International Journal of Mathematics, Statistics, and Computer Science
Vol. 1, 2023
ISSN: 2704-1077, DOI: https://doi.org/10.59543/ijmscs.v1i.7715

Severity Classification for COVID-19 Infections based on Lasso-Logistic Regression Model

Zainab Hussein Arif [1,2], Korhan Cengiz[3,4,*]
[1]Computer Technologies Engineering Department, Information Technology Collage, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq; Zhussian94@gmail.com
[2]College of Computer Science and Information Technology, University of Al-Qadisiyah, Al-Qadisiyah, Iraq
[3]Department of Computer Engineering, Istinye University, 34010, Istanbul, Turkey, korhan.cengiz@istinye.edu.tr
[4]Department of Information Technologies, Faculty of Informatics and Management, University of Hradec Kralove, Kralove, 500 03, Czech Republic

Received 25.02.2023, Revised 27.03.2023, Accepted 05.04.2023, Published 20.04.2023

ABSTRACT: The tremendous growth of the Covid19 epidemic in recent months is devastatingly affecting human civilization. Many different biomarkers are being studied to monitor the patient's health. This might mask the symptoms of various diseases, making it more challenging for a doctor to make a correct diagnosis or prognosis. Therefore, this study aimed to create several classes of prediction methods that can handle situations of varying severity (severe, moderate, and mild). Using machine learning, a Lasso-logistic regression model is developed. To create the Covid-19 clinical dataset, researchers enlisted the help of 78 patients from the Azizia main hospital sector, the Wasit Health Directorate, and the Ministry of Health. The results show that the proposed method is generally accurate to 85.9%. Deaths have been reduced thanks to the established prediction method that enables early detection of patients across three severity levels.

Keywords: Severity classification, COVID-19, Multiclass, Logistic regression.

1. INTRODUCTION

The current COVID-19 pandemic, caused by the SARS-CoV-2 virus, is a global public health emergency of unparalleled proportions. The majority of people now infected have mild to moderate symptoms, with around 2% experiencing severe symptoms. The overloading of hospitals and emergency rooms has led to a shortage of critical care beds [1]. Several recent studies have concentrated on aspects of critical care management [3] and on the severity of cases, which have a higher mortality rate than other cases [2].Overburdening medical facilities (like intensive care unit beds) with false positives for really serious conditions is a real concern. Further, individuals at high risk of death would be treated later if severe or urgent cases were not identified quickly. The earlier severe cases are recognized, the sooner resources may be mobilized and the intensity of therapy can be increased [4].

From what we know so far, COVID-19 infection can cause everything from no symptoms at all to fatal outcomes. Some individuals with mild illness rapidly worsen after initially displaying only minor symptoms, including fever, cough, and exhaustion [5]. Patients in the ICU frequently suffer from sepsis, respiratory failure, a condition known as acute respiratory distress syndrome (ARDS), cardiac failure, and septic shock [6, 7]. Patients at risk of developing a critical disease can be identified earlier, which might improve treatment delivery and decrease mortality. Several danger indicators are previously known [8, 9] from the existing literature. In the meanwhile, numerous organizations looked for ways to score and identify patients based on severity to aid clinicians in therapy and treatment. The COVID-GRAM score was suggested by Lian et al. [10] for predicting serious medical conditions among patients (death, ICU admission, or mechanical ventilation). This score is based on ten factors, and its ROC analysis yielded an Area Under the Curve (AUC) of 0.88. Ji et al. [11] used information from 208 Chinese patients to create the CALL (C = comorbidity, A = age, L = lymphocyte count, L = lactate dehydrogenase) score, which may be used to forecast the development of the disease. In their development cohort, their model with only four variables yielded an AUC of 0.91. Using data from 641, a group of researchers in the United States created a new score to foretell ICU admission or mortality. The AUC to predict ICU admission using their score was 0.74, while the AUC for predicting death was 0.82 [12]. Mortality from COVID-19-related pneumonia has been predicted using the widely used CURB-65 (C = Confusion, U = blood Urea nitrogen, R = Respiratory rate, 65 = age 65 or older) score and the Pneumonia Severity Index (PSI) in 681 laboratory-confirmed Turkish patients [13]. Multivariate regression models were used to produce all of these ratings.

