Ensembled Voting Techniques for Advanced Breast Cancer Prediction

Riska Kurnia Septiani[1], Yan Rianto[2]

[1,2] Ilmu Komputer, Universitas Nusa Mandiri, Indonesia

[1]14220021@nusamandiri.ac.id, [2*]yan.yrt@nusamandiri.ac.id

Informasi Artikel
Received: December 2024
Revised: February 2024
Accepted: March 2024
Published: June 2024

Abstract

Breast cancer is the most common type of cancer affecting women worldwide, with a significant increase in incidence rates each year. Information and Communication Technology (ICT) has made substantial contributions to the medical field, particularly through the use of Big Data and machine learning algorithms to enhance diagnostic accuracy and healthcare efficiency. This research aims to assess the performance of five breast cancer classification algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), k-Nearest Neighbors (k-NN), Logistic Regression, and Ensembled Voting, using the Breast Cancer Wisconsin (Diagnostic) dataset. The study findings indicate that all models achieved high levels of accuracy, precision, recall, and F1-Score, with Ensembled Voting reaching the highest accuracy of 98.57%. This study confirms that machine learning algorithms, particularly Ensembled Voting, can be relied upon to improve breast cancer diagnosis accuracy, thereby significantly contributing to better healthcare outcomes.

Keywords: Breast Cancer; Machine Learning; Ensembled Voting
Kata kunci: Kanker Payudara; Machine Learning; Ensembled Voting

Abstrak

Kanker payudara adalah jenis kanker paling umum yang diderita oleh perempuan di seluruh dunia, dengan peningkatan signifikan dalam angka kejadian setiap tahunnya. Teknologi Informasi dan Komunikasi (ICT) telah memberikan kontribusi besar dalam bidang medis, terutama melalui penggunaan Big Data dan algoritma machine learning untuk meningkatkan akurasi diagnosis dan efisiensi perawatan kesehatan. Penelitian ini bertujuan untuk menilai performa lima algoritma klasifikasi kanker payudara: Support Vector Machine (SVM), Decision Tree (C4.5), k-Nearest Neighbors (k-NN), Logistic Regression, dan Ensembled Voting, menggunakan dataset Breast Cancer Wisconsin. Hasil penelitian menunjukkan bahwa semua model memiliki tingkat akurasi, presisi, recall, dan F1-Score

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

yang tinggi, dengan Ensembled Voting mencapai akurasi tertinggi sebesar 98.57%. Studi ini menegaskan bahwa algoritma machine learning, terutama Ensembled Voting, dapat diandalkan untuk meningkatkan akurasi diagnosis kanker payudara, memberikan kontribusi signifikan terhadap perawatan kesehatan yang lebih baik.

1. Introduction

Breast cancer is the most common type of cancer experienced by women both in Indonesia and globally. According to the Global Observatory Cancer (GLOBOCAN) 2020 from the International Agency of Research on Cancer, breast cancer ranks first among the 10 most common cancers worldwide. Breast cancer patients account for nearly 478% of all cancer patients (WHO 2021). The incidence rate of breast cancer increases significantly every year. It is estimated that the incidence and mortality rates will continue to rise significantly in the next 5-10 years [1].

The role of Information and Communication Technology (ICT) in cancer care includes the advancements achieved by Big Data in data size and value creation. Big Data, often associated with data mining, business analytics, and business intelligence, has had a major impact on medical science by improving prediction outcomes, reducing medical costs, enhancing patient health, and improving healthcare quality and real-time decision making.

In previous research, Hiba Asri, Hajar Mousannif, Hassan Al Moatassim, and Thomas Noel compared the performance of several machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer dataset (original). The main objective of this research was to assess the accuracy in classifying data by considering the efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity, and specificity. Experimental results showed that SVM provided the highest accuracy (97.13%) with the lowest error rate. All experiments were conducted in a simulated environment using the WEKA data mining tool [2].

The study titled "An improved breast cancer disease prediction system using ML and PCA" by S. Laghmati, S. Hamida, Hicham K., and others demonstrated that the XGBoost model achieved the highest recall of over 96% for the Mammographic Mass dataset. For the WBCD (Wisconsin Breast Cancer Dataset), both the AdaBoost model and the S-LR model outperformed others with Recall reaching 95.35%. The ensemble stacking model with logistic regression achieved the highest accuracy of 93.37% for the Mammographic Mass dataset and 97.37% for the WBCD [3].

