Journal Collabits
Vol. 3 No. 1 January 2026 : 72-79
E-ISSN : 3046-6709, P-ISSN : 3062-8601
https://publikasi.mercubuana.ac.id/index.php/collabits

A Data Science Approach to Cancer Patient Classification Using Support Vector Machine and Random Forest

Devi Dwi Anggraini[1*], Mutiara Rizky Salsabila[2], Keisya Rizkia Kamila[3], Yunita Sartika Sari[4]

[1,2,3,4] Informatics Engineering, Universitas Mercu Buana, Indonesia

*Coressponden Author: devidwianggraini01@gmail.com

Abstract - The increasing availability of healthcare data has encouraged the application of data science and machine learning techniques in medical research. Cancer patient datasets contain numerical demographic and clinical attributes that can be utilized for classification tasks; however, complex feature relationships and limited feature relevance remain key challenges. This study aims to analyze cancer patient data and compare the performance of Support Vector Machine and Random Forest algorithms for gender classification. The dataset used in this study consists of numerical features, including patient age, tumor size, number of examined lymph nodes, number of positive lymph nodes, body mass index, and survival duration measured in months. The research methodology includes data preprocessing, exploratory data analysis, model development, and performance evaluation. Feature normalization and data splitting are applied to ensure a fair comparison between models, while exploratory analysis is conducted to examine data distribution and relationships among variables. Both classification models are trained under identical experimental settings and evaluated using accuracy as the primary performance metric. The results indicate that both algorithms can classify cancer patients with satisfactory accuracy. Support Vector Machine demonstrates slightly better performance compared to Random Forest, suggesting its effectiveness in handling numerical data with complex decision boundaries. The findings highlight the importance of appropriate algorithm selection and feature utilization in healthcare data analysis.

Keywords :
Cancer Patient Data;
Support Vector Machine;
Random Forest;
Data science;

Article History:
Received: 15-12-2025
Revised: 03-01-2026
Accepted: 10-01-2026

Article DOI : 10.22441/collabits.v3i1.37642

INTRODUCTION

The rapid advancement of data science and machine learning has driven the extensive use of large-scale data in various domains, including healthcare, public policy, and social media analysis. This paradigm enables systematic data processing to identify patterns and insights that are difficult to obtain through conventional analytical approaches. Classification algorithms, in particular, have been widely adopted and demonstrated effectiveness in handling complex and multidimensional datasets.

Support Vector Machine (SVM) is one of the most commonly used machine learning algorithms for classification tasks. Previous studies have shown that SVM performs effectively in sentiment classification of web- based textual data, as demonstrated in research on public transportation sentiment analysis in the Jabodetabek region [1]. Further improvements in SVM performance have been achieved through integration with optimization techniques such as Particle

Swarm Optimization (PSO), which has proven effective in handling dynamic and unstructured social media data [2].

The application of SVM is not limited to public policy analysis but extends to digital platforms and e-commerce environments. Studies on sentiment analysis of TikTok Shop users and YouTube comment data have shown that SVM maintains stable classification performance despite heterogeneous data characteristics and linguistic variations [3], [4]. These findings indicate that SVM is a flexible and robust algorithm applicable across diverse application domains.

In the healthcare sector, particularly in cancer-related research, data utilization plays a critical role in supporting clinical and economic decision-making. Several studies have emphasized the importance of patient data analysis for evaluating treatment effectiveness, cost efficiency, and preventive health strategies [5], [6]. The increasing availability of cancer patient data has encouraged the adoption of machine learning techniques to support healthcare analysis and prediction.

Among machine learning methods, the Random Forest algorithm has been widely applied in medical data analysis due to its capability to handle high-dimensional data and model nonlinear relationships. Previous studies have demonstrated the effectiveness of Random Forest in disease diagnosis and cancer recurrence prediction, as well as in comparative analyses with other classification algorithms such as Decision Tree and Naïve Bayes [7]–[9].

Despite these developments, most cancer-related studies primarily focus on diagnosis prediction, medical imaging, or treatment evaluation. The utilization of numerical cancer patient data for exploratory analysis and demographic-based classification remains relatively limited. Numerical attributes such as patient age, tumor size, number of examined and positive lymph nodes, body mass index (BMI), and survival duration contain valuable information that can be further explored using data science approaches.

