Adv Syst Sci Appl 2023; **04**; 148-155
Published online at https://ijassa.ipu.ru.

Comparison to the Proposed Hybrid Model and Machine Learning Techniques for Survival Prediction of Corona, Infected Patients

Md. Asadullah[1*], Md. Murad Hossain[1],
Md. Matiur Rahman Molla[2], Md. Matiur Rahaman[1]

[1] *Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh*

[2] *Islamic University, Kushtia, Bangladesh*

Abstract: SARS-CoV-2, a novel coronavirus discovered in Wuhan, China is spreading quickly and has a high incidence rate around the globe. As a result, everyone on the planet is having difficulty adjusting to the effects of Corona and is unable to foresee the devastation and disaster caused by COVID-19. In this work, we predict the survival status of patients infected with coronavirus using three distinct machine learning (ML) techniques: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). We also assess the classification performances of these algorithms. Here, we put out a hybrid model and evaluated it against the three previously discussed machine learning techniques. The outcomes demonstrated the 97.85% prediction accuracy of our suggested hybrid model. Aside from our suggested hybrid model, the random forest machine learning method demonstrated the highest accuracy of 94.62% among the three. Nonetheless, the prediction accuracy of the hybrid model outperforms that of the random forest and is significantly better than that of the other three ML techniques. The classification performances were assessed using the F-score, sensitivity, specificity, and precision metrics. Using 10-fold cross-validation, ROC assessments and confusion matrices produced by these machine learning algorithms were provided and examined. to assess the effectiveness of the classification. These machine learning algorithms' ROC assessments and confusion matrices are shown and examined by 10-fold cross-validation.

Keywords: COVID-19, machine learning, hybrid model, random forest, ROC Curve

1. INTRODUCTION

The entire world is struggling to adjust to Corona's effects and financial losses. They were unable to foresee the disaster and loss brought on by COVID-19. Numerous research investigations have been carried out to characterize the various models, strategies, innovations, and levels of awareness about the transmission of COVID-19. The categorization of verified COVID-19 cases associated with the epidemic diseases identified a significant obstacle to sustained advancement. The neural network's group method of knowledge handling (GMDH) form, which is linked to fake intelligence techniques, provided an explanation for binary classification modeling [1]. A Deep-learning Neural Network concept is used to provide a verification method. Using the concept of a deep-learning neural network, this system employs long short-term memory (LSTM) and a gated recurrent unit (GRU) for the last step of dataset training [2]. The XGBoost machine learning method was assisted by building a prognostic prediction model and emphasizing emphasis. As a result, 29 patients who were cleared after February 19th were tested [3]. AI-driven techniques assist in classifying COVID-19 outbreaks in a way that makes sense for their global propagation. The original goal of the research is to identify active learning-based cross-population train/test

* Corresponding author: asadullahstat@gmail.com

models that use multitudinal and multimodal data in order to detect COVID-19, similar to other healthcare difficulties [4]. WashKaro combines real sources of data with daily news and presents it in Hindi using NLP techniques, machine learning, and m-Health to provide an awareness-raising solution. Additionally, the appliance provides community-focused audio-visual content in local languages that has been hand-picked and verified by humans [5]. When implementing a mobile phone-based web survey, recommend utilizing machine learning techniques to be prepared to improve potential COVID-19 case identifications more quickly. Additionally, this will lessen the spread among those who are vulnerable [6].

An early screening model to distinguish COVID-19 pneumonia from influenza was created using deep learning techniques within a virus infection and healthy cases with lung CT pictures [7]. Using chest X-ray radiographs, three distinct convolutional neural network-based models are proposed for the identification of individuals afflicted with coronavirus pneumonia. Using 5-fold cross-validation, ROC assessments and confusion matrices using these explained models are provided and examined [8]. Described the machine learning-based classification of the extracted deep feature on chest X-ray images using patients with COVID-19 and pneumonia using ResNet152. It is also possible to anticipate the spread of the novel coronavirus (COVID-19) in likely patients using non-invasive and early forecasting of the virus by examining chest X-rays [9]. This study [10] reveals and uses machine learning-based CT radionics models to predict hospital stays for individuals with pneumonia associated to SARS-CoV-2 illness. Uses the publicly available COVID-19 chest X-ray dataset to illustrate how Bayesian Convolutional Neural Networks (BCNN) can quantify the uncertainty in Deep Learning solutions to improve the diagnostic performance of the human-machine combination and describe the uncertainty in prediction is strongly associated with the accuracy of the prediction [11]. This research assumes that Artificial Intelligence's deep learning methods could afford to extract COVID-19's specific graphical features and supply a clinical diagnosis, thus saving critical time for disease decline, to do the pathogenic test, through the support of COVID-19 radiographical changes in CT images [12]. Utilizing the recently developed Vaxign-ML machine learning technique in conjunction with Vaxign reverse vaccinology, COVID-19 vaccine candidates are predicted.