This study aims to examine how five different classifiers, namely Support Vector Machine (SVM), Decision Tree (C4.5), k Nearest Neighbors (k-NN), Logistic Regression, and Ensembled Voting, operate. These algorithms are considered some of the most influential data mining algorithms in research and are included in the top ten data mining algorithms.

The focus of this research is to assess how efficiently and effectively these algorithms are used in terms of accuracy, sensitivity, specificity, and precision. Other parts of the paper are

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

structured by discussing related literature, the context of experiments, comparison of experimental results, discussion of findings, and drawing conclusions.

This research employs the Wisconsin Breast Cancer Dataset (WBCD) for experimentation, a widely used benchmark dataset in breast cancer research. The dataset contains features computed from digitized images of fine needle aspirates (FNA) of breast masses. These features are used to predict whether a mass is benign or malignant. The classifiers' performances will be evaluated based on metrics such as accuracy, sensitivity, specificity, and precision. The outcomes will provide insights into the strengths and weaknesses of each algorithm in accurately diagnosing breast cancer cases.

The methodology involves preprocessing the dataset to handle missing values, normalize features, and possibly perform feature selection using techniques like Principal Component Analysis (PCA). Each classifier will be trained using a portion of the dataset and validated on another portion to ensure robustness and generalizability of the results. The experimental setup aims to simulate real-world scenarios to validate the applicability of these machine learning models in clinical settings for breast cancer diagnosis and prognosis.

The discussion in this study will also include a comparison of the experimental results with related research using the same or similar datasets. This aims to provide a broader context on the relative strengths and practical applicability of each algorithm in the context of early detection and management of breast cancer. It is expected that the results of this research will contribute significantly to the development of more efficient and effective prediction and diagnosis systems for breast cancer using state-of-the-art technologies in data mining and machine learning.

2. Methodology

This section details the stages carried out to design and conduct the research, starting from the selection of methods to data analysis. The research stages are shown in Figure 1.

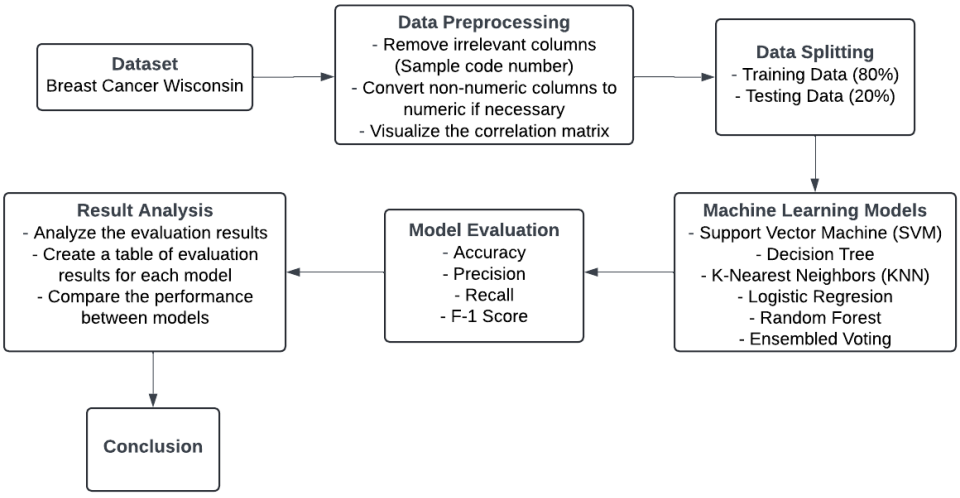


Figure 1. Research Stages

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

2.1. Dataset

The data used in this research is the Breast Cancer Wisconsin dataset, developed by Dr. William H. Wolberg. This dataset has been widely used for breast cancer research and is publicly available on the UCI Machine Learning Repository. It contains 569 instances, with 32 attributes describing various characteristics of cell nuclei present in the breast mass, including a diagnosis label (malignant or benign). Data pre-processing was carried out to handle missing values and normalize the data. Attributes that were not relevant to the classification task were removed.