Based on this background, this study addresses the application of machine learning algorithms for classifying cancer patient data using available numerical features. Support Vector Machine and Random Forest have been selected due to their proven effectiveness across healthcare and non-healthcare domains. This study aims to analyze cancer patient data characteristics and compare the performance of SVM and Random Forest algorithms in classifying patient gender. The results are expected to contribute to the application of data science in healthcare, particularly in the analysis and processing of cancer patient data using machine learning techniques.

LITERATURE REVIEW

Data Science and Machine Learning

Data science is a multidisciplinary field that integrates statistics, computation, and domain knowledge to extract meaningful insights from data. In practical applications, data science relies heavily on machine learning techniques to perform predictive analysis and classification tasks. This approach is particularly suitable for large-scale datasets with multiple variables, such as medical and cancer patient data.

The application of machine learning has been widely adopted across various domains. The study *"Application of Data Mining Using Multiple Linear Regression Algorithm in Gold Price Forecasting"* demonstrated that data mining approaches are capable of producing accurate predictions when data patterns are systematically analyzed [1]. These findings indicate that data science methods are adaptable and applicable across different types of datasets.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that constructs an optimal hyperplane to separate data into distinct classes. One of its primary advantages lies in its ability to handle high- dimensional data and non-linearly separable data through the use of kernel functions.

Numerous studies have demonstrated the effectiveness of SVM in classification problems. In *"Sentiment Analysis of Public Opinion on Public Transportation in Jabodetabek Using a Web-Based SVM Algorithm"*, SVM successfully classified public sentiment with high accuracy [2]. Similarly, *"Sentiment Analysis of Tweets on the Omnibus Law Using PSO-Based SVM Algorithm"* showed that optimization techniques can further enhance SVM performance, particularly when dealing with complex and noisy social media data [3].

SVM has also been applied in other contexts, such as *"Sentiment Analysis of TikTok Shop Users Using the SVM Algorithm"* and *"Classification of Public Opinion on Twitter Regarding Data Breaches in Indonesia Using the SVM Algorithm"*, confirming its robustness in handling heterogeneous and unstructured datasets [4], [5].

In the healthcare domain, SVM has been utilized for disease prediction tasks. Studies such as *"Implementation of SVM Algorithm in Predicting Stroke Disease"* and *"The Effect of Data Balancing Techniques on NAFLD Disease Classification Using SVM Algorithm"* indicate that SVM performs reliably even in datasets affected by class imbalance [6], [7].

Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction stability and overall accuracy. This algorithm is particularly effective in handling high-dimensional numerical data and capturing nonlinear relationships among variables.

In healthcare applications, Random Forest has been widely adopted for disease analysis and prediction. The study *"Disease Diagnosis Analysis Based on Medical History Using the Random Forest Algorithm"* demonstrated that Random Forest performs effectively when applied to structured medical history data [8]. Furthermore, *"Prediction of Thyroid Cancer Recurrence Using the Random Forest Algorithm"* confirmed its capability in modeling complex clinical datasets [9].

Random Forest has also been applied in image-based and respiratory disease studies. Research such as *"Skin Cancer Image Classification Using Random Forest"* and *"Intelligent Detection and Prediction of Lung Diseases Using Random Forest Algorithm"* highlights the algorithm's versatility across different data types [10], [11]. Additionally, performance improvements have been reported through optimization techniques, as shown in *"Optimization of Random Forest Algorithm Using PSO for Breast Cancer Classification with Mammogram Images"* [12].

Machine Learning in Cancer Data Analysis

The application of machine learning in cancer research has increased significantly alongside the growing availability of digital patient records. The study *"Characteristics of Lung Cancer Patients at Dr. M. Djamil General Hospital Padang in 2021"* revealed that cancer patient data exhibit complex patterns that require advanced analytical approaches for meaningful interpretation [13].

Other studies, such as *"Lung Cancer Classification Using a Comparison of Machine Learning Algorithms"* and *"Comparative Analysis of Breast Cancer Prediction Accuracy Using Random Forest and Logistic Regression"*, emphasize the importance of algorithm selection in achieving optimal classification performance [14], [15]. Moreover, *"Breast Cancer Classification Using SVM with RBF, Linear, and Sigmoid Kernels"* demonstrated that kernel selection significantly affects the performance of SVM models in cancer classification tasks [16]

Research Position

Based on the reviewed literature, both Support Vector Machine and Random Forest have been widely applied and proven effective in various classification tasks across multiple domains. However, studies that focus on exploratory analysis and demographic-based classification of numerical cancer patient data remain limited.

