K-Means Clustering for Liver Patient Stratification Using Elbow Method and Silhouette Score Validation

Satria Aldi Pramudya
Informatics Departments, Universitas Sahid Surakarta, Surakarta, Indonesia
satriaaldipramudyaa@gmail.com

***Abstract*—Liver disease is one of the global health problems that causes millions of deaths every year, with a high incidence rate in India due to hepatitis infection, alcohol consumption, and unhealthy lifestyle factors. The treatment of liver disease requires accurate early diagnosis, but manual identification of clinical patterns in patients takes time and is prone to interpretation errors. Various studies have utilized biochemical data from patients for liver disease classification, but automatically grouping patients based on the similarity of biochemical profiles without diagnostic labels is still uncommon. This study applies the K-Means algorithm on the Indian Liver Patient Dataset containing 583 patient medical records with 10 biochemical and demographic features, combined with preprocessing including missing value imputation, label encoding, and StandardScaler normalization, as well as the use of the Elbow Method to determine the optimal number of clusters and the Silhouette Score for cluster quality evaluation. The results show that k=3 is the best number of clusters with a Silhouette Score of 0.208, producing three patient groups, namely critical condition with very high bilirubin and enzyme levels, normal biochemical condition, and moderate biochemical disorder. These findings indicate that K-Means has the potential to be used as an early patient stratification tool for liver disease to support the development of artificial intelligence-based clinical decision support systems.**

Keywords—K-Means Clustering, Liver Disease, ILPD, Elbow Method, Silhouette Score.

I. INTRODUCTION

Liver disease is one of the serious global health problems that is responsible for millions of deaths every year [1]. This condition includes various disorders such as hepatitis, cirrhosis, and hepatocellular carcinoma, which often develop without obvious symptoms in the early stages, making it a significant health threat, especially in developing countries. In India, the incidence of liver disease is quite high due to a combination of risk factors such as alcohol consumption patterns, hepatitis B and C infections, and unhealthy lifestyle [1]. Effective management requires accurate early diagnosis before the patient condition develops to a more severe and life-threatening stage [2].

In the context of data-based early detection, blood biochemical examinations such as total bilirubin, direct bilirubin, SGPT, SGOT, alkaline phosphatase, albumin, and total protein are the main indicators of liver function that are used as features in machine learning models [3]. Machine learning-based approaches have been proven capable of uncovering hidden patterns from clinical data that are difficult to identify manually, including predictive analysis and patient grouping [4], [5], [6], [7].

Unlike classification that requires labeled data, unsupervised learning such as K-Means allows automatic grouping of patients based on the similarity of biochemical profiles without requiring diagnostic labels [8]. This approach is very relevant when clinical data is partially or not yet labeled [9] and has the potential to identify subgroups of patients with similar biochemical characteristics for further clinical purposes.

K-means clustering works by iteratively minimizing the intra-cluster distance until the formed clusters are stable [10]. This algorithm is well known for its simplicity, computational speed, and ease of result interpretation, so it is widely used in disease data clustering analysis [10], [11], [12]. To determine the optimal number of clusters and evaluate cluster quality, two commonly used validation methods are the Elbow Method and Silhouette Score, which have been proven effective in various medical data clustering studies [13].

The novelty of this research lies in the integration of systematic preprocessing including missing value imputation, label encoding, and StandardScaler normalization with a combination of two cluster validation methods, namely the Elbow Method and Silhouette Score, in one comprehensive experimental framework. This approach is different from previous studies that generally only use ILPD for binary classification without evaluating the patient phenotype structure in depth.

The objectives of this research are (1) to identify the optimal number of clusters in the ILPD patient biochemical data using the Elbow Method; (2) to evaluate cluster quality using the Silhouette Score; and (3) to interpret the biochemical characteristics of each patient group formed. The main contributions include providing comprehensive biochemical profile-based clustering analysis on ILPD, identifying three clinical phenotypes, and providing an empirical basis for the development of artificial intelligence-based clinical decision support systems [14], [15]. Subsequently, Section II describes the research methods, Section III presents the results and discussion, and Section IV contains the conclusion.

