Early Triage of COVID-19 patients exploiting Data-Driven Strategies and Machine Learning Techniques

Ji-Sung Park
Department of Applied Artificial Intelligence
Major in Bio Artificial Intelligence
Hanyang University
Ansan, Republic of Korea
jsdms316@hanyang.ac.kr

Gun-Woo Kim
Department of Computer Science
Gyeongsang National University
Jinju, Republic of Korea
kgwhsy@gmail.com

Hyeri Seok
Division of infectious diseases,
Department of Medicine
Korea University College of Medicine
Korea University Ansan Hospital
Ansan, Republic of Korea
hyeri.seok@gmail.com

Hong Ju Shin
Department of Thoracic and Cardiovascular Surgery
Korea University College of Medicine
Korea University Ansan Hospital
Ansan, Republic of Korea
babymedi@naver.com

Dong-Ho Lee
Department of Applied Artificial Intelligence
Hanyang University
Ansan, Republic of Korea
dhlee72@hanyang.ac.kr

2022 International Conference on Electronics, Information, and Communication (ICEIC) | 978-1-6654-0934-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICEIC54506.2022.9748839

***Abstract*— Since the first advent of SARS-CoV-2 in December 2019, Coronavirus disease (COVID-19) is still affecting the world. In the pandemic situation of the novel infectious disease, early detection of COVID-19 infection and severity for febrile respiratory patients is critical for efficient management of the medical system delivery system with limited medical personnel and facilities. Thus, we propose early triage exploiting data-driven strategical methods and machine learning techniques using the data of 5,628 admitted patients provided by Korea Central Disease Control Headquarters and 50 confirmed cases in Korea University Ansan Hospital. We proved validity of our data-driven strategies with machine learning models accuracy by doing 200 experiments and find out the features that affect COVID-19 through various feature selection in each medical inspection step. As a result, Stage 5 shows the results of blood test could affect to classify critical and severe cases obtaining precision of 0.2, 0.03 higher than without blood test results. But Stage 3 without blood test results achieved the highest accuracy of 0.88 showing possibility of early triage system without blood test. In conclusion, our triage system, based on data-driven strategies and machine learning techniques, can help in early detection and triage of COVID-19 patients.**

Keywords—COVID-19, Coronavirus, Triage, Data Management, Machine Learning

I. Introduction

Coronavirus Disease, which is caused by Severe Acute Respiratory Syndrome-Coronavius-2 (SARS-CoV-2), spreads through human-to-human transmission and has been spreading rapidly around the world [1]. As of August 9, 2021, the number of people worldwide has exceeded 200 million, of which more than 4 million have died [2,3]. Due to the rapid spread and risk of COVID-19, there is a shortage of medical facilities and medical personnel, which is causing many problems such as worsening of symptoms and even death because patients do not receive adequate treatment.

Therefore, we propose an early triage that can help medical personnel diagnose by using patient-centered data so that patients can receive appropriate treatment. Early triage is basically based on features that are obtained from patients' interviews or through simple examination. But most of these data are organized into categorical data type which much simplified patient information. In this situation, for more accurate classification, we adopted two data-driven strategies and utilized a machine learning technique that performed well.

Data-Driven Strategy I is to configure features according to the patients' information input order, so that the patient can quickly receive appropriate treatment in a situation of insufficient medical resources through the composition of features that affect the severity. Data-Driven Strategy II is focusing on the type of data and simplification of severity level of COVID-19. Transformation from numerical data type into categorical data type is for unifying the type of the dataset and simplification of severity level is effective to increase the high accuracy of machine learning model. Furthermore, we do four feature selection methods to get more affecting features of prediction. Finally, we have built an early triage system which predict the severity of COVID-19 patient using machine learning techniques.