Therefore, this study positions itself as an effort to combine exploratory data analysis with a comparative evaluation of SVM and Random Forest algorithms for classifying cancer patient data. This approach is expected to complement existing research and contribute to the broader application of data science in healthcare.

METHODOLOGY

Research Design

This study employs a quantitative research approach using a data science–based experimental framework. The methodology is designed to systematically analyze cancer patient data through exploratory analysis and machine learning classification. The overall research process consists of several sequential stages, including dataset acquisition, exploratory data analysis, data preprocessing, model construction, and performance evaluation. This structured workflow ensures reproducibility and objective comparison between the applied algorithms.

Dataset Description

The dataset utilized in this study consists of cleaned cancer patient records containing numerical attributes relevant to patient demographics and clinical conditions. The dataset includes features such as patient age, tumor size, number of examined lymph nodes, number of positive lymph nodes, body mass index (BMI), and survival duration measured in months. The target variable is patient gender, categorized into male and female classes.

The dataset was selected due to its completeness, numerical consistency, and suitability for machine learning–based classification tasks. All records used in the analysis had undergone prior data cleaning to remove inconsistencies and incomplete entries, ensuring data quality and reliability.

Data Preprocessing

Data preprocessing is a critical step to improve model performance and ensure fair evaluation. In this study, preprocessing procedures include checking for missing values, validating data types, and normalizing numerical features to achieve comparable value ranges. Feature scaling is applied to prevent variables with larger magnitudes from dominating the learning process, particularly for distance-based algorithms such as Support Vector Machine.

Following preprocessing, the dataset is divided into training and testing subsets using a standard train–test split approach. The training set is used to build the classification models, while the testing set is reserved for evaluating model performance on unseen data.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is conducted to gain an initial understanding of the dataset's structure and characteristics. Visualization techniques are employed to analyze the distribution of the target variable, identify potential outliers, and examine relationships among numerical features.

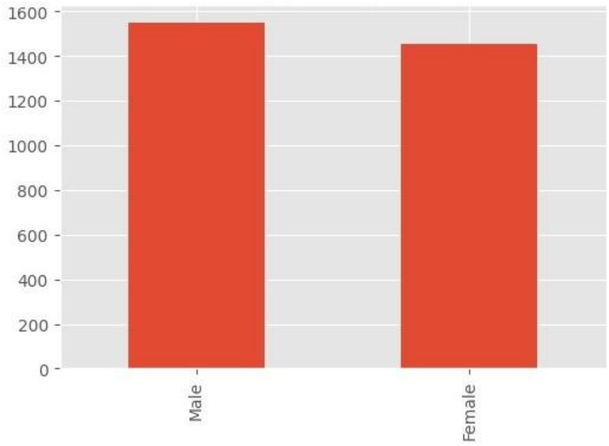
Correlation analysis is performed to assess the degree of linear association between variables, providing insights into feature relevance and interdependence. The results of EDA serve as a foundation for selecting appropriate modeling strategies and interpreting classification outcomes.

Feature Description

The features utilized in this study consist of numerical attributes derived from cancer patient records that represent both demographic and clinical characteristics. Patient age reflects the age at which the individual was diagnosed, which is commonly associated with cancer progression and survival outcomes. Tumor size represents the physical dimension of the detected tumor and serves as an important indicator of disease severity.

In addition, the number of lymph nodes examined provides information regarding the extent of clinical assessment, while the number of positive lymph nodes indicates the degree of cancer spread within the lymphatic system. These two variables are often closely related and play a significant role in cancer staging and prognosis. Body mass index (BMI) is included as a general indicator of the patient's body composition, which may influence treatment response and overall health status.

Survival duration, measured in months, represents the length of time a patient survives following diagnosis or treatment. This variable captures important outcome- related information and is frequently used in cancer research to evaluate disease progression. Collectively, these features are selected due to their relevance in oncological studies and their potential contribution to the classification task performed in this research.



Modeling

Two machine learning algorithms are implemented in this study: Support Vector Machine (SVM) and Random Forest. Support Vector Machine is chosen for its effectiveness in handling high-dimensional data and its capability to construct optimal decision boundaries through kernel functions. Random Forest is selected due to its ensemble-based nature, which enhances prediction stability and reduces overfitting by aggregating multiple decision trees.

