Informatics in Medicine Unlocked 28 (2022) 100828

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Using bi-dimensional representations to understand patterns in COVID-19 blood exam data

Vitor P. Bezzan [a], Cleber D. Rocco [b,*]

[a] Instituto de Matemática, Estatistica e Computação Científica, Universidade Estadual de Campinas, Brazil
[b] Faculdade de Ciências Aplicadas, Universidade Estadual de Campinas, Brazil

ARTICLE INFO

Keywords:
COVID-19
Blood exam
Machine learning
Unsupervised learning
Applied AI
Dimensionality reduction

ABSTRACT

Blood tests play an essential role in everyday medicine and are used by doctors in several diagnostic procedures. Moreover, this data is multivariate – and often some diseases, such as COVID-19, could have different symptom manifestations and outcomes. This study proposes a method of extracting useful information from blood tests using UMAP technique - Uniform Manifold Approximation and Projection for Dimension Reduction combined with DBSCAN clustering and statistical approaches. The analysis performed here indicates several clusters of infection prevalence varying between 2%–37%, showing that our procedure is indeed capable of finding different patterns. A possible explanation is that COVID-19 is not just a respiratory infection but a systemic disease with critical hematological implications, primarily on white-cell fractions, as indicated by relevant statistical test *p*-values in the range of 0.03–0.1. The novel analysis procedure proposed could be adopted in other data-sets of different illnesses to help researchers to discover new patterns of data that could be used in various diseases and contexts.

1. Introduction

COVID-19 (Coronavirus Disease) caused by SARS-Cov-2 virus came under intense scrutiny worldwide throughout 2020–2021. Some countries are already seeing hospitalization rates dropping due to mass vaccination campaigns, social distancing, and lockdown measures. In a sad turn of events, Brazil is one of the biggest economies still witnessing death and hospitalization rates which were high at the beginning of 2021 (especially in the North region), according to Johns Hopkins Coronavirus Resource Center [1]. More than ever all the tools available to understand the infection scientifically in a precise and structured way.

In this study, we propose an exploratory and descriptive data analysis using a bi-dimensional representation generated by UMAP - Uniform Manifold Approximation and Projection for Dimension Reduction [2], followed by a DBSCAN clustering and posterior usage of statistical tests on the clusters obtained to reveal (non-causal) relationships between different parameters of blood-test data and their diagnostic counterparts. As with any study that aims only to describe data (and therefore do not make any predictions or offer any recommendations for action), we will not perform test and training samples for our dataset.

Data was obtained from the Albert Einstein Hospital in Sao Paulo, Brazil [3]. The data consist of patients' blood tests, providing information about whether or not a given patient had COVID-19 and if the patient needed special care or not (hospitalization in standard, semi, and intensive care units).

The main purpose of this article is to show that techniques previously considered to belong to the theoretical world of data science can be successfully applied to blood test data. Furthermore, our aim is to explore where techniques such as these can be used more frequently by medical science researchers. We believe that studies organized with a greater amount of data (both in number of people and in number of variables), greater representativeness of the samples and greater variability can benefit from what we discuss here and use the proposed procedures with little modification regarding to the original experiments. We also believe that in the future our methodology can be easily extended to data that do not come from blood samples.

This article is organized as follows: in Section 2, we examine the most up-to-date literature regarding Machine Learning applied to model blood test results and explain their results. In Section 3 we present the method used to perform the two data experiments, revealing their results in Section 4. In Sections 5 and 6, we outline

[*] Corresponding author.
E-mail addresses: vitor@bezzan.com (V.P. Bezzan), cdrocco@unicamp.br (C.D. Rocco).

https://doi.org/10.1016/j.imu.2021.100828
Received 21 September 2021; Received in revised form 15 December 2021; Accepted 19 December 2021
Available online 30 December 2021

the results and discuss the limitations and possible implications of this study.

2. Literature review

In this section, some studies from the literature are presented that inspired and laid the authors' foundations to create their analysis and perspective. Some interesting results were obtained without using machine learning (ML) and should be encouraged as a first-line open-access tool available to most researchers. In [4], it can be observed that statistically significant differences were found using two-way tables based on blood test data from a hospital in Italy, which is a quick and cheap solution to detect infections. A new study is [5], which is much more focused on hematological data and sheds light on significant statistical differences and possible risk factors associated with different patients. One specific meta-analysis, including the results of 35 other studies [6], indicated factors that contribute the most to non-severe patients to develop severe diseases.

