JNATIA Volume 3, Nomor 3, Mei 2025 p-ISSN: 2986-3929
Jurnal Nasional Teknologi Informasi dan Aplikasinya e-ISSN: 3032-1948

Optimasi C4.5 Berbasis PSO untuk Prediksi Kanker Payudara dengan Data BC Wisconsin

Tun Pasek Sarwiko Dipranoto, I Gede Surya Rahayuda

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
pasekdipranoto03@gmail.com
igedesuryarahayuda@unud.ac.id

Abstract

Breast cancer is a type of cancer that often arises from the development of abnormal cells in breast tissue, which then grow uncontrollably. In Indonesia, breast cancer cases are the highest compared to other types of cancer and are one of the main causes of death. This research aims to optimize the C4.5 algorithm using Particle Swarm Optimization (PSO) to predict breast cancer using the Wisconsin Breast Cancer dataset. Breast cancer remains one of the leading causes of death in women worldwide, emphasizing the importance of early detection and accurate classification. Previous research has demonstrated the effectiveness of various algorithms, including Decision Tree, Naive Bayes, and K-Nearest Neighbors, in diagnosing breast cancer, with K-Nearest Neighbors often demonstrating superior accuracy. This research evaluates the performance of the C4.5 algorithm, both before and after being optimized with PSO. Preliminary results show that the C4.5 algorithm without optimization achieves 94% accuracy. After optimization with PSO, the accuracy increased to 96%, highlighting the potential of PSO in improving prediction models for breast cancer diagnosis.

Keywords: *Breast Cancer, C4.5 Algorithm, Particle Swarm Optimization (PSO), Machine Learning, Wisconsin Breast Cancer Dataset*

1. Pendahuluan

Mengetahui gejala penyakit adalah langkah awal yang penting dalam mencegah munculnya suatu penyakit yang dapat mengancam kesehatan dan bahkan mengakibatkan kematian. Kanker payudara adalah salah satu penyakit non-kulit yang sangat serius pada wanita, disebabkan oleh berbagai faktor, mulai dari sel dan saluran hingga jaringan penopang payudara, selain kulitnya. Kanker payudara adalah salah satu penyebab utama kematian pada wanita, dan penyakit ini menempati urutan kedua setelah kanker paru-paru [1]. Pada kanker payudara, terdapat perbedaan antara jenis ganas dan jinak. Mengetahui jenisnya sangat penting karena akan memungkinkan pencegahan dan pengobatan yang sesuai dengan jenis kanker payudara yang diderita. Hal ini bertujuan untuk mencegah terjadinya efek samping yang mungkin timbul pada pasien, bahkan mengurangi risiko kematian.

Data Globocan tahun 2020 jumlah kasus kanker dunia mencapai total 19.292.789 kasus, untuk kasus kanker payudara berjumlah 2.261.419 (11.7%). Untuk kasus kematian yang diakibatkan oleh penyakit kanker total 9.958.133, sedangkan kasus kematian yang diakibatkan oleh kanker payudara berjumlah 684.996 menyumbang (6.9%) dari kasus kematian yang diakibatkan oleh penyakit kanker [2]. Data Globocan tahun 2020, jumlah kasus baru kanker payudara mencapai 65.858 kasus (16,6%) dari total 396.914 kasus baru penyakit kanker di Indonesia. Sementara itu, untuk jumlah kematiannya mencapai lebih dari 22 ribu jiwa kasus [3], [4]. Kanker payudara terus menjadi isu kesehatan global yang memerlukan perhatian serius karena tingginya tingkat kejadian dan dampaknya pada kesejahteraan wanita di seluruh dunia. *Dataset Breast Cancer Wisconsin* (BCW) telah menjadi sumber data penting dalam usaha pemodelan dan klasifikasi kanker payudara. Data ini mencakup berbagai fitur, seperti ukuran inti sel, keberaturan sel, dan parameter lain yang memiliki peran penting dalam proses diagnosis kanker payudara.Pentingnya

p-ISSN: 2986-3929
e-ISSN: 3032-1948

deteksi dini kanker payudara tak terbantahkan dalam meningkatkan prognosis pasien serta meningkatkan tingkat kelangsungan hidup. Hal ini bertujuan untuk mengurangi kematian akibat kanker global sebesar 2,5% per tahun, sehingga mencegah 2,5 juta kematian akibat kanker payudara di seluruh dunia antara tahun 2020 dan 2040 [5].

