Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 2 Iss. 2 August 2025, pp: 94-103
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Comparison of Supervised Learning Algorithms for Cancer Prediction

Intan Adha Maharani[1*], Rifda Dwi Setiani[2], Raudhatul Khairiyah[3], Elfani Mardhatillah[4]

[1,2]Department of Information System, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[3]Departemen of Syari'ah Islamiyah, Faculty of Dirost Islamiyah, Al-azhar University, Egypt
[4]Departemen of Ushulluddin, Faculty of Dirost Islamiyah, Al-azhar University, Egypt

E-Mail: [1]12250325185@students.uin-suska.ac.id, [2]12250321086@students.uin-suska.ac.id, [3]raudhatulkhairiyyahxi7@gmail.com, [4]elfanimardhatillah@gmail.com

Received May 29th 2025; Revised Aug 18th 2025; Accepted Aug 30th 2025; Available Online Aug 31th 2025
Corresponding Author: Intan Adha Maharani

Abstract

This study focuses on the application of Machine Learning algorithms for cancer prediction using a classification dataset. Several algorithms were employed, including K-Nearest Neighbor (K-NN), Naive Bayes Classifier, Decision Tree, Random Forest, and Support Vector Machine (SVM). The primary goal of this research is to evaluate the performance of each algorithm to identify the best method for achieving high accuracy in cancer classification prediction. The experimental results reveal variations in performance among these algorithms. The evaluation was conducted using metrics such as accuracy, precision, recall, and F1-Score. Based on the analysis, Random Forest and Support Vector Machine demonstrated the best performance with the highest accuracy compared to other algorithms. Meanwhile, the Naive Bayes algorithm tended to exhibit lower predictive performance. This study emphasizes the importance of selecting the appropriate algorithm for implementing Machine Learning in medical applications, such as cancer prediction. With these findings, it is hoped that the identified methods can assist in clinical decision-making and improve the accuracy of early cancer diagnosis.

Keywords: Cancer Prediction, Decision Tree, Machine Learning, Naive Bayes, Support Vector Machine

1. INTRODUCTION

Cancer is one of the deadliest diseases in the world, characterized by the abnormal growth of cells that can damage body tissues and spread to other organs. According to the Global Cancer Observatory (GLOBOCAN) report in 2020, there were more than 19.3 million new cancer cases with 10 million deaths worldwide [1]. In Indonesia, the prevalence of cancer reaches 1.8 per 1,000 population, making it one of the most serious public health problems. This condition highlights the urgency of research in developing predictive technologies to support early cancer detection, improve recovery rates, and reduce mortality. One potential approach is supervised learning, which has been proven effective in classifying medical data. The application of supervised learning enables early cancer detection, thereby increasing the chances of recovery and reducing mortality rates [2]. In recent years, machine learning has emerged as a potential solution for cancer prediction and classification by identifying patterns in high-quality medical datasets [3].

The dataset used in this study includes demographic information (gender, age, marital status, number of children) as well as lifestyle factors (smoking habits, occupation, income level). However, cancer datasets generally face the problem of class imbalance, where the distribution of the positive class (cancer patients) is smaller compared to the negative class. This imbalance may lead the model to be more biased toward the majority class [4]. To address this issue, this study applies oversampling and the Synthetic Minority Oversampling Technique (SMOTE) to optimize the model's performance in detecting the minority class [5].

Previous studies have shown varying accuracy results in the application of supervised learning algorithms for cancer prediction. Nemade and Fegade (2023) reported accuracies of 92% for SVM and 85% for Naive Bayes on a breast cancer dataset [6][7]. Another study found that Decision Tree achieved an accuracy of 97%, while XGBoost with ensemble techniques obtained an AUC of 0.99. These findings

emphasize that algorithm performance is strongly influenced by the applied parameters. Therefore, this study highlights the importance of hyperparameter tuning to improve model accuracy and consistency [8][9].

Various supervised learning algorithms have been applied for cancer prediction, including K-Nearest Neighbor (K-NN), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). K-NN is simple and easy to understand but sensitive to the choice of parameter *k* [10]. NB performs well on datasets with simple distributions due to its probabilistic approach [19]. DT is easy to interpret but prone to overfitting [22]. RF combines multiple decision trees to improve accuracy and reduce variance [13]. SVM is effective for non-linear data with a clear margin [14]. Several previous studies have also demonstrated the effectiveness of these algorithms, such as the use of RF for lung cancer classification [4], SVM for breast cancer [7], and K-NN on medical datasets [8].

