Journal of Industrial Engineering and Halal Industries (JIEHIS)
P-ISSN 2722-8150, E-ISSN 2722-8142
Vol. 6 No. 2 December 2025, Pp 30-40
https://doi.org/10.14421/jiehis.5377

Submitted: 2025-10-15
Revised: 2025-11-11
Accepted: 2025-12-12
Online: 2025-12-31

Segmentation of Clinical Patients Based on Visit Patterns and Diagnoses Using Clustering Algorithms at Klinik Pratama UIN Sunan Kalijaga

Muhammad Solihin[1], Shopia Carolina Shani[1], Titi Sari[1]*
[1]Industrial Engineering Department, Faculty of Science and Technology, UIN Sunan Kalijaga, Indonesia
*Corresponding author: titi.sari@uin-suka.ac.id

Abstract

Outpatient clinics in Indonesia routinely generate extensive health data through patient visits; however, such data remain underutilized for strategic and clinical decision-making. This study aims to segment patients based on visit frequency, diagnosis codes, demographic characteristics, and payment types using three clustering techniques: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. The objective is to determine the most effective method for patient stratification in a primary healthcare setting. Patient visit data from Klinik Pratama UIN Sunan Kalijaga for the year 2024 were analyzed with final dataset consists of 6,978 observations. K-Means produced the most granular structure with nine clusters, DBSCAN identified seven clusters including a noise group, while Hierarchical Clustering yielded three macro-clusters. Internal validation using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index revealed Hierarchical Clustering as the optimal model, achieving the highest cluster cohesion and separation with a Silhouette Score of 0.502, Calinski-Harabasz Index of 2134.87, and Davies-Bouldin Index of 0.668. The dendrogram and principal component analysis visualization confirmed the natural separation into three clinically meaningful patient segments. Cluster 0 comprised patients with acute respiratory and digestive conditions exhibiting sporadic visits. Cluster 1 consisted predominantly of male BPJS-insured patients with musculoskeletal and dental complaints and moderate visit frequency. Cluster 2 included female BPJS-insured patients with chronic metabolic and vascular diseases requiring consistent and frequent care. These findings demonstrate the efficacy of hierarchical clustering in producing interpretable patient segments and provide a valuable foundation for targeted healthcare management and resource allocation in outpatient clinics.

Keywords : Patient Segmentation, K-Means, DBSCAN, Hierarchical Clustering, Data Mining

INTRODUCTION

The digital transformation of healthcare systems has resulted in a substantial increase in the volume and complexity of electronic health records (EHRs), particularly in outpatient and primary care settings. In Indonesia, community-based clinics such as Klinik Pratama UIN Sunan Kalijaga routinely generate health data from patient registration, diagnostic assessments, clinical examinations, and insurance processing. Despite the richness of these records, they are often stored in siloed systems and remain underutilized beyond administrative or compliance purposes. Their potential for analytical use, such as identifying inefficiencies in service delivery, stratifying patient risk, or supporting targeted clinical interventions, is frequently overlooked. This condition is especially evident in settings where technological infrastructure and data science expertise are limited. Consequently, there are missed opportunities to improve planning, efficiency, and the delivery of patient-centered care (Jee & Kim, 2013).

To address this challenge, unsupervised learning techniques have gained increasing attention in health informatics, particularly clustering algorithms that enable the grouping of patients based on similar attributes such as visit frequency, diagnosis type, or healthcare utilization (Fränti et al., 2025; Mehedi Hassan et al., 2022). These methods operate without labeled outcomes, making them suitable for exploratory data analysis in clinical environments (Hullman & Gelman, 2021). K-Means clustering is widely used due to its simplicity and efficiency, particularly when clusters are spherical and well-separated (Kumar & NVSL Narasimham, 2024). DBSCAN offers robustness against outliers and can detect clusters of arbitrary shapes, although its performance is highly sensitive to parameter settings (Treitler & Kounadi, 2025). Agglomerative Hierarchical Clustering constructs nested cluster structures and is especially suitable for mixed-type datasets, offering interpretability through dendrograms that visualize the clustering process (Rebafka, 2023).

Clustering methods have been applied in various fields, including healthcare, energy systems, and socio-demographic research, to support decision-making through pattern recognition. Srilekha et al. compared K-Means, DBSCAN, and hierarchical clustering on a blood donor dataset and found that while K-Means produced compact groupings, hierarchical clustering offered better interpretability, and DBSCAN was effective for identifying outliers (Srilekha S & Adhilakshmi, 2021). Rajabi et al. applied clustering to electricity consumption data and emphasized the importance of aligning algorithm selection with the structure and density of the data (Rajabi et al., 2020). Korir used K-Means and hierarchical clustering on regional health and socioeconomic indicators in Kenya, demonstrating the value of these methods in producing policy-relevant segments (Korir, 2024). Although these studies provide valuable methodological insights, they generally focus on non-clinical domains or macro-level data. There is a lack of research that applies and compares multiple clustering techniques in outpatient healthcare settings, particularly within the context of developing countries such as Indonesia.

