Identifying Severity of COVID-19 In Patients Using Machine Learning Methods

Dovydas Ožiūnas
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands

d.oziunas@student.utwente.nl

ABSTRACT

Applying machine learning methods can help hospitals deal with medical resource allocation. Patients' blood characteristics can be analysed to build prediction models that could signal the severity of the disease. Given this signal, hospitals can allocate more medical resources to the patient with high severity of COVID-19 over the patient with low COVID-19 severity. It would help hospitals better react to patients who are in extreme need of more intense care and reduce the mortality rate of COVID-19 patients. This research paper aims to build a classification model that could classify patients' with coronavirus disease as severe or non-severe.

Keywords

COVID-19, Machine Learning, COVID-19 severity, Cytokines, Logistic Regression, SVM, MissForest, CRISP-DM.

1.   INTRODUCTION

When The World Health Organization declared COVID-19 as a pandemic on 11 March 2020, many hospitals worldwide were starting to become overwhelmed with the number of patients they had to take in. The number of people to be hospitalised was increasing daily in various countries. Coronavirus disease presented the hospitals with serious medical resource management problems [5]. Some patients with a severe COVID-19 illness have to be transferred to the intensive care unit (further ICU) to maintain their vital body functions like breathing. For that, an ICU with lung ventilator equipment has to be available. About 5-15% of COVID-19 patients must be hospitalised in an ICU and require ventilatory support [3]. Vervoort researched ICU capacity in 182 countries and territories, and they found out that in 182 countries, ICU bed capacity ranges between 0 to 59.5 per 100,000 population [4]. Also, their findings showed that at least 96 countries and territories had a density of less than 5 ICU beds per 100,000 population. Given these numbers, it is possible that hospitals can run out of beds in intensive care units, so some patients would have to be prioritised over others to be hospitalised in an ICU. Incorrectly prioritising certain patients can lead to a higher mortality rate as some patients who need intensive care are not selected.

To tackle this problem, many scientists started to think about possible ways to help medical personnel select patients for an ICU more accurately with the help of machine learning.

The research discussed in the related work section based their studies on the data sample of 485 patients' blood tests collected in Wuhan, China. We will use the data sample of around 500 patients provided by the hospital whose name will remain confidential for this research. We want to see which machine learning methods will give the best results given the selected features.

In this research paper, we will investigate how blood biomarkers can help us foresee the severity or the progression of COVID-19. There is some already conducted research regarding this problem [1, 2]. As a starting point, we picked out specific biomarkers that were proven to be important in predicting the severity or progression of COVID-19. In the review article written by Roshanravan et al. [1], results showed that IL-6, which acts as a pro-inflammatory cytokine and an anti-inflammatory myokine, was elevated in patients with severe COVID-19 conditions. Results also showed that the level of IL-6 has a positive correlation with the severity of COVID-19. In another study conducted by Henry et al. [7], results showed that IL-10 and IL-10/lymphocyte count is significantly increased in patients with severe disease. Also, IL-10/TNF-α, IL-6/lymphocyte and IL-10/lymphocyte count were significantly elevated in patients with severe coronavirus disease. Huang et al. [8] reported elevated levels of IL-1b in severe COVID-19 cases. In their research, Yang et al. [9] have found that lower levels of lymphocyte count drive negative prediction. Their findings also showed elevation of CRP and ferritin in patients infected with COVID-19. Medical experts advised us to look into the following biomarkers; neutrophils, BDCA3, PAI1. We add these biomarkers to our selected features list, which can be found in Appendix B.

This research paper aims to answer the following research questions:

- **RQ1:** How to identify the severity of the disease with a certain level of accuracy?
- **RQ2:** How to predict COVID-19 disease progression?

The structure of the paper is as follows. First, we discuss the related work. Then the methodology described in which explains data understanding, data preparation, modelling and evaluation in detail. Lastly, the conclusions are drawn, and future work is discussed.

