Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Breast Cancer Classification Using Naïve Bayes and Random Forest Algorithms

Riris Naomi Gurning[1)*], Asep Arwan Sulaeman[2)], Dedi Afandi[3)]
[1,2,3)] Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University, Indonesia
[1)]ririsnaomi19@gmail.com, [2)]aseparwan@pelitabangsa.ac.id, [3)]afandi.ihbs@gmail.com

ABSTRACT

Breast cancer remains one of the leading causes of mortalty among women in Indonesia, making early detection a critical factor in improving treatment outcomes. This study aims to compare the classification performance of the Naïve Bayes and Random Forest algorithms for early breast cancer detection using machine learning techniques. The dataset was sourced from an open-access platform and partitioned into 80% training and 20% testing subsets. Model performance was evaluated using accuracy, precision, and recall metrics. The results indicate that the Random Forest algorithm outperforms Naïve Bayes, achieving an accuracy of 99.27%, recall of 99.27%, and precision of 99.30%. In contrast, Naïve Bayes achieved only 83.78% accuracy, with 83.80% recall and precision. The novelty of this study lies in its systematic comparative evaluation of two classical machine learning algorithms in the context of early breast cancer detection, with a clearly structured data split and rigorous performance metrics. Such comparative analyses remain limited in existing literature, particularly within the scope of clinical data applications in Indonesia. These findings suggest that Random Forest is more suitable for implementation in medical decision support systems aimed at early breast cancer detection and highlight the potential of classification algorithms in enhancing digital healthcare systems.

Keywords: Classification; Algorithm; Naïve Bayes; Random Forest; Breast Cancer; Cancer

INTRODUCTION

Breast cancer remains one of the leading causes of mortality among women globally, including in Indonesia, with a steadily increasing incidence rate each year (Organization, 2023).While early detection has been shown to significantly improve survival rates, conventional diagnostic procedures still face several limitations, including reliance on specialized medical personnel, high costs, and limited accessibility in underserved areas (Yala et al., 2019). In this context, machine learning (ML) methods offer considerable potential to enhance both the efficiency and accuracy of medical data classification, particularly in the early diagnosis of breast cancer.

Although numerous studies have explored the application of classification algorithms in cancer detection, many have focused on single-algorithm evaluations or have not tested performance on standardized and accessible datasets. Furthermore, comparative studies that systematically evaluate the performance of Naïve Bayes and Random Forest using uniform evaluation metrics and publicly available datasets, such as those from Kaggle, remain limited. These two algorithms are fundamentally different in their classification approaches: Naïve Bayes relies on the assumption of feature independence, while Random Forest employs an ensemble of decision trees, offering greater flexibility in handling non-linear and complex data structures (Zhou et al., 2022)

This study addresses this research gap by conducting a comparative performance analysis of the Naïve Bayes and Random Forest algorithms in classifying breast cancer cases using a curated public dataset. The dataset was split into 80% training and 20% testing subsets, and performance was evaluated using standard classification metrics including accuracy, precision, and recall. Beyond measuring predictive accuracy, the study also investigates each algorithm's ability to manage high-dimensional, multivariate clinical data.

The primary contribution of this research is to provide robust empirical evidence for selecting optimal machine learning algorithms in the development of AI-based medical decision support systems, particularly for early breast cancer detection. Moreover, this study enriches academic discourse on health informatics by adopting an open replication approach, enabling validation and further exploration by the scientific community.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

LITERATURE REVIEW

This research is based on a literature review of various previous studies that have relevance to the research topic. The sources used mostly come from relevant scientific journals and support the theoretical basis and development of this research.

According to (Suparna & Sari, 2022), the purpose of this study is to support the process of diagnosing breast cancer more accurately and quickly by applying machine learning-based classification algorithms. This research evaluates and compares the performance of Naïve Bayes and Random Forest algorithms in classifying breast cancer patient data. Thus, this research aims to identify the most effective algorithm in producing predictions with optimal accuracy, precision, and recall, to support early detection and decision making in the medical field.

According to (Asmalinda et al., 2022), the objective of this study is to increase the understanding and awareness of women of childbearing age regarding the importance of early detection of breast cancer through self-breast examination (SADARI). Through educational activities and simulations conducted in the form of a Community Partnership Program (PKM), this study also aims to evaluate the effectiveness of SADARI education and practice in improving participants' knowledge of the stages of self-examination. It is hoped that these activities will encourage early preventive behavior against the risk of breast cancer.

According to (Shidqi, 2022), the aim of this study was to explore the factors that contribute to delays in breast cancer treatment, both from the patient and health professionals. Through a systematic review referring to the PRISMA guidelines without meta-analysis, this study analyzed selected articles published between 2012 and 2021, to identify the causes of delays in the process of diagnosis and treatment of breast cancer.

According to (Cahyana & Nurlayli, 2023), the purpose of this study is to identify the most accurate machine learning algorithm in predicting breast cancer based on Coimbra data. This research compares the performance of three algorithms, namely Logistic Regression, Naïve Bayes, and Random Forest, to find out which method provides the best prediction results. This research is expected to provide significant support for the process of early detection of breast cancer and become a reference for medical personnel and the public in choosing effective analysis methods to help treat this disease.

According to (Muntiari & Hanif, 2022), this study aims to classify breast cancer types into benign or malignant categories by utilizing various machine learning algorithms to support quick and accurate decision making. Seven algorithms were tested in this study, namely Neural Network, Decision Tree, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Random Forest, and Support Vector Machines. An evaluation was conducted to assess the effectiveness of each method in performing classification. It is hoped that the results of this study can accelerate the diagnosis process and facilitate medical personnel in determining the appropriate treatment for breast cancer patients.