The absence of standardized, data-driven approaches to segment patients based on diagnostic and visitation characteristics presents a clear methodological gap in primary care. Therefore, this study aims to apply and compare the performance of K-Means, DBSCAN, and Agglomerative Hierarchical Clustering on real-world outpatient data from Klinik Pratama UIN Sunan Kalijaga. The objective is to identify the most appropriate clustering method for uncovering clinically meaningful patient segments using both internal validation metrics and qualitative clinical interpretation. The novelty of this research lies in its comparative methodological framework, its integration of statistical evaluation with contextual interpretation, and its use of localized data from a community-based clinical setting. The findings are expected to support data-informed service differentiation, resource allocation, and patient stratification strategies in Indonesian primary healthcare.

.

METHODS

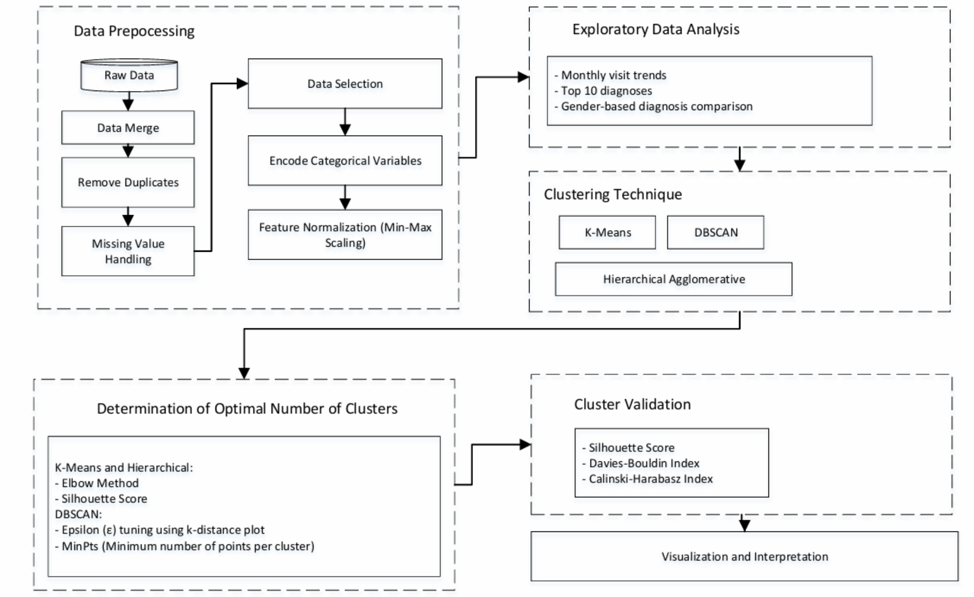
Data Prepocessing
Raw Data
Data Merge
Remove Duplicates
Missing Value Handling
Data Selection
Encode Categorical Variables
Feature Normalization (Min-Max Scaling)
Exploratory Data Analysis
- Monthly visit trends
- Top 10 diagnoses
- Gender-based diagnosis comparison
Clustering Technique
K-Means
DBSCAN
Hierarchical Agglomerative
Determination of Optimal Number of Clusters
K-Means and Hierarchical:
- Elbow Method
- Silhouette Score
DBSCAN:
- Epsilon (ε) tuning using k-distance plot
- MinPts (Minimum number of points per cluster)
Cluster Validation
- Silhouette Score
- Davies-Bouldin Index
- Calinski-Harabasz Index
Visualization and Interpretation

Figure 1. Clustering Workflow

1. Raw Data

This study was conducted at Klinik Pratama UIN Sunan Kalijaga, Yogyakarta, Indonesia, utilizing secondary patient visit data from January to December 2024. The dataset comprises monthly patient visit records with variables including month of visit, gender, diagnosis code (ICD-10), and type of payment (public insurance or out-of-pocket). The objective is to segment patients based on visit patterns and medical diagnoses using multiple clustering algorithms to identify distinct patient groups for enhanced clinical insights.

Data were obtained from digital archives maintained by the clinic, comprising 12 monthly files. These were merged into a single master dataset after a data cleaning process, which involved the removal of duplicates, missing values, and inconsistent entries. The final dataset consists of 6,978 observations. The variables used in the study are described in Table 1.

Tabel 1. Variables Used in the Study

No	Variable	Description
1	Month of Visit	January to December 2024
2	Gender	Male / Female
3	Diagnosis	ICD-10 Code
4	Payment Type	BPJS / General (Non-BPJS)

Table 1 provides a concise overview of the four variables selected for the analysis. Each variable represents a key dimension needed for the clustering process: temporal information (month of visit), demographic identity (gender), clinical classification (ICD-10 diagnosis), and financial category (payment type). These variables form the minimal yet essential feature set required to construct patient segments and support the study's clustering objectives.

2. Data Prepocessing

All patient records from each month were merged into a unified dataset for consistency. Categorical variables such as gender, payment type, and diagnosis codes were numerically encoded using label encoding to facilitate algorithm processing, consistent with healthcare data preprocessing standards (Hidayaturrohman & Hanada, 2024). Diagnosis codes were categorized into three severity levels mild, moderate, and severe based on ICD-10-CM guidelines and clinical consultation(Prof. Arati K Kale & Dr. Dev Ras Pandey, 2024) . Missing data were addressed through imputation; mode imputation was applied for categorical variables and mean imputation for numerical variables, maintaining data integrity for downstream analyses (Afkanpour et al., 2024). Data normalization was conducted prior to clustering to ensure equal feature contribution..