Careful use of AI can aid in clinical decision making, and the field has already begun to address these vexing problems in healthcare [14]. Large data sets may be used by deep learning systems to spot danger, with surprising results [15]. These methods may be used to screen for TB in chest x-rays and cancer in mammograms, as well as to predict the risk of a myocardial infarction from retinal images [16]. Decision trees, a form of predictive analytics utilized here, have been used before to forecast the likelihood of contracting pneumonia [17].

Some studies have focused on analyzing patient blood samples in an effort to identify the most useful biomarkers for providing non-invasive identification solutions that protect healthcare workers from infection and provide a severity score for further treatment[18–20]. Using a machine learning (ML) based approach, the authors of [18] were able to predict the death rate of COVID-19 patients using only three biomarkers (LDH, CRP, and lymphocytes). Similar to the machine learning approach developed in Ref.[19], however, this one uses Eleven clinical indicators to predict COVID-19 severity. Based on 18 laboratory results, which were assessed by 6 distinct deep learning models, the prediction of COVID-19 was performed with deep learning models in a research published in [20]. Many of these studies have been discussed, but mostly from a technical or medical perspective. No prior effort has attempted to develop a dependable system that integrates both technological and medical perspectives, therefore helping clinicians arrive at a decisive assessment of the severity of a patient's condition. Studies should also focus on developing tolerated well drugs that could impact indicators and the disease course of moderate and mild patients to prevent long-term pulmonary damage[21]. Despite making up the majority of cases, moderate and mild cases have been largely ignored in previous attempts at research. In addition, the dataset utilized was collected from medical testing facilities and the patient's vital functions, which can result in high technical accuracy but can also confound seasonal flu and other viral flu [22, 23]. Further, in study [24], the authors make extensive efforts to predict the severity of covid-19 patients using a combination of different observations, but the classification performance is unsatisfactory.

This work's key contributions may be summed up as follows:

- Suggest a Lasso-Logistic Regression-based multi-class case severity prediction method for early-stage COVID-19 infections.
- A rich evaluation scenario based on well-established dataset and different assessment metrics.

2. METHODOLOGY

Dataset

Dataset was extracted from studies [24, 25] The data used in the development of the models was collected between April 8th and March 12th, 2020. Azizia Primary Healthcare Sector- Wasit Governorate-Iraq confirmed the diagnosis of Covid-19 in 78 patients under the care of medical experts. The loss of smell and taste was the most prevalent complaint among 78 patients (92.3 and 91.2%, respectively). Cough (58.97%), sore throat (57.69%), sneezing (56.41%), pleuritic chest discomfort (53.84%), diarrhoea (52.56%), and nasal congestion and rhinitis (42.30%) followed fever (67.95%).For the classification task, the suggested model will make use of all 27 features from the provided dataset.

Proposed Automated Multiclass Severity Prediction model

As shown in Figure 1, this study proposes a case-severity detection system as a techno-medical assistance system to assist doctors in making decisions about their patients' conditions.

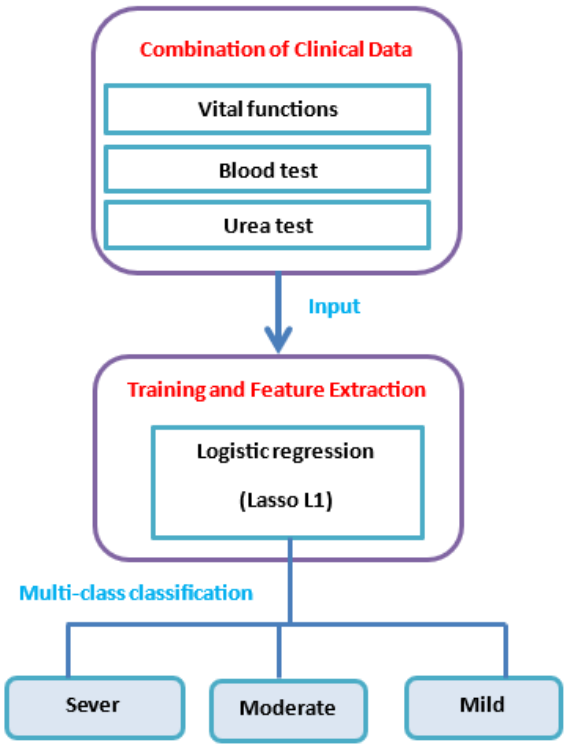

Figure 1: Lasso Logistic Regression Multiclass Severity Prediction Model.

