Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

Hybrid DBSCAN - K-Means Clustering for Financial Loss Identification in INA-CBG Claims Based on Medical Treatment Patterns

Muhammad Fajar Dianqori[1,*], Dhomas Hatta Fudholi[1], Galih Aryo Utomo[2], Irving Vitra Paputungan[1,3]

[1] Departement of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
[2] Department of Informatics and Medical Record RS Islam Yogyakarta PDHI, Yogyakarta, Indonesia
[3] Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Malaysia
Email: [1,*] muhammad.dianqori@students.uii.ac.id, [2]hatta.fudholi@uii.ac.id, [3]galih.a.utomo@gmail.com, [4]irving@uii.ac.id
Correspondence Author Email: muhammad.dianqori@students.uii.ac.id
Submitted: **05/03/2026**; Accepted: **20/03/2026**; Published**: 20/03/2020**

Abstract−Hospital financial deficits due to INA-CBG claim discrepancies pose a critical challenge to healthcare sustainability in Indonesia. The difference between hospital operating costs and INA-CBG rates often results in significant financial deficits, which can threaten the sustainability of healthcare providers, especially hospitals. However, existing studies lack a systematic approach to identify distinct patterns of financial losses based on clinical treatment characteristics. This study aims to identify clusters of patients with different financial loss characteristics using a hybrid DBSCAN-K-Means clustering approach based on medical procedure frequency patterns. The DBSCAN algorithm was employed to detect and separate noise from data, while K-Means was used to identify medical treatment patterns. The data were obtained from electronic medical records of inpatients during the 2023–2024 period at a private hospital (N = 6,021 cases). The final clustering results revealed two main clusters with a highly significant difference in deficits between clusters ($p = 6.21 \times 10^{-38}$, Cliff's Delta = −0.216). Cluster 0 represents patients with intensive care who have a higher frequency of routine procedures, with an average deficit of 1.51 times (51.3% greater) and an average length of stay of 1.76 times (76% longer) than Cluster 1. Cluster 1 represents patients with a focus on obstetrics/neonatology with a predominance of Doppler procedures. Meanwhile, the noise cluster (13.39%) represents more extreme cases with an average loss of −7.82 million IDR and high mortality. Of the total 315 treatment features, 114 were confirmed to be statistically significant. This study contributes a novel hybrid clustering framework for identifying financial loss patterns in INA-CBG claims, providing actionable insights for hospital management to optimize service utilization, adjust procedure fees for complex cases, and strengthen financial risk management strategies.

Keywords: Clinical Pathway Pattern; Financial Loss Analysis; Hospital Claim Management; INA-CBG; Hybrid Clustering

1. INTRODUCTION

Ever since its implementation in 2014, the National Health Insurance also known as Jaminan Kesehatan Nasional (JKN) program managed by BPJS Kesehatan has significantly transformed the healthcare financing landscape in Indonesia. By the end of 2023, JKN membership coverage had reached more than 275 million people, and is projected to exceed 300 million by 2045 [1-2]. Within this framework, the healthcare payment mechanism shifted from a fee-for-service system to a prospective, fee-for-service system based on Indonesia Case-Based Groups (INA-CBGs) [3].

Under the INA-CBGs system, hospitals receive fixed reimbursement rates determined by diagnosis-related groups (ICD-10) and procedure codes (ICD-9), regardless of the actual costs incurred during patient care [4]. While this system was designed to standardize payments and curb unnecessary expenditures, but in practice generated substantial financial strain on healthcare providers [5]. One of the main problems faced is the high number of pending claims, with the average claim delay rate reported to be between 22% and 25% across various healthcare facilities [3,4,6]. In addition, a persistent discrepancy exists between the actual cost of delivering care and the predetermined INA-CBG tariffs, which in many cases resulting in financial losses for hospitals [7-10].

The consequences extend beyond fiscal instability. Prolonged financial pressure directly compromises service quality and operational sustainability. A studies conducted by [11] previously revealed that up to 68% of Indonesian hospitals struggle to maintain service quality due to prolonged financial pressure. Therefore, hospitals are compelled to continuously refine their financial and operational strategies not only to mitigate losses but also to preserve operational and service sustainability [12]. Despite numerous attempts to reconcile tariff gaps to address the issue, but the outcomes remain considered as not optimal [13]. Therefore it is a necessary to have more effective and data-driven alternative approach in real-world clinical practice.

Existing research on INA-CBG claims and patient clustering can be categorized into two main streams. The first stream focuses on clustering patients based on clinical characteristics. Existing research on INA-CBG claims and patient clustering can be categorized into two main streams. The first stream focuses on clustering patients based on clinical characteristics. For example, a study conducted by Galih and Putri in 2025 applied K-Means clustering to group 995 patients based on disease severity using clinical variables such as age, gender, triage, vital signs, and laboratory results [14]. Similarly, a study conducted by Aulia in 2025 used K-Means to identify disease risk levels based on six clinical indicators, including blood pressure, blood sugar, cholesterol, and uric acid levels [15]. In the context of national health insurance, a study by Yunitaningtyas and Herianti in 2025 used PCA combined with K-Means to cluster cancer patient visits based on cost severity and ward class [16]. Other research focuses on optimizing healthcare services, such as the study conducted by Pramarta et al. in 2025, which clustered patients based on demographics and disease type for resource allocation [17], as well as the work by Prabujaya et al. in 2025, who developed a web-based K-Means clustering system for managing medical records at community health centers [18].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

The second stream examines the administrative factors that influence the claims processing. A study conducted by Maulida and Djunawan in 2022 analyzed the causes of delayed claims at a university hospital and identified four main factors: incomplete documentation, coding errors, inadequate supporting examinations, and a lack of evidence of treatment [19]. Furthermore, a study by Setiawan et al. in 2025 conducted a qualitative study on the factors causing claim delays and found that discrepancies between medical and administrative documents, delayed data entry, and poor interdepartmental coordination were the primary causes [20]. Then, a study conducted by Amalia et al. in 2023 investigated the determinants of claim rejections based on diagnosis code accuracy and reported that 81% of rejected claims were due to inaccurate diagnosis coding [21].

However, despite these valuable contributions, several critical gaps remain when comparing the current research with previous studies. Some of these studies used the K-Means or PCA+K-Means methods with clinical and demographic variables such as age, gender, vital signs, and disease type, but none analyzed patterns of financial loss or frequency of medical care as clustering features [14-18]. Meanwhile, other studies examined administrative factors affecting claims, including document completeness, coding accuracy, and interdepartmental coordination, but did not use clustering methods to identify loss patterns [19-21]. In contrast, this study employs a hybrid DBSCAN-K-Means approach using medical treatment frequency patterns to identify financial loss clusters. Unlike all previous studies [14-21], this research simultaneously addresses three aspects: financial loss analysis, treatment pattern analysis, and hybrid clustering methodology. This represents a critical gap in research, particularly in the use of Electronic Health Records (EHR) data to analyze variations in medical practices that dictate true cost and have a direct impact on claim deficits [22]. To address these gaps, this study has three main objectives. First, to identify clusters of patients with different financial loss characteristics based on medical treatment frequency patterns. Second, to apply a hybrid DBSCAN-K-Means clustering approach for outlier detection and pattern identification in INA-CBG claims data. Third, to provide clinical interpretation of the identified loss clusters for hospital management decision-making.

Unlike prior work that relied on demographic variables or diagnoses to cluster patients, in this study we employ a hybrid unsupervised learning approach based on medical treatment patterns. We use numerical representations of the frequency of medical treatment codes as feature vectors, making it possible to identify patient groups based on similarities in actual clinical practice. The combination of the DBSCAN and K-Means algorithms was chosen based on their capabilities. DBSCAN is first employed to identify and isolate outliers in the data [23]. Then K-Means is applied to cluster high-dimensional data based on pattern similarity [24,25]. By combining these two algorithms, this approach makes it possible to identify robust "Loss Clusters", which are groups of cases with specific medical characteristics that consistently generate costs above the INA-CBG claim rate.

The contributions of this research are threefold. Methodologically, we propose a novel hybrid clustering framework combining DBSCAN for outlier removal and K-Means for pattern identification in healthcare claims data. Practically, the findings provide actionable insights for hospital management to optimize service utilization, adjust procedure fees for complex cases, and strengthen financial risk management. Scientifically, this study fills the research gap by systematically analyzing medical treatment patterns (rather than demographic variables) as drivers of financial losses in INA-CBG claims.

2. RESEARCH METHODOLOGY

2.1 Research Design and Data Source

This study employs a retrospective observational design on INA-CBG claim data from Hospital X, a private hospital located in Yogyakarta, Indonesia. The data collection period spans from January 2023 to January 2024 (13 months). The inclusion criteria for this study consisted of three requirements. First, inpatient claims must have deficit status where hospital costs exceed INA-CBG reimbursement. Second, complete medical records with medical treatment procedure codes must be available. Third, patients must be discharged, excluding ongoing treatment cases. The initial dataset comprised 6,021 claim cases meeting the inclusion criteria. Given the retrospective nature of this study and the use of anonymized secondary data, formal ethical approval was waived. Permission to access and use the data was obtained from the hospital management. All data were de-identified prior to analysis to protect patient confidentiality.

Data were obtained from two integrated sources. The first source was claims data containing patient demographic and administrative information from the hospital's billing system. The second source was electronic medical records (EMR) containing medical treatment procedure codes extracted from the hospital's EMR system. The integrated dataset contains 16 variables as shown in Table 1. The key feature for clustering analysis is the Treatment List column, which contains semicolon-separated treatment codes that undergo frequency encoding transformation.