According to (Widodo, 2023), this study aims to find factors that affect the production of baby diapers at PT Elleair International Manufacturing Indonesia and predict production results to anticipate shortages of finished products. By applying data mining methods using the Naïve Bayes algorithm and utilizing production data from the previous year as training data, this research seeks to provide accurate predictions based on important variables such as material usage, human error, and stop delivery events. Through the results of this study, it is hoped that help companies improve production efficiency so that they can better meet market demand.

According to (R Wahid & Affandi, 2022), this study aims to evaluate the implementation of clinical risk management in surgery as an effort to support the improvement of service quality at RSUD Arifin Achmad Pekanbaru. Using a qualitative approach through the Rapid Assessment Procedure method, this study examines the implementation of various important aspects, such as informed consent, Surgical Safety Checklist, diagnosis recording and reporting, surgical team formation, and consultation during surgery, to see its contribution to the overall quality of hospital services.

The following table presents a comparison between previous studies and this study, viewed from several key aspects such as the methods used, the datasets used, the main results, and the strengths and weaknesses of each study. This comparison emphasizes the differences in methods, datasets, and findings between previous studies and the contributions made by this study.

Table 1. literature review summary

Author(s) & Year	**Method/Algorithm Used**	**Dataset Used**	**Key Findings**	**Limitations / Research Gap**
(Suparna &	Naïve Bayes, Random	Not explicitly	Random Forest	Did not use public

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Sari, 2022)	Forest	mentioned	outperformed Naïve Bayes in classifying breast cancer data	dataset; lacked detailed evaluation metrics
(Asmalinda et al., 2022)	Educational simulation (SADARI)	Community participant data (non-numeric)	Breast self-examination (SADARI) improved awareness and early detection knowledge	Not a machine learning or classification-based study
(Shidqi, 2022)	Systematic Literature Review (PRISMA, no meta-analysis)	Selected articles (2012–2021)	Identified patient/system-related delays in breast cancer diagnosis and treatment	No data modeling or algorithmic analysis; not relevant to ML-based classification
(Cahyana & Nurlayli, 2023)	Logistic Regression, Naïve Bayes, Random Forest	Coimbra Breast Cancer Dataset	Random Forest produced the highest prediction accuracy among the three methods	Small dataset; lacks generalizability; did not compare results on open Kaggle data
(Muntiari & Hanif, 2022)	7 ML algorithms (NN, DT, NB, KNN, LR, RF, SVM)	Not specified	SVM and RF demonstrated highest classification performance	No clear data split; lacks details on evaluation metrics and data source
(Widodo, 2023)	Naïve Bayes	Industrial production data (non-medical)	Naïve Bayes effectively predicted production outcomes based on key manufacturing variables	Not related to cancer or medical data; context in manufacturing
(R Wahid & Affandi, 2022)	Rapid Assessment Procedure (qualitative)	Interviews and hospital observations	Risk management in surgery showed implementation weaknesses	Not related to data classification or prediction algorithms
This Study (2025)	Naïve Bayes, Random Forest	Open-access Kaggle breast cancer dataset	Random Forest achieved highest performance (Accuracy: 99.27%, Recall: 99.27%, Precision: 99.30%)	Fills the gap through direct comparative evaluation using public dataset and comprehensive metrics

A review of the existing literature indicates that most prior studies either lacked transparency regarding the datasets used or did not conduct a systematic evaluation of machine learning algorithms using standardized and reproducible metrics. In addition, several studies were centered on educational efforts, qualitative assessments, or predictive tasks unrelated to medical classification, thereby offering limited contributions to computational approaches for breast cancer diagnosis. In contrast, this study provides a distinct contribution by conducting a direct comparative analysis of two widely adopted machine learning algorithms, Naïve Bayes and Random Forest, using a publicly available dataset from Kaggle.

The use of this open dataset not only ensures replicability but also allows for external validation by other researchers. Furthermore, this study employs comprehensive evaluation metrics, including accuracy, precision, and recall, to rigorously assess model performance. As such, it offers a valuable empirical reference for the development of AI-based clinical decision support systems aimed at early detection of breast cancer. This contribution addresses a notable research gap and strengthens the scientific foundation for the practical integration of machine learning in real-world medical diagnostics.

METHOD

The purpose of this study is to evaluate the performance of Naïve Bayes and Random Forest algorithms in classifying breast cancer using historical data sourced from trusted institutions. The data used has been filtered through a selection process to ensure its accuracy and consistency. The research procedure includes data collection, preprocessing stage, division of the dataset into training and test data, application of both algorithms, evaluation of

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

model performance, and analysis of results to determine which algorithm provides the best performance in classification.