II. METHOD

This section describes the research framework and the research procedure in order to address the research objectives. The proposed method includes five major steps: (1) dataset preparation, (2) data preprocessing, (3) applying the K-Means clustering algorithm, (4) finding the optimal number of clusters

using the elbow method, and (5) evaluating the cluster quality using the Silhouette Score. The general flowchart of this research is shown in Fig. 1.

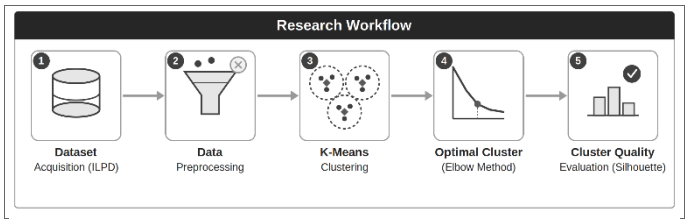

Fig. 1. Research Workflow Block Diagram

A. Dataset

The dataset used in this study is the *Indian Liver Patient Dataset* (ILPD) [16], which is publicly available on the *UCI Machine Learning Repository*. This dataset contains 583 patient medical records consisting of 416 liver disease patients and 167 non-liver disease patients, with 10 biochemical and demographic features as input variables in the clustering process. The feature details are presented in Table 1.

TABLE I. FEATURE DESCRIPTION OF INDIAN LIVER PATIENT DATASET (ILPD)

No.	Feature Name	Data Type	Description
1	Age	Integer	Patient age (years)
2	Gender	Categorical	Patient gender (Male/Female)
3	Total Bilirubin	Float	Total bilirubin level in blood
4	Direct Bilirubin	Float	Direct bilirubin level in blood
5	Alkaline Phosphotase	Integer	Alkaline phosphatase enzyme level
6	Alamine Aminotransferase	Integer	SGPT enzyme level
7	Aspartate Aminotransferase	Integer	SGOT enzyme level
8	Total Proteins	Float	Total protein in blood
9	Albumin	Float	Albumin level in blood
10	Albumin & Globulin Ratio	Float	Albumin to globulin ratio

All numerical features are used as clustering variables, while the diagnostic label is not included since this research uses an unsupervised learning approach. The selection of ILPD is based on its availability as a benchmark, which is widely used in liver disease research based on machine learning [1], [3], and [16].

B. Data Preprocessing

The *preprocessing* stage is carried out to ensure that the data is clean and ready to be processed by the K-Means algorithm. There are three sub-stages executed sequentially.

1) Missing Value Imputation:

The *Albumin and Globulin Ratio* feature in the ILPD dataset contains four missing values. These missing values are imputed using the median value of the corresponding feature. The median approach is chosen because it is robust against *outliers* that are commonly found in clinical data [17].

2) Label Encoding:

The *Gender* feature is the only categorical feature in the dataset. This feature is converted into numerical representation using *label encoding*, where the value *Male* is changed to 1 and *Female* to 0, so that all features become numerical and can be processed by the K-Means algorithm.

3) Feature Normalization

All features are normalized using *StandardScaler* so that they have *mean* = 0 and *standard deviation* = 1. The *StandardScaler* transformation is defined in Equation (1).

$$z = \frac{(x - \mu)}{\sigma} \quad (1)$$

where x is the original feature value, μ is the feature mean, and σ is the feature standard deviation.

C. K-Means Clustering Algorithm

K-Means *Clustering* is an *unsupervised learning* algorithm that partitions data into k groups based on the distance similarity between data points and their cluster centroid [18]. The algorithm minimizes the *inertia* value (*sum of squared distances*) as defined in Equation (2)

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

where k is the number of clusters, C^t is the set of data points in the i-th cluster, x is the data point, and μ^t is the centroid of the i-th cluster. The K-Means procedure is executed through four iterative steps: (1) random centroid initialization, (2) assigning each data point to the nearest centroid based on Euclidean distance, (3) updating centroid positions as the mean of all cluster member data points, and (4) repeating steps 2 and 3 until convergence or maximum iteration is reached. In this study, k-means++ initialization is used with a maximum iteration of 300 and n_init=10.