Table 1. Variable in the Integrated Dataset

No.	Column Name	Data Types	Description
1.	ID	Category (object)	Patient claim identifier
2.	Age	Numeric (int64)	Patient age (years)
3.	Sex	Numeric (int64)	Gender
4.	Class	Numeric (int64)	Room class
5.	Diaglist	Category (object)	Diagnosis code (ICD-10)

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

No.	Column Name	Data Types	Description
6.	Proclist	Category (object)	Procedure code (ICD-9)
7.	Ina-cbg	Category (object)	INA-CBG group code
8.	Description	Category (object)	INA-CBG description
9.	Discharge Status	Numeric (int64)	Discharge outcome
10.	LOS	Numeric (int64)	Length of Stay (Days)
11.	Has ICU	Boolean	ICU admission indicator
12.	Hospital Fees	Numeric (int64)	Actual hospital cost (IDR)
13.	Claim Amt	Numeric (int64)	Total insurance claimed amount (IDR)
14.	Ina-cbg Fees	Numeric (int64)	INA-CBG tariff (IDR)
15.	Deficit	Numeric (int64)	Financial loss (IDR)
16.	Treatment List	Category (object)	Medical treatment codes

2.2 Research Stages

The research methodology consists of six systematic stages:

a. Data Collection and Integration: Merging claims data with EMR treatment procedure data into a unified dataset
b. Data Preprocessing: Applying frequency encoding, log-transformation, and normalization to treatment features
c. Clusterability Testing: Validating the existence of meaningful cluster structures using Sparse PCA + Silverman's Test and Hopkins Statistic
d. Clustering: Applying the hybrid DBSCAN-K-Means framework to identify financial loss patterns
e. Interpretation of Results: Describing cluster characteristics based on treatment patterns and financial outcomes
f. Statistical Evaluation: Conducting inferential tests (Kruskal-Wallis, Mann-Whitney U, FDR correction) to validate cluster differences

These six steps are illustrated in Figure 1.

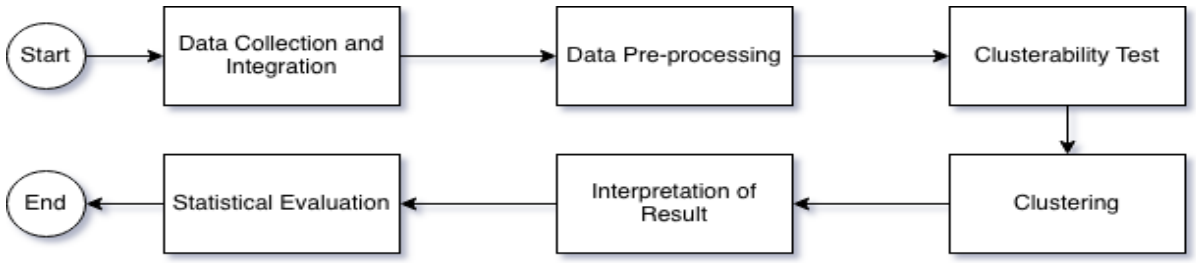

Figure 1. Research Stages

2.3 Data Collection and Integration

The data used in this study is patient data integrated from ina-cbg claims data and electronic medical records from a private hospital. This data consists of 6,021 patient records with deficit claims (the difference between the hospital rate and the claim value received is negative) for the period from January 2023 to January 2024 (one year).
The integrated dataset contains 16 variables categorized into three groups:

a. 3 demographic columns: Age, Sex, Class
b. 4 administrative columns: ID, INA-CBG, Description, Discharge Status
c. 9 clinical columns: Diaglist, Proclist, LOS, Has ICU, Hospital Fees, Claim Amt, INA-CBG Fees, Deficit, Treatment List

Table 2 presents a sample of the integrated dataset structure showing patient records with semicolon-separated treatment codes in the Treatment List column.

Table 2. Sample of Integrated Dataset

No.	ID	Age	……	Treatment List
1.	X203	2	……	A20009;….D10019
2.	X288	20	……	A20004;….D2007
3.	X207	53	……	A20009;….Q1022
……	……	……	……	……
6.019	X333	63	……	B10003;….D10021
6.020	X233	56	……	A20015;….D10015
6.021	X237	48	……	A20010;….D10021

2.3 Data Pre-processing

The Treatment List column serves as the primary input for clustering analysis. This column contains semicolon-separated treatment codes that require transformation into numerical features suitable for clustering algorithms.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

2.3.1 Frequency Encoding

Frequency Encoding is a technique that replaces category values with frequency statistics of their occurrence in a data set [26]. At this stage, medical treatment data from the treatment list column is selected as the clustering feature input and will undergo data frequency encoding transformation and normalization processes first. With this feature, treatment data that was previously a list of treatment codes separated by semicolons (;) of categorical data type will be converted into a new numeric feature that represents the frequency occurrence of each treatment code for each patient. The encoding process produced 315 treatment variables representing the variety and number of medical treatments received by each patient. Table 3 shows a sample of the encoded treatment features.

Table 3. Sample of Encoded Treatment Features

No.	B30060	B30034	……	D61036	E13022
1.	3	18	……	0	0
2.	1	6	……	0	0
3.	3	14	……	0	0
……	……	……	……	……	……
6.019	3	2	……	0	0
6.020	1	0	……	0	0
6.021	6	11	……	0	0

2.3.2 Data Log-Transformation

Because the encoded data represents the frequency of medical treatment procedures, where most patients have few procedures, but a handful of patients have a very high frequency of treatment procedures, this data generally has a right-skewed distribution and an extreme range of values [27]. To address this issue, a data transformation technique will be applied, namely log-transformation before proceeding to the next stage. The log-transformation technique aims to reduce the influence of extreme values by bringing the data distribution closer to normal [28]. Before and after log-transformation sample can be seen in Figure 2 and Figure 3.

```
Before Log-Transform:

[[ 3 18  1 ...  0  0  0]
 [ 1  6  1 ...  0  0  0]
 [ 3 14  1 ...  0  0  0]
 [ 2  7  0 ...  0  0  0]
 [ 2 19  2 ...  0  0  0]]
```

Figure 2. Sample Structure X Before Log-Transform

```
After Log-Transform:

[[1.38629436 2.94443898 0.69314718 ... 0.         0.         0.        ]
 [0.69314718 1.94591015 0.69314718 ... 0.         0.         0.        ]
 [1.38629436 2.7080502  0.69314718 ... 0.         0.         0.        ]
 [1.09861229 2.07944154 0.         ... 0.         0.         0.        ]
 [1.09861229 2.99573227 1.09861229 ... 0.         0.         0.        ]]
```

Figure 3. Sample Structure X After Log-Transform

2.3.3 Data Normalization

At this stage, the log-transformed features will be normalized. The use of normalization is crucial in data preprocessing, because machine learning models cannot produce good predictions if the input data contains noise or has inconsistent scales [29]. One commonly used method for data normalization is StandardScaler, which is used to normalize feature distributions by giving them a mean of 0 and a standard deviation of 1 [29,30].

$$X_{norm} = \frac{X-\mu}{\sigma} \quad (1)$$

Where X is the representation of the original feature value, μ is the feature mean, σ feature standard deviation, dan X_{norm} is the scaled value. This method ensures that each attribute has the same weight in the preprocessing mapping process, thereby preventing features with large scales from dominating the model calculations [30]. However, this method is highly sensitive to outliers because extreme values can significantly affect the calculation of the mean and standard deviation [25,30].

Because the encoding result feature has many columns representing the frequency occurrence of treatment, with basic medical treatment having a much higher frequency than specialized treatment, the feature is likely to contain outliers or a highly skewed distribution. Given by these conditions, a study suggests using RobustScaler, which uses the median and interquartile range (IQR) because it is much more resistant to outliers than StandardScaler [30].

$$X_{norm} = \frac{X-X_{median}}{IQR} \quad (2)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

The purpose of this entire data pre-processing series is none other than to be able to handle outliers in treatment features that are multidimensional, have skewed distributions, and contain many sparse zero values. Furthermore, the other non-treatment columns are excluded so that the analysis focuses on pure treatment patterns. The normalization using RobustScaler sample can be seen in Figure 4.

```
After RobustScaler:
[[ 1.          0.92599942  0.         ...  0.         0.
   0.        ]
 [ 0.         -0.51457317  0.         ...  0.         0.
   0.        ]
 [ 1.          0.5849625   0.         ...  0.         0.
   0.        ]
 [ 0.5849625  -0.32192809 -1.         ...  0.         0.
   0.        ]
 [ 0.5849625   1.          0.5849625  ...  0.         0.
   0.        ]]
```

Figure 4. Sample of RobustScaler Normalization Implementation

2.4 Hybrid Clustering

This study employs a two-stage hybrid clustering approach combining DBSCAN for outlier detection and K-Means for pattern identification.

2.4.1 Clusterability Test

Before clustering, data clusterability or cluster tendency will be tested to determine the existence of meaningful clusters in our dataset, thereby avoiding inappropriate grouping and misinterpretation of results. This is the ideal pipeline and a prerequisite before clustering [31].

Since the pre-processing results are high-dimensional data, an appropriate method is required. A study proposes Sparse Principal Component Analysis (SPCA) combined with Silverman's Critical Bandwidth Test as a highly recommended option for high-dimensional data such as medical or biostatistical data [32]. This method works through two main stages: first, performing smart dimensionality reduction, then applying statistical tests to detect group structures [32].

The experiment at n_components = 5 showed that the total variance explained by the five components reached 47.70%, indicating that the data structure is complex and spread across many dimensions. This condition is commonly found in medical treatment frequency-based data, which is highly heterogeneous.

Furthermore, the Hopkins statistic test was also run as additional support for clusterability testing. This is because the Hopkins statistic functions as a “gatekeeper” to prove that the data is indeed suitable for clustering before the algorithm is run [31]. In addition, the Hopkins method is also considered capable of handling high-dimensional data points by comparing the distance between random samples and their nearest neighbors [31]. A Hopkins test result of 1 indicates that the data has a strong group structure. If the value is > 0.5, the data has a strong group structure, but if the value is < 0.5, the data distribution is considered random or there are no meaningful groups, and if the value is close to 0, the data distribution is uniform.