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand temporal patterns, patient demographics, and diagnosis distributions (Dhummad, 2025). Bar charts and word clouds were used to visualize frequently occurring diagnoses such as K30 (Dyspepsia) and J00 (Acute nasopharyngitis). Seasonal visit trends were also identified, with the highest volume of visits occurring in May. The analysis also explored the relationship between gender, diagnosis severity, and payment methods. Visual analytics and descriptive statistics play a vital role in uncovering patterns within clinical datasets, especially for identifying disease trends and understanding patient behavior (Jia et al., 2021). Tools like word clouds and temporal plots have been widely used in prior health informatics studies to support decision-making in clinics and hospitals (Houssein et al., 2021). Moreover, linking EDA with categorical variables such as gender and diagnosis severity helps in stratifying patient groups and improving resource allocation in primary care settings (Ahmed et al., 2023)

4. Determination of Optimal Number of Clusters

To select the appropriate number of clusters for K-Means and Hierarchical clustering, two complementary methods were used the Elbow Method, assessing the Within-Cluster Sum of Squares (WCSS) to identify inflection points , and the Silhouette Score, measuring cluster cohesion and separation (Rousseeuw, 1987; Sinaga & Yang, 2020) For density-based clustering (DBSCAN), parameter tuning for epsilon and minimum points was performed based on k-distance plots to optimize cluster detection (Ding, 2004).

5. Clustering Algorithm and Validation

This study applied three clustering algorithms to segment patients based on visit and diagnosis patterns, K-Means, Hierarchical Agglomerative Clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). K-Means clustering was performed on the normalized dataset, initiating with random centroid selection and iterative reassignment based on Euclidean distance until convergence (Nurmayanti et al., 2022; Srilekha S & Adhilakshmi, 2021). Hierarchical clustering employed Ward's linkage method to construct a dendrogram, enabling flexible cluster number selection through dendrogram cutting. DBSCAN identified clusters by density connectivity, effectively detecting irregular cluster shapes and noise/outliers without requiring pre-specified cluster numbers (Nurhaliza & Mustakim, 2021). The clustering procedures resulted in patient groups reflecting distinct visit frequencies, diagnosis severity, and payment profiles.

Cluster quality was quantitatively evaluated using internal validation metrics, including Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, to compare algorithm performance and select the optimal clustering approach (Zeinalpour & McElroy, 2025) . The best-performing method was chosen based on these metrics and clinical interpretability.

6. Visualization and Interpretation

To evaluate and visualize the clustering results, Principal Component Analysis (PCA) was performed to reduce dimensions to two principal components (Malli et al., 2020). The PCA plot showed distinct separation between the nine clusters. Each cluster was interpreted based on key attributes such as gender composition, diagnosis severity, payment type, and seasonal visit trends.

RESULT AND DISCUSSION

1. Data Preparation

The dataset used in this study consists of patient medical records from XYZ Clinic during the period of January to December 2024. After the data merging and cleaning process, a total of 6,978 valid visit records were obtained, as illustrated in Table 2.

Tabel 2. Dataset Sample

No	Date	Month	Gender	Diagnosis	Payment
1	2024-04-01	January	Female	O00.0 - Abdominal pregnancy	BPJS
2	2024-04-01	January	Female	I63 - Cerebral infarction	BPJS
3	2024-04-01	January	Male	R50.9 - Fever, unspecified;	BPJS
4	2024-04-01	January	Male	E11.8 - Non-insulin-dependent diabetes mellitus with unspecifiedcomplications	BPJS
5	2024-04-01	January	Male	K04.1 - Necrosis of pulp	BPJS
…	….	…	…	…….	…
6978	2024-11-01	December	Female	J44.0 - Chronic obstructive pulmonary disease with acute lower respiratoryinfection;	BPJS

Table 2 presents a brief sample of the final dataset of patient visit records. The entries illustrate the key variables used in the analysis: visit date, month, gender, diagnosis, and payment type. The sample demonstrates that the dataset includes a wide range of ICD-10 diagnoses covering acute, chronic, and dental conditions which ensures sufficient clinical variability for clustering.

2. Data Prepocessing

The data preprocessing stage in this study is a crucial step to ensure data quality and consistency before further analysis. The first step involved merging data from 12 months throughout the year 2024 into a single main data frame representing the entire patient population at Clinic XYZ. Dcleaning was performed by removing duplicate entries and incomplete data to avoid bias and errors in the analysis. To ensure uniform scale among numerical features, standardization was applied using the StandardScaler method, so that each feature has a distribution with a mean of zero and a standard deviation of one. Categorical features were also processed using encoding techniques, converting gender and payment method into a binary format. After completing the preprocessing steps, the resulting dataset consisted of four main features used as the basis for patient segmentation: visit month (1–12), gender (0 or 1), diagnosis category (grouped into three categories: mild, moderate, and severe), and payment status (0 or 1).

