International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Application of Cluster Analysis Based on SHAP Values in Hemodialysis Patients Using Arteriovenous Fistula

Peng Shu*, Ling Huang*, Xia Wang, Zhuping Wen, Yiqi Luo, Fang Xu

The Central Hospital of Wuhan,Tongji Medical College,Huazhong University of Science and Technology, Wuhan, Hubei Province, People's Republic of China

*These authors contributed equally to this work

Correspondence: Peng Shu; Fang Xu, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, No. 26, Shengli Street, Jiang'an District, Wuhan, Hubei Province, People's Republic of China, Tel +8615802716692, Email 312855784@qq.com; 453328433@qq.com

International Journal of General Medicine downloaded from https://www.dovepress.com/
For personal use only.

Background: The prognosis of hemodialysis patients using arteriovenous fistula is significantly heterogeneous and influenced by various factors, including vascular conditions and underlying diseases. This study aims to reveal patient subgroup characteristics and identify key influencing factors through cluster analysis based on SHAP values.

Methods: A cohort of 974 hemodialysis patients utilizing arteriovenous fistulae was analyzed, with 55 clinical characteristics extracted for examination. Following multiple imputation, standardization, and dimensionality reduction via principal component analysis, the efficacy of K-Means, DBSCAN, and hierarchical clustering algorithms was evaluated using metrics such as the silhouette coefficient and Calinski-Harabasz index. The K-Means algorithm, with K set to 3, was chosen to develop a pseudo target variable. This was subsequently integrated with the XGBoost model, and SHAP value analysis was employed to elucidate feature contributions.

Results: The K-Means clustering algorithm demonstrated superior performance, as indicated by a Silhouette Coefficient of 0.05, effectively categorizing patients into three distinct clusters. Cluster 1 is characterized by a hemoglobin concentration range from −2 to 5, with a median of 1 and the highest variability among the clusters. Cluster 2 exhibits a hemoglobin concentration predominantly between −3 and 2, with a median of 0. Cluster 3 shows a hemoglobin concentration distribution akin to Cluster 2, albeit with slightly greater variability in the tails. SHAP analysis identified hemoglobin concentration as the most significant feature, with a SHAP value of 550, indicating that variations in its distribution are the primary drivers of the clustering process. Additionally, age, BMI, total cholesterol, and other features contribute to the clustering outcomes through complex nonlinear interactions.

Conclusion: Cluster analysis with SHAP values preliminarily identified heterogeneous subgroups in such patients, with hemoglobin concentration potentially a key driver. This approach may aid personalized treatment, but generalizability needs multicenter validation.

Keywords: hemodialysis, cluster analysis, SHAP values, unsupervised learning, personalized medicine

Introduction

The arteriovenous fistula (AVF) is the core vascular access for end-stage renal disease (ESRD) patients undergoing maintenance hemodialysis, and its functional status directly determines the adequacy of dialysis and patient survival rates.[1] Despite the advantages of AVF, such as high long-term patency and fewer complications.[2] A nationwide cohort study conducted in Japan, encompassing 183,490 patients undergoing maintenance hemodialysis, revealed that 90.7% of the participants opted for arteriovenous fistula as their dialysis modality. When compared to patients utilizing tunneled and cuffed central venous catheter and those employing arteriovenous grafts, individuals using AVF exhibited the lowest all-cause mortality rates.[3] Clinical practice shows that the failure rate within one year after surgery is still as high as 23%–46%,[4] mainly attributed to the interaction of multiple factors such as vascular calcification, thrombosis, and abnormal hemodynamics.[5] Hemoglobin concentration (HBC), as a key indicator in anemia management, its fluctuations (<100 g/L or >120 g/L) may exacerbate the risk of AVF failure by altering blood viscosity and oxidative stress responses,[6] and traditional clinical typing methods are difficult to resolve the dynamic associations of multi-dimensional characteristics, leading to insufficient identification of high-risk subgroups.[7]

Received: 6 May 2025
Accepted: 5 September 2025
Published: 13 September 2025

© 2025 Shu et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

Cluster analysis is of considerable importance in nephrological research and clinical practice, offering critical insights for the diagnosis, treatment, and management of renal diseases. Regarding disease diagnosis, cluster analysis facilitates the identification of various kidney disease subtypes and their specific characteristics. For instance, in studies involving patients with membranoproliferative glomerulonephritis, cluster analysis has been employed to discern distinct pathogenic patterns.[8,9]

However, traditional methods (such as K-means, hierarchical clustering) rely on geometric distance metrics and are unable to quantify the contribution of features and nonlinear interaction effects.[10] The study "Selective inference for k-means clustering" mainly tests mean differences between clusters in k-means clustering, without analyzing factor associations. This suggests that k-means clustering emphasizes data grouping over exploring causal relationships and associative information between factors,[11] which limits its clinical translational value. Moreover, the lack of interpretability of clustering results (the "black box" problem) is a core barrier to the application of precision medicine.[12]

SHAP values (SHapley Additive exPlanations) quantify the contribution of features to model outputs through a game-theoretic framework, providing a new paradigm for enhancing the interpretability of machine learning.[13] In medical research, SHAP and unsupervised learning methods have proven valuable. An academic investigation into the prediction of dengue fever severity employed machine learning techniques utilizing datasets from Vietnam and Bangladesh. In the analysis of the Vietnamese dataset, researchers integrated supervised learning methodologies with SHAP (SHapley Additive exPlanations) to evaluate the significance of various attributes. Conversely, for the Bangladeshi dataset, hierarchical clustering was applied to discern critical blood components associated with Dengue Hemorrhagic Fever and Dengue Shock Syndrome.[14] In a study on severe liver cirrhosis, researchers used the MIMIC-IV database and unsupervised learning methods like consensus k-means, k-means, and Self-Organizing Maps to identify clinical subtypes. They then applied the SHAP method to analyze each subtype's characteristics, offering valuable insights for clinical treatment,[15] but its application in unsupervised clustering is still in the exploratory stage. This study innovatively employed cluster analysis to construct a "pseudo-target variable", subsequently modeled this pseudo-variable using XGBoost, and visualized it through SHAP analysis. This approach revealed the characteristics of subgroups among hemodialysis patients, thereby providing new perspectives and methods for personalized medicine and precision medicine.

Methods

Data Sources and Preprocessing

This study is a retrospective cohort study, including 974 hemodialysis patients using AVF from the Blood Purification Center of The Central Hospital of Wuhan from January 2017 to March 2024. The data includes 55 clinical characteristics (Supplementary Table 1) containing the specific names of the 55 features and their abbreviations used in the statistical analysis). We collected all the blood test results of the patients before surgery through the electronic medical record system.

Inclusion and Exclusion Criteria

Inclusion Criteria: (1) Dialysis patients aged ≥18 years; (2) Patients undergoing dialysis using AVF;(3) Patients who have been on dialysis for ≥1 month. (4) Patients who underwent surgery with a radial artery-cephalic vein configuration.

Exclusion Criteria: (1) Patients undergoing dialysis using arteriovenous graft; (2) Patients with missing data ≥30%.

The features included in the study are presented in Study protocol was approved by the Ethics Committee of The Central Hospital of Wuhan (approval number: WHZXKYL2024-115), and patient data were anonymized, exempting patient informed consent. Our study fully complied with the Declaration of Helsinki.

Preprocessing procedures: (1) Missing value treatment: Categorical variables were filled with the mode; continuous variables were imputed using multiple imputation by chained equations (MICE), generating 5 complete datasets and merged through Rubin's rules.[16] We have calculated the missing values for each feature (see Supplementary Table 2 for details).

(2) Feature Standardization: Perform Z-score standardization on continuous variables, formula: $z = \frac{-\mu}{\sigma}$ χ represents the raw data, μ represents the mean of the features, σ represents the standard deviation. Standardization processing eliminates the dimensional differences among different features, preventing their impact on cluster analysis.

(3) Categorical Variable Encoding: Categorical variables such as gender, underlying diseases, and education level were one-hot encoded to generate binary dummy variables.

Feature Dimensionality Reduction and Clustering Modeling

(1) Principal Component Analysis (PCA): The number of principal components was determined by the Kaiser criterion (eigenvalue > 1) and the inflection point method of the scree plot, retaining principal components with a cumulative variance contribution rate of ≥95%.
(2) Unsupervised clustering algorithms: Three classic clustering algorithms were selected for comparative analysis:

1) K-Means clustering: The range of cluster numbers K was set from 2 to 8, and the optimal number of clusters was determined by the Elbow method. The optimal number of clusters was finally chosen as K=3.
2) DBSCAN clustering: The parameter ϵ was determined by the k-distance graph and was set to 0.3; MinPts was set to 10.
3) Hierarchical clustering: The Ward's minimum variance method was used, the distance metric was set to Euclidean distance, and the number of clusters is set to 3.

