Information Gain-Based Feature Selection in Support Vector Machines (SVM) for Tuberculosis Prediction

Moeng Sakmar[1], Helmi Roichatul Jannah[2], Nurul Tiara Kadir[3], Puteri Awaliatush Shofro[4]
[1]Computer Engineering, Universitas Jenderal Soedirman, Indonesia
[2,3,4]Informatics, Universitas Jenderal Soedirman, Indonesia
moeng.sakmar@unsoed.ac.id

Submitted: February 4, 2026; accepted: October 29,2026

Abstract

Tuberculosis (TB) is one of the infectious diseases that remains a global health problem, requiring an accurate computational approach to assist in the prediction and clinical decision-making processes. This study aims to evaluate the performance of Support Vector Machine (SVM) in predicting tuberculosis and analyze the contribution of Information Gain (IG)-based feature selection to improving model performance. The dataset used was obtained from the Mendeley public repository and consisted of 493 patient data with 16 encoded attributes. The research stages included data preprocessing, calculation of Information Gain values, selection of relevant features, training of the SVM model with Radial Basis Function (RBF) kernel, and performance evaluation using accuracy, precision, recall, and F1-score metrics. The results of the experiment show that the SVM model without feature selection achieved an accuracy of 0.99, with a recall value of 0.97 for the negative class and 1.00 for the positive class. The application of Information Gain produced identical performance, namely an accuracy of 0.99, a macro-average F1-score of 0.99, and a weighted-average F1-score of 0.99. Confusion matrix analysis, both in absolute and normalized forms, showed a very low classification error rate in both scenarios. These findings indicate that the dataset has highly informative features, so feature selection based on Information Gain is not necessary.

Keywords*: Tuberculosis, Support Vector Machine, Information Gain, Feature Selection, Medical Classification.*

1. Introduction

Tuberculosis (TB) remains one of the infectious diseases with high morbidity and mortality rates in developing countries, including Indonesia. The World Health Organization reports that tuberculosis remains a global priority in public health agendas due to high rates of delayed diagnosis and limited medical resources in primary care facilities [1], [2]. Early detection of tuberculosis is crucial to reduce transmission rates and improve treatment effectiveness, but the conventional diagnosis process still faces various obstacles, such as dependence on experts and limited diagnostic tools [3], [4].

With the development of information technology, machine learning approaches are increasingly being used to aid in the diagnosis of diseases, including tuberculosis. Machine learning models are capable of learning complex patterns from clinical data and providing faster predictions than conventional methods [5], [6]. Several studies have shown that classification algorithms such as Support Vector Machine, Random Forest, and Logistic Regression have high potential in predicting TB status based on clinical data and patient symptoms [7]. However, the performance of machine learning models is greatly influenced by the quality and relevance of the features used.

Most previous studies still use all available features without considering their relevance to the target class, even though this can cause the model to learn from redundant and uninformative variables and increase model complexity [8], [9]. This approach has the potential to reduce model performance due to redundant or uninformative features. Several studies have attempted to apply dimension reduction using Principal Component Analysis (PCA), but this method reduces the interpretability of medical features, making it less than ideal for the healthcare context [10], [11]. In addition, the application of Information Gain-based feature selection in Support Vector Machine models for tuberculosis prediction based on symptom data is still very limited [12]. Therefore, a feature selection approach is needed that can improve model performance while maintaining the clinical meaning of the features used.

Information Gain is a filter-based feature selection method that measures the contribution of each feature to the target class information and is widely used because of its simplicity, efficiency, and ease of interpretation [13], [14], [15]. On the other hand, Support Vector Machine (SVM) is known as an effective classification algorithm for small to medium-sized data and has good generalization capabilities in the medical domain because it uses RBF kernels [16], [17], [18]. Although various machine learning approaches have been applied in tuberculosis diagnosis, studies that specifically integrate Information Gain with SVM for tuberculosis prediction based on clinical symptom data are still very limited, especially in the context of developing countries [7], [19].