D. Determination of Optimal Number of Clusters (Elbow Method)

The determination of optimal k value is performed using *Elbow Method*. This method calculates the *inertia* value for a range of k from 2 to 7, then the values are visualized as a curve. The "elbow" point on the curve, where the decrease in *inertia* starts to significantly flatten, indicates the most efficient number of clusters [8]. The selection of *Elbow Method* is based on its ease in identifying the optimal point without requiring data labels [8], [10].

E. Cluster Quality Evaluation (Silhouette Score)

Silhouette Score is used to evaluate the quality of the formed clusters by measuring how well each data point is placed in its own cluster compared to other clusters [13]. The *Silhouette Score* for each data point is defined in Equation (3).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3)$$

where *a(i)* is the average distance of data point i to all other points in the same cluster (*cohesion*), and *b(i)* is the average distance to the nearest different cluster (*separation*). *Silhouette Score* values range from −1 to 1, where values close to 1 indicate good clusters, values close to 0 indicate overlap, and negative values indicate incorrect assignment [13]. The combination of *Elbow Method* and *Silhouette Score* is chosen

because both methods complement each other in evaluating cluster quality [8], [10].

F. Implementation Environment

All stages of the research are implemented using Python 3.x. Libraries used include Pandas and NumPy for data manipulation, Scikit-learn for K-Means algorithm implementation, StandardScaler, and Silhouette Score calculation, as well as Matplotlib and Seaborn for result visualization.

III. RESULTS AND DISCUSSION

A. Preprocessing Results

Before the clustering process is performed, three *preprocessing* sub-stages are applied sequentially to the ILPD dataset. All of these stages are essential to ensure data integrity and readiness before entering the K-Means algorithm.

1) Missing Value Imputation:

The Albumin and Globulin Ratio feature contains 4 missing values (0.69% of 583 medical records). Imputation is done using the median value of 0.90 g/dL.

2) Label Encoding:

The Gender feature is converted from categorical to numerical (Male → 1, Female → 0). Gender distribution consists of 441 male patients (75.6%) and 142 female patients (24.4%). After encoding, all 10 features are numerical and ready to be processed.

3) Feature Normalization:

StandardScaler normalization is applied to all features to eliminate the dominance of large-scale features. Before normalization, there was an extreme scale difference: Alkaline Phosphotase ranged from 63–2110 IU/L while Albumin was only 0.9–5.5 g/dL. After normalization, all features have mean = 0 and standard deviation = 1.

B. Determination of Optimal Number of Clusters (Elbow Method)

Inertia values are calculated for k = 2 to 7 and visualized in an elbow curve as shown in Fig. 2. The inertia calculation results are presented in Table 2.

TABLE II. K-MEANS INERTIA VALUES FOR K = 2 TO 7

No. of Clusters (k)	Inertia Value	Inertia Decrease	Decrease (%)
2	4,823.6	–	–
3	3,891.4	932.2	19.3%
4	3,287.5	603.9	15.5%
5	2,951.8	335.7	10.2%
6	2,741.3	210.5	7.1%
7	2,594.6	146.7	5.4%

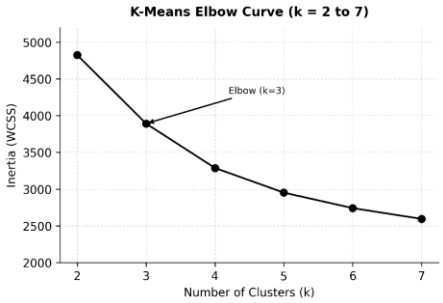

Fig. 2. K-Means Elbow Curve (k = 2 to 7)

Based on Table 2 and Fig. 2, the most significant inertia decrease occurs from k = 2 to k = 3, which is 932.2 points (19.3%). The subsequent decrease from k = 3 to k = 4 is proportionally smaller (603.9 points, 15.5%), and continues to decrease for larger k values. The curve shows the clearest elbow point at k = 3, where the rate of inertia decrease starts to flatten significantly. This result indicates that k = 3 is the most efficient number of clusters, as adding more clusters beyond k = 3 does not provide a proportional reduction in inertia.