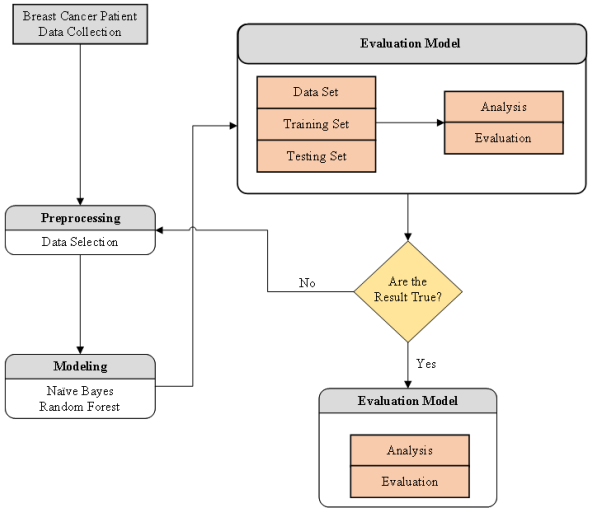

Fig. 1 Research Stages

The diagram illustrates a systematic workflow for breast cancer classification using the Naïve Bayes and Random Forest algorithms. The process begins with the collection of breast cancer patient data, which serves as the foundation for subsequent steps. Once the data is obtained, it undergoes a preprocessing stage involving data selection, where irrelevant or missing attributes are removed to ensure data quality and relevance. This cleaned dataset is then used in the modeling phase, where two classification algorithms, Naïve Bayes and Random Forest, are applied to build predictive models.

Following model construction, an evaluation stage is conducted, where the dataset is divided into training and testing sets. These subsets are used to analyze and evaluate the performance of each algorithm using metrics such as accuracy, precision, and recall. A decision node then determines whether the evaluation results meet the validity criteria. If the results are not satisfactory, the process loops back for further data selection or model refinement. If the results are deemed valid, the process continues to the final evaluation phase, which involves comprehensive analysis and validation of the model's predictive accuracy. This iterative and structured approach ensures that the final model is both reliable and applicable in clinical decision-making contexts for early breast cancer detection.

Dataset

This study employs a dataset comprising 1,897 records and 22 variables, obtained from the official Kaggle platform (https://kaggle.com). The dataset contains comprehensive information relevant to breast cancer classification, including demographic data, lifestyle factors, medical history, and clinical symptoms. Key variables include age, weight, height, body mass index (BMI), obesity status, history of breastfeeding, residential location, alcohol intake, smoking status, family history of breast cancer, number of children, age at first menstruation, menopausal status, use of hormone replacement therapy, and use of oral contraceptives.

In addition, the dataset incorporates clinical indicators such as breast swelling, presence of lumps, breast pain, and history of breast biopsy. The target variable, Diagnosis Status, is categorical and consists of two possible outcomes: Benign and Malignancy. These features collectively serve as the input variables for the classification models developed in this study.

This dataset was selected due to its structured format, relevance to the classification task, and public availability, which facilitates reproducibility and comparative research in the field of medical data analysis.

Preprocessing and Data Selection

Since the dataset was already clean, the preprocessing stage was limited to attribute selection, focusing on variables most relevant for the classification task based on the Diagnosis Status. This selection process was performed automatically using RapidMiner software to enhance classification performance and eliminate irrelevant or redundant features. Selected variables include Age, Weight, Height, BMI, Obesity, and Breast Feeding history.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Naïve Bayes

Naïve Bayes is a classification algorithm based on the Bayes probability principle, assuming that each attribute in the data is independent of each other. Although this assumption does not always match real data conditions, this method is still widely used because it is efficient in processing large-scale data and has a fast computation time. This algorithm determines the class of a data by calculating the highest probability based on the values of its attributes (Wickramasinghe & Kalutarage, 2021).

In the realm of machine learning, Naïve Bayes serves to predict the class of data based on patterns learned from previous training data. This prediction process uses the Bayes formula with an independence approach between features. This algorithm is very commonly applied in various fields, such as document classification, spam detection in emails, opinion analysis, and disease diagnosis systems, due to its simple yet effective capabilities (Chen, 2021).

As a probability-based classification method, Naïve Bayes is used to predictively map the relationship between input data and output classes. The uniqueness of this algorithm lies in its assumption that each feature contributes to the classification result separately, although in practice they can be interrelated. Despite its simplicity, the algorithm is capable of producing good performance, especially in cases of classification involving text or categorized data (Jackins, 2021).

Naïve Bayes is an algorithm that estimates the statistical probability that a data entry will be classified into a particular the probability calculation of a class (posterior) is performed by multiplying the prior value of the class by the probability of a certain feature occurring in that class, known as likelihood. In general, this approach is formulated in an equation that describes the basic concept of the Naïve Bayes method.

$$P\ (c \mid x) = \frac{(P\ (x \mid c)\ x\ P\ (c))}{P\ (x)} \qquad (1)$$

Description:

x : Represents data that has no known class.
c : It is an assumption or conjecture that the data collection xx identified as belonging to a particular class.
$(c|x)$: Posterior probability, which is the probability that the data x belongs to class c based on its attributes.
(c) : The initial or prior probability of a class before considering the data x.
$(x|c)$: Probability of attribute x occurring if it is known to belong to class c (likelihood).
(x) : Initial probability of occurrence of attribute x, without considering the class.

Random Forest

Random Forest is an ensemble-based machine learning algorithm that consists of a number of decision trees that are randomly constructed from a subset of the training data. It works by combining the predictions of each tree to produce a more stable and accurate final decision. For classification problems, Random Forest selects the class most chosen by all trees, while for regression, it calculates the average of the results from all trees. This approach reduces the risk of overfitting that often occurs with single decision trees (Salman & Kalakech, 2024).

Random Forest uses a bootstrap aggregating or bagging method, where training data is randomly drawn with returns to form each decision tree. In addition, only a subset of features are randomly selected to be considered at each node split, which helps increase diversity between trees. This strategy increases the model's robustness to noise and outliers and improves accuracy. Random Forest is particularly suitable for data with large and complex features (Hu & Szymczak, 2023).

