Predicting COVID severity using machine learning methods
-Comparison between real life and mimic dataset-

Teodora Nae, University of Twente, The Netherlands, t.nae@student.utwente.nl

ABSTRACT
After the declaration of the COVID-19 disease as a pandemic, the hospitals were overflowing with patients. Using machine learning methods to predict the severity of the disease can help the professionals from the medical field better allocate the resources in order to minimize the mortality rate. Knowing which patients to prioritize (the ones more likely to have a severe form of the disease rather than the ones with a non-severe form) would help hospitals to better respond to the needs of each individual infected with COVID-19. The CRISP-DM methodology for preparing the datasets was used in this paper to help with organizing and implementing the project. The aim of this research paper is to predict the severity of the disease based on a number of biomarkers, with the help of different machine learning algorithms. As well as to analyze the discrepancies between the results from two different datasets (a mimic dataset versus a real and accurate dataset) with the same features obtained using the same machine learning methods. For the mimic dataset a total of around 4000 entries were used for training the model, while for the real life set a total of around 700 entries matched the requirements for this study.

Keywords
COVID-19, disease, machine learning algorithms, COVID-19 severity, (blood) biomarkers, Sars-Cov-2, open dataset, real dataset, discrepancies datasets, accurate dataset/database

1. INTRODUCTION
On 11 March 2020, COVID-19 was declared a pandemic by The World Health Organization. Two years later, it is still unknown exactly why some patients get a more severe form of the disease than others. After the COVID-19 outbreak, a set of rules and regulations were implemented in order to minimize the spread of the disease. These measures were not 100% effective. People still got sick and needed medical attention, which led to an overflow of patients in the hospitals. Patients with a severe form of the disease must be transferred to the intensive care unit (ICU), where a ventilator is necessary to maintain body functions that are vital for their survival, such as breathing. About 5-15% of the COVID patients are in need of such equipment while studies show that in the 182 countries and territories studied, there are between 0 and 59.5 beds in the ICU per 100,000 population [1], which is close to 0.6%, far less than the minimum number of infected people in need of such equipment, 5%. Thus it is useful to research what combination of biomarkers make people more likely to have an advanced form of the disease, so they can receive medical attention faster in order to avoid the need of ventilators and ICU transfers as much as possible. This represents the first goal of this research paper, while the second one is performing a comparison between a real set and a mimic set.

The urgency of this problem is very high as the lives of the general population are at risk, thus many scientists tried to discover a solution for this issue with the help of machine learning algorithms as presented in the Related Work section. However, the studies presented in the latter section for certain are not enough because some studies have a very small number of samples for training the model, while others are using a maximum of six biomarkers.

1.2 PROBLEM STATEMENT
In this research paper, we will investigate the relation between different blood biomarkers and the severity of COVID-19. Some research regarding this topic has already been conducted [2,3] and the results show that some biomarkers are known to be important in predicting the severity and progression of the disease. According to Roshanravan et al. [2], IL-6 was elevated in patients with severe COVID-19 conditions. The role of IL-6 is to support immunocompetence, defined as the ability of a host to respond to infections. A study conducted by Henry et al. [4] revealed that IL-10 and IL-10/lymphocyte, IL-10/TNF- , IL-6/lymphocyte count is also increased in patients with a severe form of the disease. Another study, by Huang et al. [5] shows elevated levels of IL-1b in severe COVID-19 cases.IL-1b is essential for the host-response and resistance to pathogens. Yang et al. [6] have found elevation of CRP and ferritin in patients infected with COVID-19. Other studies have shown that a high BMI is strongly related to a severe form of COVID-19 and might even lead to a higher mortality rate [15]. Relevant literature also showed that a high level of d-dimer in the early stages of the COVID-19 disease is linked to a more severe form of the disease [16]. These biomarkers were the first step in this research, but other blood characteristics were added to the list of selected features.