The dataset includes various attributes that describe characteristics of breast cancer cells. H1: Sample Code Number serves as a unique identifier for each sample without predictive value. H2: Clump Thickness measures the density of cell clumps, with higher values indicating potential abnormalities. H3: Uniformity of Cell Size assesses the consistency in cell sizes, where greater variability can signal malignancy. H4: Uniformity of Cell Shape evaluates the regularity of cell shapes, with irregularities often associated with cancer. H5: Marginal Adhesion indicates how well cells adhere to each other, with poor adhesion suggesting cancer. H6: Single Epithelial Cell Size measures the size of individual epithelial cells, with abnormal enlargement potentially indicating cancer. H7: Bare Nuclei reflects the proportion of nuclei not surrounded by cytoplasm, with higher values often found in malignant cells. H8: Bland Chromatin assesses chromatin texture, with coarse chromatin linked to malignancy. H9: Normal Nucleoli measures nucleoli visibility, with more visible nucleoli generally associated with malignancy. H10: Mitoses describes mitotic activity, with higher values indicating increased cell division, a common feature of cancer. Finally, H11: Class is the target variable used to classify samples as benign (2) or malignant (4), which is essential for training and evaluating machine learning models.

Tabel 1. Wisconsin Breast Cancer Dataset Information

#	Attribute	Domain
H1	Sample code number	Id number
H2	Clump Thickness	1-10
H3	Uniformity of Cell Size	1-10
H4	Uniformity of Cell Shape	1-10
H5	Marginal Adhesion	1-10
H6	Single Epithelial Cell Size	1-10
H7	Bare Nuclei	1-10
H8	Bland Chromatin	1-10
H9	Normal Nucleoli	1-10
H10	Mitoses	1-10
H11	Class	(2 for benign, 4 for malignant)

Tabel 2. Description of Indicators in the Wisconsin Breast Cancer Dataset

No	Indicator	Description
1	Clump Thickness	Evaluates whether the cell is mono or multi-layered
2	Uniformity of cell size	Evaluates the consistency of cell sizes within the sample
3	Uniformity of cell shape	Evaluates the consistency of cell shapes within the sample
4	Marginal Adhesion	Calculates the proportion of cells that are adhering to each other
5	Single Epithelial Cell Size	Measures the enlargement of epithelial cell sizes
6	Bare Nuclei	Proportion of nuclei surrounded by cytoplasm versus those that are not
7	Bland Chromatin	Assesses the "texture" of the nucleus in a range from smooth to coarse
8	Normal Nucleoli	Determines whether nucleoli are small and barely visible or more visible
9	Mitoses	Describes the level of mitotic activity

2.2. Data Preprocessing

In this data preprocessing step, first, the `Sample_code_number` column is removed from the DataFrame as it is not relevant for further analysis. The next step is to identify columns that have non-numeric data types. These columns are stored in the variable `non_numeric_columns`. After that, all non-numeric columns are removed from the DataFrame to obtain a DataFrame that only contains numeric columns, which is stored in `df_numeric`. Alternatively, non-numeric columns could be converted to numeric, but this step is omitted in this code. The final step is to calculate the correlation matrix from the `df_numeric` DataFrame and visualize it in the form of a heatmap using the Seaborn library. This heatmap helps to see the relationships between each pair of numeric variables, which can be useful in further data analysis.

2.3. Dataset Splitting

The previously processed dataset is divided into training and testing sets with a ratio of 80:20. The training set is used to train the machine learning model, ensuring that the model gains an adequate understanding of the data.

2.4. Machine Learning Models

The models used include SVM, Decision Tree, Random Forest, Logistic Regression, and Ensemble Voting. SVM is effective for classification by separating data using an optimal hyperplane. Decision Tree maps features to outcomes using a tree structure, but is prone to overfitting [4]. Random Forest addresses overfitting by combining several decision trees from different data subsets, enhancing robustness and accuracy [5]. Logistic Regression predicts the probability of events for binary classification using a logistic function [6]. Ensemble Voting combines predictions from various models to produce a more accurate and stable final prediction by taking the majority vote from each model's prediction. This combination is expected to improve prediction performance and accuracy.

