# Comorbidity Based Death Risk Prediction in COVID-19 Patients Using Support Vector Machine (SVM)

**Erdhi Widyarto Nugroho**
Department of Information System, Faculty of Computer Science
Department of Accounting, Faculty of Economics and Business
Soegijapranata Catholic University, Semarang, Indonesia
erdhi@unika.ac.id

Abstract— The covid19 pandemic has hit almost all countries. Covid-19 is a disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which results in many deaths, especially for patients who have comorbidities. By using the machine learning SVM method, predict of the classification between comorbid and death.. The comorbid taken were pneumonia, diabetes, COPD, asthma, insomnia, hypertension, cardiovascular, chronic renal, obesity, and USM. This paper also added with smokers in the feature. The data were taken from Kaggle which is data from the Mexican government from 2020-2021. The SVM method uses a linear kernel and radial basis function to get the F1 value to know which these have the results better. This paper also compares the results of F1 values using other methods such as KNN, Logistic regression, Xboost, decision tree, Random Forest and Multilayer Perceptron and the last, to know the importance feature or which comorbid has the highest death rate using SVM. The Result is SVM uses linear and rbf gets almost the same F1 value. It also same with other methods and the pneumonia has the highest death rate.

Keywords— Covid-19, Comorbid, Death Risk, SVM

## I. INTRODUCTION

The covid-19 pandemic that has occurred since the end of 2019 has attacked countries. based on data from ourworldindata.org, On November 28, 2022 there were 641.56 million people in the world infected with the Covid-19 virus. Of the infected population, 6.63 million died[1]. Covid-19 is a disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which infects the respiratory tract and causes various symptoms such as fever, cough, sore throat, nausea, vomiting, dizziness, loss of ability to taste and smell, and can also not cause symptoms [2].

Symptoms will appear within 2-14 days after exposure to covid-19. The severity is influenced by age and co-morbidities (pre-existing diseases) such as hypertension, diabetes mellitus, asthma, and so on [3]. The most common co-morbidities found in Covid-19 patients are diabetes mellitus, hypertension, and obesity [4]. Obesity is a risk factor for severity of Covid-19 cases, the higher the BMI, the higher the risk of severity[5].

Patients who are confirmed positive for Covid-19 with comorbid or co-morbidities are included in the range group. In fact, comorbidities are the most common cause of death for Covid-19 patients in East Java, West Java and South Sulawesi. East Java comorbid patients as many as 95% of positive Covid-19 patients died. South Sulawesi Mortality cases in the region are almost 97% caused by comorbidities. From South Sulawesi, the number of deaths has decreased by 2.6%, which is a burden for Covid-19 patients with comorbidities. In Central Java, the cause of death for Covid-19 patients is caused by various factors, namely agent, host, environment and health service factors[6].

Machine learning, a part of artificial intelligence, is often used to predict phenomena that occur in society. For this reason, the author tries to see the relationship between co-morbidities and death rates from Covid-19 using machine

learning data taken from kaggle uploaded by MEIR NIZRI. This data is data on the covid conditions of the Mexican government from 2020-2021. The amount of data is 1,048,576 patient data. From this data predictions will be made using the Support Vector Machine (SVM) method to predict which comorbid diseases are most at risk of death. The comorbid diseases taken were pneumonia, diabetes, COPD, asthma, insomnia, hypertension, cardiovascular, chronic renal, obesity, tobacco and usmr.

The research question for this paper is the comparison of the F1 score between the SVC kernels, namely linear and radial basic function (RBC). In predicting death from comorbidities. The second is how to obtain comorbid diseases that are most at risk of death based on svm. How does the svm F1score compare to other methods such as KNN, Logistic regression, Xboost, Decision tree, Random Forest and Multilayer Perceptron

1.2 Literature review

Machine learning is often used to predict risk factors for a disease, not just cases for Covid. Ferdowsy et al conducted a study on the effect of disease susceptibility due to obesity. Data taken from cases in Bangladesh. Ferdowsy compares various prediction methods such as k-nearest neighbour (k-NN), random forest, logistic regression, multilayer perceptron (MLP), support vector machine (SVM), naïve Bayes, adaptive boosting (ADA boosting), decision tree, and gradient boosting classifier, from the prediction algorithm, Logistic Regression has a very high accuracy compared to the others[7]. Ramnya also conducted research on obesity prediction using KNN, XGB, Logistic Regression, decision trees. From this study, the accuracy of Logistic regression and Decision Tree was obtained higher[8].