2.   RELATED WORK

On 14 May 2020, Yan et al.[1] published a research paper in which they presented their solution to this problem. Their solution was a mortality prediction model that was developed using XGBoost machine learning tools. These tools selected three biomarkers that predict the mortality rate of individual patients more than ten days in advance with more than 90%

accuracy. The three predictive biomarkers were lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP).

On 8 February 2021, Sun et al. studied different approaches of computing more accurate predictions and progression states [2]. Based on their study results, they have found out that Deep learning approaches (T-LSTM, RNN, PNN and BPNN) have better prediction accuracy than non-deep learning approaches (Cox, k-NN, SVM and DT). The T-LSTM method showed predictive accuracy of more than 90% at 12 days and 98, 95 and 93% at 3, 6 and 9 days, respectively.

3. METHODOLOGY

We have chosen the Cross-Industry Standard Process for Data Mining (CRISP-DM) process model to help us plan, organize, and implement our data science project. The following subsections are the corresponding phases of the CRISP-DM methodology.

3.1 Data Understanding

In this section and its following subsections, we describe collecting, identifying, and analysing the data.

3.1.1 Collecting initial data

The required data to train a model that could predict the severity of COVID-19 comes in an XLSX file format. It is loaded using python with the help of the pandas' package.

3.1.2 Data description

The blood samples are taken between 2020-01-11 and 2020-05-07. The XLSX file contains 3608 rows and 31 columns. All values are numerical. Blood features and their characteristics can be found in Appendix A. The data is sparse; 17 features have over 94% of zero values.

3.1.3 Data exploration

To visualise the linear correlation between the biomarkers, Pearson Correlation will be used to plot a heat map found in figure 1. Heat map represents the data in the form of a map in which data values are represented as colours.
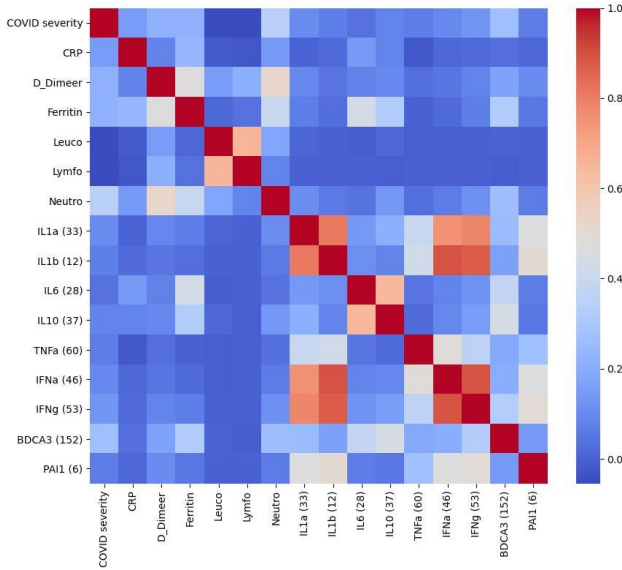

Figure 1. Pearson correlation heat map of raw data.

According to a medical expert [6], the correlation coefficients between white blood cells are not as expected. The expected correlation coefficient between leukocytes and neutrophils is to be higher than 0.9. In this case, it is 0.178 with a p-value of less than 0.01, meaning that there is less than a 1% chance that the resulting correlation coefficient occurred due to chance. P-value, which is lower than 0.05, indicates that the results are statistically significant.

Most likely, the results are not as expected because of sparse data. After we clean our data during the data preparation phase, we will plot a new correlation heatmap and investigate the results.

3.2 Data Preparation

In the following subsections, steps taken to prepare the data are explained.

3.2.1 Data selection

Below are the steps taken for data exclusion.

1. Exclude 728 rows where patients' blood sample data are missing completely at random (MCAR), meaning that the missingness has nothing to do with the patient.
2. Exclude eight features that we have not identified as significant predictors of COVID-19 severity (MIF, Gal1, LAP, Leptin, Ang1, Ang2, Adiponectin, Fibronectin).
3. Exclude 1746 rows that represent patients who tested negative for COVID-19 polymerase chain reaction (PCR) test.

We exclude the latter because that data is not relevant for answering our research question.