Figure (1) depicts a possible activity diagram for the requested task. The suggested approach suggests classifying patients into severe, moderate, and mild categories based on levels of lymphocytes, CRP, and SPO2. The severity of the patient's condition is reflected in the model's predictions. By combining results from three distinct diagnostic procedures (blood, urine, and vital functions) and evaluating their systemic significance, as established by medical literature. These traits (i.e. biomarkers) were used to teach the classifier to make the necessary distinctions between the groups.

Patients in class 1 had the most serious condition, those in class 2 had moderate illness, and those in class 3 had the least severe. Accuracy, half total error rate (HTER), false positive rate (FPR), and false negative rate (FNR) were computed to assess the efficacy of the model. Another method of seeing the system's precision graphically is through the use of receiver operating characteristics (ROC).

HTER=(FPR+FNR)/2

Accuracy= 100 – HTER

Using logistic regression, the instances are classified. In most cases, the classifier receives the features extracted from a sample and uses them to assign a score to the sample. The classifier will decide whether or not to accept a sample based on whether or not its score is higher than certain threshold. The efficiency of the entire system can be impacted by the decision of which classifier to use.

Logistic Regression:

Classification is a common application of logistic regression, a supervised learning method based on the probability function. Historically, logistic regression has been employed in the field of statistics for the purpose of analyzing data and identifying the associations between a set of independent factors and a set of dependent variables. The result of a regression analysis is a continuous numerical number. The value of a discrete dependent variable is calculated based on the value of a non-linear independent variable. For a more involved cost function, the sigmoid function is employed to give probability between 0 and 1.

The goal is to use the probabilities indicated by Eq. (1) to represent the connections among the variables that are not dependent and the class. Refer to [26] for a more in-depth explanation of how multinomial logistic regression is used to carry out the classification.

$Y \approx C0 + C1X1 + C2X2 + \ldots + CpXp$

The preceding linear regression equation shows the relationship between the dependent variable Y and the independent variables X and the coefficient estimates C. In order to fit the model, a loss function called the total of squares is used. In order to minimize the loss function, the coefficients in the equation are optimized. If there is a lot of noise in the training set, the wrong coefficients will be chosen.

The minimum absolutely shrinkage and choice operator (LASSO) [26], Ridge(L2), and Elastic-Net function are just a few of the many kernel functions available for use in logistic regression. The severity of a case is determined using Lasso (L1) logistic regression in this study. Lasso regression requires all coefficients on all except the most important variables to be zero. The model is refined until just the most important factors remain. Similar to linear regression, lasso regression employs a "shrinkage" strategy to get the coefficients of decision closer to zero. The regression coefficients obtained by linear regression are those actually seen in the data. To prevent overfitting and improve performance across datasets, you may reduce the size of these coefficients using lasso regression. When there is a lot of multicollinearity in the dataset, or if you want to speed up the process of eliminating variables and selecting features, logistic regression is the way to go.

At first, we use the (Wasit) city database as a training set for the classifier. The compatibility scores are calculated using the Lasso(L1) kernel. The following is the formula for a logistic regression classifier using a Lasso kernel:

$$\sum_{i}^{n}(y_i - \widehat{y_i})^2 + \lambda \sum_{j}^{p} \|\beta_j\|$$

The first component in the preceding formula is the sum of squared residuals we're all familiar with, and the second component is a penalty whose size is proportional to the sum of all the coefficients. The Greek letter lambda before that total is a parameter used for tuning that determines the severity of the penalty. When it is zero, the OLS regression is the same as any other.

