

Application of Principal Component Analysis (PCA) to Uncover Clinical Patterns in Patients with Heart Failure

Nur Rahmi Asnur¹, Nurul Cahaya², Lilis Mulianti³, Taufik Hidayat⁴, Khalilah Nurfadilah^{5*}

^{1,2,3,4,5} *Mathematics Study Program, Universitas Islam Negeri Alauddin Makassar, Gowa, Indonesia*

*Corresponding author: khalilah@uin-alauddin.ac.id

*Submission date: 02 January 2026, Revision: 03 February 2026, Accepted: 03 June 2026

ABSTRACT

Heart failure is a complex clinical syndrome characterized by numerous interrelated clinical variables, making the identification of essential disease patterns increasingly challenging. Advances in medical measurement technologies have generated high-dimensional clinical datasets that require efficient analytical approaches to extract meaningful information for clinical decision-making. This study aims to apply Principal Component Analysis (PCA) to reduce the dimensionality of clinical variables while preserving the underlying information structure. The analysis was conducted using the **Heart Failure Clinical Records** dataset from the UCI Machine Learning Repository, comprising 299 patients with 13 clinical variables. The results showed that the first three principal components accounted for 52.57% of the total data variance. The first principal component (PC1) contributed the largest proportion of explained variance and was primarily associated with serum creatinine, follow-up time, and serum sodium levels, all of which were closely related to patient mortality. Furthermore, biplot visualization demonstrated the ability of PCA to distinguish patient groups according to their clinical outcomes. These findings highlight the effectiveness of PCA as an exploratory statistical tool for summarizing complex clinical data, identifying underlying clinical patterns, and supporting evidence-based medical decision-making.

KEYWORDS

Principal Component Analysis, heart failure, clinical data, dimensionality reduction, multivariate analysis.

1. INTRODUCTION

Advances in measurement technologies and electronic data recording have generated unprecedented volumes of information across numerous scientific disciplines, including healthcare and clinical medicine. Clinical datasets are typically characterised by high dimensionality, whereby a large number of variables are collected from the same individual, such as biomarkers, vital signs, and laboratory test results. However, high-dimensional data introduce inherent analytical challenges, particularly multicollinearity, in which many variables exhibit strong intercorrelations and consequently contain substantial redundant information. Such redundancy not only complicates clinical interpretation but may also reduce the stability and predictive performance of statistical models through overfitting. Consequently, effective dimensionality reduction techniques have become essential for extracting the underlying data structure while improving analytical interpretability.

Among the available dimensionality reduction techniques, Principal Component Analysis (PCA) is one of the most established and widely applied multivariate statistical methods. PCA is an unsupervised learning technique that performs an orthogonal transformation of the original set of correlated variables into a new set of uncorrelated variables, known as principal components (PCs). Mathematically, the first principal component is constructed to capture the maximum possible variance present in the data, whereas each subsequent component explains the largest proportion of the remaining variance while remaining orthogonal to all preceding components. This transformation is typically achieved through the singular value decomposition (SVD) of the covariance or correlation matrix. By retaining only a subset of principal components associated

with the largest eigenvalues, PCA preserves most of the information contained in the original dataset while substantially reducing its dimensionality.

The application of PCA is particularly relevant in the analysis of clinical data. In medical research, numerous biomarkers and clinical measurements often describe the same underlying biological process, such as inflammation or organ dysfunction. The strong correlations among these variables make interpretation based on individual variables inefficient and potentially misleading. PCA addresses this issue by consolidating correlated variables into composite indicators that represent broader patient profiles. These principal components provide more stable and multicollinearity-free variables for subsequent statistical analyses, including regression, classification, and predictive modelling. Consequently, PCA enables complex clinical information to be summarised into a limited number of interpretable dimensions that capture the major sources of variability in patients' health status.

This study focuses on heart failure (HF), a complex and progressive clinical syndrome associated with substantial morbidity and mortality worldwide. The diagnosis, risk stratification, and prognostic assessment of HF require the simultaneous evaluation of multiple clinical parameters, including haemodynamic measurements such as left ventricular ejection fraction (LVEF), biochemical biomarkers such as N-terminal pro-B-type natriuretic peptide (NT-proBNP), and indicators of comorbid organ dysfunction, including renal impairment and anaemia. The diversity and complex interactions among these clinical variables make it difficult to identify distinct patient phenotypes, despite the growing emphasis on personalised medicine. Therefore, there is an urgent need for objective statistical approaches capable of integrating multivariate clinical information to reveal latent patterns underlying the heterogeneity of patients with HF.

