**Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)**
Volume 23 ; Nomor 1 ; Februari 2024 ; Page 97-104
E-ISSN : 2615-3475; P-ISSN : 1978-6603
*https://ojs.trigunadharma.ac.id/index.php/jis/index*

# Perbandingan Metode *Ensemble* Untuk Meningkatkan Akurasi Algoritma *Machine Learning* Dalam Memprediksi Penyakit *Breast Cancer* (Kanker Payudara)

**Annisa Maulana Majid[1], Ismasari Nawangsih[2]**
[1] Teknik Informatika, Universitas Pelita Bangsa, Bekasi, Indonesia
[2]Teknik Informatika, Universitas Pelita Bangsa, Bekasi, Indonesia
Email: [1]annisa.maulanamajid@pelitabangsa.ac.id , [2]ismasari.n@pelitabangsa.ac.id
Email Korespondensi: annisa.maulanamajid@pelitabangsa.ac.id

**Article History:**
Received Dec 06th, 2023
Revised Dec 28th, 2023
Accepted Jan 15th, 2024

**Abstrak**

*Machine Learning* merupakan suatu teknologi pembelajaran mesin yang dapat digunakan untuk memudahkan pekerjaan berbagai bidang, salah satunya yaitu pada bidang kesehatan. *Machine Learning* dalam bidang kesehatan dapat digunakan dalam memprediksi atau mendiagnosa suatu penyakit yang dihasilkan berdasarkan *dataset.* Kanker payudara (*breast cancer)* merupakan salah satu penyakit yang mematikan khususnya banyak diderita oleh wanita, oleh karena itu perlu adanya diagnosa dini terkait penyakit kanker payudara agar penanganan dapat dilakukan dengan tepat serta mencegah adanya penyebaran kanker pada tubuh. Penelitian sebelumnya telah membahas tentang diagnosa penyakit kanker payudara namun tingkat akurasi masih rendah sehingga perlu adanya teknik peningkatan akurasi untuk dapat memberikan informasi yang akurat. Tujuan dalam penelitian ini yaitu membandingkan metode *Ensemble* menggunakan algoritma *Machine Learning* yaitu *Decision Tree, Naïve Bayes*, dan *K-Nearest Neighbor* (KNN), untuk meningkatkan akurasi dalam memprediksi penyakit Kanker Payudara. Metode *Ensemble* yang digunakan dalam penelitian ini yaitu *Adaboost* dan *Bagging*. Hasil penelitian menunjukkan bahwa terdapat peningkatan pada algoritma klasifikasi menggunakan metode *Ensemble.* Metode paling unggul yaitu Algoritma *Decision Tree* dan Metode *Ensemble* yang menghasilkan akurasi sebesar yaitu 82.76%. Pada nilai AUC tertinggi diperoleh dari algoritma KNN yang dikombinasikan dengan metode *Bagging* yaitu sebesar 0.950 dengan kategori sangat baik.

**Kata Kunci :** *Breast cancer, Machine Learning, metode Ensemble, Adaboost, Bagging*

***Abstract***

*Machine Learning is a machine learning technology that can be used to facilitate work in various fields, one of which is the health sector. Machine Learning in the health sector can be used to predict or diagnose a disease that is generated based on a dataset. Breast cancer is a deadly disease, especially among women, therefore there is a need for early diagnosis of breast cancer so that treatment can be carried out appropriately and prevent the spread of cancer in the body. Previous research has discussed breast cancer diagnosis, but the level of accuracy is still low, so there is a need for techniques to increase accuracy to be able to provide accurate information. The aim of this research is to compare the Ensemble method using Machine Learning algorithms, namely Decision Tree, Naïve Bayes, and K-Nearest Neighbor (KNN), to increase accuracy in predicting Breast Cancer. The Ensemble methods used in this research are Adaboost and Bagging. The research results show that there is an improvement in the classification algorithm using the Ensemble method. The most superior methods are the Decision Tree Algorithm and the Ensemble method which produces an accuracy of 82.76%. The highest AUC value was obtained from the KNN algorithm combined with the Bagging method, namely 0.950 in the very good category.*

***Keyword :*** *Breast cancer, Machine Learning, metode Ensemble, Adaboost, Bagging*

## 1. PENDAHULUAN

Teknologi saat ini semakin berkembang pesat termasuk teknologi menggunakan *Machine Learning*. *Machine Learning* merupakan salah satu teknologi pembelajaran yang dapat digunakan untuk mempermudah suatu pekerjaan. *Machine learning* merupakan bagian dari kecerdasan buatan yang menerapkan algoritma dan metode yang digunakan untuk prediksi, pengenalan pola, dan klasifikasi [1]. Salah satu bidang yang dapat diterapkan menggunakan teknologi

*Machine Learning* yaitu bidang kesehatan dengan memprediksi atau mendiagnosa penyakit pasien. Prediksi penyakit menggunakan teknologi *Machine Learning* akan memudahkan tenaga medis untuk dapat mendiagnosa penyakit secara dini sehingga pasien dapat langsung ditangani dengan tepat.