The aim of this study is to compare the performance of five supervised learning algorithms in cancer classification using accuracy, precision, and recall metrics. Additionally, this study emphasizes the use of the confusion matrix as an evaluation tool to illustrate the types of prediction errors. The research gap addressed is the limitation of previous studies, which focused only on one or two algorithms without an in-depth analysis of the trade-offs between accuracy, precision, and recall. Therefore, this study presents a more comprehensive comparative analysis.

The selection of Random Forest as one of the main focuses is based on its ability to reduce overfitting, provide stable results across diverse datasets, and offer interpretability through feature importance analysis [13]. By comparing RF with other algorithms, this study aims to provide a clearer understanding of the strengths and weaknesses of each approach in the context of cancer prediction.

2. MATERIAL AND METHOD

The flowchart illustrates the research process, from the data collection stage to the prediction stage. The purpose of this explanation is to provide a structured overview and assist the research in achieving its objectives. Figure 1 depicts the entire research process.

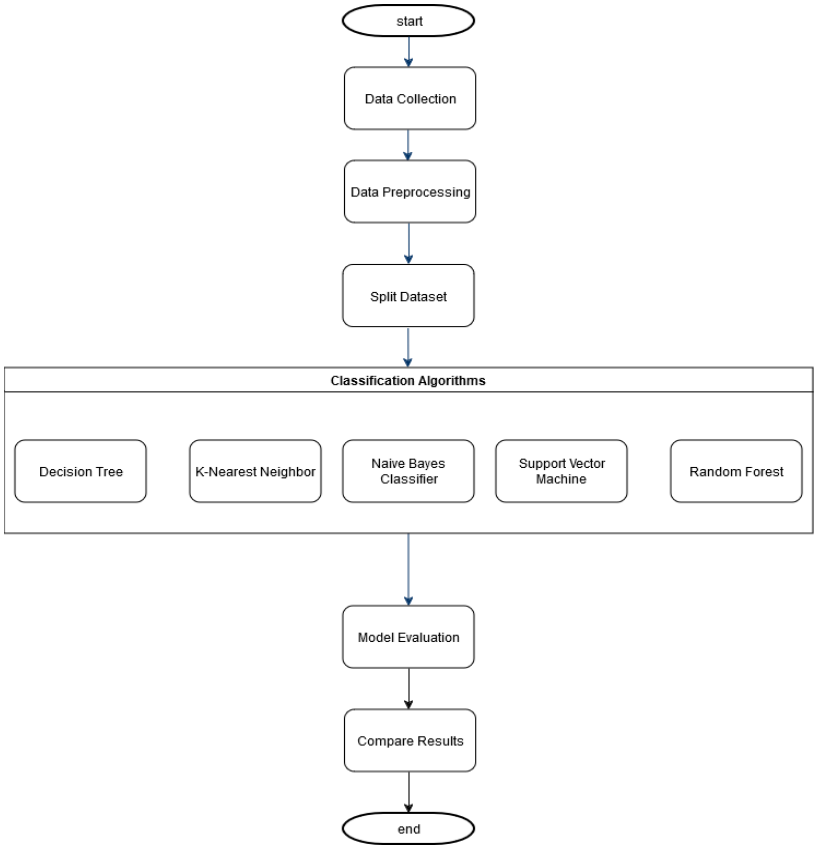

Figure 1. Flowchart of Research Methodology

2.1. Data Collecting

The cancer prediction data was obtained from the Kaggle website, where a university conducted a simulated survey to collect this dataset to study lung cancer risk factors. Demographic information, such as gender, age, marital status, and number of children, as well as lifestyle factors (e.g., smoking, occupation, years worked, income, social media use, and online gaming), and health status (e.g., cancer diagnosis), are

included in 1,000 entries. The cancer status label was virtually determined based on a combination of risk factors, and the survey was conducted in a multiple-choice format for respondent convenience. This fictitious dataset was created for educational and research purposes, featuring distribution patterns similar to the actual population.

2.2. Data Preprocessing

Preprocessing began with examining the data types to ensure no missing values were present. Duplicate data and irrelevant ID columns were removed. Categorical variables, such as Gender, Smoker, and Marital Status, were converted into numerical formats using label encoding or one-hot encoding. Numerical variables, such as Age and Years Worked, were normalized using Min-Max Scaling to range between 0 and 1 or standardized to a mean of 0 and a standard deviation of 1 if required. For the target variable (Cancer), oversampling or undersampling techniques were applied to address data imbalance. Subsequently, the dataset was divided into two subsets: 20% for testing and 80% for training. After this process, the data was ready for analysis and modeling.

