COVID-19 Patient Clustering Analysis: A Multi-Algorithm Approach for Symptom-Based Patient Stratification

Anonymous authors
Paper under double-blind review

Abstract

This study explores the application of clustering algorithms to identify COVID-19 patient susceptibility groups based on symptoms, vital signs, and demographic information. Using data from two hospital datasets comprising 26,237 patients, we compare the performance of BIRCH, DBSCAN, and K-Means algorithms. Initial analysis with BIRCH on the full feature set achieved a silhouette score of 0.977, while dimension reduction techniques identified fatigue/malaise, sore throat, headache, age, and a combined oxygen-fever metric as key discriminative features. Subsequent clustering on reduced dimensions yielded varied results: DBSCAN (silhouette score: 0.301, 8 clusters), K-Means (silhouette score: 0.440 with 10 clusters), and BIRCH (silhouette score: 0.977 with 3 highly imbalanced clusters). Our findings demonstrate that while high-dimensional clustering can achieve excellent separation metrics, dimension reduction with appropriate algorithm selection provides more interpretable and balanced patient stratification for clinical applications.

1 Introduction to Proyect

The COVID-19 pandemic has placed immense pressure on hospitals, which serve as the frontline defense against the virus. Rapid and accurate identification of COVID-positive patients is crucial for managing hospital resources, ensuring patient safety, and preventing the virus's spread within healthcare facilities. Traditional methods of diagnosing COVID-19 often rely on extensive manual testing and delayed laboratory results, which can strain hospital workflows and lead to inefficient resource allocation.

Althought this project is based on COVID-19 datasets, it seeks to explore the correlation of symptoms in clustered patients to identify susceptibility symptoms based on cleared segmentations. By leveraging patient data such as symptoms, vital signs, and demographic information, I aim to proactively correlate certain group of symptoms to explain data.

1.1 Approach and rational

In order to achive this correlation, clustering methods where used as a technique to seperate data. Initial data visualization revealed a non-spherical, continuous cascade like structure with density-based patterns. Based on these characteristics an appropiate clustering method was used to attempt to organize groups so we could later breakdown characteristics and look for the most promising fields. As an intiial approach BURCH was used for its hierarchical clustering capability, scalability, and incremental learning approach. Since is particularly suitable for large datasets, it can handle non-spherical clusters with no apparent structure.

2 Dataset Discussion

2.1 Dataset Description

The analysis utilized patient data from two hospital sources (hospital1.xlsx and hospital2.xlsx) which stay anonymous for health concerns. The datasets initially showed that records had varying naming conventions and languages (ej: Turkish column names in Hospital 1). Before attempting to merge dataset into a single standarized file we need to analyze it individually

2.1.1 Hospital 1

With a total of 54 initial columns, the dataset is organized into several distinct categories to provide a complete view of patient health and disease progression. The dataset includes demographics such as patient_id, age, sex, and nationality which can be used to establish a baseline for population based analysis.

Temporal information is also present with information such as date of first symptoms and admission tracks which let´s us remove the temporal format by subtracting the dates to a clear atemporal metric. Furthermore, the dataset contains both continious and boolean based values. The dataset is partioned into:

```
Data columns (total 54 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   patient ID                      14712 non-null  int64
 1   patient ID.1                    14712 non-null  int64
 2   nationality                     14712 non-null  object
 3   age                             14712 non-null  int64
 4   gender K=female E=male          14712 non-null  object
 5   date_of_first_symptoms          14712 non-null  datetime64[ns]
 6   BASVURUTARIHI                   14712 non-null  datetime64[ns]
 7   fever_temperature               14244 non-null  float64
 8   oxygen_saturation               14708 non-null  float64
 9   history_of_fever                14712 non-null  int64
 10  cough                           14712 non-null  int64
 11  sore_throat                     14712 non-null  int64
 12  runny_nose                      14712 non-null  int64
 13  wheezing                        14712 non-null  int64
 14  shortness_of_breath             14712 non-null  int64
 15  lower_chest_wall_indrawing      14712 non-null  int64
 16  chest_pain                      14712 non-null  int64
 17  conjunctivitis                  14712 non-null  int64
 18  lymphadenopathy                 14712 non-null  int64
 19  headache                        14712 non-null  int64
 20  loss_of_smell                   14712 non-null  int64
 21  loss_of_taste                   14712 non-null  int64
 22  fatigue_malaise                 14712 non-null  int64
 23  anorexia                        14712 non-null  int64
 24  altered_consciousness_confusion 14712 non-null  int64
 25  muscle_aches                    14712 non-null  int64
 26  joint_pain                      14712 non-null  int64
 27  inability_to_walk               14712 non-null  int64
 28  abdominal_pain                  14712 non-null  int64
 29  diarrhoea                       14712 non-null  int64
 30  vomiting_nausea                 14712 non-null  int64
 31  skin_rash                       14712 non-null  int64
 32  bleeding                        14712 non-null  int64
 33  other_symptoms                  14712 non-null  int64
 34  chronic_cardiac_disease         14712 non-null  int64
 35  hypertension                    14712 non-null  int64
 36  chronic_pulmonary_disease       14712 non-null  int64
 37  asthma                       14712 non-null  int64
 38  chronic_kidney_disease           14705 non-null  float64
 39  obesity                      14690 non-null  float64
 40  liver_disease                    14706 non-null  float64
 41  asplenia                     14690 non-null  float64
 42  chronic_neurological_disorder    14710 non-null  float64
 43  malignant_neoplasm               14712 non-null  int64
 44  chronic_hematologic_disease      14710 non-null  float64
 45  AIDS_HIV                         14710 non-null  float64
 46  diabetes_mellitus_type_1         14709 non-null  float64
 47  diabetes_mellitus_type_2         14710 non-null  float64
 48  rheumatologic_disorder           14710 non-null  float64
 49  dementia                     14710 non-null  float64
 50  tuberculosis                 14712 non-null  int64
 51  smoking                      14712 non-null  int64
 52  other_risks                  14712 non-null  int64
 53  PCR_result                    13536 non-null  object
dtypes: datetime64[ns](2), float64(13), int64(36), object(3)
memory usage: 6.1+ MB
```

