

DATA COLLECTION PROTOCOL FOR COVID-19 SEVERITY PREDICTION USING MACHINE LEARNING

Alex Fernandes de Souza

Instituto Tecnológico de Aeronáutica - ITA
Praça Marechal Eduardo Gomes, 50 - Vila das Acácias, São José dos Campos - SP
Universidade Federal de São Paulo - UNIFESP
Av. Cesare Mansueto Giulio Lattes, 1201, Eugênio de Melo - São José dos Campos - SP
alex.fernandes1989@live.com

Filipe Alves Neto Verri

Instituto Tecnológico de Aeronáutica - ITA
Praça Marechal Eduardo Gomes, 50 - Vila das Acácias, São José dos Campos - SP
verri@ita.br

RESUMO

Este trabalho investigou a eficiência do uso de um conjunto de dados para prognóstico de Covid-19 através de tarefas de Aprendizado de Máquina (AM) e propôs um método de coleta que gera um conjunto de dados arrumados para uso em AM. Foram utilizados dados do Hospital Sírio-Libanês, que demandou uma série de etapas de pré-processamentos para que os dados se tornassem adequados. Após estas etapas e aplicação em modelos de AM, observou-se um bom desempenho nos algoritmos KNN e SVM, com $AUC=0.81$. Logo, sabendo que o conjunto de dados arrumado é eficiente, este estudo indica um formato de coleta de dados que possibilita o uso rápido em modelos de AM, sem a necessidade de realização de diversas etapas de pré-processamento, potencializando o manejo no cuidado ao paciente e dos recursos de saúde.

PALAVRAS CHAVE. Coleta de dados. Covid-19. Aprendizado de Máquina.

Tópicos (Introdução. Pré-processamento de dados. Aprendizado de máquina para prognóstico Covid-19. Protocolo de coleta de dados para prognóstico de Covid-19. Conclusão. Metodologia.)

ABSTRACT

This work investigated the efficiency of using data for Covid-19 prognosis through Machine Learning (ML) models and proposed a collection method that generates a tidy dataset for use in ML. Data from Hospital Sírio-Libanês were used, which required an extensive series of pre-processing for the data to become adequate. After these steps and application in ML, a good performance was observed in the KNN and SVM algorithms, with $AUC=0.81$. Therefore, knowing that the dataset is tidy and efficient, a data collection format was determined that allows for faster use in ML models, without the need to carry out several pre-processing steps, enhancing the handling in the patient care and health resources.

KEYWORDS. Data collect. Covid-19. Machine Learning.

Paper topics (Introduction. Data pre-processing. Machine learning for Covid-19 prognosis. Data collection protocol for Covid-19 prognosis. Conclusion. Methodology.)

1. Introduction

The Coronavirus (SARS-Cov-2) pandemic has exposed a fragility of health systems on every continent [Caetano et al., 2020]. One of the biggest problems faced at the beginning of the pandemic concerns the collection of different types of data and their respective analysis [He et al., 2021].

Among the most important information, we highlight data from patients diagnosed with COVID-19, data from patients who progressed to severe cases of COVID-19, and data from patients who died [Uddin et al., 2019]. The understanding of such data aims to identify patterns that can contribute to the prediction of the health status of patients, enabling a more efficient management of hospital resources [Elaziz et al., 2020].

In the Brazilian context, the challenge in collecting COVID-19 data is in the standardization and organization of this information [Yan et al., 2020]. In Brazil, the FAPESP COVID-19 DataSharing/BR initiative groups datasets from five health institutions. Only one dataset of this repository has information indicating the outcome of each patient, making it possible to study the progression of the disease. In addition, it is important to highlight that this initiative follows a global trend of making data available on SARS-Cov-2, encouraging research in the area.

From the collection of structured data from patients with COVID-19, it is possible to use machine learning techniques to predict the health status of patients [Linssen et al., 2020]. Machine learning is a field of study that employs algorithms to identify patterns, make predictions, make decisions, and learn from data [Brnabic e Hess, 2021]. In the case of COVID-19, machine learning can be used for diagnosis [Zoabi et al., 2021], prognosis and identification of the population at greatest risk of contagion [Wang et al., 2021].