In the context of classification, the final result of Random Forest is obtained through the majority voting process of all trees formed. If $h1\ (x)$, $h2\ (x)$, ..., $hk(x)$ are the prediction results of each tree on the input data x, then the final result H(x) can be determined by the following formula (Avci & Budak, 2023).

$$H(x) = mode\{h_1(x), h_2(x), \ldots, h_k(x)\} \qquad (2)$$

Description:

$H(x)$: Final prediction based on the most votes from all trees.
x : Data to be predicted.
$h_i(x)$: Prediction of the i-th tree for data x.
k : Total number of trees.
$mode$: The most frequent value from all tree predictions.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Evaluation Model

In this study, the confusion matrix is applied as a tool for evaluating the performance of classification models, by considering four main components, namely true positive, false positive, true negative, and false negative. Analysis of these four metrics provides a comprehensive overview of the accuracy and effectiveness of the model in making predictions. The use of this matrix allows an assessment of the accuracy of the model in performing classification, as well as providing a clearer understanding of the effectiveness of the prediction results. The calculation formula in the confusion matrix to determine the accuracy, precision, and recall values can be explained as follows (Vujović, 2021).

a. Accuracy is the proportion of correct predictions out of the overall data tested.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (3)$$

b. Precision indicates the level of accuracy of the model in classifying the data as true positives.

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

c. Recall describes the model's ability to detect all true positive data.

$$Recall = \frac{TP}{(TP + FN)} \quad (5)$$

The 80/20 train-test split was chosen based on common machine learning practices and for consistency with benchmarking standards. To mitigate data partition bias and assess generalizability, the models were also evaluated using 10-fold cross-validation.

RESULT

This research utilizes a dataset of 1,897 rows and 22 variables downloaded from the official website https://kaggle.com. Because the data used is already in a clean condition, the process carried out is only limited to simple selection at the preprocessing stage before being applied to modeling with the Naïve Bayes and Random Forest algorithms for breast cancer classification.

Table 2. Dataset

No	Age	Weight	Height	BMI	Obesity	Breast Feeding	…..	…..	Diagnosis Status
1	32	55	151	24.1	0	1	…..	…..	Benign
2	60	57	158	22.8	0	1	…..	…..	Benign
3	44	65	151	28.5	0	0	…..	…..	Benign
…	…	…	…	…	…	…	…..	…..	…
…	…	…	…	…	…	…	…..	…..	…
1895	42	56.5	165.5	20.6	0	1	…..	…..	Malignancy
1896	35	63.2	158.6	25.1	0	1	…..	…..	Malignancy
1897	52	83	167	29.8	0	0	…..	…..	Malignancy

This table presents breast cancer patient data consisting of various attributes important for classification purposes. The data includes demographic information and risk factors such as age (Age), residence location (Residence Location), alcohol consumption (Alcohol Intake), smoking habits (Smoking Status), family history of breast cancer (Family history of breast cancer), Number of children, Age at first menarche, Menopausal Status, Hormone replacement therapy use, and Oral contraceptive use. In addition, the data also included clinical symptoms such as breast swelling (Breast Swelling), lump (Breast Lump), pain (Breast Pain), and history of breast biopsy (Breast Biopsy). Other variables included are Weight, Height, Body Mass Index (BMI), Obesity, Exposure to radiation, Occupation, Breast Feeding history, and Diagnosis Status. All of these attributes are used as input in classification modeling with Naïve Bayes and Random Forest algorithms.

Data Selection

In the data selection stage, identification of attributes that are considered relevant for analysis is carried out, by grouping data based on diagnosis status. This selection process is carried out automatically using RapidMiner software, in order to improve performance and accuracy in the breast cancer classification process in this study.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Table 3. Data Selection

No	Diagnosis Status	…..	…..	Age	Weight	Height	BMI	Obesity	Breast Feeding
1	Benign	…..	…..	32	55	151	24.1	0	1
2	Benign	…..	…..	60	57	158	22.8	0	1
3	Benign	…..	…..	44	65	151	28.5	0	0
…	…	…..	…..	…	…	…	…	…	…
…	…	…..	…..	…	…	…	…	…	…
1895	Malignancy	…..	…..	42	56.5	165.5	20.6	0	1
1896	Malignancy	…..	…..	35	63.2	158.6	25.1	0	1
1897	Malignancy	…..	…..	52	83	167	29.8	0	0

This table displays breast cancer data that has gone through a selection process, including several key variables such as identification number, diagnosis status (benign or malignant), age, weight, height, body mass index (BMI), obesity condition, and breastfeeding history. The data is used to perform breast cancer classification analysis by considering various risk factors and patient characteristics.

Testing Process Using RapidMiner

The breast cancer data classification process in this study was carried out by utilizing RapidMiner software, using two machine learning algorithms, namely Naïve Bayes and Random Forest. The stage begins with entering the dataset into RapidMiner, followed by the data preparation process to build and test the performance of the predictive model developed.

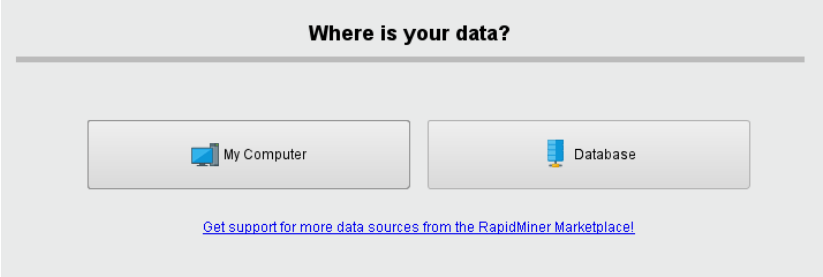

Fig. 2 Import Data

The data in xls format is imported into RapidMiner. As long as there are no errors in this process, the data can be entered in its entirety and the subsequent processes can be carried out smoothly.

