# DYNAMIC - CLUSTERING DATA DERIVED FROM CLINICAL DECISION SUPPORT SYSTEM *

**Sid Lamichhane**
*MSc semester project report*
July 2023
EPFL
sid.lamichhane@epfl.ch

**Mary-Anne Hartley & Martin Jaggi**
*Supervision*
Intelligent Global Health Research Group
EPFL
mary-anne.hartley@epfl.ch

**Alexandra Kulinkina & Zsofia Hesketh**
*Domain experts*
DYNAMIC project
Swiss TPH
alexandra.kulinkina@swisstph.ch




## ABSTRACT

**Background:** Health care services in resource-limited settings can benefit from Clinical Decision Support Systems (CDSS) to enhance quality of care. However, the potential of CDSS for syndromic surveillance remains untapped, as current surveillance methods rely on untrained staff and incomplete data. This study aims to explore the potential and limitations of clustering for syndromic surveillance, with a specific focus on the relationship between clustering and the underlying data structure of CDSS. **Methods/Findings:** The dataset consists of medical consultations conducted between December 2021 and February 2023, involving 47,886 children at 61 health facilities in Rwanda. Using demographic and medical features, consultations are clustered and analyzed spatially and temporally to assess performance and its alignment with the CDSS data structure, characterized by diagnoses and missingness. Results indicate that tested clustering approaches do not effectively cluster diagnoses, but the presence of missingness influences the clustering outcome if not properly addressed. **Conclusion:** Clustering CDSS data can reveal patterns beyond the underlying data structure, suggesting its potential for syndromic surveillance. Further exploration and refinement of clustering methods are necessary to fully exploit the capabilities of CDSS in this context.

* *Citation*: **Author.** *Title*. Year. MSc Semester Project Report. Intelligent Global Health, EPFL.

## 1 Background

Providing healthcare services in low-resource settings poses significant challenges for clinicians due to heavy workloads as well as limited training and support for serious cases [1]. Clinical Decision Support Systems (CDSS) have demonstrated their potential to enhance healthcare in such contexts by assisting clinicians in making informed decisions based on patient data [2]. In line with this objective, the DYNAMIC project of Swiss TPH developed a CDSS collaboratively with domain experts and healthcare practitioners [3]. The primary focus of DYNAMIC is to address the critical issue of antibiotic resistance, given that the average 5-year-old child in Africa has undergone 25 antibiotic treatments in their lifetime [4]. However, the untapped potential of DYNAMIC remains underutilized. One notable benefit is the system's ability to automatically collect real-time, patient-level medical data, which brings further opportunities to improve public healthcare in low-resource settings. Leveraging this data, one possible application is the utilization of CDSS data for syndromic surveillance, which this study regards as doing spatio-temporal analysis of symptoms and outbreak detection. Notably, this application offers advantages over the current standard solution for syndromic surveillance, Integrated Disease Surveillance and Response (IDSR), which relies on untrained personnel and incomplete data, leading to limitations [5].

## 2 Aim and Objectives

The aim is to explore the potential and limitations of unsupervised learning on CDSS-derived data for syndromic surveillance.

1. **Objective 1.** To preprocess DYNAMIC's CDSS data [4.1]
2. **Objective 2.** To describe the data [4.2]
3. **Objective 3.** To explore clustering approaches and their relationship to the tree structure underlying DYNAMIC's CDSS [4.3]
4. **Objective 4.** To perform syndromic surveillance [4.4]
5. **Objective 5.** To build a visualisation platform to display a selection of the above results for an interdisciplinary audience [4.5]

# 3 Methods

## 3.1 Data

### *3.1.1 Context*

This study utilized data obtained from consultations conducted between December 2021 and February 2023 in 61 outpatient facilities distributed throughout Rwanda. The data was collected using the tablet-based application of DYNAMIC's CDSS. The study included a total of 47,886 infants and children up to the age of 15 from 1,037 villages across Rwanda. The CDSS captured a wide range of patient data, encompassing demographics, clinical signs, symptoms, diagnoses suggested by clinicians, and CDSS-generated diagnoses. To give at least one of the 169 diagnoses considered by the CDSS, its underlying decision tree uses the demographics: age and gender, along with 917 signs and symptoms. Due the usage of a decision tree the CDSS data originally is stored in a tree structure but was transformed to tabular data for this study.

### *3.1.2 Features*

As features the demographics: age and gender as well as 37 medical signs and symptoms were chosen. The latter were selected by the domain experts which intend to use these 37 medical features to find clusters manually with their domain knowledge. In the future, their results can be used as an additional evaluation of this study's clustering. The used features are special due to their missingness and differentiating data types. The mixed typed data comes with several implications, most notably for the model selection and preprocessing methodology. The majority of features are binary and categorical (See Appendix A). According to domain experts all categorical features are ordinal. Missingness - as in the amount and way features are missing values – defines the features as the majority of them consists of more than 50% of NAN values (See Fig 1). Considering that the medical features originate from a decision tree, their missingness is not at random (MNAR) since e. g. a symptom is not being captured by the CDSS because the consulting clinician regards that symptom to be non-existent or of low value. Consequently, medical features are missing values whenever the underlying symptom or sign is of low value. Besides MNAR, some features are also missing at random as their missing values can be explained by other features' values (See Fig 2) or missingness (See Fig 3). For instance, the missingness of *PE212 - Respiratory rate (breaths/min) - 8469* highly correlates with existing values of *S39 - Cough - 7817* (See Fig 2) and missing values of *PE18 - Chest indrawing - 7811* (See Fig 3).

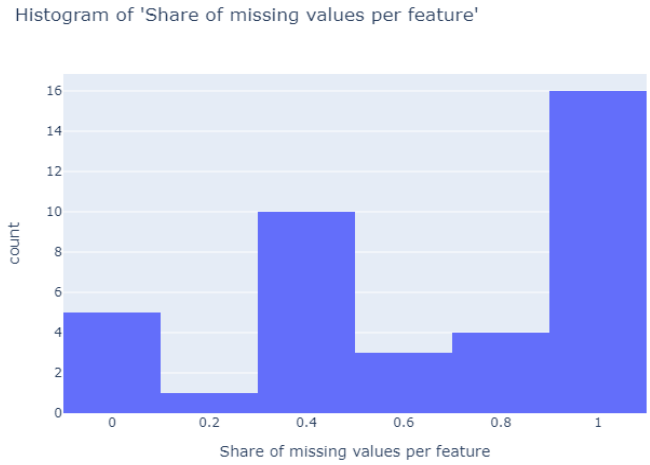



Figure 1


Figure 2


Figure 3

### *3.1.3 Labels*

As an unsupervised study, no labels are given. However, it is of interest to see how the found clusters overlap with diagnoses determined by the CDSS. Doing this, potentially reveals diagnoses missed by the CDSS; in the form of a totally new diagnosis or a subcategory of a considered diagnosis. As there are 169 possible diagnoses, this study limits itself to the 10 most common diagnoses to develop and evaluate clustering approaches [4.3]. At least one of the 10 most common diagnoses are recognised in 69.52% of all consultations (See Fig 4).To compare the found clusters with the selected diagnoses, one must consider that consultations can be assigned to more than one of the most common diagnoses (See Fig 5).

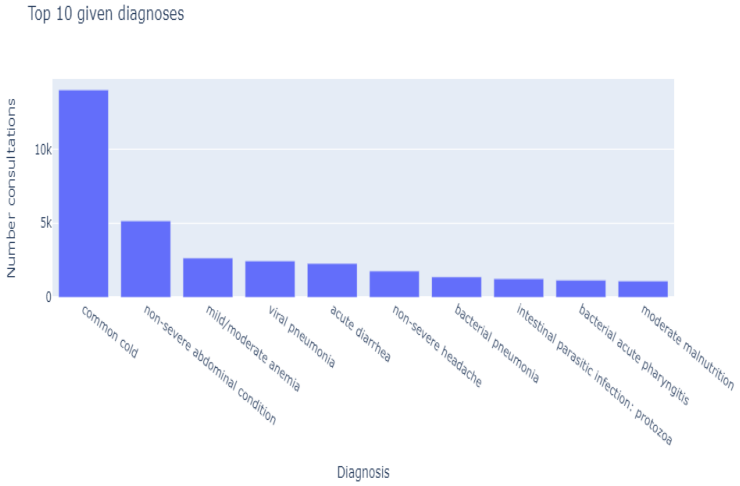



Figure 4

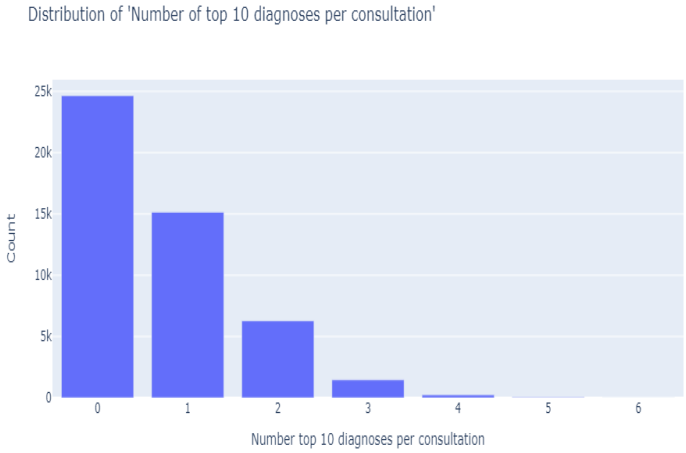



Figure 5

### 3.2 Clustering approaches

A clustering approach defines the preprocessing steps and clustering algorithm, including the setup of hyperparameters. For the purpose of syndromic surveillance, the widely used K-Means algorithm was chosen as the baseline clustering method in this study. However, as K-Means requires numerical values, an extension called K-Prototypes, designed for handling mixed-typed data, was also tested [6]. Both algorithms were implemented with the number of clusters determined by the elbow method, except during testing of various clustering approaches where 10 clusters were chosen to represent the 10 most common diagnoses.

To optimize the performance of K-Means clustering, this study experimented with different hyperparameters, particularly focusing on the initialisation method (K-Means++ vs. customized centroids) and the number of initialisations. Manual initialisation with customized centroids was found to potentially enhance clustering performance, although it introduces subjectivity [7]. In this study, clustering with a centroid for each cluster was tested to increase the likelihood of capturing all 10 most common diagnoses as individual clusters. For this purpose, each centroid was assigned the average feature values corresponding to the respective top 10 diagnoses.

Two preprocessing strategies were developed and implemented in this study. The first strategy aimed to include missingness in the features, while the second strategy focused on removing missingness through domain-based imputation. In the case of K-Prototypes, missing continuous feature values were imputed with their mean and standardized, while missing binary and categorical values were treated as an additional category. For K-Means, which requires all features to be numerical, the same preprocessing pipeline as K-Prototypes was applied, followed by one-hot encoding for binary and categorical features.

To remove missingness from the features, domain-based imputation was performed, which also maintained the ordinal scale of categorical features. In this approach, missing values were imputed with the most common state for a patient, while preserving the ordinal information through numericalization. However, this approach resulted in the loss of missingness information and raised concerns about the interpretation of distances between categories across features, particularly in the context of K-Means clustering. Given the aim of this study to explore the differences between clustering with and without missingness, examining the clustering results of all preprocessing approaches is of interest.

Table 1: Overview tested clustering approaches

| Clustering approach | Algorithm | Preprocessing | | | Hyperparameters | | |
|---|---|---|---|---|---|---|---|
| | | Numericalisation | Missingness | Scaling | Number clusters (K) | Number initialisations | Initialisation method |
| K-prototypes preprocessed features | K-Prototypes | Not needed | Included as category | Yes | 10 | 100 | As in original paper |
| K-means with K-prototypes preprocessed features | K-Means | One-hot encoding | Included as category | Yes | 10 | 100 | K-Means++ |
| K-means preprocessed with domain knowledge | K-Means | based on domain knowledge | Imputed based on domain knowledge | Yes | 10 | 100 | K-Means++ |
| K-means with K-prototypes preprocessed features and manual centroid init | K-Means | One-hot encoding | Included as category | Yes | 10 | 1 | Centroids based on diagnoses |
| K-means preprocessed with domain knowledge and manual centroid init | K-Means | based on domain knowledge | Imputed with domain knowledge | Yes | 10 | 1 | Centroids based on diagnoses |

### 3.3 Examining clustering's relationship to CDSS's tree structure

The tree structure in the given data is identifiable through diagnoses (representing the leaves) and the missingness of features (representing the branches). Therefore, analysing the relationship of clustering to both reveals its correspondence to the tree of the CDSS.

To investigate the relationship between clustering and diagnoses, this study employs multiple approaches. Firstly, overlaid visualizations are utilized by superimposing diagnoses onto features after dimensionality reduction through Uniform Manifold Approximation and Projection (UMAP) [8]. This visualization technique enables the examination of the alignment between diagnoses and clusters. Secondly, descriptive statistics are employed to analyze the distribution of diagnoses within each cluster. This provides insights into the correspondence between clustering results and the prevalence of specific diagnoses. Additionally, a supervised learning approach is employed, where a diagnoses based LightGBM (LGBM) classifier is trained to predict cluster labels. The accuracy, in the form of the cross-validated F1 score, and Shapley Additive Explanations (SHAP) values of the classifier are then utilized to assess the relevance of diagnoses in determining cluster assignments. This study aimed to address the question of whether clusters corresponded to specific diagnoses. If clusters did align with diagnoses, it would indicate a leaf-like structure in the clustering results. In this case this study would further investigate whether additional leaves, such as new diagnoses or sub-diagnoses not considered by the CDSS, were identified by the clustering. On the other hand, if clusters did not align with a tree-like structure, the study would examine the alternative patterns uncovered by the clustering. The thus resulting findings would be documented, and adjustments to the clustering approach or completely new approaches would be made to iteratively approximate the desired tree structure. The objective was to identify the factors within the data that prevented the clustering from aligning with the expected tree structure. This iterative process aimed to validate the tree-like structure and progressively approach a more accurate representation of the underlying diagnoses.

To explore the relationship between clustering and missingness, this study conducted a comparison between clustering approaches that incorporated missingness and those that did not. This comparison provided valuable insights into the impact of missingness on clustering outcomes. Additionally, analyzing the distribution of missingness within each cluster proved to be an insightful factor in understanding the clustering results. Furthermore, the same supervised learning approach as for the relationship to diagnoses was employed to delve deeper into the relationship between missingness and cluster labels. However, instead of using diagnoses this LGBM classifier takes the shadow matrix of the features as input.

### 3.4 Syndromic surveillance

This study specifically focuses on the spatio-temporal analysis of similar patients and outbreak detection as crucial components of syndromic surveillance. The analysis involves exploring clusters generated through the ultimately selected clustering approach, examining the distribution of demographics within each cluster, and tracking the number of consultations over time and space. Outlier detection is employed to identify potential outbreaks by utilizing z-scores, which are based on the proportion of consultations per week and village. Notably, this study concentrates solely on detecting outliers in the upper end, as an unusually high proportion of consultations could indicate the presence of an outbreak. Therefore, a threshold of 3 is chosen for the z-score to be classified as an outlier and thus point of interest for domain experts.

### 3.5 Visualisation platform

To help clinicians and experts in public health to perform syndromic surveillance, this study created a web based platform visualising its main results. The platform was built with Plotly Dash [9] and deployed using Heroku [10].

# 4 Results

## 4.1 Preprocessing pipeline

The preprocessing phase of this study consists of two main steps. In the first step, efforts were made to address data inconsistencies such as incorrect GPS coordinates of health facilities and inaccurate patient village information in the CDSS. Fuzzy search techniques based on the Levenshtein Distance metric were applied to correct village names, and consultations that could not be corrected were removed in consultation with domain experts. Additionally, consultations falling outside the designated intervention period were excluded as per domain expert requirements. After completing these steps, the study was left with 47,086 consultations, which accounts for 45.41% of the original raw data.

The second part of the preprocessing phase is specific to the chosen clustering approach [3.2]. This step aims to optimize the clustering process and gain insights into the relationship between the resulting clusters and the decision tree structure of the CDSS.

## 4.2 Description of data

### *4.2.1 Demographics*

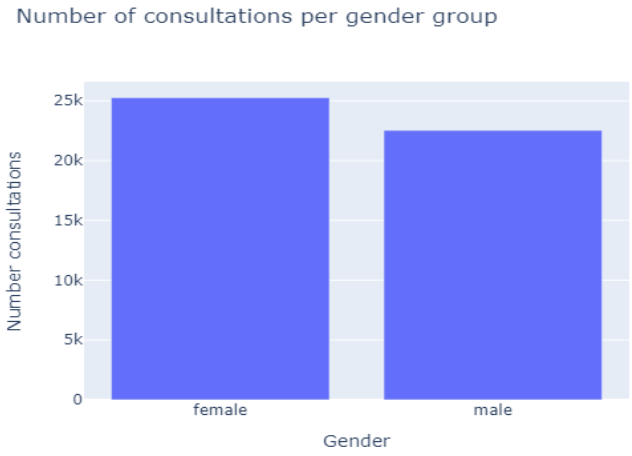



Figure 6

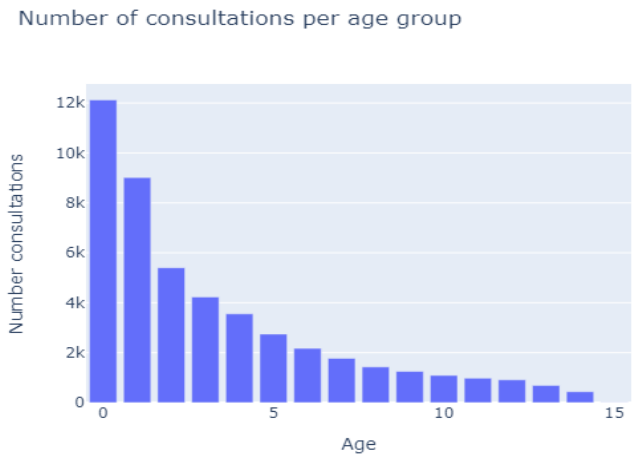



Figure 7

### *4.2.2 Temporal analysis*

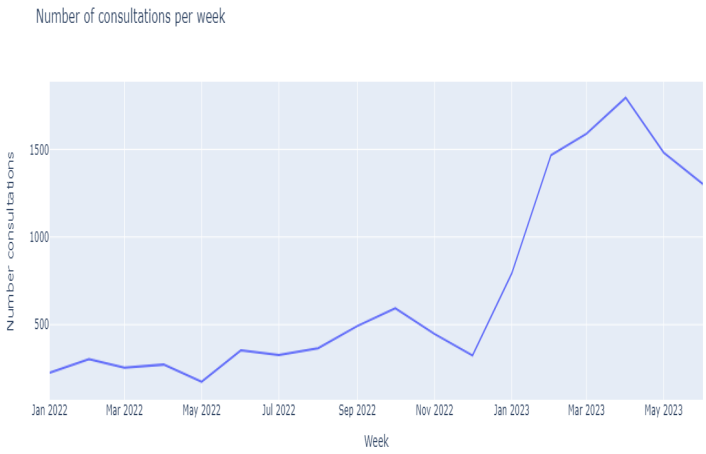



Figure 8

### *4.2.3 Spatial analysis*

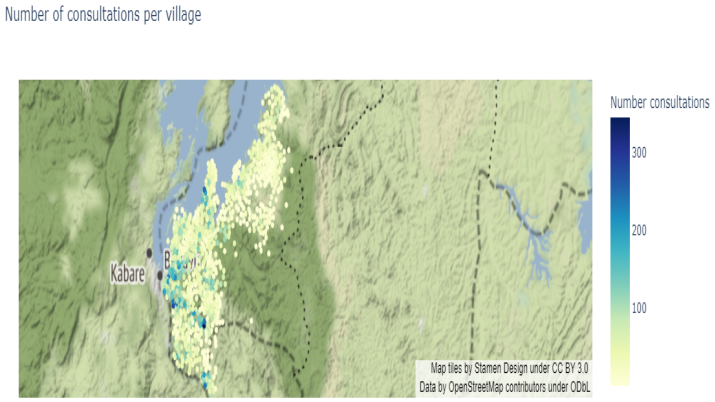



Figure 9

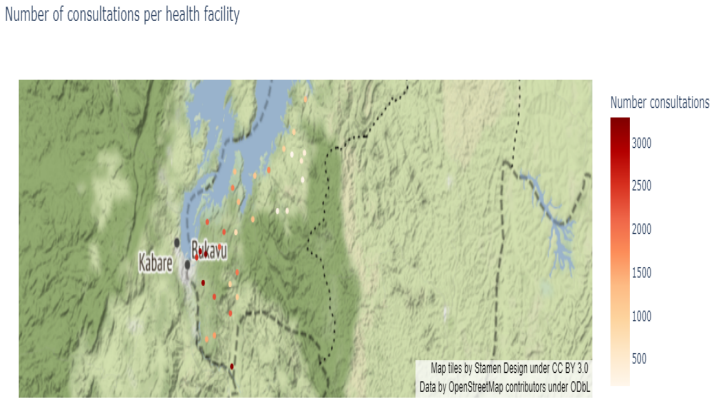



Figure 10

## 4.3 Clustering

### *4.3.1 Comparison and evaluation of clustering approaches*

In summary, the evaluation of various clustering approaches in this study has revealed that none of them can be considered optimal for the syndromic surveillance task at hand (See Appendix C). Despite this, no further optimization was pursued, and the *K-means preprocessed with domain knowledge* approach emerged as the most favorable option. This approach demonstrates a closer alignment with the desired 10 most common diagnoses, as evidenced by higher accuracy scores in the diagnosis-based classifier evaluation (See Appendix C.6.2).

It is important to acknowledge that the selected clustering approach, similar to the others examined, is not without limitations in producing distinct clusters, as indicated by its low silhouette score (See Appendix C.5). Furthermore, there exists a significant disparity in the accuracy of feature contributions to the clustering compared to alternative clustering approaches (See Appendix C).

#### 4.3.2 *Clustering's relationship to tree of CDSS*

**Diagnoses** Examining the relationship between clustering and diagnoses in CDSS data reveals that none of the clustering approaches successfully capture the diagnoses. There are potential reasons for this outcome. Firstly, the clustering algorithms may identify diagnoses other than the 10 most common ones used in this study. Secondly, it is possible that the medical features utilized in the study cannot effectively explain these diagnoses, even with domain knowledge. Consulting with domain experts would be necessary to validate this assumption. Following the methodology for investigating the clustering-diagnoses relationship, this study aimed to identify the key factors influencing the clustering results. Analyzing the SHAP values of feature-based classifiers, it was found that numerical features (such as age of patient and respiratory breath) dominantly influenced the result of all clustering approaches (See Appendix C.6.1). This observation raises concerns about the scaling of features during preprocessing, indicating a potential flaw in the process. Interestingly, manually initializing centroids, which was intended to improve the clustering performance, resulted in poorer ability to identify diagnoses (See Appendix **??**). This discrepancy may stem from the methodology used to compute the coordinates of the centroids which lacks scientific evidence. Surprisingly, despite this drawback, manual initialization of centroids still enhanced the overall clustering accuracy (See Appendix C.6.1), highlighting the need for further investigation. Lastly, it appears that clustering approaches that preserve missingness in the features encounter greater difficulty in identifying the diagnoses. This finding suggests that handling missing values presents challenges in the clustering process. These insights provide valuable directions for future research and underscore the complexities involved in aligning clustering outcomes with specific diagnoses.

**Missingness** When examining the relationship between clustering and missingness in the CDSS data, it becomes evident that the inclusion of missingness in the clustering features influences the clustering outcomes. Missingness can effectively explain the resulting clustering pattern across all approaches that incorporate missingness (See Appendix C.6.3). However, this influence is more pronounced in K-Means compared to K-Prototypes (See Appendix C.6.3). The reason behind this discrepancy could be attributed to the method used by K-Prototypes to compute distances. By considering categorical features and treating missingness as an additional category, K-Prototypes can better account for missing values. Furthermore, it is noteworthy that imputing missing values appears to diminish the impact of missingness on the clustering results. These findings highlight the importance of considering missingness in clustering analysis.

### 4.4 Syndromic surveillance

Running the selected clustering approach, *K-means preprocessed with domain knowledge*, on the complete dataset resulted in three clusters, which were utilized for syndromic surveillance and the development of the dashboard. While this study primarily focuses on establishing a data analysis pipeline for syndromic surveillance rather than conducting the surveillance itself, the disparity in clustering outcomes between the entire dataset and the subset of consultations associated with the 10 most common diagnoses is not extensively explored. Nonetheless, it remains intriguing to investigate the factors that contribute to the variation in the number of clusters when utilizing these different subsets, despite employing the same clustering approach in both cases.