Performance Analysis Criteria

There were four measures used to assess how well the machine learning method worked. The measures offered in Eq. 1–4 are accuracy, precision, recall, and F1-Score.True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) each have their own numerical value.

$$\text{Accuracy} = (TN + TP)/(TN + TP + FN + FP) \quad (1)$$

$$\text{Recall} = TP/(TP + FN) \quad (2)$$

$$\text{Precision} = TP/(TP + FP) \quad (3)$$

$$\text{F1 - Score} = 2 \times ((\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})) \quad (4)$$

3. RESULTS AND DISCUSSION

The result section is divided into different directions especially into ratio of data used for training and testing purpose. First, we have measured the performance of classification task based on random sampling with 80% have been used for training process while the 20% left for testing as presented in Table 1. Second, a cross validation scenario based on different number of folds has been also used to determine if the classification will better or worse than random sampling scenario as presented in Table 2.

Table 1: Covid-19 severity classification based random sampling 80-20%.

Model	Auc	Accuracy	F1	Precision	Recall
Lasso-Logistic regression	93	85	85	85.2	85

Table 2: Covid-19 severity classification based on K-Fold cross-validation.

Model	Number of folds	Auc	Accuracy	F1	Precision	Recall
Lasso-Logistic regression	2	89.9	79.5	79.4	79.7	79.5
Lasso-Logistic regression	**3**	**92.6**	**85.9**	**85.8**	**86.1**	**85.9**
Lasso-Logistic regression	5	93	82.1	82.1	82.1	82.1
Lasso-Logistic regression	10	93	84.6	84.7	84.9	84.6

According to the Tables 1 and 2 the highest classification performance is based on random sampling but with minimum ratio comparing to three-fold cross validation scheme in training and testing process. This is maybe due to robust characteristics for used dataset and no missing value in data. However, each of

accuracy, F1, Precision, and Recall has same score based on random sampling ratio. In the same scenario, the highest value has scored by AUC with 93% as classification performance.

On the other hand, we have utilized 2, 3, 5, and 10 folds of cross-validation to evaluate the efficacy of the suggested model. In order to ensure our model's efficacy and correctness on unseen data, we use a resampling approach called cross-validation. A number of additional Machine Learning models are trained on subsets of the current input data set, and the evaluated models are then compared to one another. Our paper makes advantage of In K-Fold cross-validating, the data is split into k subsets, or the holdout technique is repeated k times, with k-1 of the subsets used as the training set and k as the validation set. Over k iterations, we take the average of the error to determine the model's overall effectiveness. To mention we did not used ”Holdout Method” in cross validation or random sampling to avoid under-fitting or over-fitting problem.

We have observed that the best classification performance is obtained based on 3 folds cross validation where the highest value for accuracy as well as other metrics have scored. A low classification performance was obtained based on other k-folds ration almost into all evaluation criteria except AUC.

A confusion matrix is built to show how effectively a model for classification (or "classifier") operates on a dataset when the real values have already been identified. While the unpredictability matrix itself is simple, the associated concepts are depicted in Figure 2.

		Predicted			
		Mild	**Moderate**	**Sever**	**∑**
Actual	**Mild**	94.7 %	2.7 %	0.0 %	**19**
	Moderate	5.3 %	81.1 %	27.3 %	**37**
	Sever	0.0 %	16.2 %	72.7 %	**22**
	∑	**19**	**37**	**22**	**78**

Figure 2: Confusion matrix

Using data ranging from mild to moderate, RF model achieved its best classification accuracy. These two groups have fared better than the rest of the school, with a classification ratio of 17% on average. This shows that Lasso-Logistic regression performs better than other methods, both in extreme and moderate cases. Table definitions of FP-rate and TP-rate are shown along the x-axis. Complex determination values are generated via Lasso-Logistic regression, which uses a random sampling of FP and TP values and their associated outcomes to make predictions about the target variable. The Y-axis of a ROC curve represents the true positive rate (TPR), while the X-axis represents the false positive rate (FPR). Therefore, the "ideal" position on the plot is located in the upper left corner, where the FPR is zero and the TPR is one.