3.2.2 Data cleaning

In this section and its following subsections, data cleaning steps are documented.

3.2.2.1 Missing cytokine measurements

To slightly reduce the sparsity of the dataset, medical expert [6] gave advice that if there is a cytokine measurement missing while there is another cytokine measurement present for the same patient, we can substitute the missing cytokine measurements with the half lower limit of detection (LLOD) of that cytokine. In the medical field, this is an acceptable thing to do with the missing cytokine measurements [6].
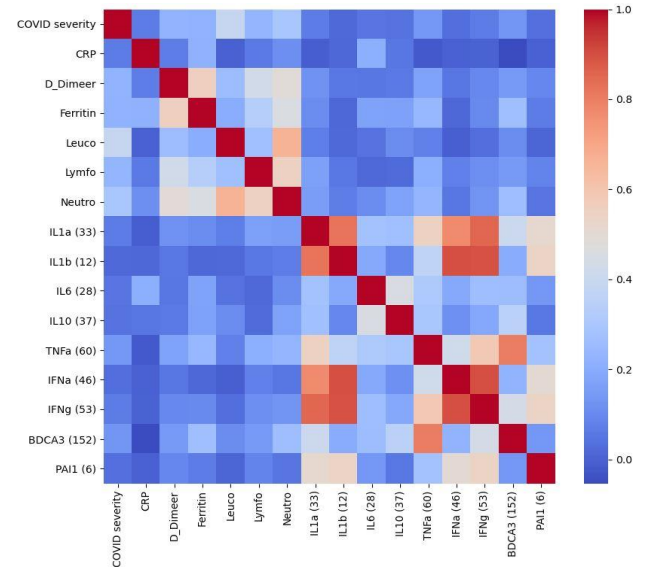
Figure 2. Pearson correlation heatmap of selected data.

Now, the correlation coefficient between leukocytes and neutrophils is 0.665 (0.001>p). The correlations between cytokines positively increased due to the substitution of zero values with LLOD.
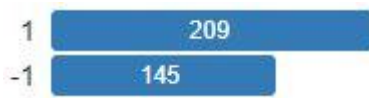
3.2.2.2 Class imbalance

If a patient is severely ill, he usually has his/her blood sample taken daily. That is the reason why 87.21% of our selected dataset consists of class 1.

Figure 4. Class imbalance

To solve this problem, we will use a random downsampling method. To our selected dataset, we will only include the first four measurements for each severely ill patient.
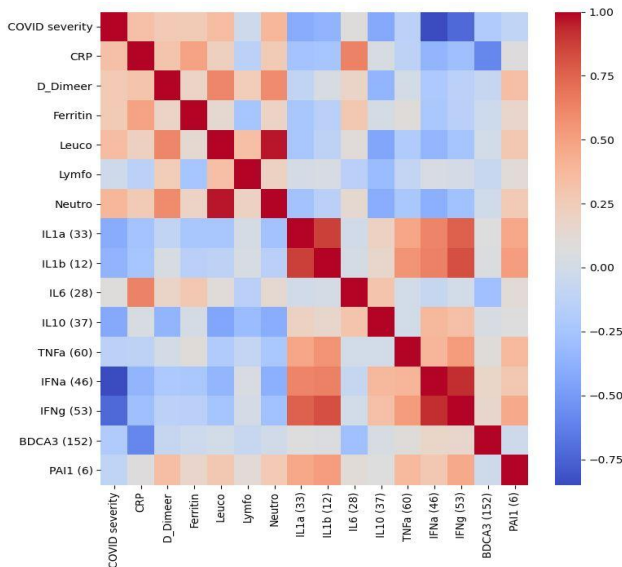

Figure 5. Class imbalance after random downsampling.

3.2.2.3 Missing value imputation

Missing value imputation

To impute the missing biomarker measurements, we will use the MissForest imputation method. MissForest is arguably the best method of imputation for handling missing laboratory measurements in medicine [10].

Figure 6. Resulting Pearson correlation after MissForest imputation.

