IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

An in-depth analysis of COVID-19 symptoms considering the co-occurrence of symptoms using clustering algorithms

D. BENITO[1], J. RUFINO[1], J.M. RAMIREZ[1], A. FERNÁNDEZ ANTA[1] and J. AGUILAR (Senior Member, IEEE)[1,2,3]

[1]IMDEA Networks Institute, Madrid, Spain
[2]CEMISID, Universidad de Los Andes, Merida, Venezuela
[3]GIDITIC, Universidad EAFIT, Medellín, Colombia

Corresponding author: J. Aguilar (e-mail: jose.aguilar@imdea.org; aguilar@ula.ve).

ABSTRACT A comprehensive analysis of the COVID-19 pandemic is necessary to prepare for future healthcare challenges. In this regard, the large number of datasets collected during the pandemic has allowed various studies on disease behavior and characteristics. For example, collected datasets can be used to extract knowledge about the symptomatic behavior of the disease. In this work, we are interested in analyzing the relationships between the different symptoms of the disease, considering various dimensions, such as countries, variants of COVID-19, and age groups. To this end, we consider the co-occurrence of symptoms as a fundamental element. More precisely, we implemented clustering techniques to discover symptomatic patterns across the various dimensions. For instance, in analyzing the dominant patterns, we observe that symptom *congestion or runny nose* almost always appears with the symptom *muscle pain* across many dimensions. Hence, the information on symptom patterns can be helpful in decision-making processes to detect and combat COVID-19 and similar diseases.

INDEX TERMS COVID-19 symptoms, Symptomatic patterns, machine learning, clustering algorithms.

I. INTRODUCTION

THE COVID-19 pandemic has had a huge impact on our society. Studies around COVID-19 have covered all disciplines, in order to understand its effects. Although the World Health Organization (WHO) has declared the end of COVID-19 as a global health emergency, studies have not stopped to continue generating knowledge that allows us to prepare for any similar event in the future. An important aspect to study is the behavior of symptoms throughout the pandemic. Being able to understand the symptoms that occurred in each country, age range, and in each variant, is a research topic.

Thus, the relationships among the symptoms associated with COVID-19 are valuable information for understanding the illness. Finding representative symptoms of COVID-19 helps support diagnosis, and pinpoint medications to treat patients efficiently. Therefore, effective diagnostic and therapeutic strategies for patients with COVID-19 rely on an analysis of symptoms and their combination. However, collecting symptom duration and analyzing their co-occurrence has not been common, and most studies typically only report the onset of COVID-19 symptoms in patients in the hospital setting.

Therefore, this information may not reflect the symptomatic behavior of patients (duration, co-occurrence, etc.) in the general population. For example, it may not capture information about non-hospitalized individuals who may experience less severe episodes of the disease. To mitigate this bias, some approaches have been employed to collect self-reported COVID-19 symptoms from the general population. These include surveys, online social networks (e.g., Twitter), and symptom checkers, among others.

A. RELATED WORKS

A wide range of AI applications have addressed the diagnosis and treatment challenges posed by COVID-19 [1]–[4]. There are numerous algorithms or knowledge models to predict, diagnose, detect, or define therapies for COVID-19 [1], [3]–[5]. Additionally, most of the current research identifies the key factors (symptoms, signals, etc.) in the diagnosis (frequency, etc.), but often without considering multidimensional aspects such as age, region, variant [4], [5].

The work of Wang et al. [5] studied the psychopathol-

ogy and psychomotor symptoms throughout the COVID-19 outbreak. The study was conducted in China during the outbreak and after the peak stages, with 2540 participants from February 6 to 16, 2020, and 2543 participants from April 25 to May 5, 2020. They considered psychopathology symptoms such as anxiety, depression, irritability, and loss of energy, and psychomotor symptoms as impaired motor skills, restlessness, and inability to relax. They concluded that symptoms of irritability and loss of energy played an important role after the peak of the pandemic. The work of Hu et al. [1] proposed a diagnosis and treatment model for COVID-19 based on complex networks and machine learning techniques. With the medical information, they constructed a heterogeneous network to discover the complex relationships among the symptoms, syndromes, and medicines. With the symptoms and medicine networks, they discovered the critical symptoms and symptom distribution for diagnosis, and the key medicines and their combinations for treatment.

The paper [2] presented Vienna's official online COVID-19 symptom checker, developed by the Vienna Social Fund (FSW, Vienna, Austria), the private company Symptoma, and the Public Health Services of the City of Vienna (MA15). The checker included 12 yes/no questions about symptoms to assess the risk for COVID-19. Users could also specify their age and sex, and whether they had contact with someone who tested positive for COVID-19. The paper analyzed several factors (symptoms, sex, or age) associated with COVID-19 positivity using a classifier model.

Guo et al. [3] explored the COVID-19 symptoms to predict its severity. Their main finding is that the severity of the illness strongly depends on the presence and severity of symptoms. The main symptom profiles can be well defined in terms of 5 or 6 symptoms. Initially, some symptoms are neurological and fatigue symptoms during the initial illness, and then neurological, gastrointestinal, and cardiopulmonary symptoms during the ongoing illness. The paper [6] explored the effects of long COVID-19 on the social life of females. They carried out semi-structured interviews via Zoom between April and June 2021 with females in the United States. They concluded that a long COVID-19 negatively affected females' social lives by causing physical limitations, economic issues, altered social relationships, social role conflicts, and social stigma.

Some incipient work exists on the co-occurrence of symptoms in patients. Now, these works have not analyzed the symptomatic behavior by mixing age, country, or variant of COVID-10, which we have called each one in this work as a dimension. For example, the goal of [4] was to identify and analyze reported symptom co-occurrences, and patient profiles. They used data extracted from Twitter and health-related forums and defined the reported symptoms using the MedDRA dictionary. Associations were assessed by computing co-occurrences in users' messages, as pairs, of the symptoms. To identify patient profiles in relation to their symptoms, user-level hierarchical clusters were created. The work of Wu et al. [7] used social media to track the co-existence of symptoms of the COVID-19 pandemic. They used a symptom lexicon containing 10 affected organs, 257 symptoms, and 1808 synonyms. Additionally, they analyzed symptoms between virus strains (Delta and Omicron) by comparing the prevalence of symptoms, to build a network of coexisting symptoms that defined the relationships between the symptoms and the affected body systems. Network analysis revealed co-occurrences between symptoms and certain systems such as cardiovascular, respiratory, and reproductive systems. In another study, Taquet et al. [8] analyzed how the incidence and co-occurrence of long-term COVID-19 symptoms varied based on demographic and disease severity differences. This involved subgroup comparisons of COVID-19 patients by sex, race, age, and hospitalization status. Regarding post-COVID conditions, Nuñez et al. [9] identified fatigue, difficulty breathing, headache, and concentration issues as predominant symptoms, often occurring together. They analyzed data from hospitalized adults in Mexico City. Danesh et al. [10] examined the post-acute sequelae of COVID-19 symptoms among patients at a COVID-19 recovery clinic, identifying clusters of patients with similar symptoms. The cohort included patients from 160 primary care clinics in Texas. Using a k-medoids algorithm, symptoms were found to cluster into two groups: neuropsychiatric (e.g., cognitive dysfunction) and pulmonary (e.g., dyspnea). Neuropsychiatric symptoms were more common in younger, female patients with longer symptom durations, while pulmonary symptoms were linked to higher comorbidity and hospitalization rates. Ghayda et al. [11] carried out a systematic review and a meta-analysis to find correlations between clinical characteristics and laboratory features of COVID-19 patients. The study provided critical insights into clinical and laboratory variables, highlighting more severe disease in elderly patients and common treatment strategies. They analyzed data from 37 studies (5196 participants) across various countries, and identified fever, cough, and fatigue/myalgia as the most common symptoms, with frequent gastrointestinal symptoms. Inflammatory markers like C-reactive protein (CRP) were elevated in 65% of cases, while lymphocyte counts were decreased in 63%. Predominant treatments included antiviral agents (79%), antibiotics (78%), and oxygen therapy (77%). Age correlated negatively with lymphocyte count and positively with dyspnea, white blood cells, neutrophils, and D-dimer.

Zhao et al. [12] examined the comorbidity between Internet addiction and depression in youth during the COVID-19 period. To do this, they used network analysis for statistical analysis considering the central symptoms of each context. In the case of Internet addiction, the central symptoms were "escape" and "irritability", and in the case of depression they were "energy" and "guilty". In particular, high correlations were observed between them, and in particular, those symptoms activate the negative feedback loop that makes them contribute even more to the comorbidity between Internet addiction and depression. Kerzhner et al. [13] conducted a systematic review and meta-analysis to evaluate persistent pain symptoms experienced by people who passed the acute

phase of COVID-19, to identify their associated functional consequences. Pain symptoms were grouped into six domains related to the chest, gastrointestinal system, musculoskeletal joints, musculoskeletal muscles, overall body, and nervous system. They conducted a meta-analysis to determine functional and quality of life impairments due to the persistence of long COVID-19 pain symptoms. Cheng et al. [14] categorized COVID-19 symptoms into four different groups of symptom combinations using text clustering methods: asymptomatic (Group 1), fever and/or dry cough (Group 2), upper respiratory symptoms (Group 3), and cardiopulmonary/systemic/gastrointestinal symptoms (Group 4). Now, the sizes of the data sets analyzed were relatively small. Finally, an incipient study has investigated the duration of COVID-19 symptoms with their co-occurrences. Gruber et al. [15] examined the frequency, duration, and patterns of long-term COVID-19 symptoms, alongside the factors contributing to prolonged effects. They utilized survival-time analysis with the Kaplan-Meier estimator to assess symptom persistence, and employed Cox regression to analyze symptom duration. Logistic regression identified risk factors associated with COVID-19 symptoms lasting over 90 days, including fatigue, headache, anosmia, and ageusia as the most common symptoms.