Using blood tests with machine learning seems to have gained attraction since the beginning of the pandemic. Theoretical justification and groundwork for supervised ML techniques can be observed in several articles. In [7], attention is paid to possible combinations of models that could be used with results varying between the values of 0.6–0.9 area under the Receiver Operating Characteristic Curve (ROC) to detect infected patients. In [8], similar results were obtained using the same dataset we adopted in this study. Amalgamating the results of these articles and some others, it can be observed that [9] which uses ensembles and achieves 99.88% accuracy in predicting infections.

Other articles with similarity but not identical purposes are available. [10] uses several ML models in a dataset provided by the Sírio Libanês Hospital, in Brazil, to predict special-care probability and the number of days under special care, obtaining a value of 0.94 area under the ROC curve for the first target. In [11], we see a prime example of how a system could be implemented to detect COVID-19 in a given patient. This study also stands out as it uses a small sample and optimization techniques to find the most important variables for the problem.

As an example of unsupervised learning techniques, a study that can be mentioned is [12], which uses a model to predict infection and compares COVID-19 manifestations with other diseases using t-distributed stochastic neighbor embedding (similar to the purpose of UMAP), concluding that blood parameters of those affected with severe COVID-19 resemble more bacterial than viral infections, which was a very surprising result.

Therefore, considering what will be shown in this article, our main contributions will be in the use of a set of computational techniques to discover hidden patterns in blood test data, using a well-known cluster technique combined with a very recent dimensionality reduction technique that has gained adherents in several more applied areas of activity. This reduction in dimensionality, studied primarily as a purely mathematical exercise by the authors of the original article, has been pivotal in discovering previously unknown patterns in several areas and we believe that the methods to be presented here can be extrapolated with broad generality to other investigations in medicine and biological sciences.

The key difference between this study and the others mentioned above is the fact that we are not pursuing the creation of a fully supervised model. Instead, we aimed to test the "manifold hypothesis" on this data to check the existence of different groups where the manifestations of the disease could be different, providing researchers a whole new set of techniques to apply in other data sets in a similar context.

Table 1 shows some articles using UMAP as a basis for dimensionality reduction in several different contexts related to medicine and biology in general over the last few years, demonstrating the versatility and power of the technique we propose to use.

Table 1
Selected references for UMAP usage in medicine and biology.

Publication year	Reference	Application/Usage
2019	[13]	Single-cell visualization using UMAP
2019	[14]	Population patterns in genomic cohorts
2021	[15]	UMAP in population genetics
2021	[16]	Artifacts in microbiome data
2021	[17]	Transfer learning on molecular fingerprints
2021	[18]	Molecular dynamics simulations

Table 2
Parameter grid and intervals used in the clustering procedure.

Parameter	Interval	Description
neighbors	$[1, +\infty)$	Balance between local and global data representation
spread	$[0, +\infty]$	Minimum distance allowed between points in representation
eps	$[0.01, 0.5]$	Maximum neighborhood distance in DBSCAN

The use of clustering techniques is widespread in medical sciences in general. In a first class of articles, patient characteristics are used to unveil some hidden data structures present for diagnosing or understanding the disease's progression, such as [19,20]. Another class of studies tends to use more comprehensive statistical analysis with clustering to separate manifestations and possible patterns arising in a more specific group of patients, as in [21,22].

Although this study offers non-causal inference, it is relevant to point out sources such as [23] that mixes up causal inference and clustering in a medical setting; something we believe that should be further explored whether any other dataset allows us to do so.

3. Method

The procedure for our analysis primarily consists of two phases. In the first phase, we project high-dimensional laboratory exam data into a two-dimensional subspace using UMAP (tuning two hyperparameters), making the dataset more amenable to clustering techniques. In the second step, we cluster the data representation using DBSCAN [24] to find any patterns that may arise. The number of clusters obtained is a consequence of the hyperparameter tuning method used. Here, we used DBSCAN as a clustering alternative because the number of clusters is not specified upfront. By doing that, we assume more neutrality when analyzing the data structure.

The "overall quality" of fit for a specific combination of hyperparameters is measured without resorting to the target's current value, using the silhouette coefficient for a given arrangement [25]. We then compare different arrangements using this metric, selecting the one with the maximum value overall. Table 2 summarizes all hyperparameters used in the cluster tuning procedure. Obtained parameter values will be discussed in Section 4.

As know in data science, high-dimensional data has fewer degrees of freedom than one might initially assume, which is known as the "Manifold Hypothesis". [26] presents a complete description of the hypothesis and several demonstrations on the subject. In Appendix, we present a small application of UMAP to a dataset well known to the general public to demonstrate what the expected results are for the type of analysis we conducted in this study.

The hypothesis and the dimensional reduction provided by UMAP allows to analyze blood test data within a new perspective: different groups with different manifestations of the disease could be traced using this technique, as these groups will tend to cluster together in the low-dimensionality representation. Moreover, more significant factors could give us some clues about the disease and its progression.