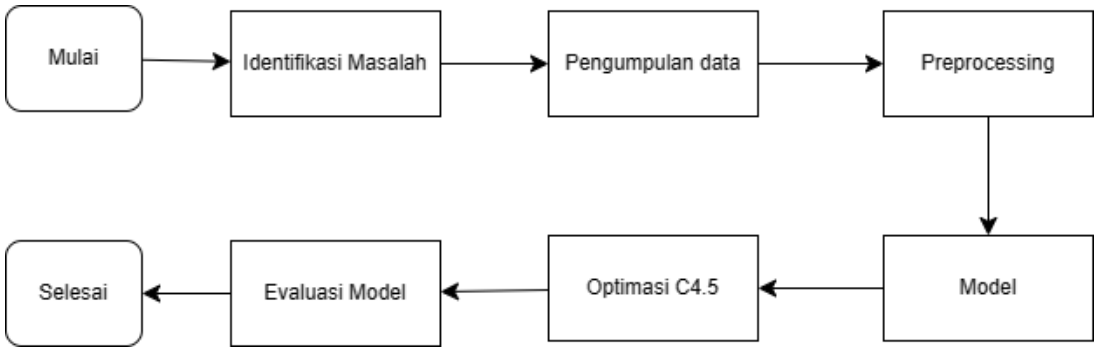
Penelitian terdahulu dalam mendeteksi dan mengklasifikasikan kanker payudara mencakup perbandingan algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors. Penelitian ini menemukan bahwa K-Nearest Neighbors memiliki akurasi yang sangat baik dibandingkan dengan Naive Bayes dan Decision Tree, dengan tingkat akurasi mencapai 98% menggunakan metode Hold-Out dan 96% menggunakan metode K-Fold [7]. Kemudian penelitian berjudul "Data Mining Techniques in Predicting Breast Cancer" juga menunjukkan hasil yang positif: K-Nearest Neighbors (KNN) mencapai tingkat akurasi sebesar 92,31%, sedangkan Support Vector Machine (SVM) mencapai akurasi 95,65% [8]. Penelitian berjudul "Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer" menunjukkan bahwa algoritma Decision Tree mencapai akurasi sebesar 91,23%, sementara K-Nearest Neighbor (KNN) mencapai akurasi 95,61% [9]. dan pada penelitian Deteksi Dini Kanker Payudara Menggunakan Algoritma *K-Nearest Neighbor* (KNN) Dan *Decision Tree C-45* perbandingan adalah algoritma *Decision Tree*, dan *K-Nearest Neighbors* mendapatkan hasil pengujian yang menunjukkan bahwa algoritma *K-Nearest Neighbors* menghasilkan performa akurasi 94,73 sedangkan algoritma *Decision Tree* 96,49% [10].

2. Metode Penelitian

Pada tahap ini, penulis akan menjelaskan secara rinci mengenai langkah-langkah dan urutan kegiatan yang akan dilakukan dalam penelitian. Ini mencakup gambaran alur lengkap mulai dari perencanaan, pengumpulan data, analisis, hingga pembuatan kesimpulan. Setiap tahap akan dibahas secara terperinci untuk memastikan penelitian berjalan sesuai dengan rencana yang telah disusun. Penelitian ini dimulai dengan memperkenalkan skema penelitian dan alat yang akan digunakan, yakni alat Machine Learning open source. Selain itu, dijelaskan juga tentang *dataset* yang akan digunakan dalam penelitian ini. Tujuan dari penelitian ini adalah untuk optimasi performa C4.5 dalam mendiagnosa kanker payudara menggunakan data *Breast Cancer Wisconsin*.

2.1 Skema Alur Penelitian

Pada bagian ini, akan diuraikan langkah-langkah dari proses penelitian yang akan dijalankan. Dalam upaya menganalisis dan mengidentifikasi pola data guna membentuk *dataset* yang memfasilitasi penelitian serta memastikan kelancaran dan pencapaian tujuan, sebuah alur tahapan penelitian akan dibentuk sebagai berikut:



Gambar 1. Skema Alur Penelitian

2.2 Identifikasi Masalah

Berdasarkan dengan latar belakang yang ada, masalah yang diidentifikasi adalah penggunaan algoritma C4.5 berbasis PSO agar algoritma ini dapat memahami model identifikasi dengan akurasi tertinggi dalam mendiagnosa kanker payudara.

