The Selection of Indicators from Initial Blood Routine Test Results to Improve the Accuracy of Early Prediction of COVID-19 Severity

Running Title: COVID-19 Severity Prediction using MCDM Algorithm

Jiaqing Luo[1#], Lingyun Zhou[3#*], Yunyu Feng[4], Bo Li[5], Shujin Guo[2*]

1. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

2. The Geriatric Respiratory department, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu 611731, China.

3. Center of Infectious Diseases, West China Hospital of Sichuan University. Chengdu 610041, China.

4. State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University and Collaborative Innovation Center. Chengdu 610041, China

5. Department of Otorhinolaryngology, Head & Neck Surgery, West China Hospital, Sichuan University, Chengdu, 610041, China

# Jiaqing Luo and Lingyun Zhou contributed equally to this work and are joint first authors.

Address for correspondence:

* Lingyun Zhou, Center of Infectious Diseases, West China Hospital of Sichuan University. Chengdu 610041, China.

E-mail: 4423925@qq.com

* Shujin Guo, The Geriatric Respiratory department, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu 611731, China.

E-mail: shujinguo@126.com

Abstract

Early prediction of disease severity is important for effective treatment of COVID-19. We determined that age is a key indicator for severity predicting of COVID-19, with an accuracy of 0.77 and an AUC of 0.92. In order to improve the accuracy of prediction, we proposed a Multi Criteria Decision Making (MCDM) algorithm, which combines the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) and Naïve Bayes (NB) classifier, to further select effective indicators from patients' initial blood test results. The MCDM algorithm selected 3 dominant feature subsets {Age, WBC, LYMC, NEUT}, {Age, WBC, LMYC} and {Age, NEUT, LYMC}. Using these feature subsets, the optimized prediction model could achieve an accuracy of 0.82 and an AUC of 0.93. This result indicated that using age and the indicators selected by the MCDM algorithm from blood routine test results can effectively predict the severity of COVID-19 at an early stage.

Keywords: Coronavirus disease 2019 (COVID-19), Severity, Blood routine test, Multiple Criteria Decision Making (MCDM).

Introduction

Currently, more than 40 million people worldwide are infected with the SARS-Cov-2 virus, and more than 10 million people are suffering from Coronavirus disease 2019 (COVID-19) and are receiving treatment. This poses a huge threat to the health and lives of people all over the world, and brings unprecedented pressure to the medical system. Many infected patients cannot receive timely and effective treatment, and it will also reduce the treatment efficiency of other emergency patients.

Patients with suspicious symptoms and epidemiological history first visit the fever clinic of the community hospital. They usually undergo 3 initial tests: SARS-Cov-2 RNA confirms SARS-Cov-2 infection (*1*), blood routine test and chest CT scan to initially assess the severity of COVID-19 (*2-4*). The timely and effective triage of COVID-19 patients based on the results of the 3 initial tests is of great significance for maintaining emergency capacity and optimizing treatment plans.

Although most COVID-19 patients are Mild-Moderate cases and can recover on their own, about 14% of patients are Severe cases, and 5% of patients are Critically Severe cases (*5*). Severe-Critically Severe cases usually develop Acute Respiratory Distress Syndrome (ARDS) or Multiple Organ Dysfunction Syndrome (MODS) within 2 weeks of infection (*6*), which consumes most of medical resources and leads to a high case fatality rate (up to 49%) (*5*). Early prediction of the severity of COVID-19 can not only help quickly triage patients (i.e., quarantine, hospital admission or ICU assignment, etc.), but also optimize the use of medical resources and timely medical intervention. Previous studies have used multiple indicators to predict the severity of COVID-19 (i.e., older age, pulmonary micro-thrombosis, increased inflammatory factors (C-reactive protein (CRP), IL-6), hyper-lactic acidemia, D-dimer progressive heightened, decreased lymphocyte count (especially CD8+ T cell count) and short-term progression of lung lesions, etc.). However, the collection of these indicators requires multiple tests and takes a lot of time.