### *4.4.1 Cluster analysis*

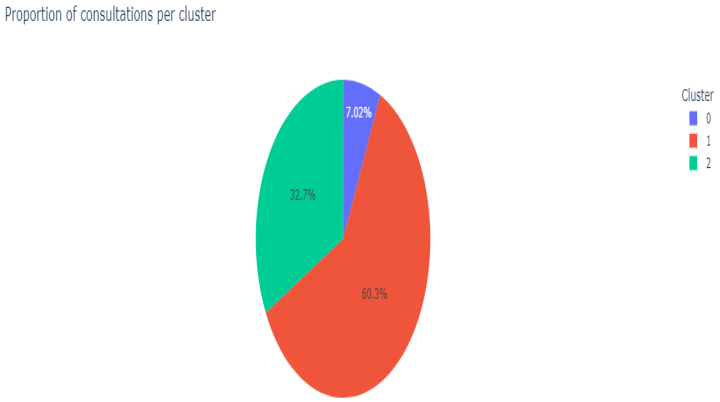



Figure 11


Figure 12

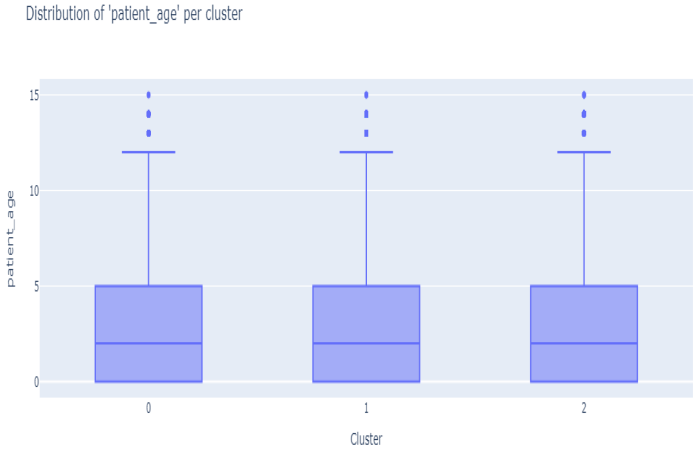



Figure 13

### *4.4.2 Spatio-temporal analysis*

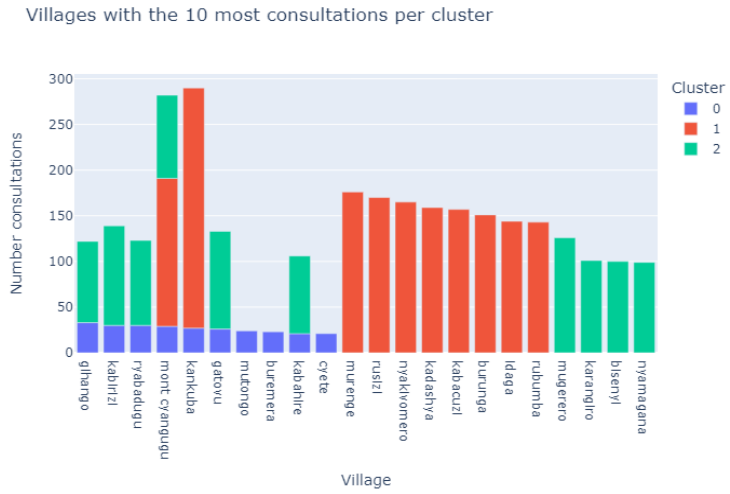



Figure 14

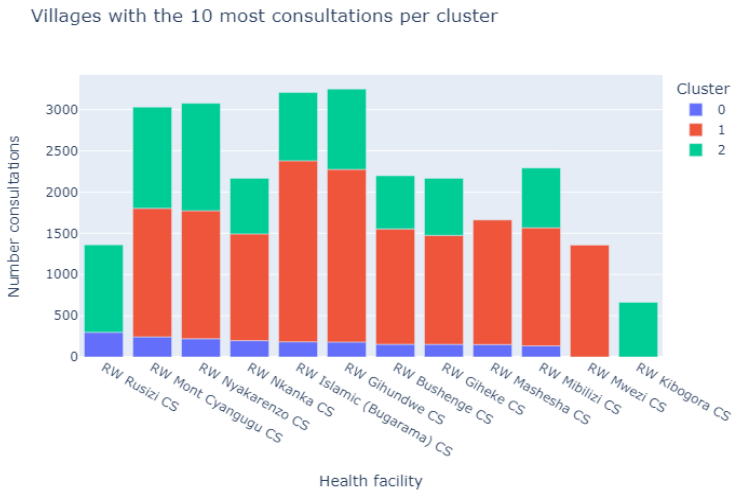



Figure 15

**Spatial analysis**

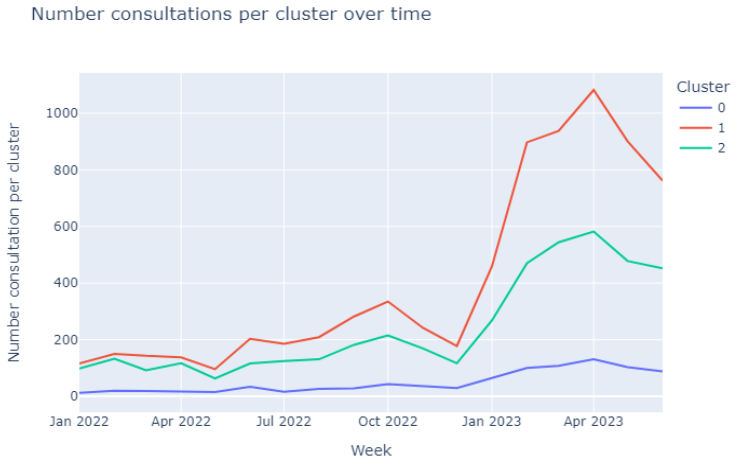



Figure 16

**Temporal analysis**

### *4.4.3 Outbreak detection*

This study found 443 outliers. To assess the potentiality of an outbreak, further investigation with domain experts is advised. Doing so, the provided syndromic surveillance dashboard can assist.

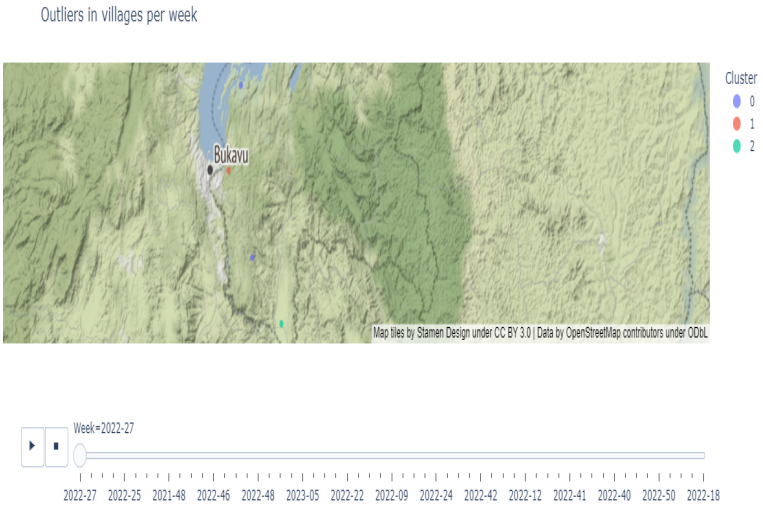



Figure 17: Outliers found in the 27th week

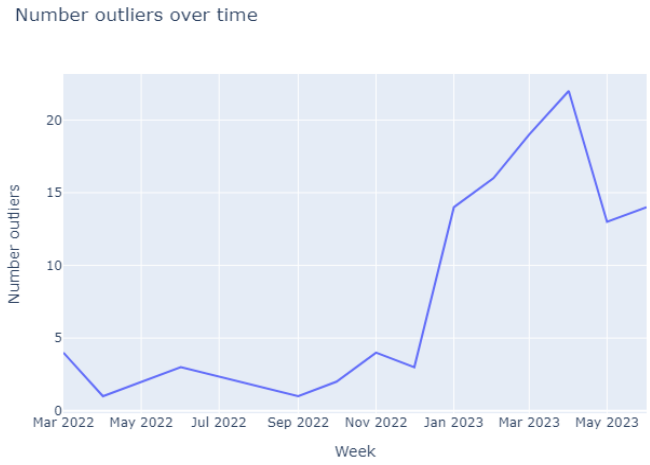



Figure 18

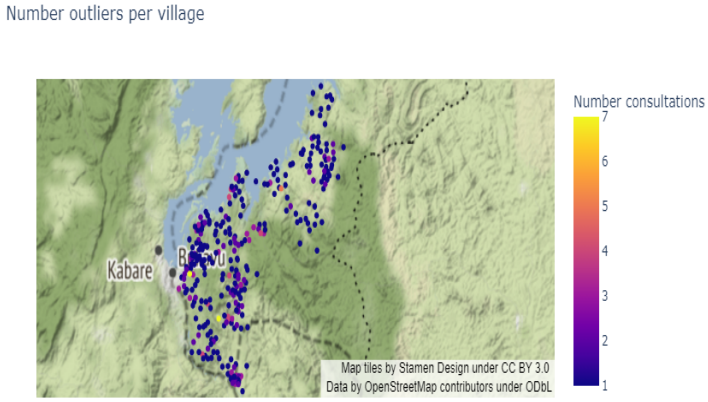



Figure 19

## 4.5 Visualisation platform

The syndromic surveillance dashboard contains four pages: Home, Data, Cluster analysis and Syndromic surveillance. The user can use the dropdown menu in the top right corner to navigate between them.

### *4.5.1 Home*

The content on this page requires adaptation to cater to the specific needs of the intended users, which is determined by the level of accessibility. If the dashboard is exclusively utilized by domain experts, a detailed and technical methodology description would be suitable. However, if the dashboard is public, considerations must be given to the limited knowledge of users in machine learning and the domain.


Figure 20: View on incomplete homepage with dropdown menu for navigation through analyses

### *4.5.2 Data*

This page informs the user about the used features and shows static visualisations about the demographics and number of consultations over space or time (See Fig 21). For the latter, the user has the option to select the unit of time (Day, Week, Month or Year) and can use a slider to inspect the number of consultations over time.

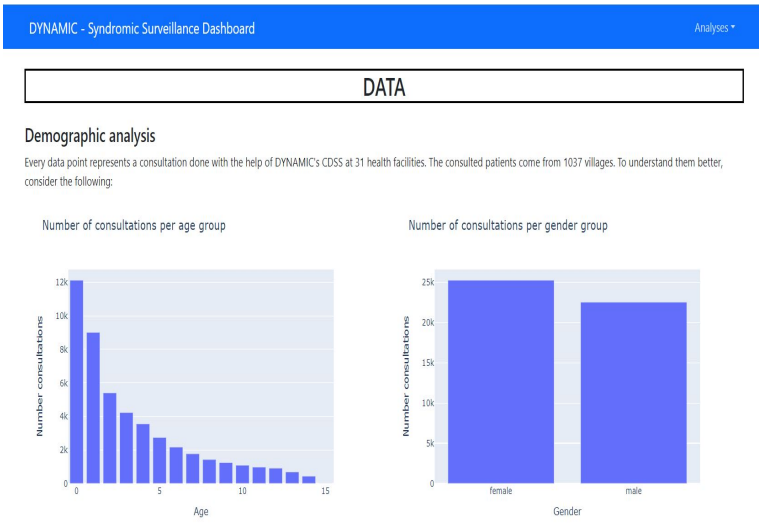



Figure 21

### *4.5.3 Cluster analysis*

With the help of this page users can explore the clustering result. They can interactively see the distribution of features, complaints and diagnoses per cluster.

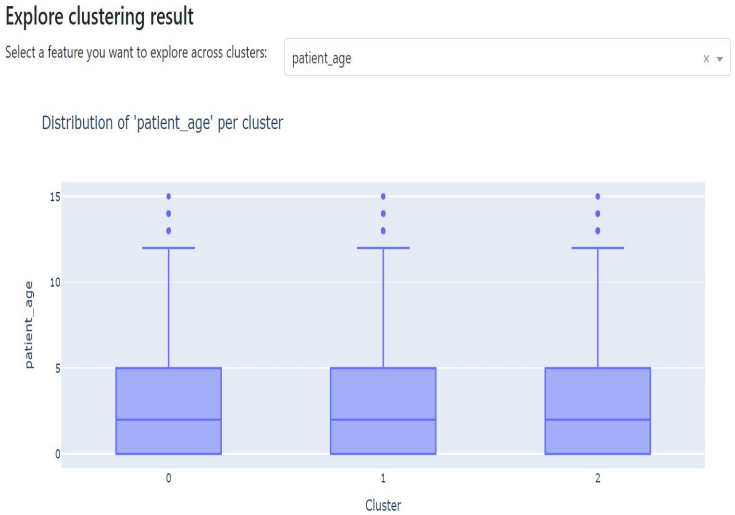



Figure 22


Figure 23

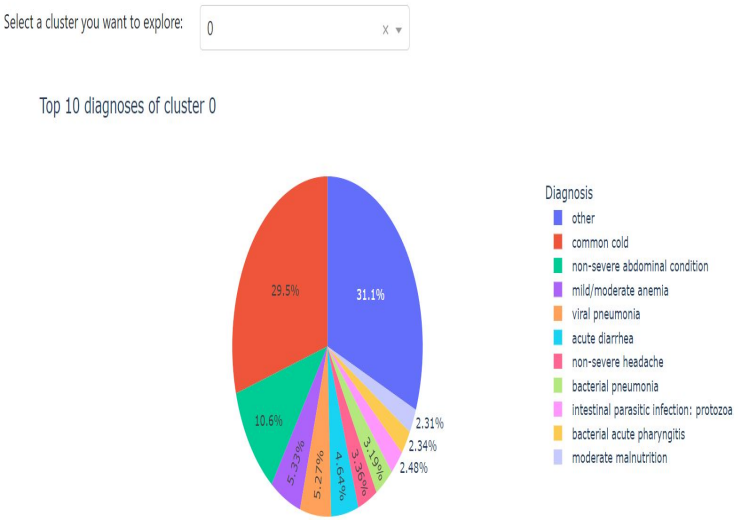



Figure 24

### *4.5.4 Syndromic surveillance*

This page provides users interactive visualizations to see outliers over space and time (See Fig 25) which can further be examined through the spatio-temporal analysis of the clusters also given on this page (See Fig 27). In combination with the cluster analysis page a holistic view about the situation can be achieved so that it can help domain experts to perform syndromic surveillance.

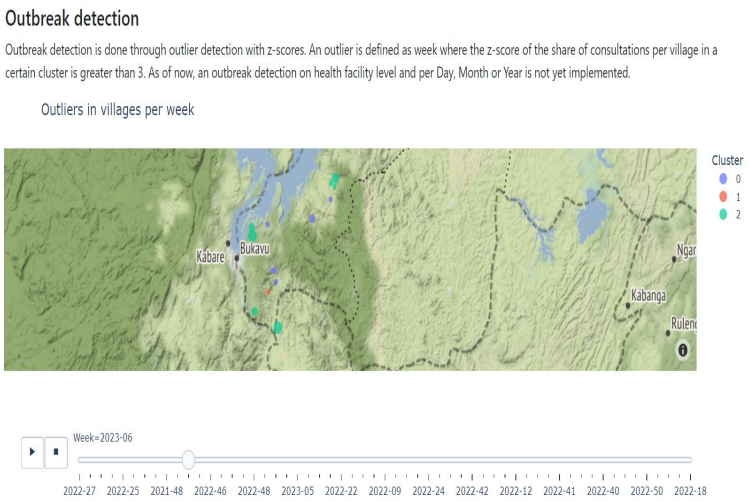



Figure 25

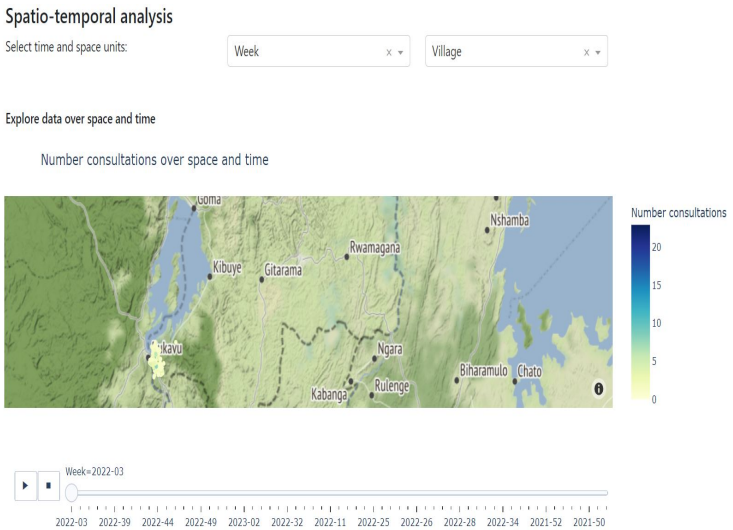



Figure 26

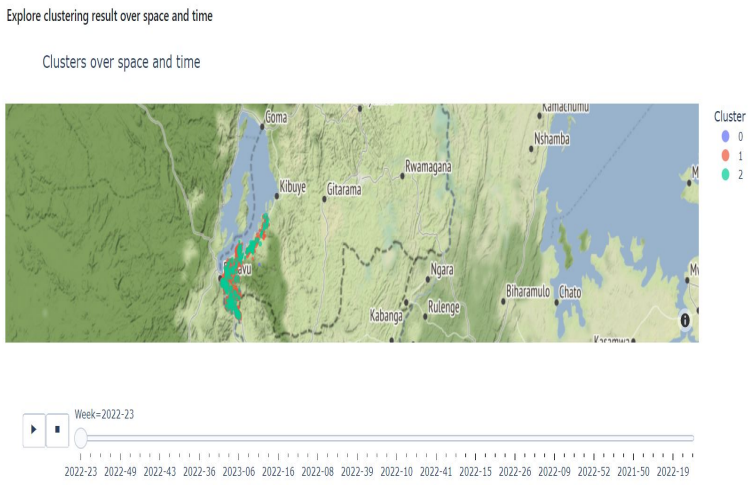

Explore clustering result over space and time
Clusters over space and time
Goma
Kibuye
Gitarama
Rwamagana
Nshamba
Ngara
Kabanga
Rulenge
Biharamulo
Chato
Cluster
0
1
2
Week=2022-23
2022-23 2022-49 2022-43 2022-36 2023-06 2022-16 2022-08 2022-39 2022-10 2022-41 2022-15 2022-26 2022-09 2022-52 2021-50 2022-19

Figure 27

# 5 Discussion

## 5.1 Limitations

This study encounters limitations stemming from the inherent constraints of the K-Means algorithm. One limitation is the algorithm's tendency to create clusters of approximately equal size, which hinders the creation of clusters that correspond to the top 10 diagnoses. These diagnoses would naturally result in clusters of varying sizes due to the differing number of consultations per diagnosis (See Fig 4). Additionally, K-Means is unable to assign a single consultation to multiple clusters, while clustering consultations for the top 10 diagnoses would require consultations to belong to more than one diagnosis cluster, considering that consultations can have multiple diagnoses (See Fig 5). Another limitation of this study originates from the fixed feature selection mandated by the domain experts. It forced using features with too many missing values (See Appendix C.5.3) and low variance (See Appendix C.5.3), which hold no information and thus do not bring any merit for the clustering regardless of the chosen algorithm. If the domain experts were to reconsider this constraint, a more comprehensive feature selection approach could potentially enhance the quality of clustering results. Various methods exist for selecting features in mixed-type data [11]. One simple way is to select features explaining the desired top 10 diagnoses best through supervised learning methods. Alternatively, in collaboration with domain experts and based on the results of this study, features can be selected using the insights gained from SHAP values and feature distributions.

## 5.2 Future work

### *5.2.1 Fix remaining errors*

Despite extensive preprocessing, several errors in the data persist. Addressing these errors should be prioritized as the initial action for any feature work. One error lies in the feature capturing the age of patients, which has been identified as incorrect. This represents a significant error, as the SHAP values obtained from the feature-based classifier clearly indicate the importance of age as a feature for multiple clustering approaches. (See Appendix C.6.1). This issue was recognized by domain experts too late to be rectified for this study. Another issue involves incorrect GPS coordinates for 77 villages. Due to the nature of the error originating from the data itself and the relatively low number of villages affected out of the total 1037, a decision was made not to correct these coordinates at present.

### *5.2.2 Adjust methodology*

To enhance the methodology of this study, several improvements can be considered.

Firstly, optimizing the chosen dimensionality reduction technique, UMAP, or exploring alternative methods can improve the quality of the reduced feature space and thus the visualization of clustering results. Principal Component Analysis (PCA) and Factor Analysis of Mixed Data (FAMD) [12], which were tested as alternatives, did not yield distinct clusters, potentially due to the data's failure to meet the linear dependency requirement of PCA [13]. On the other hand, t-distributed Stochastic Neighbor Embedding (t-SNE) and UMAP showed promising results, with UMAP ultimately being chosen for its visually separable clusters as well as higher and thus better silhouette score [14].

Secondly, improvements can be made to the diagnoses-based centroid initialization for K-Means, as the current approach was performed hastily. Investigating how this initialization method can potentially enhance clustering performance, as measured through supervised learning, would provide valuable insights.

Thirdly, examining the overall impact of the tested feature preprocessing techniques on LGBM classifier used to evaluate clustering would further strengthen the methodology. Understanding how these preprocessing methods influence the classifier's performance can contribute to a more comprehensive analysis of the clustering results.

Fourthly, it is recommended to explore additional clustering approaches, taking into account the trade-off between maintaining the ordinal scale of categorical features and preserving missingness involved in preprocessing the CDSS data. One potential solution is to numerically represent missing values based on domain knowledge, while also including a binary indicator variable for each feature to preserve missingness information. This approach may lead to an optimal clustering result with K-Means.

Lastly, considering the results of manual clustering performed by domain experts parallel to this study can serve as an additional evaluation metric for any clustering outcome. However, due to time constraints, collaboration on this matter with the other study was not feasible.

### *5.2.3 Potential follow-up projects*

This study laid the ground work for several follow-up projects which would profit from the implemented preprocessing pipeline and the insights gained on how the tested clustering approaches relate to diagnoses and missingness in the CDSS data.

**DYNAMIC's CDSS data from Tanzania** It would be of interest to extend the analysis by applying the same clustering approaches to the CDSS data from Tanzania, where DYNAMIC is also utilized. Comparing the clustering results from Tanzania with those obtained in this study would provide valuable insights and contribute to a broader understanding of the clustering performance across different settings.

**Cluster representations of patients** The iGH group has developed Modular Clinical Decision Support Networks [15], which dynamically update the patient state as new information becomes available. It would be of great interest to investigate the relationship between clustering the patient representations generated by these networks and the underlying decision tree structure of the CDSS.

**Clustering for outbreak detection** Due to this study's given resources the primary focus of this study was not on outbreak detection despite its necessity for syndromic surveillance. However, since this study demonstrated that clustering the data from DYNAMIC's CDSS reveals more than just the underlying tree structure, it also legitimised pursuing a follow-up project solely dedicated to outbreak detection. This study recommends exploring proven clustering approaches designed specifically for this purpose (For review, see [16]). Anomaly detection techniques can also be considered for outbreak detection, especially if based on FAMD [17] since this study tested and implemented FAMD for dimensionality reduction. Exploiting the given data owned by the iGH group, the host of this study, remote sensing satellite data can be additionally used for outbreak detection due to its potential to uncover risk areas for epidemic diseases by establishing connections between the environment, climate, and health [18].

## 5.3 Conclusion

In conclusion, this exploratory study highlights the important finding that clustering CDSS data does not necessarily align with the underlying decision tree. Therefore, the utilization of CDSS data for clustering purposes can lead to new insights. The presence of structural missingness within the CDSS data does impact the clustering results, but this influence can be mitigated through the application of imputation techniques guided by domain knowledge.

By implementing a comprehensive pipeline and developing a syndromic surveillance dashboard, this study facilitates the seamless continuation of the project's objective to explore the potential and limitations of CDSS data for clustering. The dashboard serves as a valuable tool for domain experts to explore the data and provide suggestions for further improvements. For instance, domain experts can propose new features that can be readily integrated into the existing data pipeline.

With an open-ended nature, this study sets the stage for future investigations into enhancing the clustering of CDSS data. Continued exploration of the data, along with expert insights and the suggested refinements, will contribute to further advancements in utilizing CDSS data for syndromic surveillance.

## Acknowledgments

Big appreciation for Swiss TPH and iGH to come up with this meaningful project and thanks to both for letting me contribute. Furthermore, I thank Mary-Anne Hartley for her supervision of my work throughout the last months due to which I learned a lot about academic writing and Machine Learning. By providing their expertise in the field of medicine and public health in Rwanda, Alexandra Kulinkina and Zsofia Hesketh from Swiss TPH were of big help for this study's aim.

## References


[1] Fabienne N. Jaeger, Mahamat Bechir, Moumini Harouna, Daugla D. Moto, and Jürg Utzinger. Challenges and opportunities for healthcare workers in a rural district of chad. *BMC Health Services Research*, 18(1), January 2018.

[2] Torsten Schmitz, Fenella Beynon, Capucine Musard, Marek Kwiatkowski, Marco Landi, Daniel Ishaya, Jeremiah Zira, Muazu Muazu, Camille Renner, Edwin Emmanuel, Solomon Gideon Bulus, and Rodolfo Rossi. Effectiveness of an electronic clinical decision support system in improving the management of childhood illness in primary care in rural nigeria: an observational study. *BMJ Open*, 12(7), 2022.

[3] Beynon F Levine GA Vaezipour N Luwanda LB et al Tan R, Cobuccio L. epoct+ and the medal-suite: Development of an electronic clinical decision support algorithm and digital platform for pediatric outpatients in low- and middle-income countries. *PLOS Digit Health*, 2023.

[4] DYNAMIC: New Clinical Decision Support Tool Reduces Antibiotic Prescription in Children by a Four-fold — swisstph.ch. `https://www.swisstph.ch/en/news/news-detail/news/dynamic-new-clinical-decision-support-tool-reduces-antibiotic-prescription-in-children-by-a-four-f` [Accessed 07-Jul-2023].

[5] Caitlin M. Wolfe, Esther L. Hamblion, Emmanuel K. Dzotsi, Franck Mboussou, Isabelle Eckerle, Antoine Flahault, Claudia T. Codeço, Jaime Corvin, Janice C. Zgibor, Olivia Keiser, and Benido Impouma. Systematic review of integrated disease surveillance and response (IDSR) implementation in the african region. *PLOS ONE*, 16(2):e0245457, February 2021.

[6] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery 2*, pages 283–304, 1998.

[7] Pasi Fränti and Sami Sieranoja. How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93:95–112, 2019.

[8] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[9] Plotly. Dash enterprise - the premier data app platform for python.

[10] Heroku. Data on heroku build data-driven apps with fully managed data services.

[11] Saúl Solorio-Fernández, Jesús Carrasco-Ochoa, and José Francisco Martínez-Trinidad. A survey on feature selection methods for mixed data. *Artificial Intelligence Review*, 55:2821–2846, 04 2022.

[12] Jérôme Pagès. *Multiple Factor Analysis by Example Using R*. 11 2014.

[13] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[14] Andrzej Dudek. Silhouette index as clustering evaluation tool. In Krzysztof Jajuga, Jacek Batóg, and Marek Walesiak, editors, *Classification and Data Analysis*, pages 19–33, Cham, 2020. Springer International Publishing.

[15] Cécile Trottet, Thijs Vogels, Kristina Keitel, Alexandra V Kulunkina, Rainer Tan, Ludovico Cobuccio, Martin Jaggi, and Mary-Anne Hartley. Modular clinical decision support networks (modn)—updatable, interpretable, and portable predictions for evolving clinical environments. *medRxiv*, 2022.

[16] Mohamad Farhan Mohamad Mohsin, Abdul Hamdan, and Azuraliza Abu Bakar. A review on anomaly detection in disease outbreak detection. *The 1st International Conference on Information Science and Management (ICoCSIM)*, pages 22–28, 01 2012.

[17] Matthew Davidow and David S. Matteson. Factor analysis of mixed data for anomaly detection, 2020.

[18] Boukaye Boubacar Traore, Bernard Kamsu-Foguem, and Fana Tangara. Data mining techniques on satellite images for discovery of risk areas. *Expert Systems With Applications*, 2017.