The Hopkins statistic value of 1 indicates a very strong clustering tendency. However, this interpretation is contextualized carefully given the characteristics of high-dimensional data. The results of these clusterability test can be seen in Figure 5

```
Shape setelah SPCA: (6021, 5)

Explained variance ratio per component: [0.16977303 0.11721249 0.05702483 0.0551506  0.07779481]

Total explained variance: 0.4769557594811652

=== SILVERMAN CRITICAL BANDWIDTH TEST ===
SPCA-1: h* = 0.320, p-value = 0.4600
SPCA-2: h* = 0.410, p-value = 0.0180
SPCA-3: h* = 0.190, p-value = 0.1940
SPCA-4: h* = 0.720, p-value = 0.6620
SPCA-5: h* = 0.360, p-value = 0.6340

Hopkins Statistic (cluster tendency): 1.0
```

Figure 5. Sample of Clusterability Test Result using SPCA Combined with Silverman Critical Bandwidth and Hopkins Statistic

2.4.2 Stage 1: DBSCAN for Outlier Removal

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm, which works by dividing areas based on data density into clusters [23], [33,34]. Where a point environment p with radius Eps, denoted as $N_{eps}\ (p)$, and is defined as follow:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

$$N_{eps}(p) = \{q \in D | dist(p,q) \leq Eps\} \quad (3)$$

Where $dist(p,q)$ is usually the Euclidean distance function between two points p and q [35]. Next, a point p is considered a core point if the number of points in its neighborhood at least meets the minimum threshold ($MinPts$) defined as follow [35]:

$$\left|N_{eps}(p)\right| \geq MinPts \quad (4)$$

Furthermore a point p can be reached directly from point q if [35]:

$$p \in N_{eps}(q)$$

$$\left|N_{eps}(p)\right| \geq MinPts(core\ point\ condition) \quad (5)$$

A cluster is formed if all points in the cluster are density-connected, while points that do not meet this rule are classified as noise [36]. This algorithm is often used to detect outliers in data [23]. Outlier or noise results detected by this algorithm will be excluded and stored. This separation will produce new data that is considered clean for further application to the k-means algorithm.

In order to find optimal parameter for DBSCAN, the $MinPts$ value is usually initialized to approximately twice the number of effective features, while ε is estimated from the k-distance plot ($k = MinPts$) [37]. In this study, as an alternative in determining the optimal DBSCAN parameters, they will be obtained through a grid search scheme on the range $eps \in \{0.5,3.1,0.2\}$ and $min\_sample \in \{5,10,15,20\}$ with selection criteria based on a composite score defined as follows:

$$CS(\epsilon, minPts) = S - (0.5 \times D) - (2.0 \times R_n)$$

$$(\epsilon^*, minPts^*) = \arg\max CS(\epsilon, minPts) \quad (6)$$

Where CS is the composite score, S is the silhouette score, D is the dbi score and R_n is the $noise\_rasio$ generated in that iteration, and $(\epsilon^*, minPts^*)$ is the optimal parameter selected from the highest score. These parameters are then applied to the DBSCAN algorithm.

2.4.3 Stage 2: K-Means for Pattern Clustering

After obtaining clean data from DBSCAN clustering, the separated data is then applied to the k-means algorithm. K-means is one of the most popular clustering algorithms due to its simplicity and efficiency in handling large data sets [24], [38]. Has the ability to minimize the distance between data points and cluster centers (centroids) and its effectiveness in grouping objects based on similar characteristics iteratively [25]. The main objective is to minimize intra-cluster variance while maximizing inter-cluster separation [39]. In the process, the assignment of data points to clusters is based on minimizing the Euclidean distance between data points and cluster centers of all clusters, where the Euclidean distance is calculated as follows:

$$d(X_i, \mu_k) = \sqrt{\sum_{j=1}^{n}(x_{ij} - \mu_{kj})^2} \quad (7)$$

X_i represent the i-th data point, μ_k is the center of the k-th, n is the number of features in the data [39]. Next is the cluster assignment rule. After the distance is calculated, each data point is assigned to the cluster with the closest cluster center, thereby minimizing the distance metric for each point. The cluster assignment rule is formulated as follows:

$$c_i = \arg\min_k, d(x_i, \mu_k) \quad (8)$$

Where c_i denotes the cluster assigned to data point x_i. This step is very important to ensure that each data point is connected to the cluster with the closest cluster center based on Euclidean distance [39]. Next is the centroid update step, where the new centroid is calculated by taking the average value of all points assigned to that cluster. The cluster center point is updated using the following equation:

$$\mu_k = \frac{1}{N_k}\sum_{x_i \in C_k} x_i \quad (9)$$

Where μ_k is the updated cluster center for cluster k. N_k represents the number of points in cluster k. C_k is the set of points assigned to cluster k. This iterative recalculation of cluster centers continues until convergence, ensuring that cluster centers are optimally placed to minimize intra-cluster variance [39]. Next, minimize the within-cluster sum of squares. This stage is represented mathematically as follows:

$$J = \sum_{k=1}^{K}\sum_{x_i \in C_k} \| x_i - \mu_k \|^2 \quad (10)$$

Where J is the objective function to be minimized, K is the number of clusters, $\| x_i - \mu_k \|^2$ represents the squared distance between each data point x_i and its cluster center μ_k. Minimizing J to ensures that the data points in

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

each cluster are as close as possible to their cluster centers, resulting in compact clusters with high cohesion [39]. Compared to other algorithms, k-means has proven to have more stable performance and is suitable for data with evenly distributed spherical clusters [40].

Next, to select the optimal number of clusters (k) is one of the main challenges in data analysis, because there is no single answer that is absolutely correct [41]. Therefore, optimal k is sought by applying four complementary methods, namely the Elbow Method (based on Within-Cluster Sum of Squares or WCSS), and the optimal k looping scheme using Silhouette Score, Davies-Bouldin Index (DBI) and Calinski-Harabasz Index (CHI) to obtain comprehensive results.

The optimal looping process k is performed through iterations on the set $K = \{2,3, \dots ,10\}$. At each iteration K, the K-Means algorithm is trained to produce a cluster label set L_H. The quality of each clustering result was then evaluated using three metrics, namely silhouette, dbi, and chi. Based on the evaluation results, the optimal K value K^* was determined through the maximum value of the Silhouette Score function, which is formally defined as:

$$K^* = \arg max_{k \epsilon K}(S_k) \tag{11}$$

These internal evaluation metrics are used to measure the extent to which the cluster results are representative of the original data structure without requiring ground-truth labels [25]. The Silhouette metric is used to measure how well a data point fits its own cluster compared to the nearest cluster, with a value range of -1 to 1, where a value close to 1 indicates good cluster separation [24,25]. The Davies-Bouldin Index metric serves to evaluate cluster quality based on the level of internal compactness and separation between clusters. The smaller the DBI value, the better the resulting cluster separation [25], [38]. Then, the Calinski-Harabasz Index metric was selected to assess the ratio of variance between clusters to variance within clusters, where a higher value indicates that the clusters are clearly and efficiently separated [24,25].

2.5 Statistical Evaluation

At this stage, to validate the significance of differences in characteristics between clusters from clustering, a comprehensive inferential statistical test was conducted on outcome variables and patterns of medical treatment. The Kruskal-Wallis non-parametric test method was chosen because of its ability to handle non-normally distributed data and assess the difference in median losses between clusters [36]. The Kruskal-Wallis test statistic is defined as follows:

$$H = \frac{12}{N(N+1)} \sum_{t=1}^{k} \frac{Rt^2}{n_t} - 3(N+1) \tag{12}$$

Where N is the total number of observations, k is the number of groups, R_t is the number of ranks in group t, and n_t is the sample size in group t. If the Kruskal-Wallis test result is significant ($p < 0{,}05$), then proceed with the Mann-Whitney U test as a pairwise post-hoc test to identify specific differences between the two clusters [42]. Next, identify treatment features that differ significantly between clusters. Each treatment frequency feature is tested using the Mann-Whitney U test individually.

Given the large number of simultaneous tests (n = 315), correction was performed using the false discovery rate (FDR) method with the Benjamini-Hochberg method [43]. FDR is defined as the average ratio of false discoveries to total reported discoveries [43]. Features with a corrected p-value < 0.05 were considered statistically significant.

3. RESULT AND DISCUSSION

3.1 Clustering Result

The hybrid DBSCAN-K-Means clustering approach was applied to 6,021 INA-CBG claim cases with financial deficits. The grid search scheme was run with an evaluation that considered three main aspects, namely the data noise ratio, the number of clusters formed, and the composite score that combined the Silhouette Score, Davies–Bouldin Index, and noise penalty. Table results of the grid search scheme can be seen in Table 4, and the graph showing the results of the grid search can be seen at Figure 6 to Figure 8.

Table 4. Encoding Result for the Treatment List Column

Eps	Sil	DBI	Noise	Cluster (n)	Score
min_sample = 5					
0.50	0.287	1.159	99.27%	5	2.278
0.70	0.212	1.196	96.55%	10	2.316
……	……	……	……	……	……
……	……	……	……	……	……
min_sample = 10					
2.90	0.424	0.705	13.39%	2	-0.196
……	……	……	……	……	……
min_sample = 20					
2.90	✕ (Cluster<2)				

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

The graph in Figure 6 below shows how changes in the epsilon (eps) and min_samples parameters affect the number of clusters formed by the DBSCAN algorithm. It can be seen that at small eps values (0.5–0.9), a relatively large number of clusters are formed, particularly when min_samples = 5, reaching up to 18 clusters at eps = 0.9. This occurs because a small search radius makes it difficult for data points to merge with other clusters, resulting in many small clusters. As the eps value increases, the number of clusters generally decreases. This is because a larger search radius allows more data points to merge into the same cluster. For example, at min_samples = 10, the number of clusters decreases from approximately 8 clusters at eps = 0.9 to just 2 clusters at eps = 2.9. Additionally, a larger min_samples value tends to produce fewer clusters. This is because more data points are required to form a cluster, so some small data groups no longer meet the criteria for a cluster and are considered noise or merge with other clusters.