Lian et al, a risk prediction model has been developed to forecast the progression of newly treated COVID-19 patients to severe conditions. This model leverages age, clinical symptoms, and underlying diseases as key predictors. The predictive algorithm employed for this model is ROC regression, a method utilized for Receiver Operating Characteristic curve analysis, aimed at assessing the predictive accuracy of a model. [9].

Odeh conducted a prediction model for patients in Jordan, from these data, according to Oden ever, shortness of breath, and diabetes need to be treated more intensively, smoking is not identified as a risk factor for complications, and hydroxychloroquine treatment is not useful. Oden makes predictions using logistic regression [10]. Cao, G et al used univariate logistic analysis in predicting the frisk actors that cause COVID-19. From the prediction results, it was found that age, fever, diabetes, hypertension, CREA, BUN, CK, LDH and neutrophils are very potentially at risk of COVID-19[11].

Alizadehsani, et al conducted research on the risk factors that cause COVID-19 and result in death. This analysis employed statistical techniques to identify the most significant symptoms associated with COVID. Fever, dyspnea, weakness, shivering, elevated C-reactive protein levels, fatigue, dry cough, anorexia, anosmia, ageusia, dizziness, sweating, and age emerged as the most crucial symptoms. However, factors such as recent travel history, asthma, corticosteroid use, liver disease, rheumatological conditions, productive cough, eczema, conjunctivitis, tobacco consumption, and chest pain did not demonstrate any significant correlation with COVID [12].

Meister from collecting data from those infected with covid in Estonia predicts risk factors in comorbid and obese patients. Apart from comorbid and obese patients, it turns out that Age and male sex factors have a higher risk. The method used by the Meister uses statistics[13]. Willette, used the cohort method to predict infection and resistance to Covid at the age of 10–14 years ago. Willete used the LDA prediction method for two separate sets of analyses to

predict either: (1) COVID-19 diagnosis (negative vs. positive); or (2) COVID-19 infection severity (mild vs. severe)[14].

Gusev conducted Predicting the Development of the Clinical Evolution of Patients Diagnosed using the Naive Bayes, Decision Trees, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM) algorithms, and the Multilayer Perceptron (MLP). MLP has specific parameters, classifying the clinical evolution of the diagnosed patient[15]. Andrade proposes a comparative approach of machine learning (ML) algorithms, predicting clinical evolution in patients diagnosed with COVID-19. Prediction Using the Multilayer Perceptron algorithm gets better results than other ML algorithms [16].

## II. RESEARCH METHODS

### A. Methods

This research begins with a literature review from journals that discusses the prediction of comorbid diseases with death. Next, search the literature on SVM for prediction of disease. Continue to search for data containing Covid-19 patients and their history of comorbid illnesses. From this data pre-processing is carried out before training is carried out. The training was carried out 3 times, using polynomial, linear and RBC SVCs. From the best taken to predict which disease has the highest mortality rate. In addition, other training was also carried out using other methods such as linear logistics, KNN and XGBOOST to get the best F1 score for this case. The following figure shows the stages of the research conducted

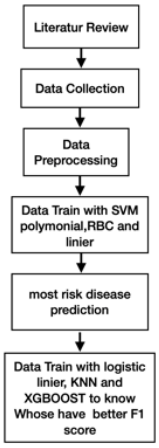


Fig 1 research stage

### B. Data collection

The dataset, sourced from the Mexican government, encompasses a vast amount of anonymized patient data, inclusive of pre-existing conditions. It comprises 21 distinct features and encompasses records for 1,048,576 individual patients. Within the Boolean features, a value of 1 signifies "yes," while 2 denotes "no," and values such as 97 and 99 indicate missing data.