Penyakit kanker merupakan salah satu penyakit mematikan. Pada wanita kanker payudara dapat menjadi hal yang rentan seiring bertambahnya usia. Pravelensi penyakit Kanker berdasarkan diagnosis dokter meningkat dari tahun 2013 ke tahun 2018 [2]. Data dari Global Cancer Observatory World Health Organization (WHO) menunjukkan bahwa kanker payudara merupakan kasus tingkat tertinggi dengan ranking ke 1 jumlah pasien sebanyak 65.858, dengan persentase sebesar 30.8% pada tahun 2020. Hal ini dapat terjadi karena adanya keterlambatan dalam penanganan kasus kanker payudara sehingga terjadi peningkatan pada jumlah pasien penderita kanker payudara.

Perlu adanya penanganan pasien sejak stadium dini untuk mengurangi angka kematian pada pasien penderita kanker payudara, sehingga perlu diagnosa dini terkait penyakit tersebut. Penelitian terkait diagnosa penyakit kanker payudara telah dilakukan menggunakan algoritma *Machine Learning* seperti *Random Forest, Neural Network, Decision Tree, Naïve Bayes, K-Nearest Neighbor, Logistic Regresion, Random Forest*, dan *Support Vector Machines*. Namun hasil akurasi menunjukkan hasil yang tidak signifikan sehingga perlu adanya peningkatan akurasi pada algoritma *Machine Learning*.

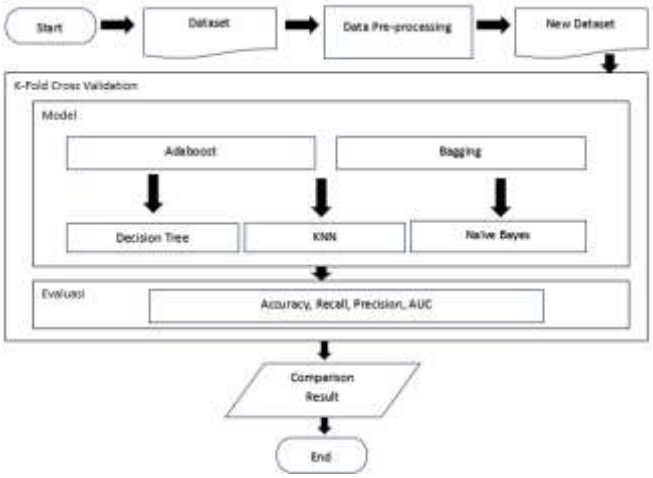
Penelitian oleh Nosayba Al-Azaam dkk. menggunakan *dataset Wisconsin Diagnostic Breast Cancer* (WDBC) menghasilkan tingkat akurasi 91% untuk algoritma *Decision Tree*, 95% untuk algoritma *Naïve Bayes*, 98% akurasi untuk algoritma KNN dan 97% untuk algoritma SVM [3]. Penelitian oleh Vincent Peter C. Magboo dan Ma. Sheila A. Magboo menggunakan *dataset Wisconsin Prognostic Breast Cancer* (WPBC) yang berasal UCI *Machine Learning Repository* menghasilkan akurasi 80% untuk algoritma *Logistic Regression*, 60% untuk algoritma *Naïve Bayes*, 60% akurasi untuk KNN, dan 75% akurasi untuk algoritma SVM [4]. Penelitian oleh Yufan Feng dkk menggunakan *dataset Asia-Pacific Metaplastic Breast Cancer* (AP-MBC) menghasilkan akurasi 83,1% untuk metode *Bagging*, 82,5% akurasi untuk algoritma *Logistic*, 81,3% untuk algoritma *Multilayer Perceptron*, 79,4% akurasi untuk algoritma *Naïve Bayes*, dan 83,8% akurasi untuk algoritma *Random Forest* [5]. Penelitian oleh Varsha Nemade dan Vishal Fegade menggunakan *dataset* WDBC dari UCI *Machine Learning Repository*, menghasilkan tingat akurasi 96% untuk algoritma KNN, 95% untuk algoritma SVM, 97% untuk algoritma *Decision Tree*, 90% untuk algoritma *Naïve Bayes*, 96% untuk algoritma *Logistic Regression*, 96% untuk teknik *Random Forest*, 96% untuk teknik *Adaboost*, dan 97% untuk algoritma *XGBoost* [6]. Penelitian oleh Nadya Meilani dan Odi Nurdiawan menggunakan *dataset* dari UCI *Machine Learning Repository*, menghasilkan tingkat akurasi 72,62% untuk algoritma KNN [7].