1.2.1 Research Question
Besides the prediction of the disease, in this research paper, an analysis of two datasets created differently will be conducted using the same classifiers. The problem statement leads to the following two main research questions:

1. What are the biomarkers that influence the severity of the disease for the people infected with COVID-19?
2. What differences can be seen between two different datasets (a mimic dataset versus a real and accurate dataset) using the same classifiers?

The first main research question can be answered by using the following sub-questions as guiding points:

RQ1: How to predict the evolution of the disease based on specific biomarkers?

RQ2: How machine learning algorithms can help recognize the severity of the disease with a particular accuracy level?

RQ3: Which person is more likely to have a severe form of the disease based on people in different age groups, gender and with different races?

2. RELATED WORK

Scopus and Google Scholar were the main tools used to gather related literature and references for this research paper. Search terms such as: “COVID biomarkers”, “ventilators”, “severity prediction” and “IL-6” were used and a number of articles were found.

Yan et al.[2] published a research paper in which they created a mortality prediction model using machine learning tools. They selected three biomarkers: lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP). With the help of this model they managed to predict the mortality of individual patients, more than 10 days in advance and with an accuracy of more than 90%.

Huyut et al. [7] discovered in their research that some biomarkers are important parameters in the diagnosis and prognosis of COVID-19 disease. With an accuracy of 65.0% in predicting the prognosis of the disease, they discovered that low ionized-calcium value was present in most patients that needed intensive care. To reach the accuracy of 68.2% in the diagnosis of the disease, they discovered that low-carboxyhemoglobin, high-pH, low-sodium, hematocrit and methemoglobin values are important biomarkers.

Kilercik et al. [8] published a research paper in which they discovered slight differences in hemoglobin or other anemia-related parameters could be observed after grouping infected patients based on the severity of the disease. The levels of ZnPP were significantly increased in the group of patients with an advanced form of the disease. They also discovered that the ratio of ZnPP to lymphocyte count (ZnPP/L) is an important parameter in determining the severity of the disease

O iūnas [9] predicted the severity of the disease based on a total of eighteen biomarkers. He used SVM and Linear Regression as classifiers for training his model, obtaining a precision of 0.78 and 0.7 respectively. The dataset used in his model was a real and accurate one obtained from a hospital which contained a small number of samples for training the model.

Xiong, Yibai, et al. [21] compared different machine learning techniques for predicting COVID-19 severity, Random Forest, SVM and Linear Regression. They discovered that Random Forest proved to be the most efficient, with the highest AUC, equal to 0.970. They used a total of 23 features for a total of 283 samples.

Kang, Jianhong, et al. developed a predictive model for the patients with a severe form of the disease. The dataset used contained a total of 151 patients from a hospital in China from a period of almost two months. The model achieved good performances, with the AUC of 0.953 and they also discovered a possible correlation between three biomarkers(a low albumin, a high globulin and a high blood urea nitrogen) and a severe form of COVID-19.

3. METHODOLOGY

The methodology will be the same one used by O iūnas [9] in his research paper, specifically, the Cross-Industry Standard Process for Data Mining (CRISP-DM) process model. This process model was published in 1999 with the scope to standardize data mining processes. Now, it is known as the most common methodology for data mining and it will be used in this paper to help with organizing and implementing the project. CRISP-DM has six sequential steps:

1. Business understanding,

2. Data understanding,

3. Data preparation,

4. Modeling,

5. Evaluation,

6. Deployment.

The first and the last step will not be discussed in this research paper, while the fifth step is in the section below, Results. The rest of the steps will have a special subsection below.

3.1 Data Understanding

The first step for this phase is to collect the data that will be used to train and test the machine learning algorithm. The csv file will be loaded in python with the help of the pandas library. The data set used will have the same format as the one used by O iūnas [9] in his research paper, 4313 rows and 53 columns containing blood features and their characteristics. The next step for this phase will be to perform data exploration with the help of different visualization tools, such as heat maps, in order to better understand the data set and its features.