Table 1 Header data

| Header | description |
|---|---|
| Sex | Male and Female |
| Age | of the patient |
| Clasification | covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive. |
| patient type | hospitalized or not hospitalized. |
| pneumonia | whether the patient already have air sacs inflammation or not |
| pregnancy | whether the patient is pregnant or not. |
| diabetes | whether the patient has diabetes or not |
| copd: | indicates whether the patient has Chronic obstructive pulmonary disease or not |
| asthma | whether the patient has asthma or not |
| inmsupr | whether the patient is immunosuppressed or not. |

| | |
|---|---|
| hypertension | whether the patient has hypertension or not |
| cardiovascular | whether the patient has heart or blood vessels related disease. |
| renal chronic | whether the patient has chronic renal disease or not. |
| other disease: | whether the patient has other disease or not. |
| obesity: | whether the patient is obese or not. |
| tobacco | whether the patient is a tobacco user |
| usmr | Indicates whether the patient treated medical units of the first, second or third level. |
| medical unit: | type of institution of the National Health System that provided the care |
| intubed: | whether the patient was connected to the ventilator. |
| icu | Indicates whether the patient had been admitted to an Intensive Care Unit. |

### C. *Data Preprocessing*

From the existing data, pre-processing is carried out, namely making the data ready for training. From these data there is no classification data that shows whether the patient died or not. There is data on the date of death. So a classification is made if the data is 99-99-9999 then the patient is alive, but if there is dead data it means the patient is dead. So the following commands are executed if off = 1 and on = 2.

df["DEATH"] = [2 if each=="9999-99-99" else 1 for each in df.DATE_DIED]

data for comorbid disease confirmed 1 and 2

Then only data on disease and death columns were selected. This data like classification, patient type, Pregnancy, other disease, medical unit, intubed, ICU and date_death. The table 2.2 provides reasons why the column was deleted.

**Table 2. the reason deleted data column**

| kolom | reason |
|---|---|
| classification | the classification shows the severity of covid-19, so it has nothing to do with the death rate |
| patient_type | hospitalized or not hospitalized, then this is not corelatiion with comorbid and death rate |
| Pregnancy | Pregnancy is not corelatiion with comorbid and death rate |
| other disease | Other disiee is not commorbid |
| medical unit | type of institution of the National Health System that provided the care. There is not correlation with commorbid |
| intubed | the patient was connected to the ventilator. Ther is not correlation with commorbid |
| ICU | ndicates whether the patient had been admitted to an Intensive Care Unit. There is correlation with commorbid |
| date_death | replaced with colomn death |

Column data is deleted so that only data columns will be used for the training process. Fig 2 shows the column data that will be used as training