(3) Model evaluation system comprehensively evaluates the quality of clustering through the following three indicators: 1. Silhouette Coefficient: Assesses the tightness within clusters and the separation between clusters, with a value range of [−1,1]. 2. Calinski-Harabasz index: A variance ratio statistical measure based on the ratio of between-cluster/dis within-cluster dispersion. 3. Davies-Bouldin index: Minimizes the similarity between clusters, with a smaller value indicating better clustering results. Grid search (Grid Search) was used to optimize hyperparameters, and 10-fold cross-validation was used to ensure the stability of the results.
(4) Model selection: The best model was selected by comparing the Silhouette Coefficient, Calinski-Harabasz index, and Davies-Bouldin index of the three models.

Constructing Interpretable SHAP Values

(1) Pseudo target variable construction: The optimal clustering results were encoded as categorical variables (cluster 1 to cluster K), serving as pseudo target variables for training the XGBoost classifier. Model parameters are set as follows: learningrate0.1, maximum depth 10, number of trees 200, and regularization term λ= 1. The Tree Explainer of SHAP is used for model interpretation.
(2) Feature contribution analysis: The contribution strength of each feature to cluster membership was calculated through SHAP values. SHAP values were based on the Shapley value theory and could quantify the contribution of each feature to the model output. Feature contributions were represented through SHAP bee swarm plots, feature importance ranking plots, partial dependence plots, etc.
(3) Feature correlation analysis: A bubble heatmap was drawn to display the correlation between variables. Positive and negative correlations were distinguished by bubble size and color, further analyzing the relationships between features.

Statistical Analysis

Data processing was conducted using Python (version: Python 3.13.2). Inter-group difference tests: For continuous variables: when they conform to a normal distribution (Shapiro–Wilk test $p > 0.05$) and have equal variances (Levene test $p > 0.05$), a one-way ANOVA was used; otherwise, the Kruskal–Wallis *H*-test was applied; for categorical variables: chi-squared test or Fisher's exact test was used (when expected frequency is <5).

Results

Data Preprocessing and Feature Processing Results

A total of 974 patients were included in this study, and the data contained 55 features. During the data preprocessing phase, the mode imputation was used for categorical variables, and multiple imputation by chained equations (MICE) was used to handle missing values for continuous variables. After data standardization, PCA was employed to reduce the number of features from 55 dimensions to 36 dimensions, retaining 95% of the variance information (Figure 1).

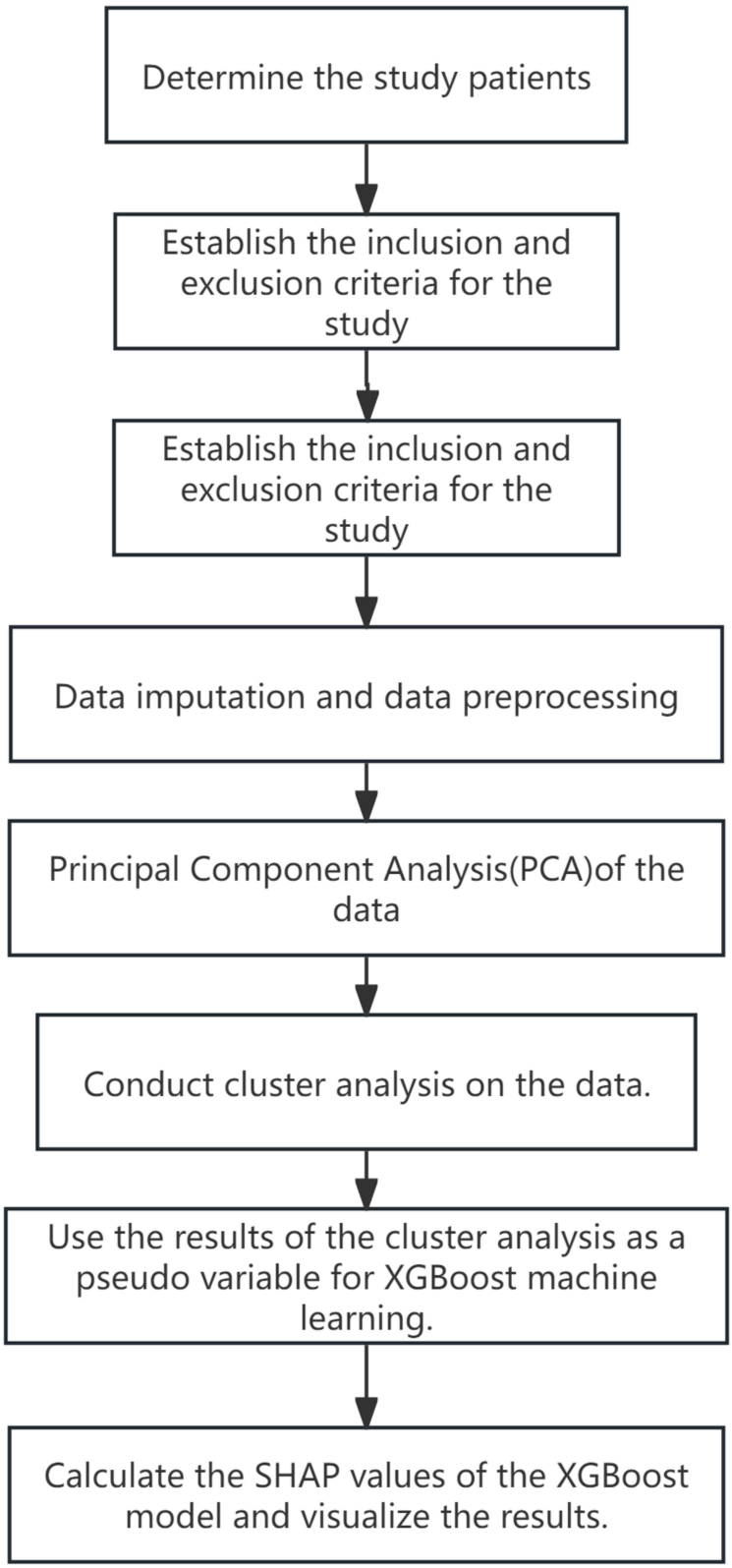

Figure 1 Flowchart of the study.

By observing the bubbles corresponding to its row and column, it can be seen that there is a certain correlation with some other variables. The larger blue bubbles indicate a strong positive correlation with these variables; the presence of red bubbles indicates a negative correlation with some variables, but the strength of the correlation can be further judged

by the size of the bubbles. BMI (Body Mass Index): Similarly, by examining the bubbles corresponding to its row and column, the correlation with other variables can be discovered. Some larger blue bubbles mean a strong positive correlation with some variables, while some red bubbles show a negative correlation with other variables (Figure 2).

Unsupervised Learning Model Comparison

Model Selection and Evaluation

This study compared three unsupervised learning models (K-Means clustering, DBSCAN clustering, and hierarchical clustering). The silhouette index for K-Means was 0.05, the Calinski-Harabasz index was 50.58,and the Davies-Bouldin index was 3.35; the DBSCAN model performed poorly, with all indices being 0; the silhouette index for hierarchical clustering was 0.01, the Calinski-Harabasz index was 32.75, and the Davies-Bouldin index was 3.75, ultimately selecting K-Means clustering as the best model (Figure 3).

Clustering Outcomes

Clustering outcomes indicate that the K-Means algorithm categorized the patients into three distinct clusters. The distribution of samples across these clusters is as follows: Cluster 1 comprises 281 samples, accounting for 28.8% of