Using the function info(), the number of not-null values for each feature is shown and this helps decide what features contain too little information and need to be removed. The function head() is also used to better understand the database as it returns the head of the dataset (the name of the columns) and the first 5 rows which gives a better understanding of the values of the features.

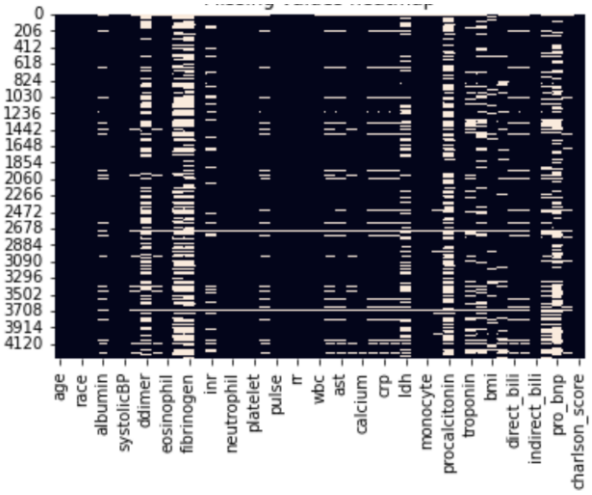

Figure 1: Heatmap of the raw dataset

In Figure 1 above is the heat map of the original dataset after the removal of unwanted features but before the imputation of new values for the features that do not contain information for all the entries. On the left there is the number of the rows in the dataset and on the bottom are the features, while in the white lines represent the values missing for the specific row and feature of the set.

4.2 Data Preparation

After the data understanding phase is complete and a good understanding of the database is achieved, we move to the next phase of the CRISP-DM model. A number of methods of data cleaning will be applied such as: remove duplicates, standardize capitalization, convert data type and the most important method, handle missing values. Two different ways for handling missing values were used: deleting and imputing missing values. For features where too much information is missing, the feature was removed completely, while for features where only some of the information is unavailable, the median between the existing values was imputed.

Four features were removed completely because of two reasons; either the feature was not important for the research such as the 'Time_from_COVID_positive_to_death_in_days' and 'death' or there was not enough information for this column, which was the case for the following two columns: 'ct_value' and 'tnf'.

Many features had missing values but for some of them, the median of the existing values could not be imputed as this method would bias the dataset. According to the doctor H. Krabbe, for the d-dimer column, it is medically accurate to add a higher value for the missing data because whenever a person is sick, this biomarker tends to increase its value. The same was implemented for the fibrinogen biomarker, for which the suggested value is 2. For the rest of the features used that had missing values in the original dataset, the mean of the existing entries was added, without influencing the accuracy of the results, using the mean() function. The figure below represents the heatmap of the dataset after the missing values were imputed for the features that were of interest for this research; biomarkers that influenced the severity of the disease.

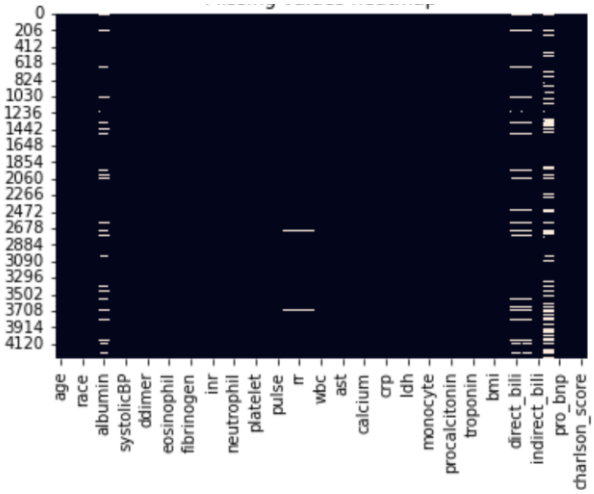

Figure 2: Heatmap of the dataset after cleaning as explained above

Two more columns were created after the cleaning of the set was done. A column called Covid Severity which has only two values: mild or severe. It was created based on an existing column called ventilator which had a 1 for people that needed a ventilator to get better and a 0 for the other patients. The people who needed a ventilator were considered to have a severe form of the disease and it was added as such in the newly created column. The second column created is a ratio between two biomarkers, interleukin 6 and lymphocytes, and it was added later as a feature for training the model.