As can be seen, to our knowledge, there are no studies that simultaneously consider the duration of COVID-19 symptoms with their co-occurrences, analyzing specific dimensions or combinations of them (variant, country, or age). Particularly, our study analyzes countries of different continents, and different moments of the pandemic (COVID-19 variants) for various social groups (according to age). This allows analyzing the most common patterns of symptomatic behavior of patients in multiple dimensions. These patterns are characterized by the co-occurrence of symptoms and their durations.

B. CONTRIBUTIONS

In this article, we try to build a diagnostic model for COVID-19 that seeks to establish symptomatic patterns by dimension (variant, country, and age). In this way, in this work, the main contributions are making an analysis of the behavior of the symptoms using three dimensions: the variant of COVID-19, the country, and the age ranges. This allows studying the behavior of the symptoms to determine the correlations between them in each case (by dimension and cross between dimensions). Thus, this work carries out an exhaustive analysis of the symptoms of COVID-19 throughout the pandemic.

To perform this study, we use as a data source the COVID-19 Trends and Impact Survey data, which includes a collection of responses to direct surveys of people with COVID-19, with their symptoms. The University of Maryland Global COVID-19 Trends and Impact Survey (UMD-CTIS), in partnership with Facebook, has built the largest COVID-19 surveillance platform to date [16]. This platform collected daily since April 2020 the responses of invited Facebook users on topics related to the COVID-19 pandemic, such as test results, vaccinations, and symptoms from more than 100 countries. We use the COVID-19 Trends and Impact Survey data to estimate the distribution of the duration of symptoms among COVID-19 symptomatic cases. We estimated symptom duration using a geometric distributed function because it fits the data collected in the survey very well. Afterward, clustering techniques were used on the dimensions of age, variant, and country, to group the data according to the symptoms present, which allows us to establish those symptoms that co-occur by dimension. Particularly, the centroids of the groups are the symptomatic patterns to be analyzed, describing the symptoms that co-occur with the same duration. Thus, the possible dimensions or combinations of the dimensions will lead us to different grouping schemes. Then, the centroids of the bigger clusters in each case will describe the specific symptomatic behaviors present

The rest of the article is organized as follows. Section II presents the dataset and our approaches to determine the symptomatic patterns using clustering algorithms. Section III contains the experiments to determine the symptomatic patterns for each case considered in this study (each dimension, or some combinations of them). Then, Section IV carries out an explainability analysis of the symptomatic patterns found. Finally, Section V presents the major conclusions and future work.

II. MATERIAL AND METHODS

A. DATASET

The University of Maryland (UMD), in partnership with Facebook, launched the Global COVID-19 Trends and Impact Survey (UMD-CTIS) in April 2020, which recorded daily responses from Facebook users about the COVID-19 pandemic, including symptoms, test results, isolation measures, and vaccination, among others [16]. This survey was launched in over 50 languages and registered millions of responses from more than 100 countries worldwide. Access to the global COVID-19 survey dataset must be requested. Academic and nonprofit researchers with specific research objectives may request access by completing Data Use Agreements. Additionally, the general public can access aggregated statistics through a publicly accessible application programming interface (API) [17].

The survey consists of a web-based questionnaire that collects information on gender, age groups, COVID-19 tests, symptoms, and vaccination, among other variables, from individuals responding to the questionnaire through Facebook. The survey consists of a web-based questionnaire that collects information on gender, age groups, COVID-19 tests, symptoms, and vaccination, among other variables. The UMD organized and stored daily microdata that is used in our project. We have used the UMD-CTIS dataset to carry out the symptom co-occurrence considering three dimensions (age, country, and variant of COVID-19). We have examined four countries that showed different behavior during the pandemic: Brazil, Canada, Japan, and South Africa. These countries were chosen because of their geographic diversity

and availability of sufficient data. In addition, we perform our analysis for two periods: 2020 and 2021. We collected samples from respondents who reported at least one symptom in the previous 24 hours. In summary, the data about symptomatic respondents for the different countries in 2020 and 2021 are shown in Table 1. In addition, the table provides information on other individual characteristics such as gender, age groups, the average number of symptoms reported per questionnaire, and frequency of symptoms among positives and negatives.

B. OUR APPROACH

Two steps were carried out on the UMD-CTIS dataset to determine the symptomatic patterns to be analyzed by dimension. In the first step, the mean duration of the symptoms is estimated. In particular, we assume that the survey is completed at a point in the symptomatic period that follows a uniform distribution, in which the reported symptom durations are assumed to be independent samples that obey a geometric distribution, and we obtain the average duration. Specifically, the estimated average duration has been calculated for each population group separately. Thus, at each instant of time, the co-occurrences of the symptoms and their durations are different. In the later step, we performed a clustering process that was carried out using some configuration of the three dimensions studied (age, country, and variant). The clusters obtained from this process undergo an explainability analysis to determine the symptomatic pattern of the dimensional configuration under consideration. The rest of this section describes the clustering process.

1) Symptomatic Patterns

To determine symptomatic patterns, a data grouping process is proposed for each dimensional configuration. Therefore, the input dataset is filtered to obtain a specific configuration of age, country, and variant of COVID-19. Afterward, a clustering algorithm is implemented with input features of symptoms and their estimated duration mean. In this context, each cluster centroid represents a symptomatic pattern.

In this way, during the clustering process, three dimensions are taken into account (variants, age, and country) such that from the initial dataset, the specific sub-dataset for the dimensional configuration to be studied is built. The following list shows the sub-dimensions considered in each dimension:

- Variant:
  - 2020 Variant.
  - Delta Variant.
  - Omicron Variant.
- Age: In the case of age, we have regrouped the ranges in Table 2 into the following values:
  - Age range [18-34].
  - Age range [35-54].
  - Age range [55-74].
  - Age range: over 74.
- Country:
  - Japan.
  - Brazil.
  - Canada.
  - South Africa.

In the clustering, a total of 25 characteristics have been used, which is shown in Table 2. In this Table, each data corresponds to a person surveyed, the duration reports how long at least one of the symptoms lasted, and the 12 symptoms are grouped into the symptoms they have had in the last 24 hours, or after 24 hours at the moment to answer the survey.

2) Clustering Methods

A study of different clustering methods was carried out. The clustering techniques used in this study were K-means [18], hierarchical [19], Gaussian mixture [13], and DBSCAN [20].

To assess the quality of the clustering algorithms, we have used self-contained performance metrics appropriate for each type of clustering algorithm. These metrics are self-contained because they indicate, according to the value obtained, how good the clustering is (the quality of the obtained clusters), where their maximum values indicate optimal clustering and their minimum values indicate bad clustering. In the case of distance-based techniques, we have used the silhouette index, which is based on quantifying the variance in the clusters combined with their separations [21]. The Silhouette index varies between -1 and 1, such that -1 indicates a bad clustering, 0 is indifferent, and 1 an optimal clustering. In the case of the density-based technique, there are two well-known metrics: CDbw and the Density-Based Clustering Validation (DBCV) index [22]. We use DBCV here, which considers the densities and shapes of the clusters. That index produces values between -1 and 1, with higher values indicating better density-based clustering solutions.

We have carried out a comparison process in three case studies for each dimension. To do this, each technique has previously been subjected to a process of optimization of its hyperparameters using a Hyperparameter Optimization scheme based on Grid Search [23], [24]. In addition, the performance metrics presented above have been used to indicate the quality of the techniques. For K-means, Gaussian, and Hierarchical mixture, the Silhouette index was used; and for DBSCAN, the DBCV index. Table 3 shows an example of the results obtained. Specifically, this Table shows the average of 20 runs obtained with each technique for each performance metric for specific values in each dimension. Thus, this table shows the results for the country Japan, individuals in the age range [35-54], and the COVID-19 variant 2000. The Silhouette index was used universally, except for DBSCAN which used the DBCV index in each analyzed dimension.

Finally, the ANOVA test [25] was carried out to evaluate two things: i) if there are statistically significant differences in the performance of each technique in its 20 runs for each case study (dimension) (see Table 4), and ii) If there is a significant difference between the techniques in each case study (see Table 5).

TABLE 1. Population used in this work by country and two non-overlapped periods (2020 and 2021).