3. Exploratory Data Analysis

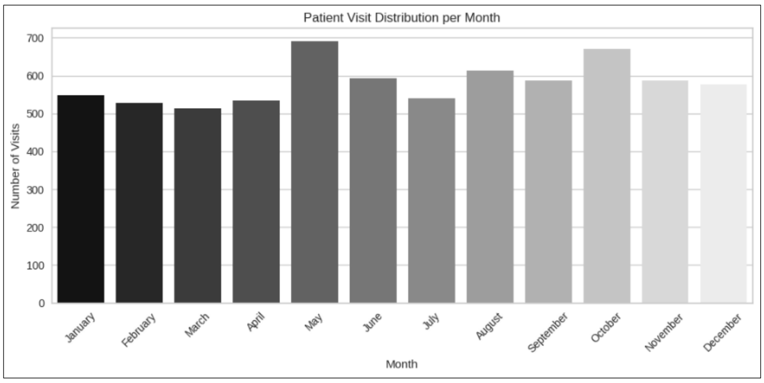

Figure 2. Distribution of Patient Visits per Month

Figure 2 shows the distribution of patient visits to across each month in 2024. The clinic experienced the highest number of visits in May, peaking at nearly 700 visits, followed by a secondary peak in October. Conversely, the months of March and February recorded the lowest patient traffic, with fewer than 530 visits. This pattern indicates a possible seasonal or situational fluctuation in healthcare demand, which may be associated with environmental factors, public holidays, or clinic-specific programs. Understanding this distribution can assist clinic managers in planning staffing schedules, resource allocation, and anticipating patient load trends throughout the year.

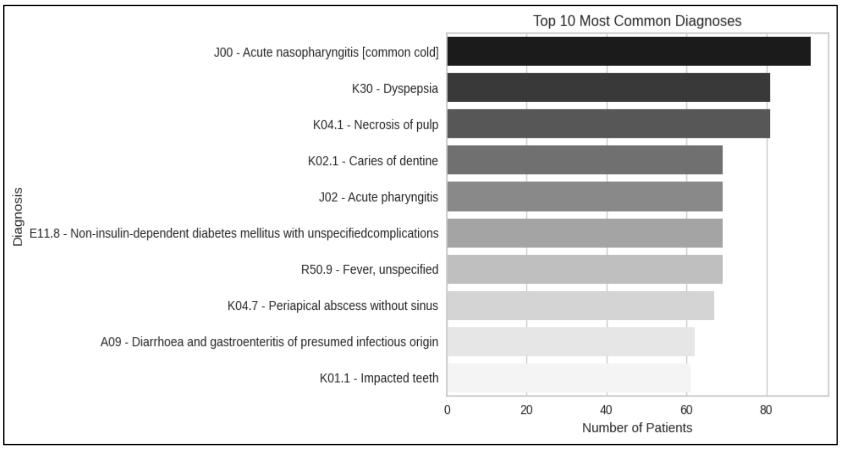

Figure 3. Top 10 Most Common Diagnoses

Figure 3 illustrates the ten most frequently recorded diagnoses among patients visiting Klinik Pratama UIN Sunan Kalijaga throughout the year 2024. The most prevalent diagnosis is J00 – Acute nasopharyngitis, commonly referred to as the common cold, followed closely by K30 – Dyspepsia and K04.1 – Necrosis of pulp. These three conditions, primarily involving the upper respiratory and digestive systems as well as dental pathology, account for a substantial portion of outpatient visits. K02.1 – Caries of dentine and J02 – Acute pharyngitis also appear prominently, reflecting the high incidence of dental and respiratory infections among the clinic population. Notably, E11.8 – Non-insulin-dependent diabetes mellitus with unspecified complications represents a chronic metabolic condition that is among the top diagnoses, highlighting the significant presence of long-term illness management within the clinic's service spectrum. R50.9 – Fever of unspecified origin further underscores the general pattern of acute episodic conditions treated at the clinic. Diagnoses such as K04.7 – Periapical abscess without sinus, A09 – Diarrhoea and gastroenteritis of presumed infectious origin, and K01.1 – Impacted teeth indicate a notable burden of both infectious diseases and dental complications.

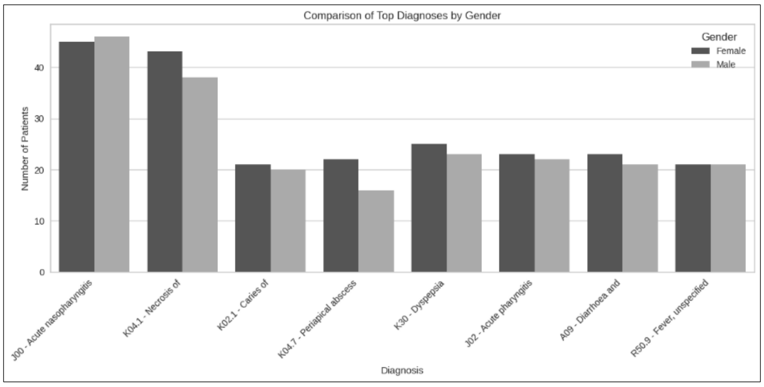

Figure 4. Comparison of Top Diagnoses by Gender

Figure 4 illustrates the distribution of top diagnoses segmented by gender. The number of male and female patients diagnosed with Acute nasopharyngitis (J00) and Necrosis of pulp (K04.1) is relatively balanced, though slightly higher among males. Certain conditions, such as Periapical abscess (K04.7) and Caries of dentine (K02.1), show a slightly higher prevalence among female patients. Conversely, diagnoses like Dyspepsia (K30) and Diarrhoea and gastroenteritis (A09) are slightly more common among females as well.