Metode *Ensemble* merupakan salah satu metode yang dapat meningkatkan akurasi. Metode *Ensemble* salah satunya yaitu *Adaboost* dan *Bagging*. *Adaboost* merupakan algoritma yang dapat fokus terhadap *misclassified tupple* sehingga dapat meningkatkan akurasi. *Bagging* merupakan salah satu metode *Ensemble* dengan teknik pembelajaran secara parallel pada tiap base model kemudian digabung.kan untuk menghasilkan hasil yang terbaik. Penelitian ini akan membandingkan metode *Ensemble* menggunakan algoritma *Machine Learning* yaitu *Decision Tree, Naïve Bayes,* dan *K-Nearest Neighbor* (KNN), untuk meningkatkan akurasi dalam memprediksi penyakit Kanker Payudara.

## 2. METODOLOGI PENELITIAN

### 2.1 Tahapan Penelitian

Penelitian ini menerapkan metode *Ensemble* yaitu *Adaboost* dan *Bagging* pada algoritma klasifikasi *Machine Learning* yaitu *Decision Tree, Naïve Bayes,* dan *K-Nearest Neighbor* (KNN). Tujuan penerapan metode *Ensemble* (*Adaboost* dan *Bagging)* yaitu untuk dapat meningkatkan akurasi dalam memprediksi penyakit kanker payudara (*breast cancer*). Tahapan dimulai dari pengambilan *dataset*, data *pre-processing*, kemudian penerapan metode dan algoritma, setelah kombinasi metode dan algoritma diterapkan, selanjutnya membandingkan model untuk mengetahui tingkat akurasi tertinggi. Tahapan-tahapan dalam penelitian sebagai berikut:




Gambar 1. Tahapan Penelitian

Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)
Volume 23 ; Nomor 1 ; Februari 2024 ; Page 97-104
E-ISSN : 2615-3475; P-ISSN : 1978-6603
https://ojs.trigunadharma.ac.id/index.php/jis/index

Tahapan penelitian dimulai dari pengambilan *dataset* melalui *dataset* public bersumber dari UCI *Machine Learning Repository*, kemudian dilakukan proses data *pre-processing* sebagai tahap pembersihan data dari *missing value* dan pengurangan data agar lebih efisien. Data *pre-processing* akan menghasilkan data baru yang siap untuk proses modelling dan penerapan metode. Sebelum tahap modelling dan penerapan metode dilakukan validasi data menggunakan *K-fold cross validation*, selanjutnya tahap modelling dengan menerapkan metode *Ensemble* yaitu *Adaboost* dan *Bagging* pada algoritma klasifikasi diantaranya *Decision Tree, Naïve Bayes*, dan KNN. Pengujian dilakukan sebanyak 9 kali diantaranya 3 metode merupakan pengujian algoritma klasifikasi secara mandiri, 3 pengujian menerapkan metode *Adaboost* pada algoritma klasifikasi dan 3 pengujian menerapkan metode *Bagging* pada algoritma klasifikasi. Pengujian akan menghasilkan nilai *Accuracy, Recall, Precision* dan AUC, selanjutnya dilakukan perbandingan pengujian untuk mendapatkan hasil yang akurat. Pengujian menggunakan aplikasi Rapidminer 10.3.

### 2.2. *Dataset*

Penelitian ini menggunakan *dataset breast cancer* yang bersumber dari UCI *Machine Learning Repository*. *Dataset breast cancer* berjumlah 286 data yang terdiri dari 10 atribut. *Dataset breast cancer* memiliki 2 class atau kategori yaitu *recurrence-events* (kambuh) dan *no-recurrence-events* (tidak kambuh). Berikut atribut dalam *dataset breast cancer*:

1. *Class: no-recurrence-events, recurrence-events*
2. *age*: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
3. menopause: lt40, ge40, premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. *inv-nodes*: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. *node-caps: yes, no*.
7. *deg-malig*: 1, 2, 3.
8. breast: left, right.
9. *breast-quad: left-up, left-low, right-up, right-low, central*.
10. *irradiat: yes, no*.