Characteristic	Brazil		Canada		Japan		South Africa	
	2020	2021	2020	2021	2020	2021	2020	2021
1. Gender								
(a) Female, N	45357	130235	5438	19472	1679	14283	3923	11291
(b) Male, N	24928	76689	2315	9824	2388	20791	2525	6730
2. Age groups								
(a) 18-24, N	8270	27474	1136	3248	179	871	739	1580
(b) 25-34, N	19596	56227	2337	7172	577	3797	2252	4889
(c) 35-44, N	21061	57452	1750	6688	997	7527	1801	4721
(d) 45-54, N	13776	39122	1210	5215	1216	10413	1141	3878
(e) 55-64, N	6968	22190	954	4478	828	8724	491	2124
(f) 65-74, N	140	6016	308	2421	479	3529	1667	799
(g) 75+, N	233	1025	126	825	66	846	27	230
3. Average number of symptoms among positive	5.37	5.16	5.25	5.27	4.38	4.45	5.51	5.61
4. Symptoms among positive								
(a) Fever, %	22.56	21.92	22.43	22.63	39.28	38.49	32.55	30.77
(b) Cough, %	54.73	57.46	63.01	67.46	61.65	64.47	58.89	65.96
(c) Difficulty breathing, %	30.72	28.17	23.74	22.80	18.79	16.62	29.03	27.61
(d) Fatigue, %	60.51	57.58	69.33	71.13	51.50	57.06	65.24	67.88
(e) Stuffy or runny nose, %	57.86	57.33	62.29	68.62	49.24	47.31	55.02	62.59
(f) Aches or muscle pain, %	58.90	58.01	55.13	53.10	41.35	44.45	57.43	58.73
(g) Sore throat, %	35.06	34.37	34.84	39.67	37.21	35.27	36.14	38.78
(h) Chest pain, %	32.00	30.03	22.19	21.52	20.67	22.88	39.25	35.57
(i) Nausea, %	29.94	28.34	26.61	25.08	11.65	10.17	27.84	28.41
(j) Loss of smell or taste, %	54.15	46.25	53.34	42.67	40.22	39.99	51.70	45.89
(k) Headache, %	65.74	63.73	60.14	58.86	41.35	44.40	64.68	65.72
(l) Chills, %	34.96	33.31	32.21	33.46	25.56	24.28	33.67	33.75
5. Average number of symptoms among negative	3.12	2.88	3.19	2.83	2.73	2.28	2.85	2.99
6. Symptoms among negative								
(a) Fever, %	6.12	5.79	4.61	4.58	19.23	11.61	10.94	12.13
(b) Cough, %	34.17	32.75	38.45	32.24	37.57	28.55	33.57	35.98
(c) Difficulty breathing, %	13.71	11.50	12.34	9.52	4.70	10.94	11.10	10.03
(d) Fatigue, %	33.46	30.02	53.05	48.95	35.29	30.48	36.06	38.81
(e) Stuffy or runny nose, %	48.86	47.88	55.09	49.82	46.35	44.60	40.82	44.61
(f) Aches or muscle pain, %	41.67	40.19	39.85	37.05	34.28	35.19	33.59	35.87
(g) Sore throat, %	23.76	21.83	27.83	21.90	28.11	20.40	22.06	22.30
(h) Chest pain, %	15.11	12.97	10.97	8.09	10.01	7.24	15.15	15.34
(i) Nausea, %	15.37	13.42	16.27	12.99	7.97	6.47	13.85	14.94
(j) Loss of smell or taste, %	10.70	5.97	4.56	3.54	3.48	2.10	8.11	7.33
(k) Headache, %	50.90	49.47	43.92	42.75	34.49	30.58	48.79	47.52
(l) Chills, %	18.15	16.31	11.82	10.77	12.00	7.78	11.36	12.66

Specifically, to determine if the differences between the means were statistically significant, the p-value was compared with the significance level to evaluate the null hypothesis, which indicates that the population means are all equal. The significance level chosen was 0.05.

In the first case (see table 4), the ANOVA test was applied to the results of the 20 runs in each case study. According to the results obtained, no statistically significant differences were observed in the performance of the techniques for each of the dimensions and their 20 runs. Thus, the null hypothesis is rejected and we conclude that the differences between the means are not statistically significant.

In the second case, we want to determine if the difference between the means of the techniques, for the different dimensions, were significantly different. In this case, as can be seen in Table 5, the differences between the means are statistically significant (the null hypothesis is accepted).

Finally, according to the results, K-means was chosen because it offered better quality in the clusters created. In the case of the Gaussian mixture and Hierarchical approaches, the quality of the clusters obtained was worse than those of the K-means algorithm. For DBSCAN, the clustering process created about 1000 clusters.

III. EXPERIMENTS

To determine the optimal number of groups for K-means in each dimension configuration, we initially use the Elbow method based on the silhouette index. The Elbow method allows for determining the number of K groups that give the best silhouette index value. The goal is to identify a point on the chart where the rate of decline of the index stops slowing sharply, resembling an "elbow" shape. These values around the elbow are generally considered the optimal value of K. Now, in our specific case, the value of K around the elbow that additionally meets the following two criteria will be chosen:

- Do not degrade the quality of the silhouette index (its value is not less than 0.4).

TABLE 2. Variables for the clustering process.

Variable type	ID Variable	Description
Duration	B2	Duration in days of at least one of the symptoms
Symptoms in the last 24h	B1_1	Fever
	B1_2	Cough
	B1_3	Difficulty breathing
	B1_4	Fatigue
	B1_5	Stuffy or runny nose
	B1_6	Aches or Muscle pain
	B1_7	Sore throat
	B1_8	Chest pain
	B1_9	Nausea
	B1_10	Loss of smell or taste
	B1_12	Headache
	B1_13	Chills
Symptoms after the last 24h	B1b_x1	Fever
	B1b_x2	Cough
	B1b_x3	Difficulty breathing
	B1b_x4	Fatigue
	B1b_x5	Stuffy or runny nose
	B1b_x6	Aches or Muscle pain
	B1b_x7	Sore throat
	B1b_x8	Chest pain
	B1b_x9	Nausea
	B1b_x10	Loss of smell or taste
	B1b_x12	Headache
	B1b_x13	Chills

TABLE 3. Average results obtained by technique for specific values in each of the dimensions.

Technique	2000 Variant	Age range [35-54]	Japan
K-means	0.49	0.49	0.51
Gaussian mixture	0.13	0.24	0.19
DBSCAN	-0.51	-0.51	-0.62
Hierarchical	0.13	0.21	0.12

TABLE 4. Results of the ANOVA test for Silhouette index for each technique

Case	K-means	Gaussian mixture	Hierarchical
Japan	0.412	0.050	0.048
2020 Variant	0.351	0.061	0.080
Age range [35-54]	0.094	0.055	0.051

TABLE 5. Results of the ANOVA test for Silhouette index for three sub-dimensions

2020 Variant	Age range [35-54]	Japan
2.83e-08	1.15e-09	7.35e-12

- The number of clusters is greater than 5 to be able to analyze various symptomatic patterns in each case.

In each of the created clustering models (for each configuration of dimensions), the following information is analyzed:

- The number of K with which the model has been created,
- The silhouette quality index.
- The centroids of each cluster, where each one represents:
  - -- Duration in days of symptoms.
  - -- Symptoms present in the last 24 hours.
  - -- Symptoms present after 24 hours.
- The percentage of records in each cluster.

Table 6 shows the K chosen in each dataset for the unidimensional cases, with its silhouette value. We can see from these results in Table 6 that the value of the Silhouette index was always around 0.5 for a K around 8.

TABLE 6. Clustering Models for each dimension using K-means.

Dimension	Case	Values of K	Silhouette value
Variants	2020 Variant	6	0.51
	Delta Variant	8	0.48
	Omicron Variant	9	0.52
Age ranges	[18-34]	6	0.51
	[35-54]	6	0.49
	[55-74]	7	0.55
	< 74	8	0.59
Country	Japan	9	0.49
	Brazil	9	0.49
	Canada	9	0.49
	South Africa	8	0.43

Additionally, we have carried out the next crossing of dimensions:

- Age Range vs Variant
  - -- >74 - Variant 2020.
  - -- >74 - Delta variant.
  - -- [55-74] - Variant 2020.
  - -- [55-74] - Delta variant.
- Age range vs Country
  - -- [55-74] - Japan.
  - -- >74 - Japan.
  - -- [55-74] - Brazil.
  - -- >74 - Brazil.
- Country vs Variant
  - -- Canada - Variant 2020.
  - -- South Africa - Variant 2020.
  - -- Canada - Variant Delta.
  - -- South Africa - Variant Delta,

Table 7 shows the K chosen for these datasets and the value of the Silhouette index. In these cases, the silhouette value in all cases is better than for the one-dimensional cases with a similar cluster number as before (around 8).

IV. STUDY OF THE OBTAINED SYMPTOMATIC PATTERNS

In this section, we are going to analyze some of the centroids (symptomatic patterns) found during the clustering process.

TABLE 7. Clustering Models for each cross dimension using K-means.