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1025152 entries, 0 to 1048574
Data columns (total 12 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   USMER           1025152 non-null  int64
 1   PNEUMONIA       1025152 non-null  int64
 2   DIABETES        1025152 non-null  int64
 3   COPD            1025152 non-null  int64
 4   ASTHMA          1025152 non-null  int64
 5   INMSUPR         1025152 non-null  int64
 6   HIPERTENSION    1025152 non-null  int64
 7   CARDIOVASCULAR  1025152 non-null  int64
 8   OBESITY         1025152 non-null  int64
 9   RENAL_CHRONIC   1025152 non-null  int64
 10  TOBACCO         1025152 non-null  int64
 11  DEATH           1025152 non-null  int64
dtypes: int64(12)
```

Fig.2. Data column used

## III. RESULTS AND DISCUSSIONS

### A. *Train data with SVM method*

From the data that has been pre-processed, the training process uses SVM, to get an F1 score. The F1 score formula is as follows

$$\text{Accuracy} = = \frac{True\ positive + True\ negative}{2x(True\ Positive + TrueNegative)} \quad (1)$$

$$Precision = \frac{True\ Positive}{TruePositive + False\ Postive} \quad (2)$$


ISSN 2442-7888 (online) DOI 10.24167/sisforma.v11i2.13159

$$\text{Recall} = \frac{True\ Positive}{True\ Posite + False\ Negative} \quad (3)$$

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{Precision} + \frac{1}{Recall}\right) \quad (4)$$

The formula for SVM linear is

K(x, xi) = sum(x * xi) (5)

And The formula for SVM RBF is

K(x,xi) = exp(-gamma * sum((x – xi^2)) (6)

After training using SVM cluster rbf, to result F1 score

rbf Accuracy : 0.9291814408552853

rbf F1 Score : 0.96297127

After evaluating the imbalance data with the confusion matrix, it was observed that the dataset utilized was indeed imbalanced. Consequently, it becomes imperative to undergo retraining using balanced data. One technique to address data imbalance is Under sampling, wherein the dataset is balanced by preserving all instances from the minority class and reducing the size of the majority class

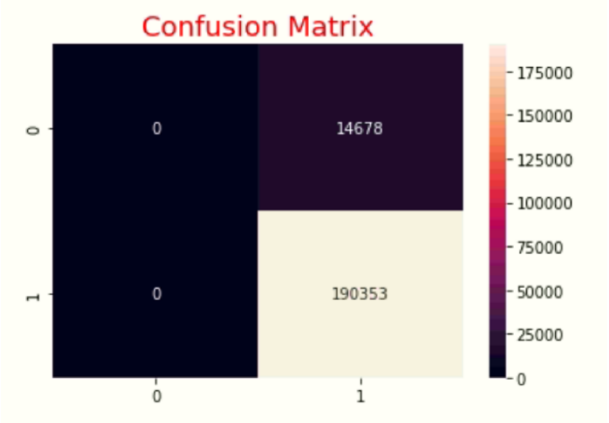


Fig 3 Confusion matrix Imbalance data

After the under sampling process is carried out, a re-training is carried out with the following results

rbf Accuracy : 0.8345379107274309

rbf F1 Score : [0.83558866]

and the matrix confusion show in fig 4

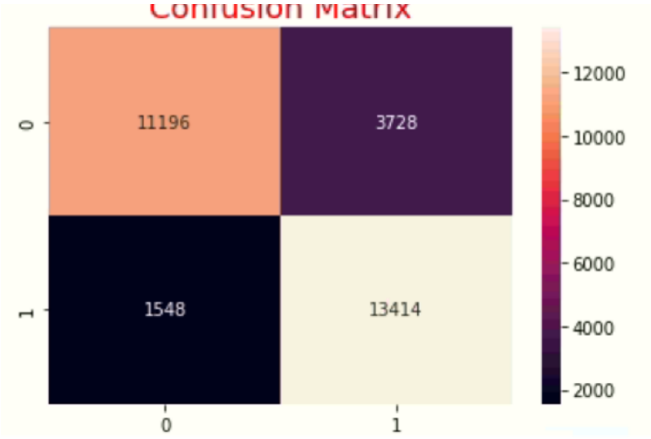


Fig 5. Confusion matrix after under sampling

TP -> 11196 Our prediction is 1 and real label is 1

TN -> 13414 Our prediction is 0 and real label is 0

FP -> 1548 Our prediction is 1 and real label is 0

FN -> 3728 Our prediction is 0 and real label is 1