II. Related Work

the COVID-19 pandemic has spread fast all over the world, several studies have proposed early triage systems recently. One study [5] proposed deep learning based early triage of critically ill COVID-19. This triage system based on survival COX model with neural networks. But for that it is necessary to build time-series dataset with expensive time cost. Another study [6] uses traditional ontology methods to support the analysis and tracking COVID-19 epidemic in Hubei, China. But this method needs the time cost work to build ontology named OntCov19 and its performance to the

COVID-19 pandemic has spread fast all over the world, several studies have proposed early triage systems recently. One study [5] proposed deep learning based early triage of critically ill COVID-19. This triage system based on survival COX model with neural networks. But for that it is necessary to build time-series dataset with expensive time cost. Another study [6] uses traditional ontology methods to support the analysis and tracking COVID-19 epidemic in

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01343, Artificial Intelligence Convergence Research Center(Hanyang University ERICA))

This research was supported by the MISP(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW) supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation)in 2021"(2018-0-00192)

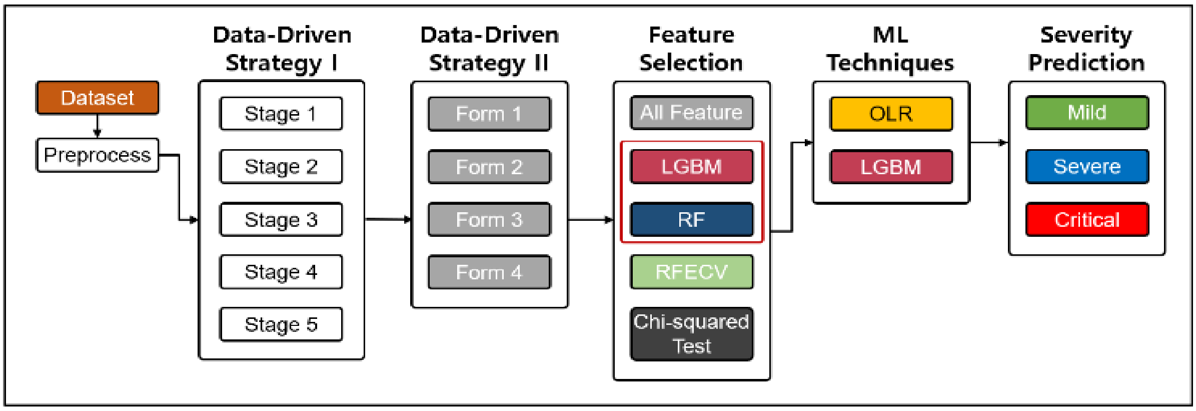

Fig. 1. Pipeline of early triage proposed

Hubei, China. But this method needs the time cost work to build ontology named OntCov19 and its performance to predict the patient is confirmed or not is lower comparing to machine learning methods.

The methods used in the studies are based on machine learning based approaches [7-9]. Even though many of confirmed cases are all over the world, it has difficulties to get data for ethical reasons. The data used in these studies collected from a consortium of few local hospitals. The data used in these studies are qualitative but require time-consuming clinical tests. Thus, it is hard to adopt these methods in early triage system. The study [10] is to find out features that can significantly affect COVID-19 called impact factors with COX's proportional hazard model with statistical analysis.

Our study is quite different from the above studies. Ours aims to focus on data-driven strategies with feature selection. In order to triage the patient's severity as quickly as possible, the clinical information that can be obtained from the patient sequentially is divided into 5 stages in sequence. We adopt form configuration, a kind of feature engineering for getting better performance of machine learning techniques.

III. Dataset

As of April 30, 2020, clinical epidemiologic information of 5,628 confirmed cases of COVID-19 in Korea who were released from quarantine is used as dataset for predicting the severity of COVID-19 at the initial stage of the examination in this paper. This dataset is provided by Korea Central Disease Control Headquarters on purpose of study. In addition, 50 confirmed cases from Korea University Ansan Hospital were included.

IV. Experiment

The pipeline of early triage of COVID-19 patients consists of 5 procedures for COVID-19 severity prediction as shown in Fig. 1. Dataset is filtered through Preprocess by elimination of useless features and missing values which may cause drop in performance of prediction. After that, dataset goes through two data-driven strategies we adopt. Strategy I is a method of changing the configuration of the feature set for each stage according to the patients' data inflow order systematically. It is to classify the severity of patients as soon as possible with high accuracy for the proper treatment. As most data in our dataset are binary type data, Strategy II is a method of transformation of input features which are not binary type into binary type to make the data types the same. Moreover, it simplifies classes of severity in early triage in accordance with Korean Response Guidelines [12]. Both of two methods in Data-Driven Strategy II are to ensure high prediction performance. Feature Selection is to classify the severity through less information by selecting the features that have a large influence on the severity classification. Finally, in the procedure of ML (Machine Learning) techniques selects the models with high performance for early triage. The explanation of each procedure is written in detail below.