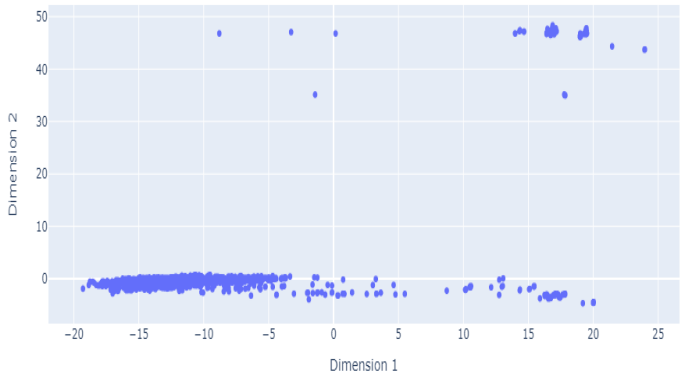
# Appendix

## A Summary of features

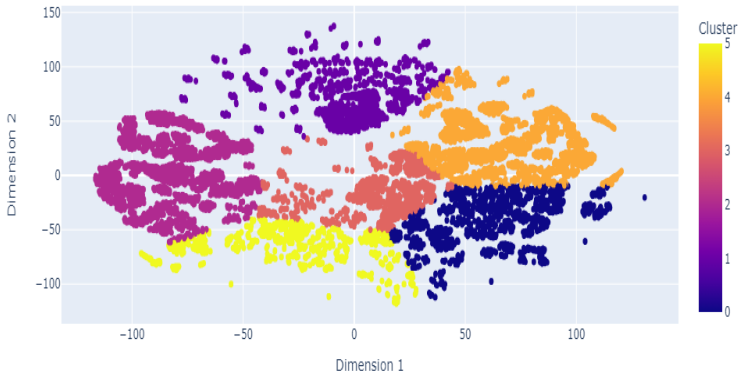
| Feature | Type | Count | Share missing values in % | Mean | Std | Min | 25% quantile | Median | 75% quantile | Max |
|---|---|---|---|---|---|---|---|---|---|---|
| patient_age | continues | 47806.0 | 0.0 | 3.3 | 3.5 | 0.0 | 0.0 | 2.0 | 5.0 | 15.0 |
| patient_gender | binary | 47806.0 | 0.0 | 0.5 | 0.5 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| PE212 - Respiratory rate (breaths/min) - 8469 | continues | 16888.0 | 64.7 | 34.8 | 10.9 | 5.0 | 28.0 | 35.0 | 40.0 | 144.0 |
| S39 - Cough - 7817 | binary | 32003.0 | 33.1 | 0.6 | 0.5 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| PE18 - Chest indrawing - 7811 | binary | 17436.0 | 63.5 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S180 - How is the infant feeding currently? - 7516 | categorical | 1401.0 | 97.1 | 0.1 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 |
| S46 - Convulsions in present illness - 8355 | binary | 32891.0 | 31.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| BC - Axillary temperature - 7823 categorical | binary | 35027.0 | 26.7 | 0.2 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| PE125 - Observation of movement - 8388 | categorical | 1396.0 | 97.1 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 |
| S96 - Unable to drink or breastfeed - 7871 | binary | 31510.0 | 34.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S157 - Vomiting everything - 8026 | binary | 31511.0 | 34.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| OS9 - Unconscious or Lethargic (Unusually sleep... | binary | 31511.0 | 34.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| PE63 - Stiff neck - 8391 | binary | 7766.0 | 83.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| PE19 - Stridor in calm child - 7812 | binary | 990.0 | 97.9 | 0.1 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S42 - Duration of fever (days) - 7819 categorical | categorical | 12589.0 | 73.7 | 0.1 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 |
| S124 - Significant weight loss - 7539 | binary | 8957.0 | 81.3 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S118 - Significant hemoptysis (>1 episode) - 7941 | binary | 1075.0 | 97.8 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S176 - Cough duration (days) - 7731 categorical | binary | 17201.0 | 64.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| E46 - Recent close contact with somebody with T... | binary | 28724.0 | 39.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| PE220 - Identifiable source of fever? - 8399 | binary | 1610.0 | 96.6 | 0.1 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| A55 - Urinary dipstick - 7648 | binary | 882.0 | 98.2 | 0.4 | 0.5 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| PE134 - Measles rash and associated signs - 8386 | binary | 1879.0 | 96.1 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| PE109 - Scarlet fever rash - 8344 | binary | 1426.0 | 97.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| PE106 - Non specific viral rash - 7930 | binary | 1743.0 | 96.3 | 0.3 | 0.4 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| PE218 - Unexplained bleeding - 8107 | binary | 780.0 | 98.4 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S153 - Diarrhea (stools are looser or more wate... | binary | 1401.0 | 97.1 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S98 - Number of loose or liquid stools over the... | categorical | 4050.0 | 91.5 | 1.2 | 1.2 | 0.0 | 0.0 | 1.0 | 2.0 | 4.0 |
| S149 - Yellow appearing skin or eyes (jaundice)... | binary | 1401.0 | 97.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| PE191 - Jaundice - 8036 | binary | 31466.0 | 34.2 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| S74 - Abnormal vaginal discharge - 8377 | binary | 240.0 | 99.5 | 0.4 | 0.5 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| S81 - Urethral discharge - 7810 | binary | 8.0 | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| S79 - Genital lesion - 7867 | binary | 29.0 | 99.9 | 0.2 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| BC7 - Weight for age (z-score) - 8434 categorical | categorical | 32958.0 | 31.1 | 0.1 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 |
| BC95 - Weight for height - 7451 categorical | categorical | 31563.0 | 34.0 | 0.1 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 |
| BC61 - MUAC for age z-score - 7839 categorical | categorical | 31371.0 | 34.4 | 0.2 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 |
| malaria | binary | 12219.0 | 74.4 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| fever without source: presumed bacterial infection | binary | 47806.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| fever without source: presumed viral illness | binary | 47806.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| suspicion of tuberculosis | binary | 47806.0 | 0.0 | 0.1 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

## B Comparison of dimensionality reduction methods

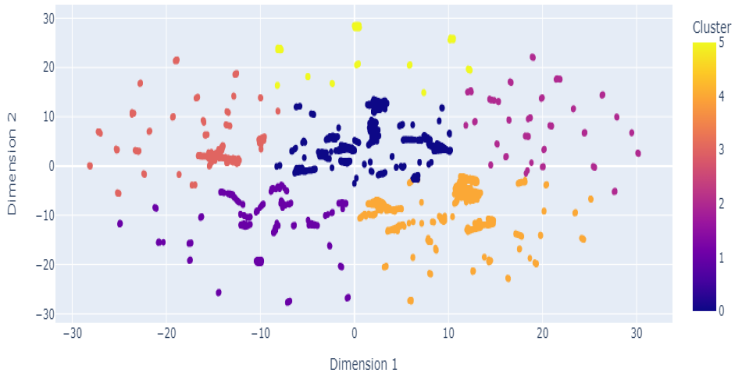
Features after FAMD




Clustering with silhouette score 0.3804728388786316 after dimensionality reduction through t-SNE

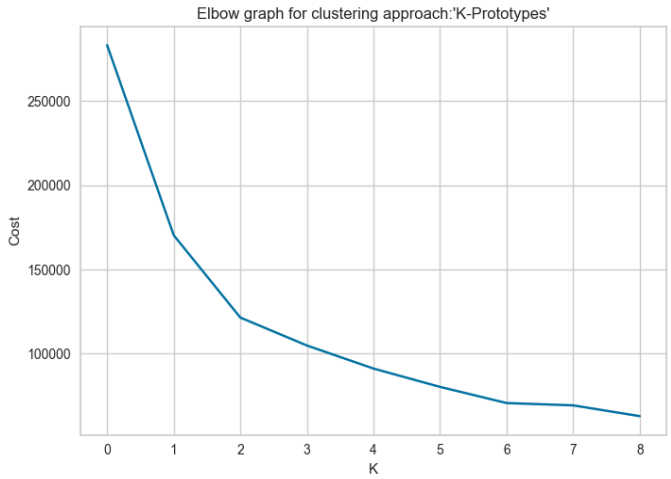


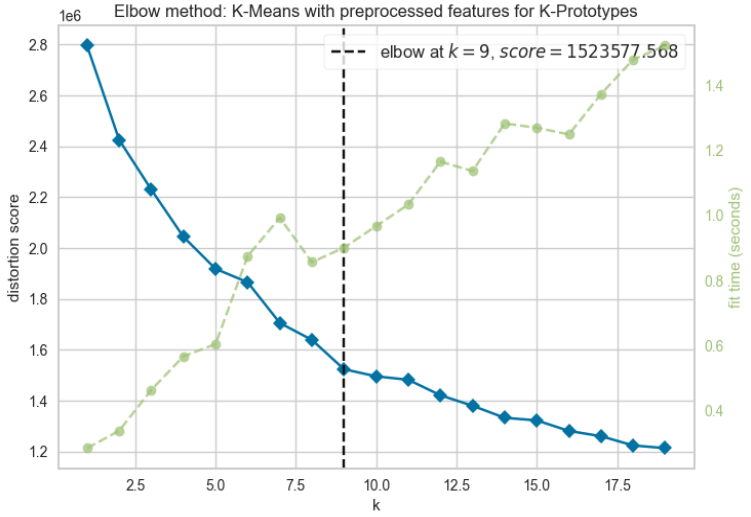

Clustering with silhouette score 0.4264075458049774 after dimensionality reduction through UMAP

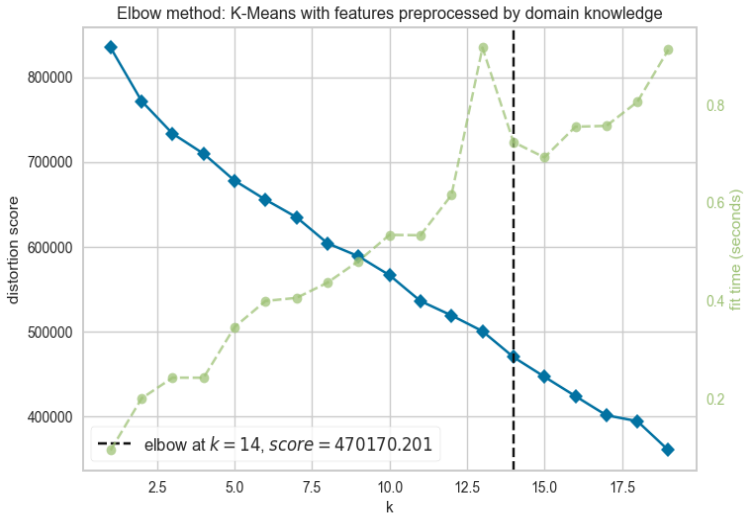


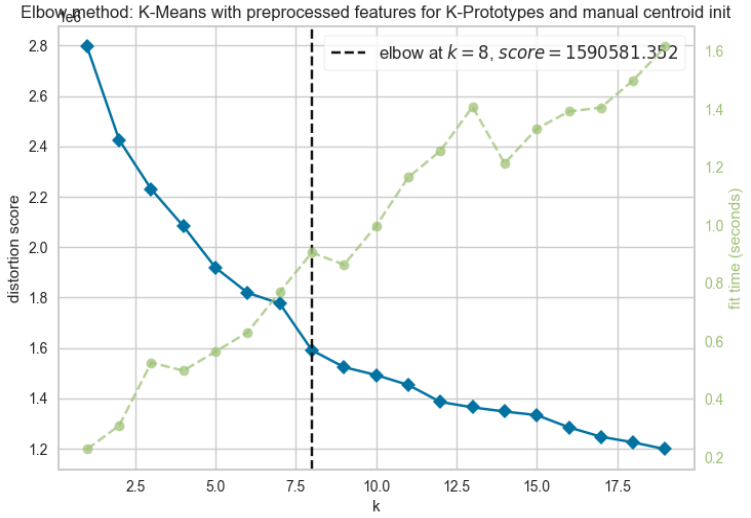
# C Evaluation of clustering approaches

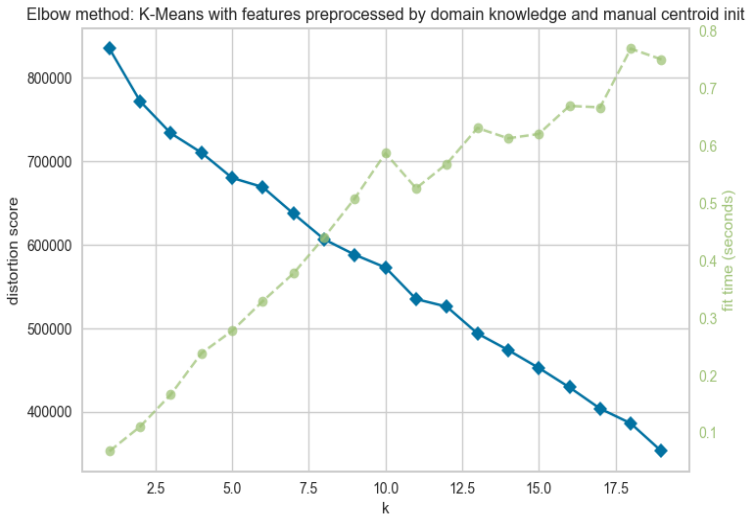
## C.1 Elbow method results



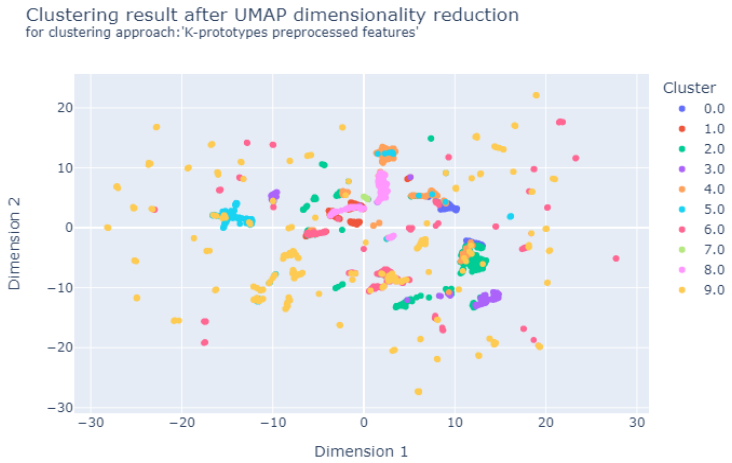



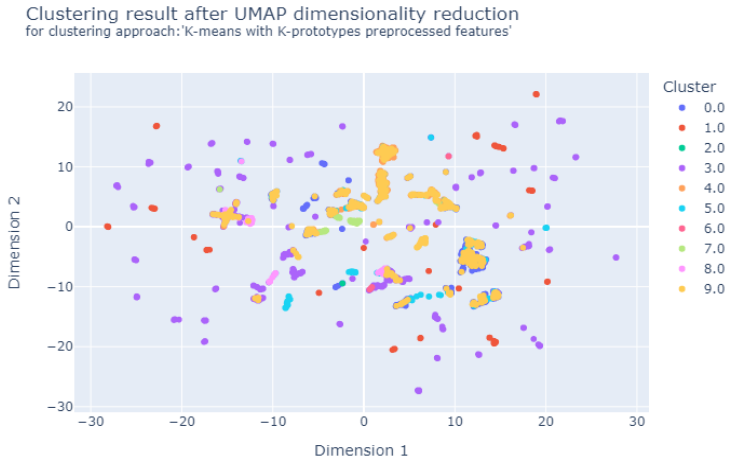

Elbow method: K-Means with features preprocessed by domain knowledge
distortion score
fit time (seconds)
k
elbow at $k = 14$, $score = 470170.201$

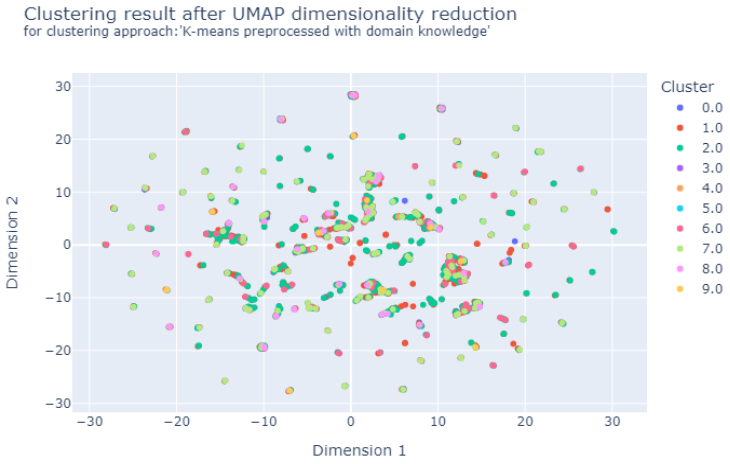

Elbow method: K-Means with preprocessed features for K-Prototypes and manual centroid init
1e6
distortion score
fit time (seconds)
k
elbow at $k = 8$, $score = 1590581.352$

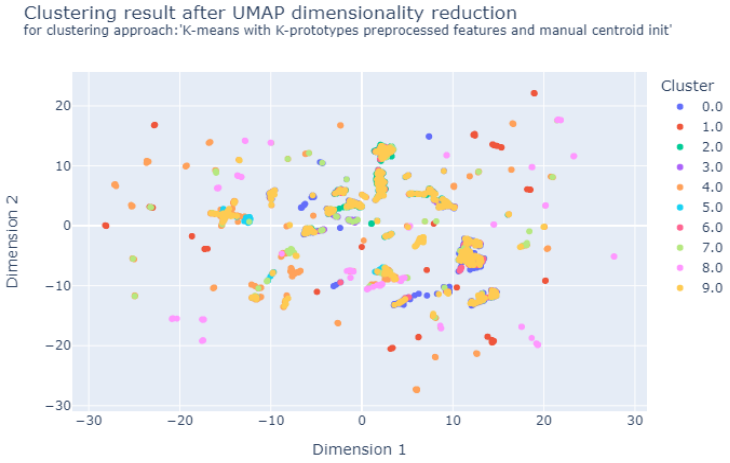

Elbow method: K-Means with features preprocessed by domain knowledge and manual centroid init
distortion score
fit time (seconds)
k

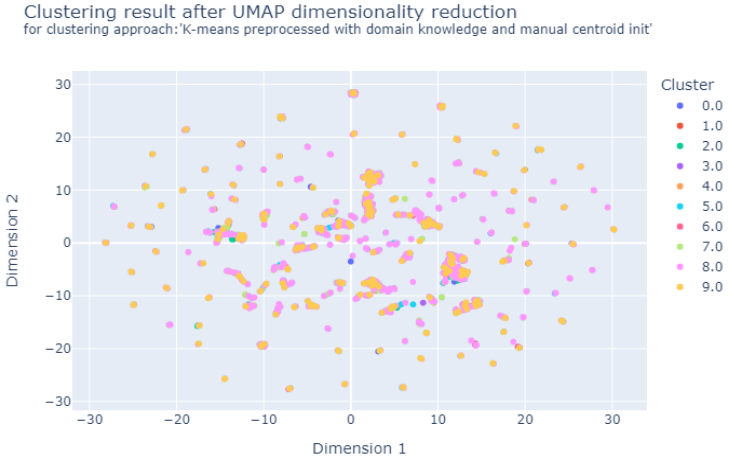
## C.2 Overlay clusters with found cluster labels



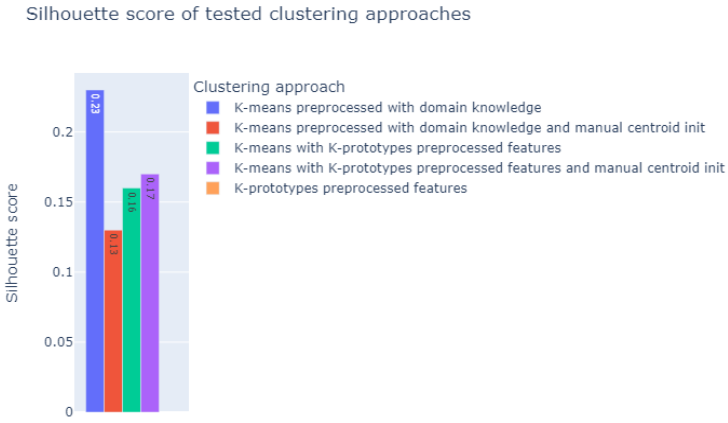




Clustering result after UMAP dimensionality reduction
for clustering approach:'K-means preprocessed with domain knowledge'
Cluster
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
Dimension 2
Dimension 1
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30


Clustering result after UMAP dimensionality reduction
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Cluster
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
Dimension 2
Dimension 1
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

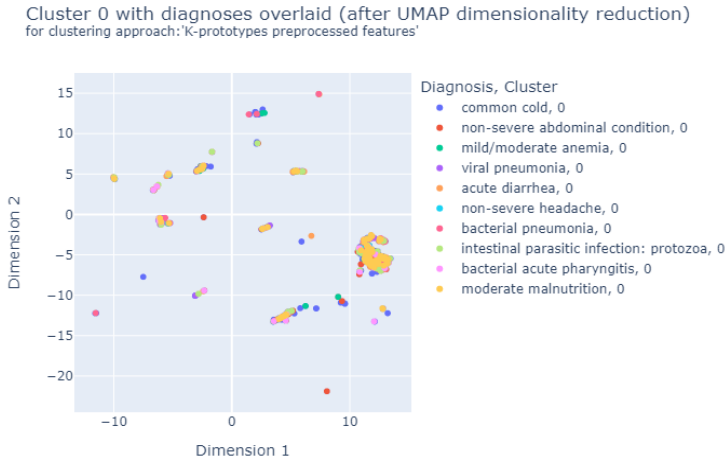

Clustering result after UMAP dimensionality reduction
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Cluster
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
Dimension 2
Dimension 1
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

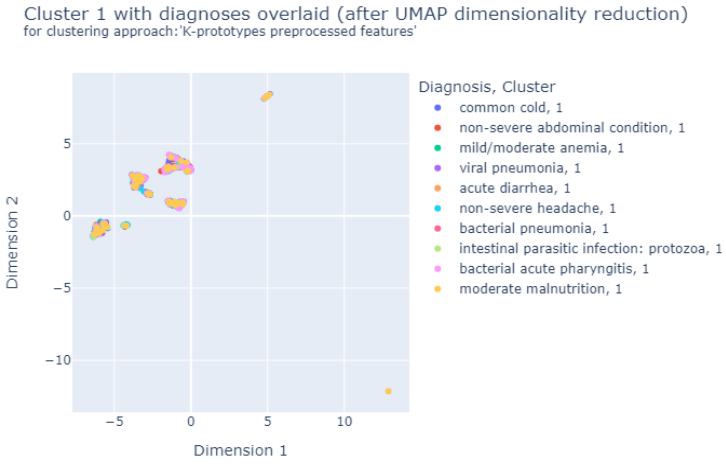
## C.3 Silhouette scores

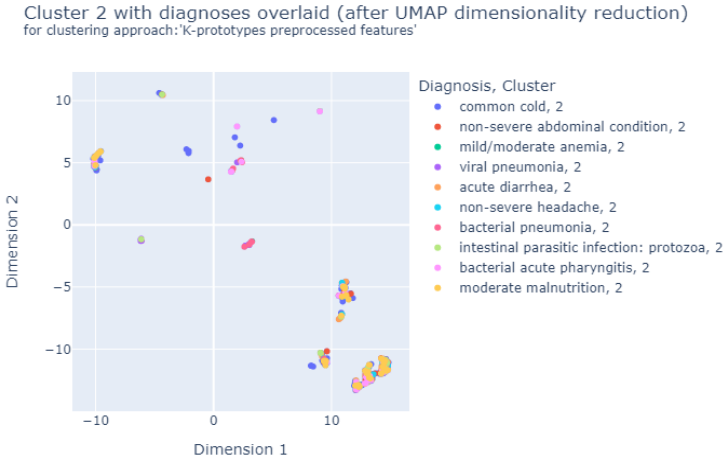


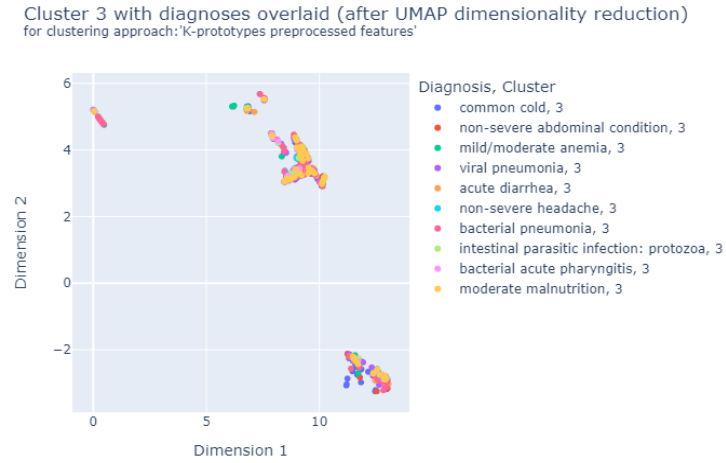

## C.4 Overlay clusters with diagnoses

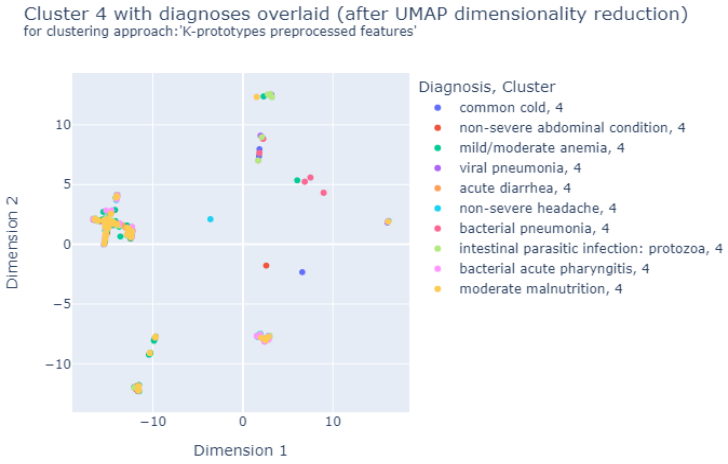
### *C.4.1 K-prototypes preprocessed features*

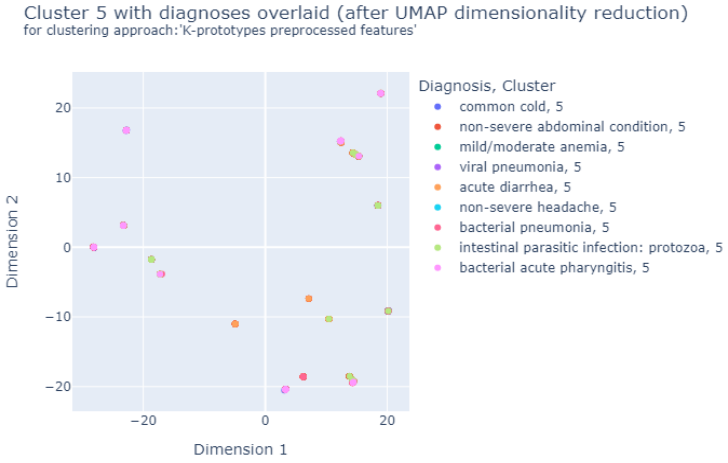


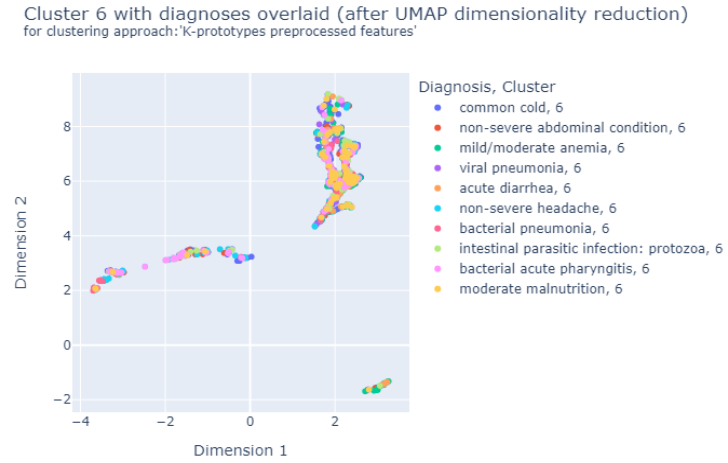

Cluster 1 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 1
non-severe abdominal condition, 1
mild/moderate anemia, 1
viral pneumonia, 1
acute diarrhea, 1
non-severe headache, 1
bacterial pneumonia, 1
intestinal parasitic infection: protozoa, 1
bacterial acute pharyngitis, 1
moderate malnutrition, 1
Dimension 2
Dimension 1