Dimension	Crosses	Values of K	Silhouette
Age Range/Variant	>74 - 2020 Variant	10	0.52
	>74 - Delta Variant	10	0.59
	[55-74] - 2020 Variant	7	0.57
	[55-74] - Delta Variant	10	0.54
Age Range/Country	[55-74] - Japan	9	0.59
	>74 - Japan	8	0.61
	[55-74] - Brazil	8	0.53
	>74 - Brazil	7	0.53
Country/Variant	Canada - 2020 Variant	7	0.56
	South Africa - 2020 Variant	6	0.51
	Canada - Delta Variant	7	0.55
	South Africa - Delta Variant	8	0.51

Additionally, we are going to discuss the correlation between the symptoms, and their relationships with the found centroids. It is important to highlight that the explainability analysis carried out on the clusters in this section is different from what is usually done with the explainability analysis methods of the Explainable artificial intelligence (XAI) area. Particularly, XAI methods seek to explain how results are obtained when using machine learning techniques. In our case, we want to interpret the clusters obtained (specifically, the dominant one), for which we will study their centroids. This is possible since a cluster can be analyzed from its centroid, its representative element (pattern).

A. ANALYSIS OF CENTROIDS

For the analysis of the centroids, the following data structure was considered, based on Table 2:

- The number of days with at least one of the symptoms (1);
- Symptoms in the last 24 hours/after 24 hours at the time of taking the survey: (2) Fever, (3) Cough, (4) Difficulty Breathing, (5) Fatigue, (6) Stuffy or runny nose, (7) Aches or muscle pain, (8) Sore throat (9) Chest pain, (10) Nausea (11) Loss of smell taste, (12) Headache, (13) Chills.

A typical centroid is:

{0.96 0.03/0.01 0.22/0.05 0.06/0.02 0.27/0.06 0.43/0.07 0.35/0.06 0.16/0.05 0.08/0.03 0.10/0.03 0.03/0.01 0.47/0.08 0.11/0.04}

That means an average duration in days of symptoms of 0.96, a value of fever of the last 24 hours of 0.03 and after 24 hours of 0.01, a value of cough of the last 24 hours of 0.22 and after 24 hours of 0.05, and so on for the rest of the symptoms.

We assume that values of the symptoms around 0.5 mean a (H)igh incidence of the symptom, lower than these values and close to 0.2 a (M)edium incidence, and (L)ow incidence from 0.05. For the rest of the values, the incidence is (N)il. These values are normalized, and this assumption is based on work [14], which considers a very relevant symptom if that normalized value is greater than 0.5.

1) Variant Dimension

Table 8 shows the symptoms that appear in each centroid for the 2020 variant. For cluster 1, the most relevant symptoms are 6 (Stuffy or runny nose) and 12 (Headache). In the case of cluster 2, the more relevant symptoms are 5 (Fatigue) and 7 (Aches or muscle pain), and so for the rest of the clusters. Symptoms 5 (Fatigue), 6 (Stuffy or runny nose), and 7 (Aches or muscle pain) appear with high relevance in almost all the clusters (in those that do not, they appear with medium relevance). Thus, these three symptoms describe the base pattern for the 2020 variant. Finally, in this case, the symptoms after 24 hours are normally not relevant.

In this case, the proportion of individuals by cluster is for cluster 1 equal to 70.18%, cluster 2 3.63%, cluster 3 1.07%, cluster 4 5.37%, cluster 5 1.27% and cluster 6 18.45%. Thus, in this case, the centroid of cluster 1 can be considered as the *dominant pattern (DP)* for this dimension, and the centroid of cluster 6 as the *subdominant pattern (SP)*. The *dominant* and *subdominant patterns* will be the patterns that will be used in each dimension (and crossing between them) to perform the comparisons of the symptomatic patterns.

TABLE 8. Incidence of the symptoms for 2020 Variant

Values vs Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13
Cluster 1	1.96	N/N	M/L	L/N	M/L	H/L	M/L	M/L	L/N	L/N	M/N	H/L	L/N
Cluster 2	30.04	N/N	M/L	L/L	H/N	M/L	H/L	M/L	L/L	L/N	L/N	N/L	N/N
Cluster 3	90.27	N/N	M/N	L/L	H/N	H/N	H/L	L/N	L/N	L/N	M/N	M/H	L/N
Cluster 4	16.24	L/N	H/L	L/L	H/L	H/L	H/L	M/L	M/L	M/L	M/L	H/L	M/L
Cluster 5	50.82	M/N	M/L	L/N	H/L	H/L	H/L	L/N	L/N	L/N	L/L	M/L	L/N
Cluster 6	6.88	L/N	H/L	L/L	H/L	H/L	H/L	M/L	M/L	M/L	M/L	H/L	M/L

In this case, the proportion of individuals by cluster is for cluster 1 equal to 70.18%, cluster 2 3.63%, cluster 3 1.07%, cluster 4 5.37%, cluster 5 1.27% and cluster 6 18.45%. Thus, in this case, the centroid of cluster 1 can be considered as the *dominant pattern (DP)* for this dimension, and the centroid of cluster 6 as the *subdominant pattern (SP)*. The *dominant* and *subdominant patterns* will be the patterns that will be used in each dimension (and crossing between them) to perform the comparisons of the symptomatic patterns.

In the case of Omicrom Variant, the DP and SP are shown in Table 9. These clusters have 64.05% and 22.12% of the individuals.

TABLE 9. Patterns for Omicrom Variant

Values vs Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13
DP	1.91	L/N	M/L	N/N	M/L	H/L	H/L	M/L	L/N	L/N	L/N	H/L	M/L
SP	5.38	L/L	H/M	L/L	H/M	H/M	H/M	M/M	L/L	L/L	L/L	H/M	L/L

For the Delta Variant, the dominant and subdominant patterns are shown in Table 10. These clusters have 58.57% and 26.90% of the individuals. If we compare the dominant patterns between these three variants, we can find that the symptoms (6) Congestion/Stuffy or runny nose, (7) Pain/Aches and (12) Headache are normally relevant in all of them, and when they do not relevant, then they have at least medium relevance. Thus, the three dominant patterns are quite similar, both in duration and in the most relevant symptoms,

TABLE 10. Patterns for Delta Variant

Values vs Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13
DP	1.47	N/N	L/L	N/N	M/L	M/L	M/L	L/N	L/N	L/N	M/N	H/L	L/L
SP	3.69	L/N	M/L	L/L	M/L	H/L	H/L	M/L	L/L	L/L	L/N	H/L	L/L

or symptoms without impact. On the other hand, for the subdominant patterns, the most relevant variation is in time, but they remain in the same order of magnitude between them for the different variants (6.88, 5.34, and 3.69, respectively). Again, the dominant symptoms are very similar. Also, there are other symptoms that in some cases appear relevant, such as (3) cough and 5 (Fatigue) (see table 9). Thus, we can see that the symptomatic pattern is very similar among the 6 patterns.

2) Country Dimension

For the case of countries, Table 11 shows the symptoms that appear in the dominant and subdominant patterns for the different countries. For Japan, the dominant pattern is 57.17% of the sample, and the subdominant pattern is 21.10%. In the case of Brazil, the dominant pattern is 46.53% of the sample, and the subdominant pattern is 30.07%. For Canada, the dominant pattern is 57.94% of the sample, and the subdominant pattern is 18.31%. Finally, for South Africa, the dominant pattern is 63.20%, and the subdominant pattern is 21.08%.

On the other hand, the common relevant symptoms of the dominant patterns are (5) Fatigue, (6) Congestion/Stuffy or runny nose, and (12) Headache, and when they are not relevant, they have at least medium relevance. In the case of subdominant patterns, they have a similar behavior. Also, it is highlighted that in the case of Japan, a phenomenon occurs that is different from the rest, because the symptoms (5) Fatigue and (12) Headache are no longer relevant. For the subdominant patterns, the behavior is very similar, and again, the duration time of the symptoms is between 3.49 and 5.58. Thus, the six dominant patterns are quite similar at the level of the most relevant symptoms, or symptoms without impact.

Finally, the symptoms after 24 hours are normally not relevant, as well (2) Fever, (4) Difficulty Breathing, (9) Chest pain, (10) Nausea, and (13) Chills.

TABLE 11. Patterns for countries

Country	Values vs Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13
Japan	DP	1.42	N/N	M/N	N/N	M/N	H/L	M/L	L/L	N/N	N/N	N/N	M/L	N/N
	SP	3.49	N/N	M/L	N/N	M/L	H/L	M/L	M/L	N/N	N/N	M/N	M/L	N/N
Brazil	DP	1.52	L/N	M/N	L/L	M/N	H/L	M/L	L/N	L/N	L/N	M/N	H/L	L/N
	SP	3.73	L/N	M/L	L/N	H/L	H/L	H/L	M/L	L/L	L/L	L/N	H/L	L/L
Canada	DP	1.77	N/N	L/N	N/N	H/L	H/L	M/N	L/N	N/N	L/N	M/N	H/L	L/N
	SP	5.58	N/N	M/L	L/N	H/L	H/L	H/L	L/L	L/N	L/N	M/N	H/L	L/N
South Africa	DP	1.84	L/N	M/L	N/N	M/L	H/L	M/L	L/N	L/N	L/N	M/N	H/L	L/N/
	SP	5.50	L/L	H/L	L/L	H/L	H/L	H/L	M/L	L/L	L/L	L/L	H/L	L/L

3) Age Dimension

For the case of age, Table 12 shows the symptoms that appear in the dominant and subdominant patterns for the different countries. For 18-34, the dominant pattern is 71.44% of the sample, and the subdominant pattern is 18.86%. In the case of 35-54, the dominant pattern is 66.64% of the sample, and the subdominant pattern is 20.25%. For 55-74, the dominant pattern is 60.41% of the sample, and the subdominant pattern is 20.24%. Finally, over 74, the dominant pattern is 59.49%, and the subdominant pattern is 25.90%.