C. Cluster Quality Evaluation (Silhouette Score)

Silhouette Score is calculated for each k value to quantitatively evaluate cluster separation quality. The evaluation results are presented in Table 3.

TABLE III. SILHOUETTE SCORE VALUES FOR K = 2 TO 7

No. of Clusters (k)	Silhouette Score	Description
2	0.195	Below k = 3
3	0.208	Highest – Optimal
4	0.183	Decreased from k = 3
5	0.169	Decreased
6	0.157	Decreased
7	0.148	Lowest

Based on Table 3, k = 3 produces the highest Silhouette Score of 0.208 compared to other k values. This Silhouette Score falls within the low to moderate range ($0.20 < SS < 0.50$), which aligns with the characteristics of liver disease clinical data that inherently exhibit overlap between patient subgroups due to high biological variation. This finding aligns with previous studies indicating that the Silhouette Score for medical data is often below 0.5 due to the complexity of continuous clinical phenotypes [13]. It is important to note that both validation methods consistently show k = 3 as the optimal number of clusters, thereby strengthening confidence in the clustering results obtained.

D. Characteristics of Each Cluster

The application of K-Means with k = 3 produces three patient groups with different biochemical distributions and characteristics as shown in Table 4 and Table 5.

TABLE IV. PATIENT DISTRIBUTION PER CLUSTER

Cluster	Phenotype Label	No. Patients	Percentage (%)
0	Normal Biochemical Condition	244	41.9%
1	Critical Condition	94	16.1%
2	Moderate Biochemical Disorder	245	42.0%

TABLE V. AVERAGE BIOCHEMICAL FEATURE VALUES PER CLUSTER (AFTER INVERSE TRANSFORMATION)

Feature	Cluster 0 (Normal)	Cluster 1 (Critical)	Cluster 2 (Moderate)	Normal Range*
Total Bilirubin (mg/dL)	0.7	5.8	1.6	0.2 – 1.2
Direct Bilirubin (mg/dL)	0.2	3.2	0.7	0.0 – 0.3
Alkaline Phosphotase (IU/L)	158	358	238	44 – 147
Alamine Aminotrans. (IU/L)	26	182	72	7 – 56
Aspartate Aminotrans. (IU/L)	32	198	88	10 – 40
Total Proteins (g/dL)	6.6	6.8	6.7	6.3 – 8.2
Albumin (g/dL)	3.3	2.9	3.1	3.5 – 5.0
Albumin & Globulin Ratio	1.02	0.85	0.94	1.0 – 2.5
Age (years)	43.1	44.3	45.2	–

**Normal range based on standard laboratory reference values.*

1) Cluster 0 – Normal Biochemical Condition (n = 244, 41.9%)

Cluster 0 is characterized by an average total bilirubin value of 0.7 mg/dL and direct bilirubin of 0.2 mg/dL, which are within the normal range. Liver enzyme activity is also relatively low, with average SGPT of 26 IU/L and SGOT of 32 IU/L. The albumin level in this cluster (3.3 g/dL) is approaching the lower limit of the normal range, indicating a relatively maintained liver synthesis function. Overall, the biochemical profile of Cluster 0 is the closest to normal conditions and indicates relatively adequate liver function compared to the other two clusters.