As expected, there is a positive, 0.969, correlation between the leukocytes and neutrophils (0.0001>p). This is a good shred of evidence that the MissForest imputation method is imputing values correctly.

3.2.3 Constructing data

A new feature indicating a severe illness is derived from the dataset. Patients who tested positive for COVID-19 and had either mild or moderate symptoms are labelled as non-severely ill. Patients with severe COVID-19 symptoms are labelled as severely ill; -1, and 1, respectively.

3.3 Modelling

Here we will build and assess different models under different data preparation states.

3.3.1 Selecting modelling techniques

To build our classification model we will test Support Vector Machine and Logistic Regression methods. The SVM model is created with a parameter kernel set to "linear". The Logistic Regression model is created with a parameter solver set to "liblinear".

3.3.2 Test design

We split our data into training and testing sets. When splitting the data, we ensure that training and testing sets maintain the same class balance as the dataset does before splitting it. 80% of data is used to train the models, and 20% is used to test them.

3.3.3 Building the models

In the following subsection, the models will be built several times to compare how different data preparation affects their performance. ROC curves, AUROC scores and confusion matrices can be found in Appendix C.

3.3.3.1 Initial data preparation

The results in Table 1 signal the overfitting of models. Given our setting, an accuracy of 97% is not realistic.

Table 1. Metric results after initial data preparation.

Method	10-fold CV avg. accuracy	Precision	Recall	F1 score
SVM	0.975	0.977	1	0.989
Logistic Regression	0.965	0.955	1	0.977

3.3.3.2 Excluding cytokines

We drop cytokines from the feature list. The new results can be found in Table 2.

Table 2. Metric results after dropping cytokines.

Method	10-fold CV avg. accuracy	Precision	Recall	F1 score
SVM	0.721	0.775	0.738	0.756
Logistic Regression	0.711	0.711	0.762	0.736

3.3.3.3 Adjusting class weights

As our training and testing sets have a class imbalance of 41:59, we build the models using matching class weights.

Table 3. Metric results after adjusting class weights.

Method	10-fold CV avg. accuracy	Precision	Recall	F1 score
SVM	0.727	0.780	0.929	0.848
Logistic Regression	0.685	0.7	1	0.824

3.4 Evaluation and discussion of results

The models were built under three different cases.

Case 1: Using the data preparation as explained in subsection 3.2.

Case 2: Same as case 1, but excluding cytokines from the feature list.

Case 3: Same as case 2, but assigning class weights that match the class imbalance of the training and testing sets.

When training the model according to case one, we ran into data overfitting. However, by dropping all cytokines from the features list, we got rid of that problem. A high number of features introduced models into overfitting. We chose cytokines to be excluded from the feature list because, initially, over 94% of their measurements were missing.

After dropping the cytokines from our list of features, our accuracy dropped significantly. It was a good sign that our model was not overfitted with data anymore. The precision and recall scores were similar, meaning that the model could predict both classes with similar success probability for each class.

We did not stop at case two because the success probabilities for both classes were similar. This classification model's goal was to predict if a person has a severe or non-severe case of COVID-19; it is better if we produce more false positives (type I errors) than false negatives (type II errors). Since a patient's health is of utmost importance, we would rather classify the patient as severely ill and run more thorough checks to confirm it than classify the patient as non-severely ill and let him heal at home in critical condition.

Building the models under the conditions of case three yielded our best results. While SVM achieved higher accuracy because of making fewer type I errors (false positives), the model of our choice for the deployment is Logistic Regression because of the reasons explained in the paragraph above.

4. CONCLUSIONS

Classification techniques are a powerful tool to categorise data. When applying these techniques in medicine, we can create a model that could help medical personnel decide on appropriate medical resource allocation. In some cases, the appropriate allocation of such resources can save a person's life. We answer our first research question by training a Logistic Regression model, predicting the severity of the coronavirus disease with an average accuracy of 68%. However, due to time constraints and limited data, the second research question is left unanswered.