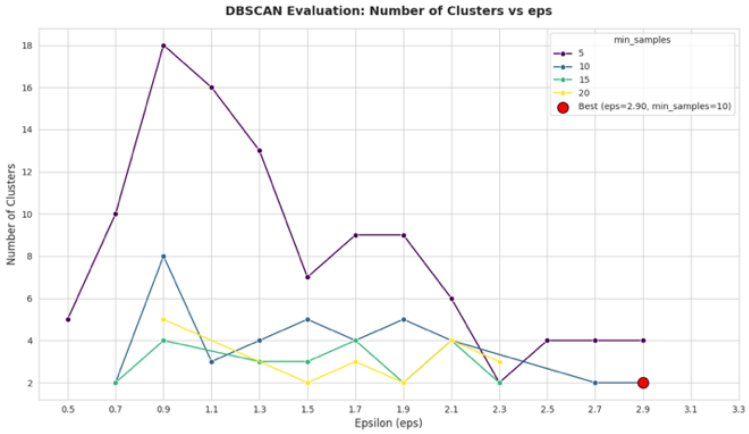

Figure 6. Grid Search (Number of Cluster vs Eps)

Next, the graph in Figure 7 shows that a higher composite score (closer to zero or positive) indicates better clustering quality, as it reflects clusters that are better separated, more compact, and have a lower proportion of noise. Based on the graph, it is evident that at small eps values (0.5–1.1), nearly all combinations of min_samples yield a low composite score (around -2.5 to -2.0). This indicates that clustering quality within the small eps range remains suboptimal. This occurs because an excessively small neighborhood radius prevents many data points from forming stable clusters, resulting in significant noise or small clusters. As the eps value increases, the composite score generally improves across all min_samples variations. This improvement indicates that the formed clusters become more stable as more data points can be connected within a single cluster. In the eps range of approximately 1.7–2.3, a significant improvement in clustering quality is observed. This indicates that the cluster structure is beginning to form more clearly. However, some parameter combinations still yield relatively lower composite scores compared to others. The best composite score on the graph is indicated by the red point, corresponding to the parameter combination eps = 2.9 and min_samples = 10. This combination yields the highest composite score, indicating that this parameter configuration provides the best balance between effective cluster separation, optimal cluster density, and minimal noise.

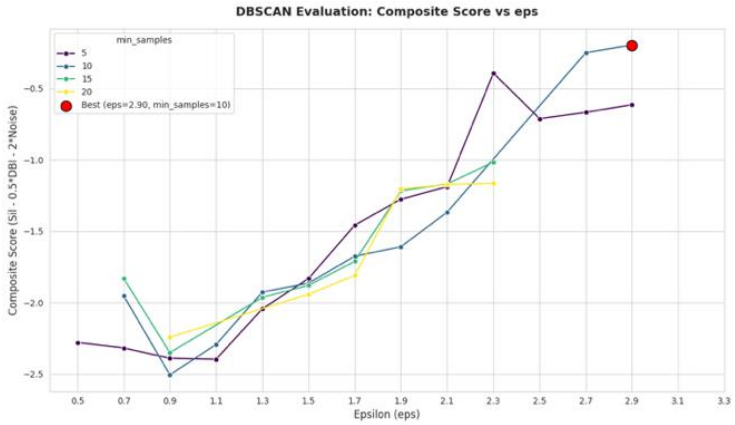

Figure 7. Grid Search (Composite Score vs Eps)

Furthermore, the graph in Figure 8 shows the relationship between the epsilon (eps) parameter value and the noise ratio produced by the DBSCAN algorithm for various values of min_samples. The noise ratio represents the proportion of data points that do not belong to any cluster and are considered outliers by the algorithm. Based on the graph, it is evident that at small eps values (around 0.5–0.9), the noise ratio is very high, even approaching 1.0 or

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

100% for some min_samples configurations. This indicates that most data points cannot form clusters. This occurs because the neighborhood search radius is too small, so many data points lack sufficient neighbors to meet the cluster formation criteria. As the eps value increases, the noise ratio decreases consistently across all min_samples variations. This decrease indicates that the larger the neighborhood radius used, the more data points can connect with each other and form clusters. In other words, increasing eps allows more data previously considered noise to become part of a cluster. It appears that larger values of min_samples tend to result in slightly higher noise ratios, particularly for small to medium values of eps. This is because more data points are required to form a cluster, so some small data groups do not meet this criterion and remain classified as noise. At an eps value of 2.9 with min_samples = 10, marked by the red point on the graph, the noise ratio reaches approximately 0.13 or 13%, which is one of the lowest values on the graph. This indicates that the majority of the data has been successfully grouped into clusters, while only a small portion remains classified as outliers. Thus, this graph shows that increasing the eps value contributes to a decrease in the noise ratio and the formation of more stable clusters. These results also support the selection of the optimal parameters eps = 2.9 and min_samples = 10, as this combination is capable of producing a relatively low noise ratio while providing good clustering quality based on other metric evaluations.

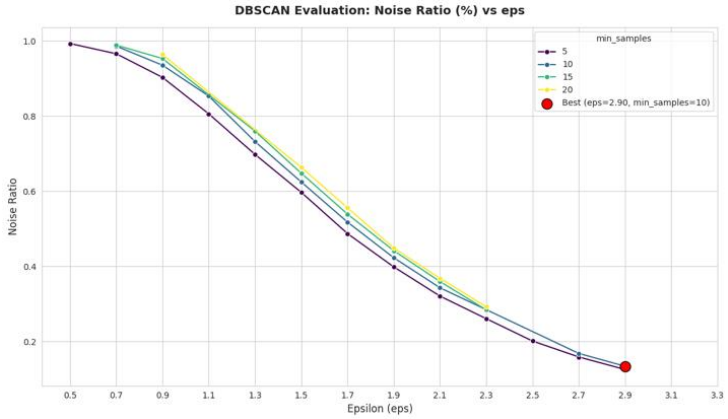

Figure 8. Grid Search Graph Result

Thus, based on these three evaluation graphs for the DBSCAN parameters (number of clusters, composite score, and noise ratio), the parameters eps = 2.9 and min_samples = 10 are the optimal parameters for the DBSCAN clustering process on this research dataset. With 806 cases (13.39%) were identified as noise/outliers, leaving 5,215 cases (86.61%) as clean data to be applied on next stage.

Next in the second stage, optimal number of clusters will determined through four complementary validation methods. First is the Elbow Method, this method showed a clear inflection point at k=2. The WCSS results graph can be seen in Figure 9.

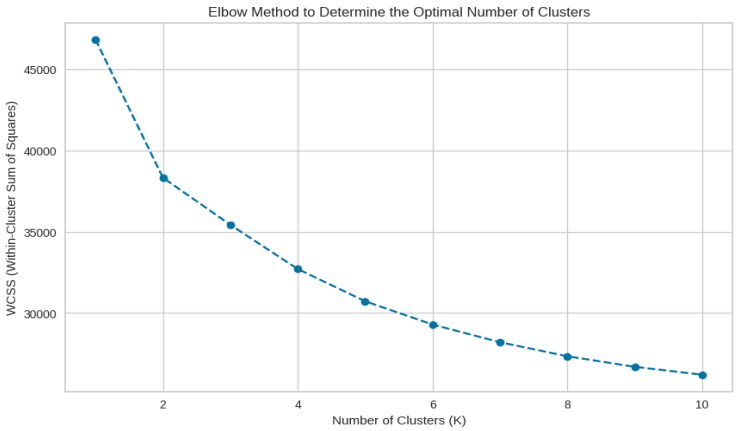

Figure 9. WCSS Graph Result

In the results shown in Figure 7 above, there is a significant decrease in the inertia score and an elbow is visible at k=2. Therefore, the optimal number of k concluded based on the “elbow” of this WCSS is k=2. These results are stored as references, then the optimal looping scheme is executed with three validation metrics in the next stage.

In this optimal looping k scheme, the number of iterations k is initialized with a range of 2 to 10. Each iteration of k will be validated using three metrics: Silhouette, DBI, and CHI. The results of this scheme's consistently confirmed k=2 as optimal, with Silhouette Score = 0.1686, DBI = 1.9167, and CHI = 1157.26. Each iterations in this scheme can be seen in Table 5, and the result of its visualization in Figure 10 to Figure 12.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

Table 5. Result of Each Iteration on the Optimal Looping k Scheme

Iteration	k	Silhouette	DBI	CHI
1.	2	0.1686	1.9167	1157.2626
2.	3	0.1211	2.0481	985.0215
3.	4	0.1350	1.9468	750.4045
4.	5	0.1235	2.1446	683.1802
5.	6	0.1300	2.0461	624.1391
6.	7	0.1260	2.0217	573.6322
7.	8	0.1134	2.0841	530.5937
8.	9	0.1158	2.0395	491.3752
9.	10	0.0927	2.2665	455.5601

Figure 10 illustrates the Silhouette Score analysis used to determine the optimal number of clusters (K) for the K-Means algorithm, ranging from K = 2 to K = 10. As shown in the figure, the highest Silhouette Score is achieved at K = 2 with a value of 0.1686. This indicates that partitioning the data into two clusters yields the best intra-cluster cohesion and inter-cluster separation compared to other configurations. When the number of clusters is increased to K = 3, the Silhouette Score drops sharply to 0.121. Although slight increases are observed at K = 4 (0.135), K = 6 (0.130), and K = 7 (0.126), all these values remain substantially lower than that of K = 2. For K ≥ 8, the score continues to decline, reaching its lowest point of 0.093 at K = 10. This pattern demonstrates that increasing the number of clusters beyond two does not enhance cluster quality; instead, it leads to poorer separation and less distinct cluster structures. Consequently, based on the Silhouette Score evaluation, the optimal number of clusters is K = 2.