Therefore, we propose two experiments. In the first, we analyze data from all patients in our dataset with blood tests measurements (red and white series) and then use the procedure outlined above. In the second one, we filter out our patient data keeping only those with confirmed COVID-19 and comparing the results using the targets for both situations. It is worth mentioning at this point that none of our analysis aims to be causal. The study was not conceived in this way, and the data are observational. For this purpose, we suggest using Causal Forests [27], which can deal with observational data and make a satisfactory causal inference whether the number of samples is high enough as the method needs several data splits.

Fig. 1 summarizes all the steps in both experiments. Silhouette coefficient is used to select the number of clusters for our experiments prior to statistical analysis.

3.1. On the dimensionality reduction technique

Dimensionality reduction techniques have already become commonplace in the data science community. Among the most usual techniques, we can mention statistical techniques already considered "classic", such as PCA [28] and ICA [29]. More modern developments include t-SNE [30] and the aforementioned UMAP, both classified as modern developments of manifold learning.

We expect data from blood tests to be highly non-linear, and their lower dimensional representations to be dominated by complex terms. PCA and ICA should therefore be disregarded as techniques to be used for dimensionality reduction in these cases as they are based on linear relationships/filters between variables. As for t-SNE, we discard its use based on scalability as the number of samples grows, as we want our methodology to be applied to arbitrary size datasets. Another major disadvantage of this method is that it does not allow the transformer object to be reused in other datasets different from the initial set.

UMAP is entirely based on Riemannian geometrical assumptions (uniform distribution, locally constant metric tensor and local connectivity). It models the data using a fuzzy topological structure. The math behind the method is fairly advanced and will not be discussed in this article. We suggest the reader consult [2] for more details.

3.2. On the clustering technique

Various clustering techniques have been widespread in the data science field in recent decades. One of the first examples that was widely discussed more since the 1960s. We can cite as examples most used by the community the k-means [31], fuzzy c-means [32], OPTICS [33] and DBSCAN [24] techniques.

The main differences among these techniques are on scalability (both in number of clusters and in number of samples), the capacity to detect clusters of different formats and detection of outliers built into the procedure.

We can consider that our data will not present trivial geometry after the dimensionality reduction process, thus invalidating the use of techniques such as k-means, which tend to separate clusters more uniformly and with more "circular" geometry.

Likewise, we can expect that our techniques can be applied to new datasets in a scalable way and with a high degree of reproducibility. This invalidates the use of the fuzzy c-means technique given the high need for components that introduce unwanted "degrees of freedom" to the method. Furthermore, we want an element to belong to a single cluster.

Thereofore, DBSCAN was the selected technique due to its scalability, detection of clusters in geometries that are not necessarily circular and the ability to filter outliers in different contexts. Furthermore, the main programming languages already have the algorithm included in their packages, which helps to implement and disseminate of the concepts presented in this article. The technique assumes that regions of clusters have a higher density of points, separated by lower density regions. The minimum density requirements are codified in the parameters shown in Table 2, when considering Euclidean distances.

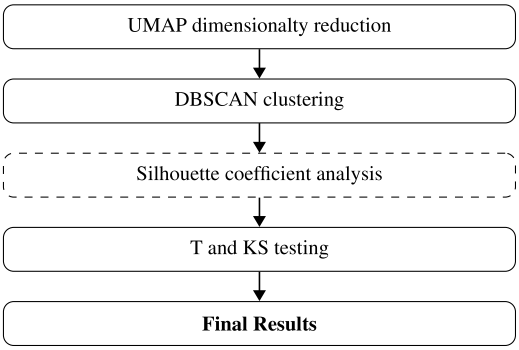

Fig. 1. Steps synthesizing our method for both experiments proposed.

4. Computational results

4.1. Data

The data contains anonymous information about 598 patients admitted to the Albert Einstein Hospital during the COVID-19 pandemic. Eighty one patients tested positive for infection (13%) and 128 patients needed special care treatment (21%, not only related to COVID-19). There are available parameters related to red and white cell counts for each patient, all of them normalized by the mean and standard deviation (z-scores). Table 3 summarizes all the variables used for the study.

To further expand on the data, Fig. 2 presents white cell distribution for all 598 patients (blue dots show negative infection whereas orange dots are positive). No univariate pattern was observed emerging in the data, which leads us to using a multivariate technique.

As mentioned above, two data experiments were performed. The first experiment consists of all 598 patients and tries to understand whether there are groups with high prevalence (greater than the average of the dataset) and to point out the main characteristics of these groups. In the second experiment, the focus is primarily on the confirmed COVID-19 diagnostic, aiming to discover any groups with more prominent special care needs than the whole dataset.

