Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology

Chapter · January 2021

DOI: 10.1007/978-3-030-93733-1_29

3 authors, including:

Aidan Cooper
HEOR
3 PUBLICATIONS **46** CITATIONS

SEE PROFILE

Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology

Aidan Cooper[1 [0000−0002−8113−3897]], Orla Doyle[1 [0000−0002−6964−0140]], and Alison Bourke[1 [0000−0002−0005−9016]]

IQVIA, London, UK
{aidan.cooper, orla.doyle, alison.bourke}@iqvia.com

Abstract. Subgroup discovery is a data mining technique that attempts to find interesting relationships between different instances in a dataset with respect to a property of interest. Cluster analysis is a popular method for extracting homogeneous groups from a heterogeneous population, however, it often yields results that are challenging to interpret and action. In this work, we propose a novel, multi-step clustering methodology based on SHAP (SHapley Additive exPlanation) values and dimensionality reduction, for the purpose of subgroup discovery. Our method produces well-separated clusters that can be readily differentiated by simple decision rules, to yield interpretable subgroups in relation to a target variable. We illustrate our approach using self-reported COVID-19 symptom data across 2,479 participants who tested positive for COVID-19, resulting in the identification of 16 distinct symptom presentations. Future work will investigate common demographic and clinical features exhibited by each cluster cohort, and map clusters to outcomes to better understand the clinical presentation, risk factors and prognosis in COVID-19, as a timely and impactful application of this methodology.

Keywords: COVID-19 · Clustering · Subgroup Discovery · SHAP

1 Introduction

One of the main goals of data mining is to discover meaningful patterns in data that can be interpreted and understood in order to extract knowledge. Subgroup discovery (SD) is a broadly applicable technique that aims to find interesting relationships between different instances in a dataset with respect to a property of interest [2, 12, 11]. Clustering is one approach to SD that involves partitioning unlabelled data into clusters or groups based on their similarity [21, 25]. It has wide-ranging applications that span subgroup analysis, outlier detection, data visualisation and numerous others. A routine problem for the clustering practitioner, however, is deriving meaningful interpretations of clustered data that can be understood and actioned.

In clinical data science, the diverse symptom presentations of many diseases are often poorly understood, and symptom cluster research is considered to be an underexplored field that could provide new targets for interventions [19]. Recently, clustering of COVID-19 symptoms has been a topic of significant interest

since the onset of the global pandemic, due to the variable symptom presentations that have been observed [9]. There have been various efforts to group COVID-19 patients based on their symptoms, although these segmentations typically have significant symptom overlap, such that the differences between clusters can be difficult to recognise [24, 22]. Previous work studying infectious disease epidemics has shown that self-reported symptom data mined from web-based surveillance systems and social media platforms can be used to monitor and detect symptom trends [13, 14, 10].

In this paper, we present a multi-step clustering methodology that is specifically designed to produce interpretable clusters in a two-dimensional space, for the purpose of SD. We describe these clusters with simple yet highly discriminative decision rules, that are human-readable and appropriate for use in practical, real-world applications.

1.1 Related Work

Our work draws upon established concepts in SD, as well as a variety of recent supervised and unsupervised machine learning techniques in application to clustering. Numerous SD algorithms have been proposed, including some specifically for clinical data analysis, which typically involve statistics-based methods for identifying 'interesting' subgroups that possess unusual distributional properties [11, 7, 4]. However, these often produce large and overlapping rule sets [2]. Hierarchical clustering has been shown as a way to mitigate high instance-overlap in sets of rules [21]. We augment this approach by conducting significant pre-processing of the data to improve the effectiveness of clustering and subsequent SD.

Central to our methodology is the use of SHAP values as the basis for clustering. This idea is conceptually similar to other approaches that attempt to remove noise-information from datasets such that only the structure important to clustering remains [23]. In their paper for efficient tree ensemble SHAP value computation, Lundberg et al. introduce the concept of "supervised clustering", where instead of applying clustering directly to feature data in a fully-unsupervised manner, clustering is run on feature SHAP values [16]. This confers a number of interesting and desirable properties, most notably weighting features by a measure of importance that emphasises the most informative features whilst minimising the effect of fluctuations in feature values that have little impact on the outcome of interest. SHAP also serves as a pre-processing step that rescales feature data into common units that are the same as the output of the supervised prediction model. Moreover, it is acknowledged as a tool for interpretable clustering by Molnar [20]. Gramegna and Giudici compared clustering on the data directly versus the SHAP values to characterise the buying behaviours of insurance customers and found that clusters were better differentiated using the SHAP-based approach [8]. While our methodology has some similarities, we augment the performance and interpretability by reducing the SHAP values to two-dimensions prior to clustering using Uniform Manifold Approximation and Projection (UMAP) which has been proposed as an effective pre-processing step

for boosting the performance of density-based clustering [18]. UMAP's major advantage is that it preserves the local and global structure of the data, which in our methodology ensures that clusters close in space are close in characteristics, further enhancing the interpretability of our approach by enabling subgroup visualisation.

Finally, rather than characterise clusters using descriptive statistics based on the outcome and feature data, we construct discriminative decision-rules that identify and differentiate the clusters, forming our subgroup descriptions. Decision rules are a popular supervised machine learning technique where interpretability is paramount, and are fundamental to most SD algorithms [20]. They have also been applied successfully in clustering applications as a way to segment data [3]. We show that the dense, well-separated clusters generated in the previous steps of our methodology yield a set of complementary decision rules that are highly discriminative in identifying and differentiating our data, whilst adhering to human-readable levels of complexity.