Six proteins, including the five non-structural proteins (nsp3, 3CL-pro, and nsp8-10) and S protein, were predicted to be adhesins after experiments with the entire proteome of SARS-CoV-2 [13]. The previous study suggested three well-known techniques: enhancing the currently limited data, using a panel selection method to select the most straightforward forecasting model among multiple models, and optimizing a personal prediction model's parameters for optimal accuracy. However, this study uses a method that supports the three benefits of knowledge mining from a small, poor dataset [14]. To ascertain the scope, duration, and end date of COVID-19 in China rather than using epidemiological models to analyze the dynamics of COVID-19 transmission in China. This work suggests real-time COVID-19 forecasting techniques influenced by fake intelligence (AI) [15]. Try testing the model's forecasting power in this article by training it on data from the 24th of January to the 3rd of March window, matching the forecasts up to the 1st of April. A one-week buffer window (March 25–April 1) was provided even for Italy and the Republic of Korea to validate the model's forecast. In the instance of the US, the model accurately depicts the progression of the infected curve and projects that the infection will stop by April 20, 2020 [16]. A case study utilizing fuzzy rule induction and CMC—which is improved by deep learning networks—is examined in order to gain more accurate stochastic insights on the evolution of the pandemic. Fuzzy rule induction approaches are used in conjunction with a deep learning-based CMC, as opposed to relying solely on uniform and uncomplicated assumptions for an MC [17]. Throughout this work, a novel bio-inspired metaheuristic simulates the transmission and infection of healthy individuals by the coronavirus. To replicate coronavirus behavior as closely as feasible, relevant parameters such as super-spreading rate, traveling rate, and re-infection probability are added to the model [18].

Incorporating city-to-city links, a mathematical framework is constructed to determine the total number of secondary cases caused by the imported cases, as well as the number of imported cases of the novel virus from epidemic resources. Additionally, to replicate outbreaks in several cities, a meta-population compartmental model was constructed and backed by a classical SIR technique [19]. This study compares the prediction outcomes of three distinct mathematical models based on many parameters and multiple locations. While the Gompertz model's fitting effect is also superior to the Bertalanffy model's, the logistic model's fitting effect is also the best [20]. Three biomarkers were chosen by machine learning methods in order to show the survival of specific patients by identifying susceptible predictive biomarkers of the severity of the disease. In order to prioritize patients at the highest risk and maybe lower the rate, this research proposes a straightforward and workable method [21]. Studies have shown that patients with cardiovascular and metabolic problems were more likely to contract COVID-19 and that the infection worsened. This investigation aims to determine the relationship between metabolic and cardiovascular illnesses using COVID-19 [22]. Nonetheless, the aforementioned studies discuss diverse mathematical strategies, artificial intelligence models, neural network models, and verification and validation methods for the prediction of patients impacted by Corona. Some articles even used a statistical, stochastic, and mathematical model to describe how corona transmission occurs. However, a small number of studies—particularly those using hybrid models—used machine learning methods. No such noteworthy research paper attempted to use machine learning tools to create a survival model for COVID-19 patients. In this work, we developed a hybrid model and focused on forecasting the survival of coronavirus-infected individuals utilizing currently available machine learning devices. We have made an effort to more closely evaluate our suggested hybrid model with the available machine learning tools. In this study, we also attempt to quantify the effectiveness of different machine learning technologies in terms of classification, validation, and accuracy. Lastly, we try to demonstrate that the employed data set is well described by our suggested hybrid model, demonstrating strong performance in terms of accuracy, precision, sensitivity, specificity, F1 Score, and AUC.