4.2. Experiment I: All patients, focusing on the confirmed COVID-19 results

In this first analysis, after performing the aforementioned dimensionality reduction with UMAP and the clustering of the resulting 2-dimensional space variables, we obtained a value of 0.12 for the silhouette coefficient (the clusters obtained are very packed together). Overall, 7 clusters were obtained with COVID-19 prevalence in the range of $3-35\%$. Moreover, 29 patients did not meet any of the DBSCAN similarity criteria and were not assigned any cluster, thus they were removed from the analysis (see Fig. 3). A close inspection of Tables 4 and 5 reveals that most extreme values reside on the first two clusters for white-cell counts. This fact could be interpreted in a two ways: patients could have comorbidities and be more susceptible to being infected by COVID-19, thus having greater white-cell counts, as pointed out by [34]. On the other hand, COVID-19 could be responsible for the values themselves. One observation is about the number of platelets, which is very low, much in line with discoveries shown in [35–38].

No extreme values were found in red cell samples for high COVID-19 prevalence clusters, but the close observation of the tables regarding the prevalence and the number of people in each cluster may help to "name" each cluster, a procedure that is made when clusters are applied in several contexts. For example, cluster 1 could be named "Non-symptomatic patients", although more data is needed to make such an affirmation.

Table 3
Variables used for study.

Fraction	Components
Red Cell	Hematocrit, Hemoglobin, Red Cells, MCHC, MCH, MCV, RDW
White Cell	Platelets, MPV, Lymphocytes, Leukocytes, Basophils, Eosinophils, Monocytes

