JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 3, July 2026, pp. 725-736

A Comparative Study of PCA-Based Dimensionality Reduction and Best Subset Selection in Disease Classification

Andreas Rony Wijaya[1*], Atika Ratna Dewi[2], Muhammah Bayu Nirwana[1], Respatiwulan[1], Sri Sulistijowati Handajani[1]
[1]Department of Statistics, Universitas Sebelas Maret, Indonesia
[2]Department of Data Sciences, Universitas Telkom, Indonesia
andreasrony@staff.uns.ac.id

Article History:
Received : 10-02-2026
Revised : 23-03-2026
Accepted : 25-03-2026
Online : 01-07-2026

Keywords:
Classification;
Dimesionality Reduction;
Feature Selection;
PCA;
Subset Regression.

ABSTRACT

Real-world datasets often contain many variables, some of which may be irrelevant or redundant. To build an effective classification model, it is important to simplify the data by keeping only the most influential features. One common approach that can be used for selecting the most influential variables is feature selection. However, when dealing with many variables, removing some may result in the loss of information. Hence, it is also necessary to consider methods that can simplify the model while retaining most of the information from the original variables. Dimensionality reduction is one such approach that effectively addresses this issue. This study employs a comparative quantitative research approach to evaluate the effectiveness of principal component analysis (PCA) as a dimensionality reduction method and best subset selection as a feature selection method in improving classification performance. The study utilizes a heart disease dataset from the UCI Machine Learning Repository consisting of 303 observations and 13 predictor variables as a case study. Both approaches are applied to reduce the number of predictor variables and make the model more interpretable. After applying both methods, three classification models — logistic regression, naïve Bayes, and linear discriminant analysis — are trained and evaluated using accuracy, recall, precision, and F1-score, and the results are further illustrated through ROC curves. Feature selection using best-subset selection yields seven variable combinations with the most significant predictors, whereas PCA requires eight principal components to explain 80% of the total variation. The best classification performance was obtained using the feature-selected dataset, achieving an accuracy of 87% and an AUC of 0.93, outperforming both the original dataset model and the PCA-reduced dataset model. These results show that feature selection using best subset selection provides a better balance between simplicity and classification performance. Furthermore, the models obtained after feature reduction, both from best subset selection and PCA, still maintain good predictive ability as indicated by their relatively high AUC values.

https://doi.org/10.31764/jtam.v10i3.38265

This is an open access article under the CC–BY-SA license

A. INTRODUCTION

Classification is recognized as one of the primary techniques applied in machine learning, used to assign observations into predefined categories (Han et al., 2023; Heaton, 2018). One common issue in building classification models is when the dataset contains too many variables, some of which may not have a significant effect on the model. This situation can make the model more complex, increase computation time, and sometimes even reduce its interpretability (Guyon & Elisseeff, 2003; Li et al., 2018). Therefore, selecting or transforming variables to

retain essential information while simplifying the model has become a key step in the modelling process. To overcome this issue, model simplification becomes important. There are two common approaches to simplifying models: the first is feature selection, which highlights the most important predictors affecting the model's performance, and the second is dimensionality reduction, which reduces the original data size to a smaller number of components that retain most of the data's information.

Feature selection plays a crucial role when we aim to keep the original interpretability of variables while reducing model complexity. Recent studies have shown the growing relevance of advanced feature selection methods, such as hybrid and ensemble-based techniques in complex data (Kuzudisli et al., 2023; Labory et al., 2024; Mohtasham et al., 2024). One of the widely used feature selection methods is best subset selection, which evaluates all possible combinations of predictors to identify which subset that gives the best balance between accuracy and simplicity based on certain evaluation criteria. Prior studies about best subset selection have shown that this approach can effectively improve both model performance and interpretability, especially in applications to medical and biological datasets (Devaraj & Paulraj, 2015).

