gomez, Raúl Azibeiro, Pedro-Ignacio Dorado-Díaz, Edgar Marcano-Millán, Carolina García-Vidal, María-Teresa Moreiro-Barroso, Noelia Cubino-Bóveda, et al. Development of a severity of disease score and classification model by machine learning for hospitalized covid-19 patients. *PloS one*, 16(4):e0240200, 2021.

[12] Cristina Menni, Ana M Valdes, Maxim B Freidin, Carole H Sudre, Long H Nguyen, David A Drew, Sajaysurya Ganesh, Thomas Varsavsky, M Jorge Cardoso, Julia S El-Sayed Moustafa, et al. Real-time tracking of self-reported symptoms to predict potential covid-19. *Nature medicine*, 26(7):1037–1040, 2020.

[13] Hannah A Burkhardt, Pascal S Brandt, Jenney R Lee, Sierramatice W Karras, Paul F Bugni, Ivan Cvitkovic, Amy Y Chen, and William B Lober. Stayhome: A fhir-native mobile covid-19 symptom tracker and public health reporting tool. *Online Journal of Public Health Informatics*, 13(1), 2021.

[14] Soo Ick Cho, Susie Yoon, and Ho-Jin Lee. Impact of comorbidity burden on mortality in patients with covid-19 using the korean health insurance database. *Scientific reports*, 11(1):1–9, 2021.

[15] Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj digital medicine*, 4(1):1–5, 2021.

[16] A Sharma, B Preece, H Swann, X Fan, RJ McKenney, KM Ori-McKenney, S Saffarian, and MD Vershinin. Structural stability of sars-cov-2 virus like particles degrades with temperature. *Biochemical and biophysical research communications*, 534:343–346, 2021.

[17] Wayne J Riley. Health disparities: gaps in access, quality and affordability of medical care. *Transactions of the American Clinical and Climatological Association*, 123:167, 2012.

[18] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

[19] Andreas Beger. Precision-recall curves. *Social Science Electronic Publishing*, 2016.

[20] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[21] Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

[22] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[23] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[24] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.