The relevant symptoms are 6 (Stuffy or runny nose), 7 (Aches or muscle pain), and 12 (Headache), with a value of the duration of symptoms around 2 for the dominant patterns. In this case, the symptoms after 24 hours are normally not relevant, as well as (2) Fever, (4) Difficulty Breathing, (9) Chest pain, (10) Nausea, and (13) Chills. Again, the behavior of the patterns by age is not very different.

TABLE 12. Patterns for Age ranges

Ages Range	Values vs Criterion	1	2	3	4	5	6	7	8	9	10	11	12	13
18-34	DP	2.01	N/N	L/L	L/N	M/L	H/L	M/L	L/N	L/N	L/N	M/N	H/L	L/N
	SP	6.75	L/L	H/M	L/L	H/L	H/L	H/L	M/L	L/L	M/L	M/L	H/L	M/L
35-54	DP	2.01	N/N	M/L	N/L	M/L	H/L	M/L	L/L	L/L	L/L	M/L	H/L	L/L
	SP	6.90	L/N	H/L	L/L	H/L	H/L	H/L	M/L	L/L	L/L	L/L	H/L	L/L
55-74	DP	1.98	L/N	M/N	L/N	M/N	H/L	H/L	L/N	N/N	N/N	L/N	M/L	N/N
	SP	7.13	N/N	M/L	L/N	M/L	H/L	H/L	L/L	L/N	L/N	L/N	M/L	L/N
Over 74	DP	2.31	N/N	M/N	N/N	M/L	H/L	H/L	L/N	N/N	N/N	M/N	L/N	N/N
	SP	8.33	N/N	M/L	L/N	M/L	H/L	H/L	L/L	L/N	L/N	L/N	L/N	N/N

4) Cross dimension: Age range vs. Variant

For the case of the combination of Age range vs. Variant, Table 13 shows the symptoms that appear in the dominant and subdominant patterns for the different combinations. For the combination >74 - Variant 2020, the dominant pattern is 65.00% of the sample, and the subdominant pattern is 17.04%. In the case of >74 - Delta variant, the dominant pattern is 66.64% of the sample, and the subdominant pattern is 20.25%. For [55-74] - 2020 Variant, the dominant pattern is 64.41% of the sample, and the subdominant pattern is 18.62%. Finally, for the combination [55-74] - Delta variant, the dominant pattern is 69.49%, and the subdominant pattern is 15.90%.

In this case, the relevant symptoms are 6 (Stuffy or runny nose) and 7 (Aches or muscle pain), with a value of the duration of symptoms around 2 for the dominant patterns. In this case, the symptoms after 24 hours are normally not relevant, as well (2) Fever, (4) Difficulty Breathing, (9) Chest pain, (10) Nausea, and (13) Chills.

TABLE 13. Patterns for Age range vs. Variant

Combination	Patterns	1	2	3	4	5	6	7	8	9	10	11	12	13
>74 - 2020 Variant	DP	1.84	N/N	M/L	L/N	M/L	H/L	H/L	L/N	N/N	N/N	M/N	L/N	N/N
	SP	5.67	N/N	M/L	L/N	M/L	H/L	H/L	L/L	L/N	L/N	M/N	M/L	L/N
>74 - Delta variant	DP	1.80	N/N	M/N	N/N	M/L	H/L	H/L	L/N	N/N	N/N	L/N	L/N	N/N
	SP	5.72	N/N	M/L	N/N	M/L	H/L	H/L	L/N	N/N	N/N	L/N	L/N	N/N
[55-74] - 2020 variant	DP	1.92	N/N	M/N	N/N	M/N	H/N	H/N	L/N	N/N	N/N	L/N	M/N	L/N
	SP	7.11	N/L	H/N	L/N	M/L	H/L	H/L	M/N	L/N	L/N	L/N	M/L	L/N
[55-74] - Delta variant	DP	2.56	N/N	M/L	N/N	M/L	H/L	H/L	L/L	N/N	N/N	M/N	H/L	N/N
	SP	0.99	N/N	L/N	N/N	M/L	H/L	H/L	L/N	N/N	N/N	L/N	H/L	L/N

5) Cross dimension: Country vs Variant

For the case of the combination of Country vs Variant, Table 14 shows the symptoms that appear in the dominant and

subdominant patterns for the different combinations. For the combination Canada - 2020 Variant, the dominant pattern is 65.32% of the sample, and the subdominant pattern is 15.40%. In the case of South Africa - 2020 Variant, the dominant pattern is 73.40% of the sample, and the subdominant pattern is 16.77%. For the Canada - Delta Variant, the dominant pattern is 68.40% of the sample, and the subdominant pattern is 19.13%. Finally, for the combination South Africa - Delta Variant, the dominant pattern is 73.64%, and the subdominant pattern is 21.61%.

In this case, the relevant symptoms are 6 (Stuffy or runny nose), 7 (Aches or muscle pain), and (12) Headache, but (3) Cough and (5) Fatigue are also important. The value of the duration of symptoms is around 1.8 for the dominant patterns and 6 for the subdominant patterns. Again, the symptoms after 24 hours are not relevant, as well (2) Fever and (4) Difficulty Breathing.

TABLE 14. Patterns for Country vs. Variant

Combination	Patterns	1	2	3	4	5	6	7	8	9	10	11	12	13
Canada-2020 Variant	DP	1.85	N/N	M/N	N/N	H/N	H/N	H/N	L/N	N/N	L/N	L/N	H/N	N/N
	SP	6.88	N/N	M/L	L/N	H/L	H/L	H/L	L/L	L/N	L/N	L/N	H/L	L/N
S. Africa-2020 Variant	DP	1.93	L/N	M/N	N/N	L/N	H/N	M/N	L/N	L/N	L/N	M/N	H/L	L/N
	SP	6.62	L/N	H/L	L/L	H/L	H/L	H/L	M/L	L/L	L/L	L/L	H/L	L/L
Canada-Delta Variant	DP	1.77	N/N	M/N	N/N	M/N	H/N	M/N	L/N	L/N	L/N	L/N	H/L	L/N
	SP	5.49	L/N	H/L	L/N	H/L	H/L	H/L	M/L	L/L	L/L	L/L	H/L	L/L
S. Africa-Delta Variant	DP	1.88	L/N	M/L	N/N	M/L	H/L	M/L	L/N	L/N	L/N	M/N	H/L	L/N
	SP	5.54	L/L	H/M	L/L	H/M	H/M	H/M	M/L	M/L	L/L	L/L	H/M	M/L

6) Cross dimension: Age range vs Country

For the case of the combination of Age range vs Country, Table 15 shows the symptoms that appear in the dominant and subdominant patterns for the different combinations. For the combination [55-74] - Japan, the dominant pattern is 62.84% of the sample, and the subdominant pattern is 19.23%. In the case of >74 - Japan, the dominant pattern is 57.14% of the sample, and the subdominant pattern is 22.14%. For [55-74] - Brazil, the dominant pattern is 59.42% of the sample, and the subdominant pattern is 22.00%. Finally, for the combination >74 - Brazil, the dominant pattern is 59.73%, and the subdominant pattern is 33.99%.

In this case, the relevant symptoms are 6 (Stuffy or runny nose) and 7 (Aches or muscle pain), and the value of the duration of symptoms is around 1.9 for the dominant patterns and close to 7 for the subdominant patterns. Again, the symptoms after 24 hours are not relevant as well (2) Fever, (4) Difficulty Breathing, (9) Chest pain, (10) Nausea, and (13) Chills. The rest are not very relevant.

B. ANALYSIS OF CORRELATIONS AMONG THE SYMPTOMS

Finally, an analysis of Pearson and Spearman correlations between symptoms has been carried out in some of the clusters created in the clustering process by dimension. Thus, the correlation analysis of the symptoms was done for each value range of each dimension, or for combinations of them, but not for the dataset in general. Previous studies made a general analysis of correlations between symptoms but also added variables of clinical characteristics and laboratory results [26], which is different from our approach. The following dimensions were considered to perform the correlation analyses:

- Variant Dimension.
- Age range dimension (in the supplementary materials).
- Cross-age vs. variants.

The criteria for selecting a variable to be analyzed in each of these dimensions were:

- It is a relevant variable.
- It has a high correlation with others regardless of its relevance.

The variables are considered to have a significant correlation if the value is greater than 0.50, and medium correlation close to this value. Finally, in the following Tables, from (2) to (13) correspond to the symptoms (Fever, Cough, Difficulty breathing, Fatigue, Nasal Congestion or runny nose, etc.) in the last 24 hours, while from (14)) to (25), the same symptoms after 24 hours.