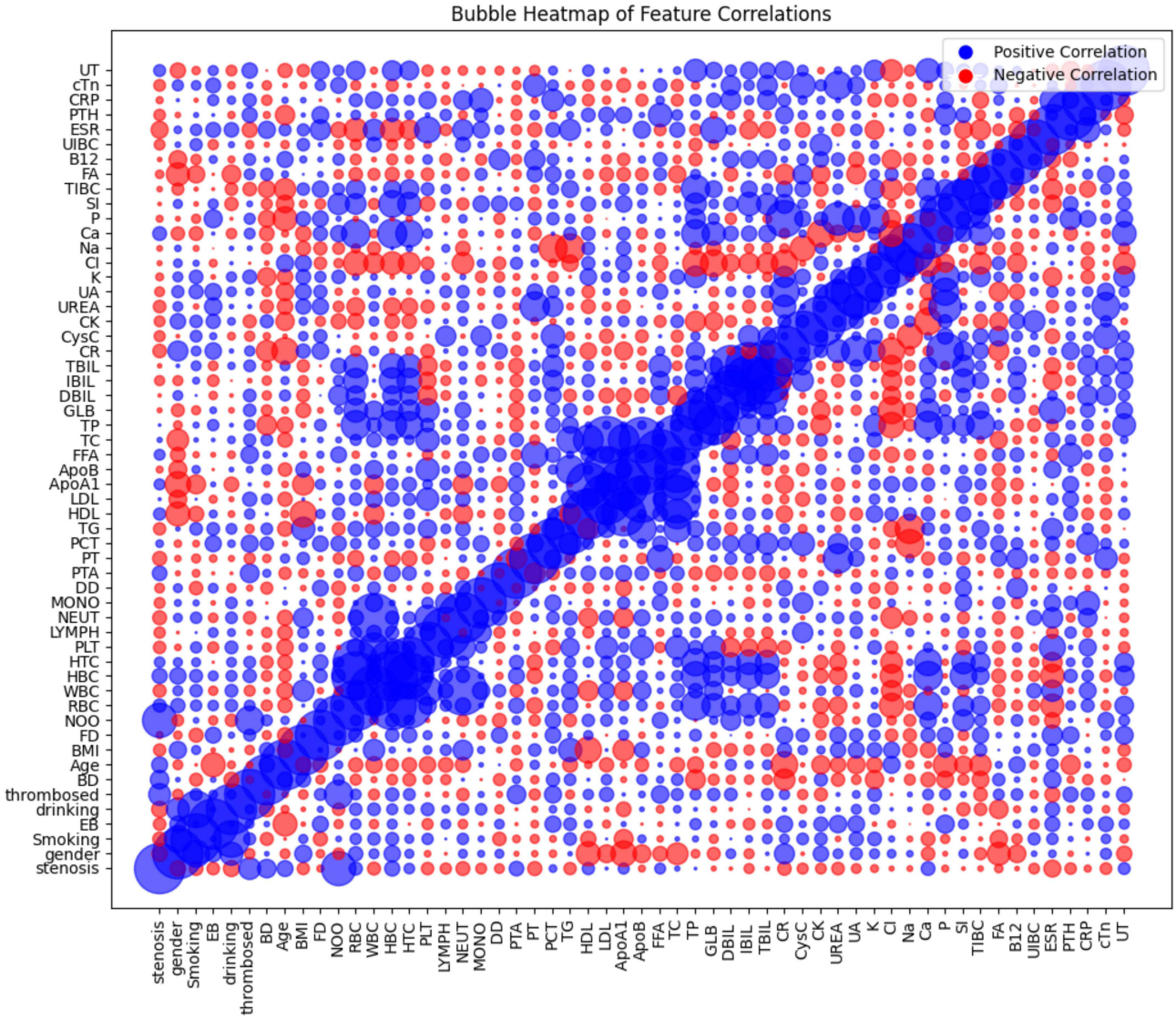

Figure 2 Bubble heatmap of all features.

https://doi.org/10.2147/IJGM.S533419

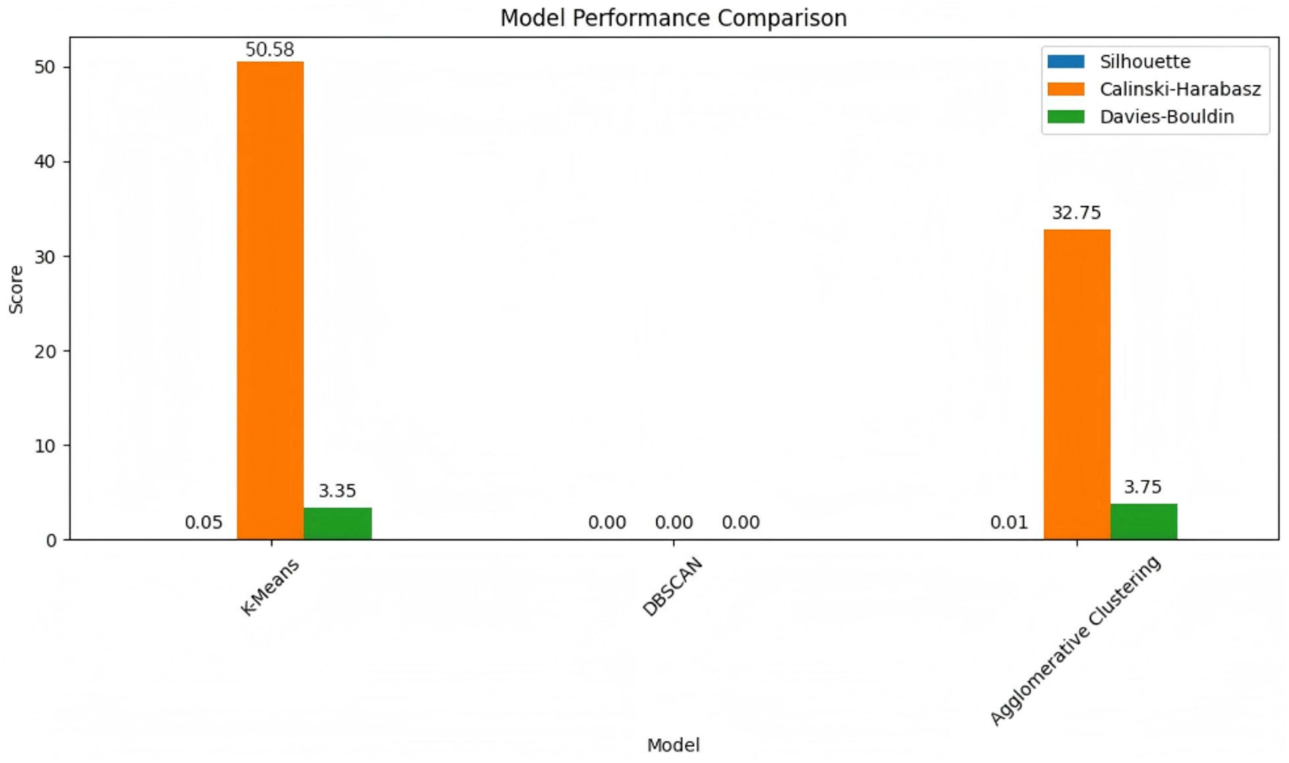

Figure 3 Comparison of Various Indicators for Three Models.

the total; Cluster 2 includes 431 samples, representing 44.2%; and Cluster 3 consists of 262 samples, which constitute 27.0% of the dataset, as illustrated in Figure 4.

Feature Importance Analysis

SHAP Value Analysis

Utilizing the outcomes of K-Means clustering, a pseudo-target variable was developed, which subsequently informed the training of an XGBoost model. The influence of each feature on the clustering results was then examined through SHAP value analysis. The results showed that HBC was the most important feature, with the most significant contribution to the clustering results.(Figure 5) The hemoglobin concentrations among patients in the three clusters exhibit distinct characteristics. In Cluster 1, the hemoglobin concentration distribution spans a broad range from −2 to 5, with a median value of 1 and a density peak near 0. This suggests considerable heterogeneity in HBC within this cluster, exemplified by the coexistence of anemia and normal levels. In Cluster 2, the distribution is more concentrated, ranging from −3 to 2, with a median of 0 and the highest density at 0, indicative of a stable HBC group. Cluster 3 shares a similar distribution range with Cluster 2 (−3 to 2) but demonstrates slightly greater variability in the tails, potentially reflecting compensatory physiological fluctuations (Figure 6).

Comparison of Feature Importance Across Clusters

Figure 7 displays a comparison chart of the SHAP values for the top 5 features in each cluster, with the SHAP value of hemoglobin in cluster 1 being significantly higher than in other clusters.

Analysis of Feature Distribution in Clusters

Distribution and Mean of Various Features Across Different Clusters

Table 1 shows significant differences in indicators such as age, red blood cell count, white blood cell count, hemoglobin, and thrombosis incidence across the three clusters ($p<0.05$).

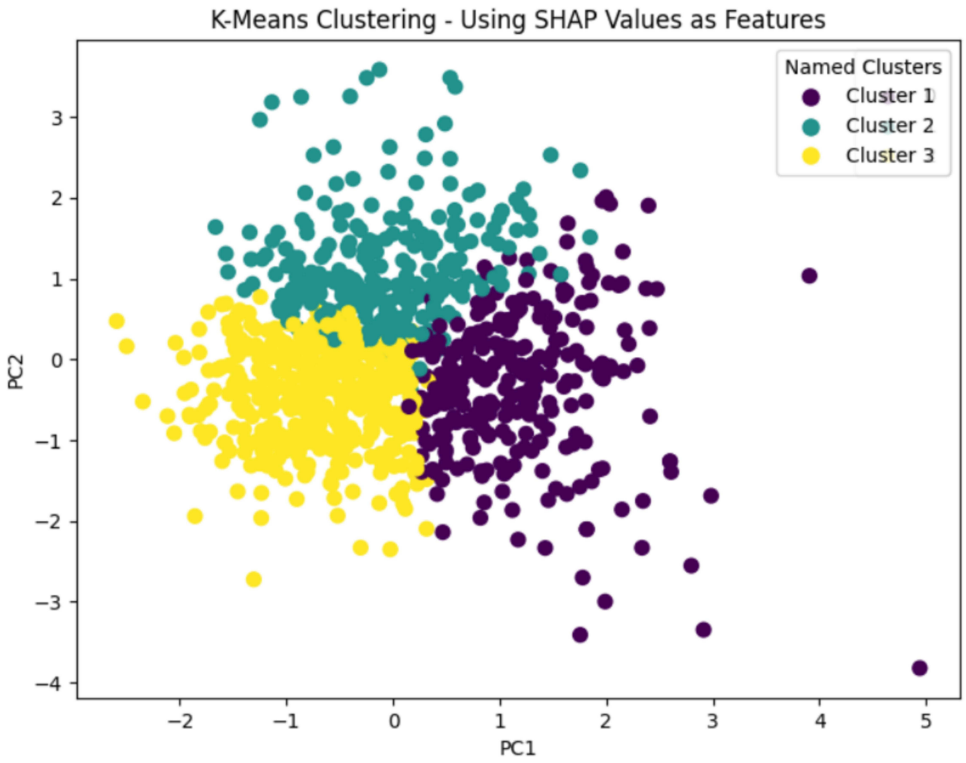

Figure 4 K-Means Clustering Diagram.