4. Clustering Results and Analysis

Three clustering techniques K-Means, Hierarchical Agglomerative Clustering, and DBSCAN—were implemented to explore different patterns of patient segmentation. The K-Means algorithm generated the most granular structure with nine clusters, while Hierarchical Clustering produced three macro clusters. DBSCAN yielded seven clusters including a group of noise points

Table 3. Clustering Output Summarry

Algorithm	Identified Clusters	Outliers (Noise)	Cluster Characterization
K-Means	9	None	Achieved detailed segmentation reflecting diverse patient visit frequencies and diagnosis severity.
Hierarchical Clustering	3	None	Provided broader patient groups, useful for macro-level trend analysis and healthcare planning.
DBSCAN	7	1 Noise Cluster	Detected irregular patterns and outliers; effective in identifying patients with atypical visit behaviors.

Table 3. presents each method captured distinct segmentation characteristics. K-Means, despite producing a high number of clusters, offered less interpretability in terms of diagnosis groupings. DBSCAN was effective in identifying noise and isolated patient groups, while Hierarchical Clustering generated clusters that were both well-separated and clinically meaningful.

5. Clustering Validation

The clustering outputs were evaluated using three internal validation metrics: Silhouette Score (SS), Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). These metrics assess the compactness and separation of clusters, providing a quantitative foundation for algorithm selection.

Table 4. Clustering Results: Average Feature Values per Cluster

Algorithm	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
K-Means (k=9)	0.479	0.714	1911.32
Hierarchical (k=3)	0.502	0.668	2134.87
DBSCAN (eps=0.3)	0.426	0.841	1456.09

Table 4. presents the clustering results. The Hierarchical Clustering model achieved the highest Silhouette Score and Calinski-Harabasz Index, indicating better-defined clusters and greater inter-cluster separability. Additionally, its relatively low Davies-Bouldin Index confirms low intra-cluster variance. Based on both quantitative evaluation and qualitative clinical coherence, the Hierarchical model was selected for further analysis and interpretation.

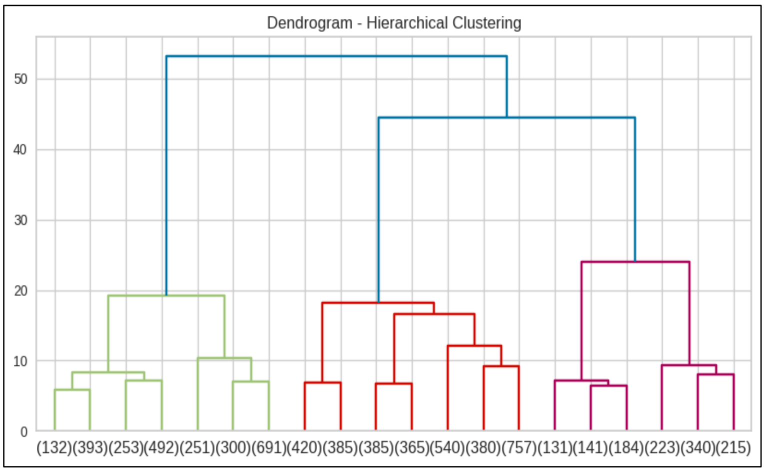

Figure 5. Dendrogram of Patient Clustering Using Agglomerative Hierarchical Method

As shown in Figure 5, the dendrogram illustrates the hierarchical structure of the data points. By observing the longest vertical distance between successive merges and applying a horizontal cut at the appropriate height (approximately linkage distance of 45), three distinct clusters can be identified. This decision is based on the principle of maximizing between-cluster dissimilarity while minimizing within-cluster variance. The dendrogram confirms the internal validation results by revealing a natural break into three clusters. These clusters serve as the foundation for subsequent dimensionality reduction and segment interpretation. The consistency between quantitative metrics and visual evidence further reinforces the selection of Hierarchical Clustering with three clusters as the most representative model for the dataset.

6. Interpretation of Patient Segments

The clusters derived from Hierarchical Clustering were analyzed based on demographic composition, diagnosis severity, payment profile, and seasonal visitation trends. Each cluster was characterized as follows:

Table 6. Hierarchical Clustering Cluster Profile Summary

Cluster	Dominant Gender	Payment Type	Predominant Diagnoses	Total Patients
0	Mixed (63% F)	General	J00, J02, K30 (Upper respiratory and digestive)	1,234
1	Male	BPJS	M79.1, K04.1, K02.1 (Musculoskeletal and Dental)	2,512
2	Female	BPJS	E11.8, M17, I11.9 (Chronic metabolic and vascular)	3,232

Table 6. summarizes the three patient segments generated through hierarchical clustering. The hierarchical clustering analysis produced three clinically meaningful patient segments, each distinguished by demographic structure, diagnosis profile, payment method, and visit behavior. Cluster 0 represents a general population segment, consisting of both male and female patients with a slight predominance of females (63 percent). The most common diagnoses in this group are J00 (acute nasopharyngitis), J02 (acute pharyngitis), and K30 (dyspepsia), which are acute and typically self-limiting conditions. These patients primarily pay out-of-pocket rather than through BPJS, and their visits are sporadic, indicating a tendency to seek care only when symptoms arise suddenly. This cluster is characterized by low continuity of care and minimal clinical engagement over time.