On the other hand, principal component analysis (PCA) represents a well-established and widely adopted approach for dimensionality reduction. PCA reconstructs the original correlated features into new components that preserve most of the variation present in the dataset. However, several studies have noted that the components capturing the highest variance do not always correspond to the best class separation (Zheng & Rakovski, 2021). Therefore, selecting an appropriate number of components is important for achieving optimal classification performance. Recent works have further explored PCA and its variants, for instance Joosse et al. (2025) using PCA for a large-scale study on routine clinical haematology data. Another study by Parman et al. (2024) applied PCA in a breast cancer classification task. The reported accuracy is increased, showing the potential benefit of dimensionality reduction for classification performance.

Many studies have shown that selecting or reducing features before running a classification model can improve its overall performance in different application areas. For example, Abdollahi & Nouri-Moghaddam (2021) applied feature selection for medical diagnosis and demonstrated improved classification accuracy compared to models using all features. Similarly, Sankarganesh & Priya (2024) used feature selection in diabetes mellitus classification, applying a random forest-based U-Net algorithm, and found that eliminating redundant variables reduced overfitting and enhanced predictive stability. Then, Andika & Dewi (2025) applied feature selection for multiple disease classification and showed better accuracy and computational efficiency. On the other hand, Esen et al. (2024) applied PCA in breast cancer data, resulting in improved classification performance when reducing dimensionality before classifying the data.

However, most existing works tend to focus on a single approach, either feature selection or dimensionality reduction, without directly comparing their relative strengths under the same modelling conditions. Several studies have investigated feature selection techniques to improve classification performance, while others have explored dimensionality reduction methods to address high-dimensional data challenges (Li et al., 2017; Morán-Fernández et al.,

2022; Sharifai & Zainol, 2020). This gap leaves open the question of which method offers the best trade-off between model simplicity, interpretability, and predictive performance, especially in complex medical datasets characterized by high dimensionality and correlated variables.

In this study, we focus on disease classification problems, specifically heart disease, as these are among the most prevalent chronic diseases worldwide and involve high-dimensional datasets. We conduct a comparative analysis between best subset selection as a feature selection technique and PCA as a dimensionality reduction technique to evaluate which approach achieves a better balance between model simplicity and classification effectiveness in medical data. This study contributes by systematically comparing these approaches across multiple classification models and demonstrating their impact on both predictive performance and model interpretability in heart disease classification.

B. METHODS

1. Research Process

This research employs a comparative quantitative research design to evaluate the performance of classification models under different data preparation techniques. The overall methodology is organized into multiple phases, beginning with data preprocessing. The dataset is first classified without incorporating any feature reduction, followed by feature reduction using both feature selection and dimensionality reduction techniques. The process continues with model training and concludes with performance evaluation.

a. Feature Selection

Feature selection is first performed using the best subset selection method. This approach identifies optimal combinations of predictor variables to identify the subset that yields the highest predictive capability. Basically, the best subset selection systematically evaluates all possible combinations of predictor variables. Here is the algorithm for best subset selection (Hanke et al., 2024):

1) Start: model with no predictors (only the intercept).
2) For each number of predictors $k = 1,2, \dots, p$: (a) Fit all possible models with k predictors; (c) Compute performance metrics for each k; and (c) In this research, the Bayesian Information Criterion (BIC) is used as the metric to select the optimal model, that the BIC formula is given by Equation (1):

$$BIC = n \cdot ln\left(\frac{RSS}{n}\right) + k \cdot ln(n) \quad (1)$$

where n is number of observations; k is count of estimated parameters (including the intercept); and RSS is residual sum of squares.

3) Choose the model corresponding to the minimum BIC among all different subset sizes. After selecting the optimal subset of predictor variable combinations, the dataset is reclassified using these feature sets.