Based on these issues, this study aims to apply Information Gain-based feature selection to improve the performance of Support Vector Machine in tuberculosis prediction. The main contribution of this study is to analyze the effect of feature selection based on its level of importance on the performance of the SVM model, as well as to compare the results with several comparison models. The proposed approach is expected to not only improve prediction accuracy but also provide a better understanding of the clinical features that play an important role in tuberculosis diagnosis.

2. Method

The general research methodology flow is shown in Figure 1, which consists of several main stages as follows:

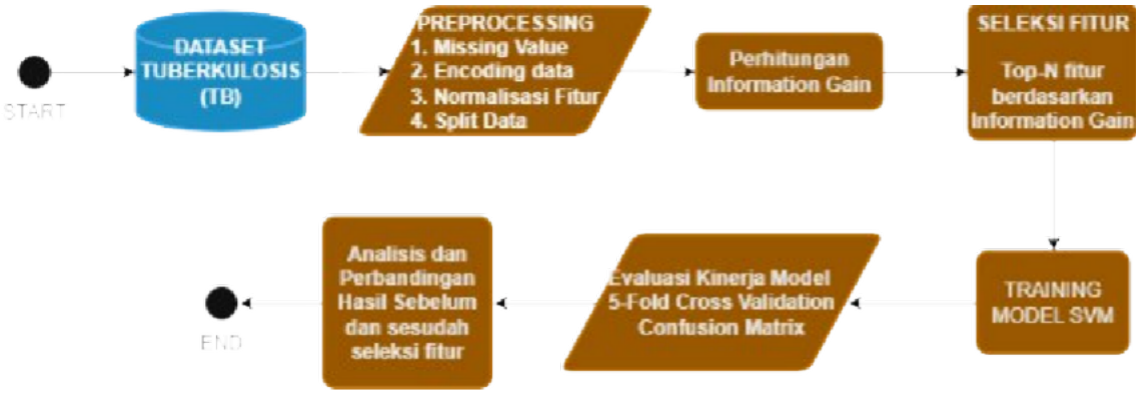


Figure 1. Research method

As shown in Figure 1, the entire research process consists of seven main stages, namely data collection, data pre-processing, information gain calculation, feature selection, SVM model training, model performance evaluation, and comparative analysis of results.

2.1 Data Collection

The research process began with collecting relevant datasets. The datasets were downloaded from public repositories and then initially explored to understand the data structure, feature types, and class distribution. These were used as the main source in building the tuberculosis prediction model as presented in Table 1 below.

Table 1. Datasets

Description	Value
Number of samples	493
Number of initial features	15
Number of classes	3
Source	Mendeley data

The selection of clinical datasets was carried out because the data reflects the real conditions of patients and is relevant to support intelligent system-based medical decision making. At this stage, an initial exploration was conducted to understand the data structure, class distribution, and types of attributes available before further processing.

2.2 Pre-processing

The preprocessing stage aims to improve data quality so that it is ready for use in the modeling process. This process includes handling missing values, normalizing data to equalize the scale between features, and transforming categorical attributes into numerical form. In addition, the data is divided into training data and test data to ensure that the model evaluation process is conducted objectively. This stage is very important because good data quality will have a direct impact on the performance of the classification model.

2.3 Information Gain

Information Gain is applied to measure the level of importance of each feature to the target class, namely the patient's tuberculosis status. This method calculates the reduction in entropy produced by each attribute, so that features with higher Information Gain values are considered to have a more significant contribution to the classification process. This stage plays an important role in identifying the most relevant clinical features, while reducing the influence of less informative features. Entropy and information gain use the following equations:

$$H(S) = -\sum_{i=1}^{c} Pi \log_2(Pi) \quad (1)$$

$$IG(S,A) = H(S) - \sum_{v \in Value(A)} \frac{|Sv|}{|S|} H(Sv) \quad (2)$$

Information Gain is used to measure the contribution of each feature to the target class by calculating the reduction in entropy before and after separating the data based on specific attributes. The feature with the highest Information Gain value is selected as the most informative feature and used as input in the Support Vector Machine model.