Considering the highly correlated and information-rich nature of HF clinical data, this study proposes PCA as an appropriate statistical approach for identifying clinically meaningful patterns. Clinical variables in HF frequently form correlated clusters that reflect common pathophysiological mechanisms. PCA can reduce these clusters into a smaller number of principal components that may represent fundamental clinical domains, such as ventricular dysfunction, volume overload, or peripheral organ dysfunction. Through this approach, clinical phenotypes can be identified and quantified objectively based on the statistical structure of the data rather than on predefined clinical classifications. Although PCA has been widely applied in biomedical research, its application to comprehensive data-driven phenotype exploration among patients with HF, particularly within the Indonesian population, remains relatively limited. Such an investigation has the potential to bridge the gap between complex multivariate clinical data and clinically meaningful interpretation.

Based on the foregoing rationale, the primary objective of this study is to apply Principal Component Analysis (PCA) to a clinical dataset of patients with heart failure in order to identify principal components that optimally represent the variability of patient phenotypes. Specifically, this study aims to: (1) provide an efficient statistical framework for reducing the dimensionality of complex HF clinical data; and (2) identify composite indicators in the form of principal components that may serve as an objective basis for risk stratification and clinical phenotype identification among patients with heart failure. The findings are expected to provide new insights into the heterogeneity of heart failure and facilitate the development of more robust statistical and predictive models for clinical decision-making.

2. LITERATURE REVIEW

Multivariate analysis is a statistical approach designed to analyze datasets containing multiple variables measured simultaneously for each observational unit. By examining several variables concurrently, multivariate methods enable researchers to identify complex relationships and underlying data structures that cannot be adequately captured through univariate analyses. Numerous studies have demonstrated the broad applicability of multivariate techniques across diverse research domains. For instance, Asfar et al [1] employed Multiple Correspondence Analysis (MCA) to identify patterns of diabetes mellitus symptoms and construct simplified risk-based patient profiles. Similarly, Zikir et al [2] applied cluster analysis to classify small and medium-sized industries, while Nurfadilah [3] utilized longitudinal cluster analysis to group gender-based human development indicators. These studies illustrate the effectiveness of multivariate techniques in simplifying complex datasets while preserving their essential information.

In clinical research, multivariate analysis plays a crucial role in understanding the complex interactions among clinical variables [4]. As the number of measured variables increases, substantial correlations among variables often arise, leading to redundancy and reduced interpretability. Consequently, dimensionality reduction techniques are required to retain the most

informative features while simplifying the overall data structure. Among these techniques, Principal Component Analysis (PCA) is one of the most widely used statistical methods. PCA transforms a set of correlated variables into a smaller number of mutually orthogonal principal components through linear combinations of the original variables, with each component ordered according to the proportion of total variance it explains [5]. As a result, PCA effectively reduces data complexity while preserving the dominant information contained in the original variables and revealing latent patterns within clinical datasets [6].

The application of PCA to clinical data has been extensively documented in the biomedical literature. Ljubicic et al [7] applied PCA to clinical and biochemical parameters in children with congenital adrenal hyperplasia and demonstrated that numerous laboratory measurements could be summarized into a limited number of components representing patients' overall clinical status. In the context of heart failure, Amir et al [8] combined PCA with k-means and hierarchical clustering to identify phenotypic subgroups of patients with heart failure with preserved ejection fraction (HFpEF) that could not be distinguished using conventional clinical variables alone. Their findings support the utility of PCA for uncovering latent phenotypic structures within heterogeneous heart failure populations. Furthermore, using the same Heart Failure Clinical Records dataset employed in the present study, Chicco and Jurman [9] adopted a machine learning-based feature-ranking approach and identified serum creatinine and ejection fraction as the two strongest predictors of survival among all available clinical variables.

Evidence supporting the effectiveness of PCA has also been reported in other clinical domains. Mekov et al [10] applied PCA to phenotype patients with chronic obstructive pulmonary disease (COPD) and reported that the first two principal components explained only 62% of the total variance. This finding suggests that cumulative explained variance within the range of 50%-60% is common in heterogeneous clinical datasets involving multiple organ systems and should not be interpreted as a limitation of PCA. More recently, Jaiswal et al [11] demonstrated that PCA can be effectively integrated with classification algorithms, including Multilayer Perceptron (MLP) models, as a preprocessing step to improve heart disease prediction performance. These studies highlight the versatility of PCA, both as an exploratory tool for identifying latent data structures and as a dimensionality reduction technique for enhancing predictive modeling.