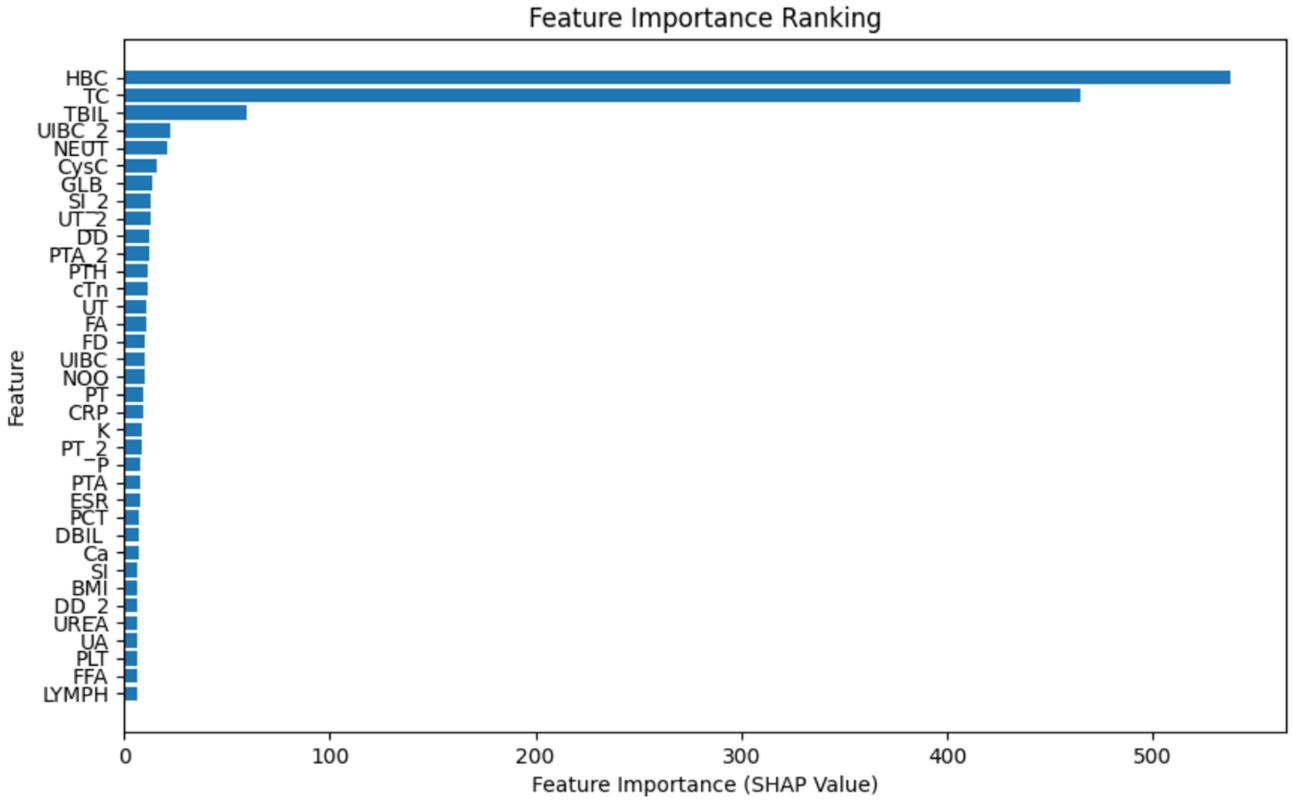

Figure 5 Feature Importance Ranking Chart.

https://doi.org/10.2147/IJGM.S533419

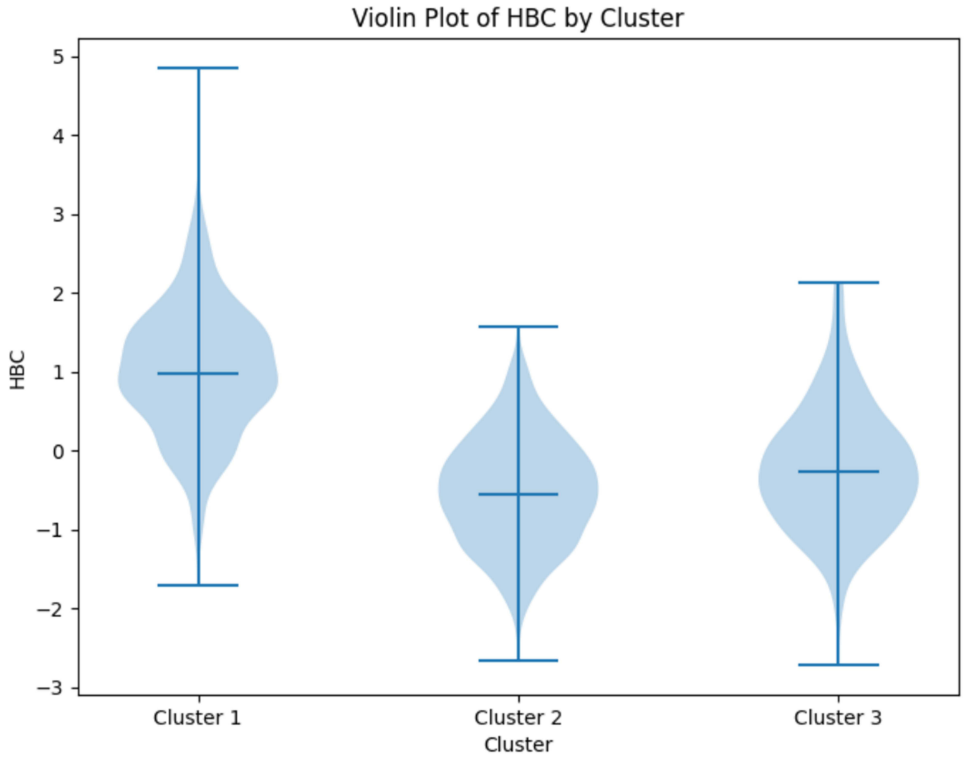

Figure 6 Violin plot of the most important feature HBC affecting clustering.

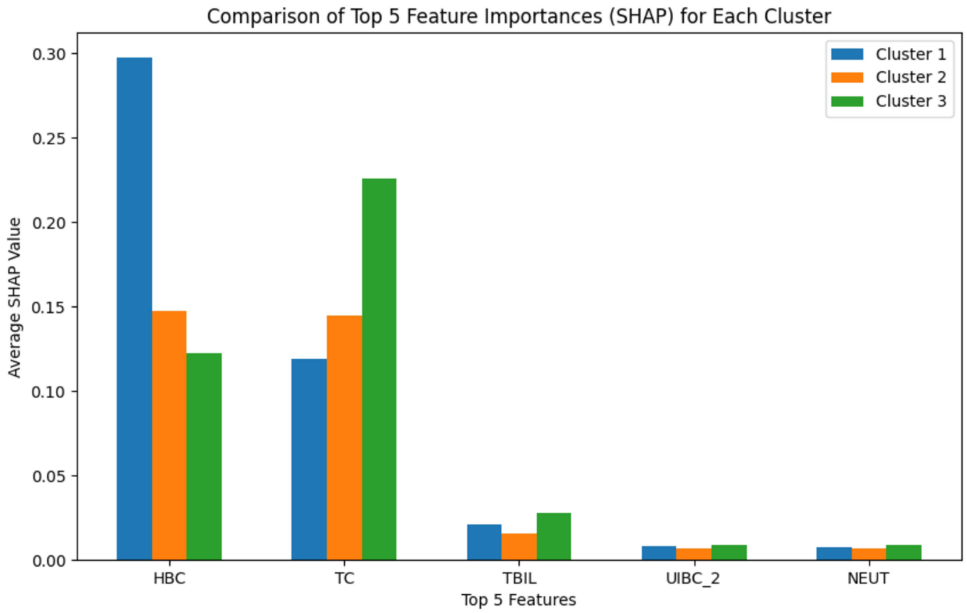

Figure 7 Comparison of SHAP values for the first 5 features.