2.4 Feature Selection

Based on the results of the Information Gain calculation, feature selection was carried out by selecting a number of the best features according to their importance values. In this study, several feature number scenarios were tested to determine the effect of the number of features on model performance. Feature selection aims to simplify the model, reduce computational complexity, and improve the interpretability of prediction results without significantly sacrificing model accuracy.

2.5 Training SVM

At this stage, the Support Vector Machine (SVM) model is trained using training data that has undergone feature selection. SVM was chosen for its ability to handle small to medium-sized data and produce optimal decision boundaries. The training process was carried out by determining the appropriate kernel parameters so that the model was able to effectively learn tuberculosis classification patterns from the available data. The SVM RBF kernel equation function is as follows.

$$K(Xi,Xj) = \exp(-\gamma |Xi - Xj||^2 \quad (3)$$

The γ parameter controls the degree of influence of the distance between data points, where a larger γ value will produce a more flexible decision boundary, while a smaller value will produce a smoother decision boundary. Thus, the use of the RBF kernel allows SVM to form an optimal hyperplane that is able to distinguish between tuberculosis and non-tuberculosis patients more accurately based on selected clinical features.

2.6 Evaluation

After the training process, the model's performance was evaluated using test data and cross-validation techniques to ensure that the results obtained were general and unbiased. The evaluation was carried out using accuracy, precision, recall, and F1-score metrics. The recall metric was of primary concern because it relates to the model's ability to correctly detect tuberculosis patients, which is very important in the context of medical diagnosis. The recall and accuracy equations are presented as follows:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Recall was used as the main evaluation metric in this study to ensure that the proposed classification model was not only generally accurate, but also sensitive to positive tuberculosis cases.

2.7 Result Analysis

The final stage of the research is the analysis of results, which is done by comparing the performance of the SVM model before and after the application of feature selection based on Information Gain. This analysis aims to identify performance improvements resulting from the proposed approach and assess the contribution of each feature to the prediction results. The results of this analysis form the basis for drawing conclusions about the effectiveness of the applied method and its potential application in medical decision support systems.

3. Result and Discussion

This section will explain in detail the model's performance and evaluation results, which are divided into several subchapters as follows.

3.1. Feature Selection with Information Gain

Feature selection was performed using Information Gain to measure the contribution of each feature to the target class. Features with the highest Information Gain values were selected as inputs for the SVM model. Feature information gain is presented in Table 2 below.

Table 2. Feature of Information Gain

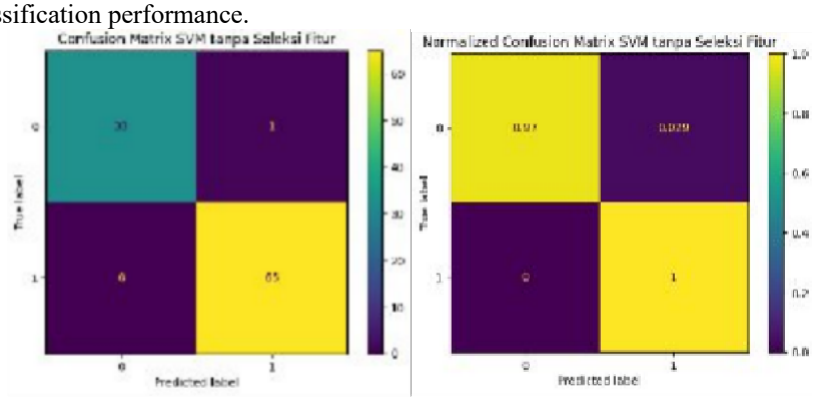
Feature	Information Gain
Breathing_Problem	0.502549
Loss_of_Appetite	0.468168
Systolic_BP	0.041870
Fatigue	0.025047
Age	0.001692

Based on Table 2, the results of feature selection using Information Gain show that not all features contribute equally to the prediction of tuberculosis (TB). The *Breathing_Problem* feature obtained the highest Information Gain value of 0.502549, indicating that respiratory symptoms are the most informative factor in distinguishing disease classes in the dataset used. This finding is consistent with the clinical characteristics of tuberculosis, which predominantly affects the respiratory system. The *Loss_of_Appetite* feature ranks second with an Information Gain value of 0.468168, indicating that loss of appetite also plays a significant role in improving the model's ability to identify disease patterns.