2.3. Split Dataset

The data was split into a training set (80%) and a testing set (20%). The training dataset comprised 800 samples, while the testing dataset included 200 samples. While the target variable (cancer status) was separated as the label to be predicted, the feature variables consisted of 10 columns. This division ensured the model could be trained and evaluated with sufficient data.

2.4. K-Nearest Neighbor (K-NN)

K-NN is a supervised learning algorithm widely used for regression and classification problems [16]. Using geometric distance metrics as a standard, this algorithm classifies unlabeled data based on its closest distance to labeled data [17]. Proses K-NN menemukan jarak terdekat antara data pelatihan dan data yang akan diuji dalam ruang fitur [18]. While K-NN is advantageous for its simplicity and practicality, it has drawbacks, such as long processing times for large datasets and sensitivity to the value of the k parameter [16]. The K-NN equation is expressed in Equation 1.

$$d_{\text{Euclidian}} = \sqrt{\sum_{i=1}^{n} (x_{i2} - x_{i1})^2} \quad (1)$$

Here, dEuclidian represents the method for calculating the distance between two points in n-dimensional space. In this case, xi1 is a sample data point, and xi2 a test data point to be compared. n represents the number of dimensions or attributes used to determine the distance between xi1 and xi2. This Euclidean distance is often used in classification or clustering tasks.

2.5. Naive Bayes Classifier

Naive Bayes applies Bayes' theorem to determine the likelihood of each class when predicting data. It is part of the generative learning group as it models the input distribution for specific classes or categories [19]. The strength of Naive Bayes lies in its low training data requirement for parameter estimation [20]. However, its limitations include unrealistic independence assumptions among variables and the potential for zero probabilities affecting classification outcomes [21]. The Naive Bayes Classifier equation is expressed in Equation 2.

$$P(C|X) = \frac{P(C|X) \times P(C)}{P(X)} \quad (2)$$

Here, X represents the data with an unknown class, and C is the hypothesis of its class in probabilistic analysis. The posterior probability P(C|X) indicates the likelihood of C after considering X, calculated using Bayes' rule. This calculation incorporates prior probability P(C), likelihood P(X|C), and predictor prior probability P(X) as a normalization factor.

2.6. Decision Tree

Decision trees are techniques used for both classification and regression. In classification, each tree node represents a decision that predicts the class based on features, while in regression, nodes predict continuous values. Building a decision tree involves identifying the optimal nodes and splitting the training data into sub-nodes [22].

Decision Trees classify objects by partitioning the data into sets based on input variables [23]. The primary characteristics of the C4.5 algorithm include its ability to generate predictive models in the form of user-friendly rules [24]. Decision trees have a root node and internal nodes, or branches, which are used for classification and prediction. [25]. The main stages of C4.5 Decision Tree construction are:

1. Identifying the root attribute with the lowest entropy and highest gain values. Entropy is determined using Equation 3:

$$\text{Entropy}(y) = \sum_{i=1}^{n} - P_i \log_2 P_i \quad (3)$$

2. Calculating gain using Equation 4:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|s_1|}{|s|} * \text{Entropy}(Si) \quad (4)$$

In data analysis, S represents the set of cases with n as the number of partitions. The probability of each partition is calculated as pi, which is the number of cases in the i-th partition (|S1|) divided by the total number of cases (|S|). For attribute A, there are n partitions based on the number of cases in the set.

3. Create branches for each value.
4. Distribute each case into the branches.
5. Repeat the process for each branch until all cases belong to the same class.

2.7. Random Forest

The Random Forest method, an evolution of the Classification and Regression Tree (CART) method, employs bootstrap aggregating (bagging) and random feature selection [26]. RF combines multiple decision trees and selects features randomly at each iteration to improve model performance and reduce overfitting [5]. Methods like tree-based models such as Random Forest are built from samples of the dataset, with only a few features selected, and then determine the values that produce the best separation within the dataset [27]. Equations for RF are expressed in Equations 5 and 6 [28]:

$$\text{Entropy}(Y) = -\sum_{i} p(c|Y) \log^2 p(c|Y) \quad (5)$$

$$\text{Information gains (Yes)} = \text{Entropy}(Y) = -\sum_{v \varepsilon \text{Values}(a)} \frac{|Y_v|}{|Y_a|} \text{Entropy}(Y_v \quad (6)$$

P(c|Y) represents the proportion of cases in Y that belong to class c, Each case a has an associated value (Values(a)), and Yv refers to the subclass of Y corresponding to class v, while Ya is the overall value associated with class a. This explains how cases are grouped and classified based on relevant attributes.