The last step in preparing the database before the training of the model is normalization. The min-max method was used because it was proven to be more efficient in other studies [18] than the z-score method. Normalization is useful for classification algorithms such as those used in this research paper to train the model because it gives all features equal weights, which leads to a better accuracy. This step in cleaning the dataset is mandatory especially for comparing the two datasets, using F-1 score and Kernel Density Estimate (KDE) plots (more details can be found in sub-section 5.1).

4.3 Modeling

The main concern for this phase is to not overfit the model. Overfitting is one of the most common problems that occur in data mining. Model overfitting appears when the model fits too close to the training set and is not able to perform well on a new data set. In order to avoid this issue as much as possible, some features may be removed from the selected feature list. Different classification algorithms will be used in order to find the one that fits the database best.

Two classifiers were selected for the training of this model. The first one is the Support Vector Machine (SVM) and Linear Regression and they will be compared with the use of the F1 score. SVM works well with unstructured data and according to [12] 'it is able to capture non-linearities in the data', while linear regression works better on identified independent variables. Another difference between the two classifiers is that linear regression is focused on statistical approaches [13] while SVM is focused on geometrical properties [14].

There were two reasons for the selection of these two classifiers. The first one, is the fact that it is expected to perform very differently on the given training set due to the different approaches of performing as well as the types of data on which the two classifiers return better accuracies. The other reason for this choice was due to the fact that O iūnas [9] , in his research paper, used the same classifiers and this way it is easier to compare the two datasets. O iūnas' [9] database is a real one, received from a hospital which means that all the information in it is real, accurate and authentic, while the dataset described in this research paper is found on the internet and it is a mimic set. SVM and Linear Regression return different accuracies on the two datasets with a significant difference between the two results and finding out why this is happening is easier if the same classifiers for training the model are used.

4. RESULTS

Research about the correlation between different biomarkers and the severity of COVID-19 disease have been made. However, the number of biomarkers is quite small in these researches, varying from three to six biomarkers, which is improved in this research.

In order to evaluate the model and its performance, we analyzed both the accuracy of the testing data set and we also created a confusion matrix for each of the three cases described below. There are two types of errors that can appear in a confusion matrix: Type 1 (false positives) and Type 2 (false negatives). The scope of the model is to predict the severity of the disease, thus it is better to wrongly classify a patient as severely ill rather than send him home with an advanced form of the disease. In order to be able to accurately compare the two models, the F1-score was calculated and the number of wrong

predictions of type 1 was also taken into account. In appendix A the confusion matrix for each of the cases described below, can be found.

$$\text{precision} = \frac{TP}{TP+FN}$$

$$\text{recall} = \frac{TP}{TP+FN}$$

$$\text{F-1 score} = \frac{2*\text{precision}*\text{recall}}{\text{precission}+\text{recall}}$$

In appendix B, there is a table with all the features used to train the model as well as some statistics for each of them and the main role of the specific biomarker. The features written in blue are the ones that are common between the two datasets explained in the following sub-section 'Comparison between the two datasets'. With the use of these biomarkers, the model was trained and three different cases appeared:

Case 1: All the features in appendix B were used with exception of the race.

Method	Precision	Recall	F1-score
SVM	0.887	1	0.94
Logistic Regression	0.90	1	0.947

Case 2: The number of samples categorized as severe or mild is uneven, thus random undersampling was performed for all the features with the exception of the race.

Method	Precision	Recall	F1-score
SVM	0.893	1	0.943
Linear Regression	0.907	1	0.951

Case 3: After random undersampling the race was added and it returned a slightly better accuracy for the SVM classifier while keeping the same accuracy for the Logistic Regression.