b. Dimensionality Reduction

Following the feature selection stage, the study then applies dimensionality reduction through principal component analysis (PCA). This method transforms the original features into a smaller number of uncorrelated components, known as principal components. The main objective of PCA is to capture as much as of the total data variance in lower dimensions (Wu et al., 2022). The reduced-dimensional dataset is subsequently classified to evaluate how dimensionality reduction affects predictive accuracy. Suppose there are q variables of the original data, using PCA will transform it into a smaller number of k principal components (Shen, 2023). Let the original data be represented as a vector of $X^T = \{X_1, X_2, \ldots, X_p\}$, where p denotes the number of original variables. The matrix variance-covariance of X can be defined as $Cov(X) = \mathbf{\Sigma}$. PCA performs an eigen decomposition of $\mathbf{\Sigma}$ to obtain a set of eigenvalue-(normalized) eigen vector pairs (λ_1, e_1), $(\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ where the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and $e_i^T e_k = 1$ if $i = k$; 0 if $i \neq k$. Then the ith principal component is given by,

$$Y_i = e_i^T \boldsymbol{X} = e_{i1}X_1 + \ldots + e_{ip}X_p; \;\; i = 1, \ldots, p$$

The variance and covariance of each principal component is

$$Var(Y_i) = e_i^T \mathbf{\Sigma} e_i = \lambda_i; \quad i = 1, \ldots, p$$

and

$$Cov(Y_i, Y_k) = e_i^T \mathbf{\Sigma} e_i = 0; \quad 1 \neq k.$$

The total variance of all principal components, represented by the sum of their eigenvalues, equals to the total variance of the original variable,

$$\lambda_1 + \lambda_2 + \ldots + \lambda_p = \Sigma_{i=1}^{p} Var(Y_i) = \Sigma_{i=1}^{p} Var(X_i) = \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}$$

The proportion of total population variance due to the kth principal component is given by Equation (2),

$$proportion_{k-th}\ PC = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \tag{2}$$

In practice, a subset of the first q principal components are selected such that the cumulative explained variance reaches a certain threshold, typically around 80%-90%. These selected components are then used as input features for the classification model, replacing the original variables without much loss of information (B. Li et al., 2022). Finally, all three datasets, the original data, the feature-selected data, and the PCA-reduced data, are used to train and test classification models. Model performance is primarily evaluated using accuracy, enabling a clear assessment of how feature selection and dimensionality reduction affect classification results.

c. Classification Methods

The classification techniques applied in this research include logistic regression, the naïve Bayes algorithm, and linear discriminant analysis. Logistic regression is a supervised learning technique that estimates the likelihood of a binary response based on a set of predictors variables. The method employs a sigmoid transformation to convert model outputs into probability values bounded between 0 and 1 (Austin & van Buuren, 2023). The model is expressed as,

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

The $P(Y = 1|X)$ denotes the probability that the dependent variable Y equal to 1 given the predictor variables, and β_i is coefficients are estimated using maximum likelihood estimation (Olowe et al., 2024; Dey et al., 2025). The second one is linear discriminant analysis, which is a statistical classification technique that seeks a linear combination of predictors that best separates two or more classes (Graf et al., 2024). The discriminant score can be written as:

$$\delta_k(x) = x\,\frac{\pi_k}{\delta^2} - \frac{\pi_k^2}{2\delta^2} + \log(\pi_k)$$

$$\Pr(Y = k|X = x) = \frac{e^{\delta_k(x)}}{\Sigma_{l=1}^{K} e^{\delta_l(x)}}$$

When $K = 2$, if $\Pr(Y = 1|\, X = x) \geq 0.5$ classify the new observation to class 1, else to class 0 (Huang et al., 2009). The third classification method employed in this study is the naïve Bayes method, which is a probabilistic classifier derived from Bayes' theorem and relies on the assumption that features are conditionally independent given the class label. Under this assumption the posterior probability of class C_k for a given feature vector X is:

$$P(C_k|X) = \frac{P(C_k)\Pi_{i=1}^{n} P(x_i|C_k)}{P(X)}$$

The predicted class is the one with the highest posterior probability (Chen et al., 2021).

d. Classification Performance Evaluation

The performance of classification models was evaluated using accuracy, precision, recall, and F1-score, which are widely used metrics for assessing classification performance in machine learning studies (Opitz, 2024; Sujon et al., 2025). In this study, TP, TN, FP and FN represent true positive, true negative, false positive, and false negative, respectively. Accuracy measures the proportion of correctly classified instances among all observations and is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the proportion of correctly predicted positive (TP) instances among all predicted positive ($TP + FP$) instances, while recall (sensitivity) measures the proportion of actual positive (TP) instances that are correctly identified by the model ($TP + TN$). The F-1 Score provides a balance between precision and recall. In addition, model performance was further assessed using the receiver operating characteristic (ROC) curve. The area under the receiver operating characteristic (ROC) curve (AUC) was used to assess the model's ability to distinguish between classes.