Figure 3: ROC for Mild Class

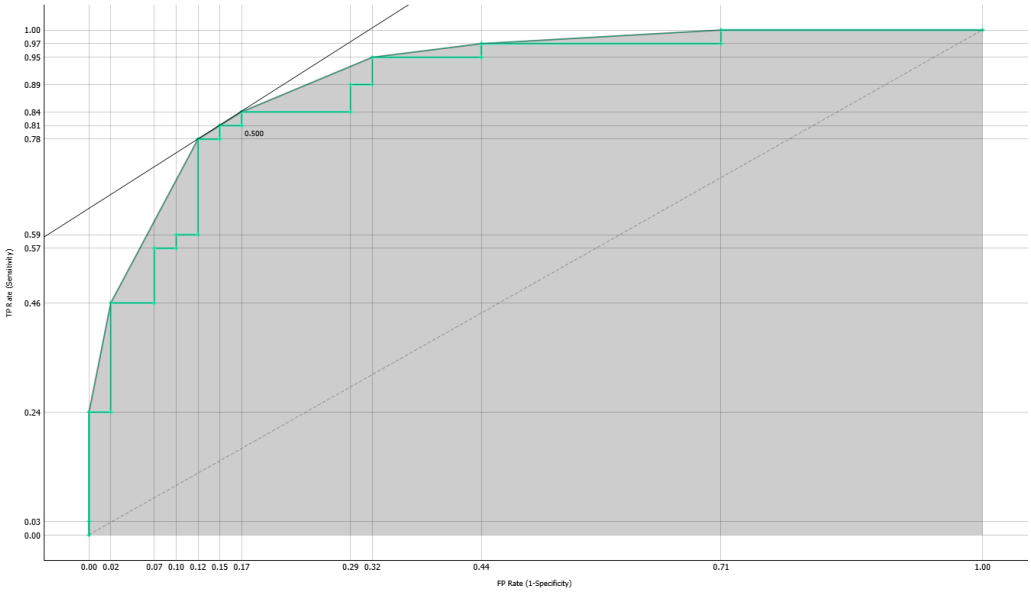

Figure 4: ROC for Moderate Class

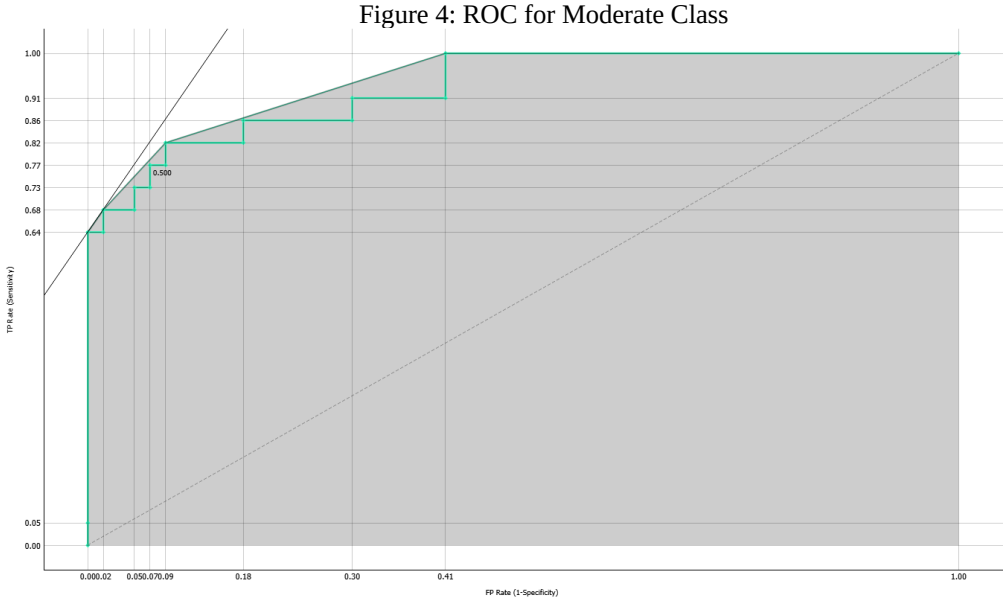

Figure 5: ROC for Sever Class

Figures 3, 4, and 5 showed that Lasso-Logistic regression predication model very significant to predict features of healthcare dataset based on Mild cases. The main reason behind that due to the predication power of proposed model. Sever class have showed better performance than moderate class but not in the same level with Mild class. However, the lowest performance was observed based on the data of moderate cases.