Despite the growing body of evidence supporting PCA in cardiovascular research, several limitations remain. Most previous studies on heart failure, including that of Amir et al [8], relied on sophisticated cardiovascular modeling parameters that may not be routinely available in primary healthcare settings. In contrast, the present study focuses on seven routinely collected continuous clinical variables that are readily accessible in standard clinical practice. By applying PCA to these commonly available measurements, this study seeks to demonstrate that meaningful clinical patterns can be identified using data obtainable during routine patient assessment. Consequently, the findings may provide a practical statistical framework for early clinical screening and risk stratification of patients with heart failure.

3. METHODOLOGY

3.1 Data source

The dataset consists of 13 clinical variables, including age, anaemia (binary), creatinine phosphokinase, diabetes (binary), ejection fraction, high blood pressure (binary), platelets, serum creatinine, serum sodium, sex (binary), smoking (binary), time (follow-up period in days), and DEATH_EVENT, which indicates patient survival status (0 = survived, 1 = deceased).

Among these variables, seven are continuous numerical variables (age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and time), whereas five variables (anaemia, diabetes, high blood pressure, sex, and smoking) are binary. Because conventional PCA based on the Pearson correlation matrix assumes continuous variables, only the seven continuous variables were included in the principal component extraction. The binary variables were excluded from the PCA, whereas DEATH_EVENT was retained solely for labeling observations in the PCA biplot and for subsequent comparisons of principal component scores between patient outcome groups.

3.2 Analytical Procedure

The study employed Principal Component Analysis (PCA) as the primary analytical method. The analytical procedure consisted of the following steps.

1. Data preprocessing

Prior to PCA, all continuous variables were standardized to eliminate differences in measurement scales. Standardization was performed using the z-score transformation,

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}, i = 1, 2, \dots, 7 \quad (1)$$

where X_i denotes the observed value of variable i , μ_i is its sample mean, and σ_i represents its standard deviation. Standardization ensures that each variable contributes equally to the covariance structure regardless of its original measurement unit.

2. Assessment of PCA suitability

The suitability of the dataset for PCA was evaluated using Bartlett's Test of Sphericity and the Kaiser–Meyer–Olkin (KMO) Measure of Sampling Adequacy. Bartlett's test was used to determine whether the correlation matrix differed significantly from an identity matrix, whereas the KMO statistic assessed the proportion of common variance shared among variables.

3. Correlation analysis

Pearson correlation coefficients were computed and visualized using a correlation heatmap to examine linear relationships among variables and identify potential redundancy before dimensionality reduction.

4. Principal component extraction

Principal components were extracted from the Pearson correlation matrix. The number of retained components was determined according to the Kaiser criterion (eigenvalue > 1), with additional confirmation provided by inspection of the scree plot.

5. Loading analysis

The loading matrix was examined to evaluate the contribution and direction of each clinical variable to the extracted principal components.

6. PCA biplot visualization

A PCA biplot was constructed using the first two principal components to visualize the distribution of patients in the reduced-dimensional space. Observations were colored according to DEATH_EVENT to facilitate visual assessment of differences between surviving and deceased patients.

7. Comparison of principal component scores

The normality of principal component scores was assessed using the Shapiro–Wilk test. Because normality assumptions were not satisfied for all principal components, differences in PC1, PC2, and PC3 scores between surviving and deceased patients were evaluated using the Mann–Whitney U test.

4. RESULT & DISCUSSION

Principal Component Analysis (PCA) was performed to reduce the dimensionality of the clinical dataset by transforming the original correlated variables into a smaller number of mutually orthogonal principal components. Components with eigenvalues greater than one were retained according to the Kaiser criterion, as they explain the largest proportion of the total variance.