Which by a simple naked eye count of column statistics we can deduce that:

- 5 columns are discreate
- 2 columns are continous values
- 47 columns are boolean variables

However, data encompassing everything from respiratory distress to 19 comorbidity features which include pre-existing conditions like hypertension, diabetes, and chronic pulmonary disease.

A simple correlation matrix shows that most of values don´t have linear dependencies between each other, except for a handfull like cought, history_of_fever, sore_thorugh, runny_nose which have a big correlation value.

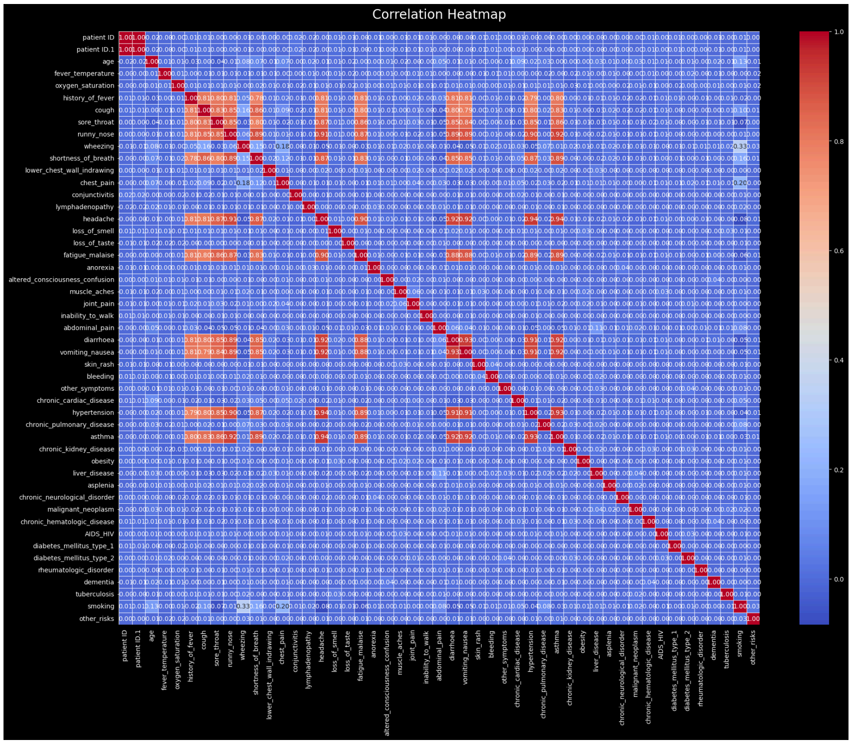

Figure 1: Hospital A Correlation Matrix.

This is expected, as this symptoms usually also group despite the disease.

2.1.2 Hospital 2

Hospital 2, has a total of 54 initial columns, the dataset is organized into several distinct categories to provide a complete view of patient health and disease progression. The dataset includes demographics such as patient_id, age, sex, and nationality which can be used to establish a baseline for population based analysis.