Method	Precision	Recall	F1-score
SVM	0.895	1	0.944
Linear Regression	0.907	1	0.951

The choice of these biomarkers was made at the recommendation of a medical professional, doctor H. Krabbe as well as other studies that confirm the importance of these biomarkers for the form of COVID-19 the patients present, as presented in the sections 'Problem Statement' and 'Related Work'.

Below, there are three pie charts representing the characteristics of the most vulnerable person to get a severe form of the disease. As it can be seen (from figures 3,4, and 5) the most vulnerable person is an African American male, between the ages 61 and 70 years.

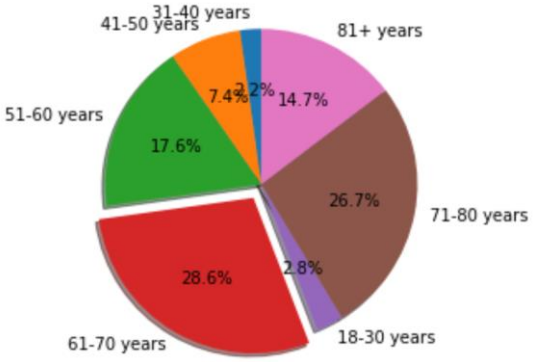

Figure 3: Pie chart for age for the patients with a severe form of the disease.

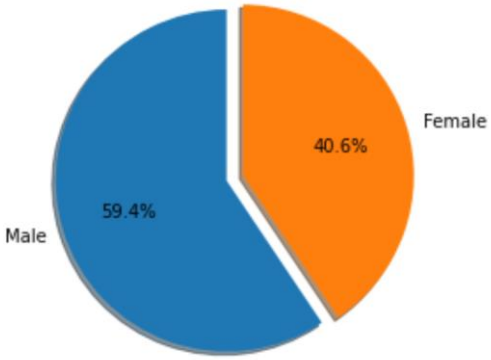

Figure 4: Pie chart for gender for the patients with a severe form of the disease.

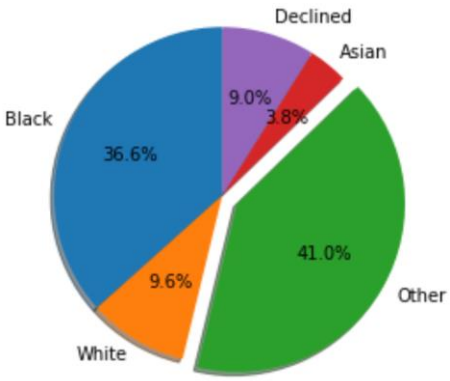

Figure 5: Pie chart for race for the patients with a severe form of the disease.

For the race pie chart, the category with the biggest percent is 'Others' which includes a variety of races like: Native-American Alaskan and Other-Pacific-Islander but since the

distribution of races that are part of this category is unknown, the race with the second highest percentage is used.

4.1 Comparison between the two datasets

After the expected results were reached and analyzed, a comparison between two datasets was performed: the dataset analyzed by O iūnas [9] and the dataset presented in this research paper up to this point.

The steps presented in the Methodology section were repeated for the real and accurate dataset, only the most important changes are presented in this paragraph. After eliminating the patients that tested negative for COVID, a total of 1472 samples out of 3608 remained. From the remaining samples, 770 rows were removed because of the lack of information present on each of these rows for the biomarkers that are common between the mimic and real datasets. The total number of samples used from the real life database is 702. Normalization was also performed on this dataset, in order to be able to properly compare the two sets.

Both datasets have the same features and the results are reached with the same machine learning techniques, SVM and Linear Regression. However, since the two datasets are created in very different ways; the one presented in this paper is a mimic set and the values of the features might not be as accurate as the values of the biomarkers of the database analyzed in O iūnas' [9] paper, which came from a hospital. Due to this, a big discrepancy in the results is expected and it can be seen in the KDE plots and the table below.