2. Dataset

The dataset used in this project is the heart disease dataset. The dataset was obtained from the UCI Machine Learning Library. The heart disease dataset contains 14 variables and 303 observations. The following Table 1 presents the details of variables in this dataset, with the target variable being the diagnosis of heart disease. The dataset was partitioned into separated training set and testing set with proportions of 2/3 and 1/3, respectively.

Table 1. Details of the dataset

Variable	Description	Type
Age	Patient's age	Numerical
Sex	Patients's sex	0=F, 1 = M
ChestPain	Category of chest pain	Typical, asymptomatic, non-anginal, atypical
RestBP	Blood pressure measured at rest (mmHg)	Numerical
Chol	Concentration of Serum cholesterol (mg/dL)	Numerical
Fbs	Indicator of fasting blood glucose above 120 mg/dL	0=False, 1=True
RestECG	Electrocardiographic condition at rest	Categorical
MaxHR	Highest heart rate reached during exercise	Numerical
ExAng	Occurrence of angina triggered by exercise	Binary (0=No, 1=Yes)
Oldpeak	Degree of ST-segment depression during exercise compared to rest	Numerical
Slope	Trend of the ST-segment	Numeric (categorical code)
Ca	Count of major coronary vessels detected via fluoroscopy	0-3
Thal	Condition of Thalassemia	7=Reversible defect, 6=Fixed defect, 3=Normal
(AHD) Target	Heart disease diagnosis indicator	1=Yes, 0=No

C. RESULT AND DISCUSSION

1. Model of the Original Dataset

The first conduct classification model of the original data, without any dimension reduction or feature selection. The models used in this analysis are linear discriminant analysis, logistic regression, and naïve Bayes. Before performing classification, the dataset was partitioned such that 2/3 of the observations were assigned to the training set and 1/3 to the test set. The model performance is shown in Table 2, and the ROC curves of those three models are shown in Figure 1.

Table 2. The model performance for classification using the original data

Method	Accuracy	F-1 Score	Precision	Recall
Logistic Regression	0.8515	0.8718	0.8448	0.8909
LDA	0.8515	0.8673	0.8226	0.9273
Naïve bayes	0.8614	0.8727	0.8727	0.8727

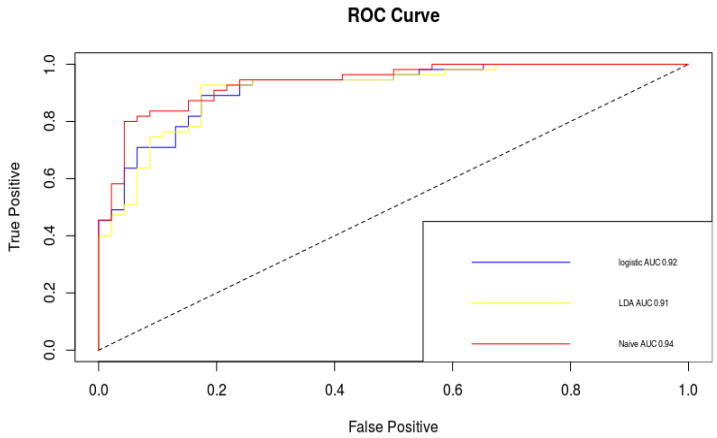


Figure 1. ROC curve of the classification using original data

Table 2 presents the performance of the original dataset classification models, using logistic regression, linear discriminant analysis (LDA), and naïve Bayes. Among them, the naïve Bayes model achieved the highest accuracy of 0.8614 and F1-score of 0.8727, indicating that this model has a better predictive performance compared to the other models. The ROC curve shown in Figure 1 supports this result, where the naïve Bayes has the largest area under the curve with an AUC of 0.94, followed by the logistic regression model with an AUC of 0.92, and LDA with an AUC of 0.91. These results indicate that all models perform well, but naïve Bayes is the highest one.