### 2.3 Data *Pre-Processing*

Data *pre-processing* digunakan untuk proses pembersihan data yang terdapat *missing value*. Data *pre-processing* merupakan tahapan untuk dapat mengevaluasi, mengidentifikasi, dan memperbaiki suatu kesalahan pada *dataset*, salah satunya yaitu memperbaiki adanya *missing value* [8]. *Missing value* dapat terjadi karena hilangnya informasi beberapa *record* data sehingga akan berpengaruh terhadap akurasi yang akan dihasilkan. Data *pre-processing* dilakukan pada *dataset breast cance*r yang terdapat beberapa *missing value* pada atribut node-caps sebanyak 8 *record*. Pada penelitian ini menggunakan teknik *data cleaning* pada tahap data *pre-processing*. Data *cleaning* merupakan suatu proses mendeteksi adanya *noisy data* atau *missing value* pada *dataset* sehingga dapat diperbaiki atau dihapus dari kumpulan *dataset* tersebut [8]. Pada penelitian ini dilakukan *pre-processing* data dengan menghapus *record* dengan nilai 0 pada atribut node-caps.

### 2.4 Validasi dan Evaluasi Data

Proses validasi dilakukan untuk dapat menguji model algoritma yang akan digunakan. *K- Fold Cross Validation* merupakan salah satu jenis pengujian yang memiliki fungsi untuk dapat menilai kinerja dari sebuah metode dan algoritma dengan cara membagi dan mengelompokkan sampel data sebanyak nilai k subset secara acak, satu kelompok subset digunakan untuk data *testing* (data uji) dan sisanya digunakan untuk data *training* (data latih) [9]. Nilai k merupakan jumlah iterasi yang akan digunakan. Proses validasi pada penelitian ini menggunakan *K-fold cross validation* dengan nilai k bernilai 10. 10 *fold cross validation* membagi data menjadi 10 subset dengan ukuran yang sama dengan ketentuan 1 bagian data menjadi data *testing* dan 9 data menjadi data *training*. Proses dilakukan berulang sebanyak 10 kali sampai dengan semua *record* data mendapatkan bagian menjadi *data testing* [10]. Alur kerja *k-fold cross validation* dimulai dari seluruh total *dataset* dibagi menjadi 10 bagian, bagian pertama menjadikan iterasi 1 sebagai data *testing* dan 9 bagian lainnya menjadi data *training*, kemudian bagian kedua menjadikan iterasi ke 2 sebagai data *testing* dan sisanya sebagai data *training*, proses akan terus berulang sampai mencapai *fold* ke 10, berikut ilustrasi proses 10-*fold cross validation*.

| *Dataset* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Test* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* |
| *Train* | *Test* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* |
| *Train* | *Train* | *Test* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* |
| *Train* | *Train* | *Train* | *Test* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* |
| *Train* | *Train* | *Train* | *Train* | *Test* | *Train* | *Train* | *Train* | *Train* | *Train* |
| *Train* | *Train* | *Train* | *Train* | *Train* | *Test* | *Train* | *Train* | *Train* | *Train* |
| *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Test* | *Train* | *Train* | *Train* |
| *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Test* | *Train* | *Train* |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Test* | *Train* |
| *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Train* | *Test* |

Gambar 2. Proses 10-*Fold Cross Validation*

Tahap validasi terdapat proses perhitungan model algoritma *Machine Learning* yaitu *Decision Tree*, *Naïve Bayes*, dan KNN yang dikombinasikan dengan metode *Ensemble* yaitu *Adaboost* dan *Bagging*. Setelah proses validasi dilakukan lanjut ke tahap evaluasi menggunakan *confusion matrix* dan ROC *curve*. *Confusion matrix* merupakan tabel untuk dapat menampilkan jumlah data uji yang akan diklasifikasikan dengan nilai benar dan salah, sehingga mudah dalam proses evaluasi hasil akurasi sistem klasifikasi. *Confusion matrix* adalah teknik analisa yang efektif dalam pengukuran kinerja sistem klasifikasi [11]. *Confusion matrix* dapat digunakan untuk menentukan model terbaik dengan mengukur kinerja dari model dan membandingkan dengan satu sama lain [12]. Berikut merupakan tabel *confusion matrix*:

Tabel 1. *Confusion Matrix*

| *Actual* | *Predictive* | |
|---|---|---|
| | *Positive* | *Negative* |
| *Positive* | *True Positive* (TP) | *False Negative* (FN) |
| *Negative* | *False Positive* (FP) | *True Negative* (TN) |