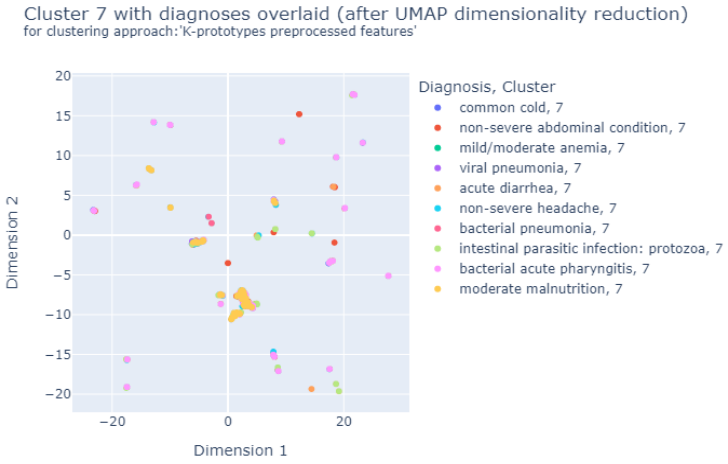

Cluster 2 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 2
non-severe abdominal condition, 2
mild/moderate anemia, 2
viral pneumonia, 2
acute diarrhea, 2
non-severe headache, 2
bacterial pneumonia, 2
intestinal parasitic infection: protozoa, 2
bacterial acute pharyngitis, 2
moderate malnutrition, 2
Dimension 2
Dimension 1

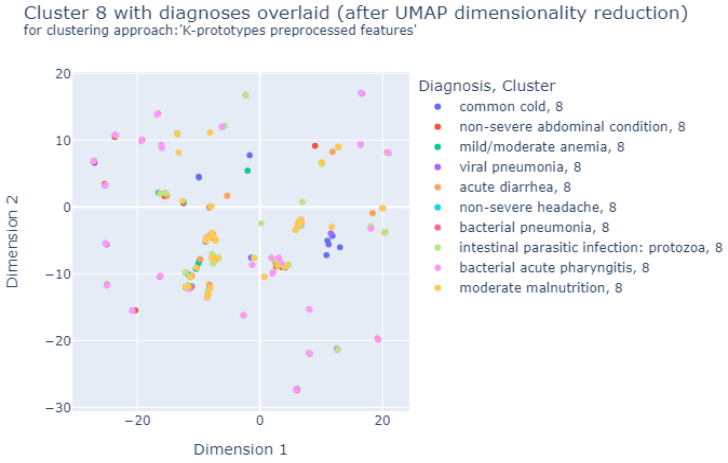

Cluster 3 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 3
non-severe abdominal condition, 3
mild/moderate anemia, 3
viral pneumonia, 3
acute diarrhea, 3
non-severe headache, 3
bacterial pneumonia, 3
intestinal parasitic infection: protozoa, 3
bacterial acute pharyngitis, 3
moderate malnutrition, 3
Dimension 2
Dimension 1

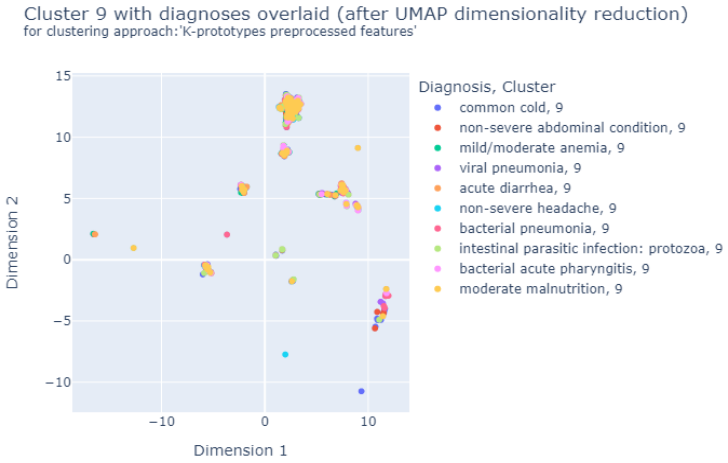

Cluster 4 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 4
non-severe abdominal condition, 4
mild/moderate anemia, 4
viral pneumonia, 4
acute diarrhea, 4
non-severe headache, 4
bacterial pneumonia, 4
intestinal parasitic infection: protozoa, 4
bacterial acute pharyngitis, 4
moderate malnutrition, 4
Dimension 1
Dimension 2


Cluster 5 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 5
non-severe abdominal condition, 5
mild/moderate anemia, 5
viral pneumonia, 5
acute diarrhea, 5
non-severe headache, 5
bacterial pneumonia, 5
intestinal parasitic infection: protozoa, 5
bacterial acute pharyngitis, 5
Dimension 1
Dimension 2

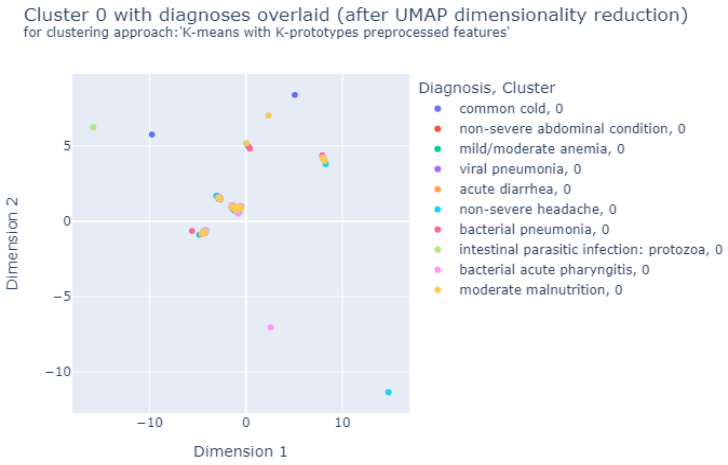

Cluster 6 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 6
non-severe abdominal condition, 6
mild/moderate anemia, 6
viral pneumonia, 6
acute diarrhea, 6
non-severe headache, 6
bacterial pneumonia, 6
intestinal parasitic infection: protozoa, 6
bacterial acute pharyngitis, 6
moderate malnutrition, 6
Dimension 1
Dimension 2

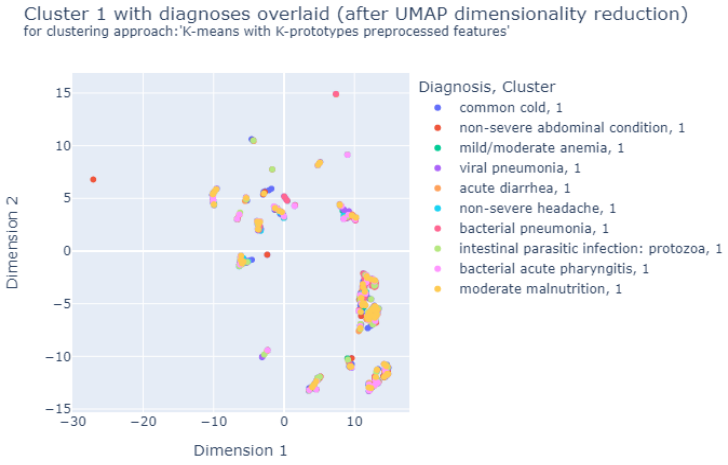

Cluster 7 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 7
non-severe abdominal condition, 7
mild/moderate anemia, 7
viral pneumonia, 7
acute diarrhea, 7
non-severe headache, 7
bacterial pneumonia, 7
intestinal parasitic infection: protozoa, 7
bacterial acute pharyngitis, 7
moderate malnutrition, 7
Dimension 1
Dimension 2

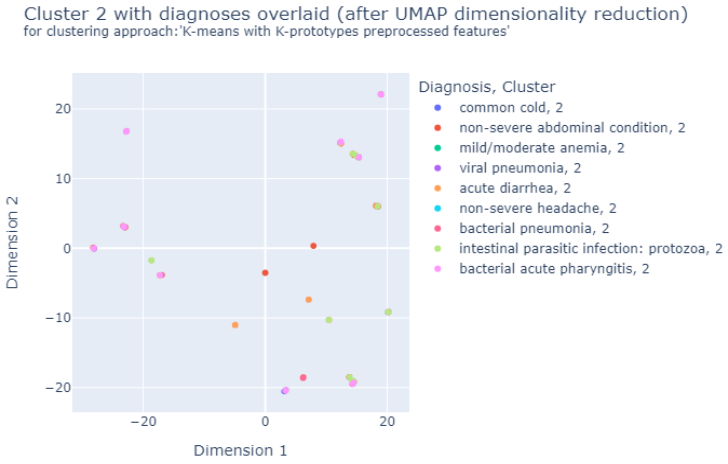

Cluster 8 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 8
non-severe abdominal condition, 8
mild/moderate anemia, 8
viral pneumonia, 8
acute diarrhea, 8
non-severe headache, 8
bacterial pneumonia, 8
intestinal parasitic infection: protozoa, 8
bacterial acute pharyngitis, 8
moderate malnutrition, 8
Dimension 1
Dimension 2

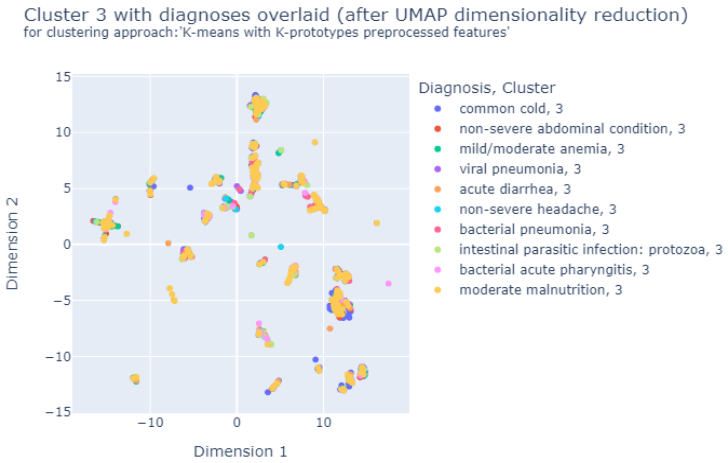

Cluster 9 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 9
non-severe abdominal condition, 9
mild/moderate anemia, 9
viral pneumonia, 9
acute diarrhea, 9
non-severe headache, 9
bacterial pneumonia, 9
intestinal parasitic infection: protozoa, 9
bacterial acute pharyngitis, 9
moderate malnutrition, 9
Dimension 1
Dimension 2

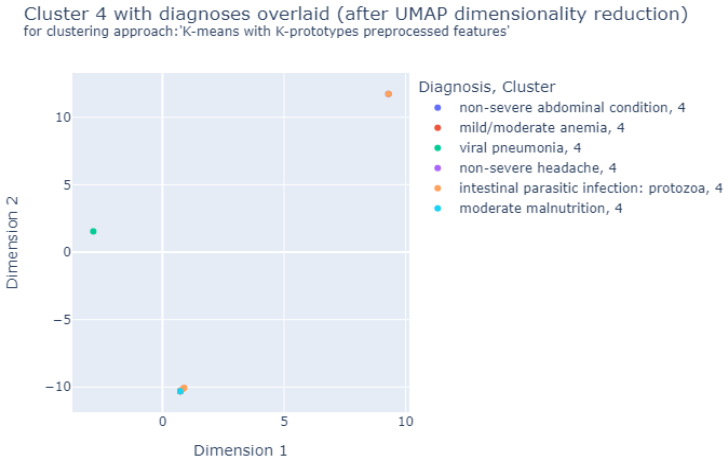
### *C.4.2 K-means with K-prototypes preprocessed features*

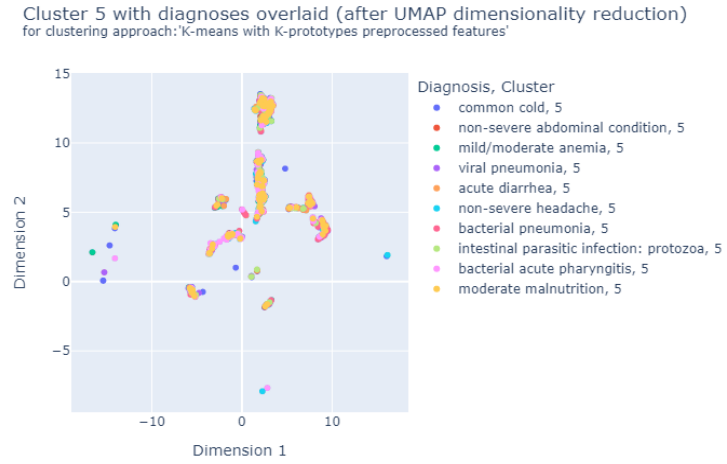


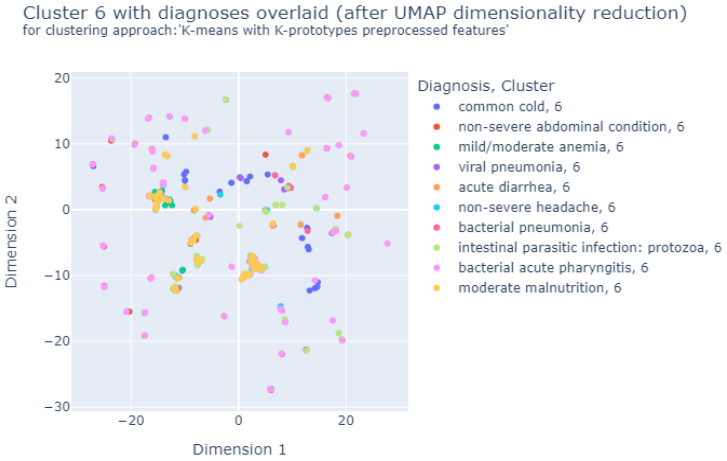



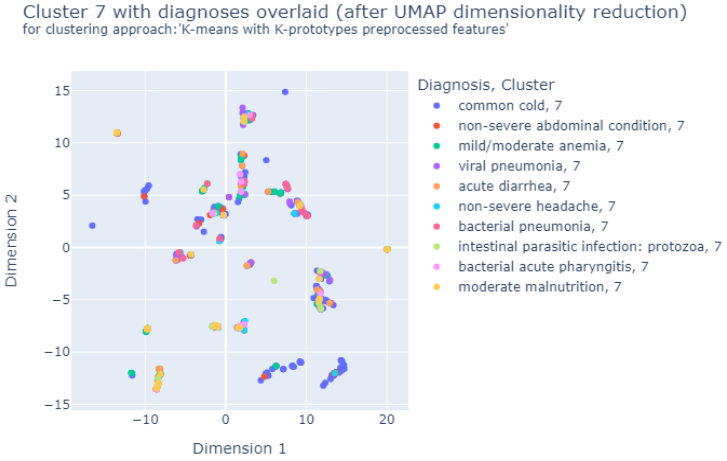

Cluster 2 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 2
non-severe abdominal condition, 2
mild/moderate anemia, 2
viral pneumonia, 2
acute diarrhea, 2
non-severe headache, 2
bacterial pneumonia, 2
intestinal parasitic infection: protozoa, 2
bacterial acute pharyngitis, 2
Dimension 2
Dimension 1
20
10
0
−10
−20
−20
0
20

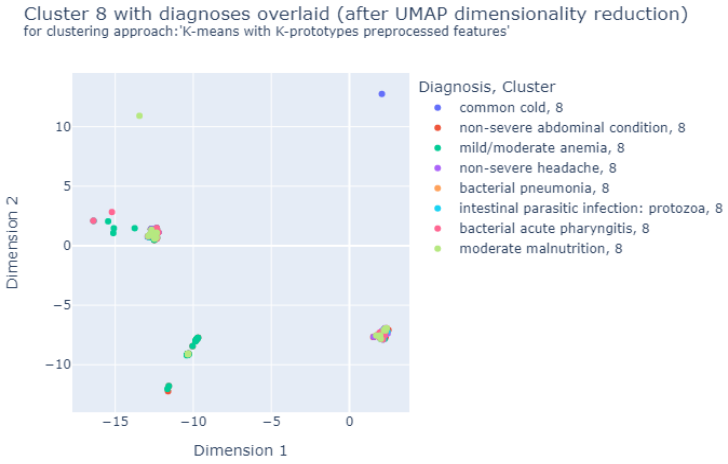

Cluster 3 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 3
non-severe abdominal condition, 3
mild/moderate anemia, 3
viral pneumonia, 3
acute diarrhea, 3
non-severe headache, 3
bacterial pneumonia, 3
intestinal parasitic infection: protozoa, 3
bacterial acute pharyngitis, 3
moderate malnutrition, 3
Dimension 2
Dimension 1
15
10
5
0
−5
−10
−15
−10
0
10

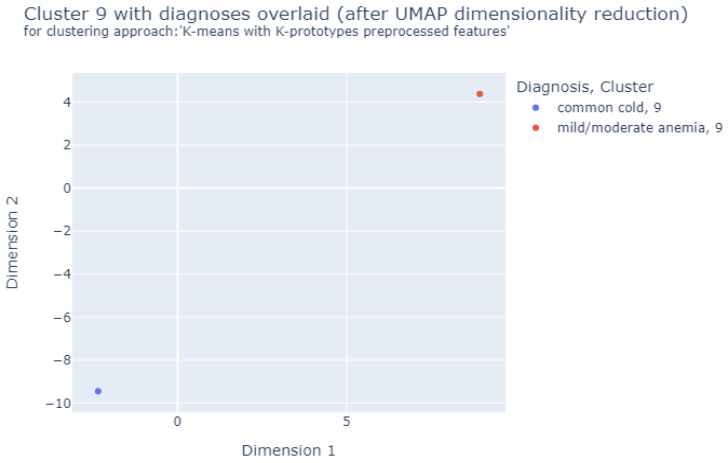

Cluster 4 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
non-severe abdominal condition, 4
mild/moderate anemia, 4
viral pneumonia, 4
non-severe headache, 4
intestinal parasitic infection: protozoa, 4
moderate malnutrition, 4
Dimension 2
Dimension 1
10
5
0
−5
−10
0
5
10


Cluster 5 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 5
non-severe abdominal condition, 5
mild/moderate anemia, 5
viral pneumonia, 5
acute diarrhea, 5
non-severe headache, 5
bacterial pneumonia, 5
intestinal parasitic infection: protozoa, 5
bacterial acute pharyngitis, 5
moderate malnutrition, 5
Dimension 1
Dimension 2

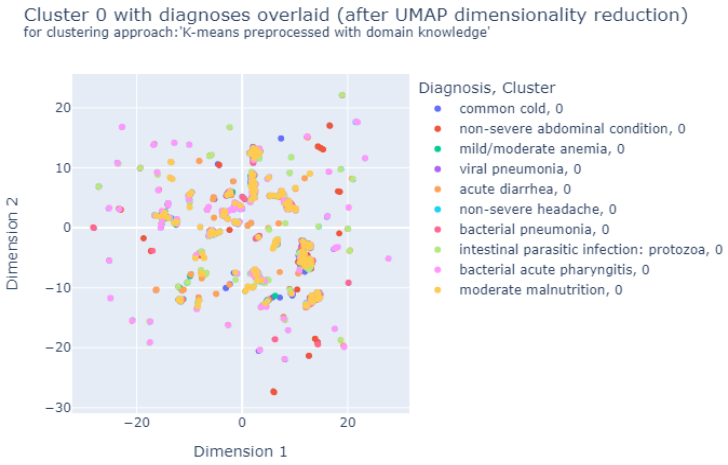

Cluster 6 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 6
non-severe abdominal condition, 6
mild/moderate anemia, 6
viral pneumonia, 6
acute diarrhea, 6
non-severe headache, 6
bacterial pneumonia, 6
intestinal parasitic infection: protozoa, 6
bacterial acute pharyngitis, 6
moderate malnutrition, 6
Dimension 1
Dimension 2

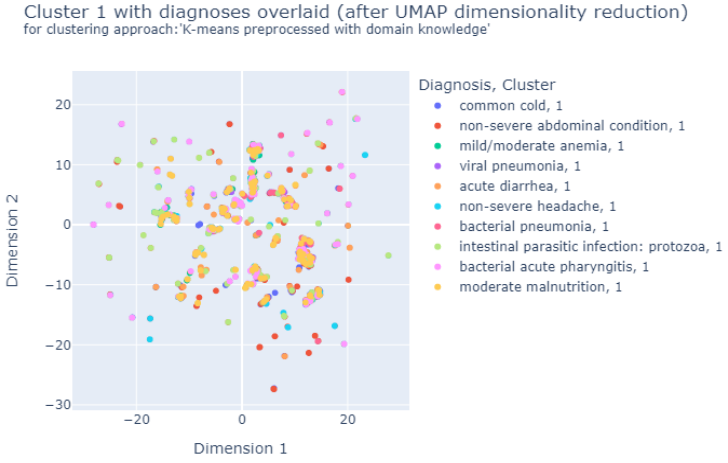

Cluster 7 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 7
non-severe abdominal condition, 7
mild/moderate anemia, 7
viral pneumonia, 7
acute diarrhea, 7
non-severe headache, 7
bacterial pneumonia, 7
intestinal parasitic infection: protozoa, 7
bacterial acute pharyngitis, 7
moderate malnutrition, 7
Dimension 1
Dimension 2

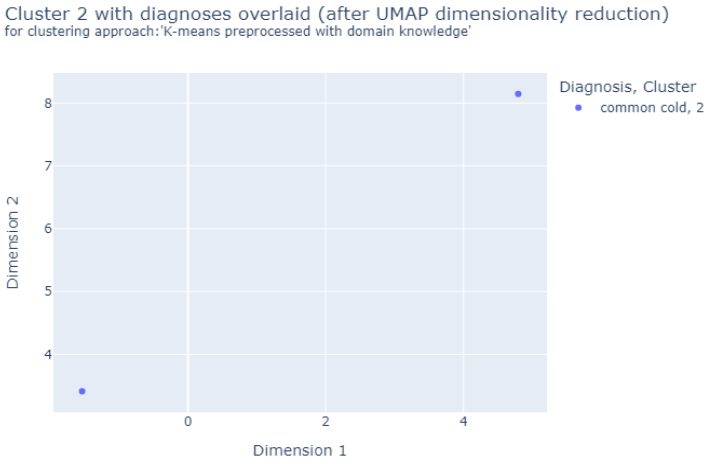

Cluster 8 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 8
non-severe abdominal condition, 8
mild/moderate anemia, 8
non-severe headache, 8
bacterial pneumonia, 8
intestinal parasitic infection: protozoa, 8
bacterial acute pharyngitis, 8
moderate malnutrition, 8
Dimension 2
Dimension 1
10
5
0
−5
−10
−15
−10
−5
0

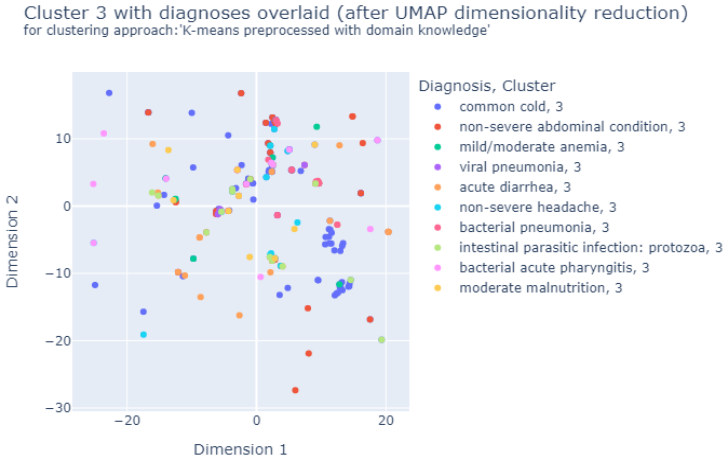

Cluster 9 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features'
Diagnosis, Cluster
common cold, 9
mild/moderate anemia, 9
Dimension 2
Dimension 1
4
2
0
−2
−4
−6
−8
−10
0
5

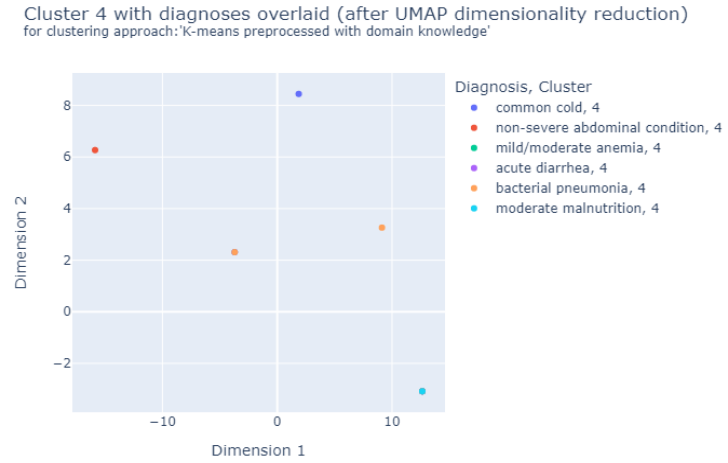
### *C.4.3 K-means preprocessed with domain knowledge*

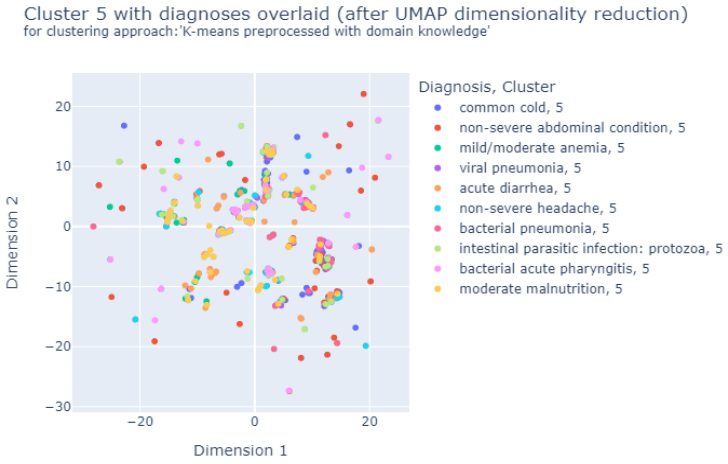


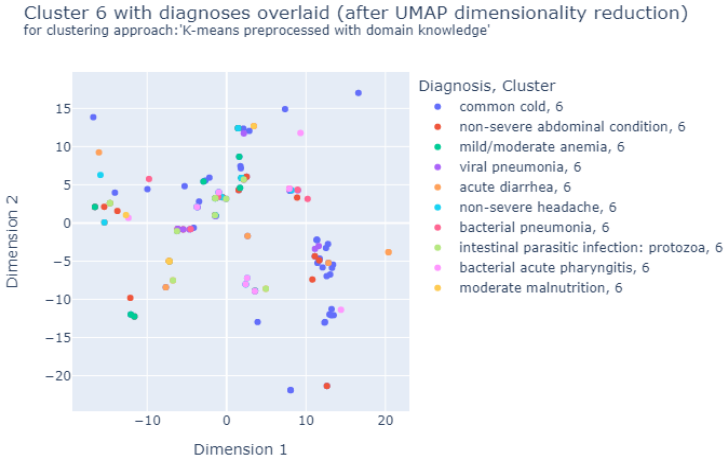



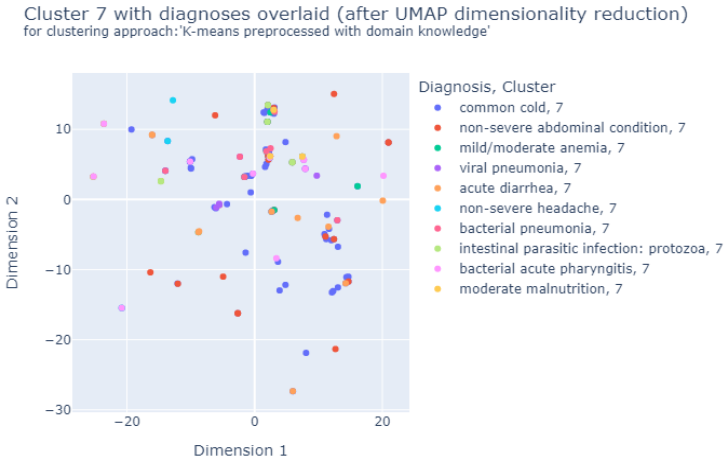

Cluster 2 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 2
Dimension 2
Dimension 1
8
7
6
5
4
0
2
4