Of all the initial tests, blood routine test is the worldwide common test, and the results are usually available within 2 hours. In this paper, we tried to select features from blood test results to predict the severity of COVID-19 quickly and accurately. Specifically, we first defined feature selection as a Multiple Criteria Decision Making (MCDM) problem that considers the correlation between input features, and the correlation between input and output features, and then combined the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) and Naïve Bayes (NB) classifier to achieve the highest prediction accuracy with the fewest features.

Our early prediction of the severity of COVID-19 based on the clinic characteristics of patients can improve the efficiency and accuracy of emergency triage of patients, thereby effectively supplementing and improving the overall management of COVID-19.

Methods

Patient enrollment and study design

This retrospective study was approved by the ethics committee of Sichuan Provincial People's Hospital. We collected 196 COVID-19 patients diagnosed according to WHO guidance (*7*) in Wuhan Red Cross Hospital from February 1, 2020 to March 15, 2020. Written or oral informed consent was obtained from patients.

Definitions

COVID-19 was confirmed by detecting SARS-CoV-2 RNA test. According to the 5th edition of the China Guidelines for the Diagnosis and Treatment Plan of COVID-19 Infection by the National Health Commission (Trial Version 5) (*8*), the cases were classified into Mild-Moderate and Severe- Critically Severe.

Data collection

The following information was extracted from each patient: Gender, Age and patients' initial blood routine test results including White Blood Cell Count (WBC), Lymphocyte Count (LYMC), Lymphocyte Ratio (LYMPH), Neutrophil Count (NEUT), Neutrophil Ratio (NEU) and Neutrophil to Lymphocyte Ratio (NLR). The dataset contained 8 input features {Gender, Age, WBC, LYMC, LYMPH, NEUT, NEU, NLR}, and 1 output feature (Severity).

Statistical Analysis

Quantitative variables were expressed as the mean ± standard deviation or the median with interquartile ranges, while categorical variables were expressed as absolute and relative frequencies. The t test or Wilcoxon-test was performed to calculate differences between quantitative data; and $\chi 2$ test was performed to calculate differences between qualitative data. According to the data characteristics, the correlation between clinic characteristics and COVID-

19 severity was calculated according to Kendall correlation coefficient (Gender-severity) or Spearman correlation coefficient. Logistic regression analysis was performed for independent variables with collinearity. Wald test was used to determine the joint significance of variables. The standard deviation was used to measure dispersion degrees. Statistical procedures were performed with R statistical software. P values of ≤0.05 were considered significant.

The MCDM algorithm design and implementation

The proposed algorithm is basically designed for predicting COVID-19 severity, either Mild-Moderate or Severe-Critically Severe case. It leads to reducing computation time, improving prediction performance, and a better understanding of the data in machine learning.

It consists of 4 major stages: preprocessing, feature ranking, feature selection and performance evaluation. Preprocessing is the process to refine the collected raw data to de-noise it. Feature ranking is the process of ordering the features by the value of some scoring function, which usually measures feature-relevance. Feature selection aims to choosing a small subset of the relevant features from the original features by removing irrelevant, redundant, or noisy features (*9*). Performance evaluation is to measure the performance of the binary classification by statistical measures, i.e., Accuracy (ACC), True Positive Rate (TPR), False Positive Rate (FPR) and F1 score.

- Preprocessing -We use stratified random sampling to divide the dataset into 2 subsets: training set (80%) and test set (20%). In these 4 stages, we only used the test set for performance evaluation. Suppose there are m input features and n output features. Let $X=\{x|1\leq x\leq m\}$ be the input feature set and $Y=\{y|m+1\leq y\leq m+n\}$ be the output feature set. Elements x and y are indexes of features. The feature set is $F=X\cup Y=\{i|1\leq i\leq m+n\}$. We calculated and visualized a $(m+n)\times(m+n)$ correlation matrix R and a $(m+n)\times(m+n)$ p-value matrix P to show the correlations between all different feature pairs. To simplify the analysis, we then preprocess R in 2 steps. STEP1: We ignored the sign of R[i,j]. Let R[i,j] = |R[i,j]| so that the range of R[i,j] changes from [-1,1] to [0,1], where $i, j \in F$. STEP2: We filtered R through P. For $x \in X$ and $y \in Y$, if P[x,y] = P[y,x] > 0.05, R[x,y] and R[y,x] are not significant. We set R[x,y] = R[y,x] = 0 and R[x,i] = R[i,x] = 1 for $i \in X$.