SHAP Values of Different Features on Three Clusters

Figure 8 illustrates the top ten features affecting the model output in Cluster 1, including Age, Body Mass Index, Red Blood Cell Count, Apolipoprotein A1, Fistula Diameter, Uric Acid, White Blood Cell Count, Low-Density Lipoprotein Cholesterol, Indirect Bilirubin, and Hematocrit. In Cluster 2, the top ten features affecting the model output are Age, Body Mass Index, Red Blood Cell Count, Apolipoprotein A1, Fistula Diameter, Hematocrit, Uric Acid, Total Protein, White Blood Cell Count, and Neutrophil Count. High age feature values are distributed on the right side of the vertical

Table 1 Distribution of Different Features in Different Clusters

Variable	Cluster 1(mean ± SD) (n=281)	Cluster 2 (mean ± SD) (n=431)	Cluster 3 (mean ± SD) (n=262)	p - values
Age(years)	58.0854±13.2568	62.1601±12.3315	59.9695 ± 13.4084	<0.001*
BMI (kg/m^2)	23.4882 ± 3.8249	23.4093 ± 3.7666	23.9741 ± 4.1825	0.161
FD (mm)	0.4589 ± 0.2523	0.4175 ± 0.1960	0.4177 ± 0.1685	0.019*
NOO (times)	1.2242 ± 1.3665	0.8167 ± 0.9795	0.8397 ± 1.0779	<0.001*
RBC (10^12/L)	3.8616 ± 0.6323	2.7437 ± 0.5568	3.0123 ± 0.6197	<0.001*
WBC (10^9/L)	6.6346 ± 2.4110	6.0038 ± 2.1954	7.2084 ± 2.8730	<0.001*
HBC (g/L)	113.7616 ± 17.8863	80.6381 ± 15.0094	86.9809 ± 16.9799	<0.001*
HTC (L/L)	35.8512 ± 7.7423	24.8545 ± 4.7186	26.8768 ± 5.3439	<0.001*
PLT (10^9/L)	185.2754 ± 73.4862	167.0464 ± 58.5609	218.5763±80.4224	<0.001*
LYMPH (10^9/L)	1.2840 ± 1.0572	1.1995 ± 1.7940	1.2164 ± 0.5441	0.708
NEUT (10^9/L)	4.7421 ± 2.2162	4.3131 ± 1.8687	5.2803 ± 2.5784	<0.001*
MONO (10^9/L)	0.5574 ± 1.3786	0.4616 ± 0.6479	0.5502 ± 0.7219	0.309
DD (mg/L FEU)	2.2823 ± 6.9175	2.1866 ± 3.0809	2.0657 ± 3.1480	0.857
PTA (%)	96.5309 ± 20.9108	93.4296 ± 21.3266	101.7747 ± 60.6632	0.014*
PT (seconds)	11.8931 ± 4.6808	12.4398 ± 6.8684	12.0009 ± 6.8607	0.469
PCT (ng/mL)	2.2755 ± 4.7421	1.2334 ± 2.9873	1.3502 ± 2.9862	<0.001*
TG (mmol/L)	1.7845 ± 1.6143	1.1966 ± 0.6993	1.9585 ± 1.3676	<0.001*
HDL (mmol/L)	1.0606 ± 0.3580	1.0032 ± 0.3533	1.1934 ± 0.3790	<0.001*
LDL(mmol/L)	2.2080 ± 0.7516	1.8503 ± 0.5745	3.2602 ± 0.8543	<0.001*
ApoA1(g/L)	1.1343 ± 0.2683	1.0819 ± 0.2464	1.2495 ± 0.2363	<0.001*
ApoB (g/L)	0.8114 ± 0.2203	0.7022 ± 0.1809	1.1088 ± 0.2171	<0.001*
FFA (mmol/L)	0.4214 ± 0.2877	0.3784 ± 0.3158	0.4013 ± 0.2603	0.155
TC (mmol/L)	3.8989 ± 0.9186	3.3842 ± 0.6939	5.3082 ± 1.0277	<0.001*
TP (g/L)	72.8192 ± 7.7272	62.0845 ± 8.9127	65.4515 ± 8.5188	<0.001*
GLB (g/L)	32.5000 ± 5.8176	27.9267 ± 5.2765	30.1599 ± 5.1477	<0.001*
DBIL(umol/L)	3.2822 ± 4.4933	1.6966 ± 1.0494	1.5298 ± 1.1762	<0.001*
IBIL (umol/L)	5.4071 ± 3.3370	3.3617 ± 1.6259	3.5065 ± 1.6814	<0.001*
TBIL (umol/L)	8.4673 ± 5.7023	5.1016 ± 2.2877	5.0934 ± 2.3006	<0.001*
CR (umol/L)	784.6092 ± 350.8913	686.4675 ± 317.6838	653.5050 ± 327.7481	<0.001*
CysC (mg/L)	6.8259 ± 5.7740	7.1547 ± 8.8457	5.7958 ± 2.2981	0.035*
CK (U/L)	114.2399 ± 115.6781	195.8708 ± 311.9338	204.2836 ± 278.9624	<0.001*
UREA (mmol/L)	20.5563 ± 8.3011	22.6124 ± 19.3169	20.5930 ± 18.8320	0.169
UA (umol/L)	393.1093 ± 126.9174	366.9332 ± 130.3121	379.5141 ± 138.2330	0.034*
K (mmol/L)	4.9006 ± 0.7809	4.6235 ± 0.7582	4.7969 ± 0.8630	<0.001*
Cl (mmol/L)	100.7815 ± 5.2134	105.1120 ± 5.4142	104.1473 ± 5.8087	<0.001*
Na (mmol/L)	138.9032 ± 7.0005	140.1912 ± 6.2397	139.8859 ± 3.6750	0.016*
Ca (mmol/L)	2.3583 ± 0.2359	2.0974 ± 0.2548	2.1400 ± 0.2535	<0.001*
P (mmol/L)	1.7339 ± 0.5197	1.6097 ± 0.5386	1.6002 ± 0.4821	0.002*
SI (umol/L)	14.5959 ± 7.0628	9.9669 ± 5.2166	10.3719 ± 5.4029	<0.001*
TIBC (umol/L)	52.0823 ± 11.5923	45.5860 ± 11.5759	46.7577 ± 11.8115	<0.001*
FA (nmol/L)	22.6684 ± 15.6984	24.0456 ± 16.4519	18.8877 ± 12.6294	<0.001*
B12 (pmol/L)	423.4445 ± 286.1892	428.3123 ± 303.4615	405.8164 ± 288.6191	0.613
UIBC (umol/L)	36.4033 ± 13.3683	57.3567 ± 252.3524	36.1000 ± 12.3777	0.152
ESR (mm/h)	52.5676 ± 39.8149	57.4944 ± 38.6590	75.8080 ± 40.1619	<0.001*
PTH (pg/mL)	133.8158 ± 220.2001	133.3974 ± 175.7682	116.7547 ± 182.0602	0.477
CRP (mm/h)	11.1086 ± 29.6032	5.5903 ± 18.0378	5.9542 ± 16.1798	0.002*
cTn (ng/mL)	0.3129 ± 2.3396	0.4348 ± 3.4832	0.1015 ± 0.3913	0.275
UT (months)	59.3559 ± 50.2183	35.7146 ± 28.0149	39.9160 ± 28.9076	<0.001*

(*Continued*)

Table 1 (Continued).

Variable	Cluster 1(mean ± SD) (n=281)	Cluster 2 (mean ± SD) (n=431)	Cluster 3 (mean ± SD) (n=262)	p - values
Stenosis				0.334
No	58.72%	63.11%	64.50%	
Yes	41.28%	36.89%	35.50%	
Gender				0.005*
Male	69.75%	71.00%	59.54%	
Female	30.25%	29.00%	40.46%	
Smoking				0.925
No	77.94%	79.12%	79.01%	
Ys	22.06%	20.88%	20.99%	
EB				0.063
Junior and Senior High School	72.60%	70.77%	63.36%	
Primary and below	18.15%	21.81%	28.24%	
Undergraduate and above	9.25%	7.43%	8.40%	
Drinking				0.388
No	90.04%	92.58%	90.08%	
Yes	9.96%	7.43%	9.92%	
Thrombosed				0.012*
No	82.21%	89.56%	88.55%	
Yes	17.79%	10.44%	11.45%	
BD				0.079
1	35.94%	23.90%	24.43%	
2	2.85%	2.55%	2.67%	
3	0.36%	0.93%	0.38%	
4	1.42%	2.09%	0.76%	
5	58.36%	69.37%	69.08%	
6	0.36%	0.00%	0.38%	
7	0.36%	0.23%	0.38%	
8	0.36%	0.93%	1.91%	

Note: 1. Hypertension; 2. Diabetes Mellitus; 3. Polycystic Kidney Disease; 4. Chronic Glomerulonephritis; 5. With 2 or more of the above; 6. Coronary Heart Disease; 7. Vasculitishe; 8. Others. The symbol (*) indicates a statistically significant difference of the feature among the three groups.