1. Assessment of PCA Suitability

The suitability of the dataset for PCA was first evaluated using Bartlett's Test of Sphericity and the Kaiser–Meyer–Olkin (KMO) Measure of Sampling Adequacy. Bartlett's test was statistically significant ($\chi^2 = 57.48, df = 21, p < 0.001$), indicating that the correlation matrix differed significantly from an identity matrix and confirming that the variables exhibited sufficient intercorrelations for principal component analysis. However, the overall KMO value of 0.554 falls

within the mediocre category according to Kaiser's classification, suggesting that the common variance shared among variables was only moderately adequate for dimensionality reduction. The Measure of Sampling Adequacy (MSA) values computed for individual variables further indicated that creatinine phosphokinase (0.477) and ejection fraction (0.490) were below the commonly accepted threshold of 0.50. These relatively low MSA values suggest that both variables contributed less strongly to the overall correlation structure than the remaining clinical variables. The first three principal

Table 1. Eigenvalues and Explained Variance of the Principal Components

PC	Eigenvalue	Proportion of Variance (%)	Cumulative Variance (%)
Dim. 1	1.48	21.07	21.07
Dim. 2	1.18	16.79	37.86
Dim. 3	1.03	14.71	52.57
Dim. 4	0.97	13.80	66.37
Dim. 5	0.89	12.68	79.05
Dim. 6	0.74	10.53	89.59
Dim. 7	0.73	10.41	100.00

components had eigenvalues greater than one (PC1 = 1.48, PC2 = 1.18, and PC3 = 1.03) and together accounted for 52.57% of the total variance in the dataset. The scree plot (**Figure 1**) further confirmed that the eigenvalues decreased gradually without a distinct inflection point (elbow), indicating that the explained variance was not concentrated in only the first few principal components. This pattern is characteristic of heterogeneous clinical datasets, where multiple variables capture different physiological processes and no single latent dimension dominates the overall variability. The

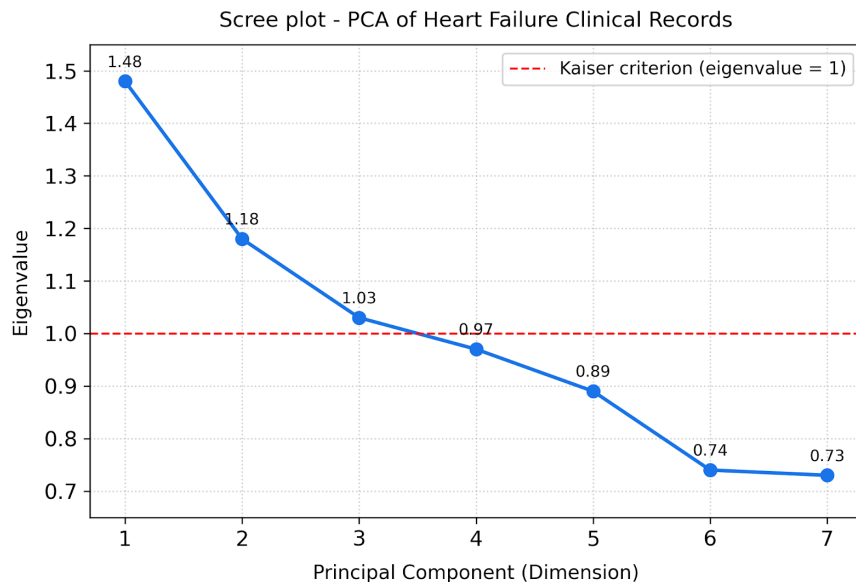


Figure 1. Scree Plot

PCA extracted three principal components with eigenvalues greater than one (PC1 = 1.48, PC2 = 1.18, and PC3 = 1.03). Collectively, these components explained 52.57% of the total variance in the dataset. According to the Kaiser criterion, retaining these three components provides an adequate representation of the multidimensional structure of the clinical data while substantially reducing the original dimensionality. Furthermore, the scree plot (**Figure 1**) shows a gradual decline in eigenvalues without a pronounced elbow, indicating that the explained variance is distributed relatively evenly across several components rather than being concentrated within the first few principal components. This pattern is commonly observed in heterogeneous clinical datasets involving variables that represent multiple physiological systems.

2. Loading Matrix

The loading matrix indicated that PC1 was primarily characterized by serum creatinine (0.51), follow-up time (-0.48), serum sodium (-0.45), and age (0.46). Collectively, these variables represent a dimension associated with overall clinical severity and cardiorenal risk, reflecting the combined influence of renal function, electrolyte balance, patient age, and survival duration. PC2 was dominated by ejection fraction (0.68), indicating that this component primarily captures cardiac functional status. Because ejection fraction is a well-established indicator of left ventricular systolic performance, PC2 can be interpreted as representing variation in cardiac function among patients with heart failure.