A. *Preprocess*

This procedure has two steps for learning techniques.

- The first step is removing useless features for three reasons, as shown in TABLE I.
- The second step is to eliminate all the missing values in the dataset.

TABLE I. Feature Removal

	Case	***Feature***
1	Useless to predict severity	ID, CURSIT
2	Too many missing values	PREG, PREGGW
3	Impossible to get features at examination step	OUTCOME, PERIOD

B. *Data-Driven Strategy I: stage configuration*

TABLE II. Stage Configuration According to Data Inflow Order

No.	**Information Groups**					
1	Basic Data	Body Index	-	-	-	-
2	Basic Data	Body Index	Early Examination Findings	-	-	-
3	Basic Data	Body Index	Early Examination Findings	Clinical Findings	-	-
4	Basic Data	Body Index	Early Examination Findings	Clinical Findings	Underlying disease	-
5	Basic Data	Body Index	Early Examination Findings	Clinical Findings	Underlying disease	General Blood Test

Strategy I is a procedure of sequentially constructing clinical epidemiologic information obtained during hospital visits. It consists of 5 stage and sequential information groups compose each stage.

- Stage 1 : This stage contains two groups Basic Data and Body Index and total three features AGE and SEX in Basic Data and BMI in Body Index. These features are the most fundamental information and possible to get easily when patients visit hospital. (Total 5651 data)

- Stage 2 : In Stage 2, Early Examination Findings group is added to Stage 1. This group includes total 4 features obtained through simple examination like heart rate, temperature, systolic blood pressure and diastolic blood pressure. (Total 5484)
- Stage 3 : Clinical Findings group consists of 12 features of accompanying symptoms such as subjective fever, cough, sputum and so on. All of these features are binary type data which only shows whether COVID-19 patients have accompanying symptoms above or not. (Total 5483)
- Stage 4 : Total 11 features including chronic cardiac disease, asthma, chronic obstructive pulmonary disease are in Underlying Disease group. All of these features are binary data, like features in Clinical Findings group. Stage 4 contains this Underlying Disease group and the groups of Stage 3. (Total 5138)
- Stage 5: The features in General Blood Test group are added to previous stage composition. General Blood Test group has 5 features such as hemoglobin, hematocrit, platelets, white blood cell. (Total 3984)

As examination of features in each stage is usually done before time-consuming PCR test or treatment, One of the main purposes of our research is to classify the severity of the confirmed patient with high accuracy as quickly as possible through various machine learning methods using the information of each stage.

C. *Data-Driven Strategy II: form configuration*

This procedure consists of a total of 4 forms.

- Form 1 uses original dataset preprocessed in data preprocess step. It means that there is no change of the dataset when Form 1 is chosen.
- Form 2 transforms numerical data typed features into categorical data typed features.
- Form 3 adjust triage from 8 classes into 3 classes, referring to Coronavirus Infectious Disease-19 Response Guideline [12].

TABLE III. KOREAN COVID-19 GUIDELINES

Severity Level	**Korean COVID-19 Response Guidelines**		
	Definition	***Current***	***Previous***
1	no limit of activity	Mild	Mild
2	limit of activity but No O2		
3	O2 with nasal prong	Severe	
4	O2 with facial mask		Severe
5	non-invasive ventilation/high flow O2	Critical	
6	Invasive ventilation		Critical
7	Multi-organ failure/ECMO/CRRT		
8	Death	Death	Death

- Form 4 adopts the ways of Form 2 and Form 3 data transformation simultaneously to check out whether both ways of data transformation can affect the performance of COVID-19 severity level.

Form Configuration is an experimental to obtain higher prediction performance through data type transformation of input features and reduction of the severity level class that is output.

D. *Feature Selection*

Feature selection is a procedure to select proper predictors which affect the performance of prediction models and prevent the curse of dimensionality. We chose three types of the supervised methods according to our purpose for COVID-19 severity prediction.

