JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 64-69
http://jurnal.unidha.ac.id/index.php/jiska

E-ISSN : 2985-9735
P-ISSN :

Penerapan Random Forest untuk Klasifikasi Diagnosis Kanker Payudara Berbasis Dataset WBCD

Naufal Aqiilah Asra[a], Maulana Al Nouri[b], Tia Risky Yasmin Saketang[c], Repi Meilani Putri[d]
[a]Ilmu Komputer, FMIPA, Universitas Negeri Medan, naufalaqiilahasra2025@gmail.com
[b]Ilmu Komputer, FMIPA, Universitas Negeri Medan, maulanaalnouri51@gmail.com
[c]Ilmu Komputer, FMIPA, Universitas Negeri Medan, tyasmin488@gmail.com
[d]Ilmu Komputer, FMIPA, Universitas Negeri Medan, prepi9861@gmail.com

Abstract
Breast cancer is one of the most critical global health challenges, with Indonesia recording 66,271 new cases in 2022 according to GLOBOCAN data published by the International Agency for Research on Cancer (IARC/WHO). Early and accurate detection is essential to improving patient survival rates, yet conventional diagnosis remains time-consuming and dependent on expert availability. This study implements the Random Forest algorithm to classify breast cancer diagnosis using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset from the UCI Machine Learning Repository. The dataset consists of 569 samples with 30 numerical features extracted from fine-needle aspirate (FNA) cell images, labeled as benign or malignant. Data preprocessing involved removing non-predictive columns, converting categorical labels to binary format, handling outliers using IQR Clipping, and applying StandardScaler normalization. The dataset was split into 80% training and 20% testing using stratified splitting, with the Random Forest Classifier configured using 100 decision trees and class_weight=balanced to handle class imbalance. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics alongside confusion matrix analysis and 5-Fold Stratified Cross Validation. The model achieved 97.37% accuracy on the test set, with zero False Positive predictions, meaning no benign patient was misdiagnosed as malignant. Cross-validation confirmed generalization ability with a mean accuracy of 96.31%, indicating no overfitting. Feature importance analysis identified area_worst, concave points_worst, and perimeter_worst as the most dominant features, consistent with the clinical morphological characteristics of malignant cancer cells. These findings demonstrate the strong potential of Random Forest as a reliable and interpretable tool for supporting breast cancer diagnosis.
Keywords: *breast cancer, classification, machine learning, Random Forest, Wisconsin Breast Cancer Diagnostic*

Abstrak
Kanker payudara merupakan salah satu tantangan kesehatan global yang paling kritis, dengan Indonesia mencatat 66.271 kasus baru pada tahun 2022 berdasarkan data GLOBOCAN yang diterbitkan oleh International Agency for Research on Cancer (IARC/WHO). Deteksi dini yang akurat merupakan faktor penting dalam meningkatkan angka kelangsungan hidup pasien, namun diagnosis konvensional masih bergantung pada ketersediaan tenaga ahli dan bersifat memakan waktu. Penelitian ini mengimplementasikan algoritma Random Forest untuk mengklasifikasikan diagnosis kanker payudara menggunakan dataset Wisconsin Breast Cancer Diagnostic (WBCD) dari UCI Machine Learning Repository. Dataset terdiri dari 569 sampel dengan 30 fitur numerik yang diekstraksi dari citra aspirasi jarum halus (FNA), berlabel jinak (benign) atau ganas (malignant). Praproses data meliputi penghapusan kolom tidak prediktif, konversi label kategorikal ke format biner, penanganan outlier menggunakan IQR Clipping, serta standardisasi menggunakan StandardScaler. Dataset dibagi menjadi 80% data latih dan 20% data uji menggunakan stratified splitting, dengan konfigurasi Random Forest Classifier menggunakan 100 pohon keputusan dan class_weight=balanced untuk menangani ketidakseimbangan kelas. Evaluasi model menggunakan metrik akurasi, presisi, recall, dan F1-score, analisis confusion matrix, serta 5-Fold Stratified Cross Validation. Model mencapai akurasi 97,37% pada data uji tanpa satu pun prediksi False Positive, artinya tidak ada pasien jinak yang salah didiagnosis sebagai ganas. Validasi silang mengkonfirmasi kemampuan generalisasi model dengan rata-rata akurasi 96,31%, menunjukkan tidak terjadi overfitting. Analisis feature importance mengidentifikasi area_worst, concave points_worst, dan perimeter_worst sebagai fitur paling dominan, konsisten dengan karakteristik morfologi klinis sel kanker ganas.
Kata Kunci: dataset kanker payudara, klasifikasi, machine learning, Random Forest, Wisconsin Breast Cancer Diagnostic

PENDAHULUAN

Kanker payudara merupakan salah satu permasalahan kesehatan global yang paling signifikan. Berdasarkan data GLOBOCAN 2022 yang diterbitkan oleh International Agency for Research on Cancer (IARC/WHO), kanker payudara tercatat sebagai jenis kanker dengan insidensi tertinggi di Indonesia, yakni mencapai 66.271 kasus baru atau sekitar 16,2% dari seluruh kasus kanker yang terdiagnosis, dengan angka kematian 22.598 jiwa. Secara global, American Cancer Society memproyeksikan 297.790 kasus baru kanker payudara invasif pada perempuan di Amerika Serikat untuk tahun 2024. Tren serupa juga terjadi di negara berkembang, termasuk Indonesia, di mana keterlambatan diagnosis sering kali menjadi faktor utama meningkatnya angka mortalitas.