Table 2. Loading Matrix (PC1–PC3)

Variable	PC1	PC2	PC3
Age	0.46	0.45	-0.01
Creatinine Phosphokinase	-0.14	-0.19	0.82
Ejection Fraction	-0.18	0.68	-0.11
Platelets	-0.20	0.25	0.40
Serum Creatinine	0.51	0.05	0.12
Serum Sodium	-0.45	0.43	0.12
Time	-0.48	-0.21	-0.37

PC3 was mainly influenced by creatinine phosphokinase (0.82) and platelet count (0.40), suggesting a dimension related to biological activity and tissue injury. Elevated creatinine phosphokinase is commonly associated with muscle or myocardial injury, whereas platelet count reflects hematological and inflammatory responses. Together, these variables may capture biological processes that complement the clinical information represented by the first two principal components. It should be noted that the orthogonality property of PCA applies to the principal component scores, meaning that the scores of PC1, PC2, and PC3 are mutually uncorrelated. In contrast, orthogonality does not apply to the variable loadings. Consequently, a single variable, such as age, may exhibit substantial loadings on more than one principal component, indicating that it contributes to multiple latent dimensions simultaneously.

3. Comparison of Principal Component Scores by Patient Survival Status

The Mann–Whitney U test revealed that PC1 scores differed significantly between surviving and deceased patients, with a large effect size ($r = 0.541$). This finding statistically confirms the separation observed in the PCA biplot (Figure 2), indicating that the first principal component effectively captures clinical variation associated with patient survival. Similarly, PC3 scores showed a statistically significant difference between the two outcome groups, with a moderate effect size ($r = 0.308$). Interestingly, this distinction was not apparent in the two-dimensional PCA biplot because the visualization was limited to PC1 and PC2. The significant contribution of PC3 therefore suggests that clinically relevant information associated with patient mortality is preserved in higher-order principal components beyond the first two dimensions.

Table 3. Results of the Mann–Whitney U Test for Principal Component Scores According to Patient Survival Status

PC	Mean Live	Mean Death	U	p-value	Effect Size (r)
PC1	-0.453	0.957	3219.0	< 0.001	0.541 (large)
PC2	0.034	-0.073	10215.0	0.500	0.039 (negligible)
PC3	-0.179	0.378	6022.0	< 0.001	0.308 (moderate)

In contrast, PC2 scores did not differ significantly between surviving and deceased patients ($p = 0.500$). Although PC2 primarily represents variation in cardiac function through its strong association with ejection fraction, the absence of a statistically significant difference suggests that this component alone was insufficient to discriminate between survival

outcomes in the present dataset. Overall, these findings demonstrate that PC1 is the principal component most strongly associated with mortality, while PC3 provides complementary prognostic information that would have been overlooked if interpretation had relied solely on the conventional two-dimensional PCA visualization. This result highlights the importance of examining higher-order principal components in clinical PCA studies, particularly when subtle yet clinically meaningful patterns may not be visible in standard biplots.

4. Clinical Interpretation of the Principal Components

(a) Clinical Interpretation of PC1

The predominance of serum creatinine, follow-up time, and serum sodium in PC1 is consistent with the findings of Chicco and Jurman (2020), who employed an independent machine learning–based feature-ranking approach and identified serum creatinine and ejection fraction as the strongest predictors of survival among patients with heart failure. This agreement provides additional support for the external validity of PC1, despite the fact that PCA is an unsupervised technique that does not incorporate DEATH_EVENT during component extraction.

From a physiological perspective, elevated serum creatinine reflects impaired renal function, a common manifestation of the cardiorenal syndrome frequently observed in patients with advanced heart failure. Likewise, shorter follow-up time is consistent with poorer survival outcomes, whereas reduced serum sodium (hyponatremia) is widely recognized as a marker of neurohormonal activation and fluid retention in decompensated heart failure. Collectively, these findings suggest that PC1 represents a latent dimension associated with overall clinical severity and cardiorenal dysfunction, explaining its strong association with mortality observed in the present study.

(b) Clinical Interpretation of PC2

Unlike PC1 and PC3, PC2 was not significantly associated with patient survival ($p = 0.500$), despite being primarily characterized by ejection fraction (loading=0.68), a variable widely recognized as an important prognostic indicator in heart failure [9]. To further investigate this apparent inconsistency, additional univariate analyses were conducted for the variables contributing most strongly to PC2. All three variables showed statistically significant differences between surviving and deceased patients: ejection fraction ($p < 0.001$), with lower mean values among deceased patients (33.5 vs 40.3); age ($p < 0.001$), with deceased patients being older on average; and serum sodium ($p < 0.001$), indicating a greater prevalence of hyponatremia among patients who died.