Using only the common biomarkers between the two datasets, one KDE plot for each of the sets was created. The main goal of these KDE plots is to see how the data is distributed for each of the databases.

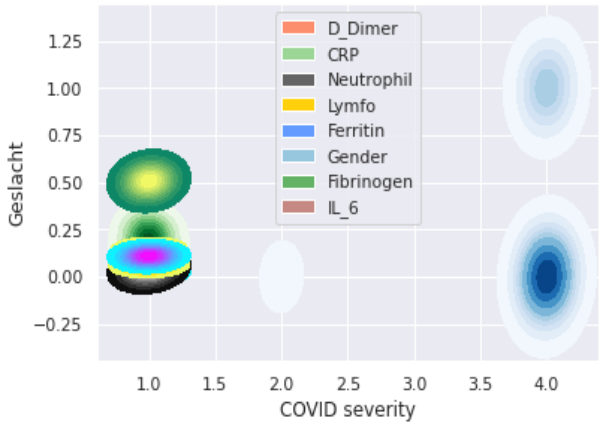

Figure 3:KDE plot for the real database from the hospital

In Figure 3, there is the KDE plot for the real dataset. All the features have values between 0 and 1 and on the OX axis that represents the severity of the disease (from 1 ~ the mildest form to 4 ~ the most severe form of COVID-19), the OY axis represents the gender (0 for men and 1 for women). Where the colors are more intense is where the biggest incidence for the specific biomarker can be seen.

In Figure 4 is the KDE plot for the mimic database. As in the previous plot, on the OX axis is the severity of the disease (0 ~ mild form and 1~ severe form), while on the OY axis is the gender (0 ~ men and 1 ~ women).

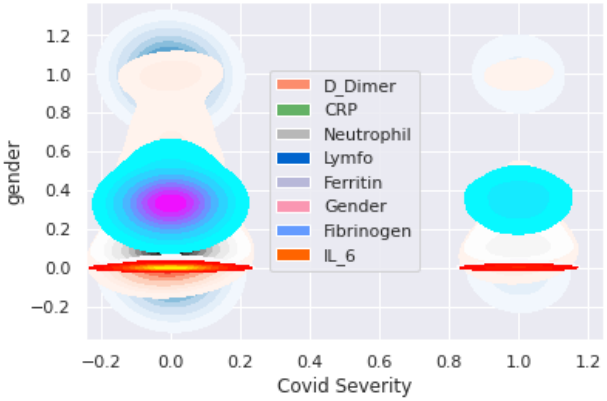

Figure 4 :KDE plot for the mimic database

In both cases, fibrinogen (green in figure 3 and blue in figure 4) has the biggest incidence for the mild form of the disease, while gender (blue in figure 3 and pink in figure 4) has the biggest incidence in both databases for the severe form of the disease. One of the main differences between the distribution of the data from the two datasets is the fact that interleukin 6 has a big incidence for the mild form of the disease in the mimic set, which is not the case for the real set.

The table below represents the results obtained on the mimic dataset by using only the parameters that are present in both sets (written with black) and the result obtained on O iūnas' [9] dataset (written with blue).

Method	Precision	Recall	F1-score
SVM	0.86	1	0.92
SVM	0.71	0.96	0.81
Logistic Regression	0.85	1	0.91
Logistic Regression	0.75	1	0.85

While these results are not as good as the ones presented above (for the mimic dataset), they are still much better than the ones obtained on the real dataset used by O iūnas [9] in his research and analyzed in comparison with the mimic dataset in this paper. This leads to the conclusion that mimic datasets do not contain accurate or real data and cannot be the sole source for research.

From the above table, it is clear that both machine learning techniques used perform better on the mimic dataset. The biggest discrepancy between the results obtained on both sets comes from the SVM classifier, with a difference of 15% in precision. This discrepancy is slightly reduced for the Logistic Regression classifier, with only 10% better precision for the mimic dataset.