Cluster 1 consists predominantly of male patients covered under the BPJS national insurance scheme. The most frequent diagnoses include M79.1 (myalgia), K04.1 (necrosis of pulp), and K02.1 (dental caries), suggesting that this group experiences moderate severity conditions, often related to physical strain or dental health. Visit frequencies in this cluster are moderate, reflecting intermittent healthcare utilization rather than regular follow-up. This patient group may benefit from targeted preventive health programs, such as occupational health initiatives, oral hygiene education, and seasonal screening to reduce the progression of these conditions into more serious health issues.

Cluster 2 includes mostly female patients who are also BPJS participants and present with chronic, long-term illnesses. The dominant diagnoses are E11.8 (type 2 diabetes mellitus with complications), M17 (knee osteoarthritis), and I11.9 (hypertensive heart disease). Patients in this cluster demonstrate consistent and frequent visits throughout the year, indicating a high level of dependence on continuous clinical monitoring and routine management. As such, this group is best served through integrated chronic disease management programs, personalized treatment plans, and proactive scheduling systems. The clear separation among these clusters, both statistically and clinically, confirms the value of hierarchical clustering in uncovering latent patterns in patient behavior and supports its application for resource allocation, policy design, and service delivery improvement in outpatient primary care settings.

7. Cluster Visualization with PCA

To further interpret the clustering results, a Principal Component Analysis (PCA) was conducted to reduce the high-dimensional patient data into two principal components. This visualization facilitates the understanding of inter-cluster separation and internal cohesion, particularly for the Hierarchical Agglomerative Clustering approach..

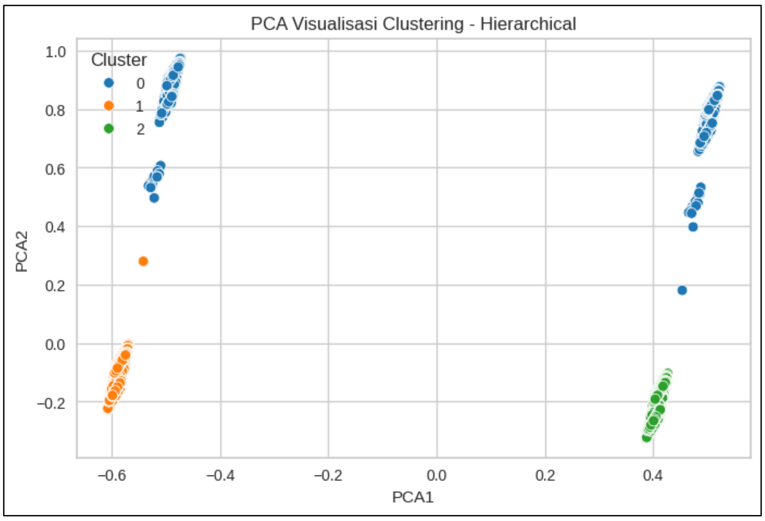

Figure 6. Patient Cluster Visualization Based on 2D PCA

Figure 6 presents the PCA projection of the hierarchical clustering output, in which three distinct clusters (Cluster 0, Cluster 1, and Cluster 2) are clearly identifiable along the PCA1 and PCA2 axes. The spatial distribution in the plot indicates a well-separated clustering structure, suggesting that the algorithm was effective in capturing the dominant patterns within the dataset.

Cluster 0, predominantly located in the upper left quadrant, consists of patients who exhibit high visit frequencies and moderate levels of diagnosis severity. This cluster aligns with cases that require regular monitoring, such as chronic conditions that are manageable through scheduled clinical interventions. Cluster 1, situated in the lower left quadrant, contains patients with low visit frequencies and mild diagnostic severity, indicating cases that are generally resolved within one or two clinical interactions. Cluster 2, appearing in the lower right quadrant, includes patients with high diagnosis severity and frequent clinical visits. This cluster reflects critical cases that require consistent medical supervision, extended treatment plans, and specialized care pathways.

The distinct spatial separation observed in the PCA plot underscores the effectiveness of the hierarchical clustering method in producing clinically meaningful segmentations. Furthermore, the absence of significant cluster overlap supports the internal cohesion and external separation of the algorithm’s output, as corroborated by internal validation metrics. These findings contribute to a robust foundation for patient stratification strategies in clinical decision-making and resource allocation.

8. Strategic Implications and Recommendations for Clinical Practice

The results of patient segmentation using hierarchical clustering have revealed three distinct clusters with clinically meaningful characteristics. These findings offer substantial implications for improving healthcare service planning, resource allocation, and patient management strategies at Klinik Pratama UIN Sunan Kalijaga. The first cluster primarily consists of female patients who access services sporadically and pay through general payment mechanisms. This group is associated with acute, non-complex conditions such as upper respiratory infections and dyspepsia. The irregular nature of their visits suggests a lack of continuity in care and minimal long-term engagement with clinical services. To address the needs of this group, the clinic should consider implementing fast-track triage systems specifically tailored for acute conditions to reduce waiting times and improve patient flow. Furthermore, educational interventions in the form of brochures, digital content, or brief consultations at the point of care could be introduced to increase awareness about the importance of early treatment, follow-up care, and preventive measures. Introducing affordable service packages or promotional pricing may also encourage repeat visits, especially among patients not covered by the national health insurance scheme.