Temporal information is also present with information such as date of first symptoms and admission tracks which let´s us remove the temporal format by subtracting the dates to a clear atemporal metric. Furthermore, the dataset contains both continious and boolean based values. The dataset is partioned into:

```
Data columns (total 54 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   patient ID                        14712 non-null  int64
 1   patient ID.1                      14712 non-null  int64
 2   nationality                       14712 non-null  object
 3   age                               14712 non-null  int64
 4   gender K=female E=male            14712 non-null  object
 5   date_of_first_symptoms            14712 non-null  datetime64[ns]
 6   BASVURUTARIHI                     14712 non-null  datetime64[ns]
 7   fever_temperature                 14244 non-null  float64
 8   oxygen_saturation                 14708 non-null  float64
 9   history_of_fever                  14712 non-null  int64
 10  cough                             14712 non-null  int64
 11  sore_throat                       14712 non-null  int64
 12  runny_nose                        14712 non-null  int64
 13  wheezing                          14712 non-null  int64
 14  shortness_of_breath               14712 non-null  int64
 15  lower_chest_wall_indrawing        14712 non-null  int64
 16  chest_pain                        14712 non-null  int64
 17  conjunctivitis                    14712 non-null  int64
 18  lymphadenopathy                   14712 non-null  int64
 19  headache                          14712 non-null  int64
 20  loss_of_smell                     14712 non-null  int64
 21  loss_of_taste                     14712 non-null  int64
 22  fatigue_malaise                   14712 non-null  int64
 23  anorexia                          14712 non-null  int64
 24  altered_consciousness_confusion   14712 non-null  int64
 25  muscle_aches                      14712 non-null  int64
 26  joint_pain                        14712 non-null  int64
 27  inability_to_walk                 14712 non-null  int64
 28  abdominal_pain                    14712 non-null  int64
 29  diarrhoea                         14712 non-null  int64
 30  vomiting_nausea                   14712 non-null  int64
 31  skin_rash                         14712 non-null  int64
 32  bleeding                          14712 non-null  int64
 33  other_symptoms                    14712 non-null  int64
 34  chronic_cardiac_disease           14712 non-null  int64
 35  hypertension                      14712 non-null  int64
 36  chronic_pulmonary_disease         14712 non-null  int64
 37  asthma                            14712 non-null  int64
 38  chronic_kidney_disease            14705 non-null  float64
 39  obesity                           14690 non-null  float64
 40  liver_disease                     14706 non-null  float64
 41  asplenia                          14690 non-null  float64
 42  chronic_neurological_disorder     14710 non-null  float64
 43  malignant_neoplasm              14712 non-null  int64
 44  chronic_hematologic_disease      14710 non-null  float64
 45  AIDS_HIV                        14710 non-null  float64
 46  diabetes_mellitus_type_1         14709 non-null  float64
 47  diabetes_mellitus_type_2         14710 non-null  float64
 48  rheumatologic_disorder           14710 non-null  float64
 49  dementia                        14710 non-null  float64
 50  tuberculosis                    14712 non-null  int64
 51  smoking                         14712 non-null  int64
 52  other_risks                     14712 non-null  int64
 53  PCR_result                      13536 non-null  object
dtypes: datetime64[ns](2), float64(13), int64(36), object(3)
memory usage: 6.1+ MB
```

Which by a simple naked eye count of column statistics we can deduce that:

- 5 columns are discreate
- 2 columns are continous values
- 47 columns are boolean variables

However, data encompassing everything from respiratory distress to 19 comorbidity features which include pre-existing conditions like hypertension, diabetes, and chronic pulmonary disease.

A simple correlation matrix shows that most of values don´t have linear dependencies between each other, except for a handfull like cought, history_of_fever, sore_thorugh, runny_nose which have a big correlation value.

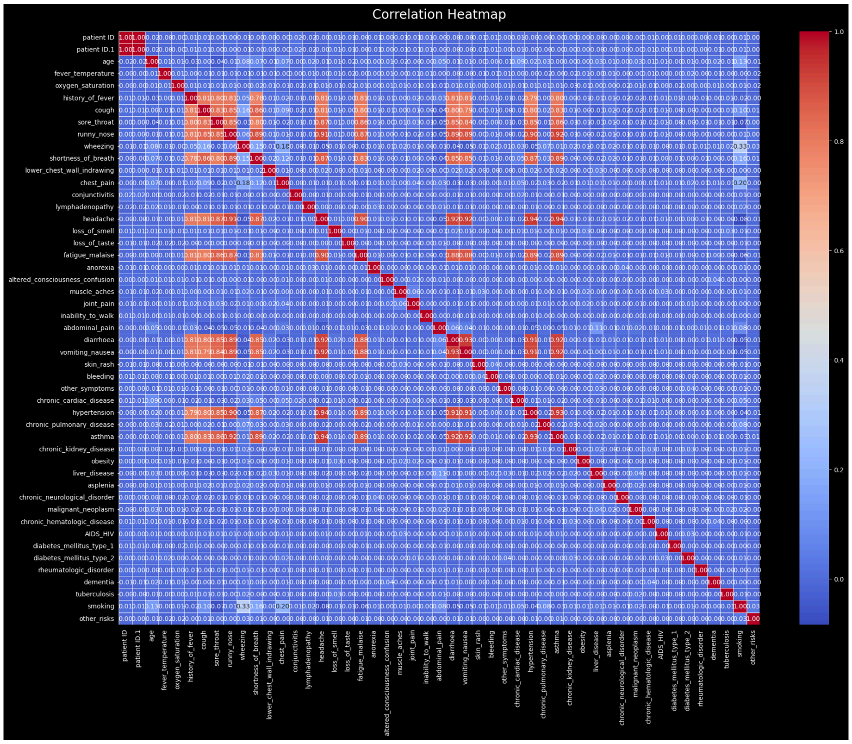

Figure 2: Hospital A Correlation Matrix.