5. FUTURE WORK

Many different directions and experiments have been left for future work due to this research's limited time and data constraints. The following ideas could be analysed:

- It should be interesting to see how different biomarker correlations differ between different age groups. Our correlation heat maps included patients aged between 28.33 and 87.24 years. Knowing that the people aged 50 and older are the most vulnerable to the infection, it could be interesting to see how correlation differs for that age group.
- A model could be created to signal the patients if they are more likely to get acute kidney injury or new bacterial cultures due to the increased IL-10 and IL-10/lymphocyte count [7].

7. REFERENCES

[1]   Yan, Li, Hai-Tao Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun et al. "An interpretable mortality prediction model for COVID-19 patients." *Nature machine intelligence* 2, no. 5 (2020): 283-288. DOI=https://doi.org/10.1038/s42256-020-0180-7

[2]   Sun, Chenxi, Shenda Hong, Moxian Song, Hongyan Li, and Zhenjie Wang. "Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning." *BMC Medical Informatics and Decision Making* 21, no. 1 (2021): 1-16. DOI=https://doi.org/10.1186/s12911-020-01359-9

[3]   Möhlenkamp, Stefan, and Holger Thiele. "Ventilation of COVID-19 patients in intensive care units." *Herz* 45, no. 4 (2020): 329-331. DOI=https://doi.org/10.1007/s00059-020-04923-1

[4]   Ma, Xiya, and Dominique Vervoort. "Critical care capacity during the COVID-19 pandemic: global availability of intensive care beds." *Journal of critical care* 58 (2020): 96. DOI=https://doi.org/10.1016/j.jcrc.2020.04.012

[5]   Yang, Xiaobo, Yuan Yu, Jiqian Xu, Huaqing Shu, Hong Liu, Yongran Wu, Lu Zhang et al. "Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study." *The Lancet Respiratory Medicine* 8, no. 5 (2020): 475-481. DOI=https://doi.org/10.1016/S2213-2600(20)30079-5

[6]   Conversation with dr. J.G. Krabbe.

[7]   Henry, Brandon Michael, Stefanie W. Benoit, Jens Vikse, Brandon A. Berger, Christina Pulvino, Jonathan Hoehn, James Rose, Maria Helena Santos de Oliveira, Giuseppe Lippi, and Justin L. Benoit. "The anti-inflammatory cytokine response characterized by elevated interleukin-10 is a stronger predictor of severe disease and poor outcomes than the pro-inflammatory cytokine response in coronavirus disease 2019 (COVID-19)." *Clinical Chemistry and Laboratory Medicine (CCLM)* 1, no. ahead-of-print (2020).

[8]   Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." *The lancet* 395, no. 10223 (2020): 497-506.

[9]   Yang, He S., Yu Hou, Ljiljana V. Vasovic, Peter AD Steel, Amy Chadburn, Sabrina E. Racine-Brzostek, Priya Velu et al. "Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning." *Clinical chemistry* 66, no. 11 (2020): 1396-1404.

[10]   Waljee, Akbar K., Ashin Mukherjee, Amit G. Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. "Comparison of imputation methods for missing laboratory data in medicine." *BMJ open* 3, no. 8 (2013).

APPENDIX

A. Demographic, and laboratory outcome information of 3608 samples in the provided dataset.

Characteristics	Statistics	
Age (Q3, Q1, IQR)	73.96, 58.09, 15.87	
Gender	2274 male, 1334 female	
Variable	**Zero value count, %**	**Mean**
CRP	850, 23.6%	88.11
D_Dimeer	2551, 70.7%	726.53
Ferritin	2618, 72.6%	382.49
Leuco	837, 23.2%	10.49
Lymfo	2298, 63.7%	1.1
Neutro	2297, 63.7%	3.45
IL1a (33)	3568, 98.9%	0.06
IL1b (12)	3566, 98.8%	0.08
IL6 (28)	3465, 96.0%	9.51
IL10 (37)	3446, 95.5%	0.93
TNFa (60)	3598, 99.7%	0.007
IFNa (46)	3553, 98.5%	0.23
IFNg (53)	3518, 97.5%	0.34
MIF (77)	3534, 97.9%	39.82
Gal1 (144)	3422, 94.8%	2,188.76
BDCA3 (152)	3420, 94.8%	282.95
LAP	3423, 94.9%	223.64
Leptin	3420, 94.8%	465.23
PAI1	3581, 99.3%	3,809.92
Ang1	3420, 94.8%	2,454.14
Ang2	3420, 94.8%	190.41
Adiponectin	3422, 94.8%	3,713,367.38
Fibronectin	3606, 99.95%	6,224.22