Recall = TP / (FN + TP)

Recall = 11196 / (3728 + 11196) = 0.75

Recall is important to us now because it shows how many of the positive values we have guessed positive correctly.

Our model has a recall of 0.75—in other words, it correctly identifies 75% of all COVID-19s.

Training was carried out using the svm cluster (linear) to produce an F1 score as follows

Linier Accuracy : 0.8234624907983671

Linier F1 Score : [0.83565911]

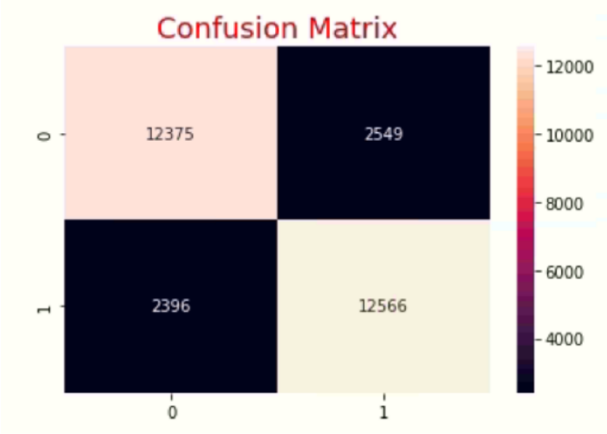


Fig 5. Confusion matrix SVM Linier

The recall is 0,82 or correctly identifies 82% of all data

3.2 Importance feature

Before getting the highest feature value from the training process. Ordinary statistical processes are carried out to get opportunities, namely deaths due to the disease / the number of sufferers of the disease

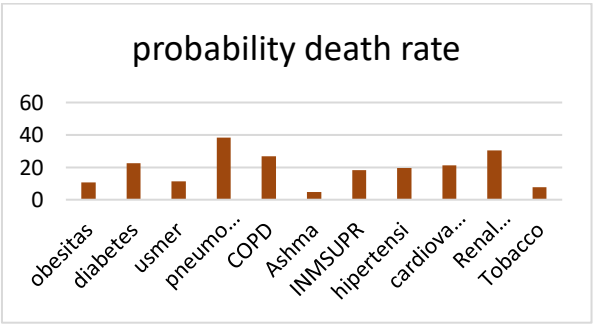


Fig 6. Probability death rate

From the above results it can be seen that pneumonia has the most influence on the mortality rate, followed by chronic renal disease and COPD. What's interesting is that the death rate of smokers from Covid is very low compared to other comorbid diseases.

In addition, a correlation process between features is also carried out to get the level of correlation with mortality. The following figure shows the level of this correlation

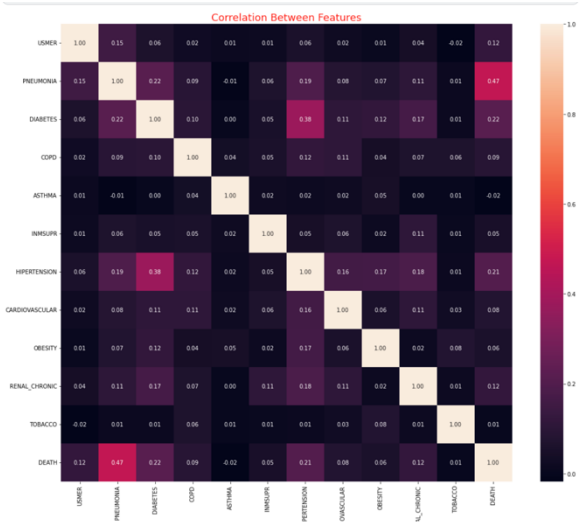


Fig 7. Correlation between feature

The highest value is pneumonia with 0.47 or 47% proximity to death. Followed by hypertension and diabetes.

If using SVM, the results obtained are as follows

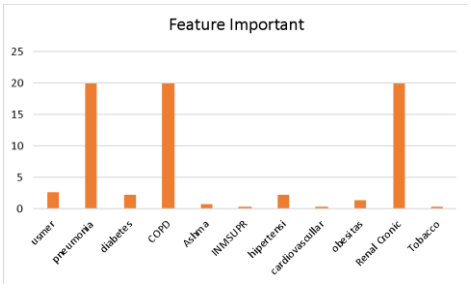


Fig 8. The Result of Importance Feature


ISSN 2442-7888 (online) DOI 10.24167/sisforma.v11i2.13159

From the results is pneumonia, COPD and renal chronic were the highest predictors. If you look at it, the highest value is still pneumonia, but the gap is not too big. This is because the prediction of Using SVM is lacking in terms of getting feature imports, in contrast to Xboost

3.3 Comparison with the other method

In order to see whether the SVM method is better than other methods, it is necessary to make predictions to get the F1 score. These methods are KNN, Logistic regression, Xboost, Decision tree, Random Forest and Multilayer Perceptron. Comparison using other prediction methods besides SVM can be shown in the following figure

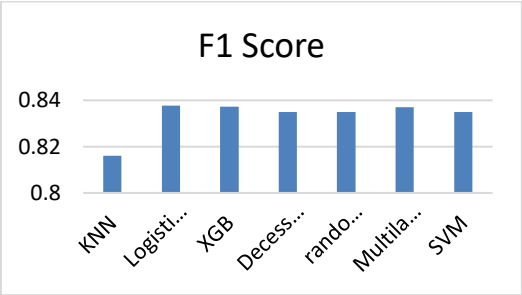


Fig 9. Comparison with other method

The highest F1 score is logistic regression and Multilayer Perceptron, while the lowest value is using KNN. Although the results obtained do not show a very large gap, there is only a slight difference and the average is 0.83 and its variance. 5.9369E-05.