2.4.1. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning technique commonly used for classification (such as Support Vector Classification) and regression (Support Vector Regression) for both linear and nonlinear data. SVM performs classification by selecting a decision boundary that optimizes the distance (maximum margin classifier) from the closest data points of each class. The decision boundary produced by SVM is known as a maximum

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

margin classifier or maximum margin hyperplane [7]. The concept of classification with SVM is to find the best hyperplane that serves as a separator between two data classes [8]. The decision function for SVM can be written as:

$$f(x) = sign(w.x + b) \quad (1)$$

w is the weight vector
x is the input feature vector
b is the bias term

2.4.2. Decision Tree (C4.5)

A Decision Tree resembles a tree structure with nodes as test data, branches as results of those tests, and leaf nodes as specific classes. The top node is known as the root node. This algorithm determines the entropy value for each attribute and then compares the Gain value for each property. Decision Trees are useful for making decisions and analyzing information from relevant attributes to produce classifications that fit the dataset. In the Decision Tree model, the node with the largest Gain value will develop into a new node. Attributes that have been processed will not be counted again. This process continues until a fact or leaf node is found [9]. Here is the equation for the Decision Tree formula:

$$\mathrm{G(S)} = 1 - \sum_{i}^{c} = 1\left(P\,{}_{i}^{2}\right) \quad (2)$$

Pi is the proportion of samples in class i
c is the number of classes

2.4.3. K-Nearest Neighbors (k-NN)

K-Nearest Neighbors (k-NN) is a classification method that utilizes training data to determine an object's class based on the nearest distance. This process involves transforming training data into a multidimensional space where each dimension represents data characteristics. The k-NN algorithm is quite simple and works by calculating the nearest distance between the query object and the training samples to determine the optimal number of neighbors. Then the majority of the neighbors found are used to predict the query object's class. To determine the optimal k value, parameter optimization such as k-fold cross-validation can be performed [10]. The formula for the Euclidean distance between two points is:

$$d(x,y) = \sqrt{\sum_{\{i=1\}}^{\{N\}} (x_i - y_i)^2} \quad (3)$$

N is the number of features

2.4.4. Logistic Regresion

Logistic Regression is a reliable classification method for predicting discrete probabilities with superior performance. In its application, logistic regression uses a logistic function to measure

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

the probability value of an event, producing binary outputs of 0 or 1. The logistic or sigmoid function transforms values from the range of minus infinity to plus infinity into the range between 0 and 1, so the output can be interpreted as the probability of a positive event. Logistic regression is used when the model needs to predict the likelihood of two different classes occurring [11]. The logistic function is given by:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

e is the base of the natural logarithm
x is the linear combination of features

2.4.5. Random Forest

Breiman first introduced the Random Forest method in 2001. The Random Forest method has two main functions to solve a case, namely classification and prediction. The basic technique used in the Random Forest method is the decision tree. In other words, the Random Forest method is a collection of decision trees used for data classification and prediction, where data is entered into the root at the top and then goes down to the leaves at the bottom. The analysis results of the Random Forest method for classification are the form of each tree formed, while the prediction results are obtained from the average value of each tree [12].

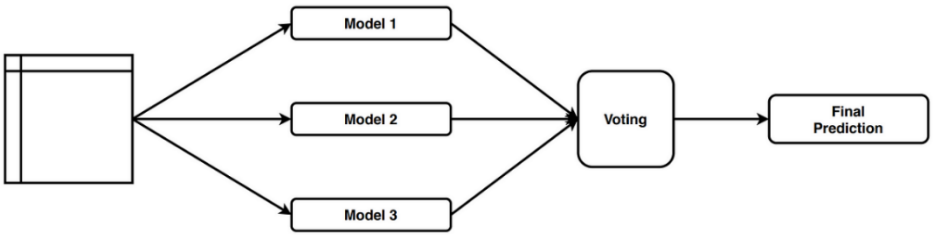
$$\widehat{\{y\}} = \frac{1}{N} \sum_{\{i=1\}}^{\{N\}} T_i(x) \tag{5}$$

$T_i$is the prediction from the i-th tree
N is the number of trees

2.4.6. Ensembled Voting

The ensemble method is a technique used to improve machine learning model performance and accuracy by combining the prediction results of several different models or algorithms. The term ensemble refers to the concept of combining many elements into one unit. Combining various machine learning approaches is expected to overcome the individual weaknesses of each model and produce more accurate and stable predictions. The way the ensemble method works involves using several diverse models in one team, where the prediction results of each model are combined to make the final decision [13].