1.2 Our Contributions

The primary contribution of our work is marrying together multiple sub-fields of clustering into a cohesive methodology for SD. Our approach produces highly discriminative subgroup rules, addressing the commonly encountered challenge in SD - particularly in clinical settings - of large and overlapping rule sets that require subsequent manual filtering and adjustment by experts. Furthermore, the ability to naturally visualise clusters distinguishes our approach from traditional statistics-based SD techniques, addressing another recognised challenge of comprehensive subgroup visualisation [2].

In application to COVID-19, we've demonstrated the feasibility of our method for producing an intuitive two-dimensional mapping of the symptom space, comprised of meaningful regions of symptomatology and well-characterised individual clusters, with respect to not only the presence or absence of different symptoms, but their severity also.

2 Methods

2.1 COVID-19 Active Research Experience (CARE)

We obtained data from participants in a community-based COVID-19 registry known as CARE (`https://www.helpstopcovid19.com/`). This registry is based in the US and is open to anyone who believes they have been exposed to COVID-19. Via a web platform, participants can report their experience, including symptoms and severity, as well as risk factors, and treatment of COVID-19. The registry also captures any results of viral COVID-19 testing. To date, over 20,000 people have enrolled into CARE, and for this analysis we used symptom reports from US participants who had COVID-19 test results and entered symptom information between 30 July 2020 and 19 January 2021.

Participants recorded the presence or absence of 21 symptoms (including *fatigue*, *cough*, *fever* and *decreased smell*) and were also able to grade the severity of each of their symptoms as "Very Mild", "Mild", "Moderate", or "Severe". For each of the 21 symptoms, we combined the presence/absence flag and severity rating into a single ordinal score between 0-5, where 0 corresponds to the absence of a symptom, 1 means the symptom was reported but not rated, and scores of 2-5 correspond to the symptom being reported and rated on the 4-point severity scale with 2 being "Very Mild" up to 5 as "Severe". In addition to the 21 individual symptom variables, we also include an overall count of the number of symptoms reported by the participants in our analysis.

2.2 Clustering and Subgroup Discovery Methodology

Our clustering methodology comprises a sequence of steps, shown in overview in Figure 1, and outlined in depth below. In summary, we use explainable supervised machine learning to emphasise contributions that help discriminate between classes (e.g., positive vs negative COVID-19 test), i.e., we transform the data to represent the discriminatory question of interest. We then embed the participants who tested positive for COVID-19 in a two-dimensional space, where we look for structure using clustering. Finally, we adopt a rules-based approach for describing and differentiating these clusters, yielding subgroups of COVID-19 symptomatology.

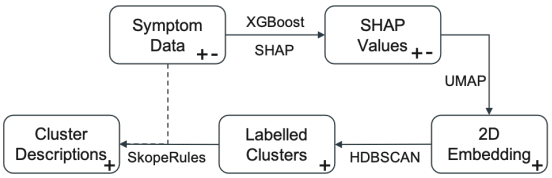

Fig. 1. Conceptual diagram of our multi-step subgroup discovery methodology. Each step is annotated with + and - icons to denote the inclusion of COVID-positive and COVID-negative participants, respectively.

To evaluate this approach of clustering in the discriminative space, we compare the results of our methodology against an equivalent clustering procedure that bypasses the supervised machine learning and SHAP stages, and instead clusters directly on the embedded symptom data in a manner akin to conventional clustering analysis.

COVID-19 Test Outcome Prediction Model The role of the supervised machine learning model is primarily to serve as a means for obtaining SHAP (SHapley Additive exPlanations) values, which describe the discriminative profile for participants with a positive test for COVID-19.

For supervised machine learning, an XGBoost classification algorithm was trained to predict the COVID-19 test outcome for the 4,063 tested participants [6]. We selected XGBoost for its non-parametric properties, ability to learn interactions between variables, and its compatibility with the SHAP methodology that provides a participant-level view of feature importance for predicting a positive COVID-19 test.

The feature set for XGBoost comprised 21 symptom scores and the count of unique symptoms as input variables, with the outcome of the COVID-19 test (positive encoded as 1, negative as 0) used as the classification label. Hyperparameters were selected by 10-fold cross-validation, using a random search methodology. Cross-validated scores were calculated for the Area Under the Curve of the Receiver Operating Characteristic (AUCROC) and accuracy.

Symptom Variable SHAP Values SHAP values were calculated for the trained XGBoost model, to determine the contribution of each variable towards the model's COVID-19 test outcome predictions for each individual instance [15]. In this analysis, instances where a variable has a SHAP value greater than zero indicates that the variable is associated with a prediction of a positive COVID-19 test result, and SHAP values less than zero indicate that the variable is associated with a negative COVID-19 test prediction. SHAP values can be thought of as a latent representation of the symptom data that emphasises the variables that are most influential for the COVID-19 test outcome predictions. The resulting SHAP values for COVID-positive participants can be clustered in terms of the similarity of the drivers that predict a positive COVID-19 test. In doing so, the clustering process is encouraged to focus on the symptom variables that are most pertinent to COVID-19, rather than consider all variables equally and potentially group participants based on similarity across irrelevant factors.

SHAP values were computed using the TreeExplainer explanation method under the *interventional* feature perturbation approach [15].