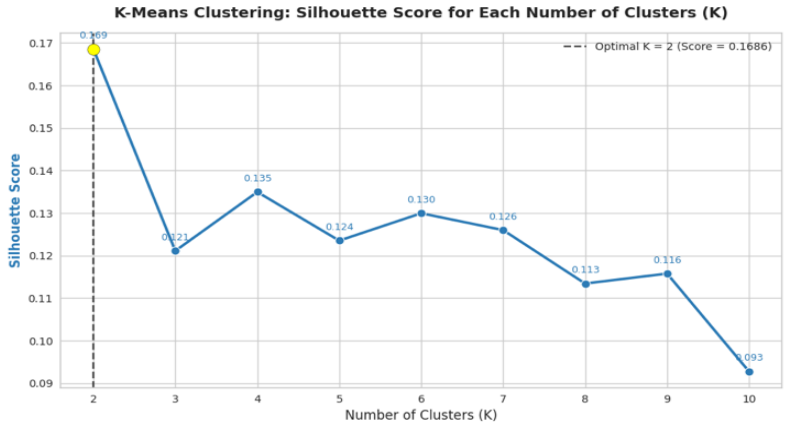

Figure 10. Graph of Optimal Looping k Scheme (Silhouette)

Figure 11 illustrates the Davies–Bouldin Index (DBI) analysis used to determine the optimal number of clusters (K) for the K-Means algorithm, ranging from K = 2 to K = 10. As shown in the figure, the lowest DBI value is achieved at K = 2 with a value of 1.917. This indicates that partitioning the data into two clusters yields the best cluster separation quality compared to other configurations. When the number of clusters is increased to K = 3, the DBI rises to 2.048. Although a slight decrease is observed at K = 4 (1.947), the value remains higher than that at K = 2. For K values ranging from 5 to 10, the DBI fluctuates but stays consistently higher, reaching 2.267 at K = 10. This pattern demonstrates that increasing the number of clusters beyond two does not improve cluster quality; instead, it results in poorer separation and greater similarity between clusters. Consequently, based on the Davies–Bouldin Index evaluation, the optimal number of clusters is K = 2. This result is consistent with the Silhouette Score analysis, reinforcing that K = 2 best represents the underlying data structure in this study.

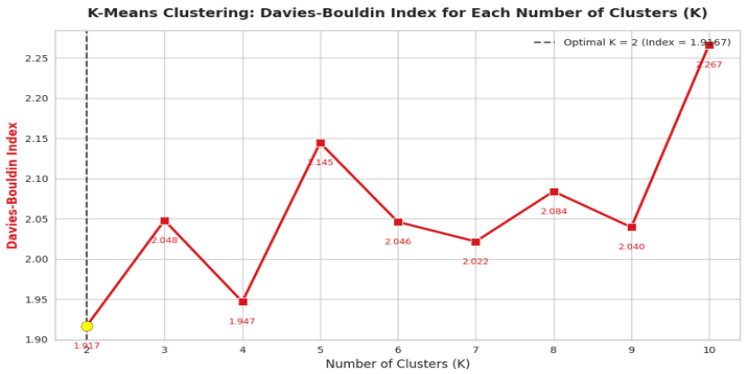

Figure 11. Graph of Optimal Looping k Scheme (DBI)

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

Figure 12 illustrates the Calinski-Harabasz Index (CHI) analysis used to determine the optimal number of clusters (K) for the K-Means algorithm, ranging from K = 2 to K = 10. As shown in the figure, the highest CHI value is achieved at K = 2 with a value of 1157.263. This indicates that partitioning the data into two clusters produces the optimal ratio of between-cluster dispersion to within-cluster dispersion compared to other configurations. When the number of clusters is increased to K = 3, the CHI drops sharply to 838.987. As the number of clusters continues to increase, the index decreases steadily, reaching its lowest point of 455.560 at K = 10. This pattern demonstrates that adding more clusters beyond two does not improve clustering quality; instead, it reduces both the compactness within clusters and the separation between them. Consequently, based on the Calinski-Harabasz Index evaluation, the optimal number of clusters is K = 2. This finding is fully consistent with the results from the Silhouette Score and Davies–Bouldin Index analyses, confirming that K = 2 best represents the underlying data structure in this study.

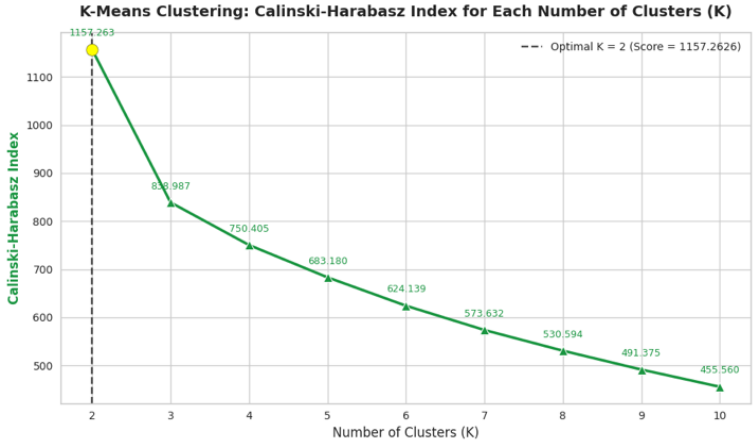

Figure 12. Graph of Optimal Looping k Scheme (CHI)

Based on the evaluation of the three internal validation metrics, it can be conclusively determined that the optimal number of clusters for this dataset is K = 2. The Silhouette Score achieved its maximum value at K = 2 (0.1686), indicating strong intra-cluster cohesion and clear inter-cluster separation for each data point. This value is substantially higher than those obtained for other K values, which exhibited an overall declining trend. Likewise, the Davies–Bouldin Index reached its minimum at K = 2 (1.917), representing the smallest ratio of average inter-cluster distance to intra-cluster distance (lower values indicate superior clustering quality). The index rose sharply for larger K values, peaking at 2.267 for K = 10. Similarly, the Calinski-Harabasz Index peaked at K = 2 (1157.263), reflecting the most favorable ratio of between-cluster to within-cluster dispersion. Collectively, all the looping optimal k scheme result show strong and consistent agreement on K = 2 as the optimal solution, with no conflicting indications or alternative elbow points. Therefore, the K-Means model with two clusters provides the most coherent, stable, and interpretable configuration.Subsequently, K = 2 was applied to the K-Means clustering algorithm. This produced two main clusters: Cluster 0, consisting of 3,393 patients (65.1%), and Cluster 1, consisting of 1,822 patients (34.9%). In addition, a noise cluster comprising 806 cases (13.39%) with extreme characteristics was identified.

Visualization of clustering results using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) was performed to provide a clear supporting picture of the separation between clusters and noise characteristics. In the 2D and 3D PCA plot results with noise, it can be seen that the noise points are widely scattered in the peripheral region, indicating that there are cases that have very different medical treatment patterns and do not match the main density structure of the data. PCA and UMAP visualization images can be seen in Figure 9 and Figure 10.

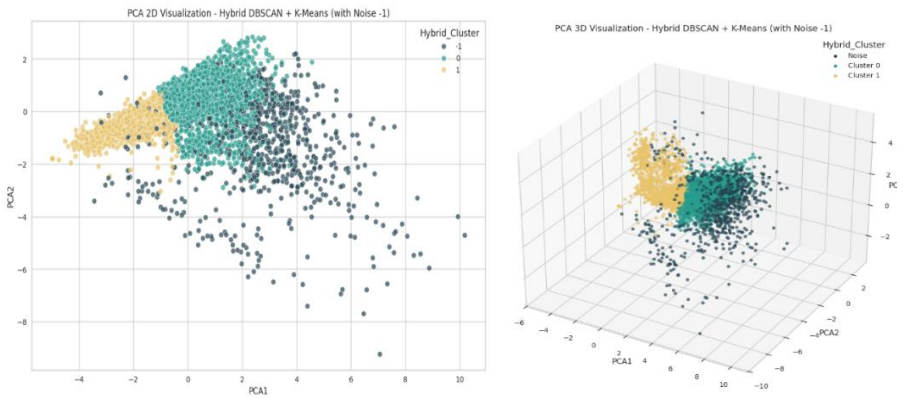

Figure 9. 2D and 3D PCA Visualization

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

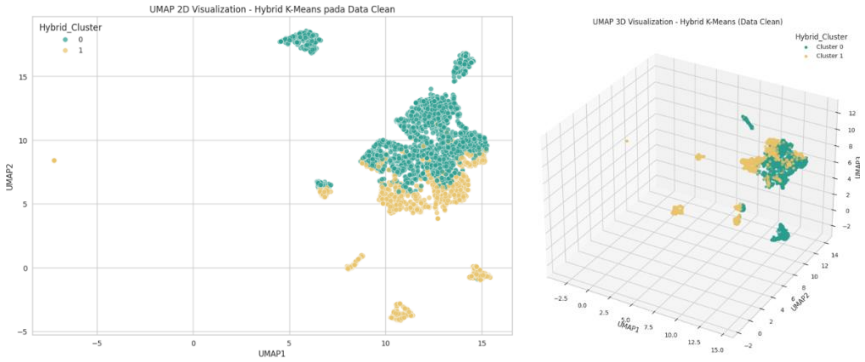

Figure 10. 2D and 3D UMAP Visualization (data clean)

3.2 Cluster Characteristics and Clinical Interpretation

The final results of the hybrid clustering between the DBSCAN and K-Means algorithms show a fairly balanced distribution in the clean data after noise removal. Cluster 0 includes 3393 patients (65.1%) and cluster 1 includes 1822 patients (34.9%), with the final evaluation metrics on clean data showing a Silhouette Score of 0.1686, Davies-Bouldin Index of 1.9167, and a Calinski-Harabasz Index of 1157.2626, indicating moderate cluster separation.