This is expected, as this symptoms usually also group despite the disease.

2.2 Data Quality Analysis

Initial exploration of both hospital datasets revealed several data quality issues requiring attention:

Hospital 1 Issues:

- Redundant columns: patient_id and patient_id.1 contained identical information
- Column naming inconsistencies: Turkish column names (e.g., 'basvurutarihi' for admission_date, 'gender_k=female_e=male' for sex)
- Data type mismatches: fever_temperature stored as string rather than float
- Missing values: 1,176 missing PCR results
- Inconsistent value representations: Gender encoded as 'k' (kadın/female) and 'e' (erkek/male)
- NaN values scattered across symptom columns

Hospital 2 Issues:

- Extensive missing data: 1,222 missing temperature values
- Categorical encoding problems: Gender field contained a third category due to data legend inconsistencies
- Column naming inconsistencies requiring standardization
- Complete rows with NaN values
- Missing symptom information

The analysis revealed systematic data collection differences between hospitals, necessitating careful harmonization strategies.

2.3 Data Merging and Transformation

2.3.1 Column Standardization

To enable dataset integration, column names were systematically standardized:

Column mappings:

- Hospital 1: $\{\text{basvurutarihi} \rightarrow \text{admission\_date}, \text{patient\_id.1} \rightarrow \text{admission\_id}, \text{gender\_k=female\_e=male} \rightarrow \text{sex}\}$
- Hospital 2: $\{\text{country\_of\_residence} \rightarrow \text{nationality}\}$

2.3.2 Dataset Integration

The two hospital datasets were concatenated row-wise to create a unified dataset:

$D_{\text{merged}} = D_{\text{hospital1}} \cup D_{\text{hospital2}}$ where rows are concatenated with reset indices.

This merging strategy resulted in a final dataset of 26,237 patient records with harmonized feature names across all sources.

2.3.3 Feature Encoding

Gender Encoding: The sex variable was converted to binary encoding (Male=1, Female=0):

$\text{sex} = \begin{cases} 'e' \rightarrow 1 \text{ (Male)} \\ 'k' \rightarrow 0 \text{ (Female)} \end{cases}$

PCR Result Encoding: The target variable was standardized to binary format:

$\text{pcr\_result} = \begin{cases} \text{positive } \rightarrow 1 \\ \text{negative } \rightarrow 0 \end{cases}$

Nationality Standardization: Country names underwent complex normalization using ISO 3166-1 numeric codes to handle various formats and spellings:

f_{std} : text → ISO 3166-1 numeric where:

$$f_{\text{std}(c)} = \begin{cases} M[c] & \text{if } c{\in}M \\ \text{ISO}(c) & \text{if ISO lookup succeeds} \\ c & \text{otherwise} \end{cases}$$

with custom mapping $M = \{\{\text{t.c.} : 792, \text{usa} : 840, \text{cyprus} : 196, ...\}\}$

Then: $\text{nationality\_numeric} = f_{\text{std(strip(lower(nationality)))}}$

This approach handles variations in country name formatting while maintaining numerical consistency for analysis.

2.4 Data Cleaning

2.4.1 Missing Value Analysis and Imputation

Temperature Data: Trimmed mean analysis was performed to assess the impact of outliers:

$\overline{T}_{\text{trimmed}} = \text{mean}\left(T_{[np]}, ..., T_{[n(1-p)]}\right)$ where $p = 0.0168$

$\overline{T}_{\text{standard}} = \frac{1}{n} \sum_{i=1}^{n} T_i$

The difference between trimmed and standard means was negligible "$(< 0.01°C)$", indicating outliers did not significantly skew the distribution. Temperature values were imputed using the mean:

$T_i = \begin{cases} T_i & \text{if } T_i{\neq} \text{ null} \\ \overline{T} & \text{if } T_i{=} \text{ null} \end{cases}$ where $\overline{T}$ is the mean temperature.

While temperatures such as 34.8-35.5°C (hypothermia) and 39.5-40.1°C (high fever) appeared unrealistic for typical cases, they were retained as potentially clinically significant observations.

Oxygen Saturation: Special consideration was given to oxygen saturation values of -1 or 0, which indicate patient death rather than measurement errors. These values were handled separately in the analysis.

Discrete Features: Symptom and comorbidity features were imputed using mode (most frequent value):

For each discrete feature $F_j \in \left\{F_{\text{symptoms}}, F_{\text{comorbidities}}\right\}$:

$$F_{ij} = \begin{cases} F_{ij} & \text{if } F_{ij}{\neq} \text{ null} \\ \text{mode}\left(F_j\right) & \text{if } F_{ij}{=} \text{ null} \end{cases}$$

2.4.2 Data Type Conversion

Boolean features were explicitly converted to integer type for computational efficiency:

Type conversion: $F_j : \{\text{boolean}\} \to \mathbb{Z}_{\geq 0}$ for all symptom and comorbidity features.