2.8. Support Vector Machine (SVM)

SVM is a supervised machine learning model used for classification and regression tasks, especially in situations where there are two classes of data. In this process, two hyperplanes are used by SVM, and quadratic programming is applied to obtain the bias and weight parameters [18]. With high accuracy and memory efficiency, SVM is effective for large datasets. However, it requires a long training time, and handling overlapping classes is a challenge. Additionally, when the number of features in the training sample exceeds the number of features in the sample, its performance decreases.
The equation for SVM is expressed in Equation 7 [29].

$$f(x_d) = \sum_{i=1}^{ns} \alpha_i y_i \overrightarrow{x_i} \overrightarrow{x_d} + b \quad (7)$$

2.9. Model Evaluation

After the model is trained using the training data, the next step is to test its performance with the testing data to see how well it generalizes to new data. This is done using several metrics, including accuracy, which measures the proportion of correct predictions to the total data, precision, which measures the accuracy of positive predictions (i.e., how many positive predictions are relevant), and recall, which measures how well the model performs in detecting true positives.

The confusion matrix contains accuracy, precision, and recall values and is used to determine classification performance based on true or false outcomes [30]. Recall, precision, accuracy, and error rate are the four outputs of this formula. The confusion matrix has four important values: true positives (TP) and true negatives (TN) indicate that the model made correct predictions, while false positives (FP) and false negatives (FN) show that the model made incorrect predictions [18]. Figure 2 is a diagram of the Confusion Matrix.

2.10. Compare Results

After calculating the evaluation metrics, the next step is to compare the performance of the algorithms based on accuracy, precision, and recall, taking into account the application context. For example, if the goal is to detect all positive cases, precision is more important, and recall is prioritized. The confusion matrix is useful for identifying the types of errors made by each algorithm, and the F1 score helps maintain a balance between precision and recall.

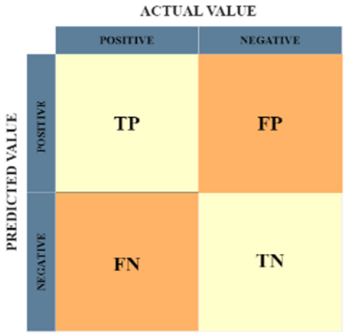

Figure 2. Confusion Matrix Diagram

3. RESULTS AND DISCUSSION

3.1. Data Preprocessing

Data preprocessing includes checking data types, removing missing values, ID columns, and duplicates, as well as converting categorical variables into numeric variables. Additionally, numeric variables are normalized or standardized, and oversampling or undersampling methods are used to address imbalances. The dataset is prepared for analysis and modeling, split into 20% test data (200 samples) and 80% training data (800 samples).

3.2. K-Nearest Neighbor (K-NN)

The dataset is loaded, checked, and cleaned by removing irrelevant columns, handling missing values, and converting categorical data to numeric format. Starting with splitting the data into training and test sets, the K-NN algorithm is used to predict class labels. The model is then tested and evaluated to assess its performance. Confusion matrix K-NN can be seen in Figure 3.

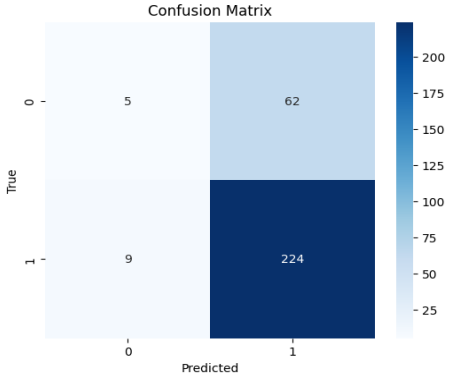

Figure 3. Confusion Matrix K-NN

Based on the confusion matrix in Figure 3, the K-NN algorithm shows good performance with an accuracy of 95.33%, recall of 97.82%, and precision of 96.14%, indicating that the model is able to correctly detect most positive cases. However, there are 9 incorrect results and 5 false negatives, indicating room for improvement.

3.3. Naïve Bayes

The Naive Bayes Classifier (NBC) model begins by cleaning and transforming the data into a numeric format. The data is then divided into labels (y) and features (X). Using hold-out validation, the data is split into training and test sets with a 70:30 ratio. The Gaussian Naive Bayes model is trained and tested using this method. To assess the model's ability to predict cancer status based on the available features, metrics such as accuracy, precision, and recall are used. Confusion matrix Naïve Bayes can be seen in Figure 4.

The model's performance is evaluated using the confusion matrix for the Naive Bayes algorithm. Although there are some errors, particularly in False Negative cases, the K-NN model successfully predicts 224 true positive cases and 62 true negative cases, with an accuracy of 95.33%, precision of 96.14%, and recall of 97.82%. This demonstrates the K-NN model's excellent ability to classify data.