While the confusion matrix for each method can be seen as follows

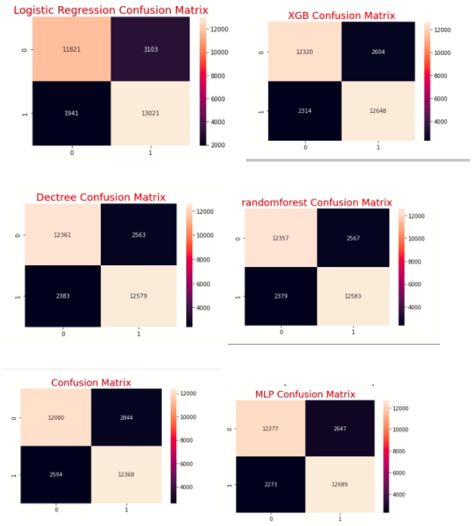


Fig 10. Matrix Confusion other method

## IV. CONCLUSION

Training method Using a linear kernel and rbf produces almost the same F1 value. It can be concluded that the data that was trained has an almost linear model. This is also seen when compared with other prediction methods such as. The F1 values of these methods are almost similar and the variance is very small. Although MLP and logistic regression get the greatest results from the others. From predictions using linear kernels, it can be seen that pneumonia, COPD and chronic renal disease have the highest mortality rates. This can also be seen from the value of the probability of death and the correlation with the greatest mortality, namely pneumonia, COPD and chronic kidney disease.

Future studies could include features of age and sex as well as the likelihood of comorbid admission to the ICU and death..

## REFERENCE

[1] Our World in Data [ cited 5 December 2022]. Available in https://ourworldindata.org/grapher/daily-covid-cases-deaths-7-day-ra

[2] Haq AD, Nugraha AP, Anggy F, Damayanti F, Wibisana IKGA,

Widhiani NPV, et al. " Faktor-Faktor Terkait Tingkat Keparahan Infeksi Coronavirus Disease 2019 (COVID-19): Sebuah Kajian Literatur". *Jurnal Ilmiah Mahasiswa Kedokteran Indonesia*. 2021;9(1).

[3] Kementiran Kesehatan Republik Indoensia. Simposium PAPDI. "Kesiapan Kemenkes Dalam Menghadapi Outbreak Novel Coronavirus" (2019-ncov). 2020; https://www.papdi.or.id/pdfs/817/dr Siti Nadia - Kemenkes RI.pd.

[4] Sanyaolu A, Okorie C, Marinkovic A, Patidar R, Younis K, Desai P, et al. "Comorbidity and its impact on patients with COVID-19". *Sn Comprehensive Clinical Medicine*.2020;2: 1-8.

[5] Aqmaria NW, Risanti ED, Mahmudah INN, Jatmiko SW. "Obesitas Sebagai Faktor Resiko Keparahan Pada COVID-19". *The 13th University Research Colloqium* 2021.

[6] Kemenkes. (2021). Situasi Terkini Perkembangan Novel Coronavirus (covid-19).

[7] Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I.,Habib, M. T. (2021). "A machine learning approach for obesity risk prediction." *Current Research in Behavioral Sciences* 2 https://doi.org/10.1016/j.crbeha.2021.100053.

[8] Ramya, A., & Rohini, K. (2021). "Comparative evaluation of machine learning classifiers with Obesity dataset". *Proceedings - 2021 International Conference on Computing Sciences, ICCS* 2021,38–41. https://doi.org/10.1109/ICCS54944.2021.00016