Abbreviations: EB, Educational Background; FD, Fistula Diameter; NOO, Number of Operations; BD, Underlying diseases; HCT, Hematocrit; LYMPH, Lymphocyte Count; MONO, Monocyte Count; NEUT, Neutrophil Count; DD, D-dimer; PTA, Prothrombin Time Activity; PT, Prothrombin Time; TG, triglycerides; HDL, high-density lipoprotein; LDL, low-density lipoprotein; ApoA1, apolipoprotein A1; Apo B, apolipoprotein B; FFA, free fatty acids; TC, total cholesterol; TP, total protein; GLB, globulin; DBIL, direct bilirubin; IBIL, indirect bilirubin; TBIL, total bilirubin; CR, creatinine; CysC, cystatin C; CK, creatine kinase; UREA, urea; UA, uric acid; SI, serum iron; TIBC, total iron-binding capacity; FA, folic acid; B12, vitamin B_{12}; UIBC, unsaturated iron-binding capacity; ESR, erythrocyte sedimentation rate; PTH, parathyroid hormone; cTn, serum troponin; UT, use time of AVF; BD, Underlying Diseases.

axis. In Cluster 3, the top ten influencing features are Body Mass Index, Age, Red Blood Cell Count, Fistula Diameter, Apolipoprotein A1, Hematocrit, Indirect Bilirubin, White Blood Cell Count, Total Protein, and Neutrophil Count. Overall, in Clusters 1 and 2, age is the primary feature affecting the model output, and the direction of age's impact on the model output is relatively consistent, but the degree and distribution differ; in Cluster 3, Body Mass Index becomes the key feature affecting the model output, and its impact pattern on the model output is similar to the impact of age in the other two clusters, yet it has its own characteristics, reflecting the intrinsic feature differences among different clusters. Although HBC is not directly listed as a significant factor in the dataset variable significance analysis, it indirectly dominates the model output through nonlinear interactions (such as synergy with Age and BMI).

The Top 5 Important Features for the Three Clustering Clusters' Model Output Results

This study analyzed the impact of five features (Hemoglobin Concentration, Total Cholesterol, Total Bilirubin, Unsaturated Iron Binding Capacity, Neutrophil Count) on the predictive probabilities of three different clusters using Partial Dependence Plots (PDP). In Cluster 1, Hemoglobin Concentration exerts a significant positive influence on the prediction outcomes,

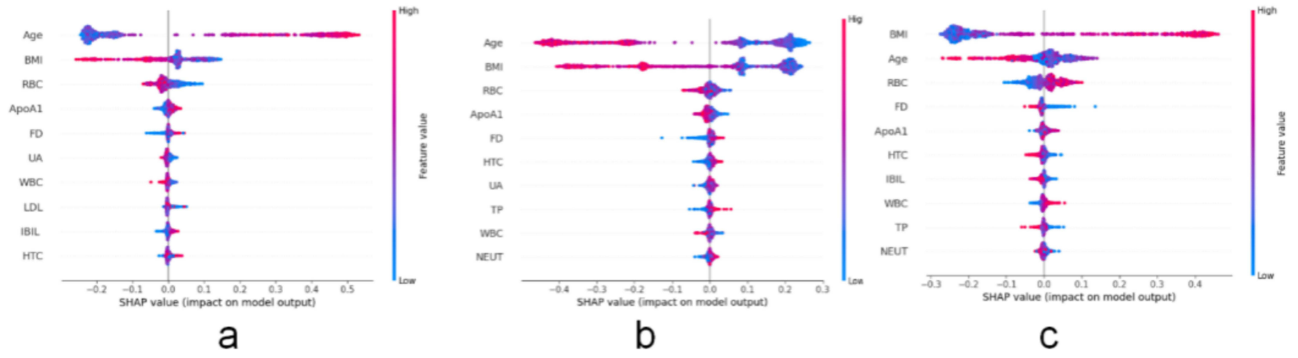

Figure 8 A ranking chart of the top ten SHAP values in the three clustering clusters.
Notes: (**a**) Ranking plot of the top 10 feature SHAP values for Cluster 1; (**b**) Ranking plot of the top 10 feature SHAP values for Cluster 2; (**c**) Ranking plot of the top 10 feature SHAP values for Cluster 3.

whereas its effect is diminished in Clusters 2 and 3. This suggests that HBC may serve as a critical distinguishing feature for Cluster 1 relative to the other clusters. Regarding Total Cholesterol and Total Bilirubin, Total Cholesterol exhibits a nonlinear influence in Cluster 1, with a reduced effect in Clusters 2 and 3. Conversely, Total Bilirubin demonstrates a negative impact in Cluster 2 and a positive impact in Cluster 3, indicating that Total Bilirubin may be associated with varying health conditions across different clusters. As for Unsaturated Iron Binding Capacity and Neutrophil Count, Unsaturated Iron Binding Capacity shows a nonlinear effect in Cluster 1, a negative impact in Cluster 2, and a minimal effect in Cluster 3. Neutrophil Count, on the other hand, has a positive impact in Cluster 2 and a negative impact in Cluster 3 (Figure 9).

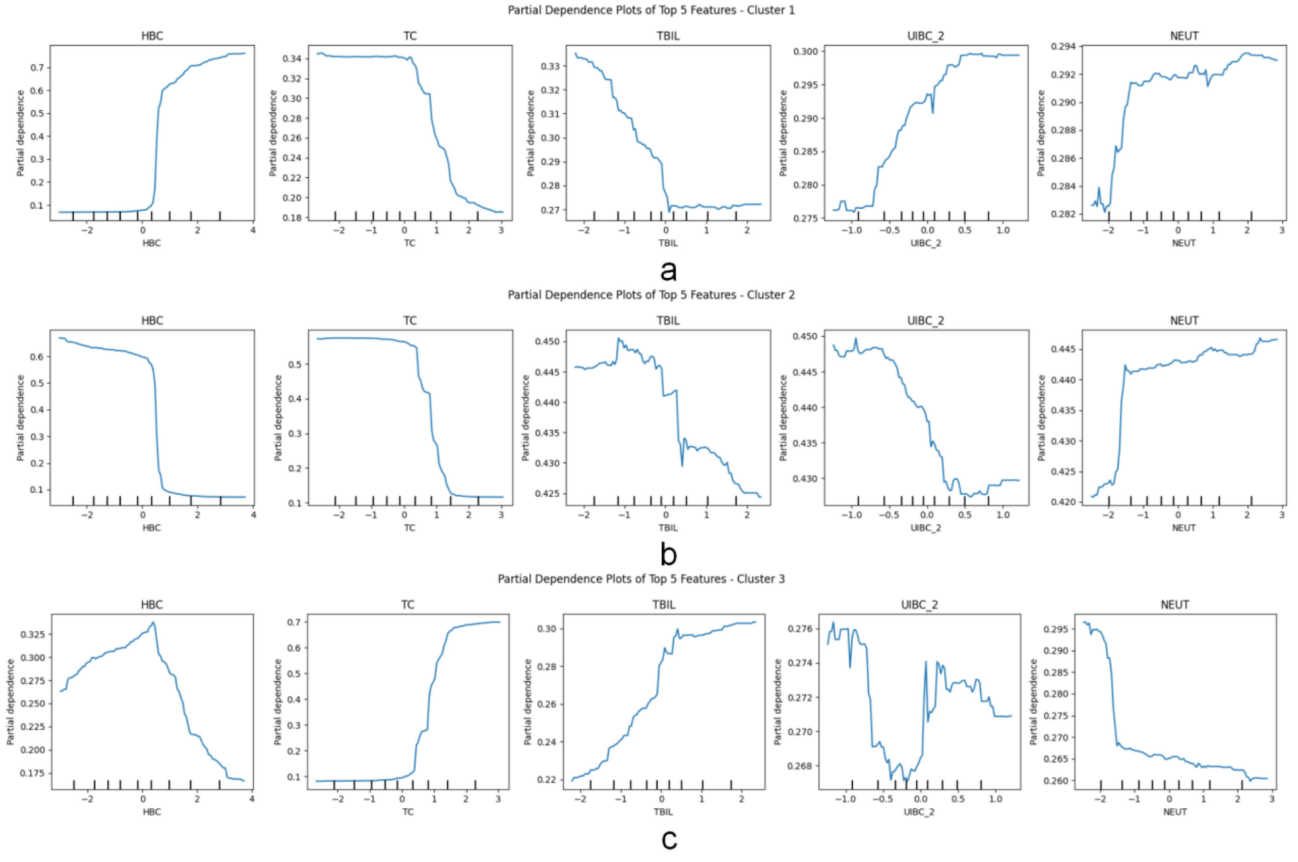

Figure 9 Partial dependence plots of the top 5 important features for the three clustering clusters.
Notes: (**a**) Partial dependence plots of the top 5 important features in Cluster 1; (**b**) Partial dependence plots of the top 5 important features in Cluster 2; (**c**) Partial dependence plots of the top 5 important features in Cluster 3.