3.4. Random Forest

The process of using the Random Forest algorithm begins with data preparation, including transforming categorical data into numerical values. The dataset is then split into features (X) and labels (y). To improve accuracy, the Random Forest model combines several decision trees. Cross-validation is used to evaluate model performance in predicting positive classes, with metrics such as accuracy, precision, and recall applied. The evaluation results show how effective the model is in making predictions. Confusion matrix Random Forest can be seen in Figure 5.

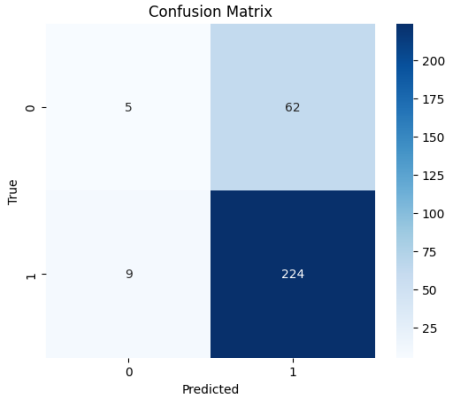

Figure 4. Confusion Matrix Naïve Bayes

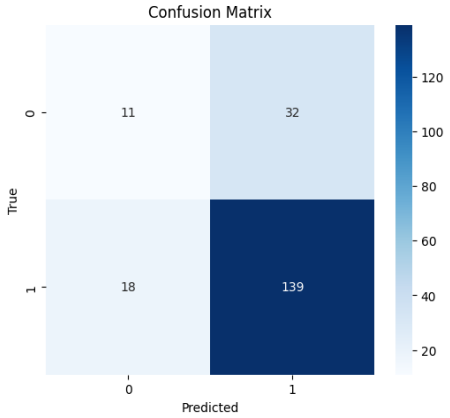

Figure 5. Confusion Matrix Random Forest

In the confusion matrix in Figure 5, the Random Forest model accurately predicted 139 positive cases and 11 negative cases; however, it incorrectly predicted 32 negative cases as positive and 18 positive cases as negative. The model shows an accuracy of 75.9%, precision of 80.93%, and recall of 90.53%, with a focus on recall for accurate cancer case detection.

3.5. Decision Tree

Data collection and preparation is the first step in building a Decision Tree. This includes handling missing values and transforming categorical data. Metrics such as the Gini index are used to select the best attributes, and the tree is built by recursively splitting the data until stopping criteria are met. Pruning is then performed to reduce complexity and prevent overfitting. After evaluating the model using a test set and metrics such as the confusion matrix, the model is used to predict new data. Confusion matrix Decision Tree can be seen in Figure 6.

In the Figure 6 Decision Tree algorithm, the confusion matrix shows an accuracy of 60.49%, precision of 59.98%, and recall of 71.67%. The model accurately predicts 226 positive cases and 15 negative cases but mispredicts 52 positive cases and 7 negative cases. To reduce errors in detecting positive cases, the evaluation focuses on improving recall. Decision Tree cancer prediction can be seen in Figure 7.