Dimensionality Reduction with UMAP Uniform Manifold Approximation and Projection (UMAP) was used to reduce the COVID-positive SHAP data to two dimensions [18]. The use of UMAP has two purposes: (i) as a pre-processing step for clustering in order to enhance the performance [1] and (ii) to aid data visualisation. One notable advantage of using UMAP for dimensionality reduction rather than alternatives such as t-SNE, is that it preserves the local and global structure of the data. That is, individual data points are maximally similar within their neighbourhood as well more similar to local neighbourhoods than those that are more distant.

For our use case, the objective is to characterise and segment COVID-19 positive participants and therefore we focus on only these participants in this and subsequent steps. An alternative formulation of the problem – which may or may not be more appropriate for other similar clinical SD tasks – would be to retain both COVID-19 positive and negative participants throughout the analysis. This would ultimately produce clusters that not only segment subgroups of COVID-19

positive participants from each other, but also distinguish them from COVID-19 negative participants. This could be preferred in settings where the intended application is for diagnosis.

The use of UMAP in this COVID-19 setting means that in addition to deriving subgroups of participants that display similar symptomatology, we also obtain regions that encompass multiple clusters with related symptom characteristics.

The UMAP embedding was computed for two components, using a local neighbourhood (*n neighbours*) of 45 data points (three times the default value), and a *minimum distance* between embedded points of zero, enabling data points to be tightly grouped to support the formation of local clusters. This higher than default *n neighbours* value encourages UMAP to produce more general, global neighbourhoods, rather than localised granular structure. Similarly, setting the *minimum distance* value to zero encourages the formation of densely packed, well-separated clusters of embedded data points [18]. These characteristics serve to optimise the clustering results in the next stage.

Density-Based Clustering with HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was used to cluster the two-dimensional UMAP embedding [17]. The HDBSCAN algorithm was selected as it does not require the number of clusters to be specified *a priori* but instead requires that the users specify the minimum size of cluster, which is potentially a more intuitive parameter to work with. Additionally, this approach doesn't force all data points to be assigned to clusters, i.e., samples can remain unassigned to all clusters. Finally, HDBSCAN has been shown to work well with UMAP [1].

HDBSCAN parameters were selected by grid search using ranges that spanned 2-200 for *minimum samples*, 2-200 for *minimum cluster size*, and 0-8 for *cluster selection epsilon*. The set of parameters chosen were those that produced clusters that were maximally separated in the embedding space as measured by the average silhouette coefficient across all instances, subject to the constraint that no more than 3% (an arbitrarily selected threshold) of instances were unassigned by HDBSCAN. Values of 10 for *minimum samples*, 40 for *minimum cluster size* and 0.5 for *cluster selection epsilon* yielded optimal results that maximised the average silhouette coefficient [5].

Rules-Based Cluster Descriptions (Subgroups) A key challenge in the use of clustering methods is the interpretation of the results. We propose to augment the interpretability of our approach by learning rules that identify clusters with high precision and coverage, thus describing our subgroups.

We use the SkopeRules package to describe and differentiate clusters of COVID-positive participants using a one-vs-all methodology [20]. Interpretable rules were found based on the underlying symptomatology data - not the computed SHAP values - for each cluster in turn, by labelling it positively and all

other clusters negatively. The resulting rule set serves as our COVID-19 symptomatology subgroup descriptions.

SkopeRules was set to use a *maximum depth* of four, to ensure the rules met the desired level of simplicity of having no more than four terms. *Minimum precision* and *minimum recall* were set at 70%, which was the highest threshold that still enabled rules to be identified for every cluster. For most clusters, the performance of the rules greatly surpassed this.

3 Results

In the following sections, we present results of the clustering and subgroup discovery methodology. We compare the performance against an equivalent procedure applied directly to the unmodified symptom data (Figure 6 and Table 2).

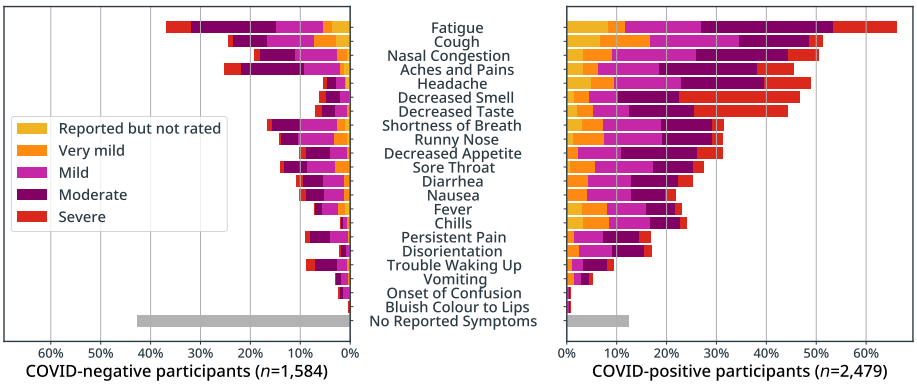

Fig. 2. Reported symptom prevalence in participants tested for COVID-19, ranked by overall prevalence (n=4,063).

4,063 participants met the criteria for inclusion (COVID-19 test and entry of symptom information between 30 July 2020 and 19 January 2021). Of these 4,063 participants, 2,479 (61.0%) received a positive COVID-19 test.