https://doi.org/10.2147/IJGM.S533419

Discussion

Principal Component Analysis

k-means clustering algorithm combined with SHAP values have certain advantages in our project. We selected the K-Means algorithm as the primary clustering method and combined it with PCA and SHAP values to enhance the model's interpretability and accuracy. The K-Means algorithm is a commonly used clustering technique that can effectively divide data into different groups. In our research, by applying PCA to the data for dimensionality reduction, we can reduce the complexity of the data while retaining the main characteristics, thereby improving the effectiveness of K-Means clustering.[17,18] In traditional clustering methods, geometric distance is usually relied upon to measure the proximity between data points, which may lead to problems when dealing with clustering tasks that have arbitrary shapes and different densities.[19] However, clustering methods based on SHAP values can better explain and differentiate data features from different omics by introducing clinical information during the feature extraction process, thereby improving the performance of clustering analysis.[20] Specifically, traditional multi-omics clustering methods may struggle with the noise or redundancy of multi-omics data when directly integrating heterogeneous features from different omics, leading to poor clustering results. Clustering methods based on SHAP values significantly enhance the clustering effect by extracting interpretable and discriminative features before data integration.[15] Moreover, experimental results show that feature extraction based on SHAP values can enhance the performance of clustering analysis and further demonstrate the effectiveness of this method by conducting enrichment analysis on gene features identified in different subtypes.[20] In summary, compared with traditional clustering methods, clustering methods based on SHAP values show higher accuracy and interpretability when dealing with complex datasets, especially in the field of cancer sub type identification, demonstrating their superiority in multi-omics data analysis.[10]

The Clinical Significance of Hemoglobin as a Key Indicator for Dialysis Patients

This study, through SHAP-driven cluster analysis, confirmed that hemoglobin was the core driver of heterogeneity typing in hemodialysis patients. The results have showed a significant difference in HBC levels between Cluster 1 and Cluster 2/3 patients (113.76±17.89 vs 80.64±15.01 g/L vs 86.9809 ± 16.98, $p<0.001$), and this heterogeneity might stem from differences in clinical practice for anemia management. Previous studies have shown that fluctuations in hemoglobin concentration are significantly associated with patient mortality. For instance, a nationwide cohort study in South Korea found that variability in hemoglobin is an important predictor of mortality, a finding that applies not only to patients with chronic kidney disease and cardiovascular disease but also to the general population.[21]

Despite current clinical practices significantly improving anemia control rates through the combined treatment of iron and erythropoietin (EPO), the average HBC values in Cluster 2 and 3 in this study still present a state of refractory anemia (HBC<100 g/L). Studies indicate that variability in hemoglobin levels not only affects patient survival rates but may also be related to other health indicators such as muscle mass. Anemia reduces blood's oxygen capacity, potentially causing cardiovascular issues like myocardial ischemia and heart failure, which are major death causes in hemodialysis patients.[22] Muscle mass is an independent predictor of erythropoietin response, suggesting that muscle mass is also a factor to be considered when managing anemia in dialysis patients, which is similar to our study results.[23] After visualizing SHAP values in our study, it was found that the BMI of patients in the three cluster groups significantly influenced the model output. Although HBC was not directly listed as a significant factor in the dataset variable significance analysis in Figure 7, it indirectly dominated the model output through nonlinear interactions (such as synergy with Age, BMI), thus indicating that BMI is also a relatively important factor in the anemia management of hemodialysis patients.

The Impact of Hemoglobin Concentration AVF-Related Complications

The results of this study suggest a significant association between hemoglobin concentration and the failure of AVF. Previous studies have also indicated that hemoglobin levels are one of the key factors affecting AVF failure. A cross-sectional study on arteriovenous fistula dysfunction in hemodialysis patients analyzed clinical data, vascular calcification scores, and lab results from 100 individuals. It found that patients with fistula dysfunction had significantly higher hemoglobin levels than those with normal function, reinforcing the link between hemoglobin levels and fistula dysfunction.[24] In particular, higher hemoglobin

levels (>120 g/L) may accelerate the process of AVF failure by increasing blood viscosity, which is consistent with the high failure rate (41.28%) observed in Cluster 1 of this study. The incidence of thrombosis in patients in Cluster 1 (17.8%) was significantly higher than in the other two clusters (10.4%/11.5%), with a significant difference ($p<0.05$). A study has shown that high-dose erythropoietin treatment is prone to cause thrombosis in arteriovenous fistulas,[25] similar to our research findings. The main reason may be that changes in hemoglobin levels are related to endothelial cell dysfunction, which is an important mechanism for AVF thrombosis.[26] High hemoglobin levels may promote the formation of AVF thrombi through various mechanisms, including increasing blood viscosity, damaging vascular endothelial function, and interacting with other biomarkers. Therefore, monitoring and managing hemoglobin levels in clinical practice is of great significance for preventing the occurrence of AVF thrombosis.

Cluster-Driven Precision Management Strategy for Hemoglobin

This study reveals the dynamic threshold effect of hemoglobin concentration and its cross-variable synergy with age and BMI through the SHAP interpretability framework, providing a new paradigm for personalized management of hemodialysis patients. In the SHAP value variable analysis, HBC did not show significance, but it dominated cluster typing through nonlinear interaction mechanisms. This finding complements the latest management techniques. Medical staff can develop decision-making systems to formulate treatment plans for lower hemoglobin levels. The dose optimization algorithm developed by Barbieri et al can increase the HBC target rate to 82% (vs.65% with traditional plans), while reducing erythropoietin usage by 30%;[27] or through pharmacist-led precision interventions, accurately calculate drug dosages, improve the achievement rate of target levels for hemoglobin and iron status, and reduce the incidence of high hemoglobin levels.[28]

Although HBC was not directly listed as a significant factor in the variable significance analysis of this study's data, it indirectly dominated model output through nonlinear interactions with variables such as age and body mass index. In the management of hemodialysis patients, age and BMI are two important considerations. Studies have shown that there are significant differences in the relationship between BMI and all-cause mortality among hemodialysis patients of different age groups, with lower BMI being associated with higher all-cause mortality, especially in younger patients.[29] Therefore, when formulating hemoglobin management strategies, it is necessary to consider the patient's age and BMI comprehensively to optimize treatment effects. Based on the above findings, a three-tier management pathway can be used in the future management of hemodialysis patients: (1) Risk stratification: Prioritize HBC management based on SHAP values (Clusters 2 and 3 as primary intervention targets); (2) Combine AI models with red blood cell lifespan prediction to achieve HBC fluctuation alerts; (3) Precision intervention: Adjust iron/EPO dosages based on the Age-BMI-HBC interaction matrix.

A study conducted by Shu Peng et al demonstrated that, among the 55 biochemical blood test characteristics of patients, hemoglobin concentration serves as a predictor for the risk of arteriovenous fistula stenosis, although it is not considered a primary influencing factor.[30] However, in the current study, after performing clustering analysis and feature dimensionality reduction on 55 characteristics from 974 patients, a significant difference was identified in the incidence of AVF thrombosis among patients with varying hemoglobin concentration levels ($p<0.05$). In conclusion, hemoglobin concentration emerges as a crucial factor influencing the patency of AVF, as well as the quality of life and prognosis of patients.

In summary, this study reveals the key role of monitoring hemoglobin concentration in optimizing the management of hemodialysis patients through cluster analysis, providing important guidance and scientific basis for clinical practice. Future research should further explore personalized hemoglobin management strategies to provide higher quality medical services for hemodialysis patients.

Conclusion

This study innovatively applies SHAP values to AVF dialysis patient cluster analysis, using a "pseudo target variable" strategy to combine unsupervised clustering with supervised interpretable models, overcoming traditional clustering's interpretability limitations. It identified patient subgroups, quantifies feature contributions, reveals key factors for group differences, offering new perspectives for personalized medicine and reference for related research.

Limitations of the study: Although this study includes a sample size of 974 cases, it is based on single-center data, and the model may have certain limitations. As this study is retrospective,certain data like surgical experience and comorbidities (eg,

peripheral vascular disease) are unavailable, potentially introducing bias into the results. Additionally, the interpretation of SHAP values may be affected by data quality and feature selection. Pseudo target variables may introduce clustering bias, which needs to be validated through external cohorts. Future studies need to verify the findings of this study in larger sample sizes and broader patient populations, and further explore the potential application of SHAP values in other medical fields.

In summary, this study successfully applies SHAP values to the cluster analysis of patients undergoing arteriovenous fistula dialysis, providing new tools and ideas for personalized medicine and precision treatment. As machine learning and artificial intelligence technologies continue to develop in the medical field, we look forward to this method playing a greater role in future clinical practice.

Data Sharing Statement

All data generated during the study, due to hospital policy requirements and patient privacy concerns, can be obtained by contacting the corresponding author and the first author if needed.