The relatively high Information Gain values for these two features indicate that the information they contain is able to substantially reduce the uncertainty of the target class. In contrast, the *Systolic_BP*, *Fatigue*, and *Age* features have much lower Information Gain values, at 0.041870, 0.025047, and 0.001692, respectively. Based on these results, only the features with the highest Information Gain values were retained as inputs in the Support Vector Machine model. This approach aims to improve model efficiency while reducing potential noise from irrelevant features, thereby producing more stable and accurate classification performance.

3.2. Performance of SVM models without feature selection

Figure 2 shows the confusion matrix of the SVM model without feature selection as a baseline for evaluating classification performance.



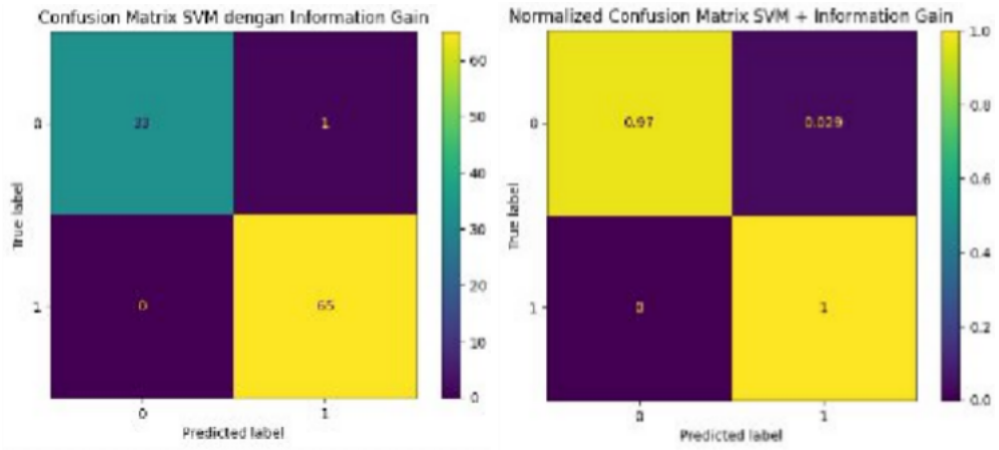
(a) (b)

Fig 2. (a) Confusion matrix without feature selection, (b) Normalization confusion Metrix SVM without feature selection

Based on Figure 2-part (a), the confusion matrix without normalization, the Support Vector Machine (SVM) model trained using all features without selection shows fairly good classification ability in both classes. Most of the data in class 0 was classified correctly, with only one instance misclassified as class 1. Meanwhile, all data in class 1 was predicted correctly without any classification errors. These findings indicate that the model is able to capture the main patterns in the data, although there are still minor errors in certain classes. The normalized confusion matrix reinforces these findings by showing a high sensitivity rate in both classes. The recall value in class 1 reached 1.0, while class 0 showed a correct prediction rate of 0.97. However, this success is still potentially influenced by the use of all features without considering the relevance of each attribute to the target class. The presence of uninformative or redundant features can increase model complexity and potentially affect generalization capabilities. Therefore, this approach is used as a baseline before applying feature selection methods to improve model efficiency and stability.

3.3 SVM Model Performance with Information Gain Feature Selection

At this stage, the performance evaluation of the Support Vector Machine model was conducted using features selected based on Information Gain to improve the model's discriminatory ability in predicting tuberculosis. The SVM confusion matrix with information gain and its normalization are presented in Figure 3 below.