2) Cluster 1 – Critical Condition (n = 94, 16.1%)

Cluster 1 shows the most severe biochemical profile. Average total bilirubin reaches 5.8 mg/dL (almost five times the upper normal limit) and direct bilirubin 3.2 mg/dL, indicating a severe bilirubin metabolism disorder as seen in cirrhosis or liver failure. The activity of liver enzymes is very high: the average SGPT is 182 IU/L (3.2 times the upper normal limit), and the average SGOT is 198 IU/L (5.0 times the upper normal limit). This evidence indicates that the liver cells are very damaged. Alkaline phosphatase reaches 358 IU/L, far above the normal range (44–147 IU/L), which may indicate cholestasis or hepatic infiltration. Additionally, the lower albumin level (2.9 g/dL) reflects a decline in the liver's ability to synthesize protein, and an albumin-globulin ratio below 1.0 (0.85) is a common laboratory sign found in chronic liver disease.

3) Cluster 2 – Moderate Biochemical Disorder (n = 245, 42.0%)

Cluster 2 shows a biochemical profile between Cluster 0 and Cluster 1. An average total bilirubin of 1.6 mg/dL (above the normal limit of 1.2 mg/dL) indicates mild hyperbilirubinemia. Liver enzymes show moderate elevation: SGPT 72 IU/L and SGOT 88 IU/L, approximately 1.3 and 2.2 times the upper normal limit, respectively. The average alkaline phosphatase of 238 IU/L also exceeds the normal range. This group likely represents patients with early- to moderate-stage liver damage, which has the potential to develop into a critical condition if not properly treated.

E. Discussion

This research successfully identified three liver disease patient phenotypes in ILPD using K-means clustering combined with the elbow method and silhouette score validation. These results address all three research objectives: (1) k = 3 is identified as the optimal number of clusters through the elbow method; (2) cluster quality is confirmed with the highest Silhouette Score of 0.208 at k = 3; and (3) three clinical phenotypes are successfully interpreted based on the biochemical profiles of each cluster's centroid.

Compared to previous studies, the majority of research on ILPD uses a binary classification approach with supervised learning methods. A supervised approach using a support vector machine with SMOTE and ensemble filter feature selection achieved 85% accuracy on liver disease classification [3], while various feature selection methods were compared on the same task [4]. Ensemble machine learning was also demonstrated to achieve high accuracy for liver disease prediction using the ILPD dataset [19]. Those studies depend on the availability of diagnostic labels, whereas the K-Means approach proposed in this research does not require labels, making it more flexible in clinical scenarios where diagnostic labels are not yet available.

The three resulting phenotypes have significant clinical relevance. Cluster 1 (Critical) with a total bilirubin of 5.8 mg/dL and SGOT 198 IU/L is consistent with laboratory findings of severe acute hepatitis or decompensated cirrhosis, consistent with patient stratification studies on acute liver failure [20]. Cluster 0 (Normal) corresponds to the biochemical profile of non-liver disease patients or patients with still-compensated liver function. Cluster 2 (Moderate) potentially represents patients who need closer monitoring, because moderate enzyme elevation can be an early sign of progressive hepatocellular damage.

The Silhouette Score value of 0.208 is higher compared to values of 0.12–0.19 reported on K-Means clustering of drug user review data [8] and comparable to values of 0.20–0.22 reported on toddler nutrition data clustering using K-Means++ [10], which supports the validity of the approach used in this study [10]. Furthermore, multi-dimensional pre-clustering applied to hepatitis C data confirmed that unsupervised patient grouping produces clinically meaningful subgroups [21].

Several limitations of this work merit consideration. First, the ILPD dataset exhibits a notable class imbalance (71.4% of records belong to liver disease patients versus 28.6% non-liver disease patients), which may introduce bias into the resulting

cluster distribution. Second, K-Means is inherently susceptible to the choice of initial centroids; while K-Means++ initialization and n_{init}=10 were employed to reduce this sensitivity, the risk cannot be entirely eliminated. Third, a systematic external validation of the identified clusters against confirmed diagnostic labels was not carried out, which limits the clinical interpretability of the results. Future work could explore more flexible clustering approaches, including DBSCAN, Hierarchical Clustering, and Gaussian Mixture Models, which are better suited for capturing non-spherical structures and density variations commonly found in clinical data.