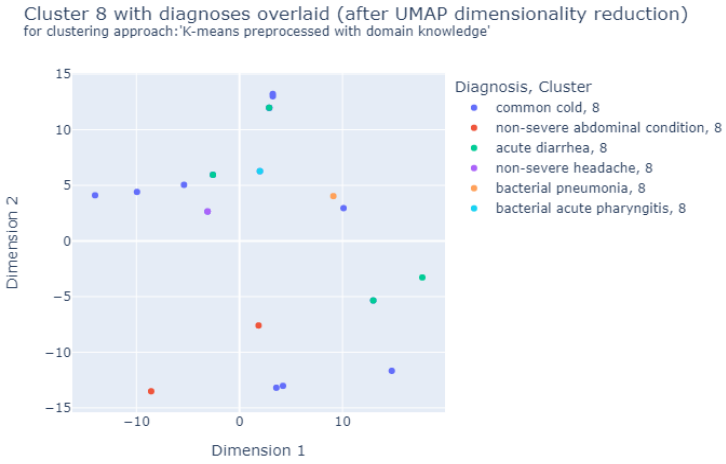

Cluster 3 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 3
non-severe abdominal condition, 3
mild/moderate anemia, 3
viral pneumonia, 3
acute diarrhea, 3
non-severe headache, 3
bacterial pneumonia, 3
intestinal parasitic infection: protozoa, 3
bacterial acute pharyngitis, 3
moderate malnutrition, 3
Dimension 2
Dimension 1
10
0
−10
−20
−30
−20
0
20

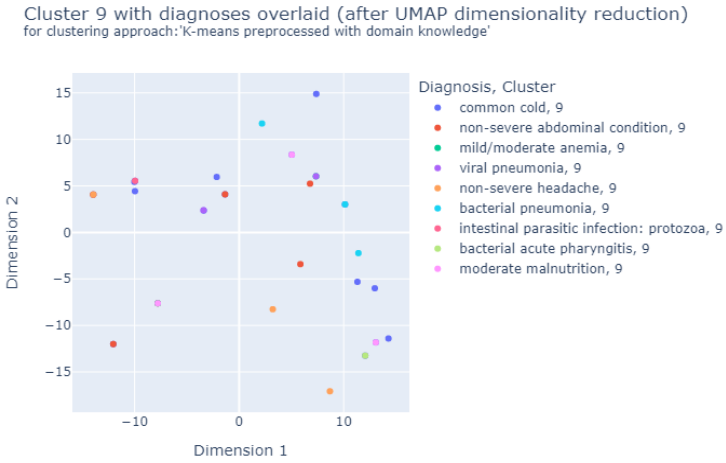

Cluster 4 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 4
non-severe abdominal condition, 4
mild/moderate anemia, 4
acute diarrhea, 4
bacterial pneumonia, 4
moderate malnutrition, 4
Dimension 2
Dimension 1
8
6
4
2
0
−2
−10
0
10


Cluster 5 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 5
non-severe abdominal condition, 5
mild/moderate anemia, 5
viral pneumonia, 5
acute diarrhea, 5
non-severe headache, 5
bacterial pneumonia, 5
intestinal parasitic infection: protozoa, 5
bacterial acute pharyngitis, 5
moderate malnutrition, 5
Dimension 2
Dimension 1

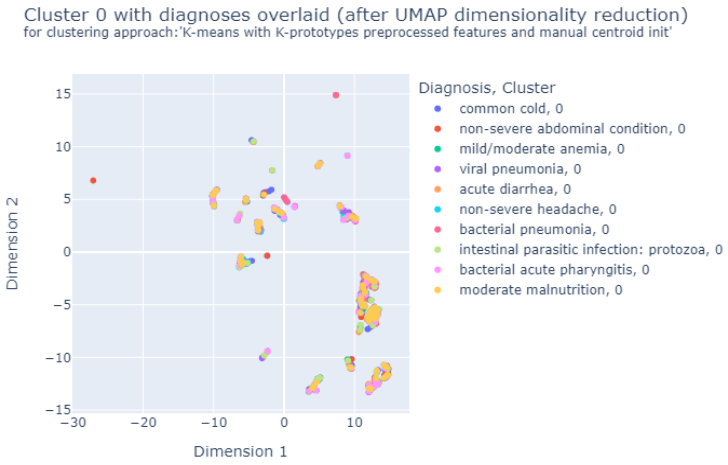

Cluster 6 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 6
non-severe abdominal condition, 6
mild/moderate anemia, 6
viral pneumonia, 6
acute diarrhea, 6
non-severe headache, 6
bacterial pneumonia, 6
intestinal parasitic infection: protozoa, 6
bacterial acute pharyngitis, 6
moderate malnutrition, 6
Dimension 2
Dimension 1

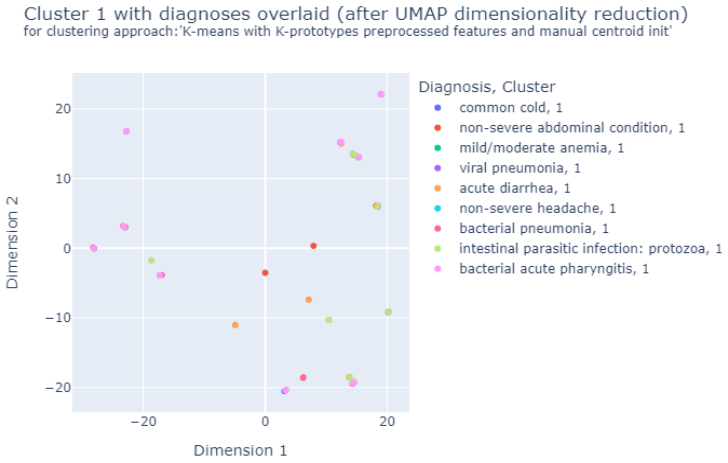

Cluster 7 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 7
non-severe abdominal condition, 7
mild/moderate anemia, 7
viral pneumonia, 7
acute diarrhea, 7
non-severe headache, 7
bacterial pneumonia, 7
intestinal parasitic infection: protozoa, 7
bacterial acute pharyngitis, 7
moderate malnutrition, 7
Dimension 2
Dimension 1

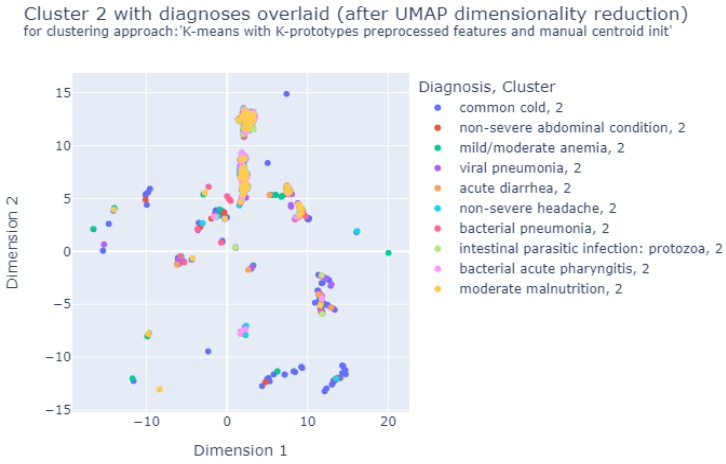

Cluster 8 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 8
non-severe abdominal condition, 8
acute diarrhea, 8
non-severe headache, 8
bacterial pneumonia, 8
bacterial acute pharyngitis, 8
Dimension 2
Dimension 1
15
10
5
0
−5
−10
−15
−10
0
10

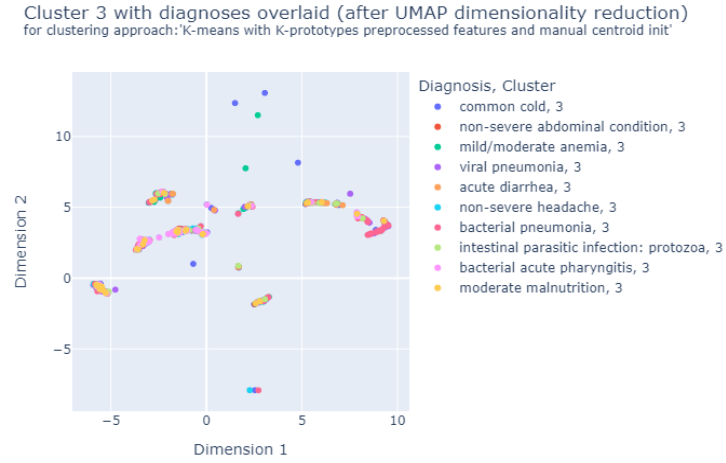

Cluster 9 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge'
Diagnosis, Cluster
common cold, 9
non-severe abdominal condition, 9
mild/moderate anemia, 9
viral pneumonia, 9
non-severe headache, 9
bacterial pneumonia, 9
intestinal parasitic infection: protozoa, 9
bacterial acute pharyngitis, 9
moderate malnutrition, 9
Dimension 2
Dimension 1
15
10
5
0
−5
−10
−15
−10
0
10

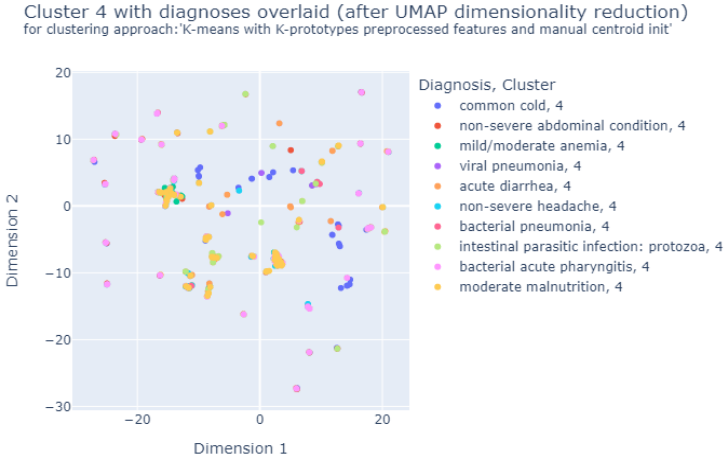
### *C.4.4 K-means with K-prototypes preprocessed features and manual centroid init*

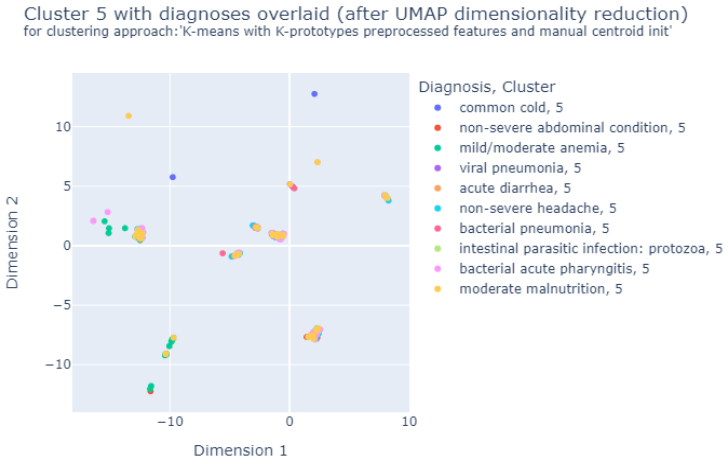


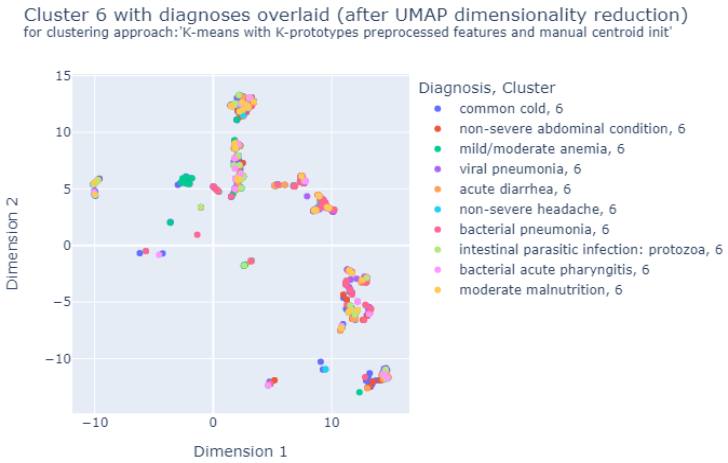



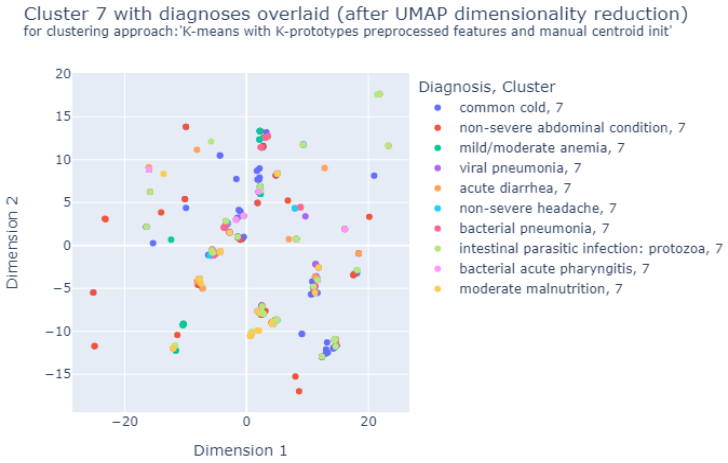

Cluster 2 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 2
non-severe abdominal condition, 2
mild/moderate anemia, 2
viral pneumonia, 2
acute diarrhea, 2
non-severe headache, 2
bacterial pneumonia, 2
intestinal parasitic infection: protozoa, 2
bacterial acute pharyngitis, 2
moderate malnutrition, 2
Dimension 2
Dimension 1

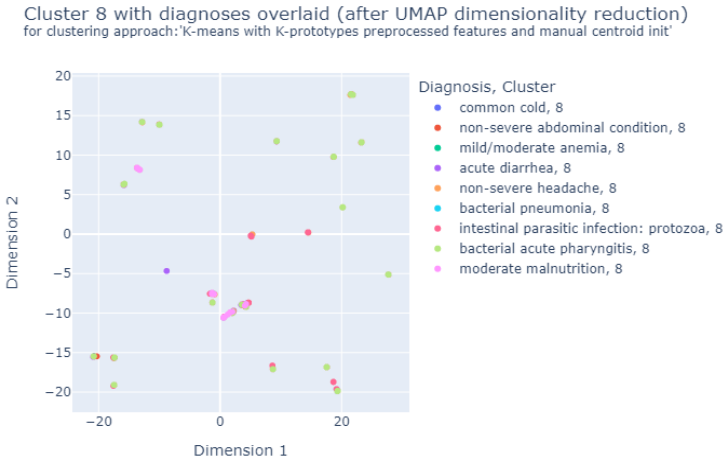

Cluster 3 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 3
non-severe abdominal condition, 3
mild/moderate anemia, 3
viral pneumonia, 3
acute diarrhea, 3
non-severe headache, 3
bacterial pneumonia, 3
intestinal parasitic infection: protozoa, 3
bacterial acute pharyngitis, 3
moderate malnutrition, 3
Dimension 2
Dimension 1

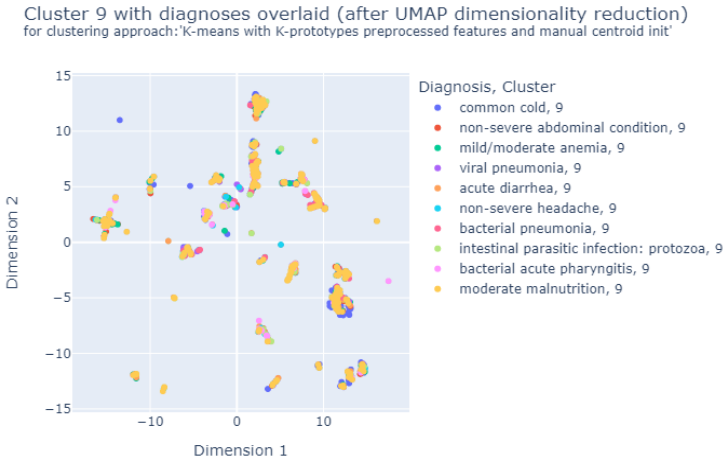

Cluster 4 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 4
non-severe abdominal condition, 4
mild/moderate anemia, 4
viral pneumonia, 4
acute diarrhea, 4
non-severe headache, 4
bacterial pneumonia, 4
intestinal parasitic infection: protozoa, 4
bacterial acute pharyngitis, 4
moderate malnutrition, 4
Dimension 2
Dimension 1


Cluster 5 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 5
non-severe abdominal condition, 5
mild/moderate anemia, 5
viral pneumonia, 5
acute diarrhea, 5
non-severe headache, 5
bacterial pneumonia, 5
intestinal parasitic infection: protozoa, 5
bacterial acute pharyngitis, 5
moderate malnutrition, 5
Dimension 1
Dimension 2

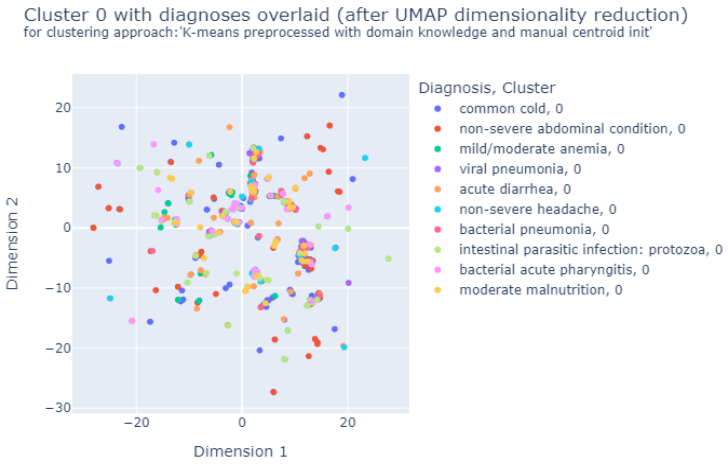

Cluster 6 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 6
non-severe abdominal condition, 6
mild/moderate anemia, 6
viral pneumonia, 6
acute diarrhea, 6
non-severe headache, 6
bacterial pneumonia, 6
intestinal parasitic infection: protozoa, 6
bacterial acute pharyngitis, 6
moderate malnutrition, 6
Dimension 1
Dimension 2

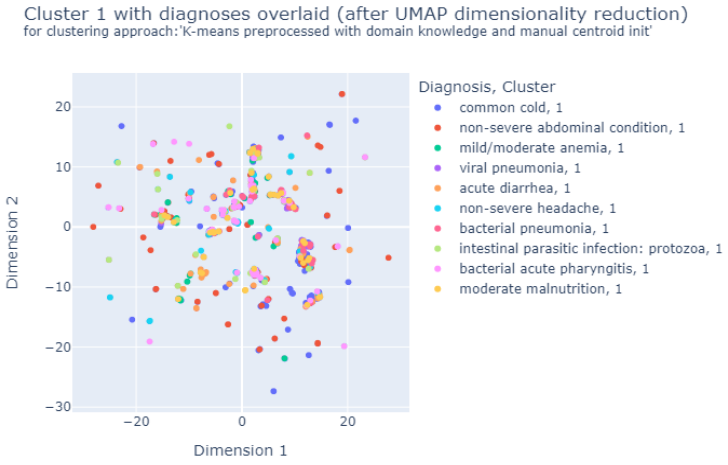

Cluster 7 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 7
non-severe abdominal condition, 7
mild/moderate anemia, 7
viral pneumonia, 7
acute diarrhea, 7
non-severe headache, 7
bacterial pneumonia, 7
intestinal parasitic infection: protozoa, 7
bacterial acute pharyngitis, 7
moderate malnutrition, 7
Dimension 1
Dimension 2

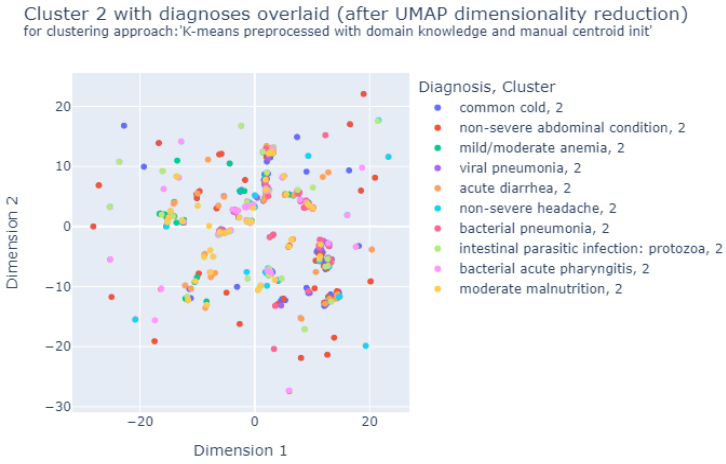

Cluster 8 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 8
non-severe abdominal condition, 8
mild/moderate anemia, 8
acute diarrhea, 8
non-severe headache, 8
bacterial pneumonia, 8
intestinal parasitic infection: protozoa, 8
bacterial acute pharyngitis, 8
moderate malnutrition, 8
Dimension 2
Dimension 1
20
15
10
5
0
−5
−10
−15
−20
−20
0
20

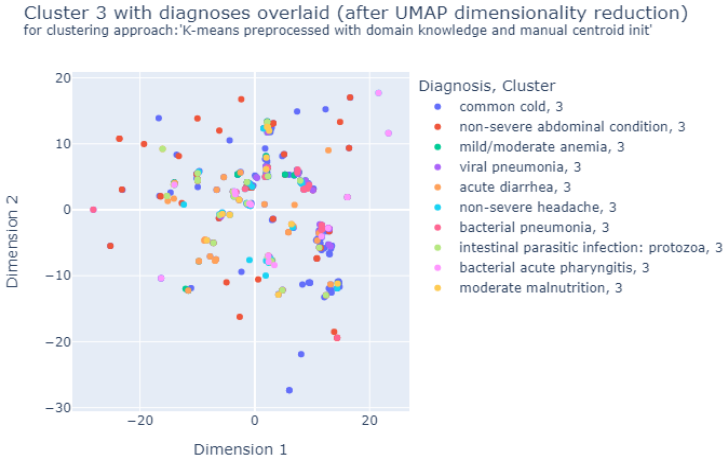

Cluster 9 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
Diagnosis, Cluster
common cold, 9
non-severe abdominal condition, 9
mild/moderate anemia, 9
viral pneumonia, 9
acute diarrhea, 9
non-severe headache, 9
bacterial pneumonia, 9
intestinal parasitic infection: protozoa, 9
bacterial acute pharyngitis, 9
moderate malnutrition, 9
Dimension 2
Dimension 1
15
10
5
0
−5
−10
−15
−10
0
10

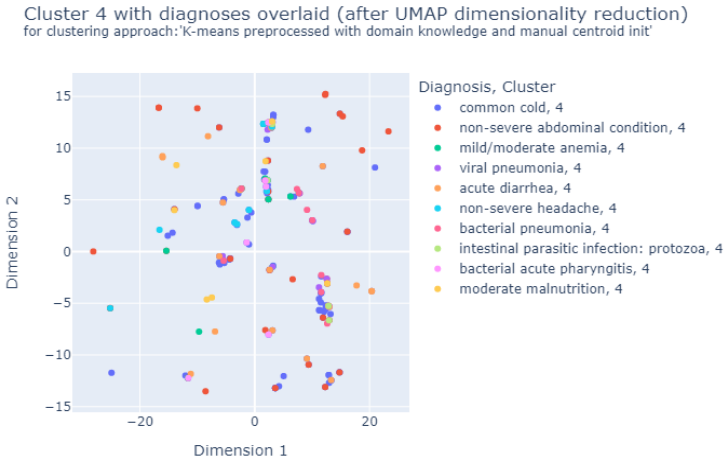
### *C.4.5 K-means preprocessed with domain knowledge and manual centroid init*

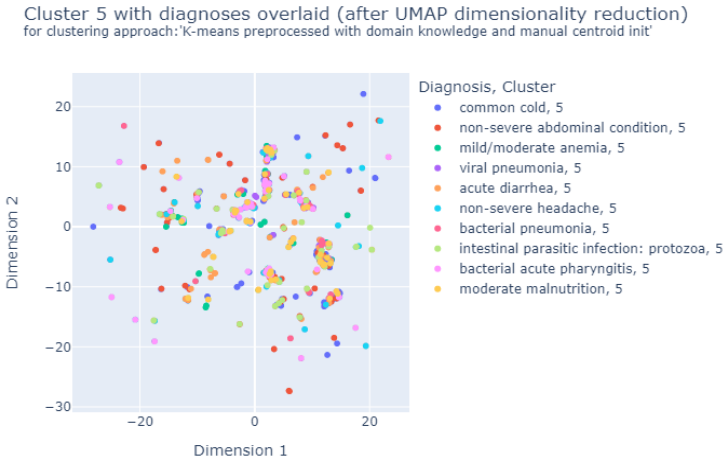


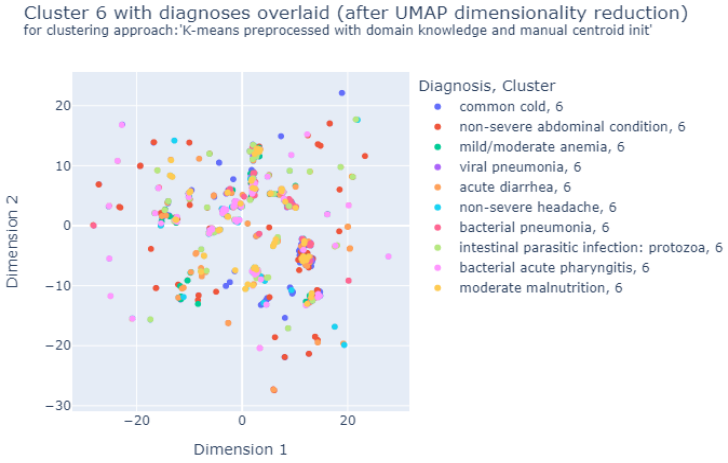



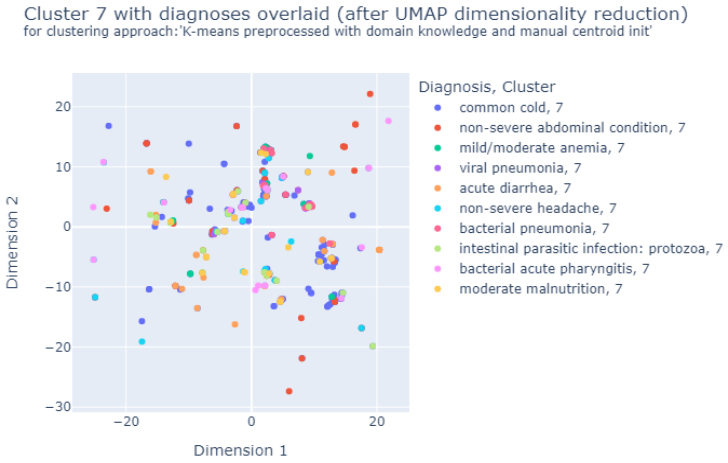

Cluster 2 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 2
non-severe abdominal condition, 2
mild/moderate anemia, 2
viral pneumonia, 2
acute diarrhea, 2
non-severe headache, 2
bacterial pneumonia, 2
intestinal parasitic infection: protozoa, 2
bacterial acute pharyngitis, 2
moderate malnutrition, 2
Dimension 2
Dimension 1