Tabel *confusion matrix* akan menghasilkan nilai *accuracy*, *recall*, dan *precision*. *Accuracy* merupakan hasil dari perbandingan total prediksi benar dengan total keseluruhan data. *Recall* merupakan perbandingan nilai benar/true positif dengan total keseluruhan nilai yang benar bernilai positif. *Precision* merupakan perbandingan nilai benar/true positif dengan total keseluruhan data positif. Rumus *accuracy*, *recall*, dan *precision* sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} X100\% \quad (1)$$

$$Recall = \frac{TP}{TP+FN} X100\% \quad (2)$$

$$Precision = \frac{TP}{TP+FP} X100\% \quad (3)$$

Kurva ROC (*Receiver Operating Characteristic*) merupakan suatu pengukuran untuk menentukan nilai dari kemampuan suatu sistem klasifikasi. Kurva ROC juga dapat digunakan untuk mengevaluasi suatu klasifikasi. Kurva ROC menggambarkan nilai AUC (*Area Under Curve*) yang merupakan perhitungan untuk memastikan klasifikasi yang unggul [13]. Berikut kategori dalam klasifikasi berdasarkan nilai AUC:

Tabel 2.2 Kategori Nilai AUC

| **Nilai AUC** | **Kategori** |
|---|---|
| 0.90-1.00 | Sangat Baik |
| 0.80-0.90 | Baik |
| 0.70-0.80 | Cukup |
| 0.60-0.70 | Buruk |
| 0.50-0.60 | Gagal |

**2.5 *Decision Tree***

*Decision Tree* merupakan penerapan dari program *Machine Learning* yang berbentuk seperti pohon keputusan dengan hasil pengujian menghasilkan nilai akurasi. *Decision Tree* merupakan algoritma yang digunakan untuk eksplorasi data untuk menemukan hubungan antara sejumlah variabel input dengan variabel target [14]. Algoritma *Decision Tree* merupakan salah satu jenis algoritma *Supervised Learning* yang terdiri dari node yang mewakili struktur cabang, kumpulan dari data yang dapat mewakili keputusan yang diberikan dari algoritma dan hasil yang diwakili oleh simpul daun [15]. Algoritma *Decision Tree* merupakan sebuah konsep dimana setiap node akan mewakili sebagai atribut dan cabang merupakan gambaran hasil pengujian disebut sebagai nilai atribut sedangkan daun akan mewakili sebagai kelas.

Algoritma *Decision Tree* merupakan salah satu algoritma klasifikasi yang populer dan dinilai efektif terkait pengklasifikasian dan prediksi. *Decision Tree* merepresentasikan sebuah kasus dengan banyaknya kriteria menggunakan pohon keputusan yang terdiri dari node dan simpul. Nilai gain tertinggi digunakan untuk menentukan atribut sebagai akar [16]. Persamaan rumus algoritma *Decision Tree* sebagai berikut:

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

Keterangan:
S = Himpunan Kasus
A = Atribut
n = Jumlah partisi atribut A
$|Si|$ = Jumlah kasus pada partisi ke -i
$|S|$ = Jumlah kasus dalam S

Perhitungan nilai entropy sebagai berikut:

$$Entropy\ (s) = \sum_{i=1}^{n} -pi * log_2 * pi \quad (5)$$

Keterangan:
S = Himpunan kasus
A = Fitur
n = Jumlah partisi S
Pi = Proporsi dari Si terhadap S

**2.6 *Naïve Bayes***

*Naïve Bayes* merupakan salah satu algoritma *Machine Learning* dengan jenis klasifikasi yang sederhana termasuk dalam kelompok probabilitas dengan dasar teorema bayes [17]. *Naïve Bayes* merupakan algoritma yang independensinya kuat, model algoritma sederhana dan dapat diimplementasikan dengan jumlah *dataset* yang besar [18]. *Naive Bayes* merupakan algoritma dengan tujuan untuk menemukan pemetaan dengan hasil terbaik antara data baru dan data dengan masalah yang tertentu [19]. *Naïve Bayes* algoritma dengan teknik yang sederhana berdasarkan algoritma *Bayesian* yang digunakan untuk membangun klasifikasi, model *Naïve Bayes* memiliki kinerja efisien yang baik dan stabil pada kumpulan data dengan skala kecil dan dapat menangani tugas klasifikasi [20].
Persamaan rumus algoritma *Naïve Bayes* sebagai berikut:

$$P(H|X) = \frac{P(X|H) * P(H)}{P(X)} \quad (6)$$

Keterangan:
X = Data yang belum diketahui nilai classnya
H = Hipotesis data X merupakan suatu class spesifik
P(H/X) = Probabilitas hipotesis H berdasarkan kondisi dari x (posteriori prob.)
P(H) = Probabilitas hipotesis H (prior prob.)
P(X/H) = Probabilitas X berdasarkan kondisi hipotesis
P(X) = Probabilitas dari X