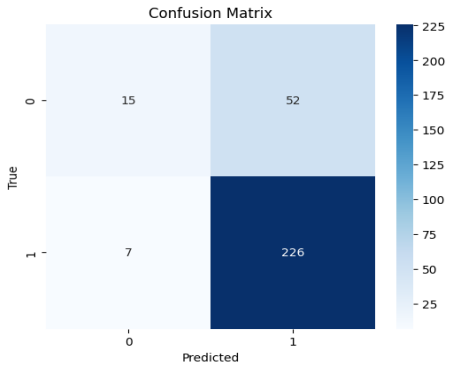

Figure 6. Confusion Matrik Decision Tree

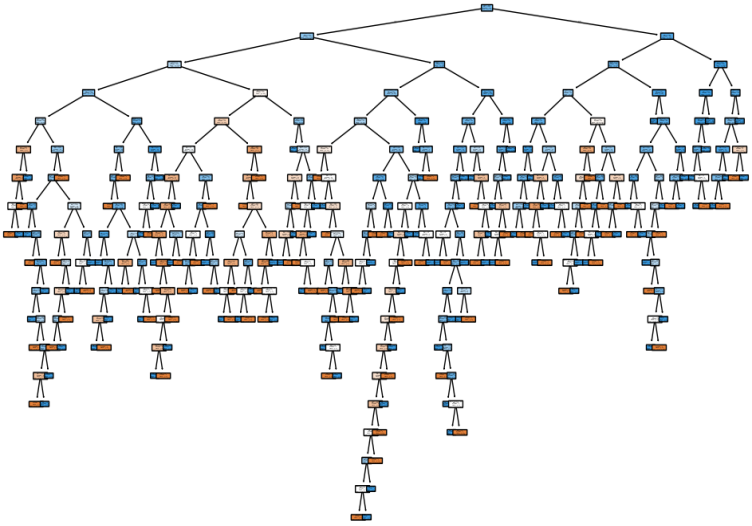

Figure 7. Decision Tree Cancer Prediction

3.6. Support Vector Machine (SVM)

The SVM process begins by importing and cleaning the data, followed by feature selection. After training the model to find the ideal hyperplane, the model is evaluated using metrics such as accuracy and precision and is then used to predict class labels on new data. Confusion matrix SVM can be seen in Figure 8.

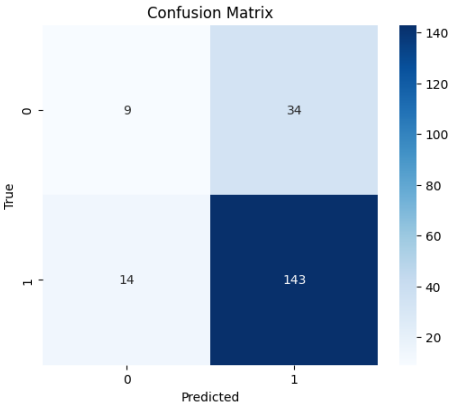

Figure 8. Confusion Matrix SVM

The confusion matrix is used in the SVM algorithm to assess the model's ability to detect cancer. With an accuracy of 77.4%, precision of 91.1%, and recall of 94.1%, the model successfully predicted 143 true positive cases and 34 true negative cases, with some errors, including 14 false positives and 9 false negatives. This indicates that the model is quite good at detecting cancer, though there are some errors, particularly with false negatives.

3.7. Algorithm Comparison

Table 1. Algoritma Comparison

Algoritma	Split Data	Recall	Precision	Akurasi
K-NN	70:30:00	96.14%	78.32 %	76.33%
	80:20:00	95.54%	79.79%	77.50%
Naïve Bayes	70:30:00	91.85%	83.27%	79.33%
	80:20:00	94.27%	84.09%	81.50%
Random Forest	70:30:00	90.65%	81.13%	75.9%
	80:20:00	90.36%	80.97%	75.7%
Decision Tree	70:30:00	60.49 %	59.98%	59.98%
	80:20:00	61.79%	59.87%	70.50%
SVM	70:30:00	77.4%	74.27%	77.4%
	80:20:00	77.4%	74.27%	77.4%

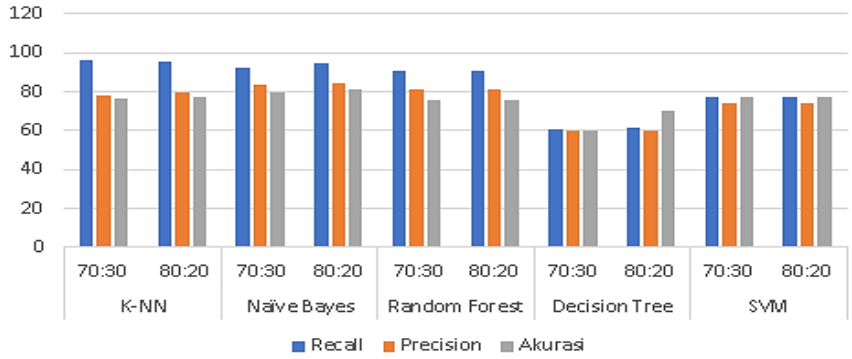

Figure 9. Accuracy Comparison

In Figure 9 and Table 1 is the experimental results show that Naive Bayes achieved the best performance with an accuracy of 81.50% on the 80:20 data split, followed by K-NN, SVM, and Random Forest. Meanwhile, the Decision Tree yielded the lowest performance.

The superior performance of Naive Bayes can be explained by the characteristics of the dataset, which are relatively simple and have feature distributions that align with the independence assumption among variables [19]. This allows the probabilistic model to generalize effectively. However, despite its high accuracy, NB has limitations in handling non-linear interactions between features, which may result in different outcomes when applied to more complex datasets [20].

In terms of interpretability, algorithms such as Decision Tree and Random Forest are more advantageous since they can display decision tree structures or feature importance [13]. The trade-off between accuracy and interpretability needs to be considered, especially in the medical context where model explainability is highly important for healthcare practitioners [12]. For instance, RF may slightly underperform compared to NB in terms of accuracy, but it is more acceptable in clinical practice due to its transparency.