- Feature Ranking-We defined a labeled feature set L and initialized with $L = \varnothing$. We iterated the procedure of ranking input features $x \in X$ and moved the first in each ranking from X to L. The ranking criteria includes 2 evaluations: EVAL1: The correlation between input feature $x \in X$ and output feature $y \in Y$, R[x,y] or R[y,x]. EVAL2: The correlation between input feature $x \in X$ and labeled feature $v \in L$, R[x,v] or R[v,x]. This explicitly evaluates multiple conflicting criteria in decision making. We proposed an algorithm to solve this Multiple Criteria Decision Making (MCDM) problem by using the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) (*10*), which is a compensatory aggregation method. The algorithm, called MCDM, creates an evaluation matrix E consisting of p criteria and q alternatives, to rank input features. According to Pareto's principle (*11*), the algorithm divide x into the following 2 types:

  TYPE1: If $|X| > min\{m-1, \lceil 0.8 \times m \rceil\}$, x to be labeled are core features (the top 20%), which should have the lowest R[v,x] from EVAL2, and the highest R[y,x] from EVAL1. The algorithm sorts the elements of sets $L \cup Y$ and X in ascending order to get sequences $(r_i)_{i=1}^{|L|+n}$and $(c_j)_{j=1}^{\text{¿}X \vee \text{¿¿}}$, respectively. Let $p = \text{¿}L \vee + n$ and $q = \text{¿}X \vee \text{¿}$, the algorithm extracts a $p \times q$ submatrix E from R such that $E[i, j] = R[r_i, c_j]$. The worst condition of $E[i, j]$ is $w_i = \text{¿}$, and the best condition of $E[i, j]$ is $b_i = \text{¿}$.

TYPE2: If $|X| \leq min\,\{\vee m-1, \lceil 0.8\times m \rceil\}$, x to be labeled are auxiliary features (the rest 80%), which only need to have the lowest R[v,x] from EVAL2. The algorithm sorts the elements of sets L and X in ascending order to get sequences $(r_i)_{i=1}^{|L|}$and $(c_j)_{j=1}^{¿X\vee¿¿}$, respectively. Let $p=¿L\vee¿$ and $q=¿X\vee¿$, E is a $p\times q$ matrix with $E[i,j]=R[r_i,c_j]$. The algorithm calculates the L2-distance between the target alternative j and the worst condition:

$$d_{wj}=\sqrt{\sum_{i=1}^{p}(E[i,j]-w_i)^2} \quad \text{Eq.1}$$

It then calculates the distance between j's condition and the best condition:

$$d_{bj}=\sqrt{\sum_{i=1}^{p}(E[i,j]-b_i)^2} \quad \text{Eq.2}$$

After that, it calculates the similarity to the worst condition:

$$s_j=\frac{d_{wj}}{d_{wj}+d_{bj}}, 0\leq s_j\leq 1 \quad \text{Eq.3}$$

$s_j=1$ if and only if alternative j has the best condition, and $s_j=0$ if and only if alternative j has the worst condition. Let $j^{¿}=arg\,\underset{j}{max}\{s_j\}$, then $X=X¿c_{j^{¿}}\}$ and $L=L\cup\backslash\{c_{j^{¿}}\}$.

The pseudocode of the MCDM algorithm is as follows:

```
Algorithm MCDM is
Input: correlation matrix R, number of input features m, number of input features n, input feature set X, output feature set Y
Output: labeled feature set L
initialize $L=\emptyset$
while $X\neq\emptyset$ do
```

```
    if $|X| > min\{m-1, \lceil 0.8 \times m \rceil\}$
        $(r_i)_{i=1}^{|L|+n} \leftarrow$ sort $L \cup Y$ and X in ascending order
    else
        $(r_i)_{i=1}^{|L|} \leftarrow$ sort L in ascending order
    $(c_j)_{j=1}^{¿X \vee ¿¿} \leftarrow$ sort X in ascending order
    extract E from R such that $E[i,j] \leftarrow R[r_i, c_j]$
    for j = 1 to q do // q is the number of columns of E
      $d_{wj} \leftarrow$ Eq.2
      $d_{bj} \leftarrow$ Eq.3
      $s_j \leftarrow$ Eq.4
    $j^{¿} \leftarrow arg\,\max_{j}\{s_j\}$
    $X \leftarrow X ¿ c_{j^{¿}}\}$
    $L \leftarrow L \cup \setminus\{c_{j^{¿}}\}$
    print L
return L
```