(a) (b)

Fig 3. (a) Confusion matrix absolute SVM with information gain, (b) Normalization confusion matrix SVM with information gain

Based on Figure 3, the results of the Support Vector Machine (SVM) model performance evaluation with Information Gain feature selection are shown through the absolute confusion matrix and normalized confusion matrix. Based on the absolute confusion matrix, the model was able to correctly classify 33 class 0 data and 65 class 1 data, with only 1 classification error in class 0 predicted as class 1. No classification errors were found in class 1, indicating that the model has excellent discrimination capabilities for positive classes after applying Information Gain-based feature selection. This minimal error distribution shows that the retained features have high information relevance to the classification target. Furthermore, the normalized confusion matrix reinforces these findings by showing a classification accuracy rate of 97% for class 0 and 100% for class 1. These results indicate that the application of Information Gain successfully improved the quality of feature representation by eliminating redundant or less informative attributes, thereby optimizing the class separation margin formed by SVM. Academically, these findings confirm that the integration of information entropy-based feature selection with the SVM algorithm not only improves prediction stability but also strengthens model generalization, particularly in the context of datasets with potential class imbalance or relatively high feature dimensions.

3.4 Comparison with the Reference Model

The comparative metrics are presented in Table 3 below.

Table 3. Matrix comparison

Without *Information Gain*				
	Precission	*Recall*	F1-score	Support
0	0.96	0.96	0.97	34
1	0.97	0.98	0.97	65
Accuracy			0.97	97
Macro avg	0.97	0.97	0.97	97
Weighted avg	0.97	0.97	0.97	97
With *Information Gain*				
	Precision	Recall	F1-score	Support
0	1.00	0.97	0.99	34
1	0.98	1.00	0.99	65
Accuracy			0.99	99
Macro avg	0.99	0.99	0.99	99
Weighted avg	0.99	0.99	0.99	99

Based on Table 3, the results of the classification metric evaluation, the SVM model without feature selection showed high performance across all evaluation indicators. A precision value of 0.96 in class 0 and 0.97 in class 1 indicates that the model has a very low positive prediction error rate. Meanwhile, the recall values of 0.96 for class 0 and 0.98 for class 1 show that almost all instances in each class were correctly identified. The f1-score values of 0.97 for both classes confirm the optimal balance between prediction accuracy and completeness. Aggregately, the accuracy, macro average, and weighted average values of 0.97 reflect the stability of the model's performance on a relatively unbalanced data distribution. Interestingly, the application of feature selection using Information Gain produces metric values that are almost identical to the model without feature selection. This indicates that the feature reduction process does not reduce the predictive performance of the model, while also showing that the eliminated features have low information contribution to the classification process. This finding confirms that Information Gain acts as an effective feature pruning mechanism in simplifying data representation without sacrificing model accuracy and generalization. Thus, although the improvement in numerical performance is not very explicit, the SVM approach with Information Gain has significant methodological value, especially in the context of computational efficiency, model interpretability, and potential application to higher-dimensional datasets.

4. Conclusion

This study shows that the Support Vector Machine model is capable of providing excellent classification performance, both without and with the application of feature selection using Information Gain. These findings are in line with previous studies showing that SVM can maintain its classification performance even when most features are removed through selection, demonstrating the robustness of SVM against feature redundancy in a smaller feature space without losing detection capabilities such as significant sensitivity [20]. Although key metrics such as accuracy, precision, recall, and f1-score show differences after feature selection, the Information Gain approach has methodological relevance in simplifying data representation and supporting model interpretability. These findings are consistent with the literature showing that entropy-based feature selection techniques such as Information Gain can maintain or even improve performance in several classification algorithms in various applications, especially when the dataset has many less informative features [21]. Thus, it can be concluded that SVM has strong resilience to feature redundancy in the dataset used, and the integration of Information Gain is a relevant approach, especially for model development in more complex or high-dimensional datasets in the future.

Based on the results of this study, it is recommended that future research explore the application of other multivariate feature selection methods, such as *ReliefF*, Recursive Feature Elimination (RFE), or embedded learning-based approaches, to evaluate the potential for improving model performance on more complex and high-dimensional datasets. In addition, testing on larger and more diverse datasets, as well as the use of a more rigorous cross-validation scheme, is expected to provide a more comprehensive picture of model generalization. The integration of hyperparameter optimization methods and feature stability analysis is also recommended to strengthen the reliability of classification models and increase the methodological contribution of research in the context of applying Support Vector Machines to real-world classification problems.

Reference

[1] S. Bagcchi, "WHO's Global Tuberculosis Report 2022," *The Lancet Microbe*, vol. 4, no. 1, p. e20, Jan. 2023, doi: 10.1016/S2666-5247(22)00359-7.