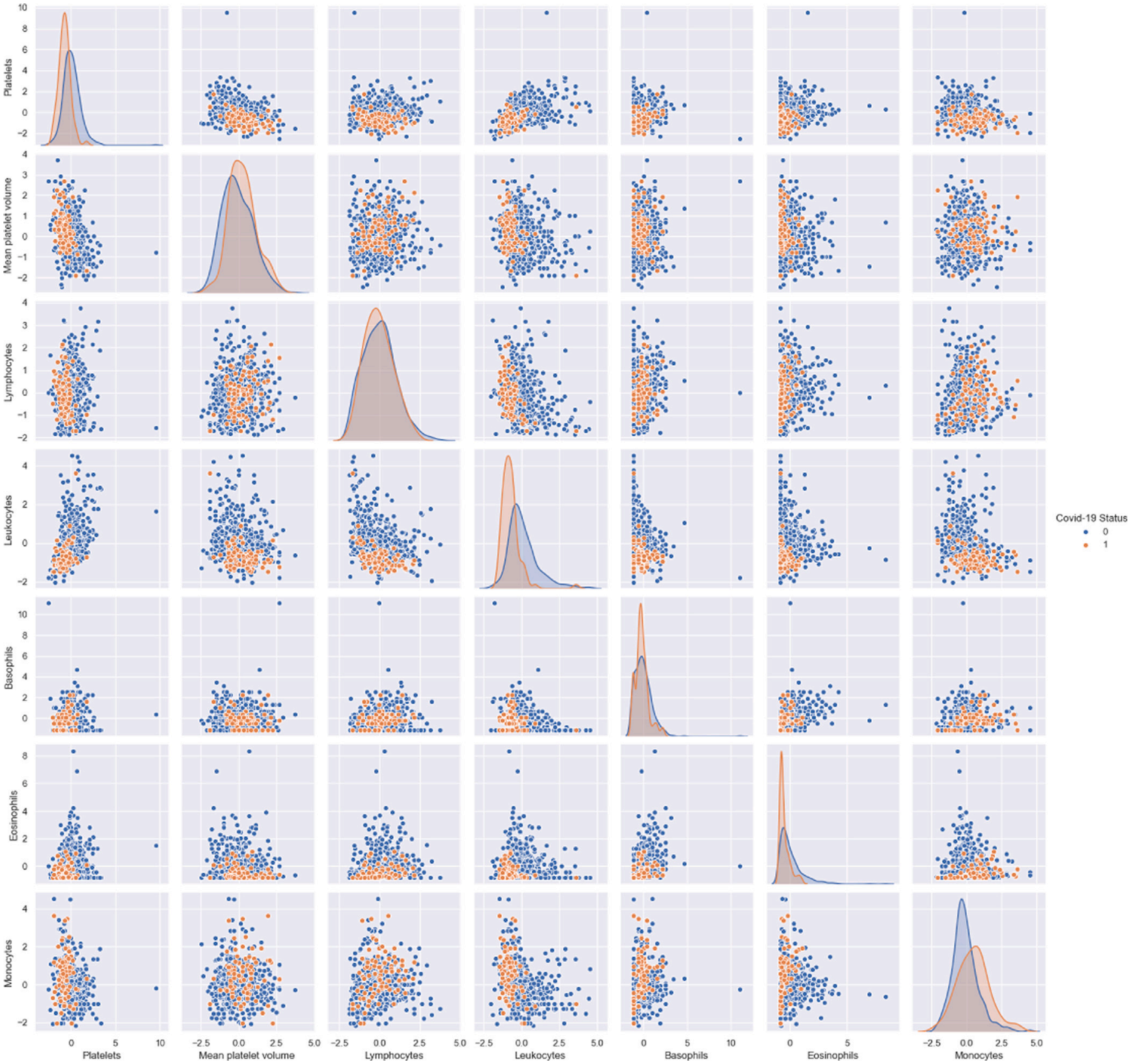

Fig. 2. White cell blood count distributions, normalized for 598 patients.

Table 4
Means for variables in clusters found in experiment I (Red components - extreme values in bold).

	Hematocrit	Hemoglobin	Red Cells	MCHC	MCH	MCV	RDW	**Covid-19** (%)	Patients
Cluster									
2	0.449555	0.360825	0.403754	−0.219273	−0.129423	−0.025629	−0.192997	34.6	26
4	0.331591	0.353596	0.177950	0.187976	0.259933	0.197007	−0.155573	23.1	39
6	**0.890685**	**0.947817**	**0.910157**	**0.404758**	−0.046087	−0.249906	**−0.234152**	19.4	31
0	−0.123704	−0.160615	−0.269449	−0.167212	0.249553	0.363449	0.330257	17.9	145
5	**−0.566416**	**−0.606294**	−0.369784	**−0.313489**	**−0.400994**	**−0.312915**	**0.680545**	16.0	25
1	−0.015429	0.021216	0.056257	0.141979	−0.079685	−0.156664	−0.216660	7.4	269
3	−0.285210	−0.324563	**−0.527503**	−0.212740	**0.398023**	**0.565605**	−0.133359	2.9	34

Table 5
Means for variables in clusters found in experiment I (White components - extreme values in bold).

	Platelets	MPV	Lymphocytes	Leukocytes	Basophils	Eosinophils	Monocytes	**Covid-19** (%)	Patients
Cluster									
2	**−0.566694**	0.092664	**0.365603**	**−0.408745**	**0.880585**	−0.018652	**0.227241**	34.6	26
4	−0.327375	−0.262615	−0.127550	−0.291689	−0.231599	**−0.303903**	0.100975	23.1	39
6	0.244400	−0.376571	0.014347	0.068407	0.130960	**0.072528**	−0.105025	19.4	31
0	−0.129383	**0.287677**	−0.154026	−0.031201	0.037454	0.068019	0.019385	17.9	145
5	−0.108903	−0.016250	**−0.803715**	**0.207712**	**−0.419260**	−0.301180	−0.465017	16.0	25
1	0.115441	−0.031031	0.160436	0.065219	−0.026183	0.058192	−0.004671	7.4	269
3	**0.555883**	**−0.477694**	−0.118372	0.242435	−0.133926	0.023392	**−0.575570**	2.9	34

Table 6
Means for variables and respective t and KS tests for clusters found in experiment II (Red components - no significant *p*-values in bold).

	Hematocrit	Hemoglobin	Red Cells	MCHC	MCH	MCV	RDW	Special Care (%)	Patients
Mean - Cluster 1	0.192373	0.228284	0.124672	0.187246	0.152039	0.078920	−0.227019	7.0	14
Mean - Cluster 2	0.276826	0.302162	0.261730	0.166864	0.034623	−0.037691	−0.194673	61.0	67
t-test	0.638796	0.619572	0.701361	0.466539	0.285301	0.295333	0.562766	–	–
KS-test	0.440488	0.675420	0.788581	0.458707	0.284728	0.348343	0.863925	–	–

Table 7
Means for variables and respective t and KS tests for clusters found in experiment II (White components - significant *p*-values in bold).