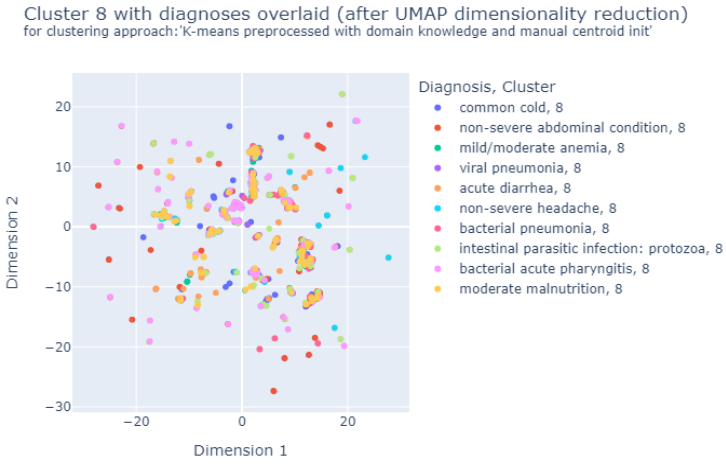

Cluster 3 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 3
non-severe abdominal condition, 3
mild/moderate anemia, 3
viral pneumonia, 3
acute diarrhea, 3
non-severe headache, 3
bacterial pneumonia, 3
intestinal parasitic infection: protozoa, 3
bacterial acute pharyngitis, 3
moderate malnutrition, 3
Dimension 2
Dimension 1

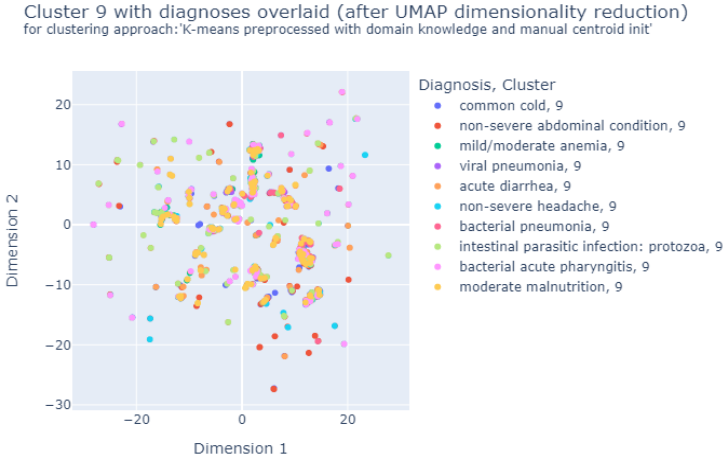

Cluster 4 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 4
non-severe abdominal condition, 4
mild/moderate anemia, 4
viral pneumonia, 4
acute diarrhea, 4
non-severe headache, 4
bacterial pneumonia, 4
intestinal parasitic infection: protozoa, 4
bacterial acute pharyngitis, 4
moderate malnutrition, 4
Dimension 2
Dimension 1


Cluster 5 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 5
non-severe abdominal condition, 5
mild/moderate anemia, 5
viral pneumonia, 5
acute diarrhea, 5
non-severe headache, 5
bacterial pneumonia, 5
intestinal parasitic infection: protozoa, 5
bacterial acute pharyngitis, 5
moderate malnutrition, 5
Dimension 2
Dimension 1
20
10
0
−10
−20
−30
−20
0
20


Cluster 6 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 6
non-severe abdominal condition, 6
mild/moderate anemia, 6
viral pneumonia, 6
acute diarrhea, 6
non-severe headache, 6
bacterial pneumonia, 6
intestinal parasitic infection: protozoa, 6
bacterial acute pharyngitis, 6
moderate malnutrition, 6
Dimension 2
Dimension 1
20
10
0
−10
−20
−30
−20
0
20


Cluster 7 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 7
non-severe abdominal condition, 7
mild/moderate anemia, 7
viral pneumonia, 7
acute diarrhea, 7
non-severe headache, 7
bacterial pneumonia, 7
intestinal parasitic infection: protozoa, 7
bacterial acute pharyngitis, 7
moderate malnutrition, 7
Dimension 2
Dimension 1
20
10
0
−10
−20
−30
−20
0
20


Cluster 8 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 8
non-severe abdominal condition, 8
mild/moderate anemia, 8
viral pneumonia, 8
acute diarrhea, 8
non-severe headache, 8
bacterial pneumonia, 8
intestinal parasitic infection: protozoa, 8
bacterial acute pharyngitis, 8
moderate malnutrition, 8
Dimension 2
Dimension 1
20
10
0
−10
−20
−30
−20
0
20

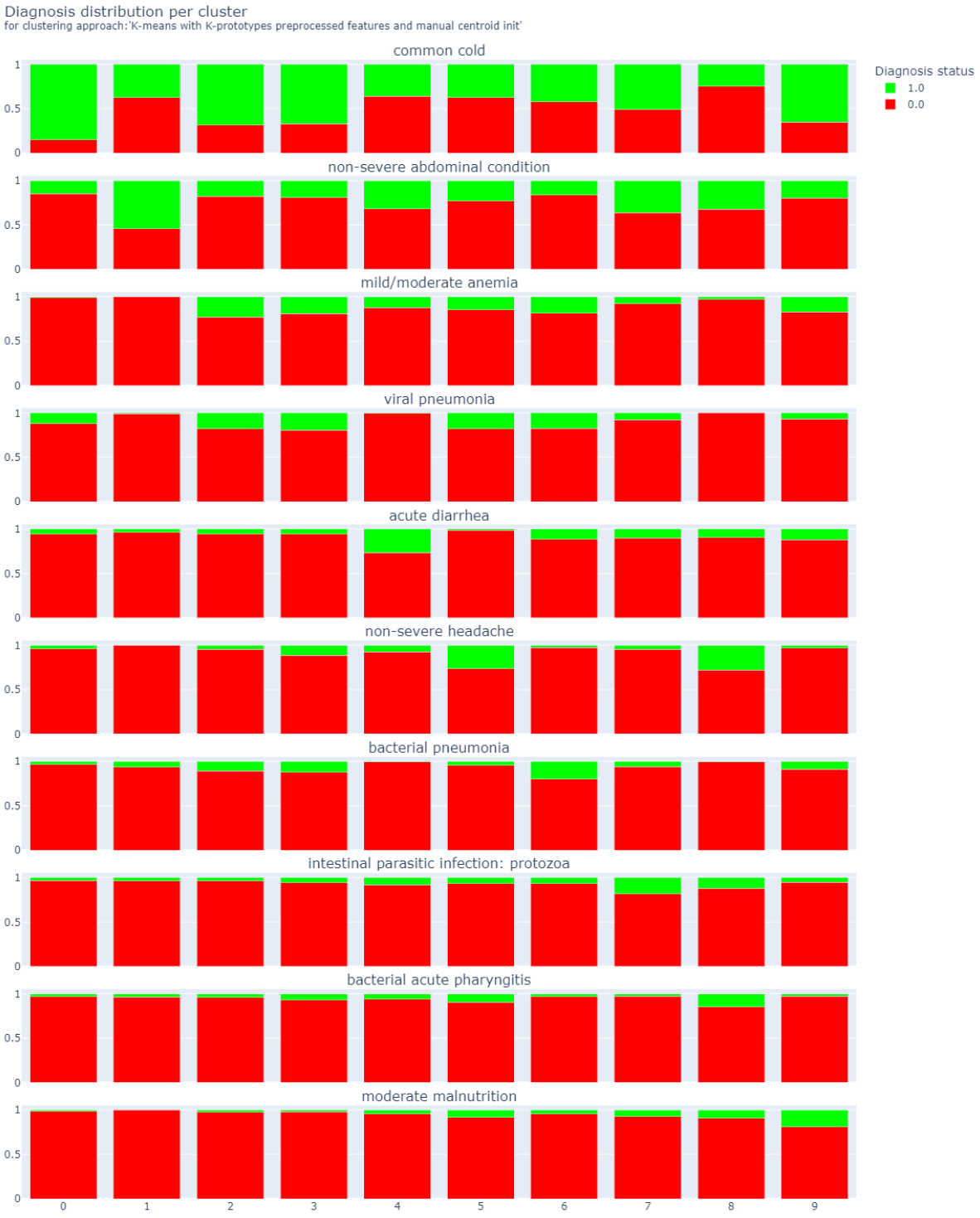

Cluster 9 with diagnoses overlaid (after UMAP dimensionality reduction)
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
Diagnosis, Cluster
common cold, 9
non-severe abdominal condition, 9
mild/moderate anemia, 9
viral pneumonia, 9
acute diarrhea, 9
non-severe headache, 9
bacterial pneumonia, 9
intestinal parasitic infection: protozoa, 9
bacterial acute pharyngitis, 9
moderate malnutrition, 9
Dimension 2
Dimension 1
20
10
0
−10
−20
−30
−20
0
20

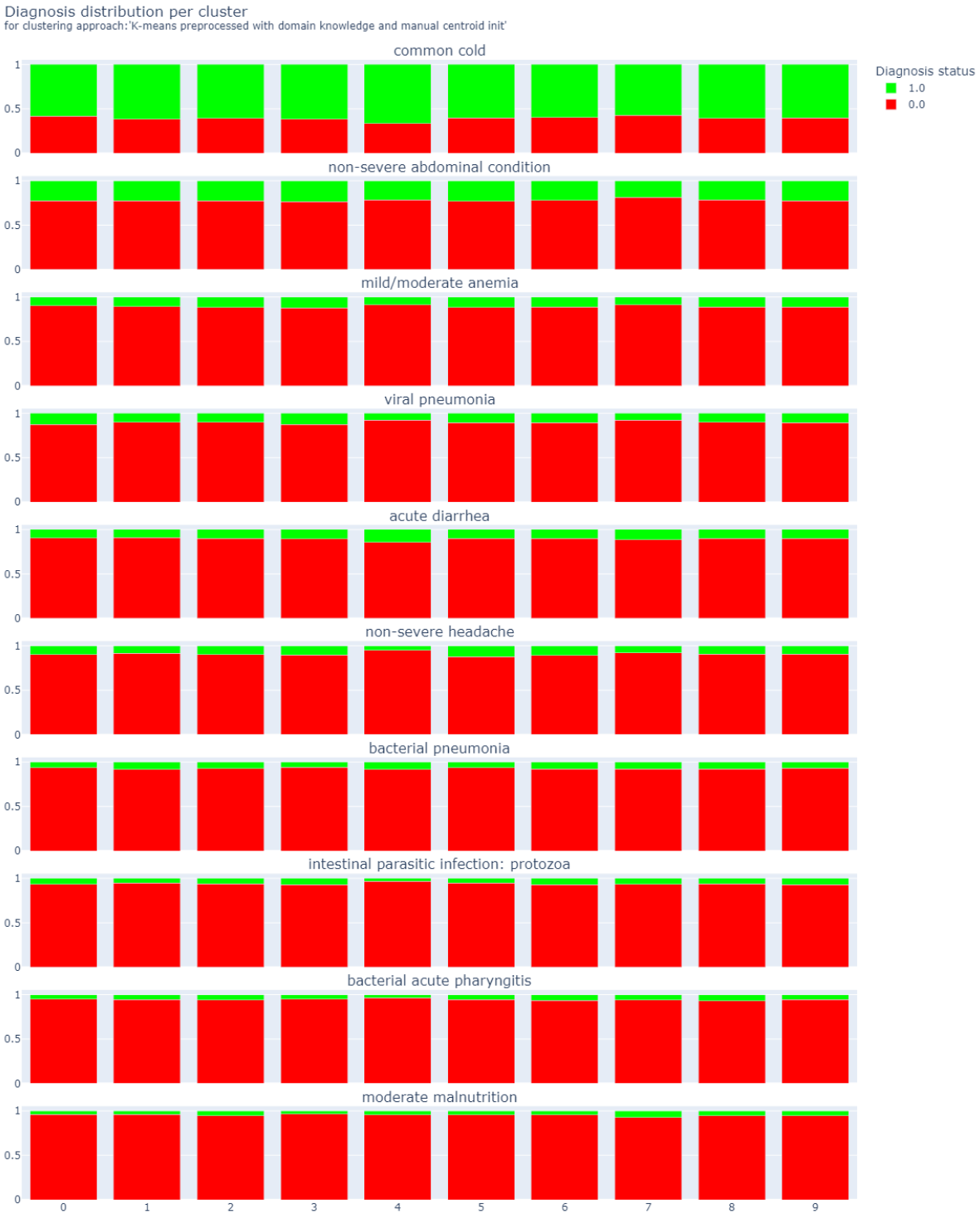
## C.5 Descriptive statistics of clusters

### *C.5.1 Diagnosis distribution per cluster*

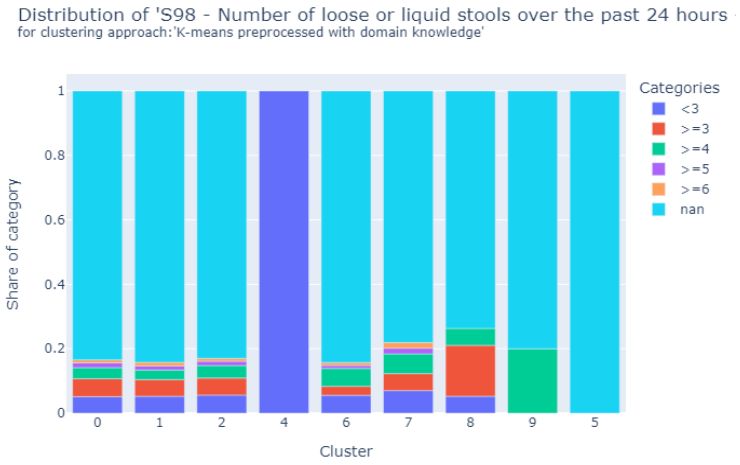

Diagnosis distribution per cluster
for clustering approach:'K-prototypes preprocessed features'
common cold
non-severe abdominal condition
mild/moderate anemia
viral pneumonia
acute diarrhea
non-severe headache
bacterial pneumonia
intestinal parasitic infection: protozoa
bacterial acute pharyngitis
moderate malnutrition
Diagnosis status
1.0
0.0
0
0.5
1
0
1
2
3
4
5
6
7
8
9

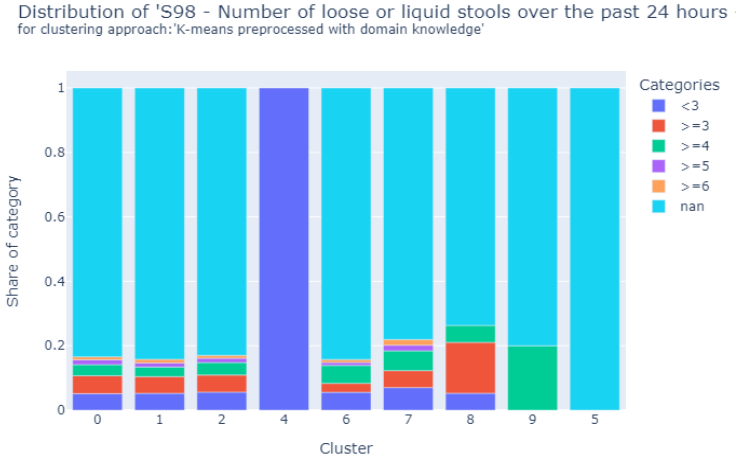

Diagnosis distribution per cluster
for clustering approach:'K-means preprocessed with domain knowledge'
common cold
non-severe abdominal condition
mild/moderate anemia
viral pneumonia
acute diarrhea
non-severe headache
bacterial pneumonia
intestinal parasitic infection: protozoa
bacterial acute pharyngitis
moderate malnutrition
Diagnosis status
1.0
0.0
1
0.5
0
0
1
3
4
5
6
7
8
9
2


Diagnosis distribution per cluster
for clustering approach:'K-means with K-prototypes preprocessed features and manual centroid init'
common cold
non-severe abdominal condition
mild/moderate anemia
viral pneumonia
acute diarrhea
non-severe headache
bacterial pneumonia
intestinal parasitic infection: protozoa
bacterial acute pharyngitis
moderate malnutrition
Diagnosis status
1.0
0.0
0
0.5
1
0
1
2
3
4
5
6
7
8
9

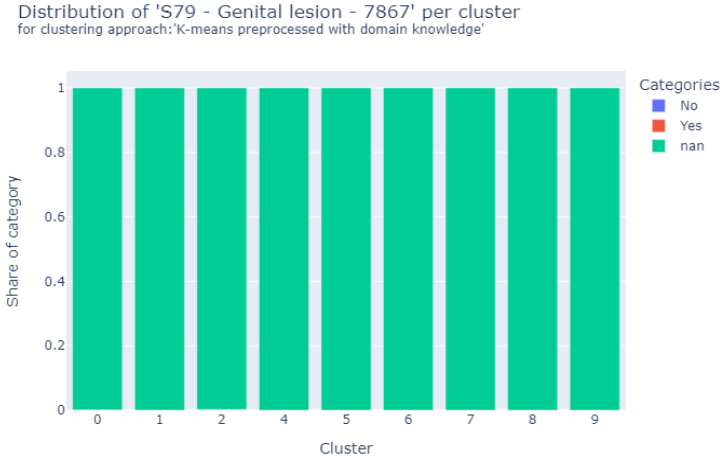

Diagnosis distribution per cluster
for clustering approach:'K-means preprocessed with domain knowledge and manual centroid init'
common cold
non-severe abdominal condition
mild/moderate anemia
viral pneumonia
acute diarrhea
non-severe headache
bacterial pneumonia
intestinal parasitic infection: protozoa
bacterial acute pharyngitis
moderate malnutrition
Diagnosis status
1.0
0.0
0
0.5
1
0
1
2
3
4
5
6
7
8
9

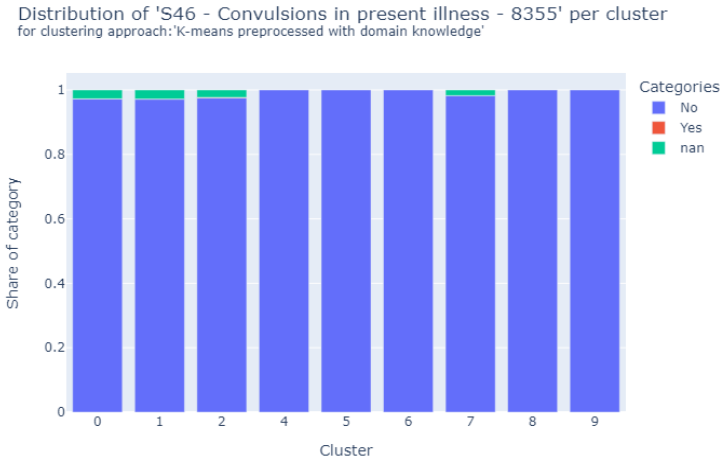
### *C.5.2 Exemplary features with distribution (partially) corresponding to clusters*