Cluster 0 exhibits significantly higher financial losses compared to Cluster 1, with mean deficit of IDR -2,015,027 versus IDR -1,331,624 (51.3% greater loss). Cluster 0 also shows longer length of stay (LOS) with mean 4.44 days versus 2.5 days (76% longer), and slightly older patient age (mean 38.6 vs. 34.4 years). Detailed characteristics of each cluster can be seen in Table 6.

Table 6. Demographic Characteristic Between Clusters

	Cluster 0	Cluster 1	Noise -1
		Deficit (IDR)	
Mean	-2,015,027E+06	-1,331,624E+06	-7,824,425E+06
Median	-1,483,189	-1,034,010	-5,185,314E+06
		LOS	
Mean	4.44	2.5	7.1
Median	4	2	6
		Age	
Mean	38.6	34.4	52.8
Median	42	31	59

a. Clinical Interpretation of Cluster 0 (High-Deficit Intensive Care Pattern):
Cluster 0 represents patients requiring intensive nursing care and multidisciplinary monitoring. The top 10 dominant treatment procedures (Table 3) reveal a pattern of high-frequency routine nursing interventions:
   1. Regular patient observation per shift (B30034): Mean frequency 12.24 vs. 5.46 in Cluster 1
   2. Fluid balance measures (B30103): Mean frequency 10.25 vs. 3.16 in Cluster 1
   3. Injection per shift (B30015b): Mean frequency 8.44 vs. 2.43 in Cluster 1
   4. Infusion replacement procedure (B30087): Mean frequency 6.61 vs. 1.67 in Cluster 1

   This pattern suggests patients with complex medical conditions requiring continuous monitoring, such as those with multiple comorbidities, post-surgical complications, or severe infections. The high frequency of specialist consultations (D10006) further supports this interpretation. From a clinical pathway perspective, these cases may benefit from early identification and targeted care protocols to optimize resource utilization while maintaining quality of care.

b. Clinical Interpretation of Cluster 1 (Obstetrics/Neonatology Pattern):
Cluster 1 shows lower treatment intensity with distinct procedural patterns. In addition to routine nursing care, this cluster exhibits procedures specific to obstetrics and neonatology, including Doppler treatment (B30062) and vaginal toucher procedures. This pattern is consistent with maternal and neonatal care cases, which typically have more predictable clinical pathways and shorter lengths of stay. The lower deficit in this cluster suggests that INA-CBG tariffs for obstetrics/neonatology cases are relatively better aligned with actual costs compared to complex medical cases in Cluster 0.

c. Clinical Interpretation of Noise Cluster (Extreme Cases):
The noise cluster (13.39% of total data) represents exceptionally complex cases with extreme financial losses (mean deficit IDR -7,824,425) and longest length of stay (mean 7.1 days). Treatment frequencies are substantially higher than both main clusters:
   1. Fluid balance measures: 19.37 (vs. 10.25 in Cluster 0)
   2. Injection per shift: 15.62 (vs. 8.44 in Cluster 0)
   3. Patient observation: 14.80 (vs. 12.24 in Cluster 0)

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

These cases likely represent critical care patients, multi-organ failure, or rare complications that fall outside typical clinical pathways. From a policy perspective, these cases may warrant special consideration for INA-CBG tariff adjustments or outlier payment mechanisms. Table and figure showing the top 10 dominant treatment procedures for each clusters can be seen in Table 7 and Figure 11.

Table 7. Comparison Top 10 Dominant Treatment Between Clusters

tim	Code	Type	Name	Mean
			Cluster 0	
1.	B30034	Nursing	Regular Patient Observation per Shift	12.24
2.	B30103	Nursing	Fluid Balance Measures	10.25
3.	B30015b	Nursing	Injection per Shift	8.44
4.	B30087	Nursing	Infusion Replacement Procedure	6.61
5.	B30082	Nursing	Oral Therapy	2.34
6.	B30060	Nursing	Verbedden	2.04
7.	D10006	Consultation	Specialist Doctor Visit	1.65
8.	B30049	Nursing	Treatment/Removal of IV	1.43
9.	A20015	Rental Equipment	Saturation Rental	0.83
10.	B30051	Nursing	Personal Hygiene	0.81
			Cluster 1	
1.	B30034	Nursing	Regular Patient Observation per Shift	5.46
2.	B30103	Nursing	Fluid Balance Measures	3.16
3.	B30015b	Nursing	Injection per Shift	2.43
4.	B30087	Nursing	Infusion Replacement Procedure	1.67
5.	B30082	Nursing	Oral Therapy	1.09
6.	B30049	Nursing	Infus Treatment/Removal	0.95
7.	B30060	Nursing	Verbedden	0.94
8.	A20015	Rental Equipment	Saturation Rental	0.71
9	B30062	Nursing	Doppler Treatment	0.60
10.	B30051	Nursing	Personal Hygiene	0.58
			Cluster Noise -1	
1.	B30103	Nursing	Fluid Balance Measures	19.37
2.	B30015b	Nursing	Injection per Shift	15.62
3.	B30034	Nursing	Regular Patient Observation per Shift	14.80
4.	B30087	Nursing	Infusion Replacement Procedure	10.23
5.	B30051	Nursing	Personal Hygiene	4.97
6.	B30068	Nursing	Syring Pump Treatment per Injection	4.86
7.	B30060	Nursing	Verbedden	4.31
8.	B30005	Nursing	Tube Feeding (3x per Shift)	4.13
9.	B30079	Nursing	GDS Check Procedure	3.82
10.	B30082	Nursing	Oral Therapy	3.68

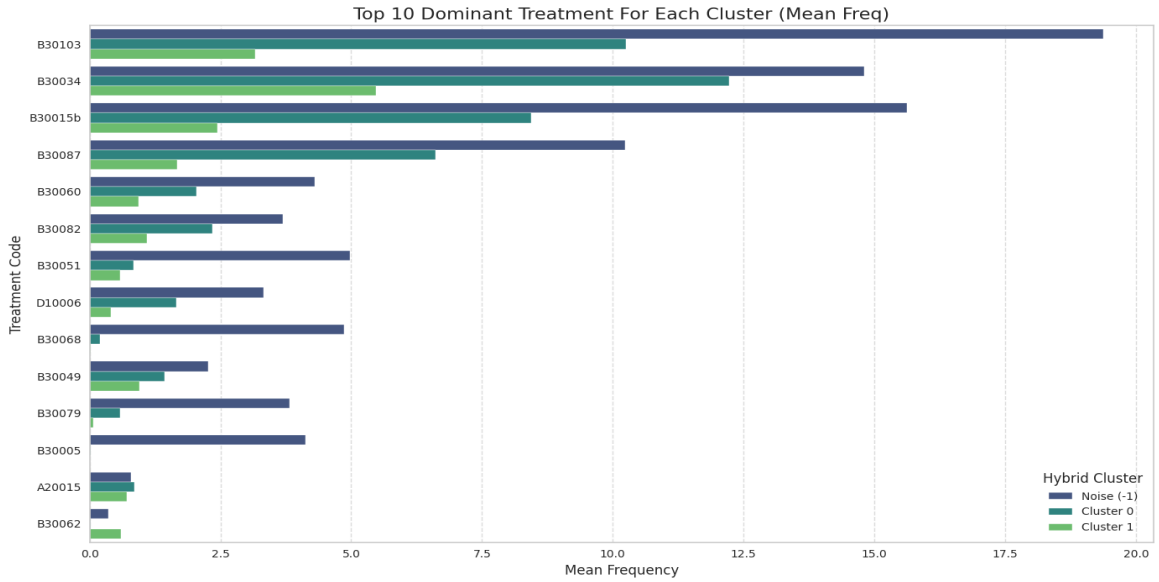

Figure 11. Boxplot Visualization of Top 10 Dominant Treatment Comparison for Each Clusters

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

3.3 Statistical Evaluation

The results of the statistical evaluation confirm that there is a very significant difference in the distribution of deficits (the difference between hospital rates and INA-CBG claim rates) between the two main clusters (0 and 1) in the clean data. The results of the statistical evaluation can be seen in Figure 12.

```
=== INFERENTIAL STATISTICAL TEST FOR DEFICIT ===

Kruskal-Wallis Test: H=165.77, p=6.21e-38

Mann-Whitney U Test: U=2423607.50, p=6.21e-38

Cliff's Delta Effect Size: -0.216

Interpretasi effect size: small
```

Figure 12. Statistical Evaluation Result for Deficit

The Kruskal-Wallis test yielded a statistic H = 165.77 with a p-value = 6.21×10^{-38}, indicating a strong rejection of H_0 that the deficit distributions between clusters are equal. The Mann-Whitney U (pairwise) follow-up test also showed consistent results with U = 2,423,607.50 and p = 6.21×10^{-38}. Cliff's Delta effect size of −0.216 indicates a small but clinically significant difference, where cluster 0 consistently shows more negative deficit values (greater loss) than cluster 1.