- Feature Selection-The goal of feature subset selection is to find the optimal input feature subset. We gradually increased the number of labeled features, and trained the model with Naïve Bayes classifier (*12*) in turn. To find the optimal subset, we sequentially tested the accuracy of trained models on the training set.
- Performance Evaluation-In order to test the stability of the algorithm and observe the influence of the dataset uncertainty on feature selection, we divided the data set 100 times (80% training set and 20% test set) and repeatedly run the algorithm. We used the test set to

analyze the performance of feature selection from Accuracy (ACC), True Positive Rate (TPR), False Positive Rate (FPR) and F1 score.

Evaluation of the predictive value of selected features

According to stratified random sampling, we divided the data set into 2 subsets: 80% of the “training set” and 20% of the “testing set”. We used Receiver Operating Characteristic (ROC) curve analysis to calculate the Area Under the Curve (AUC) and use “ROC” package in R to evaluate the prediction accuracy of our model.

Results

Baseline characteristics

We analyzed the data of 196 COVID-19 patients, of which 67 and 129 were male and female patients. After clearing the data set, there is no abnormal data (S-Figure 1). Table 1 lists the detailed baseline characteristics. The mean age of patients was 57.74±15.87 years old. The COVID-19 patients' initial blood routine test results showed that the WBC was 6.75±3.49◊109/L; LYMC was 1.12±0.58◊109/L; LYMPH was 19.91±11.52%; NEUT was 5.13±3.46◊109/L; NEU was 71.34±15.24%; the NLR was 7.45±13.08.

Difference in Age and initial blood test results between Mild-Moderate and Severe-Critically Severe groups

According to the 5th edition of the China Guidelines for the Diagnosis and Treatment Plan of COVID-19 Infection by the National Health Commission, we divided patients into 2 groups: 67 cases in the Mild-Moderate group, and 129 cases in the Severe-Critically Severe group (Table 1). Comparing Mild-Moderate and Severe-Critically Sever groups, the basal features showed no differences in Gender ($p=0.26$) (Figure1A). The Severe-Critically Severe group was significantly older than the Mild-Moderate group ($p <0.001$) (Figure 1B). The initial blood routine test seems to be important for predicting the severity of COVID-19: The Severe-Critically Severe group had a higher WBC level ($p=0.02$) (Figure1C). The Severe-Critically Severe group had extremely low LYMC ($p < 0.001$) and LYMPH ($p < 0.001$) (Figure1D, E). In contrast, NEUT ($p < 0.001$) and NEU ($p < 0.001$) in the Severe-Critically Severe group were extremely high (Figure1F, G). As a result, the Severe-Critically Severe group had a higher NLR ($p < 0.001$) (Figure1H).

Predictive value of age and initial blood test results for COVID-19 severity

By calculating the correlation between clinic characteristics and severity of COVID-19, we found that Age (r=0.73, p=0.01), WBC (r=0.24, $p < 0.01$), NEUT (r=0.34, $p < 0.01$), NLR (r=0.31, $p < 0.01$) were significantly positively correlated with the severity of COVID-19, while LYMC (r=-0.55, p=0.01) was significantly negatively correlated with the severity of COVID-19 (Figure 2A, B). These results indicated that Age and initial blood routine test results-WBC, LYMC, NEUT, NLR, might be important for predicting the severity of COVID-19.