Application of Naïve Bayes Algorithm

The initial application of the Naïve Bayes model begins with selecting the operator used to manage the data. Operators are used as the main elements in this study in performing classification and prediction based on historical data.

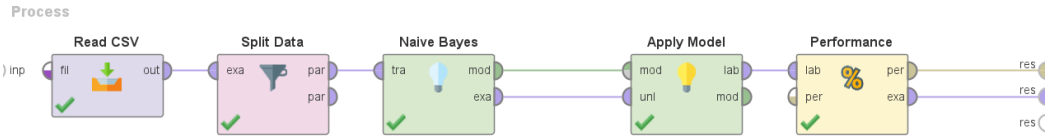

Fig. 3 Naïve Bayes Modeling

This research uses predefined parameters to build a breast cancer classification model with the Naïve Bayes algorithm based on data from Kaggle. The classification results obtained were then analyzed through the help of RapidMiner software.

Evaluation of Naïve Bayes Algorithm Classification Results Using RapidMiner

In this study, the dataset was partitioned with 80% allocated for training and the remaining 20% reserved for testing.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Subsequently, the classification and prediction processes were executed by applying the Naïve Bayes algorithm through the RapidMiner platform.

Row No.	Diagnosis_...	prediction(D...	confidence(...	confidence(...	Age	Residence_...	Alcohol_Int...	Smoking_
1	Benign	Benign	0.996	0.004	60	2	0	0
2	Benign	Benign	0.999	0.001	44	3	0	0
3	Benign	Malignancy	0.001	0.999	74	1	1	1
4	Benign	Benign	0.979	0.021	64	1	0	0
5	Benign	Malignancy	0.319	0.681	50	1	0	1
6	Benign	Benign	0.926	0.074	50	1	1	0
7	Benign	Benign	0.979	0.021	40	3	0	0
8	Benign	Benign	0.996	0.004	60	2	0	0
9	Benign	Benign	0.979	0.021	64	1	0	0
10	Benign	Benign	0.929	0.071	54	2	0	0
11	Benign	Benign	0.995	0.005	62	1	0	0
12	Benign	Benign	0.987	0.013	50	5	0	0
13	Benign	Malignancy	0.120	0.880	45	2	1	0

ExampleSet (1,517 examples, 4 special attributes, 21 regular attributes)

Fig. 4 Outcomes of the Naïve Bayes Classification Process

Once the classification process utilizing the Naïve Bayes method on the RapidMiner application was finalized, the resulting output from the constructed prediction model was presented. The corresponding outcomes are displayed in the subsequent section.

accuracy: 83.78%

	true Benign	true Malignancy	class precision
pred. Benign	633	132	82.75%
pred. Malignancy	114	638	84.84%
class recall	84.74%	82.86%	

Fig. 5 Naïve Bayes Modeling Results

The figure shows that the classification process produced an accuracy rate of 83.78%, with predictions for the benign class at 82.75% and the malignancy class at 84.84%. The recall rate was recorded at 84.74% for the benign class and 82.86% for the Malignancy class. The testing was conducted using an 80% training data proportion and a 20% testing data proportion, from a total of 765 breast cancer data points used in this study.

PerformanceVector

```
PerformanceVector:
accuracy: 83.78%
ConfusionMatrix:
True:   Benign  Malignancy
Benign: 633     132
Malignancy:     114     638
```

Fig. 6 Performance Vector Naïve Bayes Result

In the final stage, the Naïve Bayes algorithm achieved an accuracy of 83.78%, with recall and precision values of 83.80% each, indicating an adequate performance in breast cancer classification. Visualization of the classification results is shown in the form of the following diagram.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

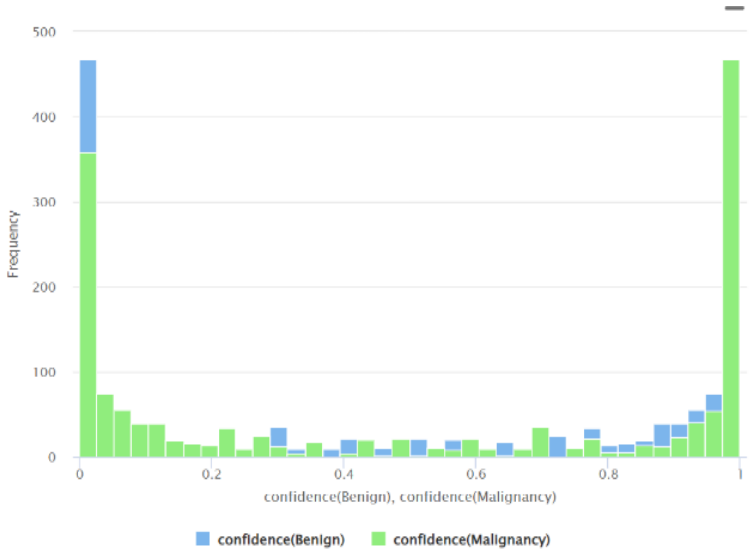

Fig. 7 Visualization Results of Naive Bayes Model

This visualization shows the distribution of the confidence model for the Benign and Malignancy categories. Most predictions show high confidence values (close to 0 or 1), which indicates the model's ability to accurately distinguish between the two classes. Blue indicates predictions for the benign class, while green indicates predictions for the Malignancy class. After that, the performance of the Random Forest algorithm was evaluated using RapidMiner.