2.4.3 Handling Missing Nationalities

Given that nationality plays a significant role in population density and disease spread patterns, records with missing nationality information were removed:

$D' = \{\boldsymbol{x}_i \in D : \text{nationality}_i \neq \text{null}\}$

2.5 Final Dataset Characteristics

The cleaned and merged dataset comprised 26,237 patient records with the following distribution:

PCR Result Distribution:
- Positive (1): 22,210 patients (84.6%)
- Negative (0): 4,027 patients (15.4%)

This substantial class imbalance (approximately 85% positive cases) reflects the dataset's focus on COVID-positive patient populations and presents considerations for clustering algorithm interpretation.

3 Initial Analysis Using BIRCH

3.1 Data Preparation for Clustering

3.1.1 Feature Selection and Scaling

Identifier columns (patient_id, admission_id, nationality) were removed as they do not contribute to clinical pattern recognition:

$D_{\text{clus}} = D \setminus \{\text{patient\_id}, \text{admission\_id}, \text{nationality}\}$

Given the presence of outliers in vital sign measurements, RobustScaler was selected over StandardScaler for feature normalization:

RobustScaler: $X'_{ij} = \frac{X_{ij} - \text{median}(X_j)}{\text{IQR}(X_j)}$

where $\text{IQR}(X_j) = Q_3(X_j) - Q_1(X_j)$ is the interquartile range.

RobustScaler uses the interquartile range (IQR) rather than mean and standard deviation, making it more resilient to extreme values in temperature and oxygen saturation data.

3.1.2 Initial Visualization

Principal Component Analysis (PCA) was employed to visualize the high-dimensional data in 2D space:

PCA projection: $\boldsymbol{X}_{\text{2D}} = \boldsymbol{X}\boldsymbol{W}_2$ where $\boldsymbol{W}_2 \in \mathbb{R}^{d \times 2}$ contains the top 2 principal components.

A visualization helper function enabled consistent cluster plotting throughout the analysis:

Scatter plot visualization: For each cluster $k \in \{0, 1, ..., K-1\}$, plot points $\{\boldsymbol{x}_i : L_i = k\}$ where $\boldsymbol{x}_i \in \mathbb{R}^2$ and L_i is the cluster label for sample i.

The initial visualization revealed continuous, non-spherical structure suggesting hierarchical organization.

3.2 BIRCH Implementation

3.2.1 Initial Model with Default Parameters

BIRCH was first applied with default hyperparameters:

BIRCH algorithm with parameters $(\tau = 0.5, B = 50, K = 3)$:

Silhouette coefficient: $s = \frac{1}{n}\sum_{i=1}^{n} \frac{b_i - a_i}{\max(a_i, b_i)}$

where a_i = average intra-cluster distance, b_i = average nearest-cluster distance.

This initial configuration achieved a silhouette score of **0.9771**, indicating excellent cluster separation.

3.2.2 Hyperparameter Tuning

To potentially improve upon this strong baseline, systematic hyperparameter optimization was conducted using randomized search:

Hyperparameter search over parameter space Θ:

$\theta^* = \arg\max_{\theta \in \Theta} s(\boldsymbol{X}, \text{BIRCH}(\boldsymbol{X}; \theta))$

where:

- $\tau \sim \text{Uniform}(0.1, 2.1)$ (threshold)
- $B \sim \text{DiscreteUniform}(20, 100)$ (branching factor)
- $K \sim \text{DiscreteUniform}(2, 10)$ (number of clusters)
- $n_{\text{iter}} = 50$ random samples from Θ

The hyperparameter search evaluated 50 different configurations across:

- **Threshold**: 0.1 to 2.1 (controls cluster radius)
- **Branching factor**: 20 to 100 (affects tree structure)
- **Number of clusters**: 2 to 10 (final cluster count)

3.2.3 Results

The optimization did not improve upon the initial silhouette score of 0.9771, though it produced different cluster assignments. The best configuration maintained three clusters with the following distribution:

Cluster 0:	503 patients (1.9%)
Cluster 1:	25,397 patients (96.8%)
Cluster 2:	337 patients (1.3%)

This highly skewed distribution, with one cluster containing 96.8% of patients, raised concerns about the practical utility of the clustering despite the excellent silhouette score. High silhouette scores can sometimes indicate that one cluster dominates the dataset rather than meaningful separation.

3.3 Post-Analysis for Dimension Reduction

To improve cluster balance and interpretability, feature selection analysis was conducted to identify the most discriminative variables.

3.3.1 Feature Categorization

Features were systematically separated into boolean and continuous types:

Feature partitioning:

$F_{\text{bool}} = \left\{ f_j : f_j \in \{0, 1\} \wedge f_j \neq \text{labels} \right\}$

$F_{\text{cont}} = F \setminus (F_{\text{bool}} \cup \{\text{labels}\})$

3.3.2 Temporal Feature Analysis

Date-encoded features were examined for variability:

Temporal analysis: $\Delta t_i = t^i_{\text{admission}} - t^i_{\text{symptoms}}$

Result: $\Delta t_i = 0 \forall i$, thus date features removed.