1) Intrinsic methods which are also called embedded methods select features for prediction as a result of learning based models like LGBM (Light Gradient Boosting Machine) and RF (Random Forest) by training each feature directly.

2) Wrapper methods consists of forward selection, backward elimination, stepwise selection methods. Backward elimination method starts with all features of the dataset and eliminates a less important feature from the feature set. One of backward elimination methods RFECV (Recursive Feature Elimination and Cross-validated Selection) is selected for the experiment.

3) Filter methods select features through the ranks of correlation coefficients calculated using statistical analysis. Because most data of the dataset are binary type, we choose the chi-squared test as a representative method for experiment.

These feature selection methods are to find out which feature is the most important feature for predicting severity. If we can get less features with feature selection with better performance comparing to all features used, it can shorten the time it takes for a patient to receive treatment, reducing time for less significant feature examination.

E. *Machine Learning Techniques*

After Feature Selection, data were split into 70% of training set to train a supervised learning model using machine learning techniques and 30% of test set for 10-fold stratified cross validation. We choose LGBM, RF, OLR (Ordinal Logistic Regression), KNN (K Nearest Neighborhood), SVM (Support Vector Machine) and MLP (Multi-Layer Perceptron) models at first as test models. LGBM and OLR which shows better performance according to two strategies and feature selection comparing to the rest of models are selected to conduct the experiment.

V. EXPERIMENTAL RESULT

As shown in pipeline of triage, a total of 200 configurations were tested to find the optimal features according to time, type, and composition. The performance was evaluated based on the classification accuracy. We find that reduction of severity level into 3 Mild, Severe, Critical classes of Data-Driven Strategy II get better performance in the overall experimental results. Without simplification, the machine learning model cannot detect most confirmed cases which need treatments, but by reducing the level of severity, it is possible to detect cases. It is because of skewness of the dataset. About 80% of cases are belonged to the case with no limit of activity, thus other cases are very hard to predict. By grouping similar classes, increasement of each class leads to performance. On the other hands, transformation of numerical type into binary data type has no significant effect on early triage system.

TABLE IV. BEST PERFORMANCE OF OLR WITH STAGE3 AND FORM4 STRATEGIES AND NO FEATURE SELECTION

S3, F4	**No Feature Selection, Model: OLR**			
	Precision	***Recall***	***F1-Score***	***Support***
1	0.90	0.98	0.94	1135
2	0.35	0.10	0.15	121
3	0.53	0.45	0.49	64
			Accuracy	0.88

The best performance obtained by performing the entire experiment is a configuration of Stage 3, Form4 and using all features without feature selection. It achieved 0.88 accuracy. In this configuration, the 4163 train data is used and test with 1320. Interestingly, this composition includes only basic information and general symptoms, unlike many studies that show that blood components are closely related to COVID-19.

TABLE V. PERFORMANCE OF LGBM WITH STAGE3, FORM4 STRATEGIES AND CHI-SQUARED TEST FEATURE SELECTION

S3, F4	**Feature Selection: Chi-squared Test, Model: LGBM**			
	Precision	***Recall***	***F1-Score***	***Support***
1	0.90	0.98	0.94	1135
2	0.37	0.08	0.14	121
3	0.56	0.44	0.49	64
			Accuracy	0.87

Comparing to using all the features, it was meaningful to do feature selection. As shown in TABLE V., it has almost the same performance with the best performance achieving 0.87 accuracy. Through Chi-squared test feature selection, we select 13 among 19 features like age, sex, BMI, blood pressure, heart rate, fever, sputum, sore throat, rhinorrhea and short of breath. Features such as muscle ache, altered consciousness, vomiting, diarrhea, fatigue and headache were deprecated. Each feature selection method showed performance enhancement according to Data-Driven Strategy, but they commonly select features like age, sex, BMI, blood pressure, heart rate, cough, sputum, sore throat, muscle ache, rhinorrhea, hypertension, headache, diabetes, chronic cardiac disease, dementia and results of blood test.