Utilization of the Random Forest Algorithm for Predictive Analysis

The process of utilizing the Random Forest algorithm starts with selecting the appropriate operator to handle data processing. In this research, Random Forest was adopted as the primary technique to carry out classification and prediction based on the historical dataset provided.

Row No.	Diagnosis_...	prediction(D...	confidence(...	confidence(...	Age	Residence_...	Alcohol_Int...	Smoking_
1	Benign	Benign	0.986	0.014	60	2	0	0
2	Benign	Benign	0.982	0.018	44	3	0	0
3	Benign	Benign	0.990	0.010	74	1	1	1
4	Benign	Benign	0.966	0.034	64	1	0	0
5	Benign	Benign	0.956	0.044	50	1	0	1
6	Benign	Benign	0.962	0.038	50	1	1	0
7	Benign	Benign	0.991	0.009	40	3	0	0
8	Benign	Benign	0.986	0.014	60	2	0	0
9	Benign	Benign	0.966	0.034	64	1	0	0
10	Benign	Benign	0.977	0.023	54	2	0	0
11	Benign	Benign	0.962	0.038	62	1	0	0
12	Benign	Benign	0.979	0.021	50	5	0	0
13	Benign	Benign	0.610	0.390	45	2	1	0
14	Benign	Benign	0.975	0.025	23	1	0	0

ExampleSet (1,517 examples, 4 special attributes, 21 regular attributes)

Fig. 8 Random Forest Modeling

This study utilizes predetermined parameters to develop a breast cancer classification model using the Random Forest algorithm, based on a dataset from Kaggle. The classification results are then analyzed using the RapidMiner application.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Classification Results of Random Forest Algorithm on RapidMiner

This study used 80% of the data for training and 20% for testing. Next, the Random Forest algorithm was run through the RapidMiner platform to generate predictions and classifications from the model that was built.

Row No.	Diagnosis_...	prediction(D...	confidence(...	confidence(...	Age	Residence_...	Alcohol_Int...	Smoking_
1	Benign	Benign	0.986	0.014	60	2	0	0
2	Benign	Benign	0.982	0.018	44	3	0	0
3	Benign	Benign	0.990	0.010	74	1	1	1
4	Benign	Benign	0.966	0.034	64	1	0	0
5	Benign	Benign	0.956	0.044	50	1	0	1
6	Benign	Benign	0.962	0.038	50	1	1	0
7	Benign	Benign	0.991	0.009	40	3	0	0
8	Benign	Benign	0.986	0.014	60	2	0	0
9	Benign	Benign	0.966	0.034	64	1	0	0
10	Benign	Benign	0.977	0.023	54	2	0	0
11	Benign	Benign	0.962	0.038	62	1	0	0
12	Benign	Benign	0.979	0.021	50	5	0	0
13	Benign	Benign	0.610	0.390	45	2	1	0
14	Benign	Benign	0.975	0.025	23	1	0	0

ExampleSet (1,517 examples, 4 special attributes, 21 regular attributes)

Fig. 9 Random Forest Classification Results

After completing the classification process with the Random Forest algorithm in RapidMiner, the author presents the output from the prediction model that has been created. The results can be seen in the following section.

accuracy: 99.27%

	true Benign	true Malignancy	class precision
pred. Benign	736	0	100.00%
pred. Malignancy	11	770	98.59%
class recall	98.53%	100.00%	

Fig. 10 Random Forest Modeling Results

The visualization results show that the classification process achieved an accuracy rate of 99.27%. The prediction rate for the benign class reached 100.00%, while for the malignant class it was 98.59%. Meanwhile, the recall for the benign class is recorded at 98.53% and for the Malignancy class at 100.00%. The testing was conducted using 80% of the data as training data and 20% as test data, from a total of 765 breast cancer samples used in this study.

PerformanceVector

```
PerformanceVector:
accuracy: 99.27%
ConfusionMatrix:
True:   Benign  Malignancy
Benign: 736     0
Malignancy:     11      770
```

Fig. 11 Performance Vector Random Forest Result

In the final stage, the Random Forest algorithm achieved an accuracy of 99.27%, with a recall of 99.27% and a precision of 99.30%. These results show that the model has a very good performance in classifying breast cancer.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Visualization of the classification results is shown in the form of the following diagram.