Ensemble voting is a technique in machine learning where several models or classifiers are combined to improve overall prediction performance. In ensemble voting, each model casts a



ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

vote or prediction for the desired output (e.g., class or value), and the final result is decided based on the majority vote or pre-determined weights. This approach can reduce overfitting and increase model stability, especially when used in combination with various types of different algorithms.

Figure 2. Ensembled Votting

2.4.7. Confusion Matrix

A confusion matrix is a method used to measure the accuracy of a classifier. This method can determine the accuracy, specificity, and sensitivity of the resulting classes [14]. The table below shows a confusion matrix indicating 4 different combinations of predicted and actual values.

Table 3. Confusion Matrix

	Positive	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FP)	True Negative (TN)

The performance metrics used include F1 score, Precision, Recall, and Accuracy. The F1 score, by definition, is the harmonic mean of precision and recall, thus combining both aspects. F1-score computes the average value of Recall and Precision with balanced weighting [15].

$$F1\ Score = 2\ \times \frac{recall \times precission}{recall + precission} \tag{6}$$

Recall indicates the ratio of true positive predictions to the total number of actual positive data.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{7}$$

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive} \tag{8}$$

Meanwhile, Precision reflects the ratio of true positive predictions to the total positive predictions made.

$$Precission = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{9}$$

$$Precission = \frac{True\ Positive}{Total\ Predicted\ Positive} \tag{10}$$

Next, to evaluate good results, we look at the level of accuracy. Accuracy is the ratio of correct predictions to the total predictions made by an algorithm.

$$Accuracy = \frac{(TP + TN)}{(TP + FN + TN + FN)} \tag{11}$$

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

3. Results and Discussion

The performance of the machine learning models for binary classification on the Breast Cancer dataset is summarized in the table 4 below. Each model is evaluated using accuracy, precision, recall, and F1-Score. Accuracy indicates the proportion of correct predictions out of all predictions made by the model. Precision measures the proportion of true positive predictions, which is important for reducing false positives in the diagnosis of serious diseases. Recall measures the proportion of actual positive cases that are correctly identified, crucial for detecting all cancer cases. The F1-Score, as the harmonic mean of precision and recall, provides a balanced evaluation, especially on datasets with class imbalance, thus giving a more holistic view of the model's performance.

Table 4. Performance of the Support Vector Machine (SVM) Model

Metric	SVM	Decision Tree	k-NN	Logistic Regression	Random Forest	Ensembled Voting
Precision (2)	0.98	0.96	0.98	0.97	0.98	0.99
Precision (4)	0.93	0.91	0.96	0.98	0.96	0.98
Recall (2)	0.97	0.96	0.98	0.99	0.98	0.99
Recall (4)	0.96	0.91	0.96	0.93	0.96	0.98
F1-Score (2)	0.97	0.96	0.98	0.98	0.98	0.99
F1-Score (4)	0.95	0.91	0.96	0.95	0.96	0.98
Accuracy	0.9642	0.9428	0.9714	0.9714	0.9714	0.9857
Macro Avg (P)	0.96	0.93	0.97	0.97	0.97	0.98
Macro Avg (R)	0.96	0.93	0.97	0.96	0.97	0.98
Macro Avg (F1)	0.96	0.93	0.97	0.97	0.97	0.98
Weighted Avg (P)	0.96	0.94	0.97	0.97	0.97	0.99
Weighted Avg (R)	0.96	0.94	0.97	0.97	0.97	0.99
Weighted Avg (F1)	0.96	0.94	0.97	0.97	0.97	0.99
Support (2)	95	95	95	95	95	95
Support (4)	45	45	45	45	45	45

The evaluation results of various classification models for breast cancer diagnosis reveal notable performance across the board. The Support Vector Machine (SVM) model shows impressive metrics with a precision of 0.98, recall of 0.97, and an F1-Score of 0.97 for class 2 (benign), while achieving a precision of 0.93, recall of 0.96, and an F1-Score of 0.95 for class 4 (malignant). The overall accuracy is 0.9642, indicating that 96% of predictions are correct. The macro and weighted averages for precision, recall, and F1-Score are consistently 0.96, reflecting reliable performance across classes.