Wald test showed that only Age was the key indicator in predicting the severity of COVID-19 (Table2). Using stratified random sampling, we generated the Receiver Operating Characteristic (ROC) curve to evaluate the predictive values: 80% for the "training set" and 20% for the "testing set". Using {Age} for prediction, we can obtain an accuracy of 0.77, and an Area Under the Curve (AUC) of 0.92 (Figure2C). Through dispersion analysis, we found that WBC, LYMC and LYMPH may be able to optimize prediction performance (Table3, Table4). The ROC curve showed that {Age, WBC, LYMC} had an accuracy of 0.82 and an AUC of 0.93 (Figure2D).

Details of the MCDM algorithm to predict the severity of COVID-19

The MCDM algorithm and Logistic regression analysis have obtained consistent results: Age was a key indicator in predicting the severity of COVID-19. In addition, the MCDM algorithm verified that the {Age, WBC, LYMC} subset is one of the index sets with the highest prediction accuracy.

Preprocessing (Figure3A) - In the COVID-19 data set, $m=8$ and $n=1$. The 9 × 9 correlation matrix R, The 9 × 9 p-value matrix P and the range of R[i,j] for $i, j \in F$ becomes [0,1]. Since P[1,9]=P[9,1]=0.1442>0.05 , R[1,9] and R[9,1] are not significant, R[1,9]=R[9,1]=0, R[1,1:8]=ones(1,8) and R[1:8,1]=ones(8,1).

Feature Ranking (Figure3B) - When $|X|=8>min\{8-1,\lceil 0.8\times 8\rceil\}=7$, $L\cup Y=\varnothing\cup\{9\}=\{9\}$ and $X=\{1,\ldots,8\}$. Then, we have, $(r_i)_{i=1}^{1}=(9)$ and $(c_j)_{j=1}^{8}=(1,\ldots,8)$. Since $p=|L|+n=1$ and $q=|X|=8$, E is a 1×8 submatrix of R. When $|X|=5<7$, $L=\{2,3,4\}$ and $X=\{1,5,6,7,8\}$. Then, we have $(r_i)_{i=1}^{3}=(2,3,4)$ and $(c_j)_{j=1}^{5}=(1,5,6,7,8)$. Since $p=|L|=3$ and $q=|X|=5$, E is a 3×5 submatrix of R. When $|X|=8>7$, $w_i=1$ and $b_j=0$. By Eq. 1 and Eq.2, we calculated $d_{w2}=0.5913$ and $d_{b2}=0.4087$. By Eq. 3, we have $s_2=0.5913$. When $|X|=5<7$, $w_i=1$ and $b_i=0$. By Eq.1 and Eq.2, we calculated $d_{w6}=1.1871$ and $d_{b6}=0.9912$. By Eq. 3, we got $s_6=0.5450$.

Feature Selection (Figure3C) - When 4 features {2,5,8,4} are selected, the accuracy of EVAL1 reached a peak of 0.803. Interestingly, with less features {2,3,4}, the accuracy of EVAL1+EVAL2 can reach a higher 0.815.

Performance Evaluation (Figure3D) - {2,3,4} has the lowest number of features, but the highest score among multiple performance metrics. We can see that the accuracy of {2,5,8,4,7,6,3}, {2,5,8,4} and {2,3,4} are 0.74, 0.82 and 0.87, respectively. We can also see that the F1 score of {2,5,8,4,7,6,3}, {2,5,8,4} and {2,3,4} are 0.67, 0.72 and 0.78, respectively.

Influence of dataset uncertainty on the feature selection of the MCDM algorithm

To test the stability of the algorithm and observe the influence of the dataset uncertainty on feature selection, we divided the data set 100 times (80% training set and 20% test set) and repeatedly run the algorithm. The average number of features selected by 3 different criteria, EVAL1, EVAL1 (subset) and EVAL1+EVAL2 (subset) are 6.58 (95% CI: 6.48 - 6.68), 3.26 (95% CI: 3.01 - 3.51) and 3.52 (95% CI: 3.40 - 3.64), respectively (Figure4A). The criteria, EVAL1+EVAL2 (subset), adopted by the MCDM algorithm improved most performance metrics. The metrics (ACC, TPR, FPR and F1 score) of EVAL1+EVAL2 (subset) are 0.81 (95%