TABLE 15. Patterns for Age range vs. Country

Age range/Country	Patterns	1	2	3	4	5	6	7	8	9	10	11	12	13
[55-74]-Japan	DP	1.91	N/N	M/N	N/N	M/N	H/L	H/L	L/N	N/N	N/N	L/N	M/N	N/N
	SP	7.30	N/N	M/N	N/N	M/N	H/L	H/L	L/N	N/N	N/N	L/N	L/N	N/N
>74-Japan	DP	1.96	N/N	M/N	N/N	M/N	H/L	H/L	L/N	N/N	N/N	M/N	L/N	N/N
	SP	7.56	N/N	M/N	N/N	M/N	H/N	H/N	L/N	N/N	N/N	L/N	L/N	N/N
[55-74]-Brazil	DP	1.90	N/N	M/N	N/N	N/N	H/L	H/L	L/N	N/N	N/N	L/N	H/L	L/N
	SP	5.40	N/N	H/L	L/N	H/L	H/L	H/L	M/L	L/N	L/N	L/N	H/L	L/L
>74-Brazil	DP	2.07	N/N	M/N	N/N	M/N	H/L	H/L	L/N	N/N	N/N	L/N	M/L	L/N
	SP	7.11	N/N	H/L	L/N	H/L	H/L	H/L	L/L	L/N	L/N	M/N	L/L	N/N

1) Variant Dimension

For the relevant variables (see Table 16), the Spearman's correlations between the most relevant variables, such as (6) Stuffy or runny nose and (7) Aches or muscle pain is -0.04, or (6) Stuffy or runny nose and (12) Headache is 0.05, or (7) Aches or muscle pain and (12) Headache is 0.21, for the case of 2020 variant. Pearson are -0.18, 0, and 0.17, respectively. With respect to Delta and Omicron Variants, only (7) Aches or muscle pain and (12) Headache have a medium Pearson correlation, the rest of the relevant variables are not correlated.

On the other hand, the most relevant correlations are between the same variables in the last 24 hours and after 24 hours, but not between different symptoms (see Table 17). Thus, we see that there are no correlations between different symptoms.

Table 18 shows the high correlations between different symptoms for this dimension. We can see the repetition of a good correlation between (5) Fatigue and (12) Headache symptoms in several cases. Also, it is observed that the highest correlations are between these symptoms.

2) Cross dimension: Age range vs. Variant

In this case, Table 19 shows the correlations. In all cases, symptoms (6) Stuffy or runny nose and (7) Aches or muscle

TABLE 16. Correlations between the relevant variables in the Variant Dimension

Variant	Variables	Spearman	Pearson
2020	3 and 5	0.04	0.09
	3 and 6	0.02	0.12
	3 and 7	-0.06	0.03
	3 and 12	0.05	0.09
	5 and 6	-0.10	0.05
	5 and 7	0.24	0.29
	5 and 12	0.26	0.27
	6 and 7	-0.04	-0.18
	6 and 12	0.05	0
	7 and 12	0.21	0.17
Delta	6 and 7	-0.15	-0.09
	6 and 12	0.05	0.06
	7 and 12	0.15	0.21
Omicron	3 and 5	0.01	0.08
	3 and 6	0.13	0.18
	3 and 7	-0.08	0.01
	3 and 12	0.08	0.13
	5 and 6	-0.08	0.09
	5 and 7	0.12	0.30
	5 and 12	0.24	0.29
	6 and 7	-0.16	0
	6 and 12	0.03	0.13
	7 and 12	0.18	0.27

TABLE 17. The Highest Correlations in Variant Dimension

Variant	Variables	Spearman	Pearson
2020	2 and 14	0.58	0.59
	4 and 16	0.56	0.56
	8 and 20	0.50	0.52
	9 and 21	0.56	0.61
	10 and 22	0.57	0.58
	11 and 23	0.70	0.66
	13 and 25	0.56	0.63
Delta	2 and 14	0.72	0.82
	3 and 15	0.45	0.61
	4 and 16	0.67	0.74
	5 and 17	0.47	0.56
	7 and 19	0.39	0.51
	8 and 20	0.60	0.72
	9 and 21	0.71	0.77
	10 and 22	0.64	0.72
	11 and 23	0.78	0.85
	12 and 24	0.49	0.52
	13 and 25	0.69	
Omicron	2 and 14	0.74	0.80
	3 and 15	0.50	0.61
	4 and 16	0.69	0.76
	5 and 17	0.47	0.59
	6 and 18	0.36	0.51
	7 and 19	0.46	0.56
	8 and 20	0.58	0.74
	9 and 21	0.66	0.79
	10 and 22	0.59	0.74
	11 and 23	0.77	0.82
	12 and 24	0.50	0.55
	13 and 25	0.65	0.78

TABLE 18. The Highest Correlations between different symptoms in Variant Dimension

Variant	Variables	Spearman	Pearson
2020	5 and 12	0.42	0.45
Delta	3 and 8	0.35	0.36
	5 and 7	0.34	0.34
Omicron	5 and 12	0.35	0.39

pain have a medium negative correlation between them. The rest of the relevant symptoms are not correlated.

TABLE 19. Correlations between the relevant variables in the Cross dimension: Age range vs. Variant

Variant	Variables	Spearman	Pearson
>74-2020 Variant	6 and 7	-0.31	-0.21
>74-Delta Variant	6 and 7	-0.29	-0.16
[55-74]-2020 Variant	3 and 6	0.01	0.02
	3 and 7	-0.13	-0.11
	6 and 7	-0.28	-0.23
[55-74]-Delta Variant	6 and 7	-0.29	-0.25
	6 and 12	0.05	0.14
	7 and 12	0.12	0.07

Table 20 shows the high correlations between different symptoms for this dimension. It is relevant to appreciate several things, that there are correlations between different symptoms in the last 24 hours and symptoms of more than 24 hours (for example, between (2) Fever and (22) Nausea symptoms for >74 vs Delta Variant, or between (10) Nausea and (21) muscle pain symptoms for [55-74] vs 2020 Variant). We can also see the repetition of a good correlation between (4) Difficulty Breathing and (13) Chills symptoms in several crosses. Finally, it is observed that the highest correlations are between between (14) Fever and (25) Chills symptoms with 0.68 (Pearson), and between (4) Difficulty Breathing and (13) Chills symptoms with 0.59 (Pearson).

TABLE 20. High Correlations between different symptoms in the Cross dimension: Age range vs. Variant

Variant	Variables	Spearman	Pearson
>74-2020 Variant	16 and 25	0.54	0.57
	4 and 13	0.44	0.41
	4 and 9	0.40	0.42
	5 and 13	0.38	0.41
	2 and 10	0.37	0.39
	16 and 17	0.36	0.38
>74-Delta Variant	14 and 25	0.64	0.68
	4 and 13	0.57	0.59
	2 and 13	0.50	0.52
	2 and 22	0.49	0.46
	14 and 23	0.46	0.45
[55-74]-2020 Variant	8 and 9	0.39	0.40
	10 and 21	0.41	0.44
	5 and 7	0.39	0.38
[55-74]-Delta Variant	2 and 12	0.34	0.35
	16 and 21	0.34	0.34

C. STATISTICAL ANALYSIS OF THE CENTROIDS

In this section, we analyze whether there is a significant difference between the centroids of the clusters obtained, particularly, for the dominant and subdominant patterns. In our specific scenario, we have applied the t-test to compare these two centroids taking into account the 25 variables of which the centroids are composed. Two cases were selected for the centroid analysis: one linked to the dimension of country for Japan with 9 clusters (case 1), and another related to the combination of the dimensions of age greater than 74 and the Delta variant with 10 clusters (case 2).

Thus, in this case, the goal of the t-test is to evaluate if the means of the centroids are the same (null hypothesis). In this context, we have set the p-value for the null hypothesis at 0.05 (it is used by default), acting as a threshold to determine the significance of the observed differences. Specifically, the t-test evaluates whether the difference between the centroids between dominant and subdominant patterns is statistically significant [27]. The p-value is calculated from individuals of the dominant and subdominant patterns for each case study. To perform the t test, we analyze each multidimensional pattern of the dominant and subdominant individuals, considering that each one is defined by 25 variables.

In the first case, the dominant pattern is cluster 5 with 117,945 individuals, and the subdominant pattern is cluster 0 with 52,756 individuals. We obtain a p-value of $0.01 < 0.05$ leads to the rejection of the null hypothesis. Consequently, the means of the centroids are deemed significantly different. In the second case, the dominant pattern is cluster 0 with 5,524 individuals, and the subdominant pattern is cluster 6 with 1,900 individuals. The resulting p-value is $0.02 < 0.05$, mirroring the first scenario's outcome. Once again, the null hypothesis is rejected. In conclusion, the means of the dominant and subdominant patterns are significantly different.