2.3 Pengumpulan Data

Data diambil dari *dataset* pada laman uci yang dapat diakses menggunakan link berikut:

https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
https://archive.ics.uci.edu/static/public/17/data.csv

Dataset ini merupakan data penderita kanker payudara yang berjumlah 570 data dan juga 32 atribut
.

2.4 Preprocessing

Data dipersiapkan dan diolah menggunakan platform Rapidminer dengan proses normalisasi, eliminasi fitur yang tidak relevan, serta penanganan data yang hilang. Setelah proses pengolahan selesai, data yang telah dipersiapkan siap digunakan untuk melatih dan menguji model.

2.5 Model

Pada proses pemodelan dimulai dengan *dataset*, di mana algoritma klasifikasi akan diterapkan untuk membuat model klasifikasi. Selanjutnya, parameter evaluasi akan dihasilkan. Model yang dieksplorasi dalam penelitian ini adalah Algoritma C4.5 dengan optimasi PSO *(Particle Swarm Optimization)*.

a. Algoritma C4.5

Algoritma C4.5 adalah salah satu teknik yang umum digunakan dalam data mining untuk melakukan klasifikasi atau pengelompokan berdasarkan aturan-aturan yang dihasilkan dari struktur pohon keputusan. Persamaan 1 merupakan perhitungan entropy dan persamaan 2 adalah perhitungan Gain. Sebelum melakukan perhitungan Gain untuk atribut, terlebih dahulu dihitung nilai entropy, dengan persamaan

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

Keterangan:

S : Himpunan Kasus
A : Atribut
n : Jumlah partisi himpunan atribut A
pi : proporsi s_i terhadap s

Selanjutnya nilai gain tertinggi dari atribut digunakan sebagai penentu atribut yang akan dijadikan akar. Cara menghitung gain digunakan persamaan berikut:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \quad \frac{|s_i|}{|s|} \quad * Entropy(s_i) \tag{2}$$

Keterangan:

S : Himpunan Kasus
A : Atribut
n : Jumlah partisi himpunan atribut A
$|s_i|$: Jumlah kasus pada partisi ke-i

$|s|$: Jumlah kasus dalam S

b. PSO *(Particle Swarm Optimization)*

Particle Swarm Optimization (PSO) merupakan sebuah teknik optimisasi yang diperkenalkan oleh Dr. James Kennedy dan Dr. Russell C Eberhart pada tahun 1995. Metode ini terinspirasi oleh perilaku kolektif burung dan lebah. Kelebihan dari pendekatan ini adalah kesederhanaan konsepnya, kemudahan dalam implementasinya, dan efisiensi perhitungan yang lebih tinggi dibandingkan dengan metode matematis dan teknik heuristik lainnya.[11]. Adapun persamaannya:

$$V_{i,m} = W.V_{i,w} + C_1 * R * (pbest_{i,m} - X_{i,m}) + C_2 * R * (gbest_m - X_{i,m}) \quad (3)$$

Perhitungan kecepatan baru untuk setiap partikel, yang merupakan solusi potensial, didasarkan pada kecepatan sebelumnya (Vi,m), lokasi dimana nilai keunggulan pribadi terbaik (pbest) telah dicapai, dan lokasi populasi global (gbest untuk versi global, lbest untuk versi lokal) atau lingkungan lokal pada versi lokal dari algoritma dimana nilai keunggulan terbaik telah dicapai.

$$X_{id} = X_{i,w} + V_{i,m} \quad (4)$$

Posisi setiap partikel dalam ruang solusi diperbaharui. Dua bilangan acak, c1 dan c2, dihasilkan secara independen. Penggunaan berat inersia w telah terbukti meningkatkan kinerja dalam berbagai aplikasi. Hasil perhitungan partikel adalah kecepatan partikel yang berada dalam interval [0,1].