2. Feature Selection and Its Model

Feature selection was performed using the best subset selection method, with the Bayesian Information Criterion (BIC) used to evaluate model performance for each possible subset of predictors. The BIC values were calculated using Equation (1), as illustrated in Figure 2.

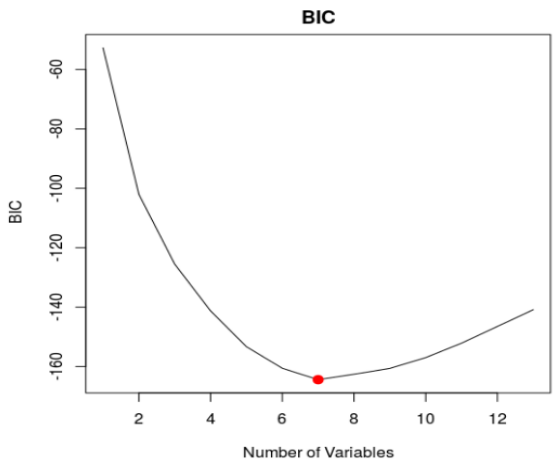


Figure 2. The value of BIC for each number of predictors

In Figure 2, the minimum BIC value is indicated by the red dot. Therefore, for this case, seven predictors can be used to conduct the classification model. These seven variables are cp, thalach, sex, oldpeak, ca, thal, and exang. The classification models were then constructed using the seven selected variables. The dataset was partitioned into separated training set and testing set with proportions of 2/3 and 1/3, respectively. The performance results for each model are summarized in Table 3. The resulting ROC curve is depicted in Figure 3.

Table 3. The model performance after feature selection

Method	Accuracy	F-1 Score	Precision	Recall
Logistic Regression	0.8515	0.8673	0.8448	0.8909
LDA	0.8416	0.8596	0.8305	0.8909
Naïve bayes	0.8713	0.8829	0.8750	0.8909

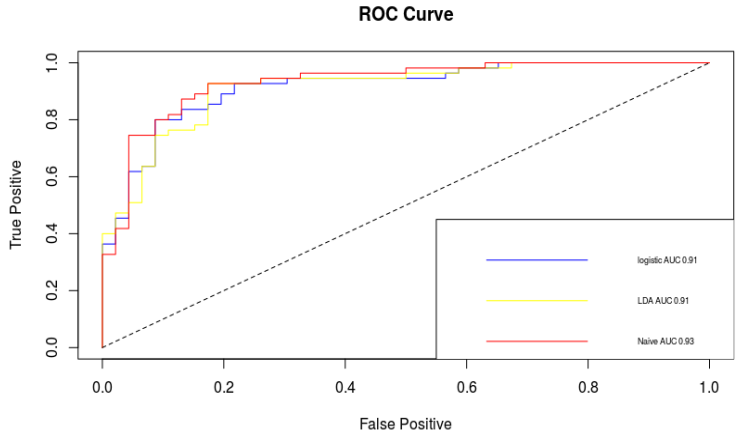


Figure 3. ROC curve of the model after feature selection

Table 3 presents the classification performance after feature selection using the best subset selection method. The naïve Bayes model also achieves the highest performance with an accuracy of 0.8713 and an F1-score of 0.8829. The ROC curve in Figure 3 shows that all three models maintain high discriminative power, with naïve Bayes having the highest AUC (0.93), followed by logistic regression and LDA, which have AUCs of 0.91.

3. Principal Component Analysis and Its Model

PCA is a widely used approach for simplifying high-dimensional data by transforming multiple original variables into fewer components that retain the majority of the dataset's variance (Johnson & Wichern, 2014). To decide how many principal components are needed in our analysis, the cumulative proportion of explained variance was calculated using Equation (2), as shown in Figure 4. Based on the results, eight principal components are needed to explain 80% of the total data variance.