Along these lines, Linssen *et al* 2020 showed that it is possible to use the change in ten complete blood count parameters to predict the worsening of COVID-19 clinical conditions. For this, the authors used data from 982 adult patients, COVID-19 positive, and performed a prognostic score to predict during the first three days after care which patients recover without mechanical ventilation or deteriorate within a period of two weeks. The research obtained a AUC curve of 0.875 (IC 0.95).

The research by Yan *et al* [2020] showed COVID-19 prognosis through a model to identify predictive biomarkers of disease mortality, using data from 485 infected patients. Machine learning tools selected three biomarkers (lactic dehydrogenase - LDH, lymphocytes and C-reactive protein) capable of predicting mortality more than 10 days in advance, with an accuracy of 0.90. Predictive studies are important, as they can help in decision making about which patients need to be prioritized, aiming to reduce the mortality rate.

The objective of this article was to indicate the main challenges regarding data collection for the prognosis of COVID-19, indicating a protocol that can be adopted to reduce errors and enhance the use of this information in machine learning models. We used data from Hospital Sírio-Libanês, located in the city of São Paulo in Brazil, in this study. We select this database because it is the most complete among the open access databases available. Based on this, the available data allow, among other tasks, to point out variables related to the prognosis of COVID-19.

2. Methodology

As explained, in this study we used data provided by Hospital Sírio-Libanês available in the repository FAPESP COVID-19 DataSharing/BR¹. The data is available in 4 files in comma-separated values (csv) format. The first file contains information about the final state of health of each patient, the second file contains a dictionary about the information, the third file is composed

¹<https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/97>

of information obtained from the exams performed by each patient and the fourth file contains information about the characteristics of the patients, such as sex and age, for example.

For data processing, we use pandas², numpy³ and matplotlib⁴ libraries from Python, in addition to the Tidyverse⁵ library from R.

The first step of the data analysis process considers the grouping of this information into a single dataset. This grouping makes it possible to identify in a single dataset which types of exams were performed, exam dates, outcome dates and patient characteristics.

From joining, a dataset was obtained with the characteristics described in Table 1.

Tabela 1: Statistics for the Sírio-Libanês hospital dataset considering data on clinical examinations and patient outcome.

Characteristics	
Number of observations	2952999
Number of features	21
Categorical features	10
Numerical features	1
Textual features	10
Missing cells	0%
Duplicate rows	53426 (1.8%)

For the information presented in Table 1, we have that: Number of observations refers to the number of lines in the dataset; Number of features is the total number of columns; Categorical features are columns in categorical format; Numerical features are columns in numeric format; and Textual features are columns with data in string format.

The main information contained in this dataset and which were used are: ID_PACIENTE (patient identification), DE_EXAME (type of medical exam performed), DE_ANALITO (variables analyzed in the exams), DT_COLETA (date the exam was performed).

After viewing the characteristics of the respective dataset, the data was processed. All characters were converted to uppercase, the graphic accents were removed and double spacing was eliminated. So, we identified 167 types of analytes in the dataset. Analyte is a substance or component obtained in clinical examinations. The analytes were selected according to their occurrence in the dataset. For this study, 52 analytes with 18,000 entries or more were selected respectively. The other analytes were discarded as they are in a smaller number in the dataset.

To ease statistical analysis and the application of machine-learning methods, we convert textual resources into numeric and date format.

Another important aspect considers the selection and organization of the variables for diagnostic and prognostic tasks of COVID-19. According to experts, to indicate a diagnosis and/or prognosis of COVID-19 it is necessary to analyze the behavior of different analytes obtained from different tests.

²<https://pandas.pydata.org/>

³<https://numpy.org/>

⁴<https://matplotlib.org/>

⁵<https://www.tidyverse.org/>

Thus, a suitable dataset must have in each column a respective analyte originated from a certain medical examination. This organization follows the guidelines by Wickham [2014], who points out in his study the ideal format that a dataset should have in order to be used for statistical analysis and also in machine learning tasks.

It is worth noting that each selected analyte has a specific unit of measurement. Then, the analytes associated with their measurement units were transposed into columns, becoming the study variables, generating a structured final dataset.