The uneven distribution of the information from the two datasets, as presented in the KDE plots, is one of the reasons for

the big discrepancy in the accuracies of the two sets. Another reason is that the real database contains less entries (around 700) after the cleaning steps (including the normalization) were performed, while the mimic one has much more (around 4000).

5. CONCLUSION

Machine learning algorithms are powerful classifying tools that used in medicine can create models that are helpful for the medical professionals. With the help of these models the medical personnel can better allocate their resources and focus on the more ill patient. Sometimes, these models are the difference between life and death.

The first main research question is answered by training a Logistic regression model with an accuracy of 90.7% and a SVM model with an accuracy of 89.5. The features can be found in appendix B and how these biomarkers were chosen is explained in the 'Results' section.

The second main research question is answered by presenting the difference of results between the two datasets when the model is trained with the same features and using the same classifiers. Both machine learning techniques used return a better precision for the mimic dataset. A difference of 15% in the accuracy between the databases for the SVM classifier and only 10% difference for the Logistic Regression classifier can be seen.

6. FUTURE WORK

Due to the limited time for this research, there are many directions for future work not discussed. A few direction for future research are:

- Analyzing the effects of long COVID on the major organs like kidneys, lungs and liver. Long COVID is the persistence of the symptoms after the end of the infection [17]
- Predicting the severity of COVID for people that are vaccinated with one, two or three doses as well as researching the changes in the biomarkers for vaccinated people.

7. ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisors, dr. F. Bukhsh and dr. F. Ahmed, for being near me at every step of this journey and for guiding me in reaching the best results possible. I would also like to extend my sincere thanks to dr. H. Krabbe for giving me the medical advice without which this project would not have been possible.

8. REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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).

[5] 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.

[6] 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

[7] Huyut, M., Üstündağ, H "Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study" Medical Gas Research Volume 12, Issue 2, Pages 60 - 66, 2022

[8] Kilercik, M., Ucal, Y., Serdar M et al "Zinc protoporphyrin levels in COVID-19 are indicative of iron deficiency and potential predictor of disease severity" PLoS ONE Volume 17, Issue 2 February 2022

[9] O iūnas, D "Identifying Severity of COVID-19 In Patients Using Machine Learning Methods" June 2021

[10] Wirth, R., & Hipp, J. "CRISP-DM: Towards a Standard Process Model for Data Mining" 2000 DOI = 10.4103/2045-9912.326002

[11] Mihara, Masahiko, et al. "IL-6/IL-6 Receptor System and Its Role in Physiological and Pathological Conditions." Clinical Science, vol. 122, no. 4, 1 Feb. 2012, pp. 143–159

[12] Martens, David, et al. "Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring." *Rule Extraction from Support Vector Machines*, 2008, pp. 33–63

[13] Su, Xiaogang, et al. "Linear Regression." *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, 10 Feb. 2012, pp. 275–294

[14] Chen, Pai-Hsuen, et al. "A Tutorial Onv-Support Vector Machines." *Applied Stochastic Models in Business and Industry*, vol. 21, no. 2, Mar. 2005, pp. 111–136.

[15] Du, Yanbin, et al. "Association of Body Mass Index (BMI) with Critical COVID-19 and In-Hospital Mortality: A Dose-Response Meta-Analysis." *Metabolism*, vol. 117, Sept. 2020

[16] Rostami, Mehrdad, and Hassan Mansouritorghabeh. "D-Dimer Level in COVID-19 Infection: A Systematic Review." *Expert Review of Hematology*, vol. 13, no. 11, 1 Nov. 2020, pp. 1265–1275

[17] Raveendran, A.V., et al. "Long COVID: An Overview." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 3, May 2021, pp. 869–875

[18] Henderi, Henderi, et al. "Comparison of Min-Max Normalization and Z-Score Normalization in the K-Nearest Neighbor (KNN) Algorithm to Test the Accuracy of Types of Breast Cancer." *IJIIS: International Journal of Informatics and Information Systems*, vol. 4, no. 1, 1 Mar. 2021, pp. 13–20.