CI: 0.80 - 0.82), 0.69 (95% CI: 0.67 - 0.71), 0.09 (95% CI: 0.08 - 0.11) and 0.75 (95% CI: 0.73 - 0.77) respectively, while those of EVAL1 are 0.75 (95% CI: 0.74 - 0.77), 0.60 (95% CI: 0.58 - 0.62), 0.07 (95% CI: 0.06 - 0.09) and 0.71(95% CI: 0.70 - 0.73) respectively (Figure4B). Although dataset uncertainties have an influence on feature selection, there were still 3 subsets: {Age, WBC, LYMC, NEUT} with a selection rate of 44%, {Age, NEUT, LYMC} with a selection rate of 38%, and {Age, WBC, LYMC} with a selection rate of 9%, which dominated EVAL1+EVAL2 (subset) feature selection. These 3 subsets can achieve high accuracy with a small number of features (Figure4C).

Predictive value of the features selected by the MCDM algorithm

Using stratified random sampling, we generated ROC curves to evaluate the predictive values of the subsets selected by the MCDM algorithm: 80% for the “training set” and 20% for the “testing set”. Our analysis results showed that {Age, WBC, LYMC, NEUT} (Figure5A), {Age, NEUT, LYMC} (Figure5B) and {Age, WBC, LYMC} (Figure5C) all achieved 0.82 accuracy and 0.93 AUC. The MCDM algorithm can steadily and accurately select Age and other features from initial blood routine test results to predict the severity of COVID-19.

Discussion

In this paper, we determined that age was the most critical indicator for predicting the severity of COVID-19. To improve the prediction accuracy, we proposed an MCDM algorithm, which combines the TOPSIS and NB classifier, to further select the indicators of patients' initial blood routine test. By ranking features, the MCDM algorithm selected 3 subsets including {Age, WBC, LYMC, NEUT}, {AGE WBC, LMYC} and {Age, NEUT, LYMC}, all of which can achieve 0.82 accuracy and 0.93 AUC.

Previous studies have shown that elderly COVID-19 patients with multiple concomitant diseases tend to develop Multiple Organ Failure (MOFE), which may lead to high morality in elderly patients infected by SARS-CoV-2. According to the latest meta-analysis of the elderly in the European community, the prevalence of frailty is around 15% for the elderly 65 years and older (*13*), and the case fatality rate of patients over 85 years old is 1,000 times that of patients aged 5-17 years (*14*). Our research indicated that age was the most important indicator for predicting the severity of COVID-19, with an accuracy 0.77 and an AUC of 0.92. However, some elderly patients had a good prognosis, so prognostic evaluation and medical decision-making based on age alone might not be accurate enough.

We found that WBC, LYMC and NEUT in initial blood routine test results other than age are also important for predicting the severity of COVID-19. Guo et al. (*15*) pointed out that the MuLBSTA score revealed that multi-lobar infiltrates, lymphocytes ≤0.8×109/L, bacterial infection, smoking status, hypertension, and age ≥60 years could help prognosticate outcomes in COVID-19 patients. The elevated WBC/NEUT is a basic sign of bacterial infection. Bacterial co-infection in COVID-19 patients may develop severe form of disease, complicating the clinical situation (*16-18*). The control and elimination of viruses depends on humoral immunity. Viral infections usually lead to abnormal changes in the levels of lymphocyte subsets which further impaired immune system functionality. The decrease of LYMC is the simplest and most intuitive indicator to predict the humoral immune response, indicating that the patient's T cell function is defective (*19-21*). The count of lymphocyte subsets (CD4+ and CD8+ T cell), especially CD8+ T cell, is directly proportional to the severity of COVID-19 (*22,23*).

Although logistic regression can determine the key indicator {Age}, and discrete analysis can find a better subset {Age, WBC LYMC}, it is difficult to determine the best subset due to the small sample size or multicollinearity. Previous studies used the MCDM algorithm to evaluate diagnostic tests (*24*) and help doctors hasten COVID-19 treatment (*25*). As far as we know, this is the first time the MCDM algorithm has been used to predict the severity of COVID-19. It first uses TOPSIS for feature ranking, and then combines the NB classifier for feature selection. Even if the sample size is small, the MCDM algorithm can select 3 effective subsets {Age, WBC, LYMC}, {Age, WBC, LYMC, NEUT} and {Age, NEUT, LYMC}. The selection process is visual and interpretable helping doctors find the features of the progress of emerging infectious diseases early, to make faster and better prevention and treatment plans. We used the ROC curve to evaluate the predictive value of the features selected by the MCDM algorithm. The results showed that the MCDM algorithm can not only find all effective subsets, but also predict stably and accurately.