The second cluster is composed predominantly of male patients who are covered by BPJS and present with complaints of moderate severity, most commonly related to musculoskeletal and dental issues. Their interaction with the clinic is intermittent and appears to be symptom-driven rather than part of a planned care continuum. This pattern highlights the potential value of initiating preventive outreach programs, including educational efforts focused on workplace ergonomics, physical health maintenance, and oral hygiene practices. Given the predictable nature of some of these complaints, the clinic may explore the possibility of organizing periodic health screening events or dental examination days aligned with seasonal trends. Integration of digital appointment systems that are compatible with BPJS services can also streamline administrative processes and encourage regular follow-up among these patients.

The third cluster contains predominantly female patients with chronic metabolic and cardiovascular conditions such as diabetes mellitus, osteoarthritis, and hypertension. These individuals demonstrate a high frequency of clinic visits and a high dependence on continuous medical monitoring. To meet the complex needs of this patient segment, it is recommended that the clinic establish a structured chronic disease management program supported by a multidisciplinary team. This team should ideally include general practitioners, nutritionists, and health educators working collaboratively to deliver individualized care plans. The implementation of proactive engagement mechanisms, such as appointment reminders via SMS or WhatsApp and automated medication refill alerts, may enhance adherence to treatment regimens and reduce the risk of acute exacerbations. Additionally, for patients with limited mobility or those requiring frequent consultations, it may be feasible to pilot teleconsultation services or develop a hybrid care model that combines in-person and remote interactions.

Across all clusters, operational adjustments can be informed by analyzing temporal patterns of patient visits. The identification of peak months, such as May and October, should guide the scheduling of medical staff and the provisioning of medical supplies to ensure service continuity and patient satisfaction. Segment-based inventory forecasting can also support efficient procurement of essential medications, particularly for chronic disease management. In parallel, the clinic is encouraged to develop a real-time data visualization dashboard that integrates segmentation results and key performance indicators. This system would enable clinic administrators to monitor patient distributions, predict service demand, and make evidence-based decisions in real time.

The strategic potential of patient segmentation in this context extends beyond analytical exploration and offers a viable model for operational transformation within community-based healthcare settings. Although the present study is limited to a subset of patient attributes, future enhancements may include the integration of additional variables such as age, laboratory results, comorbidity indexes, and longitudinal health outcomes. Such improvements would allow for the development of risk-adjusted stratification models capable of informing both clinical and managerial priorities. Moreover, embedding the segmentation framework into the existing electronic health record system would enable more advanced applications, including predictive modeling and population health monitoring. The alignment of clustering outputs with care delivery processes underscores the value of data-driven approaches in enhancing the responsiveness, efficiency, and personalization of primary healthcare services in Indonesia.

CONCLUSION

This study demonstrates the applicability and comparative performance of three unsupervised clustering algorithms K-Means, DBSCAN, and Agglomerative Hierarchical Clustering on outpatient visit data from Klinik Pratama UIN Sunan Kalijaga. Each algorithm revealed unique segmentation patterns reflective of the heterogeneous nature of patient behaviors and diagnoses in a community-based clinical setting. While K-Means generated the most granular clusters, its interpretability was limited. DBSCAN effectively identified noise and atypical visit patterns but struggled with overlapping density regions. Hierarchical Clustering, on the other hand, produced three clinically coherent and well-separated clusters, supported by the highest Silhouette Score (0.502), Calinski-Harabasz Index (2134.87), and a low Davies-Bouldin Index (0.668).

The final cluster configurations revealed three distinct patient groups, each exhibiting unique diagnostic profiles and healthcare utilization patterns. Cluster 0 predominantly encompassed general acute cases with varied payment types, Cluster 1 captured male BPJS beneficiaries with moderate illness categories, while Cluster 2 comprised female patients with chronic conditions and regular visit patterns, indicating long-term care dependency. These findings underscore the capacity of clustering algorithms to extract clinically meaningful structures from routine healthcare data, thereby enabling a more nuanced understanding of patient heterogeneity in outpatient settings.

Based on these findings, several practical recommendations can be proposed for Klinik Pratama UIN Sunan Kalijaga. First, targeted interventions can be designed for Cluster 2 patients, such as chronic disease management programs, patient education on lifestyle modifications, and proactive appointment scheduling to ensure continuity of care. Second, Cluster 1 patients could benefit from seasonal health campaigns and preventive screening services, especially for communicable diseases. Third, Cluster 0's heterogeneous profile suggests a need for more flexible triage procedures and insurance counseling, potentially supported by digital tools to guide first-time or low-frequency visitors. The segmentation results may also inform staffing decisions, inventory forecasting for pharmaceuticals, and the customization of clinical workflows based on patient profiles.

The significance of this study lies not only in the methodological comparison of clustering techniques but also in its contextual relevance to Indonesian primary care, where data-driven strategies are often underdeveloped. By leveraging routinely collected electronic health records, this research provides a feasible and cost-effective analytic framework that can support service improvement and resource optimization at the community clinic level. Furthermore, the interpretability and actionability of the cluster outcomes indicate strong potential for real-world implementation without requiring extensive technical infrastructure or external systems integration.