2. MATERIALS AND METHODS

2.1 Dataset Collection and Processing

We used metadata from "https://github.com/ieee8023/covid-chestxray-dataset" in this paper. There are 312 patients in our dataset, and each patient has additional data (referred to as a variable). The following unique characteristics are included in our dataset: patient ID, finding/type of pneumonia, age (years of the patient), sex (male or female), survival status (yes/no), offset (number of days), Went_icu (Yes (Y) if the patient was in the intensive care unit (ICU) or critical care unit (CCU) at any point during this sickness or No (N) or blank if unknown), Iintubated (Yes (Y) if the patient was intubated (or ventilated) at any time during this illness, etc. There are 109 female patients and 203 male patients among the 312 patient records in this collection. This is a list of all the metadata fields along with COVID-19-related explanations. We forecast the survival status of patients infected with Corona based on the characteristics. Version 3.6.3 of the R program is used for data processing and analysis.

2.2 Proposed Hybrid Model

The idea behind hybrid or ensemble techniques is that a combination of several single models can often produce effective discriminatory rules. We are going to build ensembles of machine learning algorithms. We integrate predictions from four caret models via stacking. This may indicate that the models are proficient in multiple domains, enabling a novel

classifier to determine how to extract the best performance and accuracy from each model. In this instance, we will utilize the random forest method to combine the predictions, which is why the hybrid model shows the highest accuracy when compared to all or any machine learning software alone.

2.3 Experimental Setup

Figure 2.1 outlines the several stages that our investigation covered. We pre-processed the dataset for our predictive model in the first stage. The dataset is then divided into training and testing datasets as the next phase. Three well-known classification methods as well as a hybrid model were used in our investigation. As a result, the most accurate prediction model gets approved to be used as a future model.

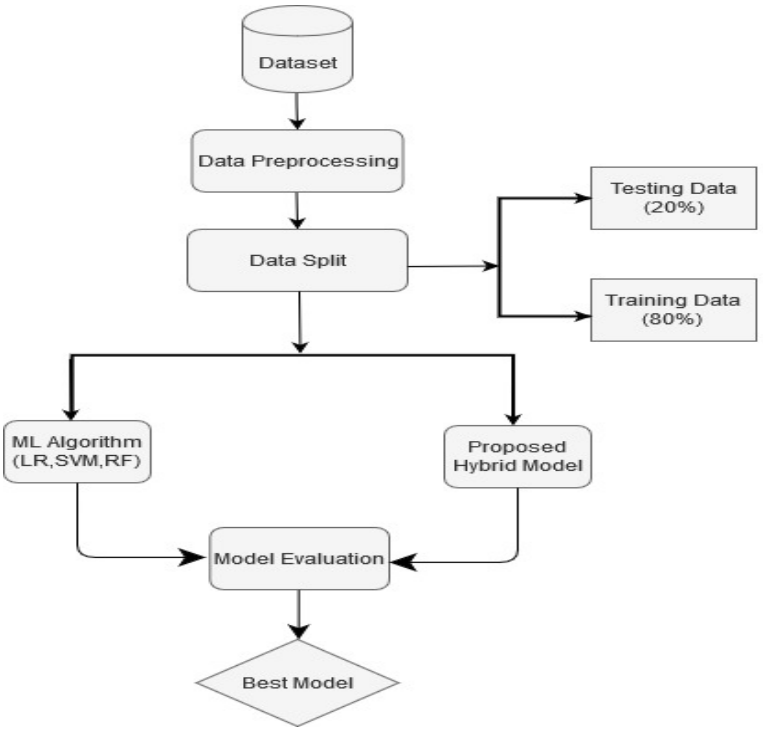

Fig.2.1. Flowchart for Proposed Hybrid Model

2.4. Machine Learning Technique

2.4.1 Support vector machines (SVM)

Support-vector machines (SVMs, support vector networks) are supervised learning models with corresponding learning algorithmic programs used for regression analysis and compartmentalization through simulated data in machine learning techniques. In addition to performing linear variety, Support Vector Machines (SVMs) can also do nonlinear classification by implicitly transforming their inputs into high-dimensional feature spaces, a technique known as the kernel trick. The support vector machine category was created by Vladimir Vapnik and Alexey Chervonenkis in an effort to transfer a linearly separable hyperplane and split the dataset into two groups. Assuming perfect data separation, let's proceed. After that, we can maximize the subsequent:

Minimize $\| \omega \|^2$, subject to:

$$(w.x_i + b) \geq 1, \text{if } y_i = 1$$

$$(w.x_i + b) \leq -1, \text{if } y_i = -1$$

The last two constraints can be compacted to:

$$y_i(w.x_i + b) \geq 1$$

A quadratic algorithm called linear SVM works effectively with datasets that are easily divided into two sections by a hyper-plane. However, datasets can occasionally be complex and challenging to categorize with a linear kernel.