Age (*26-29*), underlying diseases (*30*), systemic immune status (*31*), and blood test results can be used as key features to predict the severity of COVID-19. Although these features can improve the accuracy of prediction (84%~93%), the tests are time-consuming, expensive, and labor-intensive. Our algorithm can select features from blood test results to achieve a prediction accuracy of 82%. During the COVID-19 pandemic, it is more in line with clinical needs and is easy to promote and use in areas with different medical levels.

The feature selection may have some limitations, because there were only 196 cases and all were from China. In future, we would like to collect more data and conduct multi-center evaluations.

Conclusion

We defined feature selection as a MCDM problem so that the algorithm can provide a reference for clinical practice. The concise features {Age, WBC/NEUT, LYMC} and high accuracy are very conducive to rapid triage of COVID-19 patients. Using the most common blood routine test, medical institutions could better determine the quarantine, hospital admission, ICU assignment of COVID-19 patients. The MCDM algorithm can be used for small sample data sets, and the prediction is accurate and stable, which will help establish a rapid response mechanism in the early stage of emerging infectious disease outbreaks.

Acknowledgments

This work was funded by Natural Science Foundation of China 81802468, Sichuan Science and Technology Program 2019YFS0207 and China Postdoctoral Science Foundation 2020M670062ZX to Dr. Lingyun Zhou; Grant 81700044 from Natural Science Foundation of China to Dr. Shujin Guo. The funds had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interests: The authors declare that they have no conflict of interests.

Author Contributions: Jiaqing Luo, Lingyun Zhou, and Shujin Guo designed the study. Shujin Guo collected data. Jiaqing Luo, Bo Li and Yunyu Feng developed the algorithm. Jiaqing Luo, Yunyu Feng and Lingyun Zhou edited the manuscript. Lingyun Zhou and Shujin Guo reviewed the manuscript.

Ethics: The study was approved by the ethics committee of Sichuan Provincial People's Hospital. Participant consent was not required.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. World Health Organization. Coronavirus disease 2019 (COVID-19) situation report–51. Geneva, Switzerland: World Health Organization; 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10
2. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. JAMA. 2020;323:1061-9. doi:10.1001/jama.2020.1585
3. Yang X, Yu Y, Xu J, Shu H, Liu H, Wu Y, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med. 2020;8:475-81. doi:10.1016/S2213-2600(20)30079-5
4. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, 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; e202033. doi:10.1001/jamainternmed.2020.2033
5. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. JAMA. 2020; 323:1239-42.
6. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China [Chinese]. Zhonghua Liu Xing Bing Xue Za Zhi. 2020; 41:145–51. doi:10.3760/cma.j.issn.0254-6450.2020.02.003
7. World Health Organization (2020). Clinical Management of Severe Acute Respiratory Infection When Novel Coronavirus (nCoV) Infection is Suspected: Interim Guidance.

Available online at :https://www.who.int/docs/default-source/coronaviruse/clinical-management-of-novel-cov. pdf (accessed January20, 2020)