[9] Lian, Z., Li, Y., Wang, W., Ding, W., Niu, Z., Yang, X.,; Wu, C. (2021). "The Prediction Model of Risk Factors for COVID-19 Developing into Severe Illness Based on 1046 Patients with COVID-19". *Emergency Medicine International*, 2021, 1–11. https://doi.org/10.1155/2021/7711056.

[10] Odeh, M. M., al Qaissieh, R., Tarifi, A. A., Kilani, M. M., Tadros, R. E., al khashman, A. I., & Alzoubi, K. H. (2021). "A prediction model of risk factors for complications among SARS-CoV2 positive patients: Cases from Jordan". *Journal of Infection and Public Health*, 14(6), 689–695. https://doi.org/10.1016/j.jiph.2021.02.010.

[11] Cao, G., Li, P., Chen, Y., Fang, K., Chen, B., Wang, S., Feng, X., Wang, Z., Xiong, M., Zheng, R., Guo, M.,; Sun, Q. (2020). "A Risk Prediction Model for Evaluating the Disease Progression of COVID-19 Pneumonia". *Frontiers in Medicine*, 7. https://doi.org/10.3389/fmed.2020.556886

[12] Alizadehsani, R., Alizadeh Sani, Z., Behjati, M., Roshanzamir, Z., Hussain, S., Abedini, N., Hasanzadeh, F., Khosravi, A., Shoeibi, A., Roshanzamir, M., Moradnejad, P., Nahavandi, S., Khozeimeh, F., Zare, A., Panahiazar, M., Acharya, U. R., & Islam, S. M. S. (2021). "Risk factors prediction, clinical outcomes, and mortality in COVID-19 patients". *Journal of Medical Virology*,93(4),2307–2320. https://doi.org/10.1002/jmv.26699

[13] Meister, T., Pisarev, H., Kolde, R., Kalda, R., Suija, K., Milani, L., Karo-Astover, L., Piirsoo, M., & Uuskula, A. (2022). "Clinical characteristics and risk factors for COVID-19 infection and disease severity: A nationwide observational study in Estonia." *PLoS ONE*, *17*(6 June).

https://doi.org/10.1371/journal.pone.0270192

[14] Willette, A. A., Willette, S. A., Wang, Q., Pappas, C., Klinedinst, B. S., Le, S., Larsen, B., Pollpeter, A., Li, T., Mochel, J. P., Allenspach, K., Brenner, N., & Waterboer, T. (2022). "Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study". *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-07307-z

[15] Gusev, E., Sarapultsev, A., Solomatina, L., & Chereshnev, V. (2022). "Sars-Cov-2-Specific Immune Response and the Pathogenesis of COVID-19". In *International Journal of Molecular Sciences* (Vol. 23, Issue 3). MDPI. https://doi.org/10.3390/ijms23031716

[16] Andrade, E. C. de, Pinheiro, P. R., Barros, A. L. B. de P., Nunes, L. C., Pinheiro, L. I. C. C., Pinheiro, P. G. C. D., & Holanda Filho, R. (2022). Towards Machine Learning Algorithms in Predicting the Clinical Evolution of Patients Diagnosed with COVID-19. *Applied Sciences (Switzerland)*, *12*(18). https://doi.org/10.3390/app12188939