2.4.2 Random Forest model

Even without hyper-parameter adjustment, the random forest method effectively employs a machine-learning algorithm that produces excellent results most of the time. Because of its ease of use and versatility, it's also one of the most used algorithms (used for classification and regression jobs). The random forest has a great advantage when used to classification and regression tasks, which make up the majority of machine learning systems in use today. The hyperparameters of random forests are identical to those of decision trees and bagging classifiers. We may also use the regressor algorithm to carry out regression operations for random forest. The irregular forest classifier is made up of a combination of tree classifiers. Each tree casts a unit vote to identify the most prevalent input vector. Each classifier is constructed using an arbitrary vector that may be freely examined from the input vector.

2.4.3 Logistic Regression

There are significant similarities between logistic regression and linear regression. Nevertheless, their intended application remains the primary distinction. While logistic regression is utilized for classification tasks, linear regression methods are employed for value prediction. Let $Y_i, i = 1,2, \ldots \ldots N$ be a binary outcome $(0|1)$ from Bernoulli $(1, \pi_i)$ with $\pi_i = Pr[Y_i = 1]$. The logistic regression model can be defined as:

$$logit = [Pr(Y_i = 1|x_i)] = log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta^T x_i$$

2.5 Performance Evaluation

We evaluate the machine learning methods for classification based on the following criteria.

Technique Names	Formula
Accuracy =	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision =	$\frac{TP}{TP + FP}$
Sensitivity/Recall =	$\frac{TP}{TP + FN}$
Specificity =	$\frac{TN}{TN + FP}$
F1-Score =	$2 * \frac{Precision * Recall}{Precision + Recall}$

3. RESULTS AND DISCUSSION

To test the three machine learning classification algorithms and one hybrid model to predict the survival status of patients infected with coronavirus, we conducted several analyses. The prerequisites for machine learning algorithms' specialized performance evaluation are shown in Table 3.1.

Table 3.1. Performance Measurements for Classification Technique

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
RF	0.9462	0.6923	0.9875	0.9875	0.8432	0.9514

SVM	0.9355	0.7692	0.9625	0.9625	0.8449	0.8683
LR	0.7634	0.1538	0.8625	0.8625	0.2611	0.8587
Hybrid	0.9785	0.8462	1.0000	1.0000	0.9167	0.9793

The results of the performance evaluations for the three machine learning methods used to predict survival are shown in Table 3.1. It also explains how a suggested hybrid model was developed. Out of the three machine learning methods, random forest performs the best with an accuracy of 94.62%, while logistic regression has the lowest accuracy at 76.34%. We evaluate the performance of three machine learning algorithms using the current standards, and the random forest model consistently outperforms the other two. However, compared to Random Forest, our hybrid model produced superior accuracy (97.85%). Out of all the machine learning techniques, our suggested hybrid model produced better results when compared to the other evaluation criteria. We also use the ROC curve to illustrate the performance of the various ML methods. The discriminant power in four distinct algorithms is shown in graph 2 for various colors. The color black represents our suggested hybrid model. Our suggested model has a strong ability to discriminate. The suggested model has an AUC value of.9793, which indicates that the predictions of the proposed hybrid model are 97.93% accurate, according to table 3.1's AUC value. If we take into account logistic regression as well, the AUC=.8587 indicates that logistic regression accurately predicts 85.87% of patients with corona infection who do not survive.