The absence of statistical significance at the level of the composite component may be explained by the opposing clinical implications of these variables. Although age, ejection fraction, and serum sodium all exhibited positive loadings on PC2, their relationships with mortality differ substantially. Lower values of ejection fraction and serum sodium are associated with increased clinical risk, whereas higher age is likewise associated with poorer prognosis. Consequently, reductions in PC2 scores attributable to diminished cardiac function and hyponatremia may be partially offset by increases associated with older age, thereby attenuating the overall discriminatory ability of the composite component.

This finding illustrates an important limitation of PCA in clinical applications. Although PCA effectively summarizes correlated variables into latent dimensions, a composite component may obscure the prognostic contribution of individual variables when they exhibit different or even opposing relationships with the clinical outcome. Accordingly, PC2 should be interpreted primarily as a descriptive dimension reflecting cardiac function and the age-related metabolic profile of patients, rather than as a direct prognostic indicator of mortality.

(c) Clinical Interpretation of PC3

The significant association between PC3 and patient mortality warrants careful interpretation. Classical cardiovascular literature suggests that creatinine phosphokinase (CPK) is not consistently elevated in congestive heart failure itself, in contrast to acute myocardial infarction, where CPK serves as a well-established biomarker of myocardial injury. Consequently, CPK has traditionally been regarded as a marker for differentiating acute myocardial infarction from uncomplicated heart failure rather than as an indicator of heart failure severity.

Nevertheless, studies involving critically ill patients have reported that elevated CPK levels are associated with increased mortality and a greater requirement for inotropic support, suggesting that CPK reflects the extent of

systemic tissue injury rather than cardiac dysfunction alone. Therefore, the significance of PC3 observed in this study is likely to represent the presence of acute complications or comorbid conditions among patients with more severe heart failure, rather than a dimension directly attributable to cardiac function itself. Importantly, this clinically relevant information would not have been identified if interpretation had relied exclusively on the conventional two-dimensional PCA biplot.

(d) Overall Interpretation

Overall, the PCA identified three major latent dimensions underlying the clinical characteristics of patients with heart failure. PC1 primarily represents overall clinical severity and cardiorenal dysfunction, PC2 reflects cardiac function and age-related physiological variation, and PC3 captures biological activity and systemic tissue injury that may be associated with severe comorbid conditions. These findings emphasize that the interpretation of PCA should be based primarily on the magnitude of variable loadings, which indicates the relative contribution of each variable to a given principal component. At the same time, the sign of the loading should not be overlooked, as it conveys the direction of the relationship between the original variable and the corresponding latent dimension.

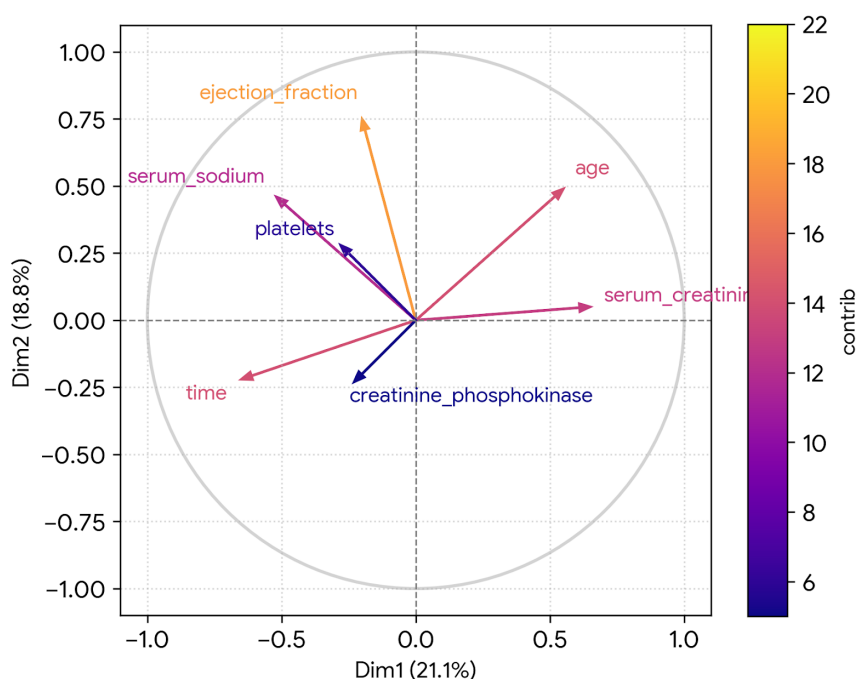


Figure 2. Biplot of the first two principal components (PC1 and PC2)