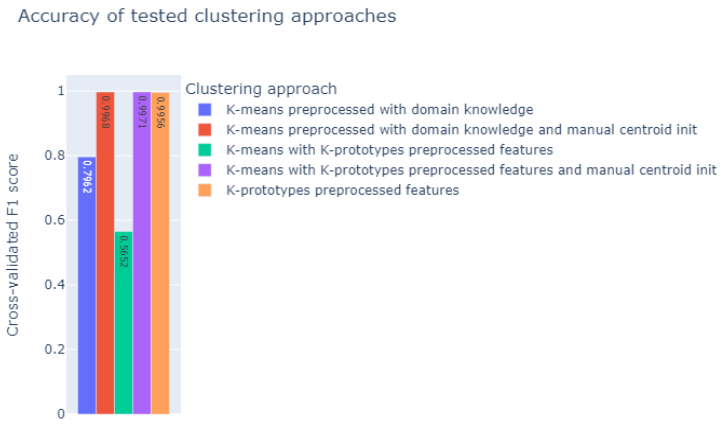

### *C.5.3 Exemplary redundant features*

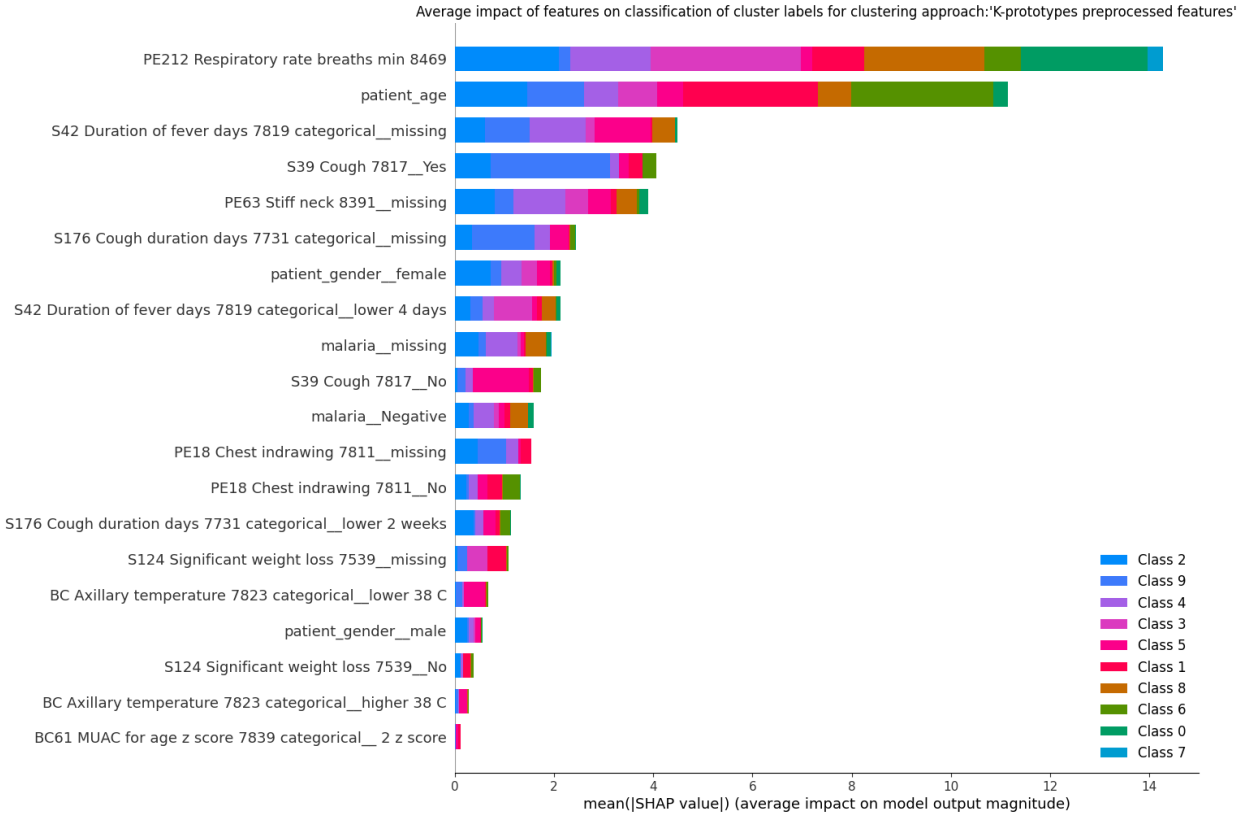


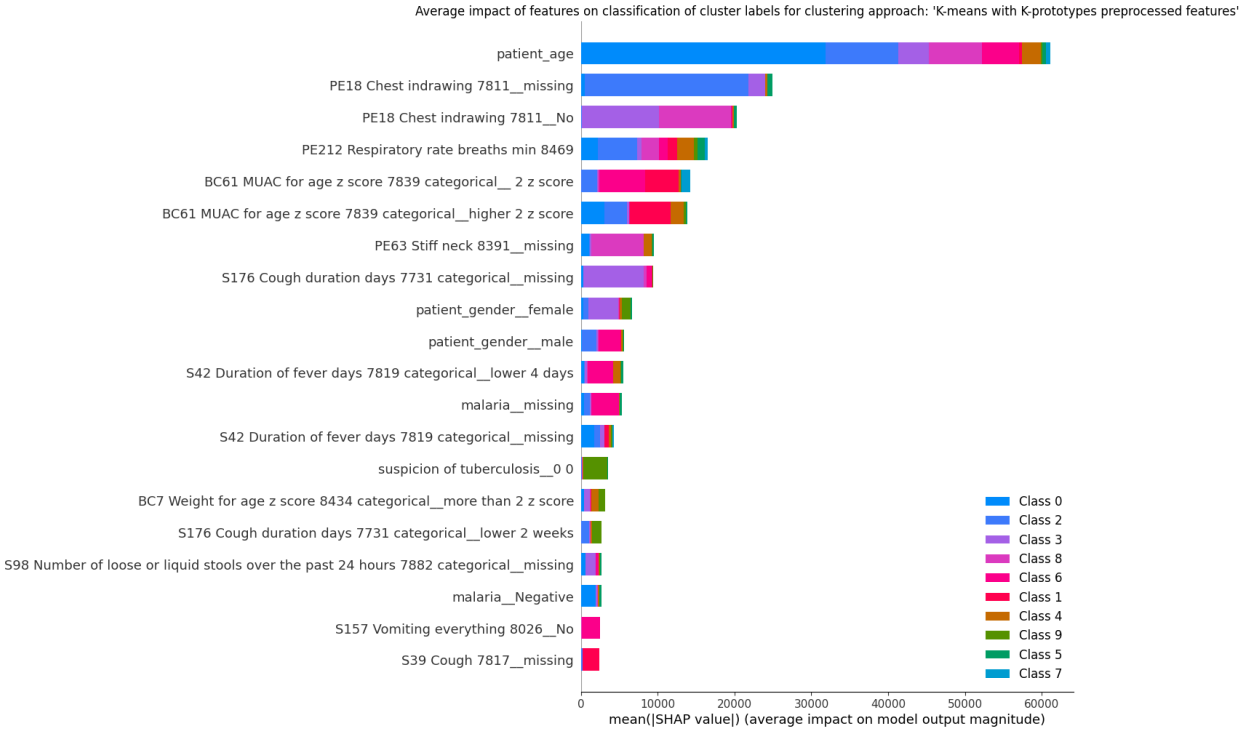


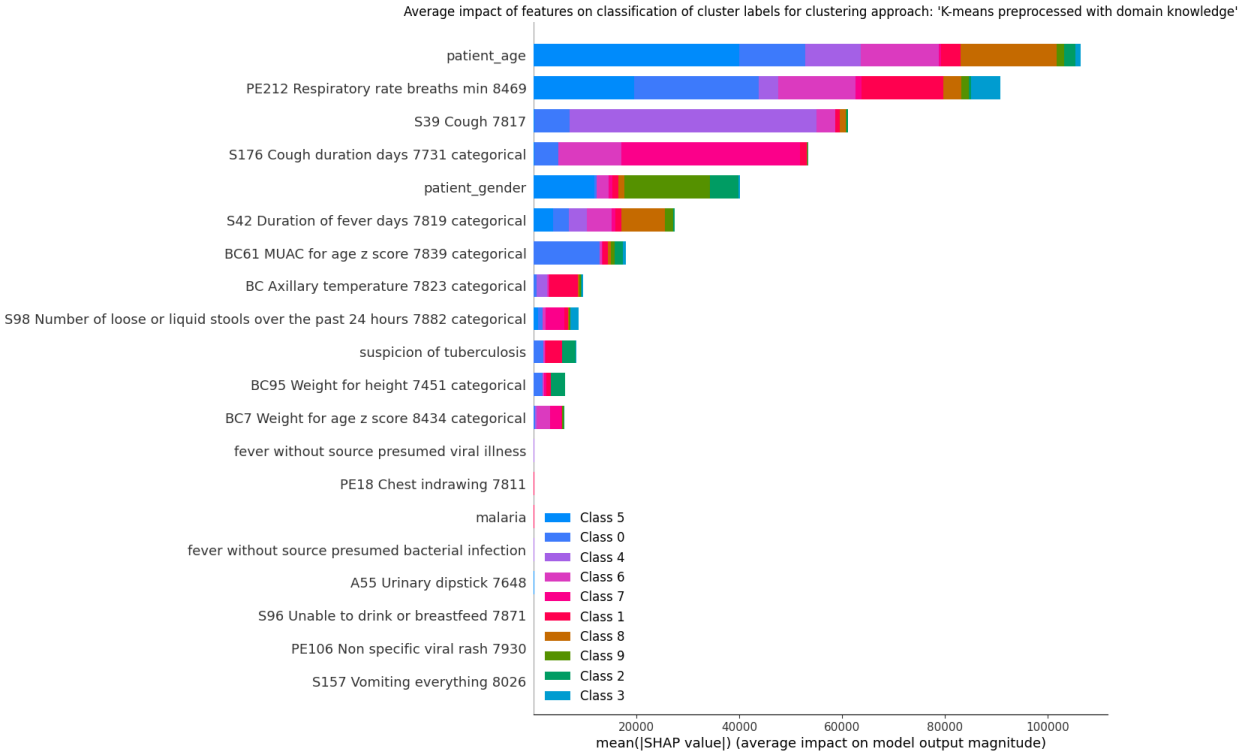

## C.6 Accuracy and SHAP values of classifier for cluster labels

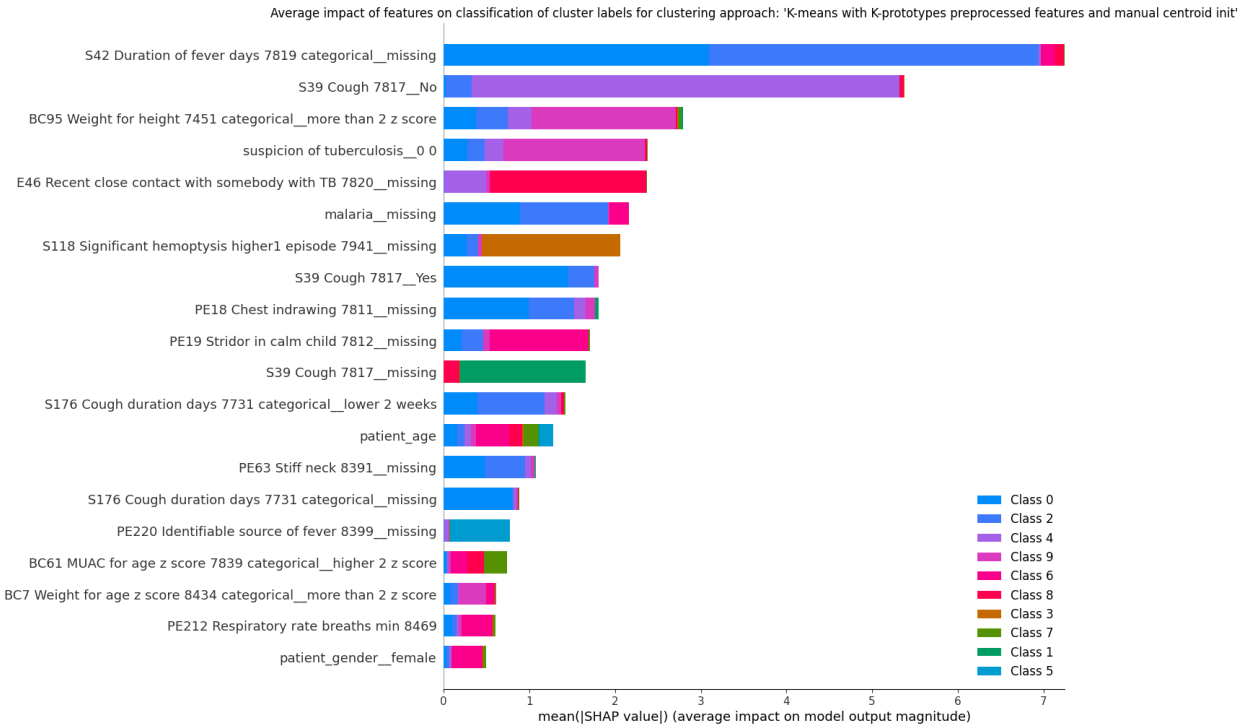
### *C.6.1 Feature based classifier*

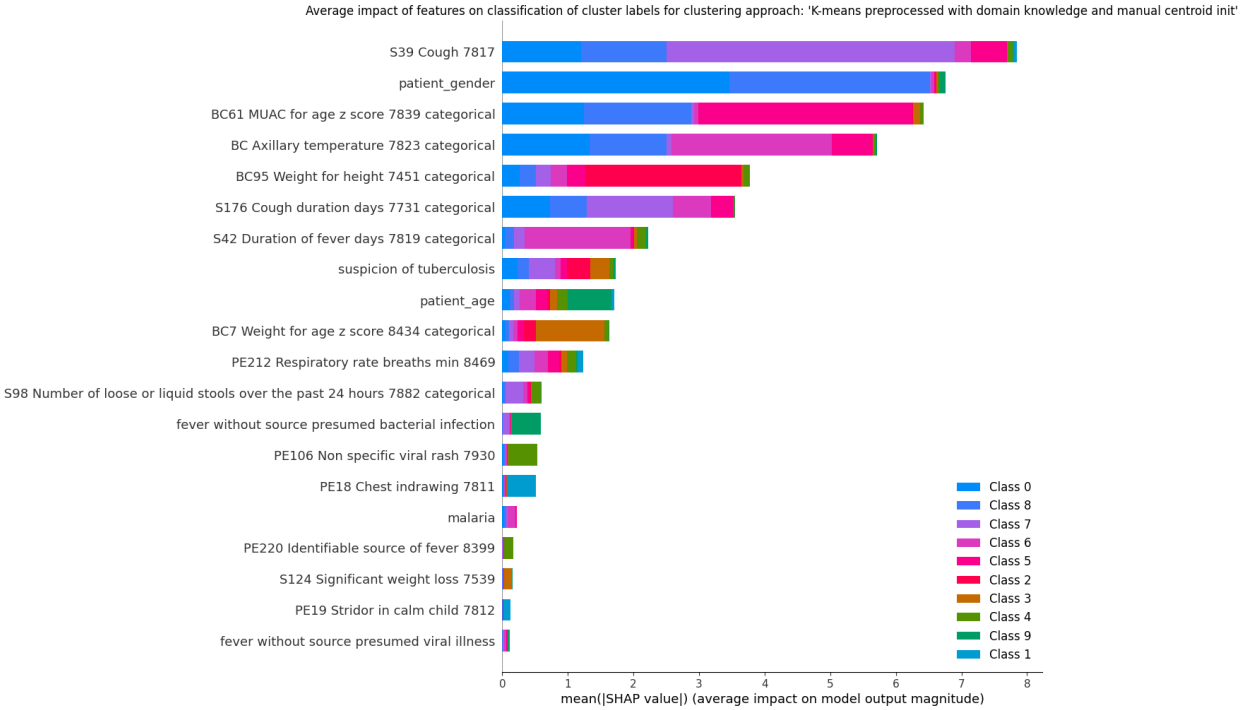



Average impact of features on classification of cluster labels for clustering approach:'K-prototypes preprocessed features'
PE212 Respiratory rate breaths min 8469
patient_age
S42 Duration of fever days 7819 categorical__missing
S39 Cough 7817__Yes
PE63 Stiff neck 8391__missing
S176 Cough duration days 7731 categorical__missing
patient_gender__female
S42 Duration of fever days 7819 categorical__lower 4 days
malaria__missing
S39 Cough 7817__No
malaria__Negative
PE18 Chest indrawing 7811__missing
PE18 Chest indrawing 7811__No
S176 Cough duration days 7731 categorical__lower 2 weeks
S124 Significant weight loss 7539__missing
BC Axillary temperature 7823 categorical__lower 38 C
patient_gender__male
S124 Significant weight loss 7539__No
BC Axillary temperature 7823 categorical__higher 38 C
BC61 MUAC for age z score 7839 categorical__ 2 z score
Class 2
Class 9
Class 4
Class 3
Class 5
Class 1
Class 8
Class 6
Class 0
Class 7
0
2
4
6
8
10
12
14
mean(|SHAP value|) (average impact on model output magnitude)

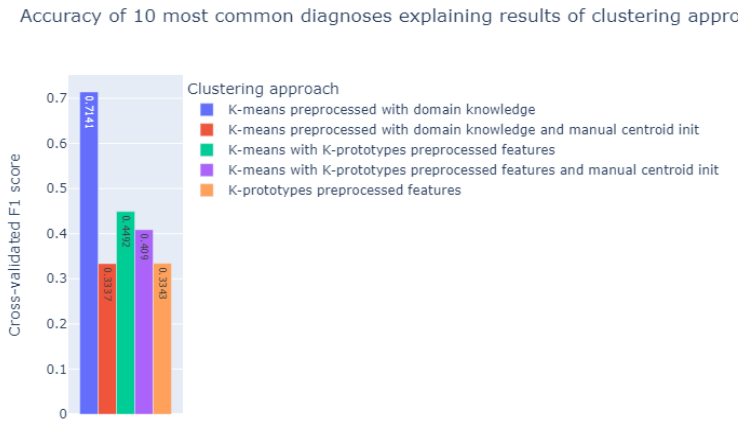

Average impact of features on classification of cluster labels for clustering approach: 'K-means with K-prototypes preprocessed features'
patient_age
PE18 Chest indrawing 7811__missing
PE18 Chest indrawing 7811__No
PE212 Respiratory rate breaths min 8469
BC61 MUAC for age z score 7839 categorical__ 2 z score
BC61 MUAC for age z score 7839 categorical__higher 2 z score
PE63 Stiff neck 8391__missing
S176 Cough duration days 7731 categorical__missing
patient_gender__female
patient_gender__male
S42 Duration of fever days 7819 categorical__lower 4 days
malaria__missing
S42 Duration of fever days 7819 categorical__missing
suspicion of tuberculosis__0 0
BC7 Weight for age z score 8434 categorical__more than 2 z score
S176 Cough duration days 7731 categorical__lower 2 weeks
S98 Number of loose or liquid stools over the past 24 hours 7882 categorical__missing
malaria__Negative
S157 Vomiting everything 8026__No
S39 Cough 7817__missing
Class 0
Class 2
Class 3
Class 8
Class 6
Class 1
Class 4
Class 9
Class 5
Class 7
0
10000
20000
30000
40000
50000
60000
mean(|SHAP value|) (average impact on model output magnitude)

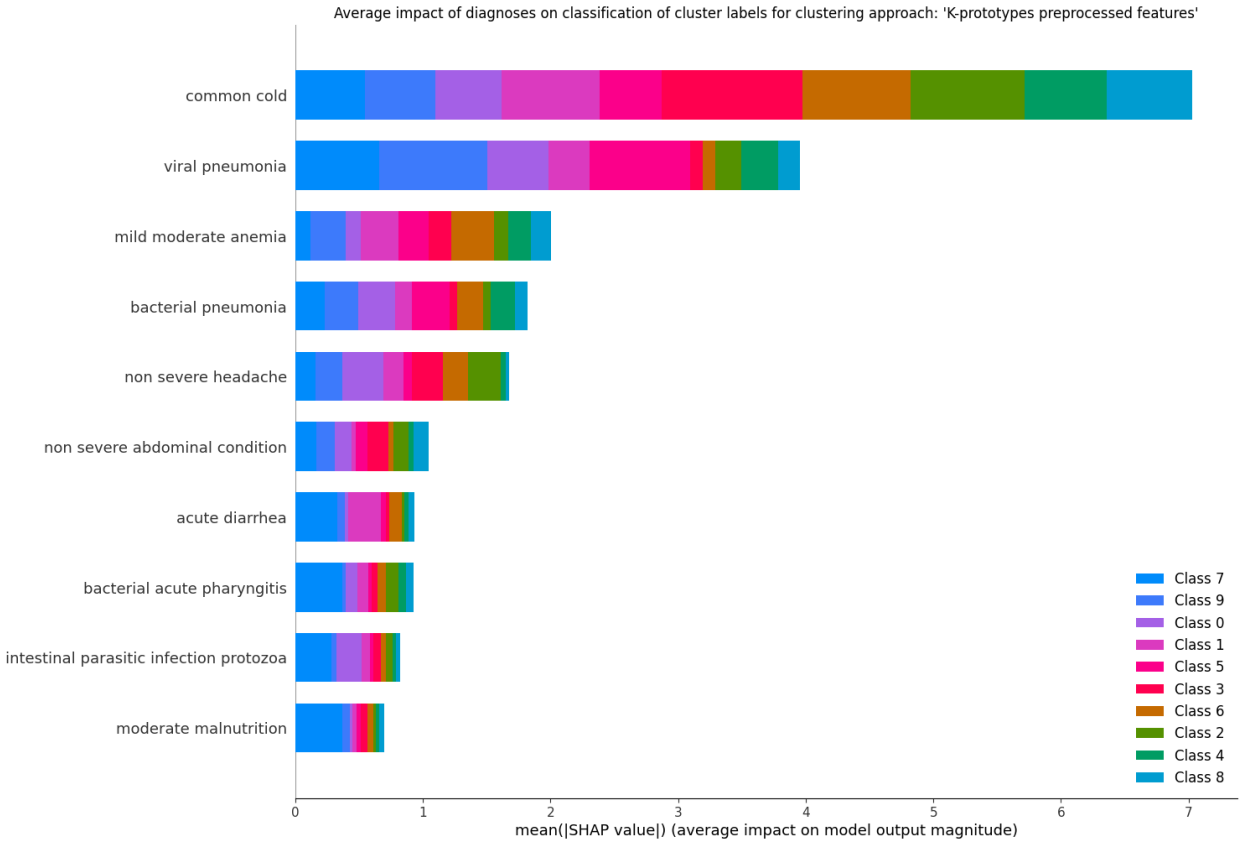

Average impact of features on classification of cluster labels for clustering approach: 'K-means preprocessed with domain knowledge'
patient_age
PE212 Respiratory rate breaths min 8469
S39 Cough 7817
S176 Cough duration days 7731 categorical
patient_gender
S42 Duration of fever days 7819 categorical
BC61 MUAC for age z score 7839 categorical
BC Axillary temperature 7823 categorical
S98 Number of loose or liquid stools over the past 24 hours 7882 categorical
suspicion of tuberculosis
BC95 Weight for height 7451 categorical
BC7 Weight for age z score 8434 categorical
fever without source presumed viral illness
PE18 Chest indrawing 7811
malaria
fever without source presumed bacterial infection
A55 Urinary dipstick 7648
S96 Unable to drink or breastfeed 7871
PE106 Non specific viral rash 7930
S157 Vomiting everything 8026
Class 5
Class 0
Class 4
Class 6
Class 7
Class 1
Class 8
Class 9
Class 2
Class 3
20000
40000
60000
80000
100000
mean(|SHAP value|) (average impact on model output magnitude)

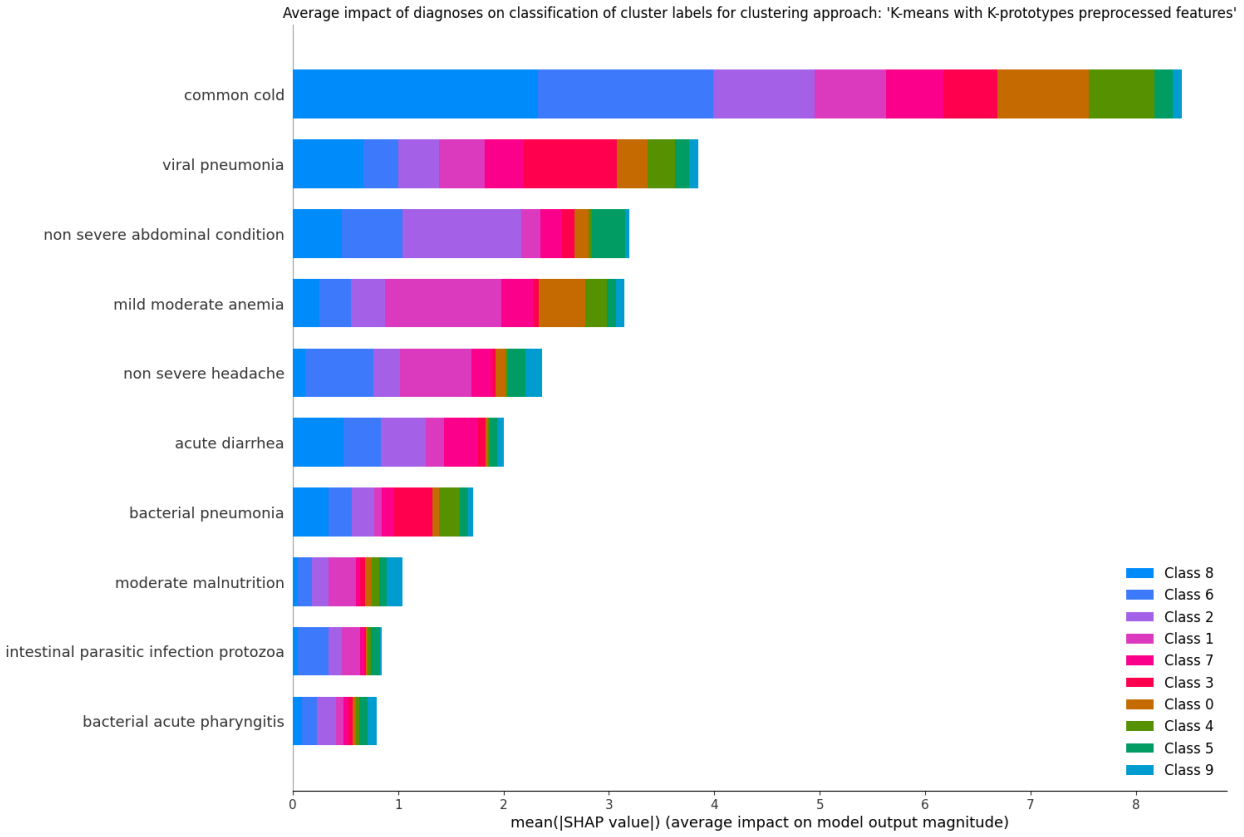

Average impact of features on classification of cluster labels for clustering approach: 'K-means with K-prototypes preprocessed features and manual centroid init'
S42 Duration of fever days 7819 categorical__missing
S39 Cough 7817__No
BC95 Weight for height 7451 categorical__more than 2 z score
suspicion of tuberculosis__0 0
E46 Recent close contact with somebody with TB 7820__missing
malaria__missing
S118 Significant hemoptysis higher1 episode 7941__missing
S39 Cough 7817__Yes
PE18 Chest indrawing 7811__missing
PE19 Stridor in calm child 7812__missing
S39 Cough 7817__missing
S176 Cough duration days 7731 categorical__lower 2 weeks
patient_age
PE63 Stiff neck 8391__missing
S176 Cough duration days 7731 categorical__missing
PE220 Identifiable source of fever 8399__missing
BC61 MUAC for age z score 7839 categorical__higher 2 z score
BC7 Weight for age z score 8434 categorical__more than 2 z score
PE212 Respiratory rate breaths min 8469
patient_gender__female
Class 0
Class 2
Class 4
Class 9
Class 6
Class 8
Class 3
Class 7
Class 1
Class 5
0
1
2
3
4
5
6
7
mean(|SHAP value|) (average impact on model output magnitude)