Keterangan:

n : jumlah partikel dalam kelompok
d : dimensi
vi,m : kecepatan partikel ke-i pada iterasi ke-i
w : faktor bobot inersia
c1, c2 : konstanta akselerasi (learning rate)
R : bilangan random (0-1)
xi,d : posisi saat ini dari partikel ke-i pada iterasi ke-i
pbest : posisi terbaik sebelumnya dari partikel ke-i
gbest : partikel terbaik diantara semua partikel dalam satu kelompok atau populasi

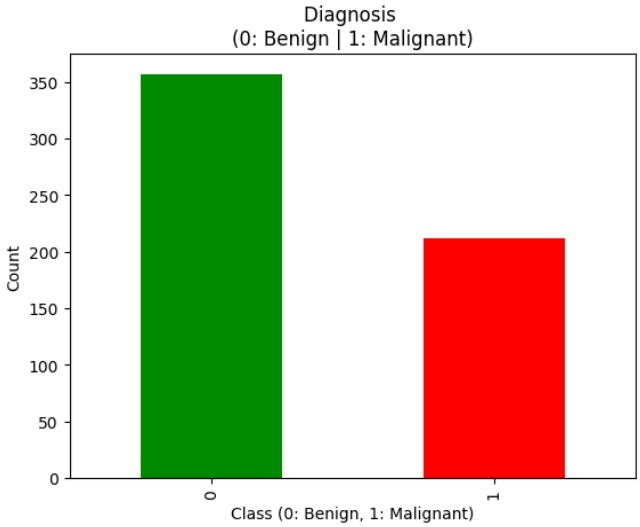
3. Hasil dan Diskusi

Penelitian ini bertujuan untuk mendapatkan hasil prediksi dan akurasi terbaik dengan menganalisis penerapan model algoritma C4.5 dengan model algoritma C4.5 berbasis bagging yang dioptimalkan menggunakan *Particle Swarm Optimization* (PSO) dalam memprediksi penyakit kanker payudara.

3.1. *Dataset*

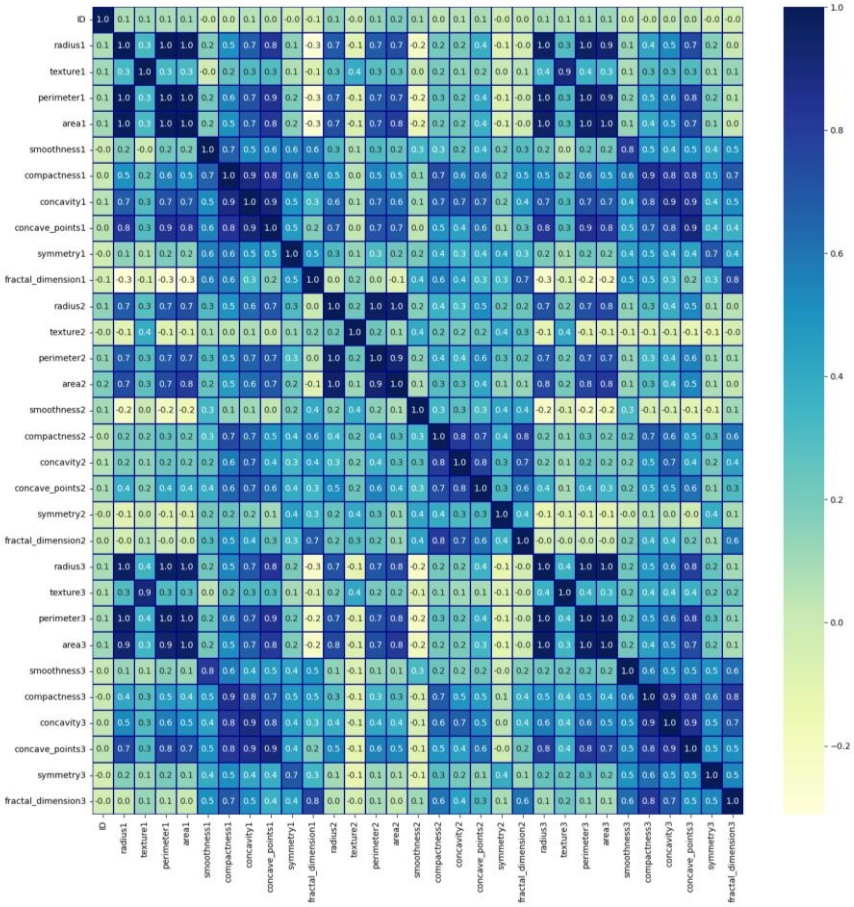
Pada tahap awal dari penelitian dan eksperimen diawali dengan melakukan analisis terhadap *dataset* yang telah divalidasi melalui proses preprocessing menggunakan sebuah algoritma yang dijalankan pada *jupyter notebook*. Maka didapatkan gambaran umum dari *dataset* yang berupa grafik visual,dapat dilihat di Gambar 2:

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 2. Grafik Diagnosis dari *Dataset*