Finally, we performed a final analysis to compare the dominant and non-dominant patterns in each case study (we chose the centroids of each cluster). To do this, we have carried out Factor Analysis of Mixed Data (FAMD), which is a version of Principal Components Analysis (PCA) but considering both quantitative and qualitative variables. Tables 21 and 22 show the variables that make up the principal component for cases 1 and 2, respectively.

It can be seen in both cases that the subdominant pattern considers more variables in its first principal component, and that all the variables contained in the dominant pattern are also contained in the subdominant pattern. Now, it is observed that the order of relevance of the variables in both cases between the dominant and the subdominant patterns is completely different. This, together with the t-test, reaffirms that the dominant and subdominant patterns are significantly different, therefore, each one provides us with useful information that describes the behavior of COVID-19 in each dimension analyzed.

TABLE 21. The first principal component for case 1

Case 1			
Dominant pattern		Subdominant pattern	
Variable	Component	Variable	Component
B1_1.1	0.486891	B1_1.1	0.420670
B1_13.1	0.376665	B1_13.1	0.376592
B1b_x1.1	0.517459	B1b_x1.1	0.438134
B1b_x13.1	0.443558	B1b_x13.1	0.424239
B1b_x4.1	0.365003	B1b_x4.1	0.345144
B1_4.1	0.202884	B1_4.1	0.213059
B1b_x12.1	0.158630	B1b_x12.1	0.255816
B1b_x6.1	0.194916	B1b_x6.1	0.192100
B1b_x9.1	0.110935	B1b_x9.1	0.172413
		B1b_x8.1	0.155017
		B1b_x7.1	0.118350

TABLE 22. The first principal component for case 2

Case 2			
Dominant pattern		Subdominant pattern	
Variable	Component	Variable	Component
B1_1.1	0.325541	B1_1.1	0.379275
B1_13.1	0.274914	B1_13.1	0.255738
B1b_x1.1	0.303095	B1b_x1.1	0.379278
B1b_x13.1	0.325890	B1b_x13.1	0.304434
		B1b_x12.1	0.293829
		B1b_x10.1	0.265207
		B1b_x10.1	0.321557

D. RESULT DISCUSSION

Regarding the symptomatic patterns, we see that the symptoms that normally happen relevant are (5) Fatigue, (6) Stuffy or runny nose, (7) Aches or muscle pain, and (12) Headache, with high or medium relevant (but frequently with high relevance). We also see that the symptoms after 24 hours are not relevant as well (2) Fever and (4) Difficulty Breathing, never appear or appear with low relevance. Other symptoms sometimes do not appear, or appear, but usually with low relevance. We also see that the symptomatic patterns of the cases of a single dimension with respect to the cases of crossed dimensions are practically similar. So, if we wanted to get a general symptomatic pattern of COVID-19, then it would be with the high relevance of symptoms (5), (6), (7), and (12), and the non-relevance of symptoms (2) and (4).

Regarding the correlations, we do not see any high relationship between the relevant symptoms, only in some cases a medium correlation (for example, between symptoms (5) and (7) a correlation of Pearson of 0.29 for [35-54] range age, or between symptoms (6) and (7) a correlation of Spearman of -0.29 for [55-74] range age). Also, we see that the only high correlations are between the same variables for the cases of the last 24 hours/after 24 hours (see Table 17), or between non-relevant symptoms (for example, for >74 vs Delta Variant, the case of (14) Fever and (25) Chills symptoms with 0.68 (Pearson)).

With respect to the duration in days of symptoms, the dominant symptomatic patterns have a value of around 2 days, while the subdominant ones are around 6.5 days, for all the

cases studied.

V. CONCLUSION

This study sought to carry out a comprehensive analysis of the behavior of the symptoms, based on the surveys carried out on Facebook on the behavior of COVID-19. For this, three dimensions were considered: countries, ages, and variants of the disease. During the study, based on the responses of the respondents, an estimate of the duration of the symptoms was made using a geometric distributed function, and a grouping of the information based on the behavior of this variable together with the symptoms declared in the surveys. Based on the grouping process, the dominant or subdominant patterns were identified (those groups that grouped more individuals surveyed), and from there, an analysis of their centroids was made. This allowed for an exhaustive analysis of the behavior of symptoms over time and age groups or regions.

Thus, it was possible to determine by dimension, or crossed dimensions, the behavior of the symptoms, in order to build a general symptomatic pattern of COVID-19. In our case, this pattern is made up of the presence of symptoms of Fatigue (5), Stuffy or runny nose (6), Aches or Muscle pain(7) and Headache (12), and the absence of symptoms of fever (2) and Difficulty breathing (4), with a duration of symptoms of around 2 days.

A limitation of this work is that a single data source was used for the analysis, so this analysis will have to be extended to other data sources. Also, it was assumed that the estimated duration was common for all symptoms reported at a given time.

ACKNOWLEDGMENT

This work was partially funded by the CoronaSurveys-CM grant, funded by IMDEA Networks and Comunidad de Madrid, Spain, COMODIN-CM and PredCov-CM grants funded by the Comunidad de Madrid and the European Union through the European Regional Development Fund (FEDER), grants TED2021-131264B-I00 (SocialProbing) and PID2019-104901RB-I00 funded by the Ministry of Science and Innovation, Spain and the European Union (NextGenerationEU/PRTR).

REFERENCES

[1] F. Hu, M. Huang, J. Sun, X. Zhang, and J. Liu, "An analysis model of diagnosis and treatment for covid-19 pandemic based on medical information fusion," *Information Fusion*, vol. 73, pp. 11–21, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521000415

[2] N. Munsch, S. Gruarin, J. Nateqi, T. Lutz, M. Binder, J. H. Aberle, A. Martin, and B. Knapp, "Symptoms associated with a COVID-19 infection in the general population of Vienna," *medRxiv*, 2021. [Online]. Available: https://www.medrxiv.org/content/early/2021/12/15/2021.02.24.21252426

[3] P. Guo, A. B. Ballesteros, S. P. Yeung, R. Liu, A. Saha, L. Curtis, M. Kaser, M. P. Haggard, and L. G. Cheke, "COVCOG 1: Factors predicting Cognitive Symptoms in Long COVID. A First Publication from the COVID and Cognition Study," *medRxiv*, 2021. [Online]. Available: https://www.medrxiv.org/content/early/2021/10/27/2021.10.26.21265525

[4] A. Déguilhem, J. Malaab, M. Talmatkadi, S. Renner, P. Foulquié, G. Fagherazzi, P. Loussikian, T. Marty, A. Mebarki, N. Texier, and S. Schuck, "Identifying profiles and symptoms of patients with long covid in france: Data mining infodemiology study based on social media," *JMIR Infodemiology*, vol. 2, no. 2, p. e39849, Nov 2022. [Online]. Available: https://infodemiology.jmir.org/2022/2/e39849

[5] Y. Wang, Z. Hu, Y. Feng, A. Wilson, and R. Chen, "Changes in network centrality of psychopathology symptoms between the COVID-19 outbreak and after peak," *Mol Psychiatry*, vol. 25, p. 3140–3149, 2020.

[6] A. Aghaei, R. Zhang, S. Taylor, C.-C. Tam, C.-H. Yang, X. Li, and S. Qiao, "Social Life of Females with Persistent COVID-19 Symptoms: A Qualitative Study," *International Journal of Environmental Research and Public Health*, vol. 19, no. 15, 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/15/9076

[7] J. Wu, L. Wang, Y. Hua, M. Li, L. Zhou, D. W. Bates, and J. Yang, "Trend and co-occurrence network of covid-19 symptoms from large-scale social media data: Infoveillance study," *J Med Internet Res*, vol. 25, p. e45419, Mar 2023. [Online]. Available: https://www.jmir.org/2023/1/e45419

[8] M. Taquet, Q. Dercon, S. Luciano, J. Geddes, M. Husain, and P. Harrison, "Incidence, co-occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19," *PLoS Med*, vol. 18, 2021.

[9] I. Núñez, J. Gillard, S. Fragoso-Saavedra, D. Feyaerts, L. Islas-Weinstein, A. Gallegos-Guzmán, U. Valente-García, J. Meyerowitz, J. Kelly, H. Chen, E. Ganio, A. Benkendorff, J. Flores-Gouyonnet, P. Dammann-Beltrán, J. Heredia-González, G. Rangel-Gutiérrez, C. Blish, K. Nadeau, G. Nolan, and S. Valdés-Ferrer, "Longitudinal clinical phenotyping of post COVID condition in Mexican adults recovering from severe COVID-19: a prospective cohort study," *Frontiers in Medicine*, vol. 10, 08 2023.

[10] V. Danesh, A. Arroliga, J. Bourgeois, B. M., M. M., W. J., M. T., and K. S., "Symptom Clusters Seen in Adult COVID-19 Recovery Clinic Care Seekers," *Journal of General Internal Medicine*, vol. 38, no. 442–449, 2023.

[11] R. Ghayda, J. Lee, J. Lee, D. Kim, K. Lee, S. Hong, Y. Han, J. Kim, J. Yang, A. Kronbichler, L. Smith, A. Koyanagi, L. Jacob, and J. Shin, "Correlations of Clinical and Laboratory Characteristics of COVID-19: A Systematic Review and Meta-Analysis," *Int J Environ Res Public Health*, vol. 17, 2020.