All differences equaled zero, indicating patients were admitted on the day of first symptom onset. Consequently, both date features were removed from the continuous feature set as they provided no discriminative power.

3.3.3 Continuous Variable Discrimination Analysis

Three metrics assessed the discriminative power of continuous features:

Between-Cluster Mean Separation: Measures how far apart cluster centers are (larger is better):

Between-cluster mean separation:

$$\sigma_{\text{between}(f_j)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\mu_{kj} - \overline{\mu}_j\right)^2}$$

where $\mu_{kj} = \text{mean}\left(\left\{f_{ij} : L_i = k\right\}\right)$ and $\overline{\mu}_j = \frac{1}{K}\sum_{k=1}^{K}\mu_{kj}$

Within-Cluster Standard Deviation: Measures cluster tightness (smaller is better):

Within-cluster standard deviation:

$$\overline{\sigma}_{\text{within}(f_j)} = \frac{1}{K}\sum_{k=1}^{K}\sigma_{kj}$$

where $\sigma_{kj} = \sqrt{\frac{1}{n_k - 1}\sum_{i:L_i=k}\left(f_{ij} - \mu_{kj}\right)^2}$

Discriminative Ratio: Combines both metrics to assess overall separation quality:

Discriminative ratio:

$$\rho\left(f_j\right) = \frac{\sigma_{\text{between}\left(f_j\right)}}{\overline{\sigma}_{\text{within}\left(f_j\right)}}$$

Results revealed:

Feature	Discriminative Ratio	Interpretation
oxygen_saturation	0.016	Almost total overlap
fever_temperature	0.045	Almost total overlap
age	0.280	Partial separation
nationality_numeric	8.930	Suspiciously high

The interpretation scale used:

- < 0.05: Centers tiny compared to spread
- 0.05–0.10: Almost total overlap
- 0.10–0.30: Partial separation
- 0.30–0.50: Clear but overlapping
- 0.50–1.00: Strong separation
- > 1.00: Very strong/suspicious

Analysis Conclusions:

Oxygen saturation and **fever temperature** showed poor discrimination (ratios < 0.05), with cluster centers barely separable relative to within-cluster variation. However, both features hold critical medical significance.

Age demonstrated moderate discriminative power (ratio: 0.28), with clusters showing partial separation by patient age.

Nationality_numeric exhibited suspiciously high separation (ratio: 8.93). This occurred because hot-encoded nationality labels lack true numerical ordering—the numeric codes are arbitrary identifiers rather than meaningful continuous values.

Given the medical importance of vital signs despite their low statistical discrimination, we decided to combine oxygen_saturation and fever_temperature into a single feature using PCA rather than discarding them entirely.

3.3.4 Boolean Feature Discrimination Analysis

Two complementary metrics evaluated boolean (symptom and comorbidity) features:

Delta P (Effect Size): Measures the maximum difference in symptom prevalence across clusters:

Effect size (Delta P):

$\Delta p(f_j) = \max_k p_{kj} - \min_k p_{kj}$

where $p_{kj} = \frac{1}{n_k} \sum_{i:L_i=k} f_{ij}$ is the prevalence in cluster k.

Cramér's V (Association Strength): Quantifies statistical association between feature and cluster assignment:

Cramér's V statistic:

$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$

where:

- $\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is the chi-squared statistic
- n = total sample size
- $k = \min(r, c)$ for contingency table with r rows and c columns

The Cramér's V interpretation scale:

- < 0.05: No discrimination
- 0.05–0.10: Weak
- 0.10–0.20: Moderate
- 0.20–0.30: Strong
- > 0.30: Very strong

Statistical analysis identified the top discriminative features as:

pcr_result, history_of_fever, fatigue_malaise, sore_throat

Refined Feature Selection:

Despite strong statistical associations, **pcr_result** and **history_of_fever** were excluded from the final feature set. PCR result represents the diagnostic outcome rather than a symptom predictor, and history of fever largely overlaps with the fever_temperature measurement.

Headache was added based on medical domain knowledge despite moderate statistical scores, as it represents a distinctive COVID-19 symptom pattern.

Final selected boolean features: $F_{\text{selected}} = \{\text{fatigue\_malaise}, \text{sore\_throat}, \text{headache}\}$

4 Second Clustering Attempt Using Reduced Dimensions

4.1 Dimension Reduction Strategy

Based on the post-analysis findings, a reduced feature set was constructed combining statistically and clinically significant variables:

Dimension reduction: $\boldsymbol{X}_{\text{reduced}} \in \mathbb{R}^{n \times 5}$ with features:

$F_{\text{reduced}} = F_{\text{selected}} \cup \{\text{age}\} \cup \{\text{oxygen\_fever}\}$

where oxygen_fever is the first principal component:

$f_{\text{oxygen\_fever}} = \boldsymbol{w}_1^T \cdot \left[(f_{\text{oxygen}}, f_{\text{temperature}})^T - \boldsymbol{\mu} \right]$

with $\boldsymbol{w}_1 = \arg\max_{\|\boldsymbol{w}\| = 1} \text{Var}(\boldsymbol{X}_{\text{vital}} \boldsymbol{w})$

Final scaling: $\boldsymbol{X}' = \text{RobustScaler}(\boldsymbol{X}_{\text{reduced}})$

The final reduced feature set comprised five dimensions:

- **fatigue_malaise** (boolean symptom)
- **sore_throat** (boolean symptom)
- **headache** (boolean symptom)
- **age** (continuous demographic)
- **oxygen_fever** (continuous vital sign composite)

4.2 Visualization of Reduced Data

Three-dimensional PCA projection enabled visualization of the reduced feature space:

3D PCA projection: $\boldsymbol{X}_{\text{3D}} = \boldsymbol{X}_{\text{reduced}}\boldsymbol{W}_3$ where $\boldsymbol{W}_3 \in \mathbb{R}^{5\times 3}$ contains the top 3 principal components.

Plot: $\{\boldsymbol{x}_i : L_i = k\}$ for each cluster k, colored by cluster assignment.

4.3 DBSCAN on Reduced Data

DBSCAN was applied to the reduced feature space with systematic hyperparameter optimization:

DBSCAN hyperparameter optimization:

$(\varepsilon^*, m^*) = \arg\max_{(\varepsilon,m)} s(\boldsymbol{X}, \text{DBSCAN}(\boldsymbol{X}; \varepsilon, m))$

subject to: $K \geq 2$ and $|N| < n$

Parameter grid:

- $\varepsilon \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ (neighborhood radius)
- $m \in \{5, 10, 15, 20\}$ (minimum points)

where $N = \{i : L_i = -1\}$ is the noise set.

4.3.1 DBSCAN Results

The optimal DBSCAN configuration achieved:

- **Best parameters:** eps=0.5, min_samples=5
- **Silhouette score:** 0.301 (moderate separation)
- **Number of clusters:** 8
- **Noise points:** 0

DBSCAN successfully identified eight distinct patient groups without classifying any samples as noise. The moderate silhouette score (0.301) indicates overlapping but distinguishable clusters, suggesting genuine structure in the reduced feature space. Unlike BIRCH on full features, DBSCAN produced more balanced cluster sizes, enhancing practical interpretability.

4.4 K-Means on Reduced Data

K-Means clustering was evaluated across multiple values of k to establish a centroid-based baseline:

```
from sklearn.cluster import KMeans

best_score = -1
best_k = None
best_labels = None
```

```
k_values = range(2, 11)

for k in k_values:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=100)
    labels = kmeans.fit_predict(X_reduced)
    score = silhouette_score(X_reduced, labels)

    if score > best_score:
        best_score = score
        best_k = k
        best_labels = labels
```

4.4.1 K-Means Results

Silhouette scores across different k values:

```
k=2:  0.377
k=3:  0.368
k=4:  0.320
k=5:  0.342
k=6:  0.385
k=7:  0.412
k=8:  0.389
k=9:  0.418
k=10: 0.440 (best)
```

The optimal K-Means configuration identified:

- **Best k:** 10 clusters
- **Silhouette score:** 0.440 (moderate-good separation)

K-Means demonstrated progressive improvement with increasing k, achieving the highest silhouette score at k=10. This result outperformed DBSCAN (0.301) on the reduced feature set, suggesting that spherical cluster assumptions reasonably approximate the data structure after dimension reduction.

4.5 BIRCH on Reduced Data

BIRCH was re-applied to the reduced feature space with hyperparameter optimization:

```
param_distributions = {
    "threshold": uniform(0.1, 1.0),
    "branching_factor": randint(20, 100),
    "n_clusters": randint(2, 9)
}

results = compute_birch_with_hyperparams(X_reduced, param_distributions)
df_birch_redu_res = pd.DataFrame(results).sort_values(
    "silhouette", ascending=False
)

df_birch_reduced = df_birch_redu_res.iloc[0]
best_birch_labels = df_birch_reduced.labels
best_silhouette = df_birch_reduced.silhouette
```

4.5.1 BIRCH Reduced Results

BIRCH on reduced features achieved:

- **Silhouette score:** 0.977 (excellent, matching full-feature performance)

- **Number of clusters:** 3
- **Cluster distribution:**
    - Cluster 0: 503 patients (1.9%)
    - Cluster 1: 25,397 patients (96.8%)
    - Cluster 2: 337 patients (1.3%)

Notably, BIRCH maintained its exceptionally high silhouette score even after dimension reduction, but the cluster distribution remained severely imbalanced. The near-identical performance on both full and reduced feature sets suggests BIRCH is primarily identifying the same dominant patient subgroup (96.8% in Cluster 1) regardless of feature space dimensionality.

This persistent imbalance, despite excellent silhouette metrics, indicates that BIRCH may not be the optimal algorithm for this dataset when seeking balanced, clinically actionable patient stratification.