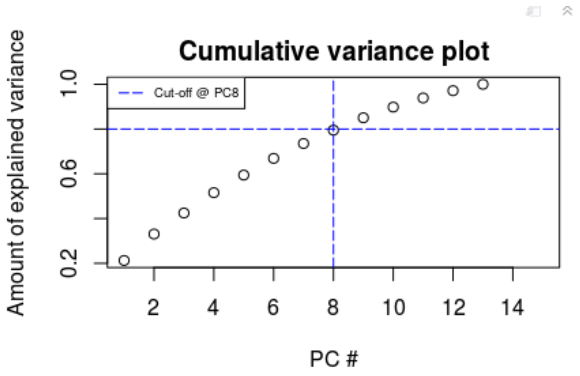


Figure 4. Plot of cumulative variance of component PCA

Eight principal components were used to perform the classification analysis. Prior to modelling, the components were partitioned using the same proportion as in the previous step. The performance comparison of the three classification models can be found in Table 4, and their ROC curves are shown in Figure 5.

Table 4. The model performance of classification using 8 PCA

Method	Accuracy	F-1 Score	Precision	Recall
Logistic Regression	0.8218	0.8257	0.8137	0.8182
LDA	0.8119	0.8257	0.8182	0.8182
Naïve bayes	0.8317	0.8468	0.8393	0.8545

Table 4 shows the classification results after applying dimensionality reduction using eight principal components obtained from PCA. Among the three models, again, naïve Bayes achieved the best performance with an accuracy of 0.8317 and an F1-score of 0.8468. Figure 5 presents the ROC curves, demonstrating that all models maintain a strong discriminative ability, with AUC values ranging from 0.89 to 0.91.

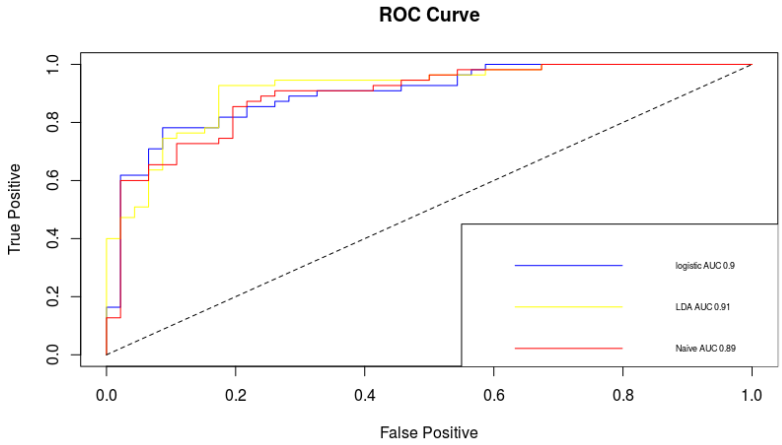


Figure 5. ROC curve of classification using 8-component PCA

4. Model Comparison

From the results and analysis presented in the previous section, nine classification models were obtained: three from the original dataset (without feature reduction), three after feature selection, and three after dimensionality reduction using PCA. The comparison of these nine models enables a clearer understanding of how the effects of feature selection and dimensionality reduction impact classification performance. Overall, models trained on the feature-selected dataset achieved the best performance, particularly for the naïve Bayes classifier, which recorded the highest accuracy and F-1 score among all models. This approach reduces the number of predictors from 13 variables to 7 variables. This result indicates that removing irrelevant or redundant variables can improve model performance.

In contrast, the PCA-based approach reduced the 13 predictor variables into 8 principal components that capture approximately 80% of the total variance. However, models trained on the PCA-reduced dataset showed there is a decrease in classification performance compared to other approaches. This decrease suggests that when applying PCA to reduce the dimensionality of the data, some information may have been lost during the transformation process. Nevertheless, the ROC curves indicate that the PCA-based models still retain acceptable classification capability. These results highlight that feature selection using best subset selection provides a better balance between simplicity and predictive accuracy compared to PCA-based dimensionality reduction in the heart disease dataset as a case study.