The Decision Tree model also demonstrates strong performance, with precision values of 0.96 for class 2 and 0.91 for class 4, and an overall accuracy of 0.9428. This model’s balanced macro and weighted averages suggest effectiveness in classifying breast cancer. Similarly, the K-Nearest Neighbors (KNN) model excels with precision, recall, and F1-Score around 0.98 for class 2 and 0.96 for class 4, achieving an overall accuracy of 0.9714, which underscores its high precision and reliability.

The Logistic Regression model achieves an overall accuracy of 0.9714, with precision scores of 0.97 for class 2 and 0.98 for class 4, and recalls of 0.99 and 0.93, respectively. The F1-Scores are 0.98 for class 2 and 0.95 for class 4, indicating a strong balance between precision and recall.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

Macro averages for precision, recall, and F1-Score are 0.97, 0.96, and 0.97 respectively, although there is a class imbalance with 95 examples for class 2 and 45 for class 4.

The Random Forest model performs excellently, with precision, recall, and F1-Score metrics of 0.98 for class 2 and 0.96 for class 4. Its overall accuracy is 0.9714, with macro and weighted averages for precision, recall, and F1-Score at 0.97, demonstrating high accuracy and strong consistency.

Finally, the Ensembled Voting model delivers outstanding results with precision, recall, and F1-Score of 0.99 for class 2 and 0.98 for class 4, achieving an overall accuracy of 0.9857. The macro average for precision, recall, and F1-Score is 0.98, while the weighted averages are all 0.99, indicating exceptional performance and balance across classes.

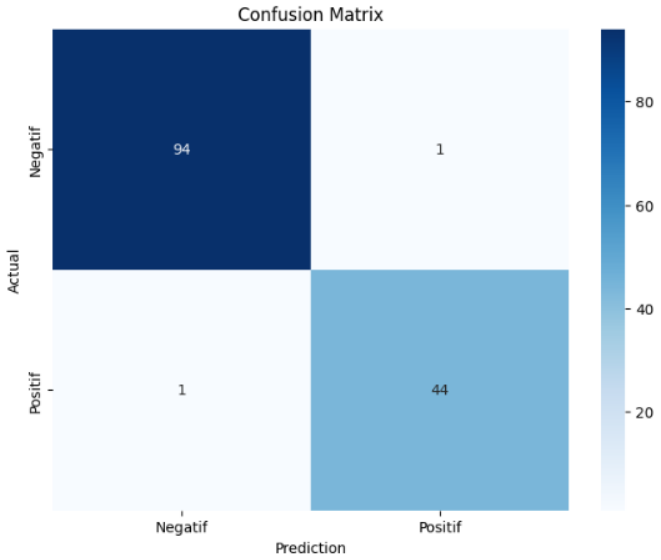


Figure 3. Confusion Matrix

The confusion matrix indicates that the model correctly identified 94 instances as True Negatives (TN), meaning these were correctly predicted as negative cases. It made 1 False Positive (FP) error, where a negative case was incorrectly predicted as positive. Similarly, it had 1 False Negative (FN) error, where a positive case was incorrectly predicted as negative. Additionally, the model accurately identified 44 True Positives (TP), correctly predicting these positive cases. This suggests the model performs well with high accuracy, correctly classifying the majority of cases.

4. Conclusion and Suggestions

The evaluation results of various breast cancer classification models reveal consistently high performance across all models. SVM, Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Ensembled Voting models achieve impressive overall accuracies ranging from 0.94 to 0.99, coupled with high precision, recall, and F1-Score for both cancer classes (class 2 and class 4). Among them, the Ensembled Voting model stands out with the highest overall accuracy of 0.9857 and a robust balance between precision, recall, and F1-Score across all classes, establishing it as the most reliable model for breast cancer diagnosis.

Each model demonstrates strong classification capabilities, with Ensembled Voting particularly excelling in the comprehensive analysis of the tested data.