The symptom variables are visualised in Figure 2 for both COVID-positive and negative participants. Non-zero symptom scores are somewhat normally distributed, with ratings of "Mild" or "Moderate" being the most commonly reported severities for all symptoms, with the exception of *decreased smell* and *decreased taste* for which "Severe" was the most frequent severity rating. Scores of 1 (symptom reported but severity not rated) are relatively infrequent. *Fatigue*, *cough*, and *nasal congestion* were the most common symptoms (n=2,222, 1,659, 1,556), whereas *bluish colour to lips*, *onset of confusion*, and *vomiting* were the least reported symptoms (n=24, 62, 180). The mean number of symptoms reported for the 3,080 participants who reported at least one symptom was 6.71

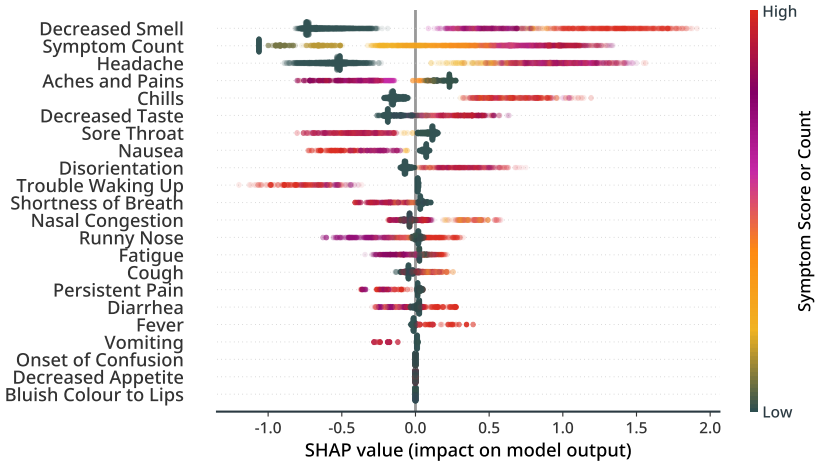

Fig. 3. SHAP values summary plot for the 22 symptom variables used for clustering, ranked by their mean absolute value across all participants with a test result. *Decreased smell*, *symptom count*, and *headache*, were the three most influential variables for the XGBoost model's predictions, whereas *bluish colour to lips* was the least informative variable.

(SD 4.36). 983 participants reported no symptoms, of whom 307 (31.2%) had positive COVID-19 tests.

3.1 COVID-19 Test Outcome Prediction Model

The XGBoost model achieved mean cross-validated scores of 0.825 for Area Under the Curve of the Receiver Operating Characteristic (AUCROC) and 77.1% for accuracy. Qualitatively speaking, these scores indicate that there is a sufficient signal in the symptom scores alone to differentiate between positive and negative participants, albeit with moderate levels of predictive performance.

3.2 Symptom Variable SHAP Values

Figure 3 shows the SHAP values derived from the XGBoost model. For each of the 22 symptom variables, each participant is represented as an individual data point. The symptoms are ordered according to their mean absolute SHAP value (contribution to prediction) and the colour scale represents the symptom severity level. *Decreased smell*, *headache*, and *aches and pains* were the three most informative symptoms for predicting the COVID-19 test results, as ranked by mean absolute SHAP value. Examining the colours of the individual points reveals that in the first two cases, higher symptom scores are typically associated with positive test result predictions, whereas scores of zero encourage the opposite. A similar trend is seen for other highly ranking symptoms, such as

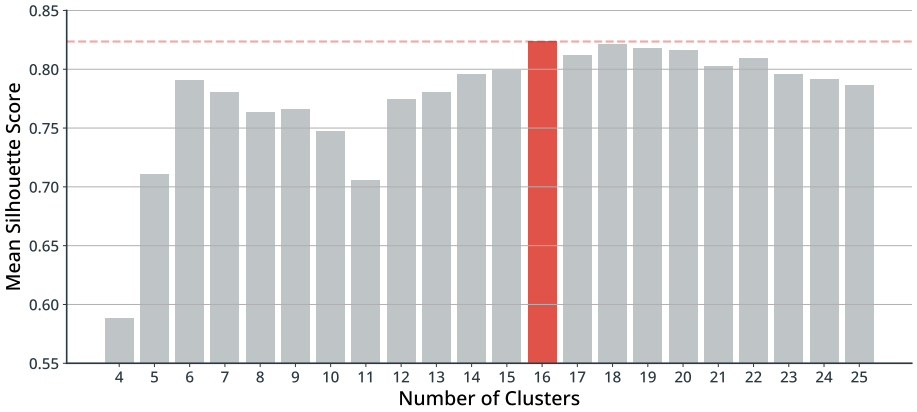

Fig. 4. Maximum mean silhouette scores observed for different counts of clusters during the HDBSCAN parameter grid search. The global maximum is at 16 clusters, which we visualise in Figure 5 and define subgroups for in Table 1.

chills and *decreased taste*, although the opposite is observed for *aches and pains*, *sore throat*, and *nausea*. *Runny nose* and *fatigue* are examples of a mixed cases, where high scores can encourage predictions of positive or negative test outcomes dependent on the presentation of other symptoms.

Comparison of Figure 3 with Figure 2 indicates some correlation between the overall prevalence of symptoms and their influence over the XGBoost model's predictions, although the symptoms that top the SHAP rankings are those that are disproportionately represented in one of the test outcome groups. *Decreased smell* is one such symptom, and has been recognised in other studies as one of the strongest predictors of COVID-19 [10]. Another notable example is *chills*, which ranks 15$^{\text{th}}$ by prevalence yet 4$^{\text{th}}$ by influence.