These findings are consistent with previous studies showing that feature selection methods can improve classification performance by retaining the most relevant features while removing irrelevant features. For instance, recent research demonstrates that feature selection techniques significantly enhance model accuracy and interpretability in classification tasks, particularly in high-dimensional dataset (Wang et al., 2024). However, other studies indicate that dimensionality reduction techniques such as PCA remain effective in capturing the overall structure of the data and improving computational efficiency, although they may lead to a slight decrease in classification performance due to information loss during transformation (Atluri et al., 2024).

D. CONCLUSION AND SUGGESTIONS

This study conducted a comparative analysis of two model simplification approaches: feature selection using best subset selection and dimensionality reduction using principal component analysis (PCA), in the case of heart disease classification. Best subset selection reduced the predictors from 13 to 7 variables and produced higher accuracy, particularly for the naïve Bayes model, achieving the best overall performance in terms of accuracy and F1-score among the evaluated models, while PCA reduced the features to 8 principal components that explain around 80% of total variance but showed slightly lower accuracy. Therefore, feature selection with best subset selection provided a better balance between simplicity, interpretability, and predictive performance than PCA-based dimensionality reduction.

For future work, this study can be extended by testing other feature selection and dimensionality reduction approaches, such as LASSO or recursive feature elimination (RFE). Then the dataset can also be extended to a larger dataset to validate the consistency of these findings. Future studies are also encouraged to evaluate the robustness of these approaches

using different classifiers and cross-validation strategies to enhance the generalizability of the results.

ACKNOWLEDGEMENT

We extend our gratitude to the Research and Community Services Agency (LPPM) at Universitas Sebelas Maret (UNS). This work was supported by RKAT Universitas Sebelas Maret under the Penelitian Penguatan Kapasitas Grup Riset (PKGR-UNS) B scheme, grant number 371/UN27.22/PT.01.03/2025.

REFERENCES

Abdollahi, J., & Nouri-Moghaddam, B. (2021). Feature selection for medical diagnosis: Evaluation for using a hybrid Stacked-Genetic approach in the diagnosis of heart disease. *ArXiv*. https://arxiv.org/abs/2103.08175

Andika, R. A., & Dewi, C. (2025). Importance of Feature Selection for Multiple Disease Classification. *Jurnal Buana Informatika*, *16*(1), 34–45.

Austin, P. C., & van Buuren, S. (2023). Logistic regression vs. predictive mean matching for imputing binary covariates. *Statistical Methods in Medical Research*, *32*(11), 2172–2183. https://doi.org/10.1177/09622802231198795

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *Eurasip Journal on Advances in Signal Processing*, *30*(2021), 1–12. https://doi.org/10.1186/s13634-021-00742-6

Devaraj, S., & Paulraj, S. (2015). An Efficient Feature Subset Selection Algorithm for Classification of Multidimensional Dataset. *Scientific World Journal*, *2015*. https://doi.org/10.1155/2015/821798

Dey, D., Haque, M. S., Islam, M. M., Aishi, U. I., Shammy, S. S., Mayen, M. S. A., Noor, S. T. A., & Uddin, M. J. (2025). The proper application of logistic regression model in complex survey data: a systematic review. *BMC Medical Research Methodology*, *25*(15). https://doi.org/10.1186/s12874-024-02454-5

Esen, G., Altaibek, A., Amankulov, J., Matkerim, B., & Nurtas, M. (2024). Enhancing Breast Cancer Detection with Dimensionality Reduction Techniques: A Study Using PCA and LDA on Wisconsin Breast Cancer Data. *Procedia Computer Science*, *251*, 414–421. https://doi.org/10.1016/j.procs.2024.11.128

Graf, R., Zeldovich, M., & Friedrich, S. (2024). Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study. *Biometrical Journal*, *66*(1). https://doi.org/10.1002/bimj.202200098

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection André Elisseeff. *Journal of Machine Learning Research*, *3*, 1157–1182.

Han, J., Pei, J., & Tong, H. (2023). *Data Mining: Concepts and Techniques*.