Then, an analysis of each variable was performed, seeking to identify errors, such as whether two or more columns represent the same analyte, thus correcting any input errors and making the base more suitable for use in machine learning tasks.

3. Results and discussion

3.1. Data pre-processing

Based on the proposed organization of the data, a new dataset was obtained⁶. The characteristics of the respective dataset are described in Table 2:

Tabela 2: Statistics for the dataset obtained from the selection of 52 analytes from the Hospital Sírío-Libanês and in a tidy format.

Characteristics	
Number of observations	235061
Number of features	66
Categorical features	10
Numerical features	49
Text features	7
Missing cells	91,2%
Duplicate rows	0 (0%)

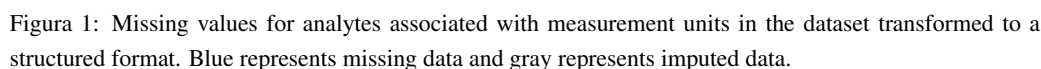
In Table 2 we have the Number of observations which are the rows and the Number of features are the columns. We have in this new dataset 10 Categorical features, 49 Numerical features and seven features in text format. Each column, with the exception of the patient identification, service identification and collection date columns, are formed by an association between analyte and measurement unit.

One of the problems of organizing the data for the dataset used is the generation of a huge amount of missing values. Regarding the analyzed analytes, we have the proportion of absent ones shown in Fig. 1. It is important to note that the large amount of missing values occurs because the dataset was tidied up, since it was not possible to use it efficiently in its original format.

Considering the selected analytes, it was proposed to combine some analytes in order to reduce the number of missing values. These analytes that were grouped into a single column can be seen in Table 3. This combination was made because two attributes of similar names had the same type of information. Soon, these two attributes were combined in a single and new attribute, described in 3.

From the union of the columns mentioned in Table 3, the final dataset became composed of 62 columns.

⁶<https://drive.google.com/file/d/1tAKOe7QoIbNEzFhHBDBhDTGYUZqXq3wU/view>



Considering these variables and the challenges reported so far, there is a serious problem of lack of standardization in the medical information entered in the Hospital Sírío-Libanês database. It was also noted that the filling of these two analytes does not follow a clear pattern, making it

Tabela 3: Association of columns with the same analytes identified in the structured dataset.

Unified analytics		New column
EOSINOFILOS (%) -	EOSINOFILOS (%) %	EOSINOFILOS %
BASOFILOS (%) -	BASOFILOS (%) (%)	BASOFILOS %
LINFOCITOS (%) -	LINFOCITOS (%) %	LINFOCITOS %
MONOCITOS (%) -	MONOCITOS (%) %	MONOCITOS %
NEUTROFILOS (%) -	NEUTROFILOS (%) %	NEUTROFILOS %

Tabela 4: String to Numeric Transformation

Analyte	String	New Value
CALCULO		
P/AFRODESCENDENTE CKD-EPI ML/MINUTO	SUPERIOR A 90	91
CALCULO		
P/AFRODESCENDENTE MDRD ML/MINUTO	SUPERIOR A 60	61
CALCULO		
P/NAO AFRODESCENDENTE CKDEPI ML/MINUTO	SUPERIOR A 90	91
CALCULO		
P/NAO AFRODESCENDENTE MDRD ML/MINUTO	SUPERIOR A 60	61
DIMEROS D. QUANT NG/ML	INFERIOR A 215	214

difficult to process the data to obtain useful information.

Thus, the new dataset for the Sírío-Libanês hospital, theoretically ideal for machine learning tasks, is composed of 60 variables, 3 of which are categorical (patient identification, service identification and collection date) and 57 numerical (analytes and measurement units). It is worth mentioning that each observation is formed by the patient identification (ID_PACIENTE), the service identification (ID_ATENDIMENTO), the date of collection of the respective analyte (DT_COLETA) and values obtained from the combinations of analytes and measurement units.

After organizing the data, another important step in the analysis was to verify the behavior of the values of each variable. This analysis can be seen in 2, which shows that all clinical attributes have a large number of outliers. These outliers were not removed, as they may be a specific characteristic of the disease or the patient.

In addition to the reported standardization errors, it is also possible to note the presence of variables with extreme outliers, raising doubts about the accuracy of the data collected and entered into the system. Thus, the dataset generated from these processing results in a large amount of missing data associated with outliers.