	Platelets	MPV	Lymphocytes	Leukocytes	Basophils	Eosinophils	Monocytes	Special Care (%)	Patients
Mean - Cluster 1	−0.445631	0.331228	0.063713	−0.537869	0.016237	−0.305755	0.858424	7.0	14
Mean - Cluster 2	−0.734901	0.263530	−0.049911	−0.741464	−0.205530	−0.516632	0.406545	61.0	67
t-test	**0.061341**	0.399595	0.331979	0.150230	0.156617	**0.056762**	**0.088217**	–	–
KS-test	**0.034455**	0.689187	0.272100	0.564482	0.284728	**0.030776**	0.105875	–	–

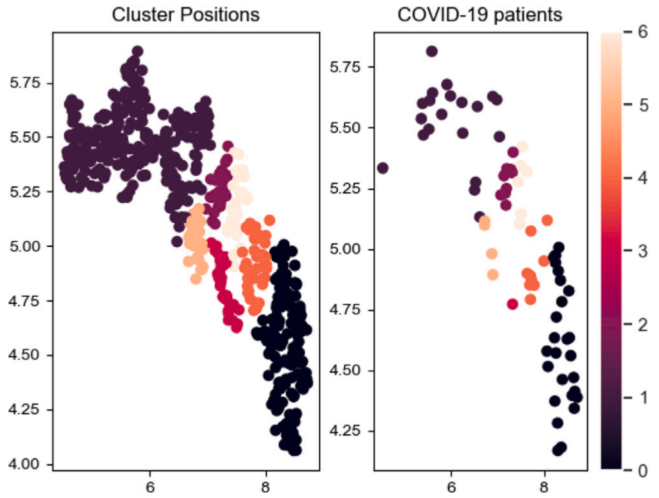

Fig. 3. DBSCAN cluster results for Experiment I. On the right, all COVID-19 patients with clusters associated.

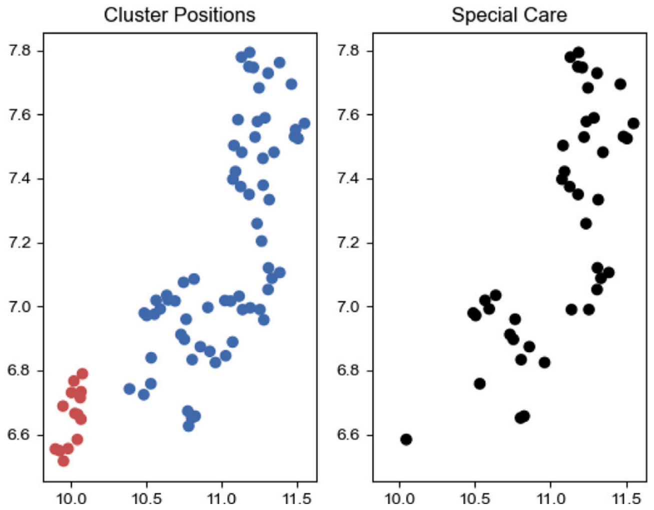

Fig. 4. DBSCAN cluster results for Experiment II. On the right, all special-care patients.

4.3. Experiment II: COVID-19 patients, focus on special care

In this analysis, we obtained a value of 0.40 for the silhouette coefficient (the clusters obtained seem very separated, as shown in Fig. 4). Overall, two clusters were obtained, with COVID-19 prevalence in the range of 7–61%. No patients without clusters were obtained in this analysis.

The number of clusters obtained allows us to go one step further in the analysis. We conducted two-sample one-sided (lower) t- and KS-statistical tests. Tables 6 and 7 show the *p*-values associated with one of these tests in every parameter. The result is very similar to Experiment I. Red cell components do not display any statistical differences between the two groups, however white cell components show statistical differences. Once more, platelets appear as a significant factor, once again indicating a relationship between coagulation factors, COVID-19 and a possible patient prognostic.

5. Limitations and possible extensions

There are two limitations to this study. The first one is data: the variables to be analyzed ("wider": more columns) and the number of patients ("higher": more rows) could lead to a substantial improvement in the outcomes achieved so far, allowing us to separate the clusters better.

More variables for each patient also mean that different representations could be obtained. In medical terms, more complex relationships could be extracted. Restricting ourselves only to blood exams, C-reactive protein, AST, ALT, GGT, and LDH could be excellent additions to the analysis. Other data sources could be leveraged: social and economic data could help to trace relationships between infection severity and social strata. Genetic markers could help to understand whether some populations are more susceptible to infections than others. Medical imaging data could help to associate blood parameters with physiological changes in organs and tissues, and so on.

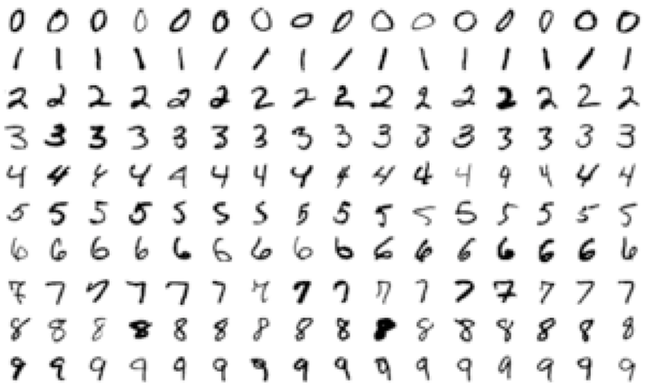

Fig. 5. MNIST data [39] examples. Each example is a 28 × 28 pixel image.