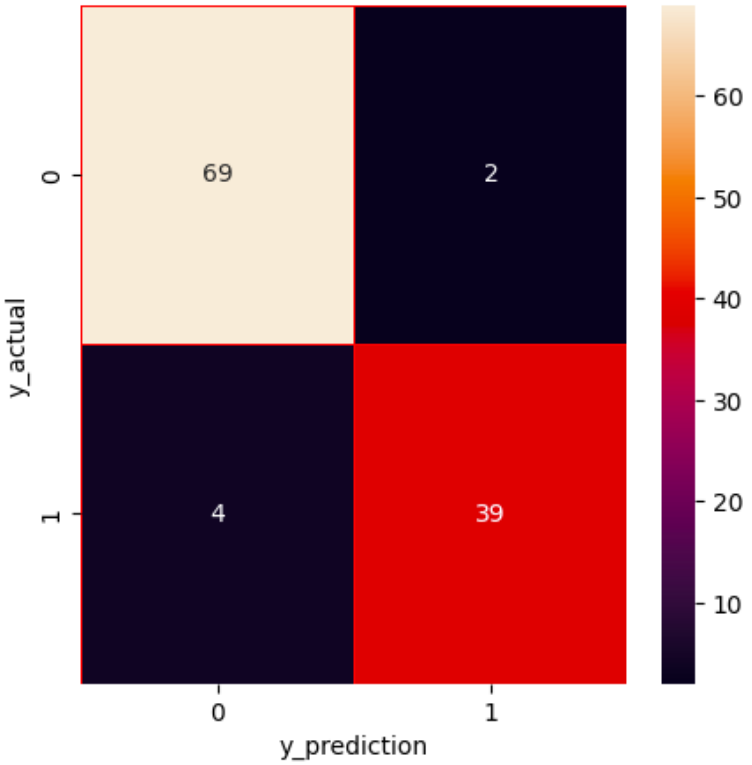
Setelah *dataset* ditampilkan dalam bentuk grafik batang *dataset* juga dibuat dalam bentuk *heatmap* matriks korelasi untuk menggambarkan nilai koefisien korelasi yang berbeda dari hubungan 31 variabel pada *dataset* dapat dilihat pada Gambar 3



Gambar 3. Korelasi Map

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Dalam korelasi map di atas memiliki nilai korelasi setiap pasang variabel yang diwakili oleh kotak berwarna pada grafik. nilai korelasi berkisaran dari -1 hingga 1, di mana 1 adalah korelasi positif sempurna, -1 adalah korelasi negatif sempurna, dan 0 menunjukkan tidak ada korelasi. Setelah data di buatkan korelasi maka *dataset* juga dibuatkan dalam bentuk *heatmap confusion matrix* seperti Gambar 4.



Gambar 4. *Heatmap Confusion Matrix*

Pada *Heatmap Confusion Matrix* dapat dilihat pada sumbu y (y_actual), terdapat label aktual (Benign dan Malignant) dan sama seperti sumbu x dalam *heatmap* terdapat empat kotak yang mewakili jumlah true *positives*, *false negatives, false positives*, dan *true negatives*. Kotak kiri atas berwarna beige dengan nilai 69, menunjukkan prediksi true *positive* dimana baik nilai aktual maupun prediksi adalah 0. Kotak kanan atas berwarna biru tua dengan nilai 2, mewakili prediksi *false negative* dimana nilai aktual adalah 0 tetapi diprediksi sebagai 1. Kotak kiri bawah berwarna merah dengan nilai 4, menunjukkan prediksi *false positive* dimana nilai aktual adalah 1 tetapi diprediksi sebagai 0. Kotak kanan bawah berwarna merah tua dengan nilai 39 untuk prediksi *true negative* dimana baik nilai aktual maupun prediksi adalah 1.