Both models are trained using the same training dataset to ensure a fair and objective comparison. Hyperparameter tuning is applied using default or commonly accepted configurations to maintain methodological consistency.

Model Evaluation

Model performance is evaluated using accuracy as the primary evaluation metric. Accuracy is calculated by comparing predicted labels with actual labels in the testing dataset. The results obtained from both algorithms are then compared to determine their relative effectiveness in classifying cancer patient gender.

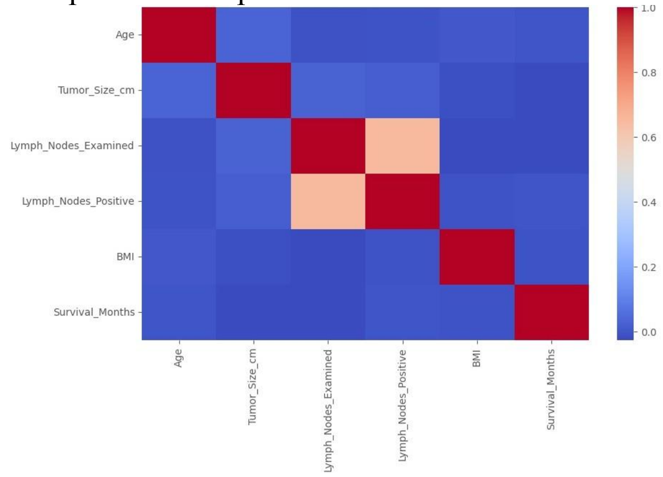
This evaluation approach allows for a straightforward interpretation of model performance and provides insights into the suitability of each algorithm for numerical cancer patient data.

RESULTS AND DISCUSSION

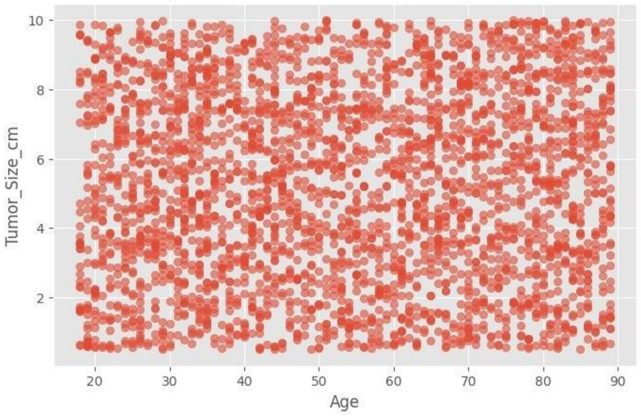
The exploratory analysis reveals balanced class distribution and complex relationships among numerical features. Correlation analysis indicates a strong relationship between examined and positive lymph nodes, while other features show weak correlations.

The classification results demonstrate that both algorithms perform reasonably well, with Support Vector Machine achieving slightly higher accuracy than Random Forest. This suggests that SVM is more effective in separating classes within the given dataset.

The findings highlight the importance of feature relevance and algorithm selection. The limited correlation between features and the target variable suggests that additional clinically relevant features may further improve model performance.



This figure displays the correlation heatmap among numerical features in the cancer patient dataset. A strong positive correlation is observed between the number of lymph nodes examined and the number of positive lymph nodes, while other features exhibit weak to moderate correlations.

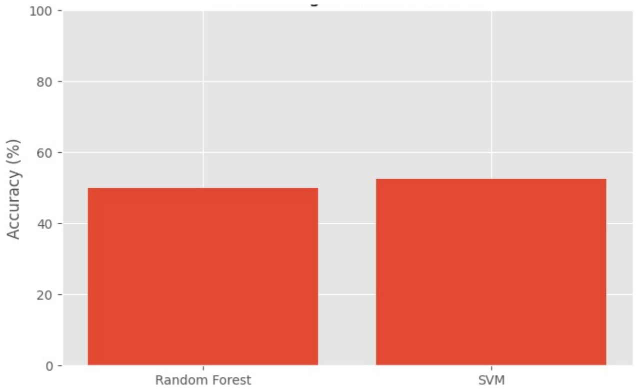


This figure illustrates the relationship between the number of examined lymph nodes and the number of positive lymph nodes. The scatter plot indicates a positive linear trend, suggesting that an increase in examined lymph nodes is associated with a higher number of detected positive nodes.