These findings underscore the effectiveness of machine learning algorithms in breast cancer diagnosis, highlighting Ensembled Voting as a leading choice due to its superior performance metrics. Future research could focus on further enhancing these models' interpretability and scalability in clinical settings, thereby advancing their utility in real-world medical applications for improved patient care and outcomes.

Daftar Pustaka

[1] Y. R. Putri *et al.*, "Konsep Analisis Adaptasi Psikologis Pada Fase Awal Kanker Payudara," *J. Endur.*, vol. 7, no. 1, pp. 192–198, 2022, doi: 10.22216/jen.v7i1.839.

[2] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[3] S. Laghmati, S. Hamida, K. Hicham, B. Cherradi, and A. Tmiri, *An improved breast cancer disease prediction system using ML and PCA*, vol. 83, no. 11. 2024. doi: 10.1007/s11042-023-16874-w.

[4] D. Muriyatmoko, A. Musthafa, and M. H. Wijaya, "Klasifikasi Profil Kelulusan Nilai AKPAM Dengan Metode Decision Tree," no. April, 2024.

[5] P. Simamora, S. A. Pasaribu, and V. Wijaya, "Peningkatan dan Optimalisasi Prediksi Harga Emas Menggunakan Metode Combine Machine Learning Random Forest dan Gradient Boosting," vol. 01, pp. 42–51, 2024.

[6] J. Ilmiah and T. Informasi, "SUBMIT PREDICTION OF PATIENTS INDICATED WITH HEART DISEASE USING," vol. 4, no. 1, pp. 19–23, 2024.

[7] K. Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 1, pp. 35–46, 2021, doi: 10.28932/jutisi.v7i1.3182.

[8] S. Y. Pangestu, Y. Astuti, and L. D. Farida, "Algoritma Support Vector Machine Untuk Klasifikasi Sikap Politik Terhadap Partai Politik Indonesia," *J. Mantik Penusa*, vol. 3, no. 1, pp. 236–241, 2019, [Online]. Available: https://t.co/eF

[9] S. A. Pratiwi, A. Fauzi, S. Arum, P. Lestari, and Y. Cahyana, "KLIK: Kajian Ilmiah Informatika dan Komputer Prediksi Persediaan Obat Pada Apotek Menggunakan Algoritma Decision Tree," *Media Online*, vol. 4, no. 4, pp. 2381–2388, 2024, doi: 10.30865/klik.v4i4.1681.

[10] A. Naufal Hilmi *et al.*, "Implementasi Algoritma K-Nearest Neighbor (KNN) untuk Identifikasi Penyakit pada Tanaman Jeruk Berdasarkan Citra Daun," no. 2, pp. 107–117, 2024, [Online]. Available: https://doi.org/10.62951/router.v2i2.78

[11] Brury Barth Tangkere, "Analisis Performa Logistic Regression dan Support Vector Classification untuk Klasifikasi Email Phising," *J. Ekon. Manaj. Sist. Inf.*, vol. 5, no. 4, pp. 442–450, 2024, doi: 10.31933/jemsi.v5i4.1916.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v21i2. 13004

[12] S. Mahmuda, "Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube," *J. Jendela Mat.*, vol. 2, no. 01, pp. 21–31, 2024, doi: 10.57008/jjm.v2i01.633.

[13] M. Fadel, Z. Arifin, G. Triyono, T. Faculty, and U. Budi, "Application of Ensemble Method for Employee Turnover Penerapan Metode Ensemble Untuk Prediksi Turnover," vol. 5, no. 3, pp. 767–775, 2024.

[14] I. T. Batam, I. T. Batam, A. Jl, G. Mada, K. V. City, and T. A. Sekupang, "Analisis Ramalan Cuaca di Sekupang , Kota Batam Menggunakan Algoritma Decision Tree dan Confusion Matrix Fitrianingsih (KBBI), keadaan , udara (tentang suhu , cahaya matahari , kelembaban , kecepatan angin , dan," no. 4, 2024.

[15] A. Muhaimin, M. Amin Hariyadi, and M. I. Imamudin, "Klasifikasi Prestasi Akademik Siswa Berdasarkan Nilai Rapor dan Kedisiplinan dengan Metode K-Nearest Neighbor," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 193–202, 2024, doi: 10.55338/jikomsi.v7i1.2865.