A comparison with the study by Ozcan et al. (2021) revealed different findings. In their study, SVM was reported as the best algorithm for breast cancer classification, whereas in this study NB outperformed the others [31]. This discrepancy is likely due to variations in dataset characteristics, preprocessing techniques, and parameter tuning strategies. These findings reinforce the argument that the performance of machine learning algorithms strongly depends on the dataset context. To enrich the interpretation, this study also presents ROC/AUC curves and per-algorithm *confusion matrices*. These visualizations assist in understanding the trade-off between the true positive rate and false positive rate, while also emphasizing that although NB achieved the highest accuracy, other models such as RF and SVM remain relevant when interpretability and stability are prioritized.

4. CONCLUSION

This study examines the performance of five supervised learning algorithms, namely K-NN, Naive Bayes, Decision Tree, Random Forest, and SVM, for cancer prediction using a relevant dataset. The analysis results indicate that Naive Bayes achieves the highest accuracy at 79.33%, followed by SVM (77.4%), K-NN (76.33%), Random Forest (75.9%), and Decision Tree with the lowest accuracy of 60.49%. In addition to accuracy, Naive Bayes demonstrates a good balance between precision and recall, which is crucial for accurately detecting positive cases. Each algorithm has unique characteristics: K-NN is simple but requires parameter optimization, Naive Bayes is effective but struggles with feature independence assumptions,

Decision Tree is easy to interpret but prone to overfitting, while Random Forest and SVM are more accurate but require more computational resources. The main contribution of this study is to provide a comparative overview of the effectiveness of supervised learning algorithms for cancer prediction, considering accuracy, precision, and recall. The study is limited by the relatively small dataset and the lack of feature selection and ensemble optimization techniques. Future research is recommended to use larger datasets, apply data balancing methods such as SMOTE, undersampling, and oversampling, and include evaluation using AUC/ROC metrics to enrich result interpretation. This comparative analysis is expected to support the development of effective machine learning-based early cancer detection systems as decision-support tools in medical practice.

REFERENCES

[1] I. Ahmad and F. Alqurashi, "Early cancer detection using deep learning and medical imaging: A survey," *Crit. Rev. Oncol. Hematol.*, vol. 204, no. October, p. 104528, 2024, doi: 10.1016/j.critrevonc.2024.104528.

[2] L. Liu *et al.*, "Machine learning protocols in early cancer detection based on liquid biopsy: A survey," *Life*, vol. 11, no. 7, pp. 1–39, 2021, doi: 10.3390/life11070638.

[3] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017, doi: 10.1016/j.neucom.2017.01.026.

[4] L. Sari, A. Romadloni, and R. Listyaningrum, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest," *Infotekmesin*, vol. 14, no. 1, pp. 155–162, 2023, doi: 10.35970/infotekmesin.v14i1.1751.

[5] H. Suryono, H. Kuswanto, and N. Iriawan, "Rice phenology classification based on random forest algorithm for data imbalance using Google Earth engine," *Procedia Comput. Sci.*, vol. 197, no. 2021, pp. 668–676, 2021, doi: 10.1016/j.procs.2021.12.201.

[6] V. Nemade and V. Fegade, "Machine Learning Techniques for Breast Cancer Prediction," *Procedia Comput. Sci.*, vol. 218, no. 2022, pp. 1314–1320, 2022, doi: 10.1016/j.procs.2023.01.110.

[7] A. Bilal, A. Imran, T. I. Baig, X. Liu, E. Abouel Nasr, and H. Long, "Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization," *Sci. Rep.*, vol. 14, no. 1, pp. 1–25, 2024, doi: 10.1038/s41598-024-61322-w.

[8] M. Tiara *et al.*, "Pemanfaatan Algoritma Adasyn Dan Support Vector Machine Dalam Meningkatkan Akurasi Prediksi Kanker Paru-Paru," vol. 8, no. 5, pp. 8773–8778, 2024.

[9] C. A. Ul Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," *ICAC 2018 - 2018 24th IEEE Int. Conf. Autom. Comput. Improv. Product. through Autom. Comput.*, no. September, pp. 1–6, 2018, doi: 10.23919/IConAC.2018.8748995.

[10] A. Eleyan, "Breast cancer classification using moments," *2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet.*, pp. 1–4, 2012, doi: 10.1109/siu.2012.6204778.

[11] N. Manjunathan, N. Gomathi, and S. Muthulingam, "Early Detection of Breast Cancer using Machine Learning," *Int. Conf. Sustain. Comput. Smart Syst. ICSCSS 2023 - Proc.*, vol. 10, no. 3, pp. 165–169, 2023, doi: 10.1109/ICSCSS57650.2023.10169777.