The top twenty features based on absolute effect size (Cliff's Delta) were dominated by routine nursing treatment, with the highest effect size values in B30103 (Fluid Balance Treatment) with an effect size of 0.95, B30015b (Injection per Shift) with an effect size of 0.91, B30034 (Regular Patient Observation per Shift) with an effect size of 0.85, and B30087 (Infusion Change Treatment) with an effect size of 0.85. Details of the significance difference test results for each feature can be seen in Figure 13.

```
=== FEATURE IMPORTANCE: SIGNIFICANT DIFFERENCE TEST FOR FEATURE ===

Number of significant features (FDR < 0.05): 114
Top 10 most significant features:

    feature  p_value_corrected  effect_size  significant
6    B30103       0.000000e+00     0.958807         True
5   B30015b       0.000000e+00     0.910175         True
1    B30034       0.000000e+00     0.858068         True
10   B30087       0.000000e+00     0.850529         True
12   D10006      3.058514e-290     0.585706         True
0    B30060      4.702414e-257     0.537077         True
9    D10015      3.093025e-184     0.425056         True
15   B30049      8.155641e-154     0.365035         True
7    B30082       2.021153e-66     0.281689         True
27   B10002       1.132505e-79     0.251647         True
60   D62002      6.718532e-186    -0.238008         True
17   A20004      1.248681e-154    -0.206719         True
18   B30062      3.644510e-154    -0.206144         True
21   B30043       1.793232e-44     0.187310         True
44   B10016       3.672702e-73     0.181676         True
16   D62007      1.598298e-104    -0.168720         True
58   B20122      3.119565e-124    -0.162104         True
3    A20009       5.468108e-58     0.135506         True
11   B30079       1.793232e-44     0.133949         True
4    B30065       2.092145e-57     0.133612         True
```

Figure 13. Significance Difference Test on Top 20 Treatment Features

Feature importance analysis using the Mann-Whitney U test on each medical action frequency feature, followed by multiple testing correction using the Benjamini-Hochberg method (FDR < 0.05), identified 114 statistically significant features. These features had a significantly higher total frequency in cluster 0 than in cluster 1, where cluster 0 was the cluster with a higher deficit than cluster 1. This shows that the intensity and frequency of basic nursing treatment, especially observation, fluid management, injections, and infusion replacement, were the main distinguishing factors between clusters. In addition, basic nursing treatment, specialist consultations (D10006, D10015), and several non-surgical procedures (B10002) also contributed significantly, even tough with a lower effect size.

3.4 Discussion

a. Comparison with Previous Studies:
   Our findings align with and extend several previous studies on INA-CBG claims and healthcare clustering: First, unlike clustering studies that used demographic variables [14-18], our approach based on medical treatment frequency patterns provides more actionable insights for clinical pathway optimization. For instance, [14] clustered patients by disease severity using clinical variables but did not address financial outcomes. Similarly, [15] focused on disease risk classification without considering claim deficits. Our study demonstrates that treatment pattern clustering directly identifies financial loss drivers, enabling targeted cost management interventions. Second, compared to claims analysis studies [19-21] that examined administrative factors (coding accuracy, document completeness), our hybrid clustering approach can identify clinical patterns underlying financial losses.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

b. Practical Implications for Hospital Management:
Based on our findings, we recommend the following interventions:
1. For cluster 0 (High-Deficit Intensive Care): Implement early warning systems to identify patients likely to require intensive nursing care. Consider clinical pathway redesign for complex medical cases with multiple comorbidities.
2. For Noise Cluster (Extreme Cases): Develop outlier payment negotiation protocols with BPJS Kesehatan. Document these cases separately for tariff adjustment advocacy.
3. For Overall Claim Management: Establish treatment frequency monitoring dashboards focusing on the 114 significant features identified. Flag cases exceeding threshold frequencies for early financial review.

c. Policy Implications:
Our findings have implications for INA-CBG tariff policy:
1. Nursing care intensity should be considered in tariff calculations, not just diagnosis and procedures.
2. Outlier payment mechanisms are needed for extreme cases (13.39% of deficit claims).
3. Clinical specialty adjustments may be warranted, as obstetrics/neonatology shows better tariff alignment than complex medical cases.

d. Limitations:
This study has several limitations. First, data sourced from a single private hospital, which may limit generalizability to public hospitals or other regions. Second, the retrospective design precludes causal inference. Third, treatment frequency does not capture treatment complexity or resource intensity per procedure. Future studies should incorporate multi-hospital data and consider procedure complexity weights.

4. CONCLUSION

This study successfully identified significant differences in patterns of medical treatment utilization among inpatients using a hybrid clustering approach combining DBSCAN and K-Means algorithms, with preprocessing including log-transformation, RobustScaler normalization, and removal of 13.39% noise (extreme cases). Supported by highly significant differences in deficits between clusters ($p = 6.21 \times 10^{-38}$, Cliff's Delta = −0.216), with 114 significant treatment features identified after FDR correction and the dominance of routine nursing codes (effect size > 0.85 in the top four features), the findings indicate that patterns of high-frequency service utilization in basic nursing interventions are a dominant factor driving financial losses in the INA-CBG system. This study makes three key contributions: (1) methodologically, we propose a novel hybrid DBSCAN - K-Means clustering framework for healthcare claims analysis that effectively handles outliers and high-dimensional treatment data; (2) practically, the findings provide actionable insights for hospital management to optimize service utilization, develop evidence-based clinical guidelines, audit nursing action utilization to identify potential over-utilization, and strengthen financial risk management strategies; and (3) scientifically, this research fills a critical gap by systematically analyzing medical treatment patterns (rather than demographic variables) as drivers of financial losses in INA-CBG claims. The identified clusters may reflect cost accumulation not fully covered by package rates, variations in clinical practice patterns, or disease complexity, especially in patients with longer hospital stays or intensive monitoring needs. The results of inferential tests and feature importance analysis validate the quality of the clusters produced and provide sufficient empirical evidence to support intervention measures, including potential adjustment of INA-CBG rates as a financial risk management measure for hospitals and policy considerations for outlier payment mechanisms.

REFERENCES

[1] G. Widjaja, Wagiman, D. Ersita Yustanti, H. H. Sijabat, and H. Dhanudibroto, "Evaluasi Implementasi Kebijakan Jaminan Kesehatan Nasional (JKN) dalam Meningkatkan Akses Layanan Kesehatan Masyarakat di Indonesia," *JK: Jurnal Kesehatan*, vol. 3, no. 2, 2025, doi: 10.62668/jukesah.v3i02.1311.

[2] M. N. Faiz, R. S. R. Kandau, and F. P. Gurning, "Evaluation of JKN Implementation in Improving Access to Health Services in Medan Artikel Review," *Jurnal Kolaboratif Sains*, vol. 8, no. 7, 2025, doi: 10.56338/jks.v8i7.8206.

[3] E. N. Oktoriani, A. Rosarini, and I. A. Lubis, "Analysis of ICD-9-CM Coding Accuracy in the Reimbursement Claim Process for INA-CBGs Compliance," *ICISTECH*, vol. 5, no. 1, 2025, doi: 10.62951/icistech.v5i1.215.

[4] G. A. E. Sutrisnawati, F. Manuaba, and S. P. M. E. Purwani, "Analysis of the Suitable of INA-CBG's Claim Coding at BPJS Kesehatan Wasin at RSUP Prof. Dr. I.G.N.G.Ngoerah Denpasar," *Jurnal Health Sains*, vol. 04, no. 12, 2023, doi: 10.46799/jhs.v4i12.1162.

[5] W. P. Nugraheni, A. H. Zahroh, and R. K. Hartono, "Best Practice of Hospital Management Strategy to Thrive in the National Health Insurance Era," *Indonesian Journal of Health Administration*, vol. 9, no. 1, 2021, doi: 10.20473/jaki.v9i1.2021.9-22.

[6] P. Maulina, I. Sartika, N. Ani, and B. Rahardjo, "Educational Interventions and Digital Innovation for Improving BPJS Inpatient Claims," *The Journal of Educational Development*, vol. 13, no. 1, 2025, doi: https://doi.org/10.15294/jed.v13i1.24372.

[7] I. Gumala Andirasdini, D. Saputra, M. Rivai, and S. E. M. Putra, "Analysis of the Health Social Security Administration (BPJS Kesehatan) Claim Amount using Random Forest Regression," *IAJ Indonesian Actuarial Journal Persatuan Aktuaris Indonesia (PAI)*, vol. 01, no. 01, 2025, doi: https://doi.org/10.65689/iajvol01no1pp001-008.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

[8] N. Igusti, A. R. Amalia AP, and Arman, "Optimizing Hospital Tariffs and Resource Allocation through Unit Cost Analysis: Lessons from a Major Indonesian Public Hospital," *Journal of Current Health Sciences*, vol. 5, no. 3, 2025, doi: 10.47679/jchs.2025127.

[9] R. Nurul Fathah and T. Anggita Safitri, "Analisis Perbandingan Tarif Rumah Sakit dan Tarif INA CBG Pada Pelayanan Rawat Inap di RS PKU Muhammadiyah Yogyakarta," *EKOMA : Jurnal Ekonomi*, vol. 3, no. 3, 2024, doi: https://doi.org/10.56799/ekoma.v3i3.3126.

[10] R. D. Monica, F. M. Firdaus, I. P. Lestari, Y. Suryati, D. Rohmayani, and A. Hendrati, "Analisis Perbedaan Tarif Riil Rumah Sakit dengan Tarif Ina-CBG's Berdasarkan Kelengkapan Medis Pasien Rawat Inap pada Kasus Persalinan Sectio Caesarea guna Pengendalian Biaya Rumah Sakit TNI AU Dr. M. Salamun Bandung," *Jurnal Manajemen Informasi Kesehatan Indonesia*, vol. 9, no. 1, 2021, doi: 10.33560/jmiki.v9i1.289.

[11] F. N. Laila, V. Paramarta, F. Yulianty, K. Kosasih, and A. Febriani, "Hospital Finances, Service Quality in Hospital Care and the Indonesian National Health Insurance System," *Malahayati International Journal of Nursing and Health Science*, vol. 8, no. 6, 2025, doi: 10.33024/minh.v8i6.874.