5 Results and Discussion

5.1 Algorithm Performance Comparison

The clustering experiments yielded contrasting results across algorithms and feature spaces:

Table 1: Clustering algorithm performance summary

Algorithm	**Features**	**Silhouette**	**Clusters**	**Key Observation**
BIRCH	Full	0.977	3	Highly imbalanced (96.8% in one cluster)
DBSCAN	Reduced	0.301	8	Balanced clusters, no noise
K-Means	Reduced	0.440	10	Best on reduced features
BIRCH	Reduced	0.977	3	Same imbalance as full features

5.2 Feature Importance Findings

The dimension reduction analysis revealed significant insights into COVID-19 symptom discrimination:

Continuous Features:

- **Age:** Emerged as the most discriminative continuous variable (ratio: 0.28), indicating partial cluster separation by patient age groups
- **Oxygen saturation & fever temperature:** Individually showed poor discrimination (ratios < 0.05), but their combination via PCA captured essential vital sign variation
- **Nationality:** Demonstrated statistically high separation (ratio: 8.93) but was excluded due to arbitrary hot-encoding rather than true numerical relationships

Boolean Features:

- **Top discriminative symptoms:** fatigue_malaise, sore_throat, and headache
- **Excluded despite statistical significance:** pcr_result (outcome rather than predictor) and history_of_fever (redundant with temperature measurement)
- Medical domain knowledge guided final feature selection, balancing statistical and clinical considerations

5.3 Cluster Interpretation

Analysis of the initial BIRCH clustering with three clusters revealed distinct patient profiles:

Cluster 0 (503 patients, 1.9%): Younger patients (mean age: 38.5 years) with milder symptoms—lower fever (37.3°C), reduced oxygen saturation (93.9%), and lower rates of fever history (22.5%) and cough (15.9%). This group likely represents early-stage COVID or mild presentations.

Cluster 1 (25,397 patients, 96.8%): The dominant cluster with mean age 43.1 years, moderate symptoms including 51% fever history and 29.6% cough rate. This represents the standard COVID-19 patient profile.

Cluster 2 (337 patients, 1.3%): Youngest group (mean age: 34.9 years) with mixed symptom presentation—41.2% fever history, 24.3% cough, and 25.8% sore throat. May represent a distinct symptomatic subgroup.

5.4 Methodological Insights

Silhouette Score Limitations: High silhouette scores (0.977 for BIRCH) do not guarantee clinically useful clustering. The severely imbalanced distribution suggests the metric captured one dominant group's homogeneity rather than meaningful patient stratification.

Algorithm-Data Interaction: BIRCH's hierarchical structure may be overly sensitive to the dataset's inherent imbalance (85% COVID-positive). DBSCAN and K-Means, operating on density and centroid principles respectively, produced more balanced groupings on reduced features.

Feature Engineering Value: Combining weakly discriminative but medically critical features (oxygen saturation and fever temperature) into a single PCA component preserved clinical information while reducing dimensionality.

Scaling Considerations: RobustScaler proved appropriate given outliers in vital sign measurements, though additional outlier investigation could further refine the analysis.

5.5 Clinical Implications

The clustering results offer several potential applications for COVID-19 patient management:

Risk Stratification: The identification of distinct symptom profiles (particularly the mild symptom cluster) could support early triage decisions and resource allocation.

Symptom Monitoring: The key discriminative features—fatigue/malaise, sore throat, headache, age, and vital sign composites—provide a focused set of indicators for population-level surveillance.

Algorithm Selection for Healthcare: When deploying clustering in clinical settings, algorithm choice should prioritize balanced, interpretable groups (favoring K-Means or DBSCAN here) over purely statistical metrics (which favor BIRCH).

Data Collection Priorities: The poor discrimination of individual vital sign measurements suggests value in multi-parameter vital sign scoring systems rather than isolated readings.

6 Conclusion

This study demonstrates the application of multiple clustering algorithms to COVID-19 patient symptom data, revealing important insights about algorithm selection and feature engineering for healthcare analytics. While BIRCH achieved excellent silhouette scores (0.977), the resulting cluster imbalance limits clinical utility. K-Means (silhouette: 0.440, 10 clusters) and DBSCAN (silhouette: 0.301, 8 clusters) on reduced feature sets provided more balanced and interpretable patient stratifications.

The dimension reduction process identified fatigue/malaise, sore throat, headache, age, and combined oxygen-fever metrics as key discriminative features. This focused feature set enables efficient symptom-based patient monitoring while maintaining clinically relevant information.

Future work should explore:

- Supervised learning approaches using PCR results as labels
- Temporal clustering tracking symptom evolution over admission duration
- Integration of comorbidity features for risk-adjusted stratification
- Validation on additional hospital datasets to assess generalizability
- Investigation of the severely imbalanced underlying data distribution

The findings underscore that effective healthcare clustering requires balancing statistical performance with clinical interpretability, domain knowledge integration, and practical deployment considerations.