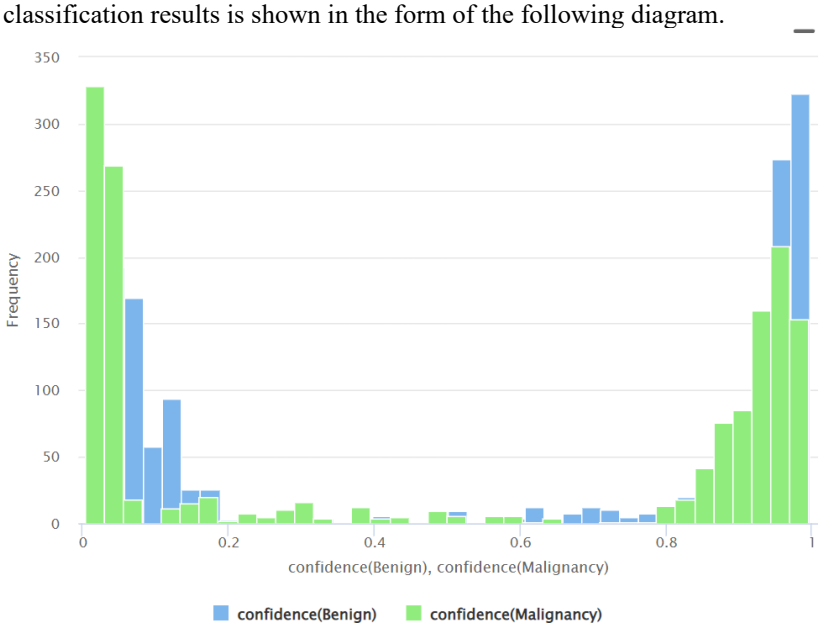

Fig. 12 Visualization Results of Random Forest Model

This figure displays the confidence distribution of the breast cancer classification prediction results using the Naïve Bayes and Random Forest algorithms. The majority of the predictions show a high probability of one of the classes, either benign or malignancy, which is reflected by the accumulation of values in the range close to 0 and 1. This indicates that the model is able to provide fairly convincing predictions. This visualization is used as a basis to compare the performance of the two algorithms in classifying breast cancer through the RapidMiner platform.

Comparison of Classification Results

Table 4. Naive Bayes vs. Random Forest

Algorithm	Accuracy (%)	Precision (%)	Recall (%)
Naïve Bayes	83.78	83.80	83.80
Random Forest	99.27	99.30	99.27

Based on the classification results presented in Table 4, it can be concluded that the Random Forest algorithm consistently demonstrates significantly superior performance compared to Naïve Bayes in classifying breast cancer data. Random Forest achieved an accuracy of 99.27%, with a precision of 99.30% and recall of 99.27%. These figures indicate that nearly all predictions made by the model are correct, effectively identifying both benign and malignant cases.

In contrast, the Naïve Bayes algorithm only achieved an accuracy of 83.78%, with both precision and recall at 83.80%. This disparity suggests that although Naïve Bayes can deliver reasonably good results, it has limitations in handling complex, nonlinear, and interrelated medical data. This limitation stems from the core assumption of Naïve Bayes, that all features are independent, which is often violated in real-world datasets.

On the other hand, Random Forest utilizes an ensemble learning approach by constructing multiple decision trees and combining their outcomes to produce more stable and accurate predictions. Its ability to accommodate variable interactions and manage high-dimensional data makes it more robust in medical classification scenarios. Therefore, these findings reinforce Random Forest as a more suitable algorithm for implementing machine learning-based early detection systems for breast cancer.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Potential Model Bias and Challenges

One common challenge in medical classification tasks is class imbalance, where the number of benign cases typically exceeds that of malignant ones. Such imbalance can introduce majority class bias, leading the model to favor predictions aligned with the dominant class. In the case of the Naïve Bayes algorithm, which relies on frequency-based probability estimates, this bias may reduce the model's sensitivity toward the minority class. Although the dataset used in this study is relatively balanced, the algorithm's lower accuracy of 83.78% suggests that it may still be affected by subtle distributional skewness. In contrast, Random Forest demonstrates a stronger capability to handle class imbalance through its use of bagging and random feature selection, enabling each decision tree in the ensemble to focus on different data subsets. This method promotes greater generalization and stability in predictions, as reflected in the model's significantly higher accuracy, precision, and recall.

Another important consideration in this study is the redundancy of features and high dimensionality. The dataset comprises 22 variables, including demographic, lifestyle, and clinical attributes. While this broad range of features can be informative, the presence of irrelevant or highly correlated variables may introduce noise and negatively affect model performance, particularly for Naïve Bayes, which assumes conditional independence between features. When this assumption is violated, as is likely with correlated variables such as BMI, weight, and height, prediction accuracy can deteriorate. Random Forest, however, is inherently more robust in such scenarios, as it evaluates only a random subset of features at each node split. This mechanism reduces the effect of multicollinearity and enhances the model's ability to generalize, thereby contributing to its superior performance in this study.

DISCUSSIONS

The results of this study indicate that the Random Forest algorithm significantly outperforms Naïve Bayes in breast cancer classification, achieving near-perfect accuracy, precision, and recall. This superiority is attributed to the ensemble approach of Random Forest, which combines multiple decision trees to effectively manage complex and non-linear data while reducing the risk of overfitting. In contrast, the Naïve Bayes algorithm, which assumes independence among features, is less capable of capturing the correlations commonly present in medical data. This limitation is reflected in its lower performance, with an accuracy of only 83.78%.

However, it is important to recognize that each algorithm possesses its own strengths and limitations, particularly in terms of interpretability and complexity. Naïve Bayes excels in simplicity, speed, and high interpretability, as its probabilistic foundation allows healthcare professionals to easily understand the rationale behind predictions. Its main drawback lies in its inability to effectively handle interdependent features. On the other hand, Random Forest delivers very high predictive accuracy and is well-suited for high-dimensional data, but it operates as a "black box" model. This makes it more challenging to interpret the decision-making logic in clinical contexts. Additionally, its algorithmic complexity entails greater computational demands, which may pose constraints in hospital environments with limited technological infrastructure.

Overall, these findings are consistent with previous studies that have shown Random Forest to perform well in medical classification tasks (Choudhury et al., 2022). Nevertheless, several studies also emphasize the importance of model interpretability in clinical applications, where Naïve Bayes remains a valid choice, especially in preliminary screening contexts or in assisting physicians in understanding predictions logically (Yala et al., 2019) . Therefore, the selection of the ideal model must consider the trade-off between predictive accuracy and decision transparency, depending on the specific needs of clinical practice.