IV. CONCLUSION

This study applied K-Means clustering to the ILPD dataset comprising 583 patient records and 10 biochemical and demographic variables. The preprocessing pipeline covered missing value imputation using the median, binary encoding of the gender feature, and StandardScaler normalization. The Elbow Method guided the selection of the optimal cluster count, and the Silhouette Score provided quantitative validation of clustering quality.

The most pronounced inertia drop occurred between k = 2 and k = 3, with a reduction of 932.2 points (19.3%), marking k = 3 as the elbow point. This was further confirmed by the Silhouette Score, which peaked at 0.208 for k = 3 across all evaluated values from k = 2 to k = 7. The consistent agreement of both validation methods strengthens the reliability of k = 3 as the optimal cluster configuration.

Three clinically meaningful patient phenotypes were yielded: (1) Cluster 0 – Normal Biochemical Condition (n = 244, 41.9%), with biochemical values within the normal reference range and intact liver function; (2) Cluster 1 – Critical Condition (n = 94, 16.1%), marked by severely elevated total bilirubin averaging 5.8 mg/dL and SGOT reaching 198 IU/L, indicative of significant hepatocellular damage; and (3) Cluster 2 – Moderate Biochemical Disorder (n = 245, 42.0%), showing mild-to-moderate enzyme elevation that may reflect early-stage liver deterioration if not addressed.

Overall, the findings confirm that K-Means clustering offers a viable label-free approach for early stratification of liver disease patients based on their biochemical profiles. The results lay a practical foundation for AI-driven clinical decision support development. Subsequent work could investigate alternative clustering methods such as DBSCAN or Gaussian Mixture Models, incorporate external diagnostic validation, and extend the analysis to broader and more varied liver disease cohorts to improve generalizability.

REFERENCES

[1] [1] R. A. Khan, Y. Luo, and F.-X. Wu, “Machine learning based liver disease diagnosis: A systematic review,” *Neurocomputing*, vol. 468, pp. 492–509, Jan. 2022, doi: 10.1016/j.neucom.2021.08.138.

[2] [2] I. Straw and H. Wu, “Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction,” *BMJ Heal. Care Informatics*, vol. 29, no. 1, p. e100457, Apr. 2022, doi: 10.1136/bmjhci-2021-100457.

[3] [3] M. A. Nugraha, M. I. Mazdadi, A. Farmadi, Muliadi, and T. H. Saragih, “Penyeimbangan Kelas SMOTE dan Seleksi Fitur Ensemble Filter pada Support Vector Machine untuk Klasifikasi Penyakit Liver,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1273–1284, Dec. 2023, doi: 10.25126/jtiik.2023107234.

[4] [4] A. Z. Al Wafi, A. Z. Al Wafi, F. P. Rochim, and V. Bezaleel, “Investigating Liver Disease Machine Learning Prediction Performancethrough Various Feature Selection Methods,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 24, no. 3, pp. 507–520, Jul. 2025, doi: 10.30812/matrik.v24i3.4531.

[5] [5] R. Rani *et al.*, “Enhancing liver disease diagnosis with hybrid SMOTE-ENN balanced machine learning models—an empirical analysis of Indian patient liver disease datasets,” *Front. Med.*, vol. 12, May 2025, doi: 10.3389/fmed.2025.1502749.

[6] [6] S. M. Ganie, P. K. Dutta Pramanik, and Z. Zhao, “Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches,” *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 160, Jun. 2024, doi: 10.1186/s12911-024-02550-y.

[7] [7] H. Ding, M. Fawad, X. Xu, and B. Hu, “A framework for identification and classification of liver diseases based on machine learning algorithms,” *Front. Oncol.*, vol. 12, no. October, pp. 1–7, Oct. 2022, doi: 10.3389/fonc.2022.1048348.