4. COMPARISON WITH STATE OF THE ART

Most studies involving medical data and imaging analysis need to start with benchmarking so researchers can evaluate how well new methods stack up against the status quo. Most benchmarks are run on a standard dataset or on methods designed for a related problem space.

For this reason, we used the most advanced and reliable COVID-19 classification strategies from the extant literature to conduct our benchmarking. The performance accuracy for severity prognosis has been fairly compared with state-of-the-art approaches in this study.

Table 3: Comparison with state of the art

Ref	Model	Accuracy
Proposed	Lasso-Logistic regression	**85.9%**

[27]	SVM	80%
[27]	Decision tree	70%
[24]	SVM	81.5
[24]	Decision tree	75

Table 3 displays the results of a comparison of each class and technique. Additionally, all prediction categories, including Mild, Moderate, and Severe, have been included in the comparisons. Everything considered state-of-the-art has been trained and evaluated using the same dataset. A severity prediction model using SVM and decision tree was used to determine the level of severity in [27, 24]. In contrast to [27] and [24], the suggested work achieved better performance when executing the benchmark workloads on the system. In spite of this, the suggested model outperforms all precedent research in terms of accuracy and precision optimisation rates attained by lightweight iteration in experimental settings.

5. CONCLUSIONS

Prediction algorithms based on Covid-19 are now one of the most used tools for the continuing epidemic. Using a total of 27 pathologically and technically validated characteristics from a variety of medical sources, we were able to zero down on three key markers for our study. Researchers in the healthcare industry used previous medical results to inform the development of a helpful technological system that can forecast the severity of an illness. A sample of 78 affected patients was obtained by the researchers themselves. Specialized doctors in accordance with the standards used by Iraqi hospitals classified patients as either "severe," "moderate," or "mild" upon admission. The integration of these technical and medical resources will help decrease health hazards and fatalities. We have created Lass0-logistic regression for COVID-19 prediction, which can predict multi-class of case severity (severe, moderate, and mild) with more than 85% accuracy, allowing for early intervention, diagnosis, and maybe a reduction in mortality for COVID-19 afflicted individuals. This work paves the way for further investigations into the effects of COVID-19 using alternative machine learning algorithms and the database.

Funding: This research received no external funding.
Conflict of interest: The authors declare no conflicts of interest.