The *symptom count* ranks 2$^{\text{nd}}$ most informative and shows a linear relationship with test outcome prediction, where the absence of any symptoms leads to a negative test prediction and the presence of many symptoms leads to a positive test prediction.

As is to be expected, the top three most informative variables feature prominently within the rules that describe the resulting clusters.

3.3 Dimensionality Reduction and Clustering

Table 1 shows the counts, descriptions, and rules of the COVID-positive clusters (visualised in Figure 5). The grid search of HDBSCAN parameters found that a maximum mean silhouette score of 0.822 was achieved by a 16-cluster segmentation. HDBSCAN is not constrained to assign each data points to clusters, and 62 (2.5%) of participants were unassigned. Counts of participants in each cluster vary between 315 in the largest cluster (2, *"headache without chills or decreased smell"*) and 54 in the smallest (8, *"chills without headache or decreased smell"*).

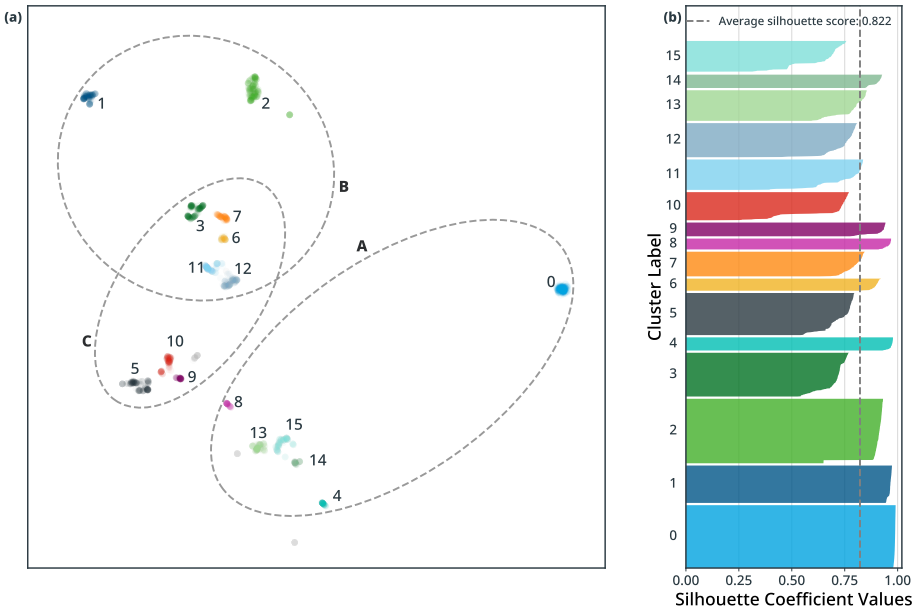

Fig. 5. (a) Visual representation of the clustered, two-dimensional UMAP embedding of COVID-positive participant SHAP values. Regions **A**, **B**, and **C** - identified by examining the rules in Table 1, and selected for their differences with respect to the two most influential symptoms - surround distinct spatial areas of key symptom combinations: no *decreased smell* or *headache* (**A**); *headache* (**B**); *decreased smell* (**C**). (b) Silhouette plot of the 16 identified clusters, with micro-average silhouette score 0.822.

Although it was found that 16 clusters was the optimal number for maximising the mean silhouette score, from the grid search results in Figure 4, it can be seen that a local maximum also exists for 6 clusters. If a smaller number of coarser clusters with higher instance counts was desired for reasons of clinical utility, this alternative segmentation could be characterised further. This illustrates the flexibility of our methodology for producing subgroups at the desired level of granularity, by subsuming or dividing clusters within the two-dimensional symptom space through tuning of the HDBSCAN parameters.

Figure 6 shows the equivalent results after clustering on a two-dimensional UMAP embedding obtained directly from the underlying symptom data, bypassing the SHAP value computation stage. Clustering with HDBSCAN on this embedding is highly unstable with respect to small changes in parameters, with results fluctuating between two and over twenty distinct clusters. After careful fine-tuning of parameters, this procedure produces seven clusters with a considerably lower mean silhouette score of 0.140. The majority (68.1%) of the data is assigned to a single, amorphous cluster, largely comprised of data points with negative silhouette coefficients. There were 234 participants (9.4%) without cluster membership, and the five smallest clusters account for only 9.4% of the total

data, further evidencing this procedure's limited effectiveness for deriving an interpretable segmentation of this data.

3.4 Rules-Based Cluster Descriptions

The rules, comprised of between one and four terms, had mean weighted scores of 95.3% precision and 97.9% recall (Table 1). In this context, precision refers to the percentage of participants identified by a rule that belong to its respective cluster. Similarly, recall refers to the percentage of participants belonging to a cluster that are correctly identified by its respective rule.

By comparison, the mean weighted precision and recall of the decision rules derived for the seven clusters obtained via clustering directly on the symptom data are 99.1% and 52.5%, respectively (Table 2). In order to find rules for all clusters, *minimum precision* and *minimum recall* thresholds need to be lowered to 40%. The reduced discriminative performance as compared to the 16-cluster SHAP-based variant, and the imbalanced distribution of participants across the clusters, means that this set of rules is of limited practical utility, where the goal is to identify well-separated groups of symptom presentations that can be characterised meaningfully.