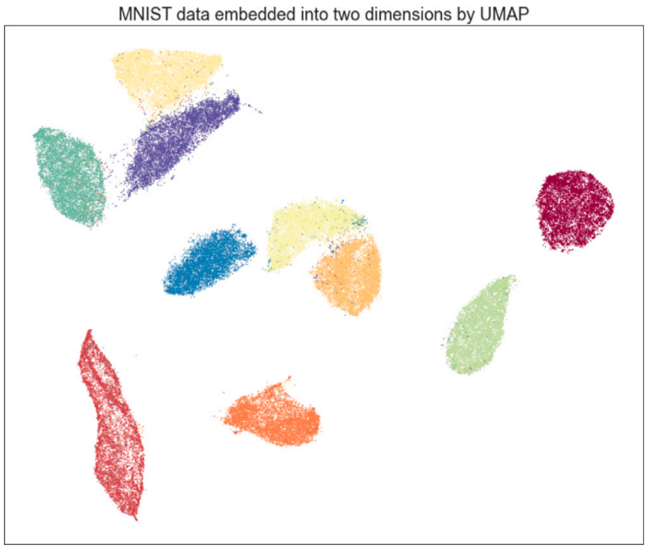

Fig. 6. UMAP dimensionality reduction results on MNIST data. Each one of the colors represents a different number (the coordinates were omitted). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The second point is the non-causality of analysis. None of this study's conclusions are causal for two reasons: the data is observational, and the number of patients and parameters is not large. This reveals an excellent opportunity for researchers because the procedure applied here could be used to control the experiment data without any modifications. There are some studies in the literature that combine cluster analysis with causal inference but they are still very sparse [23]. Statistically significant samples and more parameters could help to create groups of patients where a treatment (or protective measures) could be tailored for each group. Other diseases could also benefit from the same approach presented here.

Considering the nature of this research, other epidemics (e.g. Dengue fever, Zika Virus, Ebola) could be an excellent investigation opportunity, as the primary source of data used here is inexpensive and could be collected even in developing and emerging countries.

6. Final remarks

Using only data science methods, we were able to demonstrate that different prevalence subgroups exist, and that these groups have different medical interpretations that make sense. This study opens a window of opportunity for those with access to individual and more granular blood data for patients, paving the way for a more comprehensive analysis with more factors to be analyzed. Moreover, we aim to help to demonstrate that COVID-19 is not only "a simple flu" with only respiratory effects but a more complex disease with several potential implications and outcomes, particularly hematological as described by relevant statistical testing.

Special implications in platelets (which control coagulation), eosinophils and monocytes (related to infection control and adaptive immunity) further disclose that COVID-19 is a multi-systemic, multi-implication disease that must be analyzed from a multi-disciplinary perspective and the clusters found can be the first indication that several approaches must be taken by medical staff, policymakers and governments. In the future, we can use similar techniques with augmented data to address different problems related to COVID-19 such as vaccine distribution, field hospital construction, disease spread analysis and other issues. The approach presented here can be also easily adapted to other diseases.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the anonymous referees for their useful comments. This project partially benefited from the CNPq, Brazil grant (process number: 400868/2016-4).

Appendix. UMAP example on MNIST data

A good way to visualize the dimensional reduction performed by UMAP is by comparing Figs. 5 and 6. Fig. 5 shows elements of the so-called MNIST dataset [39], which consists of 28 × 28 pixel images (784 dimensions) of thousands of handwritten digits. In Fig. 6, after the UMAP algorithm, we can see that similar points tend to cluster closely, and non-similar digits tend to be more distant. The overall distance is controlled by the parameters' neighbors and spread in Table 2. We selected here a specific two-dimensional representation of our data for viewing purposes, knowing that high-dimensional representations could be necessary to deal with very-high dimensionality data/ or complex behaviors.

References

[1] JHCRC. Johns hopkins coronavirus resource center. 2021, Accessed: 17 Feb 2021.

[2] McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. 2020, arXiv:1802.03426.

[3] HIAE. Diagnosis of COVID-19 and its clinical spectrum. 2020, https://www.kaggle.com/einsteindata4u/covid19/ [Accessed 17 Feb 2021].

[4] Ferrari D, Motta A, et al. Routine blood tests as a potential diagnostic tool for COVID-19. Clin Chem Lab Med 2020.