**2.7 *K- Nearst Neighbor* (KNN)**

KNN adalah algoritma klasifikasi yang sederhana dengan mengklasifikasikan menurut data pelatihan terdekat, titik data ditentukan pada K terdekat yang diberikan tugas untuk klasifikasi data yang belum berlabel, [21]. Algoritma KNN salah satu dari bagian *Supervised Learning* yang mengklasifikasikan data berdasarkan kedekatan atau suatu jarak dari data ke data lainnya [22]. Algoritma KNN menghasilkan keputusan yang fleksibel bersifat nonparametik dengan tanpa memberikan asumsi secara distribusi, dalam pengklasifikasian objek dilihat berdasarkan pada data pelatihan yang memiliki jarak paling terdekat pada objek tersebut [23]. Algoritma yang mengatur tentang generalisasi sesuai dengan aturan tetangga terdekat dengan arti k-sample diuji dengan kemiripan terdekat [24]. Persamaan rumus algoritma KNN sebagai berikut:

$$d_i = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}^{\ 2} \quad (7)$$

Keterangan:

Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)
Volume 23 ; Nomor 1 ; Februari 2024 ; Page 97-104
E-ISSN : 2615-3475; P-ISSN : 1978-6603
https://ojs.trigunadharma.ac.id/index.php/jis/index
d = Jarak
x = Data Testing
y = Data Training
n = Dimensi data
i = Variable Data

**2.8 Metode *Ensemble***

Metode *Ensemble* merupakan salah satu algoritma *Machine Learning* untuk menghasilkan prediksi terbaik dengan menggunakan kombinasi beberapa algoritma untuk pengklasifikasian daripada menggunakan satu algoritma saja [25]. Metode *Ensemble* merupakan metode yang dapat meningkatkan hasil kinerja atau akurasi dengan mengkombinasikan beberapa algoritma klasifikasi daripada menggunakan singgle model klasifikasi. Metode *Ensemble* terdapat beberapa jenis salah satunya yaitu *Adaboost* dan *Bagging*. *Adaboost* merupakan metode yang dapat meningkatkan ketelitian pada kelas yang tidak seimbang atau dapat meningkatkan identifikasi dari kelas minoritas yang sulit seimbang dengan kelas mayoritas [26]. *Adaboost* merupakan metode yang dilakukan secara berulang dengan fokus utama yaitu untuk mendapatkan hasil pengklasifikasian yang lemah yang dihasilkan dari pembelajaran kumpulan data yang sama secara berulang kemudian menggabungkan beberapa hasil klasifikasi lemah yang diperoleh dari data pelatihan untuk menghasilkan pengklasifikasian yang kuat [27]. *Bootstrap Aggregation* atau *Bagging* merupakan salah satu metode *Ensemble* dengan klasifikasi terbaik, dapat digunakan untuk mengurangi tingkat kesalahan yang terdapat dalam klasifikasi, dan dapat meningkatkan kecepatan dalam klasifikasi dengan kebutuhan memori yang kecil [28].Metode *Bagging* merupakan metode *Ensemble yang* dapat membantu dalam mengurangi *overfitting* data [29].

## 3. HASIL DAN PEMBAHASAN

Penelitian dilakukan sebanyak 9 kali pengujian, diantaranya yaitu 3 kali pengujian menggunakan algoritma klasifikasi secara mandiri atau tunggal, 3 kali pengujian menggunakan algoritma klasifikasi yang dikombinasikan dengan metode *Adaboost*, dan 3 kali menggunakan algoritma klasifikasi dikombinasikan dengan metode *Bagging*.

Hasil pengujian akan menghasilkan nilai yang dapat digunakan untuk perbandingan yaitu nilai *accuracy, recall, precision* dan AUC.