Closer inspection of the rules and their constituent terms shows that they're highly complementary as a set, cleanly segmenting groups of instances by the presence or absence of specific symptoms and their differing levels of severity. The second largest cluster, 0, corresponds to the 12.4% of COVID-positive participants with no reported symptoms.

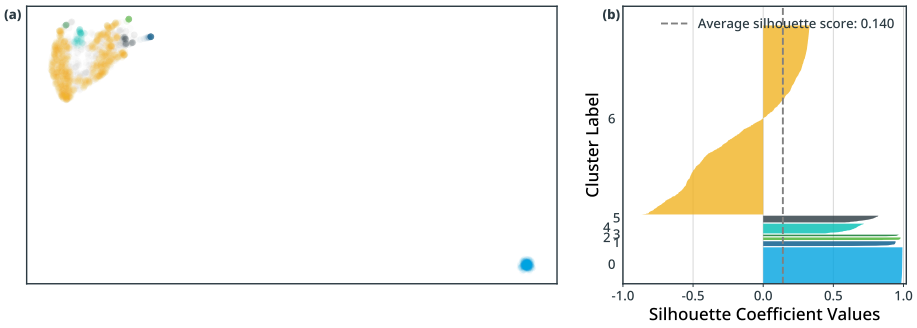

Fig. 6. (a) Visual representation of the clustered, two-dimensional UMAP embedding of COVID-19-positive participants' symptom data (as opposed to SHAP values). (b) Silhouette plot of the 7 identified clusters, with micro-average silhouette score 0.140.

A benefit of UMAP not afforded by related techniques such as t-SNE, is that clusters close together in the embedded space are more similar than those that are far apart. We illustrate this with three *regions*, A, B, C, which surround clusters that share similar characteristics with respect to the two most influential

Table 1. Subgroup descriptions and rules for COVID-positive participants, and corresponding rule precision and recall scores.

Cluster	Count (%)	Description	Rule	Precision	Recall
0	307 (12.4%)	Asymptomatic	symptom count = 0	100.0%	100.0%
1	183 (7.4%)	Headache and chills, without decreased smell	headache ≥ 1 chills ≥ 1 decr. smell = 0	100.0%	100.0%
2	315 (12.7%)	Headache, without chills or decreased smell	headache ≥ 1 chills = 0 decr. smell = 0	100.0%	100.0%
3	214 (8.6%)	Headache, severe decreased smell, no/unrated trouble waking up, no chills	headache ≥ 1 trouble waking up ≤ 1 chills = 0 decr. smell = 5	100.0%	100.0%
4	65 (2.6%)	1 symptom that isn't decreased smell or mod./severe fatigue	symptom count = 1 decr. smell = 0 fatigue ≤ 3	77.1%	98.5%
5	206 (8.3%)	Severe decreased smell, without headache	decr. smell = 5 headache = 0	100.0%	99.5%
6	60 (2.4%)	Severe decreased smell, rated trouble waking up, and headache	decr. smell = 5 headache ≥ 1 trouble waking up ≥ 2	100.0%	100.0%
7	121 (4.9%)	Severe decreased smell, headache, chills, and no/unrated trouble waking up	decr. smell = 5 headache ≥ 1 chills ≥ 1 trouble waking up ≤ 1	100.0%	100.0%
8	54 (2.2%)	Chills without headache or decreased smell	chills ≥ 1 headache = 0 decr. smell = 0	100.0%	100.0%
9	68 (2.7%)	Non-severe decreased smell, mild-severe aches and pains, without headache	$1 \leq$ decr. smell ≤ 4 aches and pains ≥ 3 headache = 0	91.9%	100.0%
10	138 (5.6%)	Non-severe decreased smell, no/unrated/very mild aches and pains without headache	$1 \leq$ decr. smell ≤ 4 aches and pains ≤ 2 headache = 0	79.6%	96.4%
11	149 (6.0%)	Non-severe decreased smell, with headache and chills	$1 \leq$ decr. smell ≤ 4 headache ≥ 1 chills ≥ 1	98.6%	96.0%
12	167 (6.7%)	Non-severe decreased smell, headache, without chills	$1 \leq$ decr. smell ≤ 4 headache ≥ 1 chills = 0	95.4%	98.8%
13	150 (6.1%)	3+ symptoms that aren't headache, decreased smell, or mild-severe aches and pains	symptom count ≥ 3 headache = 0 decr. smell = 0 aches and pains ≤ 2	73.4%	99.3%
14	66 (2.7%)	2 symptoms that aren't headache or decreased smell	symptom count = 2 headache = 0 decr. smell = 0	75.9%	100.0%
15	60 (2.4%)	Mild-severe aches and pains, without headache, decreased smell, or rated chills	aches and pains ≥ 3 headache = 0 chills ≤ 1 decr. smell = 0	100.0%	77.9%
N/A	62 (2.5%)	unassigned	-	-	-

symptom variables, *decreased smell* and *headache* (Figure 5). Clusters within Region A explicitly exclude both symptoms, clusters within Region B explicitly include *headache*, and clusters within Region C explicitly include *headache*. Clusters within the overlap of Regions B and C consist of both *decreased taste* and *headache*. Trivially, Region A also encompasses the most universally distant cluster, 0, which corresponds to participants who reported no symptoms. This aligns with clinical intuition that symptomless patients should be far removed from the symptomatic groups. Similar regions can be identified for other prominent symptoms such as *chills*.