[19]Merriam-Webster. *Merriam Webster's Collegiate Dictionary.* Springfield, Merriam-Webster, 2014.

[20]Davies, Julie. "Procalcitonin." *Journal of Clinical Pathology*, vol. 68, no. 9, 29 June 2015, pp. 675–679.

[21]Xiong, Yibai, et al. "Comparing Different Machine Learning Techniques for Predicting COVID-19 Severity." *Infectious Diseases of Poverty*, vol. 11, no. 1, 17 Feb. 2022.

[22]Kang, Jianhong, et al. "Machine Learning Predictive Model for Severe COVID-19." *Infection, Genetics and Evolution*, vol. 90, June 2021.

10. APPENDIX

Appendix A: Confusion Matrix for each of the cases described in the 'Results' section.

A1: Case 1 ~ Using all the features in appendix B without 'race' and before random undersampling.

N=188	Predicted: MILD	Predicted: SEVERE
Actual: MILD	TN= 100	FP= 4
Actual: SEVERE	FN= 18	TP=66

SVM

N=188	Predicted: MILD	Predicted: SEVERE
Actual: MILD	TN= 93	FP= 11
Actual: SEVERE	FN= 8	TP=76

Logistic Regression

A2: Case 2 ~ Using all the features in appendix B without 'race' and after random undersampling.

N=188	Predicted: MILD	Predicted: SEVERE
Actual: MILD	TN= 94	FP= 10
Actual: SEVERE	FN= 10	TP=74

SVM

N=188	Predicted: MILD	Predicted: SEVERE
Actual: MILD	TN= 91	FP= 13
Actual: SEVERE	FN= 6	TP=78

Logistic Regression

A3: Case 3 ~ Using all the features in appendix B including 'race'.

N=188	Predicted: MILD	Predicted: SEVERE
Actual: MILD	TN=92	FP=12
Actual: SEVERE	FN= 7	TP=77

SVM

N=188	Predicted: MILD	Predicted: SEVERE
Actual: MILD	TN=94	FP=10
Actual: SEVERE	FN= 6	TP=78

Logistic Regression

Appendix B: Laboratory and demographic information of the 4313 samples from the dataset used in training the models.

Feature	Statistics (Mean-Min-Max)	Role/Description
Age	18-89	The ages of the patients from the databases are between 18 and 89
Gender	Males: 2289; Females: 2024	The division of the sample data based on gender
Race	African American: 1560; Other: 1794; Asian: 113; White: 428; Declined: 418	The division of the sample data based on race
Lymphocyte	1.34 (0.1 - 209.1)	"Any of the white blood cells of the immune system that play a role in recognizing and destroying foreign cells, particles, or substances that have invaded the body" [19]
Fibrinogen	3.7 (1.3-6.2)	"A plasma protein that is produced in the liver and is converted into fibrin during blood clot formation" [19]
Neutrophil	6.74 (0.1 - 48.1)	"A granulocyte that is the chief phagocytic white blood cell of the blood" [19]
Interleukin 6	0.3 (0.19 - 20)	"Is produced by various cells, stimulates synthesis of plasma proteins, and plays a role in producing fever" [19]
D-dimer	4 (0.22 - 20)	"A compound formed by the union of two radicals or two molecules of a simpler compound" [19]
Ferritin	1359.71 (1.9 - 100000)	"A protein that functions in the storage of iron and is found especially in the liver and spleen" [19]
Interleukin 6/ Lymphocyte	0.37 (0.00095 - 25)	The ratio between interleukin 6 and lymphocytes.
BMI	30.29 (9.9 - 3069.26)	The body mass index.
PTT	35.37 (19 - 200)	"A complex enzyme found especially in platelets with role in the clotting of blood" [19]
Procalcitonin	1.29 (0.05 - 2.37)	"Is produced by the C cells of the thyroid. Its synthesis is in bacterial infection" [20]