Berikut table hasil pengujian:

Tabel 2. Hasil Pengujian

| Pengujian | *Accuracy* | *Recall* | *Precision* | AUC |
|---|---|---|---|---|
| *Decision Tree* | 74.46% | *38.04%* | *64.15%* | 0.685 |
| *Naïve Bayes* | *73.41%* | *54.40%* | *55.82%* | 0.725 |
| KNN | *75.57%* | *19.66%* | *91.67%* | 0.651 |
| *Decision Tree+Adaboost* | 82.76% | 45.58% | 93% | 0.720 |
| *Naïve Bayes+Adaboost* | 75.94% | 59.27% | 61.77% | 0.712 |
| KNN+*Adaboost* | 78.11% | 28.49% | 92.50% | 0.637 |
| *Decision Tree+Bagging* | 82.76% | 45.58% | 93% | 0.864 |
| *Naïve Bayes+Bagging* | 75.94% | 59.27% | 61.77% | 0.796 |
| KNN+*Bagging* | 77.39% | 26.13% | 91.67% | 0.950 |

Berdasarkan hasil yang didapatkan menunjukkan adanya peningkatan algoritma klasifikasi pada nilai *accuracy, recall, precision,* dan AUC apabila dikombinasikan menggunakan metode *Ensemble* dari pada menggunakan algoritma tunggal. Hasil paling unggul yaitu menggunakan algoritma klasifikasi *Decision Tree* dengan hasil *accuracy* 74.46% jika menggunakan algoritma tunggal, dan hasil paling unggul apabila dikombinasikan dengan *Adaboost* dan *Bagging* menggunakan algoritma klasifikasi *Decision Tree* yaitu 82.76%. Pada nilai AUC tertinggi diperoleh dari algoritma KNN yang dikombinasikan dengan metode *Bagging* yaitu sebesar 0.950 dengan kategori sangat baik.

## 4. KESIMPULAN

Model metode yang menghasilkan *accuracy* paling unggul dari hasil pengujian yaitu menggunakan algoritma klasifikasi *Decision Tree* yang dikombinasikan menggunakan metode *Ensemble* yaitu *Adaboost* dan *Bagging* dengan hasil yaitu 82.76%, berdasarkan hasil tersebut, meningkat sebesar 8.3% dari hasil algoritma *Decision Tree* yang diuji secara tunggal yaitu sebesar 74.46%. Nilai AUC dengan hasil pengujian tertinggi diperoleh dari algoritma KNN yang digabungkan dengan metode *Bagging* yaitu sebesar 0.950 dengan kategori sangat baik. Berdasarkan hasil pengujian maka metode algoritma *Decision Tree* dan metode *Ensemble* dapat meningkatkan akurasi dalam mendiagnosa penyakit kanker payudara (*breast cancer)* secara dini, sehingga dapat membantu tenaga medis untuk menangani pasien secara optimal.

Metode tersebut dapat digunakan bagi penelitian mendatang untuk meningkatkan akurasi menggunakan *dataset* lain atau teknik lain untuk dapat menghasilkan akurasi yang lebih meningkat.

## UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih banyak kepada Universitas Pelita Bangsa yang telah mendukung kami baik secara finansial sehingga penelitian ini dapat terlaksana dengan baik. Kami juga mengucapkan terima kasih kepada semua pihak yang telah membantu sehingga penelitian ini dapat terwujud.

## DAFTAR PUSTAKA

[1] N. R. Muntiari and K. H. Hanif, "Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning," *J. Ilmu Komput. dan Teknol.*, vol. 3, no. 1, pp. 1–6, 2022, doi: 10.35960/ikomti.v3i1.766.

[2] Kemenkes RI, "Hasil Riset Kesehatan Dasar Tahun 2018," *Kementrian Kesehat. RI*, vol. 53, no. 9, pp. 1689–1699, 2018.

[3] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer," *Ann. Med. Surg.*, vol. 62, no. November 2020, pp. 53–64, 2021, doi: 10.1016/j.amsu.2020.12.043.

[4] V. P. C. Magboo and M. S. Magboo, "Machine learning classifiers on breast cancer recurrences," *Procedia Comput. Sci.*, vol. 192, pp. 2742–2752, 2021, doi: 10.1016/j.procs.2021.09.044.

[5] Y. Feng *et al.*, "Predicting breast cancer-specific survival in metaplastic breast cancer patients using machine learning algorithms," *J. Pathol. Inform.*, vol. 14, no. August, p. 100329, 2023, doi: 10.1016/j.jpi.2023.100329.

[6] V. Nemade, V. Fegade, V. Nemade, and V. Fegade, "Machine Learning Techniques for Breast Cancer Prediction," *Procedia Comput. Sci.*, vol. 218, no. 2022, pp. 1314–1320, 2023, doi: 10.1016/j.procs.2023.01.110.

[7] N. Meilani and O. Nurdiawan, "Data Mining untuk Klasifikasi Penderita Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor," vol. 2, no. 1, pp. 177–187, 2023.