Table 2. Cluster results derived from symptom data (as opposed to SHAP values). The *Cluster Descriptions* are interpretations of the *Cluster Rules* derived by SkopeRules, translated into natural language. The *Cluster Rules* are combinations of terms that dictate the presence or absence, and severity of different symptoms. *Rule Precision* quantifies the percentage of COVID-19-positive participants identified by the *Cluster Rule* that truly belong to the respective cluster. *Rule Recall* quantifies the percentage of COVID-19-positive participants that truly belong to a cluster that are successfully identified by its corresponding *Cluster Rule*.

Cluster	Count (%)	Description	Rule	Precision	Recall
0	326 (13.2%)	Asymptomatic	symptom count = 0	100.0%	94.2%
1	44 (1.8%)	Only fatigue	symptom count = 1 fatigue $\geq$ 2	100.0%	61.4%
2	24 (1.0%)	Only cough	symptom count = 1 cough $\geq$ 1	100.0%	62.5%
3	15 (0.6%)	2-3 symptoms that include decreased smell and (very) mild decreased taste	2 $\geq$ symptom count $\leq$ 3 decr. smell $\geq$ 1 2 $\geq$ decr. taste $\leq$ 3	68.2%	100.0%
4	88 (3.5%)	2-6 symptoms that include rated decreased smell and nasal congestion, but not rated decreased taste	2 $\geq$ symptom count $\leq$ 6 decr. smell $\geq$ 2 nasal congestion $\geq$ 2 decr. taste $\leq$ 1	97.7%	47.7%
5	61 (2.5%)	1-3 symptoms that include rated shortness of breath, but not rated nasal congestion	1 $\geq$ symptom count $\leq$ 3 shortness of breath $\geq$ 2 nasal congestion $\leq$ 1	88.6%	50.8%
6	1687 (68.1%)	3+ symptoms that include moderate-severe decreased smell, rated decreased taste, but not severe trouble waking up	symptom count $\geq$ 3 decr. smell $\geq$ 4 decr. taste $\geq$ 2 trouble waking up $\leq$ 4	99.6%	44.0%
N/A	234 (9.4%)	Unassigned	-	-	-

4 Conclusions and Future Directions

Our methods augment clustering approaches for SD by incorporating SHAP values as a latent representation of the data that is more amenable to clustering with respect to the target of interest. We have shown that compared to a standard methodology that clusters directly on the feature data, our approach yields subgroups that are better distributed and described by interpretable rules with superior discriminative performance. In doing so, we address a common pitfall of SD, by generating a compact and highly accurate rule set with minimal overlap.

We also demonstrate the utility of visualising and identifying subgroups derived from a 2D embedding, as a means for understanding not only the commonalities of instances within a subgroup, but also the similarities of subgroups proximally located in the 2D space. This addresses another challenge in SD of comprehensively visualising subgroups, and for practical purposes, can help elucidate 'regions' that encompass multiple subgroups and encapsulate broader trends in the data.

Having established the potential of this methodology on this experimental dataset, in future work we intend to investigate the different demographic and clinical features exhibited by the clusters produced by this methodology, and map clusters to outcomes to better understand the clinical presentation, risk factors and prognosis in COVID-19. We believe this novel approach to SD, based on a multi-step supervised clustering pipeline, produces a cleaner and better characterised segmentation of COVID-19 symptomatology than a conventional clustering approach, and that this methodology can be applied more widely to other SD problems where interpretability and differentiability of clusters is paramount.

References

1. Allaoui, M., Kherfi, M.L., Cheriet, A.: Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In: El Moataz, A., Mammass, D., Mansouri, A., Nouboud, F. (eds.) Image and Signal Processing. pp. 317–325. Springer International Publishing, Cham (2020)
2. Atzmueller, M.: Subgroup discovery. WIREs Data Mining and Knowledge Discovery **5**(1), 35–49 (2015). https://doi.org/https://doi.org/10.1002/widm.1144, https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1144
3. Barbado, A., Corcho, Ó., Benjamins, R.: Rule Extraction in Unsupervised Anomaly Detection for Model Explainability: Application to OneClass SVM. arXiv e-prints arXiv:1911.09315 (2019)
4. Belfodil, A., Belfodil, A., Bendimerad, A., Lamarre, P., Robardet, C., Kaytoue, M., Plantevit, M.: Fssd - a fast and efficient algorithm for subgroup set discovery. In: 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 91–99 (2019). https://doi.org/10.1109/DSAA.2019.00023
5. Chen, G., Jaradat, S., Banerjee, N., Tanaka, T., Ko, M., Zhang, M.: Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. Statistica Sinica **12**, 241–262 (2002)

6. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. arXiv e-prints arXiv:1603.02754 (2016)
7. Esnault, C., Gadonna, M.L., Queyrel, M., Templier, A., Zucker, J.D.: Q-finder: An algorithm for credible subgroup discovery in clinical data analysis — an application to the international diabetes management practice study. Frontiers in Artificial Intelligence **3**, 83 (2020). https://doi.org/10.3389/frai.2020.559927, https://www.frontiersin.org/article/10.3389/frai.2020.559927
8. Gramegna, A., Giudici, P.: Why to buy insurance? an explainable artificial intelligence approach. Risks **8**(4) (2020). https://doi.org/10.3390/risks8040137, https://www.mdpi.com/2227-9091/8/4/137
9. Grant, M.C., Geoghegan, L., Arbyn, M., Mohammed, Z., McGuinness, L., Clarke, E.L., Wade, R.G.: The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries. PLoS One **15**(6), e0234765 (2020)
10. Güemes, A., Ray, S., Aboumerhi, K., Desjardins, M.R., Kvit, A., Corrigan, A.E., Fries, B., Shields, T., Stevens, R.D., Curriero, F.C., Etienne-Cummings, R.: A syndromic surveillance tool to detect anomalous clusters of covid-19 symptoms in the united states. Scientific Reports **11**(1), 4660 (2021). https://doi.org/10.1038/s41598-021-84145-5, https://doi.org/10.1038/s41598-021-84145-5
11. Helal, S.: Subgroup discovery algorithms: A survey and empirical evaluation. Journal of Computer Science and Technology **31**, 561–576 (2016). https://doi.org/10.1007/s11390-016-1647-1
12. Herrera, F., Carmona, C.J., González, P., Del Jesus, M.J.: An overview on subgroup discovery: Foundations and applications. Knowledge and Information Systems **29**, 495–525 (2011). https://doi.org/10.1007/s10115-010-0356-2
13. Kalimeri, K., Delfino, M., Cattuto, C., Perrotta, D., Colizza, V., Guerrisi, C., Turbelin, C., Duggan, J., Edmunds, J., Obi, C., Pebody, R., Franco, A.O., Moreno, Y., Meloni, S., Koppeschaar, C., Kjelsø, C., Mexia, R., Paolotti, D.: Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms. PLOS Computational Biology **15**(4), 1–21 (2019). https://doi.org/10.1371/journal.pcbi.1006173, https://doi.org/10.1371/journal.pcbi.1006173
14. Lim, S., Tucker, C.S., Kumara, S.: An unsupervised machine learning model for discovering latent infectious diseases using social media data. Journal of Biomedical Informatics **66**, 82–94 (2017). https://doi.org/https://doi.org/10.1016/j.jbi.2016.12.007, https://www.sciencedirect.com/science/article/pii/S1532046416301812
15. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence **2**(1), 2522–5839 (2020)
16. Lundberg, S.M., Erion, G.G., Lee, S.I.: Consistent Individualized Feature Attribution for Tree Ensembles. arXiv e-prints arXiv:1802.03888 (2018)
17. McInnes, L., Healy, J., Astels, S.: hdbscan: Hierarchical density based clustering. The Journal of Open Source Software **2**(11) (2017), https://doi.org/10.21105/joss.00205
18. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv e-prints arXiv:1802.03426 (2018)

19. Miaskowski, C., Barsevick, A., Berger, A., Casagrande, R., Grady, P., Jacobsen, P., Kutner, J., Patrick, D., Zimmerman, L., Xiao, C., Matocha, M., Marden, S.: Advancing symptom science through symptom cluster research: Expert panel proceedings and recommendations. Journal of the National Cancer Institute **109** (2017). https://doi.org/10.1093/jnci/djw253
20. Molnar, C.: Interpretable Machine Learning (2019), https://christophm.github.io/interpretable-ml-book/
21. Niemann, U., Spiliopoulou, M., Preim, B., Ittermann, T., Völzke, H.: Combining subgroup discovery and clustering to identify diverse subpopulations in cohort study data. In: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). pp. 582–587 (2017). https://doi.org/10.1109/CBMS.2017.15
22. Rubio-Rivas, M., Corbella, X., Mora-Luján, J.M., Loureiro-Amigo, J., López Sampalo, A., Yera Bergua, C., Esteve Atiénzar, P.J., Díez García, L.F., Gonzalez Ferrer, R., Plaza Canteli, S., et al.: Predicting clinical outcome with phenotypic clusters in covid-19 pneumonia: An analysis of 12,066 hospitalized patients from the spanish registry semi-covid-19. Journal of Clinical Medicine **9**(11), 3488 (2020). https://doi.org/10.3390/jcm9113488, http://dx.doi.org/10.3390/jcm9113488
23. Schelling, B., Bauer, L.G.M., Behzadi, S., Plant, C.: Utilizing structure-rich features to improve clustering. In: The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2020 (2020), http://eprints.cs.univie.ac.at/6416/
24. Sudre, C.H., Lee, K.A., Lochlainn, M.N., Varsavsky, T., Murray, B., Graham, M.S., Menni, C., Modat, M., Bowyer, R.C.E., Nguyen, L.H., Drew, D.A., Joshi, A.D., Ma, W., Guo, C.G., Lo, C.H., Ganesh, S., Buwe, A., Pujol, J.C., du Cadet, J.L., Visconti, A., Freidin, M.B., El-Sayed Moustafa, J.S., Falchi, M., Davies, R., Gomez, M.F., Fall, T., Cardoso, M.J., Wolf, J., Franks, P.W., Chan, A.T., Spector, T.D., Steves, C.J., Ourselin, S.: Symptom clusters in covid-19: A potential clinical prediction tool from the covid symptom study app. Science Advances **7**(12) (2021). https://doi.org/10.1126/sciadv.abd4177, https://advances.sciencemag.org/content/7/12/eabd4177
25. Zimmermann, A., De Raedt, L.: Cluster-grouping: From subgroup discovery to clustering. Machine Learning **77**, 125–159 (2009). https://doi.org/10.1007/s10994-009-5121-y