Considering what was observed and from a deep understanding of the dataset, a possible alternative for the solution of missing values is to obtain as many exams on the same date, so that

they also present several outliers, highlighting the importance of greater control in data entry and validation of outliers with experts.

In this way, by selecting a specific exam and transforming the datasets into a tidy format, it is possible to obtain a more complete and ideal dataset for application in machine learning tasks, in this case, in the prognosis of COVID-19.

3.2. Machine learning for Covid-19 prognosis

For the elaboration of machine learning for the prognosis of Covid-19, we selected the attributes obtained from the blood test called “HEMOGRAMA, sangue total”, as described above. These attributes have few missing values.

It is worth noting that several pieces of information were lost, reducing the dataset used. Patients who did not have a test collection date and/or outcome date were excluded from the base, as it was not possible to infer their hospital stay.

To determine the severity factor, an arbitrary method was chosen where patients who were hospitalized for more than 10 days or who died are considered as “SEVERE CASES”. Patients with a period of hospitalization shorter than this are considered as “NOT SEVERE CASES”. Regarding the determination of severity, this was decided by our research group at a meeting where physicians and infectious disease specialists helped to determine some parameters.

It was also noted that some patients had tests collected on more than one date, which raises questions about which information is more relevant in the medical context. In this study, the last date on which the exams were collected was considered. It is important to note that a health professional can request exams from different dates according to their needs, such as comparing exams on different dates, a situation that does not apply to this work. In addition, it is noteworthy that no selection of exams was made based on the dates of collection and date of service, in order to obtain the greatest amount of data possible.

After the exam was selected, the dataset was organized associating the analytes with their measurement units, the value being filled in by the column OF_RESULT. Thus, at the end, a dataset of 23 columns and 12,703 lines was obtained, with 20 columns referring to the selected analytes, and the others containing information about sex, age, severity. Each row corresponds to a single patient. Of this total, 9227 patients were classified as non-severe and 3476 patients were classified as severe. In this case, it is important to point out that this approach reduced the number of analytes from 63 to 21, however, there is no generation of missing values, making the dataset more suitable for use in AM.

When analyzing some of the selected analytes, a clear separation was observed between the plot classified as “SEVERE” and “NOT SEVERE”, as shown in Fig. 3.

It is also possible to observe in Fig. 3 that there is a difference between severe and non-severe patients for the selected analytes. This feature is important as it enhances the performance of machine learning models, which can classify patients more accurately.

For machine learning, the techniques *K Nearest Neighbor* (KNN), *Support Vector Machine* (SVM) and *Decision Tree* were used. These are classic techniques and extensively used in diagnostic and prognostic studies of diseases. In addition, they are easy to use techniques and do not need a lot of processing power. For the generation of models, we used technique of cross-validation with 10 folders (k-fold=10). Then, the data were normalized, as they have different scales.

We use two metrics to evaluate the models: AUC (*Area under de Curve*) and F1-Score. For each of the metrics, ten results were obtained, with the average representing the final result presented followed by the standard deviation. The AUC and F1-score values obtained can be verified in Table 5.