Figure 2 presents the PCA biplot of the first two principal components, illustrating the relationships between the clinical variables and the reduced-dimensional space defined by PC1 (horizontal axis) and PC2 (vertical axis). The first principal component explained 21.1% of the total variance, whereas the second accounted for 16.8%, yielding a cumulative explained variance of 37.9%. In the biplot, each arrow represents an original clinical variable, where longer vectors indicate stronger contributions to the principal components. The color gradient reflects the relative magnitude of each variable's contribution, with darker red indicating greater importance.

Among the variables, ejection fraction and age contributed most strongly to PC2, as indicated by their relatively long vectors oriented toward the positive direction of the second principal component. In contrast, serum creatinine and serum sodium exhibited the strongest contributions to PC1, suggesting that this component primarily reflects variation associated with cardiorenal status and overall clinical severity. Meanwhile, follow-up time and creatinine phosphokinase were projected toward the negative side of PC1, indicating negative loadings on the first principal component.

The relative orientation of the vectors also provides insight into the relationships among the original variables.

Variables with similar vector directions, such as age and ejection fraction, tend to exhibit positive correlations, whereas variables pointing in opposite directions, such as follow-up time and serum creatinine, indicate negative correlations. Variables forming approximately right angles are interpreted as having weak or negligible linear associations.

Overall, the biplot supports the interpretation derived from the loading matrix by demonstrating that PC1 is primarily associated with clinical severity and cardiorenal dysfunction, whereas PC2 predominantly reflects cardiac functional status together with age-related physiological variation. As an exploratory visualization, the biplot provides an intuitive representation of the multivariate relationships among clinical variables while facilitating the interpretation of the latent dimensions identified by PCA

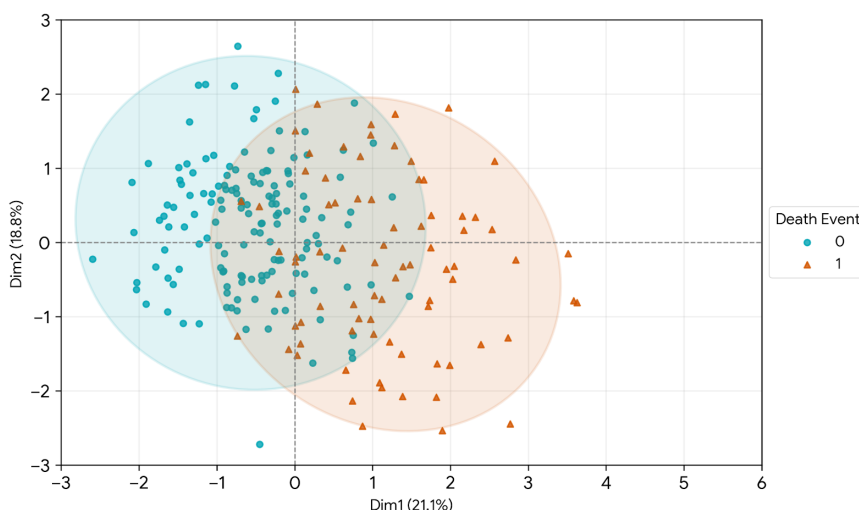


Figure 3. PCA Biplot

5. CONCLUSION

This study applied Principal Component Analysis (PCA) to seven continuous clinical variables obtained from 299 patients with heart failure, resulting in three principal components that collectively explained 52.57% of the total variance. Although the suitability assessment indicated a statistically significant correlation structure among the variables (Bartlett's test, $p < 0.001$), the overall Kaiser–Meyer–Olkin (KMO) value of 0.554 suggested only moderate sampling adequacy, reflecting the heterogeneous nature of clinical variables representing different physiological systems.

Among the extracted components, PC1, primarily characterized by serum creatinine, follow-up time, and serum sodium, demonstrated the strongest association with patient survival and significantly discriminated between surviving and deceased patients with a large effect size. This finding is consistent with previous studies identifying renal function and electrolyte balance as key prognostic indicators in heart failure. In addition, the significant association between PC3, dominated by creatinine phosphokinase, and mortality represents an important finding of this study. Because this relationship was not apparent in the conventional two-dimensional PCA biplot, it suggests that higher-order principal components may contain clinically meaningful information that could otherwise remain undetected using visual inspection alone. In contrast, PC2, representing cardiac functional status, did not independently differentiate survival outcomes despite containing clinically important variables such as ejection fraction.