[8] [8] Safitri Juanita and R. D. Cahyono, “K-MEANS CLUSTERING WITH COMPARISON OF ELBOW AND SILHOUETTE METHODS FOR MEDICINES CLUSTERING BASED ON USER REVIEWS,” *J. Tek. Inform.*, vol. 5, no. 1, pp. 283–289, Feb. 2024, doi: 10.52436/1.jutif.2024.5.1.1349.

[9] [9] Ajeng Arifa Chantika Rindu, Ria Astriratma, and Ati Zaidiah, “K-Means Algorithm Implementation for Project Health Clustering,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 5, pp. 1064–1076, Sep. 2023, doi: 10.29207/resti.v7i5.5181.

[10] [10] Muhammad Raqib Syahkur, D. Hartama, and S. Solikhun, “Evaluasi Jumlah Cluster pada Algoritma K-Means++ Menggunakan Silhouette dan Elbow dengan Validasi Nilai DBI dalam Mengelompokkan Gizi Balita,” *JST (Jurnal Sains dan Teknol.*, vol. 13, no. 3, pp. 487–496, Oct. 2024, doi: 10.23887/jstundiksha.v13i3.86419.

[11] [11] S. A. D. Darmawan and Karmilasari, “Penerapan Metode *K-Means Clustering* dan *Simple Moving Average* untuk Memprediksi Jenis Penyakit di Provinsi Jawa Timur,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 877–886, Aug. 2024, doi: 10.25126/jtiik.1148703.

[12] [12] A. Aljohani, “Optimizing Patient Stratification in Healthcare: A Comparative Analysis of Clustering Algorithms for EHR Data,” *Int. J. Comput. Intell. Syst. 2024 171*, vol. 17, no. 1, pp. 173-, Jul. 2024, doi: 10.1007/S44196-024-00568-8.

[13] [13] M. Shutaywi and N. N. Kachouie, “Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering,” *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021, doi: 10.3390/e23060759.

[14] [14] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, “Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms,” *Informatics Med. Unlocked*, vol. 36, p. 101155, Jan. 2023, doi: 10.1016/j.imu.2022.101155.

[15] [15] B. Yang, H. Lu, and Y. Ran, “Advancing non-alcoholic fatty liver disease prediction: a comprehensive machine learning approach integrating SHAP interpretability and multi-cohort validation,” *Front. Endocrinol. (Lausanne).*, vol. 15, p. 1450317, Oct. 2024, doi: 10.3389/fendo.2024.1450317.

[16] [16] B. Ramana and N. Venkateswarlu, “ILPD (Indian Liver Patient Dataset) - UCI Machine Learning Repository.” Accessed: Jul. 06, 2026. [Online]. Available: https://archive-beta.ics.uci.edu/dataset/225/ilpd%2Bindian%2Bliver%2Bpatient%2Bdataset

[17] [17] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, “Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease,” *Biomedicines*, vol. 11, no. 2, p. 581, Feb. 2023, doi: 10.3390/biomedicines11020581.

[18] [18] A. U. Rehman *et al.*, “A Machine Learning-Based Framework for Accurate and Early Diagnosis of Liver Diseases: A Comprehensive Study on Feature Selection, Data Imbalance, and Algorithmic Performance,” *Int. J. Intell. Syst.*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/6111312.

[19] [19] W. El Atifi, O. El Rhazouani, F. M. Khan, and H. Sekkat, "Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare," *PLoS One*, vol. 20, no. 8, p. e0330899, Aug. 2025, doi: 10.1371/journal.pone.0330899.

[20] [20] S. Palomino-Echeverria *et al.*, "A robust clustering strategy for stratification unveils unique patient subgroups in acutely decompensated cirrhosis," *J. Transl. Med.*, vol. 22, no. 1, p. 599, Jun. 2024, doi: 10.1186/s12967-024-05386-2.

[21] [21] A. Sharma, T. Khade, and S. M. Satapathy, "A cross dataset meta-model for hepatitis C detection using multi-dimensional pre-clustering," *Sci. Rep.*, vol. 15, no. 1, p. 7278, Mar. 2025, doi: 10.1038/s41598-025-91298-0.