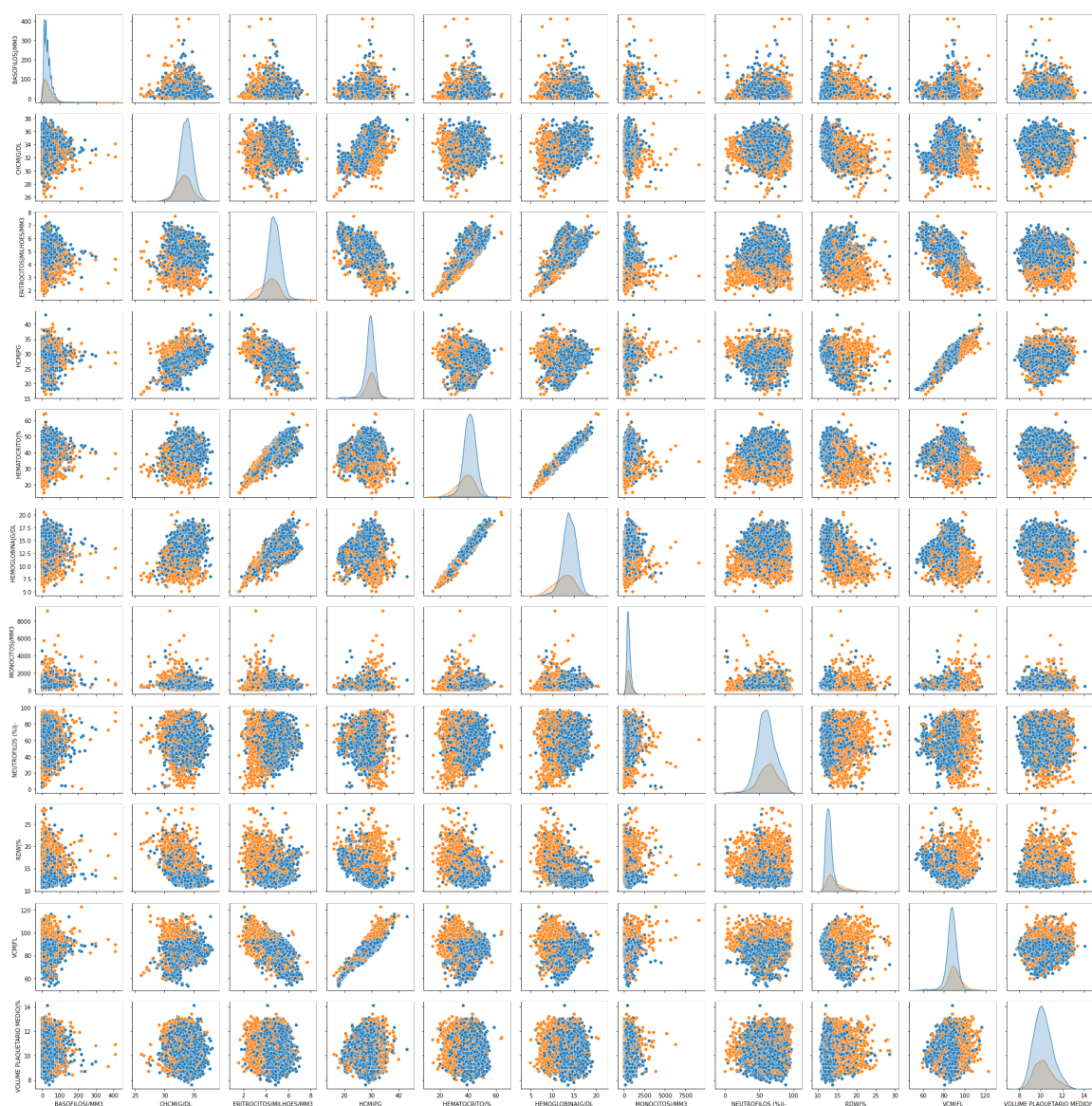


Figura 3: Comparative dispersion analysis of analytes for critically ill patients (orange color) and non-severe patients (blue color) obtained from Hospital Sírío-Libanês data.

Considering the information in Table 5, it can be said that KNN performed better when compared to the other two methods, both for AUC and F1 values. Bearing in mind that in these two metrics, the closer the output is to one, the better the performance of the algorithm. In this way, from the attributes collected from the blood test, it is possible to determine the severity of a patient with a certain degree of precision.

In addition, it is important to remember that in order to use these data, several data pre-processing steps were necessary, which made it possible to use them to point out a possible clinical prognosis using AM techniques, indicating the importance of having organized data for clinical analysis. Code for machine learning is available on GitHub⁸

⁸https://github.com/alexsouza1989/COVID-19_PROGNOSIS/blob/main/SBPO.ipynb

Tabela 5: Results for machine learning prognosis for data from Hospital S rio-Liban s.

Algoritmo	AUC (Mean e SD)	F1-Score (Mean e SD)
KNN	0.81 (0.01)	0.60 (0.02)
Decision Tree	0.72 (0.01)	0.50 (0.02)
SVM	0.81 (0.01)	0.54 (0.03)

3.3. Data collection protocol for covid-19 prognosis

In order to use the data in the prognosis of Covid-19 through machine learning, several types of transformations and modifications were necessary for the data to be adequate.