TABLE VI. PERFORMANCE OF LGBM WITH STAGE5, FORM4 STRATEGIES AND RF FEATURE SELECTION

S5, F3	**Feature Selection: RF, Model: LGBM**			
	Precision	***Recall***	***F1-Score***	***Support***
1	0.88	0.99	0.93	735
2	0.57	0.11	0.19	107
3	0.59	0.56	0.57	61
			Accuracy	0.85

As mentioned above, the results of blood test are related to COVID-19 closely. TABLE VI. shows slightly low performance with accuracy 0.85 comparing to above two results but by influence of results of blood test, it shows improved performance in predicting severe and critical cases of COVID-19. At last, it was confirmed that valid test results were shown at least stage 3, and further, stage 5 was able to increase performance of early triage in predicting severe and critical cases. Through the results of overall experiment, we can prove the validity of Data-Driven Strategy I.

VI. CONCLUSION

For early diagnosis of COVID-19 among patients with febrile respiratory illness in COVID-19 pandemic status with limited medical resources, we proposed a COVID-19 Early Triage using two data-driven strategies and machine learning techniques. The results demonstrated the effectiveness of our Data-Driven Strategies and exploiting machine learning techniques. Stage configuration in Data-Driven Strategy I may help predict the severity of COVID-19 in advance even in the absence of the blood test results and transfer the patients to the appropriate medical institutions. The reliability increases when the blood test results are added. Form configuration in Data-Driven Strategy II was also useful for predicting the severity of COVID-19 disease. The categorization of data to overcome the limitation of the binary data types reduced the classes and implemented the prediction of severity. This study has limitation in that its performance to predict severe cases and critical cases is lower than expected because of extremely skewed data. Further studies are needed, adding more severe cases and adding featured data of COVID-19.

REFERENCES

[1] "Information of COVID-19", Coronavirus Disease-19, Republic of Korea, last modified 25 February 2021, accessed 9 August 2021, http://ncov.mohw.go.kr/baroView.do?brdId=4&brdGubun=41

[2] "WHO Coronavirus (COVID-19) Dashboard", Word Health Organization, last modified 6 August 2021, accessed 9 August 2021, https://covid19.who.int/?adgroupsurvey={adgroupsurvey}&gclid=CjwKCAjwpMOIBhBAEiwAy5M6YFTNNPAGuwUWzynO24t-Eds8qI5e9vUr-1T6rJayBKCfcxwK68ptnxoCiqkQAvD_BwE

[3] "Cases in Korea", Coronavirus disease-19, Republic of Korea, last modified 9 August 2021, accessed 9 August

[4] "COVID-19 Response", Coronavirus Disease-19, Republic of Korea, last modified 2 March 2021, accessed 9 August 2021, http://ncov.mohw.go.kr/en/baroView.do?brdId=11&brdGubun=111&dataGubun=&ncvContSeq=&contSeq=&board_id=&gubun=

[5] Liang, Wenhua, et al. "Early triage of critically ill COVID-19 patients using deep learning." *Nature communications* 11.1 (2020): 1-7.

[6] Çelik Ertuğrul, Duygu, and Demet Çelik Ulusoy. "A knowledge-based self-pre-diagnosis system to predict Covid-19 in smartphone users using personal data and observed symptoms." *Expert Systems* (2021).

[7] An, Chansik, et al. "Early triage of patients diagnosed with COVID-19 based on predicted prognosis: A Korean national cohort study." (2020).

[8] Patel, Dhruv, et al. "Machine learning based predictors for COVID-19 disease severity." *Scientific Reports* 11.1 (2021): 1-7.

[9] Kar, Sujoy, et al. "Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID)." *Scientific reports* 11.1 (2021): 1-11.

[10] Kim, Hyung-Jun, et al. "An Easy-to-Use Machine Learning Model to Predict the Prognosis of Patients With COVID-19: Retrospective Cohort Study." *Journal of medical Internet research* 22.11 (2020): e24225.

[11] Kim, Yu-Rin, Seoul-Hee Nam, and Seon-Rye Kim. "Impact Factors and Validity of Blood Variables on Death in COVID-19 patient: Using Data of Korea Disease Control and Prevention Agency." Journal of the Korea Society of Computer and Information 25.11 (2020): 179-185.

[12] "Coronavirus Infectious Disease-19 Response Guidelines (for local governments) 10th Edition", Central Disaster Management Headquarters · Central Disease Control Headquarters, 5 May 2021