[12] Y. Zhao, D. Qu, S. Chen, and X. Chi, "Network analysis of internet addiction and depression among Chinese college students during the COVID-19 pandemic: A longitudinal study," *Computers in Human Behavior*, vol. 138, p. 107424, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563222002461

[13] O. Kerzhner, E. Berla, M. Har-Even, M. Ratmansky, and I. Goor-Aryeh, "Consistency of inconsistency in long-COVID-19 pain symptoms persistency: A systematic review and meta-analysis," *Pain Practice*, vol. n/a, no. n/a. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/papr.13277

[14] X. Cheng, H. Wan, H. Yuan, L. Zhou, C. Xiao, S. Mao, Z. Li, F. Hu, C. Yang, W. Zhu, J. Zhou, and T. Zhang, "Symptom Clustering Patterns and Population Characteristics of COVID-19 Based on Text Clustering Method," *Front. Public Health*, 2022.

[15] R. Gruber, M. Montilva, C. Weßels, G. Schlang, M. Svenj, S. Jedhoff, S. Herbrandt, and F. Mattner, " Long-term symptoms after SARS-CoV-2 infection in a cohort of hospital employees: duration and predictive factors," *BMC Infect Dis*, vol. 119, 2024.

[16] F. Kreuter, N. Barkay, A. Bilinski, A. Bradford, S. Chiu, R. Eliat, J. Fan, T. Galili, D. Haimovich, B. Kim *et al.*, "Partnering with Facebook on a university-based rapid turn-around global survey," *Survey Research Methods: SRM*, vol. 14, no. 2, pp. 159–163, 2020.

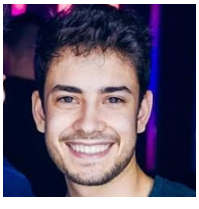
[17] J. Fan, Y. Li, K. Stewart, A. R. Kommareddy, A. Garcia, J. O'Brien, A. Bradford, X. Deng, S. Chiu, F. Kreuter, N. Barkay, A. Bilinski, B. Kim, T. Galili, D. Haimovich, S. LaRocca, S. Presser, K. Morris, J. A. Salomon, E. A. Stuart, R. Tibshirani, T. A. Barash, C. Cobb, A. Gros, A. Isa, A. Kaess, F. Karim, R. Eliat, O. E. Kedosha, S. Matskel, R. Melamed, A. Patankar, I. Rutenberg, T. Salmona, and D. Vannette, "The University of Maryland Social Data Science Center Global COVID-19 Trends and Impact Survey, in partnership with Facebook," https://covidmap.umd.edu/api.html, 2020.

[18] S. Zahi and B. Achchab, "Clustering of the population benefiting from health insurance using k-means," in *Proceedings of the 4th International Conference on Smart City Applications*, ser. SCA '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3368756.3369103

[19] G. A. Sgarro, L. Grilli, A. A. Valenzano, F. Moscatelli, D. Monacis, G. Toto, A. De Maria, G. Messina, and R. Polito, "The Role of BIA Analysis in Osteoporosis Risk Development: Hierarchical Clustering Approach," *Diagnostics*, vol. 13, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/13/2292

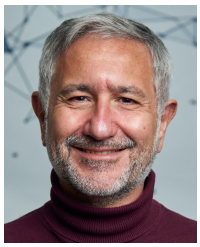
[20] Q. Zhu, X. Tang, and A. Elahi, "Application of the novel harmony search optimization algorithm for DBSCAN clustering," *Expert Systems with Applications*, vol. 178, p. 115054, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421004954
[21] L. Morales, C. Ouedraogo, J. Aguilar, C. Chassot, M. S., and K. Drira, "Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform," *Service Oriented Computing and Applications*, vol. 13, no. 199-–219, 2019.
[22] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander, "Density-based clustering validation," in *4th SIAM International Conference on Data Mining (SDM)*, 2014.
[23] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A. Boulesteix, D. Deng, and M. Lindauer, "Hyperparameter optimization: foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, 2023.
[24] Y. Quintero, D. Ardila, E. Camargo, F. Rivas, and J. Aguilar, "Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables," *Computers in Biology and Medicine*, vol. 134, p. 104500, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521002948
[25] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.
[26] R. A. Ghayda, J. Lee, J. Y. Lee, D. K. Kim, K. H. Lee, S. H. Hong, Y. J. Han, J. S. Kim, J. W. Yang, A. Kronbichler, L. Smith, A. Koyanagi, L. Jacob, and J. I. Shin, "Correlations of clinical and laboratory characteristics of covid-19: A systematic review and meta-analysis," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, 2020. [Online]. Available: https://www.mdpi.com/1660-4601/17/14/5026
[27] T. Chen, M. Xu, J. Tu, H. Wang, and X. Niu, " Relationship between Omnibus and Post-hoc Tests: An Investigation of performance of the F test in ANOVA," *Shanghai Arch Psychiatry*, vol. 25, pp. 60–64, 2018.

DIEGO JAVIER BENITO received the B.S. degree in Computer Science from the University of Alcalá, Alcalá de Henares, Spain, in 2024. Currently, he is pursuing an M.S. in Data Analytics and Big Data at the University of Alcalá. He worked as an intern at IMDEA Network, Spain, during 2022-2023. Since 2023, he work as a research engineer in the Global Computing group at IMDEA Network. His research interests include machine learning, deep learning, and expert systems, with a particular focus on explainability analysis of artificial intelligence.



JESÚS RUFINO ROBLES has a Master's degree in Mathematical Engineering and a Bachelor's degree in Mechanical Engineering, offering expertise in data analytics, machine learning, and industrial engineering. Currently serving as a Data Scientist at Continental in Hannover, Germany, driving data analytics and reporting initiatives for global industrial operations. Prior experience includes roles as a Research Engineer at IMDEA Networks in Madrid, where I conducted advanced research and statistical analysis for telecom networks, and as an Engineer at IMP Cloner in Granada, leading Research and Development studies and contributing to various engineering projects. Skilled in predictive modeling, statistical analysis, and system administration, with proficiency in programming languages and AWS cloud technologies.

JUAN RAMÍREZ received a B.S. in electrical engineering, an M.S. in biomedical engineering, and a Doctor degree in applied sciences from the Universidad de Los Andes, Venezuela, in 2002, 2007, and 2017, respectively. From 2004 to 2019, he was an Associate Professor at the Electrical Engineering Department at Universidad de Los Andes, Venezuela. In addition, He spent a postdoctoral internship at the High Dimensional Signal Processing (HDSP) group at the Universidad Industrial de Santander, Colombia. He also was a Got Energy Talent GET-COFUND Marie Skłodowska-Curie Actions fellow at the Department of Computer Science from Universidad Rey Juan Carlos, Spain, from 2019 to 2021. He is currently a postdoctoral researcher at the IMDEA Networks Institute in Spain.



DR. ANTONIO FERNÁNDEZ ANTA is a Research Professor at IMDEA Networks. Previously he was a Full Professor at the Universidad Rey Juan Carlos (URJC) and was on the Faculty of the Universidad Politécnica de Madrid (UPM), where he received an award for his research productivity. He was a postdoc at MIT from 1995 to 1997, and spent sabbatical years at Bell Labs Murray Hill and MIT Media Lab.

He has more than 30 years of research experience, and more than 250 scientific publications. He has been awarded the Premio Nacional de Informática "Aritmel" in 2019 and is a Mercator Fellow of the SFB MAKI in Germany since 2018. Recently his scientific papers have received awards such as the 2nd Best Paper Award (Honorary Mention) at the AAAI 2024 Social Impact Track, the Mario Gerla Best Paper Award at MedComNet 2022 or the Best Teaser Award at WoWMoM 2021.

He was the Chair of the Steering Committee of DISC and has served in the TPC of numerous conferences and workshops. He received his M.Sc. and Ph.D. from the University of SW Louisiana in 1992 and 1994, respectively. He completed his undergraduate studies at the UPM, having received awards at the university and national level for his academic performance. He is a Senior Member of ACM and IEEE.

DR. JOSE AGUILAR (Senior Member, IEEE) received the Systems Engineer degree from the Universidad de Los Andes, Mérida, Venezuela, in 1987; the M.Sc. degree in computer science from Universite Paul Sabatier-France, in 1991; and the Ph.D. degree in computer science from Universite Rene Descartes-France, in 1995. He has carried out several Postdoctorals at the Department of Computer Science, University of Houston, 199-2000; the Laboratoire d'analyse et d'architecture des systèmes (LAAS), CNRS, Toulouse, France, 2011-2012; and the University of Alcala, Spain, 2020-2022, with a Marie Skłodowska-Curie Fellowship.

He is currently a Senior Research at the IMDEA Networks Institute. He is also a member of the Mérida Science Academy, Full Professor at the Department of Computer Science, Universidad de Los Andes, and was member of the IEEE CIS Technical Committee on Neural Networks. He has published more than 650 articles and ten books, in the field of parallel and distributed computing, computer intelligence, and science and technology management. His research interests include artificial intelligence, semantic mining, big data, emerging computing, and intelligent environments.