REFERENCES

[1] Alyasseri, Z.A.A., et al., Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. 2022. 39(3): p. e12759.
[2] Kumar, N.M., et al., Artificial intelligence-based solution for sorting COVID related medical waste streams and supporting data-driven decisions for smart circular economy practice. Process Safety and Environmental Protection, 2021. 152: p. 482-494.
[3] Allioui, H., et al., A Multi-Agent Deep Reinforcement Learning Approach for Enhancement of COVID-19 CT Image Segmentation. 2022. 12(2): p. 309.
[4] Quiroz JC, Feng Y, Cheng Z, Rezazadegan D, Chen P, Lin Q, Qian L, Liu X, Berkovsky S, Coiera E, Song L, Qiu X, Liu S, Cai X. Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study JMIR Med Inform 2021;9(2):e24572. doi: 10.2196/24572
[5] CDC. Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19). Available online: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidancemanagement-patients.html (accessed on 28 August 2020).
[6] Zhou, F.; Yu, T.; Du, R.; Fan, G.; Liu, Y.; Liu, Z.; Xiang, J.; Wang, Y.; Song, B.; Gu, X.; et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. Lancet 2020, 395, 1054–1062.
[7] Grasselli, G.; Zangrillo, A.; Zanella, A.; Antonelli, M.; Cabrini, L.; Castelli, A.; Cereda, D.; Coluccello, A.; Foti, G.; Fumagalli, R.; et al. Baseline Characteristics and Outcomes of 1591 Patients InfectedWith SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. JAMA 2020.
[8] Rodriguez-Morales, A.J.; Cardona-Ospina, J.A.; Gutiérrez-Ocampo, E.; Villamizar-Peña, R.; Holguin-Rivera, Y.; Escalera-Antezana, J.P.; Alvarado-Arnez, L.E.; Bonilla-Aldana, D.K.; Franco-Paredes, C.; Henao-Martinez, A.F.; et al. Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis. Travel Med. Infect. Dis. 2020, 34, 101623.
[9] Yang, J.; Zheng, Y.; Gou, X.; Pu, K.; Chen, Z.; Guo, Q.; Ji, R.; Wang, H.; Wang, Y.; Zhou, Y. Prevalence of comorbidities and its e_ects in patients infected with SARS-CoV-2: A systematic review and meta-analysis. Int. J. Infect. Dis. 2020, 94, 91–95.
[10] Liang,W.; Liang, H.; Ou, L.; Chen, B.; Chen, A.; Li, C.; Li, Y.; Guan,W.; Sang, L.; Lu, J.; et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. JAMA Intern. Med. 2020.
[11] Ji, D.; Zhang, D.; Xu, J.; Chen, Z.; Yang, T.; Zhao, P.; Chen, G.; Cheng, G.; Wang, Y.; Bi, J.; et al. Prediction for Progression Risk in Patients with COVID-19 Pneumonia: The CALL Score. Clin. Infect. Dis. 2020.
[12] Zhao, Z.; Chen, A.; Hou, W.; Graham, J.M.; Li, H.; Richman, P.S.; Thode, H.C.; Singer, A.J.; Duong, T.Q. Prediction model and risk scores of ICU admission and mortality in COVID-19. PLoS ONE 2020, 15, e0236618.
[13] Satici, C.; Demirkol, M.A.; Altunok, E.S.; Gursoy, B.; Alkan, M.; Kamat, S.; Demirok, B.; Surmeli, C.D.; Calik, M.; Cavus, Z.; et al. Performance of pneumonia severity index and CURB-65 in predicting 30-day mortality in patients with COVID-19. Int. J. Infect. Dis. 2020, 98, 84–89.

[14] Shortliffe, E. H.; Sepulveda, M. J. (2018): Clinical decision support in the era of artificial intelligence. Journal of American Medical Association, vol. 320, no. 21, pp. 2199-2200.
[15] Sun, H.; McIntosh, S. (2018): Analyzing cross-domain transportation big data of New York City with semi-supervised and active learning. Computers, Materials & Continua, vol. 57, no. 1, pp. 1-9.
[16] Bai, Y.; Yao, L.; Wei, T.; Tian, F.; Jin, D. et al. (2020): Presumed asymptomatic carrier transmission of COVID-19. Journal of the American Medical Association.
[17] Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M. et al. (2015): Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721-1730.
[18] Yan, L., et al., An interpretable mortality prediction model for COVID-19 patients. 2020. 2(5): p. 283-288.
[19] Zhou, K., et al., Eleven Routine Clinical Features Predict COVID-19 Severity. 2020.
[20] Schwartz, D.A.J.A.o.p. and l. medicine, An analysis of 38 pregnant women with COVID-19, their newborn infants, and maternal-fetal transmission of SARS-CoV-2: maternal coronavirus infections and pregnancy outcomes. 2020. 144(7): p. 799-805.
[21] Xia, L., et al. The course of mild and moderate COVID-19 infections—the unexpected long-lasting challenge. in Open Forum Infectious Diseases. 2020. Oxford University Press US.
[22] Bai, X., et al., Predicting COVID-19 malignant progression with AI techniques. 2020.
[23] Pourhomayoun, M. and M.J.S.H. Shakibi, Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. 2021. 20: p. 100178.
[24] Dinar, A.M., et al., Towards Automated Multiclass Severity Prediction Approach for COVID-19 Infections Based on Combinations of Clinical Data. 2022.
[25] Hameed Abdulkareem, K., et al., Smart Healthcare System for Severity Prediction and Critical Tasks Management of COVID-19 Patients in IoT-Fog Computing Environments. 2022.
[26] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. J. of the Royal Stat. Soc., B, 58(1):267-288, 1996.
[27] Jiang, X., et al., Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. 2020. 63(1): p. 537-551.