Hanke, M., Dijkstra, L., Foraita, R., & Didelez, V. (2024). Variable selection in linear regression models: Choosing the best subset is not always the best choice. *Biometrical Journal*, *66*(1). https://doi.org/10.1002/bimj.202200209

Heaton, J. (2018). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. *Genetic Programming and Evolvable Machines*, *19*(1–2), 305–307. https://doi.org/10.1007/s10710-017-9314-z

Huang, D., Quan, Y., He, M., & Zhou, B. (2009). Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *Journal of Experimental and Clinical Cancer Research*, *28*(1). https://doi.org/10.1186/1756-9966-28-149

Johnson, R. A. ., & Wichern, D. W. . (2014). *Applied multivariate statistical analysis*. Pearson Educated Limited.

Joosse, H. J., Chumsaeng-Reijers, C., Huisman, A., Hoefer, I. E., van Solinge, W. W., Haitjema, S., & van Es, B. (2025). Haematology dimension reduction, a large scale application to regular care

haematology data. *BMC Medical Informatics and Decision Making*, *25*(1). https://doi.org/10.1186/s12911-025-02899-8

Kehinde Josephine Olowe, Ngozi Linda Edoh, Stephane Jean Christophe Zouo, & Jeremiah Olamijuwon. (2024). Comprehensive review of logistic regression techniques in predicting health outcomes and trends. *World Journal of Advanced Pharmaceutical and Life Sciences*, *7*(2), 016–026. https://doi.org/10.53346/wjapls.2024.7.2.0039

Kuzudisli, C., Bakir-Gungor, B., Bulut, N., Qaqish, B., & Yousef, M. (2023). Review of feature selection approaches based on grouping of features. In *PeerJ* (Vol. 11). PeerJ Inc. https://doi.org/10.7717/peerj.15666

Labory, J., Njomgue-Fotso, E., & Bottini, S. (2024). Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. *Computational and Structural Biotechnology Journal*, *23*, 1274–1287. https://doi.org/10.1016/j.csbj.2024.03.016

Li, B., Gui, X., & Zhou, Q. (2022). Construction of Development Momentum Index of Financial Technology by Principal Component Analysis in the Era of Digital Economy. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/2244960

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. In *ACM Computing Surveys* (Vol. 50, Number 6). Association for Computing Machinery. https://doi.org/10.1145/3136625

Mohtasham, F., Pourhoseingholi, M. A., Hashemi Nazari, S. S., Kavousi, K., & Zali, M. R. (2024). Comparative analysis of feature selection techniques for COVID-19 dataset. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-69209-6

Opitz, J. (2024). A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice. *Transactions of the Association for Computational Linguistics*, *12*, 820–836. https://doi.org/https://doi.org/10.1162/tacl_a_00675

Parman, N. H., Hassan, R., & Zakaria, N. H. (2024). Breast Cancer Prediction Using Support Vector Machine Ensemble with PCA Feature Selection Method. *International Journal of Innovative Computing*, *14*(1), 15–19. https://doi.org/10.11113/ijic.v14n1.461

Sankarganesh, P. V, & Priya, D. R. (2024). Improved Feature Selection and Classification for Diabetes Mellitus Using Random Forest-Based U-Net Classifier. *International Journal of Intelligent Systems and Applications in Engineering IJISAE*, *12*(4), 1772–1780. www.ijisae.org

Shen, Z. (2023). Comparison and Evaluation of Classical Dimensionality Reduction Methods. *Highlights in Science, Engineering and Technology ICMEA*, *70*(2023), 411–418. https://doi.org/https://doi.org/10.54097/hset.v70i.13890

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data*, *12*(1). https://doi.org/10.1186/s40537-025-01313-4

Wu, R. M. X., Zhang, Z., Yan, W., Fan, J., Gou, J., Liu, B., Gide, E., Soar, J., Shen, B., Fazal-E-Hasan, S., Liu, Z., Zhang, P., Wang, P., Cui, X., Peng, Z., & Wang, Y. (2022). A comparative analysis of the principal component analysis and entropy weight methods to establish the indexing measurement. *PLoS ONE*, *17*(1 January), 1–26. https://doi.org/10.1371/journal.pone.0262261

Zheng, J., & Rakovski, C. (2021). On the application of principal component analysis to classification problems. *Data Science Journal*, *20*(1). https://doi.org/10.5334/dsj-2021-026