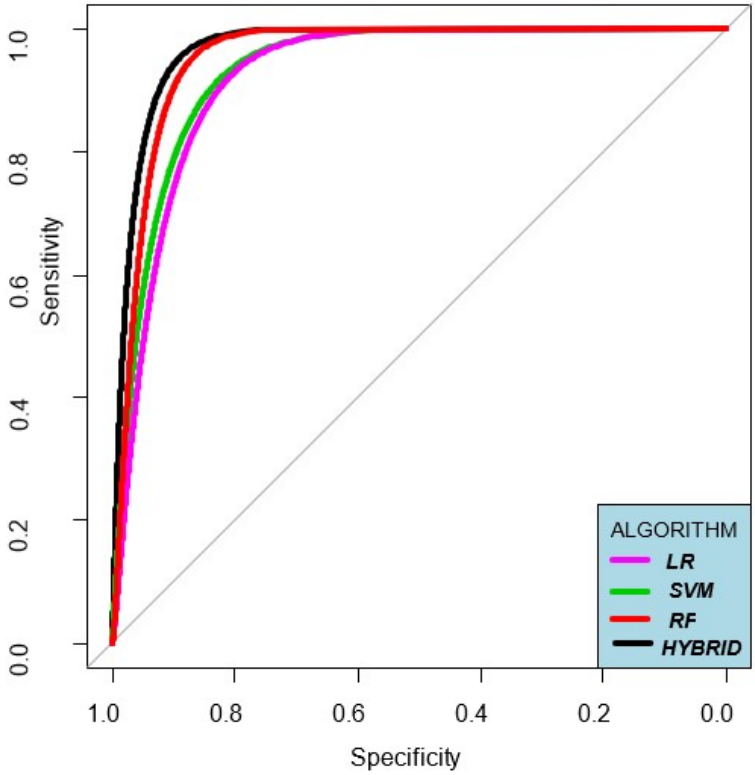

Fig. 2. ROC Curve for different machine learning algorithm compare to hybrid model

4. CONCLUSION

After examining the trial data, we came to the conclusion that our suggested hybrid model achieves the best accuracy, 97.85%, and AUC, 97.93%. As a result, our suggested approach can identify corona-infected patients who do not survive more precisely. In terms of classification accuracy or area under the ROC curve, our hybrid model performed the best. The suggested hybrid model's AUC score is close to 1, indicating a more accurate prediction. Our suggested hybrid model appears to perform better across the board for all performance assessment factors, according to the experimental results. We observe that the suggested

model with the lowest error has the highest accuracy. As the literature study has shown, we conclude that the criteria of a predictive model for corona-infected patient survival prediction has only been partially met. In conclusion, we can state that, when compared to the three machine learning approaches, our suggested hybrid model is the most predictive model for all machine learning methods and has a lower classification error.

ACKNOWLEDGEMENT

We thank all the co-authors who contributed to preparing this manuscript – especially the author Md. Asadullah has does analysis and contributed to interpretation. Secondly Md. Murad Hossain prepared an introduction and contributed to overall manuscript writing. Remaining co-authors help us reviewing all write-ups.

REFERENCES

[1] Asadullah, M., Hossain, M. M., Rahaman, S., Amin, M. S., Sumy, M. S. A., et al. (2023). Evaluation of machine learning techniques for hypertension risk prediction based on medical data in Bangladesh, *Indonesian Journal of Electrical Engineering and Computer Science*, **31**(3), 1794–1802.

[2] Bandyopadhyay, S. K. & Dutta, S. (2020). Machine Learning Approach for Confirmation of COVID-19 Cases: Positive, Negative, Death and Release, *Iberoamerican Journal of* Medicine, **03**, 172–177.

[3] Dandekar, R. & Barbastathis, G. (2020). Quantifying the effect of quarantine control in Covid-19 infectious spread using machine learning, *medRxiv*: 2020.04.03.20052084, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.04.03.20052084v1.

[4] Fong, S. J., Li, G., Dey, N., Gonzalez-Crespo, R. & Herrera-Viedma, E. (2020). Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak, *Int. J. Interact. Multimed. Artif. Intell.*, **6**(1) 132.

[5] Ghoshal, B. & Tucker, A. (2020). Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection, *arXiv*: 2003.10769, [Online]. Available: http://arxiv.org/abs/2003.10769.

[6] Hu, Z., Ge, Q., Li, S., Jin, L. & Xiong, M. (2020). Artificial intelligence forecasting of COVID-19 in China, *arXiv*:2002.07112, [Online]. Available: http://arxiv.org/abs/2002.07112.

[7] Hossain, M. M., Asadullah, M., Hossain, M. A., & Amin, M. S. (2022). Prediction of depression using machine learning tools taking consideration of oversampling. *Malaysian Journal of Public Health Medicine*, **22**(2), 244–253.