[12] A. Yaqoob, R. Musheer Aziz, and N. K. verma, "Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review," *Human-Centric Intell. Syst.*, vol. 3, no. 4, pp. 588–615, 2023, doi: 10.1007/s44230-023-00041-3.

[13] E. Asamoah, G. B. M. Heuvelink, I. Chairi, P. S. Bindraban, and V. Logah, "Random forest machine learning for maize yield and agronomic efficiency prediction in Ghana," *Heliyon*, vol. 10, no. 17, p. e37065, 2024, doi: 10.1016/j.heliyon.2024.e37065.

[14] P. P. Sengar, M. J. Gaikwad, and A. S. Nagdive, "Comparative study of machine learning algorithms for breast cancer prediction," *Proc. 3rd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2020*, no. December 2016, pp. 796–801, 2020, doi: 10.1109/ICSSIT48917.2020.9214267.

[15] M. M. Hassan *et al.*, "A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction," *Decis. Anal. J.*, vol. 7, no. May, p. 100245, 2023, doi: 10.1016/j.dajour.2023.100245.

[16] A. Almomany, W. R. Ayyad, and A. Jarrah, "Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3815–3827, 2022, doi: 10.1016/j.jksuci.2022.04.006.

[17] A. Hernandes, S. Kurnia Gusti, F. Syafria, L. Handayani, and S. Ramadhani, "Klasifikasi Data Penerimaan Zakat dengan Algoritma K-Nearest Neighbor," *Media Online*, vol. 4, no. 3, pp. 1632–1640, 2023, doi: 10.30865/klik.v4i3.1528.

[18] Z. C. Dwinnie, L. Khairani, M. A. M. Putri, J. Adhiva, and M. I. F. Tsamarah, "Application of the Supervised Learning Algorithm for Classification of Pregnancy Risk Levels," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 1, no. 1, pp. 26–33, 2023, doi: 10.57152/predatecs.v1i1.806.

[19] Y. Shang, "Prevention and detection of DDOS attack in virtual cloud computing environment using

Naive Bayes algorithm of machine learning," *Meas. Sensors*, vol. 31, no. December 2023, p. 100991, 2024, doi: 10.1016/j.measen.2023.100991.

[20] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 504–510, 2021, doi: 10.29207/resti.v5i3.3067.

[21] Rayuwati, Husna Gemasih, and Irma Nizar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid," *Jural Ris. Rumpun Ilmu Tek.*, vol. 1, no. 1, pp. 38–46, 2022, doi: 10.55606/jurritek.v1i1.127.

[22] A. Al Nasseri, A. Tucker, and S. De Cesare, "Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms," *Expert Syst. Appl.*, vol. 42, no. 23, pp. 9192–9210, 2015, doi: 10.1016/j.eswa.2015.08.008.

[23] M. R. Anugrah, N. A. Al-Qadr, N. Nazira, and N. Ihza, "Implementation of C4.5 and Support Vector Machine (SVM) Algorithm for Classification of Coronary Heart Disease," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 1, no. 1, pp. 20–25, 2023, doi: 10.57152/predatecs.v1i1.805.

[24] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-2941-7.

[25] A. F. Lubis *et al.*, "Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors, Naïve Bayes and Decission Tree," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. 1, pp. 44–51, 2024, doi: 10.57152/predatecs.v2i1.1103.

[26] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[27] D. Ananda, S. Nurhidayarnis, T. A. Afifah, M. A. Ramadhan, and I. Mahendra, "Text Classification of Translated Qur'anic Verses Using Supervised Learning Algorithm," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 1, no. 2, pp. 78–84, 2024, doi: 10.57152/predatecs.v1i2.870.

[28] A. Rahmah, N. Sepriyanti, M. H. Zikri, I. Ambarani, and M. Y. bin Shahar, "Implementation of Support Vector Machine and Random Forest for Heart Failure Disease Classification," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 1, no. 1, pp. 34–40, 2023, doi: 10.57152/predatecs.v1i1.816.

[29] H. Apriyani and K. Kurniati, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," *J. Inf. Technol. Ampera*, vol. 1, no. 3, pp. 133–143, 2020, doi: 10.51519/journalita.volume1.isssue3.year2020.page133-143.

[30] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," 2020, [Online]. Available: http://arxiv.org/abs/2001.09636

[31] I. Ozcan, H. Aydin, and A. Cetinkaya, "Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer," *Asian Pacific J. Cancer Prev.*, vol. 23, no. 10, pp. 3287–3297, 2022, doi: 10.31557/APJCP.2022.23.10.3287.