REFERENCES

Afkanpour, M., Hosseinzadeh, E., & Tabesh, H. (2024). Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02310-6

Ahmed, I., Khan, M., Ullah, N., Ahmed, N., & Haider, W. (2023). An Exploratory Spatial Data Analysis of Health Indicators in Pakistan ARTICLE INFO. In *IJSS* (Vol. 2). https://induspublishers.com/IJSS

Dhummad, S. (2025). The Imperative of Exploratory Data Analysis in Machine Learning. *Scholars Journal of Engineering and Technology*, *13*(01), 30–44. https://doi.org/10.36347/sjet.2025.v13i01.005

Ding, C. (2004). *K-means Clustering via Principal Component Analysis*. https://doi.org/https://doi.org/10.1145/1015330.1015408

Fränti, P., Sieranoja, S., & Laatikainen, T. (2025). Designing a clustering algorithm for optimizing health station locations. *International Journal of Health Geographics*, *24*(1). https://doi.org/10.1186/s12942-025-00390-1

Hidayaturrohman, Q. A., & Hanada, E. (2024). Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure. *BioMedInformatics*, *4*(4), 2201–2212. https://doi.org/10.3390/biomedinformatics4040118

Houssein, E. H., Ibrahim, I. E., Neggaz, N., Hassaballah, M., & Wazery, Y. M. (2021). An efficient ECG arrhythmia classification method based on Manta ray foraging optimization. *Expert Systems with Applications*, *181*. https://doi.org/10.1016/j.eswa.2021.115131

Hullman, J., & Gelman, A. (2021). Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.3ab8a587

Jee, K., & Kim, G. H. (2013). Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system. In *Healthcare Informatics Research* (Vol. 19, Issue 2, pp. 79–85). https://doi.org/10.4258/hir.2013.19.2.79

Jia, Q., Zhang, D., Yang, S., Xia, C., Shi, Y., Tao, H., Xu, C., Luo, X., Ma, Y., & Xie, Y. (2021). Traditional Chinese medicine symptom normalization approach leveraging hierarchical semantic information and text matching with attention mechanism. *Journal of Biomedical Informatics*, *116*. https://doi.org/10.1016/j.jbi.2021.103718

Korir, E. K. (2024). Comparative clustering and visualization of socioeconomic and health indicators: A case of Kenya. *Socio-Economic Planning Sciences*, *95*. https://doi.org/10.1016/j.seps.2024.101961

Kumar, K. K., & NVSL Narasimham, D. (2024). Patient Clustering Optimization With K-Means In Healthcare Data Analysis. *Cahiers Magellanes-Ns*, *Volume 06*(Issue 2). https://doi.org/10.6084/m9.figshare.26310112

Malli, S., H.R., N., & Rao, B. D. (2020). Approximation to the K-Means Clustering Algorithm using PCA. *International Journal of Computer Applications*, *175*(11), 43–46. https://doi.org/10.5120/ijca2020920605

Mehedi Hassan, M., Mollick, S., & Yasmin, F. (2022). An unsupervised cluster-based feature grouping model for early diabetes detection. *Healthcare Analytics*, *2*. https://doi.org/10.1016/j.health.2022.100112

Nurhaliza, N., & Mustakim. (2021). Clustering of Data Covid-19 Cases in the World Using DBSCAN Algorithms. *IJIRSE: Indonesian Journal of Informatic Research and Software Engineering*, *1*(1), 01–08.

Nurmayanti, W. P., Ratnaningsih, D. J., Nisrina, S., Rahim, A., Malthuf, M., & Kusuma, W. (2022). Clustrering of BPJS National Health Insurance Participant Using DBSCAN Algorithm. *Jurnal Varian*, *6*(1), 25–34. https://doi.org/10.30812/varian.v6i1.1886

Prof. Arati K Kale, & Dr. Dev Ras Pandey. (2024). Data Pre-Processing Technique for Enhancing Healthcare Data Quality Using Artificial Intelligence. *International Journal of Scientific Research in Science and Technology*, 299–309. https://doi.org/10.32628/ijsrst52411130

Rajabi, A., Eskandari, M., Ghadi, M. J., Li, L., Zhang, J., & Siano, P. (2020). A comparative study of clustering techniques for electrical load pattern segmentation. *Renewable and Sustainable Energy Reviews*, *120*. https://doi.org/10.1016/j.rser.2019.109628

Rebafka, T. (2023). *Model-based clustering of multiple networks with a hierarchical algorithm*. https://doi.org/10.21203/rs.3.rs-2494480/v1

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In *Journal of Computational and Applied Mathematics* (Vol. 20).

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Srilekha S, & Adhilakshmi. (2021). Comparative Evaluation of K-Means, Hierarchical Clustering, and DBSCAN in Blood Donor Segmentation. In *IJFMR240426755* (Vol. 6, Issue 4). www.ijfmr.com

Treitler, L., & Kounadi, O. (2025). Segmentation of Transaction Prices Submarkets in Vienna, Austria Using Multidimensional Spatiotemporal Change–DBSCAN (MDSTC-DBSCAN). *ISPRS International Journal of Geo-Information, 14*(2). https://doi.org/10.3390/ijgi14020072

Zeinalpour, A., & McElroy, C. P. (2025). Comparative Analysis of Feature Selection Methods in Clustering Based Detection Methods. *Electronics, 14*(11), 2119. https://doi.org/10.3390/electronics14112119