8. Lin L, Li TS. Interpretation of “guidelines for the diagnosis and treatment of novel coronavirus (2019-ncov) infection by the national health commission (trial version 5)”. Zhonghua Yi Xue Za Zhi. 2020;100: E001. doi:10.3760/cma.j.issn.0376-2491.2020.0001
9. Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng. 2014; 40:16-28.
10. Yoon K. "A reconciliation among discrete compromise situations". J Oper Res Soc. 1987; 38:277-86. doi:10.1057/jors.1987.44.
11. Bunkley N. "Joseph Juran, 103, Pioneer in Quality Control, Dies". The New York Times. 2008.
12. McCallum A. "Graphical Models, Lecture2: Bayesian Network Represention". Retrieved 22 October, 2019.
13. Lippi G, Plebani M. Procalcitonin in patients with severe coronavirus disease 2019 (COVID-19): a meta-analysis. CLIN CHIM ACTA. 2020;505:190.
14. Mo P, Xing Y, Xiao Y, Deng L, Zhao Q, Wang H, et al. Clinical characteristics of refractory COVID-19 pneumonia in Wuhan, China. Clinical Infectious Diseases.2020.
15. Guo L, Wei D, WU Y, ZHOU M, ZHANG X, Li Q, et al. Clinical features predicting mortality risk in patients with viral pneumonia: the MuLBSTA score. FRONT MICROBIOL. 2019;10:2752.
16. Ma Y, Hou L, Yang X, Huang Z, Yang X, Zhao N, et al. The association between frailty and severe disease among COVID-19 patients aged over 60 years in China: a prospective cohort study. BMC MED. 2020; 18:1-8.

17. Stawicki SP, Jeanmonod R, Miller AC, Paladino L, Gaieski DF, Yaffee AQ, et al. The 2019–2020 novel coronavirus (severe acute respiratory syndrome coronavirus 2) pandemic: A joint american college of academic international medicine-world academic council of emergency medicine multidisciplinary COVID-19 working group consensus paper. Journal of global infectious diseases. 2020;12:47.
18. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): The epidemic and the challenges. Int J Antimicrobial Agents. 2020;105924
19. Wang F, Nie J, Wang H, Zhao Q, Xiong Y, Deng L, .et al. Characteristics of peripheral lymphocyte subset alteration in COVID-19 pneumonia. J INFECT DIS. 2020;221:1762-9.
20. Diao B, Wang C, Tan Y, Chen X, Liu Y, Ning L, et al. Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). FRONT IMMUNO, 2020;11:827.
21. Mathew D, Giles JR, Baxter AE, Oldridge DA, Greenplate AR, Wu JE, .et al. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implication. Science.2020;369(6508).
22. Pallotto C, Suardi LR, Esperti S, Tarquini R, Grifoni E, Meini S, et al. Increased CD4/CD8 ratio as a risk factor for critical illness in coronavirus disease 2019 (COVID-19): a retrospective multicentre study. INFECT DIS-NOR. 2020; 52: 675-7.
23. Ganji A, Farahani I, Khansarinejad B, Ghazavi A, Mosayebi G. Increased expression of CD8 marker on T-cells in COVID-19 patients. BLOOD CELL MOL DIS. 2020;102437.
24. Sayan, M., Sarigul Yildirim, F., Sanlidag, T., Uzun, B., Uzun Ozsahin, D., & Ozsahin, I. (2020). Capacity Evaluation of Diagnostic Tests For COVID-19 Using Multicriteria

Decision-Making Techniques. Computational and Mathematical Methods in Medicine, 2020.

25. Albahri OS, Al-Obaidi JR, Zaidan AA, Albahri AS, Zaidan BB, Salih MM, et al. Helping doctors hasten COVID-19 treatment: Towards a rescue framework for the transfusion of best convalescent plasma to the most critical patients based on biological requirements via ml and novel MCDM methods. COMPUT METH PROG BIO. 2020;196:105617.

26. Guan WJ, Zhong NS. Clinical Characteristics of Covid-19 in China. Reply. N Engl J Med. 2020;382:1861-62. doi:10.1056/NEJMc2005203

27. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395:507-13. doi:10.1016/S0140-6736(20)30211-7

28. Chen R, Liang W, Jiang M, Guan W, Zhan C, Wang T, et al. Risk Factors of Fatal Outcome in Hospitalized Subjects With Coronavirus Disease 2019 From a Nationwide Analysis in China. Chest. 2020;158:97-105. doi:10.1016/j.chest.2020.04.010

29. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, 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. doi:10.1016/S0140-6736(20)30566-3

30. Wu C, Chen X, Cai Y, Zhou X, Xu S, Huang H, et al. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. JAMA internal medicine. JAMA Intern Med. 2020;180:1-11. doi:10.1001/jamainternmed.2020.0994

31. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395:497-506. doi:10.1016/S0140-6736(20)30183-5