[8] Hossain, M. M., Asadullah, M., Rahaman, A., Miah, M. S., Hasan, M. Z., et al. (2021). Prediction on domestic violence in Bangladesh during the COVID-19 outbreak using machine learning methods, *Applied System Innovation*, **4**(4), 77.

[9] Jia, L., Li, K., Jiang, Y., Guo, X. & Zhao, T. (2020). Prediction and analysis of Coronavirus Disease 2019, *arXiv*: 2003.05447, [Online]. Available: http://arxiv.org/abs/2003.05447.

[10] Kumar, R., Arora, R., Bansal, V. & Sahayasheela, V. J. (2020). Accurate prediction of COVID-19 using chest x-ray images through deep feature learning model with smote and machine learning classifiers, *medRxiv*: 2020.04.13.20063461, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.04.13.20063461v1.

[11] Korkut, S., Göl, S., & Kilic, M. S. (2020). poly (pyrrole-co-pyrrole-2-carboxylic acid) / pyruvate oxidase based biosensor for phosphate: determination of the potential, and application in streams, *Electroanalysis*, **32**(2), 271–280.

[12] Li, B., Yang, J., Zhao, F., Zhi, L., Wang, X., et al. (2020). Prevalence and impact of

cardiovascular metabolic diseases on COVID-19 in China, *Clin Res Cardiol*, **109**(5), 531–538.

[13] Mamani, S., Nolan, D. A., Shi, L. & Alfano, R. R. (2020). Special classes of optical vector vortex beams are Majorana-like photons, *Opt. Commun.*, **464**, 125425.

[14] Martínez-Álvarez, F., Asencio-Cortés, G., Torres, J. F., Gutiérrez-Avilés, D., Melgar-García, L., et al. (2020). Coronavirus Optimization Algorithm: A bioinspired metaheuristic based on the COVID-19 propagation model, *arXiv*: 2003.13633, [Online]. Available: http://arxiv.org/abs/2003.13633.

[15] Ong, E., Wong, M. U., Huffman, A. & He, Y. (2020). COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning, *bioRxiv*: 2020.03.20.000141, [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.03.20.000141v2.

[16] Pandey, R., Gautam, V., Bhagat, K. & Sethi, T. (2020). A machine learning application for raising wash awareness in the times of COVID-19 pandemic, *arXiv:* 2003.07074, [Online]. Available: http://arxiv.org/abs/2003.07074.

[17] Pirouz. B., Haghshenas, S. S., & Piro, P. (2020). Investigating a serious challenge in the sustainable development process: Analysis of confirmed cases of COVID-19 (new type of Coronavirus) through a binary classification using artificial intelligence and regression analysis, *Sustain*, **12**(6), 2427.

[18] Qi, X., Jiang, Z., Yu, Q., Shao, C., Zhang, H., et al. (2020). Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study, *medRxiv*: 2020.02.29.20029603, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.29.20029603v1.

[19] Rao, A. S. R. S. & Vazquez, J. A. (2020). Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey in the populations when cities/towns are under quarantine, *Infect. Control Hosp. Epidemiol.*, **1400**.

[20] Santosh, K. C. (2020). AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data, *J. Med. Syst.*, **44**(5), 1–5.

[21] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., et al. (2020). A deep learning algorithm using CT images to screen for corona virus disease (COVID-19), *medRxiv*: 2020.02.14.20023028, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.14.20023028v5.

[22] Xu, X., Jiang, X., Ma, C., Du, P., Li, X., et al., deep learning system to screen Coronavirus Disease 2019 Pneumonia, *arXiv*: 2002.09334, [Online]. Available: http://arxiv.org/abs/2002.09334.

[23] Yuan, H.-Y., Hossain, M. P., Tsegaye, M., Zhu, X., Jia, P., et al. (2020). Estimating the risk on outbreak spreading of 2019-nCoV in China using transportation data, *medRxiv*: 2020.02.01.20019984, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.01.20019984v1.

[24] Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., et al. (2020). A machine learning-based model for survival prediction in patients with severe COVID-19 infection, *medRxiv*: 2020.02.27.20028027, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v3.

[25] Yan, L., Zhang, H.-T., Xiao, Y., Wang, M., Guo, Y., et al. (2020). Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan, *medRxiv*: 2020.02.27.20028027, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v2.