[5] Liao D, Zhou F, et al. Haematological characteristics and risk factors in the classification and prognosis evaluation of COVID-19: a retrospective cohort study. Lancet 2020;7.

[6] Bao J, Li C, et al. Comparative analysis of laboratory indexes of severe and non-severe patients infected with COVID-19. Clin Chim Acta 2020;509:180–94.

[7] Brinati D, Campagner A, et al. Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. J Med Syst 2020;(135).

[8] de Moraes Batista AF, ao Luiz Miraglia J, et al. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. MedRxiv 2020. http://dx.doi.org/10.1101/2020.04.04.20052092, URL https://www.medrxiv.org/content/early/2020/04/14/2020.04.04.20052092.

[9] AlJame M, Ahmad I, et al. Ensemble learning model for diagnosing COVID-19 from routine blood tests. Inform Med Unlocked 2020;21.

[10] Bezzan V, Rocco CD. Predicting special care during the COVID-19 pandemic: A machine learning approach. Health Inf Sci Syst 2021;(34).

[11] de Freitas Barbosa VA, Gomes JC, et al. Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests. Res Biomed Eng 2021.

[12] Kukar M, Gunčar G, et al. COVID-19 diagnosis by routine blood tests using machine learning. 2020, arXiv:2006.03476.

[13] Becht E, McInnes L, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnol 2019;37:38–44.

[14] Diaz-Papkovich A, Anderson-Trocmé L, et al. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLOS Genet 2019.

[15] Diaz-Papkovich A, Anderson-Trocmé L, et al. A review of UMAP in population genetics. J Human Genet 2021;66:85–91.

[16] Armstrong G, Martino C, et al. Uniform manifold approximation and projection (UMAP) reveals composite patterns and resolves visualization artifacts in microbiome data. MSystems 2021;6.

[17] Lovric M, Duricic T, et al. Should we embed in chemistry - a comparison of unsupervised transfer learning with PCA, UMAP, and VAE on molecular fingerprints. Pharmaceuticals 2021;14.

[18] Trozzi F, Wang X, et al. UMAP as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: A comparison study. J Phys Chem B 2021;125:5022–34.

[19] McLachlan G. Cluster analysis and related techniques in medical research. Stat Methods Med Res 1992;1.

[20] Skerman HM, Yates PM, et al. Multivariate methods to identify cancer-related symptom clusters. Res Nurs Health 2009;32:345–60.

[21] Paul R, Sayed A. Clustering medical data to predict the likelihood of diseases. In: 2010 fifth international conference on digital information management. 2010.

[22] Alashwal H, Halaby ME, et al. The application of unsupervised clustering methods to Alzheimer's disease. Front Comput Neurosci 2019;13.

[23] Oltean H, Gagnier JJ. Use of clustering analysis in randomized controlled trials in orthopaedic surgery. BMC Med Res Methodol 2015;15.

[24] Schubert E, Sander J, et al. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. ACM Trans Database Syst 2017;19.

[25] J.Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.

[26] Fefferman C, Mitter S, et al. Testing the manifold hypothesis. J Amer Math Soc 2016;29(4):983–1049.

[27] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Amer Statist Assoc 2018;113:1228–42.

[28] Tipping ME, Bishop CM. Mixtures of probabilistic principal component analysers. Neural Comput 2006;11(2):443–82.

[29] Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. Neural Netw 2006;13(4):401–30.

[30] van der Maaten L, Hinton G. Visualizing high-dimensional data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[31] MacQueen J. Some methods for classification and analysis of multivariate observations.In: Berkeley symposium on mathematical statistics and probability. 1967. p. 281–97.

[32] Dunn J. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybern 1973;3.

[33] Ankerst M, Breunig MM, et al. OPTICS: ordering points to identify the clustering structure. ACM Sigmod Rec 1999;28(2):49–60.

[34] de Souza WM, Buss LF, et al. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat Hum Behav 2020;4:856–65.

[35] G.D. W, J.L. M. The impact of COVID-19 disease on platelets and coagulation. Pathobiology 2021;88:15–27.

[36] Güçlü E, Kocayiğit H, et al. Effect of COVID-19 on platelet count and its indices. Revista Da AssociaÇÃo Médica Brasileira 2020;66.

[37] Battinelli EM. COVID-19 concerns aggregate around platelets. Blood 2020;136:1221–3.

[38] Mei H, Luo L, et al. Thrombocytopenia and thrombosis in hospitalized patients with COVID-19. J Hematol Oncol 2020;13.

[39] LeCun Y, Cortes C. MNIST handwritten digit database. 2010, URL http://yann.lecun.com/exdb/mnist/.