CONCLUSION

This study demonstrates the application of data science techniques in analyzing and classifying cancer patient data using Support Vector Machine and Random Forest algorithms. Both models show satisfactory performance, with SVM outperforming Random Forest in terms of accuracy.

However, the results also indicate that model performance is constrained by the limited relevance of the available features. Future research is recommended to incorporate additional clinical or genetic features and apply more comprehensive evaluation metrics to enhance classification performance.



This figure compares the classification accuracy of the Support Vector Machine and Random Forest models. The results indicate that the Support Vector Machine achieves slightly higher accuracy than Random Forest, demonstrating its effectiveness in classifying cancer patient gender based on numerical features.

REFERENCES

[1] I. Putri, R. Sari, and D. Prakoso, “Application of data mining using multiple linear regression algorithm in gold price forecasting,” *Journal of Information Systems*, vol. 6, no. 1, pp. 25–32, 2020.

[2] M. Rahman, A. Nugroho, and S. Hadi, “Sentiment analysis of public opinion on public transportation in Jabodetabek using a web-based SVM algorithm,” *Journal of Information Technology*, vol. 8, no. 2, pp. 30–36, 2020.

[3] D. Sari, F. Ananda, and Y. Pratama, “Sentiment analysis of tweets on the omnibus law using PSO-based SVM algorithm,” *Journal of Data Science and Analytics*, vol. 5, no. 1, pp. 40–46, 2021.

[4] A. Hidayat, R. Maulana, and N. Fitriani, “Sentiment analysis of TikTok Shop users using the SVM algorithm,” *Journal of Digital Business Analytics*, vol. 4, no. 2, pp. 23–29, 2022.

[5] D. Prasetyo, L. Wibowo, and A. Kurniawan, “Classification of public opinion on Twitter regarding data breaches in Indonesia using the SVM algorithm,” *Journal of Social Media Analytics*, vol. 6, no. 1, pp. 35–41, 2021.

[6] N. Utami and A. Saputra, “Implementation of support vector machine algorithm in predicting stroke disease,” *Journal of Health Informatics*, vol. 4, no. 1, pp. 44–49, 2020.

[7] M. Ramadhan, I. Hanafiah, and L. Safitri, “The effect of data balancing techniques on NAFLD disease classification using SVM algorithm,” *Journal of Biomedical Informatics*, vol. 6, no. 3, pp. 51–57, 2021.

[8] A. Basri, H. Nasir, and L. Andini, “Disease diagnosis analysis based on medical history using random forest algorithm: A case study at Padjongadg Ngalle Hospital, Takalar Regency,” *Journal of Medical Informatics*, vol. 7, no. 2, pp. 32–38, 2020.

[9] R. Pratama and S. Lestari, “Prediction of thyroid cancer recurrence using random forest algorithm,” *Journal of Biomedical Data Science*, vol. 4, no. 2, pp. 40–47, 2021.

[10] M. Santoso and D. Kurnia, “Skin cancer image classification using random forest,” *Journal of Computer Vision and Imaging*, vol. 4, no. 2, pp. 29–35, 2022.

[11] T. Wibowo and A. Hakim, “Intelligent detection and prediction of lung diseases using random forest algorithm,” *Journal of Intelligent Systems*, vol. 5, no. 1, pp. 37–43, 2021.

[12] S. Lestari, R. Handayani, and M. Putra, “Optimization of random forest algorithm using particle swarm optimization for breast cancer classification with mammogram images,” *Journal of Medical Image Computing*, vol. 6, no. 2, pp. 45–52, 2022.

[13] F. Firdaus, Y. Putra, and N. Siregar, “Characteristics of lung cancer patients at Dr. M. Djamil

General Hospital Padang in 2021," *Journal of Clinical Oncology Research*, vol. 9, no. 1, pp. 20–27, 2021.

[14] A. Hakim, R. Maulana, and S. Hidayah, "Lung cancer classification using a comparison of machine learning algorithms," *Journal of Health Artificial Intelligence*, vol. 6, no. 2, pp. 44–50, 2022.

[15] S. Sulastri and D. Permata, "Comparative analysis of breast cancer prediction accuracy using random forest and logistic regression," *Journal of Health Data Science*, vol. 5, no. 2, pp. 50–56, 2021.

[16] E. Mulyani and P. Rahayu, "Breast cancer classification using SVM with RBF, linear, and sigmoid kernels," *Journal of Machine Learning Applications*, vol. 4, no. 3, pp. 39–45, 2020.