Overall, the findings demonstrate that PCA is an effective exploratory statistical tool for summarizing multidimensional clinical data and identifying latent clinical patterns in patients with heart failure. At the same time, the results emphasize that principal components should be interpreted cautiously, as composite components may obscure the prognostic contribution of individual variables when their clinical associations differ. Therefore, PCA should be regarded primarily as an exploratory technique that complements, rather than replaces, subsequent predictive modeling approaches.

Several limitations should be acknowledged. First, the analysis included only seven continuous clinical variables because classical PCA based on the Pearson correlation matrix is not directly applicable to binary variables. Consequently, five binary variables contained in the original dataset were excluded from the dimensionality reduction process. Second, the findings are specific to the study population comprising patients with New York Heart Association (NYHA) class III–IV heart failure from Faisalabad, Pakistan, and therefore require external validation before being generalized to other populations. Finally, the association observed between PC3 and mortality remains exploratory because the dataset did not contain detailed information regarding comorbidities or acute clinical events that could explain the underlying biological mechanisms.

Future studies should consider dimensionality reduction techniques specifically designed for mixed data, such as Factor Analysis of Mixed Data (FAMD) or hybrid approaches combining Multiple Correspondence Analysis (MCA) with PCA, thereby allowing both continuous and categorical clinical variables to be analyzed simultaneously. Furthermore, incorporating additional biomarkers, such as inflammatory markers or N-terminal pro-B-type natriuretic peptide (NT-proBNP), may strengthen the underlying correlation structure and improve the clinical interpretability of the extracted components. Validation using independent heart failure cohorts, particularly in Indonesian populations, is also warranted to assess the generalizability of the identified clinical dimensions.

REFERENCES

- [1] F. Asfar, Ambarwati, M. Jannah, D. Sartika, A. Setiawan, and K. Nurfadilah, “Exploring diabetes mellitus risk patterns with multiple correspondence analysis at torabelo hospital, central sulawesi,” *MSA*, vol. 13, no. 2, pp. 1–12, 2025.
- [2] A. Zikir, K. Nurfadilah, and A. Irwan, “Perbandingan metode clustering dengan menggunakan metode average linkage dan metode k-means pada industri kecil dan menengah di kabupaten wajo,” *MSA*, vol. 10, no. 2, pp. 57–62, 2022.
- [3] K. Nurfadilah, “Analisis cluster longitudinal pada pembangunan manusia di sulawesi selatan berbasis gender,” *MSA*, vol. 9, no. 1, pp. 89–93, 2021.
- [4] I. Ghozali, *Aplikasi Analisis Multivariate dengan Program IBM SPSS 23*. Semarang: Badan Penerbit Universitas Diponegoro, 2016.
- [5] B. Santosa and I. Setiawan, *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2006.
- [6] Kusrini and L. Luthfi, *Algoritma Data Mining*. Yogyakarta: Andi, 2009.
- [7] M. L. Ljubicic, A. Madsen, A. Juul, K. Almstrup, and T. H. Johannsen, “The application of principal component analysis on clinical and biochemical parameters exemplified in children with congenital adrenal hyperplasia,” *Frontiers in Endocrinology*, vol. 12, p. Article 652888, 2021.
- [8] O. A. et al., “Phenotyping heart failure using model-based analysis and physiology-informed machine learning,” *Frontiers in Physiology / bioRxiv preprint*, 2021.
- [9] D. Chicco and G. Jurman, “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 16, 2020.
- [10] E. V. Mekov, N. A. Yanev, N. Kurtelova, T. Mihalova, A. Tsakova, Y. Yamakova, and R. E. Petkov, “Phenotyping chronic obstructive pulmonary disease through principal component analysis: Identification of clinical clusters,” *Cureus*, vol. 17, no. 4, p. e82811, 2025.
- [11] V. Jaiswal, S. Gupta, S. Gupta, D. Kumar, and A. A. Hameed, “Enhancing the heart disease classification using multi-level perceptron and principal component analysis,” in *Recent Trends and Advances in Artificial Intelligence (ICAETA 2024)*, ser. LNNS, vol. 1138. Springer, 2024.