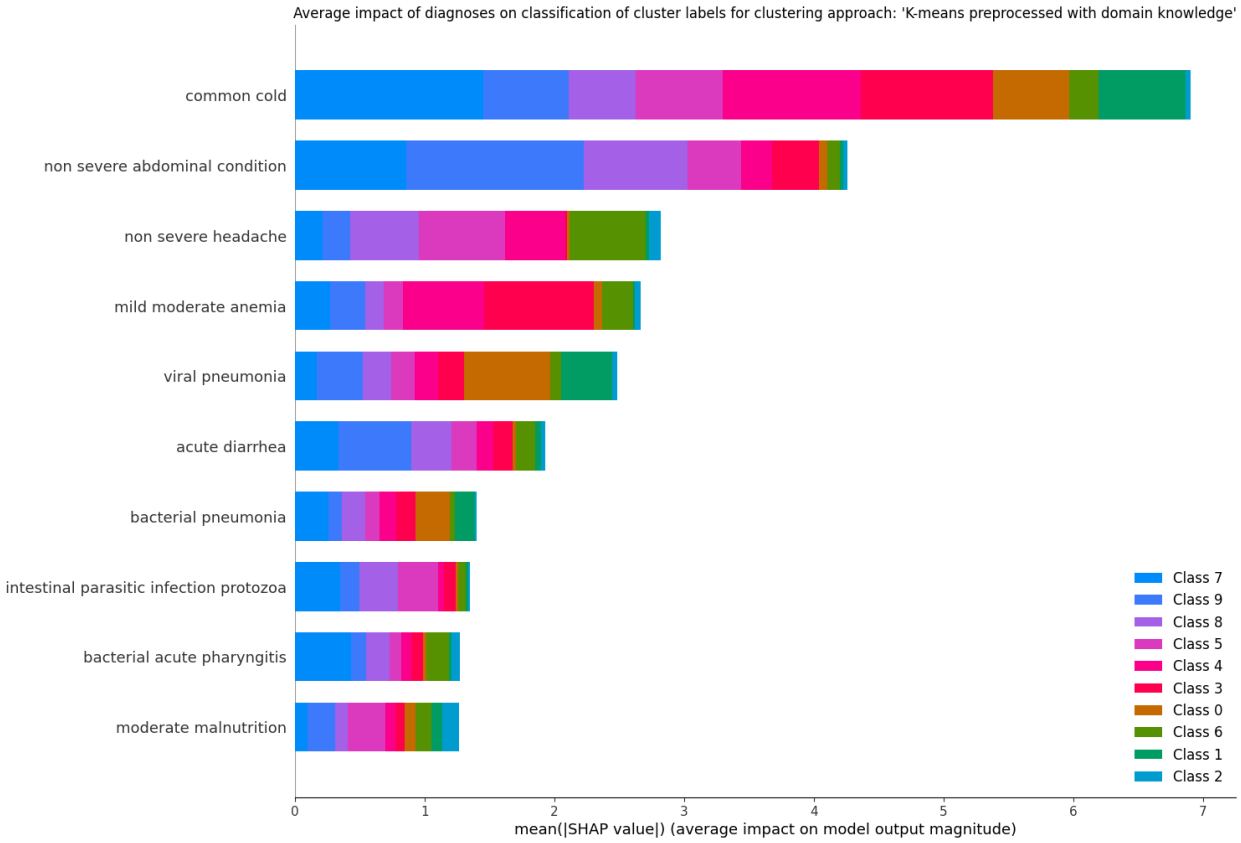


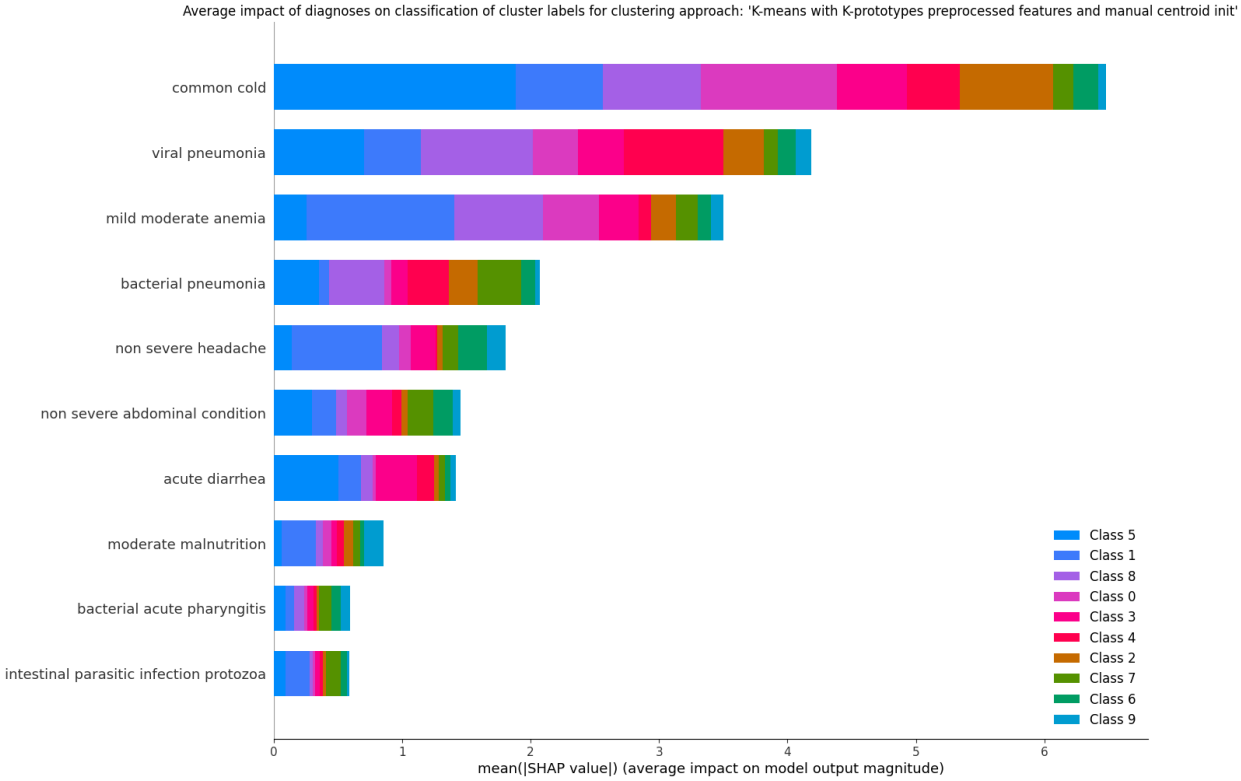

### *C.6.2 Diagnosis based classifier*

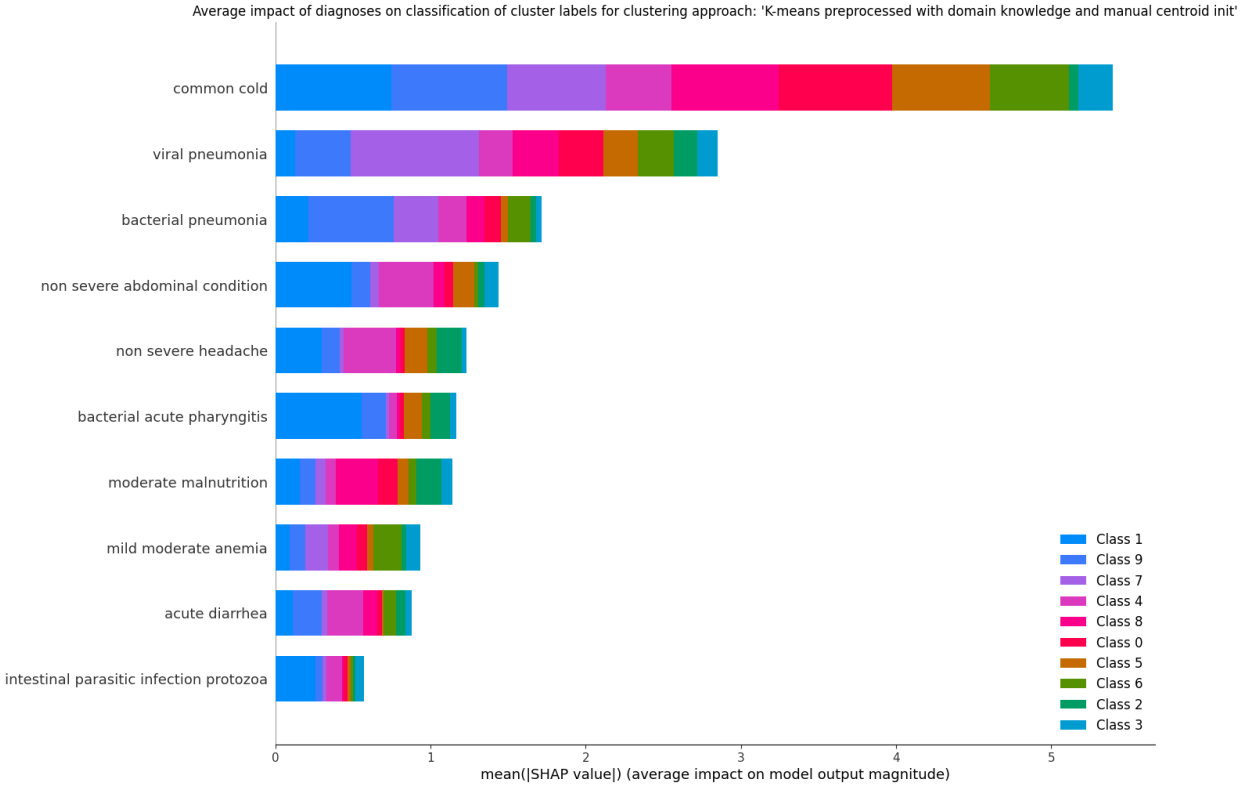



Average impact of diagnoses on classification of cluster labels for clustering approach: 'K-prototypes preprocessed features'
common cold
viral pneumonia
mild moderate anemia
bacterial pneumonia
non severe headache
non severe abdominal condition
acute diarrhea
bacterial acute pharyngitis
intestinal parasitic infection protozoa
moderate malnutrition
Class 7
Class 9
Class 0
Class 1
Class 5
Class 3
Class 6
Class 2
Class 4
Class 8
0
1
2
3
4
5
6
7
mean(|SHAP value|) (average impact on model output magnitude)

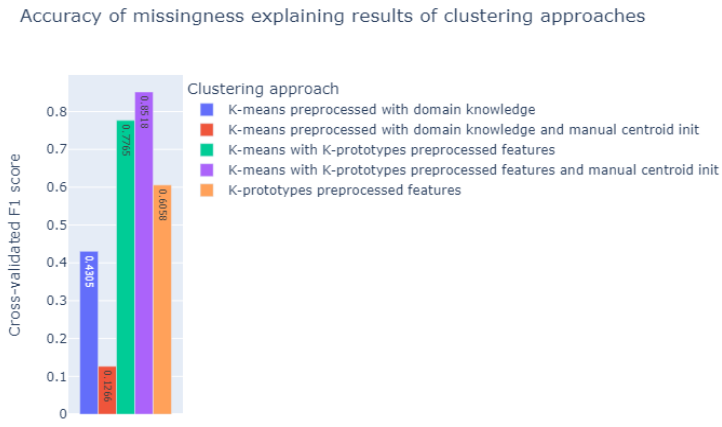

Average impact of diagnoses on classification of cluster labels for clustering approach: 'K-means with K-prototypes preprocessed features'
common cold
viral pneumonia
non severe abdominal condition
mild moderate anemia
non severe headache
acute diarrhea
bacterial pneumonia
moderate malnutrition
intestinal parasitic infection protozoa
bacterial acute pharyngitis
Class 8
Class 6
Class 2
Class 1
Class 7
Class 3
Class 0
Class 4
Class 5
Class 9
0
1
2
3
4
5
6
7
8
mean(|SHAP value|) (average impact on model output magnitude)

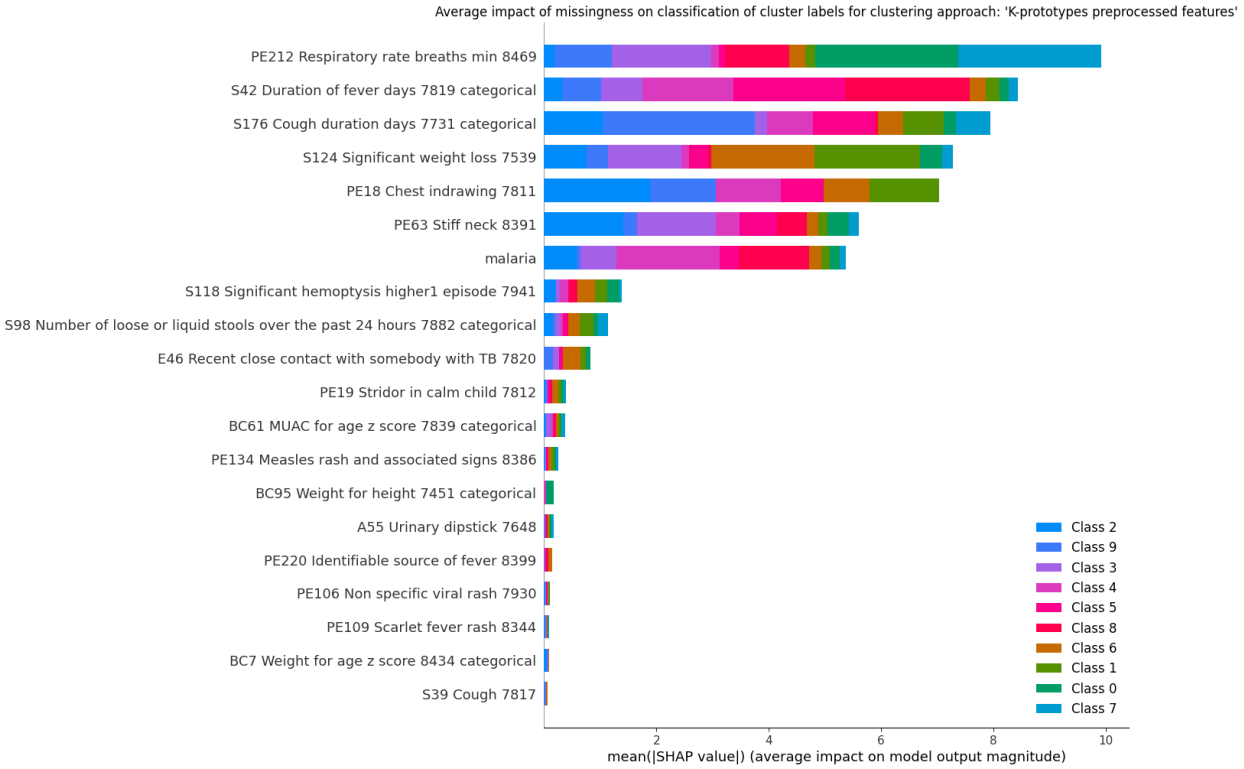

Average impact of diagnoses on classification of cluster labels for clustering approach: 'K-means preprocessed with domain knowledge'
common cold
non severe abdominal condition
non severe headache
mild moderate anemia
viral pneumonia
acute diarrhea
bacterial pneumonia
intestinal parasitic infection protozoa
bacterial acute pharyngitis
moderate malnutrition
Class 7
Class 9
Class 8
Class 5
Class 4
Class 3
Class 0
Class 6
Class 1
Class 2
0
1
2
3
4
5
6
7
mean(|SHAP value|) (average impact on model output magnitude)

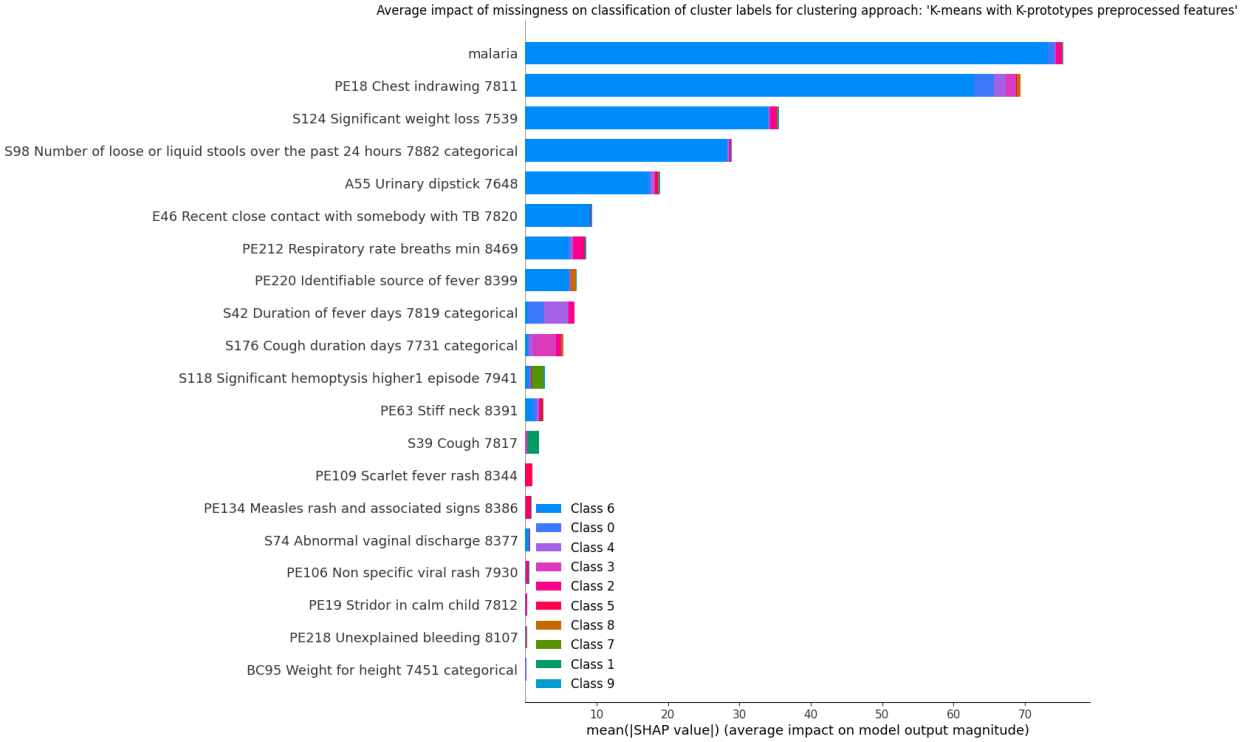

Average impact of diagnoses on classification of cluster labels for clustering approach: 'K-means with K-prototypes preprocessed features and manual centroid init'
common cold
viral pneumonia
mild moderate anemia
bacterial pneumonia
non severe headache
non severe abdominal condition
acute diarrhea
moderate malnutrition
bacterial acute pharyngitis
intestinal parasitic infection protozoa
Class 5
Class 1
Class 8
Class 0
Class 3
Class 4
Class 2
Class 7
Class 6
Class 9
0
1
2
3
4
5
6
mean(|SHAP value|) (average impact on model output magnitude)

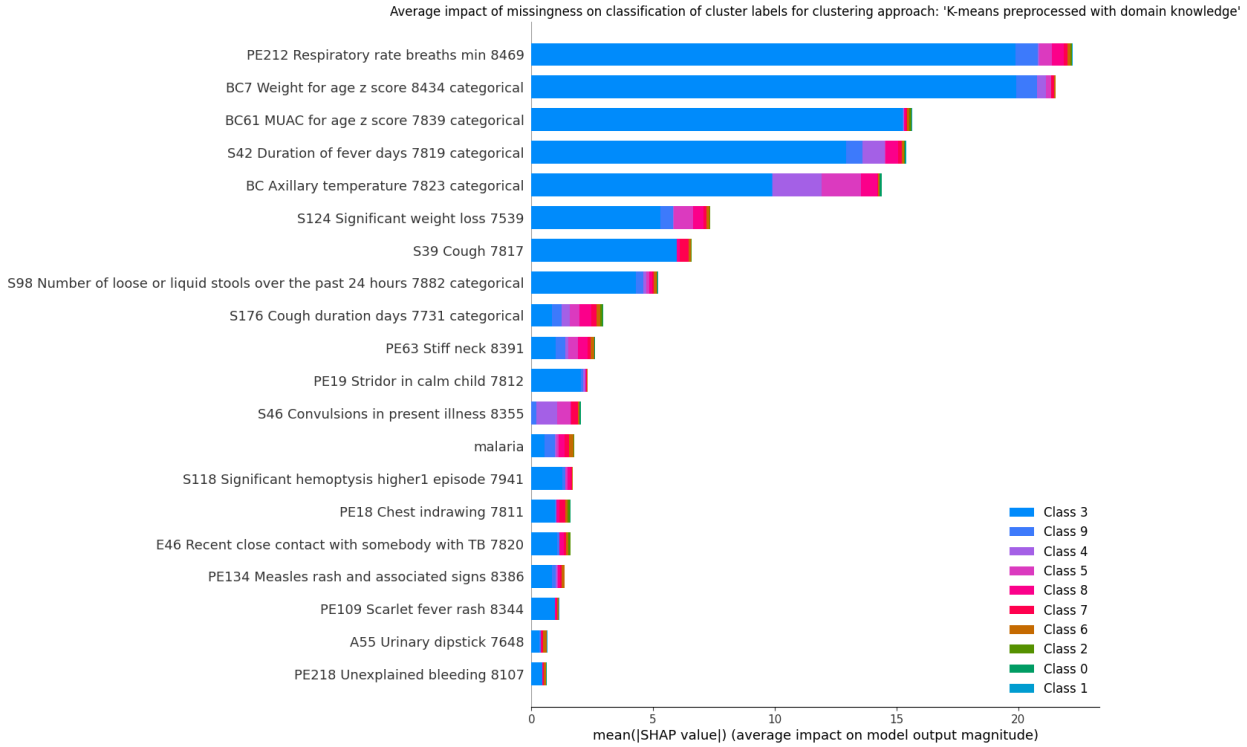


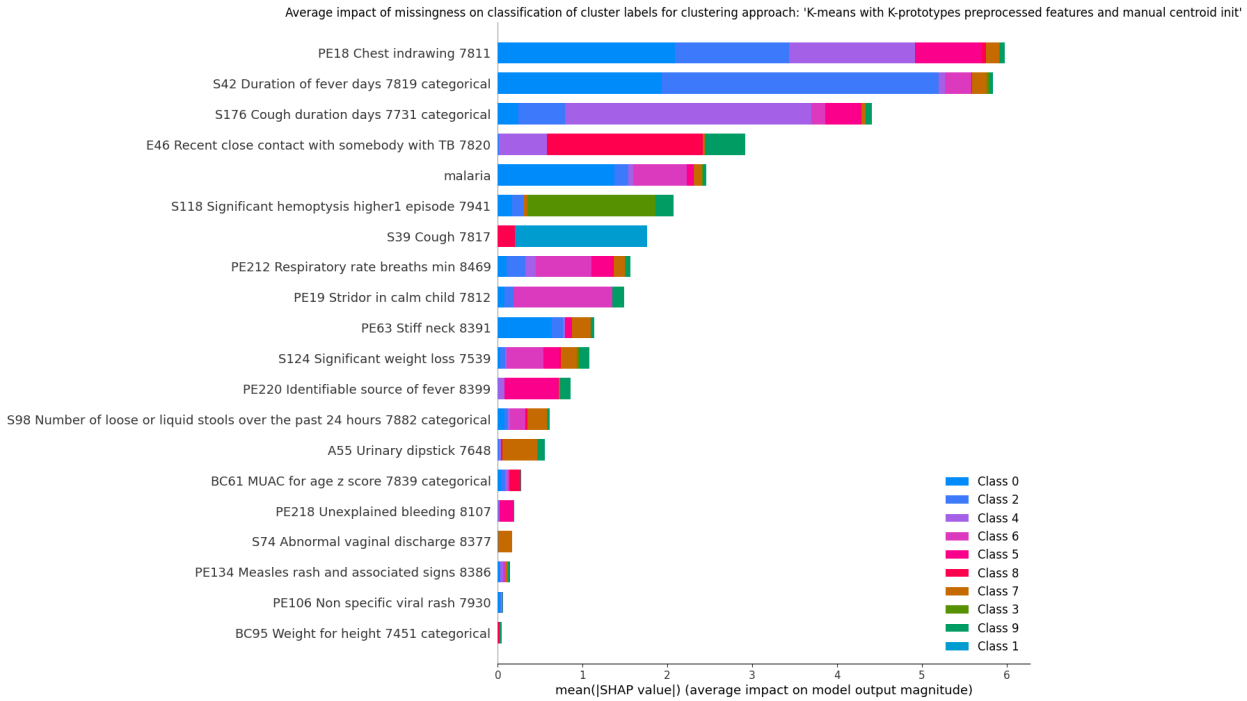

### *C.6.3 Missingness based classifier*

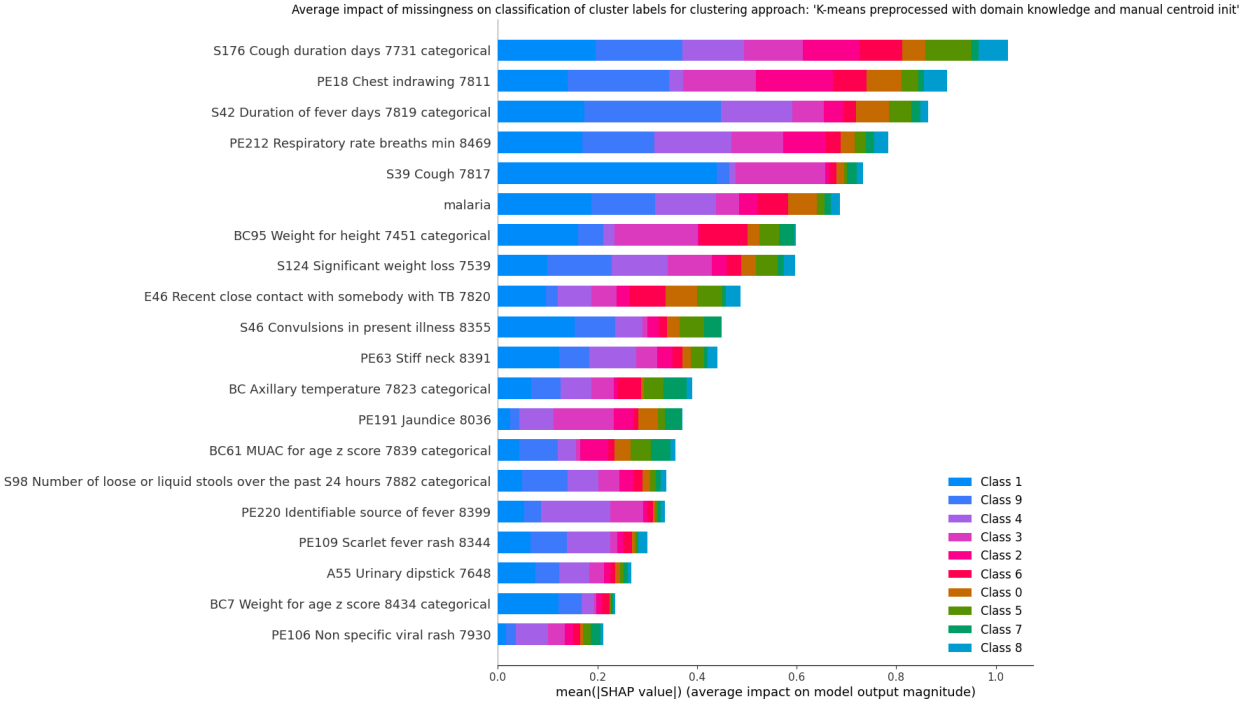



Average impact of missingness on classification of cluster labels for clustering approach: 'K-prototypes preprocessed features'
PE212 Respiratory rate breaths min 8469
S42 Duration of fever days 7819 categorical
S176 Cough duration days 7731 categorical
S124 Significant weight loss 7539
PE18 Chest indrawing 7811
PE63 Stiff neck 8391
malaria
S118 Significant hemoptysis higher1 episode 7941
S98 Number of loose or liquid stools over the past 24 hours 7882 categorical
E46 Recent close contact with somebody with TB 7820
PE19 Stridor in calm child 7812
BC61 MUAC for age z score 7839 categorical
PE134 Measles rash and associated signs 8386
BC95 Weight for height 7451 categorical
A55 Urinary dipstick 7648
PE220 Identifiable source of fever 8399
PE106 Non specific viral rash 7930
PE109 Scarlet fever rash 8344
BC7 Weight for age z score 8434 categorical
S39 Cough 7817
Class 2
Class 9
Class 3
Class 4
Class 5
Class 8
Class 6
Class 1
Class 0
Class 7
2
4
6
8
10
mean(|SHAP value|) (average impact on model output magnitude)


Average impact of missingness on classification of cluster labels for clustering approach: 'K-means with K-prototypes preprocessed features'
malaria
PE18 Chest indrawing 7811
S124 Significant weight loss 7539
S98 Number of loose or liquid stools over the past 24 hours 7882 categorical
A55 Urinary dipstick 7648
E46 Recent close contact with somebody with TB 7820
PE212 Respiratory rate breaths min 8469
PE220 Identifiable source of fever 8399
S42 Duration of fever days 7819 categorical
S176 Cough duration days 7731 categorical
S118 Significant hemoptysis higher1 episode 7941
PE63 Stiff neck 8391
S39 Cough 7817
PE109 Scarlet fever rash 8344
PE134 Measles rash and associated signs 8386
S74 Abnormal vaginal discharge 8377
PE106 Non specific viral rash 7930
PE19 Stridor in calm child 7812
PE218 Unexplained bleeding 8107
BC95 Weight for height 7451 categorical
Class 6
Class 0
Class 4
Class 3
Class 2
Class 5
Class 8
Class 7
Class 1
Class 9
10
20
30
40
50
60
70
mean(|SHAP value|) (average impact on model output magnitude)


Average impact of missingness on classification of cluster labels for clustering approach: 'K-means preprocessed with domain knowledge'
PE212 Respiratory rate breaths min 8469
BC7 Weight for age z score 8434 categorical
BC61 MUAC for age z score 7839 categorical
S42 Duration of fever days 7819 categorical
BC Axillary temperature 7823 categorical
S124 Significant weight loss 7539
S39 Cough 7817
S98 Number of loose or liquid stools over the past 24 hours 7882 categorical
S176 Cough duration days 7731 categorical
PE63 Stiff neck 8391
PE19 Stridor in calm child 7812
S46 Convulsions in present illness 8355
malaria
S118 Significant hemoptysis higher1 episode 7941
PE18 Chest indrawing 7811
E46 Recent close contact with somebody with TB 7820
PE134 Measles rash and associated signs 8386
PE109 Scarlet fever rash 8344
A55 Urinary dipstick 7648
PE218 Unexplained bleeding 8107
Class 3
Class 9
Class 4
Class 5
Class 8
Class 7
Class 6
Class 2
Class 0
Class 1
0
5
10
15
20
mean(|SHAP value|) (average impact on model output magnitude)


Average impact of missingness on classification of cluster labels for clustering approach: 'K-means with K-prototypes preprocessed features and manual centroid init'
PE18 Chest indrawing 7811
S42 Duration of fever days 7819 categorical
S176 Cough duration days 7731 categorical
E46 Recent close contact with somebody with TB 7820
malaria
S118 Significant hemoptysis higher1 episode 7941
S39 Cough 7817
PE212 Respiratory rate breaths min 8469
PE19 Stridor in calm child 7812
PE63 Stiff neck 8391
S124 Significant weight loss 7539
PE220 Identifiable source of fever 8399
S98 Number of loose or liquid stools over the past 24 hours 7882 categorical
A55 Urinary dipstick 7648
BC61 MUAC for age z score 7839 categorical
PE218 Unexplained bleeding 8107
S74 Abnormal vaginal discharge 8377
PE134 Measles rash and associated signs 8386
PE106 Non specific viral rash 7930
BC95 Weight for height 7451 categorical
Class 0
Class 2
Class 4
Class 6
Class 5
Class 8
Class 7
Class 3
Class 9
Class 1
0
1
2
3
4
5
6
mean(|SHAP value|) (average impact on model output magnitude)


Average impact of missingness on classification of cluster labels for clustering approach: 'K-means preprocessed with domain knowledge and manual centroid init'
S176 Cough duration days 7731 categorical
PE18 Chest indrawing 7811
S42 Duration of fever days 7819 categorical
PE212 Respiratory rate breaths min 8469
S39 Cough 7817
malaria
BC95 Weight for height 7451 categorical
S124 Significant weight loss 7539
E46 Recent close contact with somebody with TB 7820
S46 Convulsions in present illness 8355
PE63 Stiff neck 8391
BC Axillary temperature 7823 categorical
PE191 Jaundice 8036
BC61 MUAC for age z score 7839 categorical
S98 Number of loose or liquid stools over the past 24 hours 7882 categorical
PE220 Identifiable source of fever 8399
PE109 Scarlet fever rash 8344
A55 Urinary dipstick 7648
BC7 Weight for age z score 8434 categorical
PE106 Non specific viral rash 7930
Class 1
Class 9
Class 4
Class 3
Class 2
Class 6
Class 0
Class 5
Class 7
Class 8
0.0
0.2
0.4
0.6
0.8
1.0
mean(|SHAP value|) (average impact on model output magnitude)