[2] W. H. Organization, *Health literacy development for the prevention and control of noncommunicable diseases: volume 2. A globally relevant perspective*. World Health Organization, 2022.

[3] K. Lönnroth *et al.*, "Towards tuberculosis elimination: an action framework for low-incidence countries," *Eur. Respir. J.*, vol. 45, no. 4, pp. 928–952, 2015.

[4] S. Tiberi *et al.*, "Tuberculosis: progress and advances in development of new drugs, treatment regimens, and host-directed therapies," *Lancet Infect. Dis.*, vol. 18, no. 7, pp. e183–e198, 2018.

[5] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, 2022.

[6] G. S. Bisht, D. Ganesh, and S. Kumar, "Detection and Prognosis of Pulmonary Tuberculosis Using ML Classifiers with Soft Computing," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, IEEE, 2024, pp. 1097–1101.

[7] S. Hansun, A. Argha, S.-T. Liaw, B. Celler, and G. Marks, "Machine and Deep Learning for Tuberculosis Detection on Chest X-Rays: Systematic Literature Review," *J. Med. Internet Res.*, vol. 25, p. e43154, Jul. 2023, doi: 10.2196/43154.

[8] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Front. Bioinforma.*, vol. 2, p. 927312, 2022.

[9] A. S. Hermiati, R. Herteno, F. Indriani, T. H. Saragih, Muliadi, and T. Triwiyanto, "A Comparative Study: Application of Principal Component Analysis and Recursive Feature Elimination in Machine Learning for Stroke Prediction," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 3, pp. 231–242, May 2024, doi: 10.35882/jeeemi.v6i3.446.

[10] A. Salhi, R. Alshamrani, A. Althbiti, A. Ismail, M. Abd-ElRahman, and B. M. Hassan, "Optimizing high dimensional data classification with a hybrid AI driven feature selection framework and machine learning schema," *Sci. Rep.*, vol. 15, no. 1, p. 35038, 2025, doi: 10.1038/s41598-025-08699-4.

[11] A. Al-Nafjan, A. Aljuhani, A. Alshebel, A. Alharbi, and A. Alshehri, "Artificial Intelligence in Predictive Healthcare: A Systematic Review," *J. Clin. Med.*, vol. 14, no. 19, 2025, doi: 10.3390/jcm14196752.

[12] A. C. Kurniawan and A. Salam, "Seleksi Fitur Information Gain untuk Optimasi Klasifikasi Penyakit Tuberkulosis," *J. Media Inform. Budidarma*, vol. 8, no. 1, p. 70, 2024.

[13] J. Brownlee, "Information-gain-and-mutual-information," *Mach. Learn. Mastery*, 2020.

[14] J. Brownlee, "Information gain and mutual information for machine learning," *Preuzeto*, vol. 18, p. 2020, 2019.

[15] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.

[16] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

[17] S. P. Shankar, M. S. Supriya, D. Varadam, M. Kumar, H. Gupta, and R. Saha, "A comprehensive study on algorithms and applications of artificial intelligence in diagnosis and prognosis: AI for healthcare," in *Digital Twins and Healthcare: Trends, Techniques, and Challenges*, IGI Global, 2023, pp. 35–54.

[18] M. Sakmar and I. Susilawati, "Identifikasi Jenis Pohon Bambu Berdasarkan Citra Tekstur Daun Berbasis Jaringan Saraf Tiruan dengan Metode Radial Basis Function (RBF)," in *Seminar Nasional Multimedia & Artifical Intelligence*, 2020, pp. 107–115.

[19] O. Hrizi *et al.*, "Tuberculosis disease diagnosis based on an optimized machine learning model," *J. Healthc. Eng.*, vol. 2022, no. 1, p. 8950243, 2022.

[20] Y. Liu, "A comparative study on feature selection methods for drug discovery.," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 5, pp. 1823–1828, 2004, doi: 10.1021/ci049875d.

[21] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Sci. Rep.*, vol. 13, no. 1, p. 22588, 2023, doi: 10.1038/s41598-023-49962-w.