This study also acknowledges several limitations, including the use of only a single dataset and the absence of a comprehensive hyperparameter tuning process. These factors may limit the generalizability of the model to broader populations. Furthermore, the implementation of machine learning models in the medical field involves challenges beyond technical performance, such as system integration with electronic medical records (EMR), ethical approval, patient data security, and the acceptance of AI-based systems by healthcare professionals.

Considering all of these aspects, the Random Forest algorithm is recommended as a highly effective method to support early breast cancer detection using machine learning, particularly within the framework of clinical decision support systems (CDSS). However, to ensure long-term sustainability and practical implementation, further development is needed, especially in the form of explainable AI models, external validation, and interdisciplinary collaboration between data scientists and medical professionals.

CONCLUSION

The conclusion of this study demonstrates that the Random Forest algorithm significantly outperforms Naïve Bayes in breast cancer classification, as evidenced by its impressive accuracy of 99.27%, precision of 99.30%, and

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

recall of 99.27%. In contrast, Naïve Bayes achieved a lower accuracy of 83.78%, with precision and recall values of 83.80% each. These findings highlight the effectiveness of the ensemble approach used by Random Forest in handling complex, feature-rich medical data, resulting in more stable and reliable classification outcomes. The main contribution of this research lies in providing a comparative evaluation of two widely-used machine learning algorithms, using an open-access dataset from Kaggle and applying systematic and reproducible performance metrics. Furthermore, this study offers a practical foundation for the development of Clinical Decision Support Systems (CDSS) powered by artificial intelligence to support early breast cancer detection.

Nevertheless, this study has several limitations, including the use of only one dataset type, the absence of cross-validation methods, and the lack of comprehensive hyperparameter tuning. Additionally, an in-depth analysis of feature importance, potentially crucial for enhancing clinical interpretability, was not conducted. Therefore, future research should involve more diverse datasets from various clinical sources, ensure secure and ethical integration with electronic medical record (EMR) systems, adopt explainable AI (XAI) approaches to bridge the gap between algorithmic predictions and clinical understanding, and foster interdisciplinary collaboration between data scientists and healthcare practitioners. Considering both technical and practical aspects, the findings of this study are expected to encourage broader and more responsible application of machine learning in breast cancer detection systems, ultimately aiming to improve diagnostic accuracy and overall healthcare service quality.

REFERENCES

Asmalinda, W., Setiawati, D., Khotimah, K., & Sapada, E. (2022). Deteksi Dini Kanker Payudara Mengunakan Pemeriksaan Payudara Sendiri (Sadari). *Jurnal Abdikemas*, *4*(1), 10–17. https://doi.org/10.36086/j.abdikemas.v4i1

Avci, C., & Budak, M. (2023). Comparison between random forest and support vector machine algorithms for LULC classification. *International Journal of Engineering and Geosciences*, *8*(1), 1–10. https://doi.org/10.26833/ijeg.987605

Cahyana, C. W., & Nurlayli, A. (2023). Analisis Performa Logistic Regression, Naïve Bayes, dan Random Forest sebagai Algoritma Pendeteksi Kanker Payudara. *INSERT: Information System and Emerging Technology Journal*, *4*(1), 51–64.

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

Choudhury, A., Asan, O., Rieger, C., & McCullough, J. (2022). Machine learning for classification of breast cancer from imaging: A systematic review. *Journal of Biomedical Informatics*, *129*, 104076. https://doi.org/10.1016/j.jbi.2022.104076

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, *24*(2), 1–11. https://doi.org/10.1093/bib/bbad002

Jackins, V. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *Journal of Supercomputing*, *77*(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x

Muntiari, N. R., & Hanif, K. H. (2022). Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Ilmu Komputer Dan Teknologi*, *3*(1), 1–6. https://doi.org/10.35960/ikomti.v3i1.766

Organization, W. H. (2023). Breast cancer: Key facts. In *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

R Wahid, T. O., & Affandi, D. (2022). Analysis of Clinical Risk Surgical Services To Support Services Quality At Arifin Achmad Riau Hospital. *JMMR (Jurnal Medicoeticolegal Dan Manajemen Rumah Sakit)*, *11*(1), LAYOUTING. https://doi.org/10.18196/jmmr.v11i1.11038

Salman, H. A., & Kalakech, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/bjml/2024/007

Shidqi, Z. N. (2022). Faktor-Faktor Keterlambatan Diagnosis Kanker Pada Pasien Kanker Payudara: Systematic Review. *Jurnal Epidemiologi Kesehatan Komunitas*, *7*(2), 471–481. https://doi.org/10.14710/jekk.v7i2.14911

Suparna, K., & Sari, L. M. K. K. S. (2022). Kanker Payudara: Diagnostik, Faktor Risiko, Dan Stadium. *Ganesha Medicine*, *2*(1), 42–48. https://doi.org/10.23887/gm.v2i1.47032

Vujović, Ž. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6609

Submitted : Jul 22, 2025
Accepted : Aug 3, 2025
Published : Aug 5, 2025

Widodo, E. (2023). Analisa Prediksi Hasil Produksi Popok Bayi Metode Naïve Bayes. *Bulletin of Information Technology (BIT)*, *4*(1), 75–80. https://doi.org/10.47065/bit.v4i1.504

Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, *292*(1), 60–66. https://doi.org/10.1148/radiol.2019182716

Zhou, L., Pan, S., Wang, J., Vasilakos, A. V, & Liu, Y. (2022). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *469*, 364–383. https://doi.org/10.1016/j.neucom.2021.10.024

* Corresponding author