[8] A. M. A. Rahim, I. Y. R. Pratiwi, and M. A. Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over- Sampling Technique Dan Random Forest Clasifier," vol. 12, no. 1, pp. 2995–3011, 2023.

[9] D. Cahyanti, A. Rahmayani, and S. Ainy, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," vol. 1, no. 2, pp. 39–43, 2020.

[10] R. Y. Nugroho Agung, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," vol. 1, no. 10, pp. 504–510, 2021.

[11] R. Nurhidayat and K. E. Dewi, "PENERAPAN ALGORITMA K-NEAREST NEIGHBOR DAN FITUR EKSTRAKSI N-GRAM DALAM ANALISIS SENTIMEN BERBASIS ASPEK," vol. 12, no. 1, pp. 91–100, 2023.

[12] A. Miftahusalam, H. Pratiwi, I. Slamet, P. S. Statistika, and U. S. Maret, "Perbandingan Metode Random Forest dan Naive Bayes pada Analisis Sentimen Review Aplikasi BCA Mobile," pp. 1–8, 2023.

[13] L. Qadrini, A. Seppewali, and A. Aina, "DECISION TREE DAN ADABOOST PADA KLASIFIKASI PENERIMA PROGRAM BANTUAN SOSIAL," vol. 2, no. 7, 2021.

[14] M. Ula, A. F. Ulva, M. Mauliza, M. A. Ali, and Y. R. Said, "Application of Machine Learning in Determining the Classification of Children'S Nutrition With Decision Tree," *J. Tek. Inform.*, vol. 3, no. 5, pp. 1457–1465, 2022, doi: 10.20884/1.jutif.2022.3.5.599.

[15] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, no. November 2021, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.

[16] D. Septhya *et al.*, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 15–19, 2023.

[17] S. Bhatia and J. Malhotra, "Naïve bayes classifier for predicting the novel coronavirus," *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2021*, no. Icicv, pp. 880–883, 2021, doi: 10.1109/ICICV50876.2021.9388410.

[18] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052068.

[19] F. J. Yang, "An implementation of naive bayes classifier," *Proc. - 2018 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2018*, pp. 301–306, 2018, doi: 10.1109/CSCI46756.2018.00065.

[20] Z. Ye, P. Song, D. Zheng, X. Zhang, and J. Wu, "A Naive Bayes model on lung adenocarcinoma projection based on tumor microenvironment and weighted gene co-expression network analysis," *Infect. Dis. Model.*, vol. 7, no. 3, pp. 498–509, 2022, doi: 10.1016/j.idm.2022.07.009.

[21] B. Srinivas and G. Sasibhushana Rao, "A hybrid CNN-KNN model for MRI brain tumor classification," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 5230–5235, 2019, doi: 10.35940/ijrte.B1051.078219.

[22] S. K. P. Loka and A. Marsal, "Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Classifier untuk Klasifikasi Status Gizi Pada Balita," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 8–14, 2023, doi: 10.57152/malcom.v3i1.474.
[23] Yulianto Anggi Priliani and S. Darwis, "Penerapan Metode K-Nearest Neighbors (kNN) pada Bearing," *J. Ris. Stat.*, vol. 1, no. 1, pp. 10–18, 2021, doi: 10.29313/jrs.v1i1.16.
[24] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.
[25] E. Mardiani *et al.*, "Komparasi Metode Knn , Naive Bayes , Decision Tree , Ensemble , Linear Regression Terhadap Analisis Performa Pelajar Sma," vol. 3, pp. 13880–13892, 2023.
[26] R. I. Arumnisaa and A. W. Wijayanto, "Perbandingan Metode Ensemble Learning : Random Forest , Support Vector Machine , AdaBoost pada Klasifikasi Indeks Pembangunan Manusia ( IPM ) Comparison of Ensemble Learning Method : Random Forest , Support Vector," vol. 12, pp. 206–218, 2023.
[27] D. Hu *et al.*, "Demand response-oriented virtual power plant evaluation based on AdaBoost and BP neural network," *Energy Reports*, vol. 9, pp. 922–931, 2023, doi: 10.1016/j.egyr.2023.05.012.
[28] Y. Religia, A. Nugroho, and W. Hadikristanto, "Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," vol. 1, no. 10, pp. 187–192, 2021.
[29] B. S. Ju, S. Kwag, and S. Lee, "Performance-based drift prediction of reinforced concrete shear wall using bagging ensemble method," *Nucl. Eng. Technol.*, vol. 55, no. 8, pp. 2747–2756, 2023, doi: 10.1016/j.net.2023.05.008.