[12] W. L. M. Manopo and N. Susanti, "Prediksi Kebangkrutan Rumah Sakit Akibat Ketidaksesuaian Sistem Tarif INA-CBGS," *JAFM: Journal of Accounting and Finance Management*, vol. 6, no. 2, 2025, doi: 10.38035/jafm.v6i2.

[13] Y. Firmanto, "Early Warning System: Solusi Klaim Negatif Rumah Sakit Program Jaminan Kesehatan Nasional," *JEAM: Jurnal Ekonomi Akutansi Manajemen*, vol. 20, no. 1, 2021, doi: https://doi.org/10.19184/jeam.v20i1.19303.

[14] Aliasi Galih and Putri Raissa Amanda, "Penerapan K-Means Clustering untuk Pengelompokan Pasien Rumah Sakit berdasarkan Penerapan K-Means Clustering untuk Pengelompokan Pasien Rumah Sakit berdasarkan Tingkat Keparahan Penyakit," *Jatilima: Jurnal Multimedia dan Teknologi Informasi*, vol. 07, no. 3, 2025, doi: 10.54209/jatilima.v7i03.1598.

[15] W. Aulia, A. Putera Utama Siahaan, L. Marlina, and M. Iqbal, "Analisis Algoritma K-Means Clustering dalam Identifikasi Tingkat Risiko Penyakit Berdasarkan Data Rekam Medis Pasien," *Journal of Science and Social Research*, vol. 8, no. 3, 2025, Accessed: Mar. 13, 2026. [Online]. Available: https://garuda.kemdiktisaintek.go.id/documents/detail/5310778

[16] K. Yunitaningtyas and Herianti, "Identifying Stratifications of Cancer Patient Visits: Approach of Clustering Using PCA of Mixed Data," *ICDSOS*, vol. 2025, no. 1, 2025, doi: 10.34123/icdsos.v2025i1.622.

[17] Paramarta VIP, Johan Siti Hantina, Eralsyah M. Nabil Sulthoni, Abdulah Mohamad, Kena Erna, and Takbiranda Milano Wibi, "Penerapan Metode Clustering Dalam Peningkatan Pelayanan Kesehatan Di Puskesmas Binanga, Sulawesi Barat," *J-CEKI: Jurnal Cendekia Ilmiah*, vol. 4, no. 3, 2025, doi: 10.56799/jceki.v4i3.8021.

[18] I. Prabujaya, Rahman, F. Ibrahim, A. Azhari Muin, and S. Wahyuni, "Sistem Klasterisasi Data Rekam Medis pada Puskesmas Kampili Menggunakan Algoritma K-Means," *SYNTECH: System Information and Computer Technology*, vol. 1, no. 2, 2025.

[19] E. S. Maulida and A. Djunawan, "Analisis Penyebab Pending Claim Berkas BPJS Kesehatan Pelayanan Rawat Inap Rumah Sakit Universitas Airlangga," *Media Kesehatan Masyarakat Indonesia*, vol. 21, no. 6, 2022, doi: 10.14710/mkmi.21.6.374-379.

[20] A. P. Setiawan, S. Setiatin, and Y. A. Nuraeni, "Analisis Penyebab Penundaan Klaim BPJS Kesehatan Pasien Rawat Inap Di Rumah Sakit X," *Jurnal Ilmiah Perekam dan Informasi Kesehatan Imelda*, vol. 10, no. 2, 2025, doi: 10.52943/jipiki.v10i2.1939.

[21] R. Amalia, S. Lestari, A. Ferdianto, P. S. Akbar, and N. Fardilan, "Determinan Pengembalian Berkas Klaim Berdasarkan Akurasi Kode Diagnosis Pasien Rawat Inap BPJS Kesehatan di RSI Sultan Agung Semarang)," *Jurnal Rekam Medis dan Informasi Kesehatan*, vol. 6, no. 2, 2023, doi: 10.31983/jrmik.v6i2.10633.

[22] S. Sethi, "Feature Engineering for Healthcare Big Data: Approaches to Missing Data Imputation, Dimensionality Reduction, and Time-Series Analysis," *International Journal of Multidisciplinary Research and Growth Evaluation.*, vol. 1, no. 1, 2020, doi: 10.54660/.ijmrge.2020.1.1.120-124.

[23] E. Akbar, "Perbandingan Algoritma DBSCAN - K-Means dan K-Means untuk Pengelompokan Madrasah Aliyah Provinsi Jawa Timur," Universitas Islam Negeri Syarif Hidayatullah, 2023. Accessed: Sep. 11, 2025. [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/68812

[24] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, "The Silhouette Coefficient and the Davies-Bouldin Index are more Informative than Dunn Index, Calinski-Harabasz Index, Shannon Entropy, and Gap Statistic for Unsupervised Clustering Internal Evaluation of Two Convex Clusters," *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/peerj-cs.3309.

[25] P. M. Hasugian, D. Mathelinea, S. Simamora, and P. B. N. Simangunsong, "Comparative Evaluation of Data Clustering Accuracy through Integration of Dimensionality Reduction and Distance Metric," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 24, no. 3, 2025, doi: 10.30812/matrik.v24i3.5057.

[26] T. Johnson, A. J. Liu, S. Raza, and A. McGuire, "A Comparison of Modeling Preprocessing Techniques," *arXiv preprint arXiv:2302.12042*, 2023, doi: 10.48550/arXiv.2302.12042.

[27] M. Yu and Y. Zhang, "Comparative Models on Low Multiplier DRG Classification for Advanced Lung Cancer," *Front. Public Health*, vol. 13, 2025, doi: 10.3389/fpubh.2025.1614938.

[28] D. Arifuddin, K. Kusrini, and K. Kusnawi, "Perbandingan Performansi Algoritma Multiple Linear Regression dan Multi Layer Perceptron Neural Network dalam Memprediksi Penjualan Obat," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 2, 2025, doi: 10.57152/malcom.v5i2.1952.

[29] N. M. Nayan, A. M. Rana, M. M. Islam, J. Uddin, T. Yasmin, and J. Uddin, "An Interpretable and Balanced Machine Learning Framework for Parkinson's Disease Prediction using Feature Engineering and Explainable AI," *PLoS One*, vol. 20, no. 10, 2025, doi: 10.1371/journal.pone.0333418.

[30] J. M. H. Pinheiro *et al.*, "The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3635541.

[31] A. F. Diallo and P. Patras, "Deciphering Clusters with a Deterministic Measure of Clustering Tendency," *IEEE*, vol. 36, no. 4, 2024, doi: 10.1109/TKDE.2023.3306024.

[32] J. Laborde, P. A. Stewart, Z. Chen, Y. A. Chen, and N. C. Brownstein, "Sparse Clusterability: Testing for Cluster Structure in High Dimensions," *BMC Bioinformatics*, vol. 24, no. 1, 2023, doi: 10.1186/s12859-023-05210-6.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2640−2656
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9483

[33] Muh. Y. Fauzan, Garno, and Y. Umaidah, "Penerapan Algoritma DBSCAN dan K-Means untuk Clustering Penderita Pneumonia di Kabupaten Karawang," *JATI: Jurnal Mahasiswa Teknik Informatika*, vol. 9, no. 3, 2025, doi: 10.36040/jati.v9i3.13410.

[34] A. Saputra and R. Yusuf, "Perbandingan Algoritma DBSCAN dan K-MEANS dalam Segmentasi Pelanggan Pengguna Transportasi Publik Transjakarta Menggunakan Metode RFM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, 2024, doi: 10.57152/malcom.v4i4.1516.

[35] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 1996. doi: 10.5555/3001460.3001507

[36] N. S. S. Oktavia, A. Iriany, and A. B. Astuti, "Algoritma DBSCAN dan Shared Nearest Neighbor dalam Pengelompokan Spasial Produktivitas Jeruk Siam di Indonesia," *Jurnal Ilmiah Matematika*, vol. 13, no. 3, 2025, doi: 10.26740/mathunesa.v13n3.p45-58.

[37] F. Razaq and R. Muliono, "Deteksi Pola Kunjungan Pasien Berdasarkan Status Kesehatan Menggunakan Algoritma DBSCAN," *INCODING: Journal of Informatics and Computer Science Engineering*, vol. 5, no. 2, 2025, doi: 10.34007/incoding.v5i2.979.

[38] M. Almaripat, A. Faqih, and A. R. Rinaldy, "Sales Data Classterization Analysis Using K-Means Method for Marketing Strategy Development," *JAIEA: Jurnal of Artificial and Engineering Application*, vol. 4, no. 2, 2025, doi: 10.59934/jaiea.v4i2.792.

[39] S. Arham Akheel, "Understanding Learning Styles for Adaptive Learning Systems Using K-Means Clustering," *IJFMR: International Jurnal for Multidisciplinary Research*, vol. 1, no. 3, 2019, doi: 10.36948/ijfmr.2019.v01i03.10061.

[40] R. Wulandari, "Perbandingan Performa Algoritma K-Means dan DBSCAN dalam Clustering pada Data Berdimensi Tinggi," Undergraduate Thesis, Universitas Sumatera Utara, 2025. [Online]. Available: https://repositori.usu.ac.id/handle/123456789/108486

[41] E. Schubert, "Stop Using the Elbow Criterionfor K-Means," *Association for Computing Machinery*, vol. 25, no. 2023, 2023, doi: 10.1145/3606274.3606278.

[42] Z. M. Milenovic, "Application of Mann-Whitney U Test in Research of Professional Training of Primary School Teacher," vol. 6, no. 1, 2011, doi: 10.32728/mo.06.1.2011.06.

[43] M. Rostami and O. Saarela, "A Feature Selection Method that Controls the False Discovery Rate," *arXiv preprint arXiv:2208.02948*, 2023, doi: 10.48550/arXiv.2208.02948.