These preprocessing steps can be avoided by containing the data in an efficient way. Fig. 4 shows a data collection suggestion that generates a dataset that can be quickly applied in statistical analysis and machine learning.

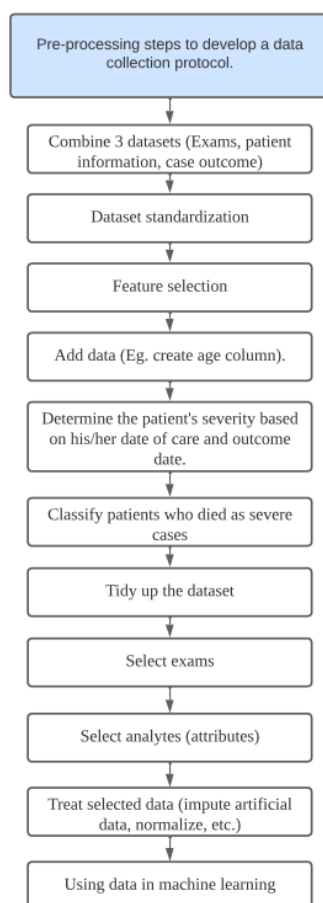


Figura 4: Data processing steps to develop an efficient data collection protocol.

Considering all the transformations that were necessary to use the data from the Hospital S rio-Liban s, as shown in the Fig. 4, this study sought to indicate the most appropriate way to collect this data in order to generate a tidy dataset, avoiding the need for several steps of pre-processing. The proposed collection method can be seen in the Fig. 5.

Therefore, knowing that the data set after being tidy proved to be efficient, a collection protocol was proposed that aims to generate a set of data in a tidy format, allowing statistical analyzes and the use in aiding the prognosis to be more efficient. fast and efficient. That way, in future and unknown diseases, using this protocol, it is possible to collect tidy data and use it for better management of patients and healthcare resources.

Referências

- Brnabic, A. e Hess, L. M. (2021). Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. *BMC medical informatics and decision making*, 21(1):1–19.
- Caetano, R., Silva, A. B., Guedes, A. C. C. M., Paiva, C. C. N. d., Ribeiro, G. d. R., Santos, D. L., e Silva, R. M. d. (2020). Challenges and opportunities for telehealth during the covid-19 pandemic: ideas on spaces and initiatives in the brazilian context. *Cadernos de Saúde Pública*, 36.
- Dorn, M., Grisci, B. I., Narloch, P. H., Feltes, B. C., Avila, E., Kahmann, A., e Alho, C. S. (2021). Comparison of machine learning techniques to handle imbalanced covid-19 cbc datasets. *PeerJ Computer Science*, 7:e670.
- Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., e Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of covid-19. *Plos one*, 15(6):e0235187.
- Hadley, W. (2014). Tidy data. *The Journal of Statistical Software*, 59(1).
- He, F., Page, J. H., Weinberg, K. R., e Mishra, A. (2021). Development and validation of simplified machine learning algorithms to predict prognosis of hospitalized covid-19 patients: a multi-center, retrospective study. *Journal of medical Internet research*.
- Linssen, J., Ermens, A., Berrevoets, M., Seghezzi, M., Previtali, G., Russcher, H., Verbon, A., Gillis, J., Riedl, J., de Jongh, E., et al. (2020). A novel haemocytometric covid-19 prognostic score developed and validated in an observational multicentre european hospital-based study. *Elife*, 9:e63195.
- Silveira, E. C. (2020). Prediction of covid-19 from hemogram results and age using machine learning. *Frontiers in Health Informatics*, 9(1):39.
- Uddin, S., Khan, A., Hossain, M. E., e Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16.
- Wang, Q., Xu, R., e Volkow, N. D. (2021). Increased risk of covid-19 infection and mortality in people with mental disorders: analysis from electronic health records in the united states. *World Psychiatry*, 20(1):124–130.
- Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., et al. (2020). An interpretable mortality prediction model for covid-19 patients. *Nature machine intelligence*, 2(5):283–288.
- Zoabi, Y., Deri-Rozov, S., e Shomron, N. (2021). Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj digital medicine*, 4(1):1–5.