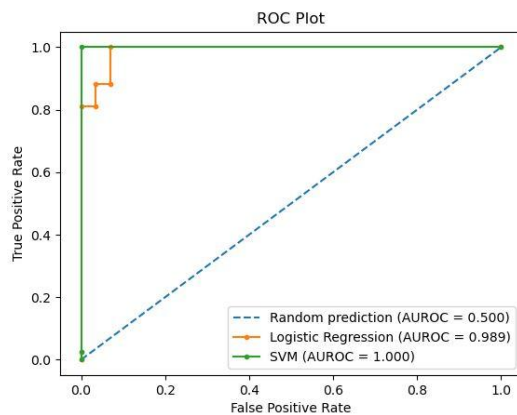**B. Demographic, and laboratory outcome information of 354 samples in the selected dataset.**

Characteristics	Statistics
Age, mean (min, max) (years)	64.54 (28.33, 87.24)
Gender	Male 241; Female 113
COVID severity	Severe 209; Non-Severe 145
Feature	**Mean (min, max)**
CRP	169.49 (8.3, 549)
D-dimer	2494.11 (177, 8819)
Ferritin	2013.36 (128, 10200)
Leukocytes	8.98 (0.5, 36.75)
Lymphocytes	0.87 (0.02, 8.46)
Neutrophils	8.23 (0.51, 32)
IL1-α	2.12 (1.05, 30.67)
IL1-β	2.77 (0.6, 62.52)
IL-6	189.38 (1.4, 1174)
IL-10	20.38 (1.7, 68.81)
TNF-α	1.16 (1.1, 2.62)
IFN-α	6.72 (0.7, 126.03)
IFN-γ	10.41 (0.29, 122.96)
BDCA-3	5140.65 (1074.79, 11687.26)
PAI-1	62281.89 (403.5, 1192100)

C. Receiving Operating Characteristic curve plots and confusion matrices.

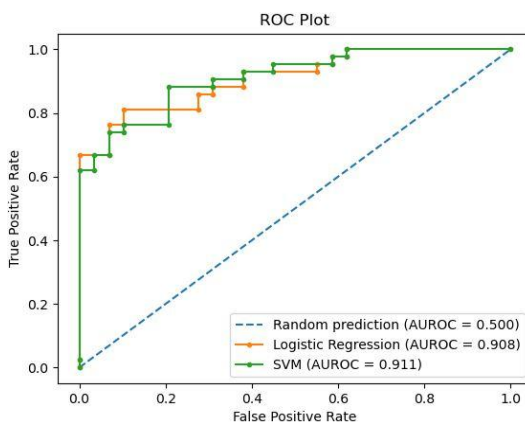### C1. Metric results with data prepared as explained in 3.2.

C2. Results of metrics after excluding cytokines from the feature list 3.3.3.2.
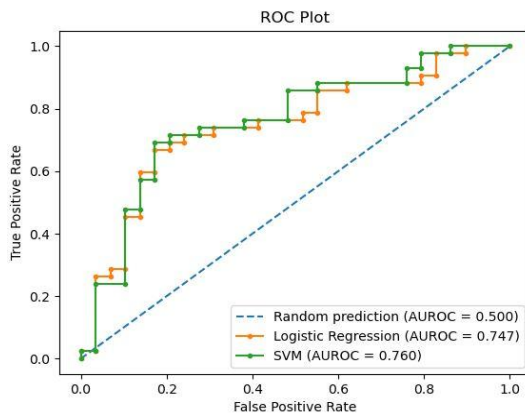
C3. Results of metrics with adjusted class weights in model creation 3.3.3.3.