3.2. Algoritma C4.5

Pada penggunaan algoritma C4.5 yang belum dioptimasi dengan asumsi parameter adalah 100 dan menghasilkan akurasi di angka 94%, seperti pada Gambar 5.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

```
Confusion Matrix:
[[69  2]
 [ 5 38]]
Max Depth yang digunakan: 100
Akurasi sebelum optimasi: 0.9385964912280702
              precision    recall  f1-score   support

           B       0.93      0.97      0.95        71
           M       0.95      0.88      0.92        43

    accuracy                           0.94       114
   macro avg       0.94      0.93      0.93       114
weighted avg       0.94      0.94      0.94       114
```

Gambar 5. Akurasi Algoritma C4.5 Sebelum Optimasi

3.3. Algoritma C4.5 Optimasi

Pada pengujian ini algoritma C4.5 yang dioptimasi dengan *Particle Swarm Optimization* (PSO) yang menghasilkan nilai akurasi 96% dengan parameter 100, seperti pada Gambar 6

```
Confusion Matrix setelah optimasi:
[[69  2]
 [ 5 38]]
Stopping search: maximum iterations reached --> 100
Akurasi setelah optimasi: 0.956140350877193
              precision    recall  f1-score   support

           B       0.95      0.99      0.97        71
           M       0.97      0.91      0.94        43

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Gambar 6. akurasi Algoritma C4.5 Setelah Dioptimasi dengan PSO

4. Kesimpulan

Berdasarkan uraian dari pembahasan yang dilakukan sebelumnya,dapat disimpulkan bahwa prediksi kanker payudara menggunakan algoritma C4.5 yaitu sebesar 94%, sedangkan setelah algoritma C4.5 yang dioptimasi menggunakan *Particle Swarm Optimization* (PSO) menjadi 96%. Proses optimasi ini tidak terlalu berdampak pada hasil akurasi pada algoritma C4.5 dengan selisih 0,02%, namun dapat dilihat akurasi yang dihasilkan setelah dioptimasi terdapat peningkatan.

Daftar Pustaka

[1] H. Oktavianto and R. P. Handri, “Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes,” *INFORMAL Informatics* J., vol. 4, no. 3, p. 117, 2020, doi: 10.19184/isj. v4i3.14170.
[2] The Global Cancer Observatory, “Kasus Kanker Payudara Dunia”, Doi: 10.8.
[3] Kemenkes Ri, “Kanker Payudara Paling Banyak Di Indonesia,” 2024 https://kemkes.go.id/id/kanker-payudaya-paling-banyak-di-indonesia-kemenkes-

targetkan-pemerataan-layanan-kesehatan.
[4] Globocan, “Kasus Kanker Payudara Indonesia.” Accessed: Mei. 7, 2024. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/populations/360-indonesia-fact-sheet.pdf
[5] Who, “Breast Cancer,” 2024. [Online]. Available: https://www.who.int/News-Room/Fact-Sheets/Detail/Breast-Cancer
[6] Fahrurrozi, W. (2023). Deteksi Dini Kanker Payudara Menggunakan Algoritma K-Nearest Neighbour (KNN) dan Decision Tree C-45. *Zenodo (CERN European Organization for Nuclear Research)*. doi: https://doi.org/10.5281/zenodo.8412264.
[7] M. A. Jabbar, E. Hasmin, Sunardi, And W. Musu, *“*Komparasi Algoritma Decision Tree, Naive Bayes, Dan K- Nearest Neighbors Dalam KlasifikasiKanker Payudara*,*” *Csrid Journal*, Vol. 14, No. 3, Pp. 258–270, 2022.
[8] F. K. Nasser and S. F. Behadili, “Breast Cancer Detection Using Decision Tree and KNearest Neighbour Classifiers,” *Iraqi Journal of Science*, Vol.63, No. 11, Pp. 4987–5003, 2022, doi: https://doi.org/10.24996/Ijs.2022.63.11.34.
[9] H. Rajaguru and S. R. Sannasi Chakravarthy, “Analysis Of Decision Tree AndK-Nearest Neighbor Algorithm In the Classification Of Breast Cancer,”*Asian Pacific Journal Of Cancer Prevention*, Vol. 20, No. 12, Pp. 3777–3781, 2019, doi: https://doi.org/10.31557/Apjcp.2019.20.12.3777.
[10] Fahrurrozi, W. (2023). Deteksi Dini Kanker Payudara Menggunakan Algoritma K-Nearest Neighbour (KNN) dan Decision Tree C-45. Zenodo (*CERN European Organization for Nuclear Research)*. doi: https://doi.org/10.5281/zenodo.8412264.
[11] Husniah, H.F. and Arifin, T. (2021). Implementasi Algoritma Naïve Bayes Berbasis Particle Swarm Optimization Untuk Memprediksi Penyakit Hepatitis. *Jurnal Ilmu Komputer, 14(1)*, p.36. doi: https://doi.org/10.24843/jik.2021.v14.i01.p05 .

.