Ethical Statement

This study received approval from the Ethics Committee of The Central Hospital of Wuhan, under approval number WHZXKYL2024-115. Given the retrospective nature of the study, the Ethics Committee granted an exemption from obtaining informed consent.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was funded by the Chen Xiaoping Foundation for the development of SCIENCE and Technology of Hubei province, No.CXPJJH124001-2404.

Disclosure

All authors declare no conflicts of interest.

References

1. Hill K, Xu Q, Jaensch A, et al. Outcomes of arteriovenous fistulae cannulation in the first 6 weeks of use: a retrospective multicenter observational study. *J Vasc Access*. 2021;22(5):726–732. doi:10.1177/1129729820954717
2. Zhao J, Jourd'heuil FL, Xue M, et al. Dual function for mature vascular smooth muscle cells during arteriovenous fistula remodeling. *J Am Heart Assoc*. 2017;6(4):e004891. doi:10.1161/JAHA.116.004891
3. Murakami M, Fujii N, Kanda E, et al. Association of four types of vascular access including arterial superficialization with mortality in maintenance hemodialysis patients: a nationwide cohort study in Japan. *Am J Nephrol*. 2023;54(3–4):83–94. doi:10.1159/0005299913
4. Disease K. Improving Global Outcomes (KDIGO).KDIGO clinical practice guideline for anemia in chronic kidney disease. *Kidney Int Suppl*. 2012;2(4):279–335. doi:10.1038/kisup.2012.37
5. Waldrop TI, Graham C, Gard W, et al. Biomimetic cardiac tissue chip and murine arteriovenous fistula models for recapitulating clinically relevant cardiac remodeling under volume overload conditions. *Front Bioeng Biotechnol*. 2023;16(11):1101622. PMID:36873372;PMCID:PMC9978753. doi:10.3389/fbioe.2023.1101622
6. Purushothaman M, Krishnan P, KR P, et al. Genotype-dependent impairment of hemoglobin clearance increases oxidative and inflammatory response in human diabetic atherosclerosis. *Arterioscler Thromb Vasc Biol*. 2012;32(11):2769–2775. PMID: 22982461. doi:10.1161/ATVBAHA.112.252122
7. Zhang J, Xue Y, Liu X, et al. Identification of 4 subgroups in juvenile dermatomyositis by principal component analysis-based cluster analysis. *Clin Exp Rheumatol*. 2022;40(2):443–449. doi:10.55563/clinexprheumatol/t2hxjd
8. Elisabeth Stømer U, Klopstad Wahl A, Gunnar Gøransson L, Hjorthaug Urstad K. Health literacy in kidney disease: associations with quality of life and adherence. *J Ren Care*. 2020;46(2):85–94. PMID: 31950601. doi:10.1111/jorc.12314
9. Garam N, Prohászka Z, Szilágyi Á, et al. Validation of distinct pathogenic patterns in a cohort of membranoproliferative glomerulonephritis patients by cluster analysis. *Clin Kidney J*. 2019;13(2):225–234. PMID: 32296528; PMCID: PMC7147314. doi:10.1093/ckj/sfz073
10. Al-Daoud M, Al-Jarrah O, Al-Ayyoub M, et al. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. *Inf Sci*. 2023;571:510–534. doi:10.1016/j.ins.2023.01

11. Chen YT, Witten DM. Selective inference for k-means clustering. *J Mach Learn Res*. 2023;24:152. PMID: 38264325; PMCID: PMC10805457. doi:10.3150/18-BEJ1040A
12. Yang B, Lu H, Ran Y. Advancing non-alcoholic fatty liver disease prediction: a comprehensive machine learning approach integrating SHAP interpretability and multi-cohort validation. *Front Endocrinol*. 2024;15:1450317. doi:10.3389/fendo.2024.1450317
13. Prendin F, Pavan J, Cappon G, et al. The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP. *Sci Rep*. 2023;13(1):16865. doi:10.1038/s41598-023-44155-x
14. Chowdhury SU, Sayeed S, Rashid I, Alam MGR, Masum AKM, Dewan MAA. Shapley-additive-explanations-based factor analysis for dengue severity prediction using machine learning. *J Imaging*. 2022;8(9):229. PMID: 36135395; PMCID: PMC9506144. doi:10.3390/jimaging8090229
15. Zhang S, Li J, Chen Y, Xu S. Relationship prediction between clinical subtypes and prognosis of critically ill patients with cirrhosis based on unsupervised learning methods: a study from two critical care databases. *Int J Med Inform*. 2025;201:105952. PMID: 40328059. doi:10.1016/j.ijmedinf.2025.105952
16. Zhao Y, AH H, Zhou H, MW A, GG K. A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring. *J Biopharm Stat*. 2014;24(2):229–253. PMID:24605967;PMCID:PMC4009741. doi:10.1080/10543406.2013.860769
17. Selim SZ, Ismail MA. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Trans Pattern Anal Mach Intell*. 1984;6(1):81–87. PMID:21869168. doi:10.1109/tpami.1984.4767478
18. JF V, Macías R. On the behaviour of K-means clustering of a dissimilarity matrix by means of full multidimensional scaling. *Psychometrika*. 2021;86(2):489–513. PMID: 34008128. doi:10.1007/s11336-021-09757-2
19. Zhao H, Song N, Feng H, et al. Construction and validation of a prognostic model for gastrointestinal stromal tumors based on copy number alterations and clinicopathological characteristics. *Front Oncol*. 2022;12:1055174. PMID: 36620561;PMCID:PMC9811389. doi:10.3389/fonc.2022.1055174
20. Dugan EL, Barbuto AE, Masterson CM, Shilt J. Multivariate functional principal component analysis and k-means clustering to identify kinematic foot types during gait in children with cerebral palsy. *Gait Posture*. 2024;113:40–45. PMID: 38838379. doi:10.1016/j.gaitpost.2024.05.032
21. Son M, Yang S. Association between long-term hemoglobin variability and mortality in Korean adults: a nationwide population-based cohort study. *Sci Rep*. 2019;9(1):17285. PMID:31754187;PMCID:PMC6872712. doi:10.1038/s41598-019-53709-x
22. Ciceri P, Cozzolino M. The emerging role of iron in heart failure and vascular calcification in CKD. *Clin Kidney J*. 2020;14(3):739–745. PMID: 33777358; PMCID: PMC7986369. doi:10.1093/ckj/sfaa135
23. Takata T, Mae Y, Yamada K, et al. Skeletal muscle mass is associated with erythropoietin response in hemodialysis patients. *BMC Nephrol*. 2021;22(1):134. PMID:33863297;PMCID:PMC8052822. doi:10.1186/s12882-021-02346-6
24. Zhang F, Yu J, Li G, et al. The risk factors for arteriovenous fistula dysfunction in maintenance hemodialysis patients: a cross-sectional study. *Hemodial Int*. 2024;28(2):170–177. PMID: 38448796. doi:10.1111/hdi.13145
25. Wärme A, Hadimeri H, Nasic S, Stegmayr B. The association of erythropoietin-stimulating agents and increased risk for AV-fistula dysfunction in hemodialysis patients.A retrospective analysis. *BMC Nephrol*. 2021;22(1):30. doi:10.1186/s12882-020-02209-6
26. Taghavi M, Jacobs L, Demulder A, et al. Antiphospholipid antibody positivity is associated with maturation failure and thrombosis of native arteriovenous fistula: a retrospective study in HD patients. *Clin Kidney J*. 2024;17(11):sfae308. PMID:39512379;PMCID:PMC11540859. doi:10.1093/ckj/sfae308
27. Barbieri C, Molina M, Ponce P, et al. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int*. 2016;90(2):422–429. PMID: 27262365. doi:10.1016/j.kint.2016.03.036
28. van den Oever FJ, Heetman-Meijer CFM, Birnie E, Vasbinder EC, Swart EL, Schrama YC. A pharmacist-managed dosing algorithm for darbepoetin alfa and iron sucrose in hemodialysis patients: a randomized,controlled trial. *Pharmacol Res Perspect*. 2020;8(4):e00628. PMID:32715653;PMCID:PMC7383089. doi:10.1002/prp2.628
29. Kim H, Jeong SA, Cho Y, et al. The impact of body mass index on mortality according to age in hemodialysis patients: an analysis of the Korean renal data system. *Kidney Res Clin Prac*. 2025;44(2):217–227. MID: 39815798; PMCID: PMC11985308. doi:10.23876/j.krcp.24.160
30. Shu P, Huang L, Huo S, et al. Machine learning-based risk prediction model for arteriovenous fistula stenosis. *Eur J Med Res*. 2025;30(1):217. PMID: 40156016; PMCID: PMC11954292. doi:10.1186/s40001-025-02490-x

International Journal of General Medicine

Dovepress
Taylor & Francis Group

Publish your work in this journal

The International Journal of General Medicine is an international, peer-reviewed open-access journal that focuses on general and internal medicine, pathogenesis, epidemiology, diagnosis, monitoring and treatment protocols. The journal is characterized by the rapid reporting of reviews, original research and clinical studies across all disease areas. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/international-journal-of-general-medicine-journal