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 64-69
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :

Studi longitudinal menggunakan data Population-Based Cancer Registry (PBCR) Yogyakarta mencatat bahwa insidensi kanker payudara di Indonesia menunjukkan variasi spasial dan temporal yang signifikan antara tahun 2008 hingga 2019. Selain itu, analisis data BPJS Kesehatan periode 2017-2020 menemukan bahwa insidensi, mortalitas, dan disability-adjusted life years (DALYs) meningkat seiring bertambahnya usia, dengan puncak pada kelompok usia 55-59 tahun [1].

Deteksi dini dan diagnosis yang akurat merupakan faktor krusial dalam meningkatkan angka kelangsungan hidup penderita kanker payudara. Diagnosis konvensional yang mengandalkan pemeriksaan manual bersifat memakan waktu, rentan terhadap variabilitas antar-pengamat, serta bergantung sepenuhnya pada ketersediaan tenaga ahli yang terdistribusi tidak merata, terutama di wilayah terpencil seperti Indonesia [2].

Dalam konteks inilah, perkembangan pesat kecerdasan buatan (artificial intelligence/AI), khususnya machine learning (ML) dan deep learning (DL), menawarkan solusi yang menjanjikan. Sistem Computer-Aided Diagnosis (CAD) berbasis ML telah menunjukkan kemampuan luar biasa dalam mengidentifikasi pola-pola halus pada data klinis yang bahkan tidak dapat ditangkap oleh mata manusia [3].

Wisconsin Diagnostic Breast Cancer (WDBC) dataset dari UCI Machine Learning Repository menjadi tolok ukur (benchmark) yang paling banyak digunakan dalam penelitian klasifikasi kanker payudara. Dataset ini memuat 569 sampel dengan 30 fitur numerik yang diekstraksi dari citra fine-needle aspirate (FNA), diklasifikasikan menjadi dua label: jinak (benign) dan ganas (malignant). Berbagai algoritma ML mulai dari Support Vector Machine (SVM), Random Forest (RF), hingga Neural Network telah diuji pada dataset ini dengan akurasi kompetitif berkisar 96% hingga 99% [4].

Kanker payudara didefinisikan sebagai pertumbuhan sel abnormal yang tidak terkendali pada jaringan payudara. Berdasarkan karakteristik histopatologis, kanker payudara diklasifikasikan menjadi dua kategori utama: karsinoma duktal invasif (IDC) yang merupakan jenis paling umum (sekitar 70-80%), dan karsinoma lobular invasif (ILC). Histopatologi tetap menjadi standar emas dalam diagnosis, namun interpretasi manualnya bersifat memakan waktu dan bergantung pada keahlian patologis [5].

Machine learning (ML) adalah cabang kecerdasan buatan yang memungkinkan sistem komputer untuk belajar dan membuat prediksi dari data tanpa diprogram secara eksplisit. Dalam konteks medis, ML telah merevolusi cara deteksi dan diagnosis penyakit, terutama untuk kondisi kompleks seperti kanker [6]. Penelitian dari Chandra et al. (2025)[7], yang dipublikasikan di Jurnal Sistem Informasi Dan Informatika (JISKA) menunjukkan bahwa Random Forest unggul dibanding Decision Tree dalam tugas klasifikasi berbasis ML, dengan akurasi mencapai 92%. Sementara itu, Desriansyah et al. (2025)[8], dalam jurnal yang sama melaporkan bahwa SVM dan MLP memberikan performa terbaik dalam klasifikasi teks, sementara Random Forest menunjukkan performa kuat dan stabil.

Support Vector Machine (SVM) bekerja dengan mencari hyperplane optimal yang memaksimalkan margin antara dua kelas data menggunakan fungsi kernel seperti RBF, polinomial, dan linear. Penelitian Elsadig et al. (2023) menunjukkan bahwa SVM dengan kernel RBF mencapai akurasi hingga 99,51% pada Wisconsin Breast Cancer Dataset. Random Forest (RF) adalah metode ensemble learning yang membangun sejumlah besar decision tree secara paralel dengan mekanisme majority voting. RF juga dikenal karena kemampuannya menghasilkan feature importance yang mendukung prinsip Explainable AI [9].

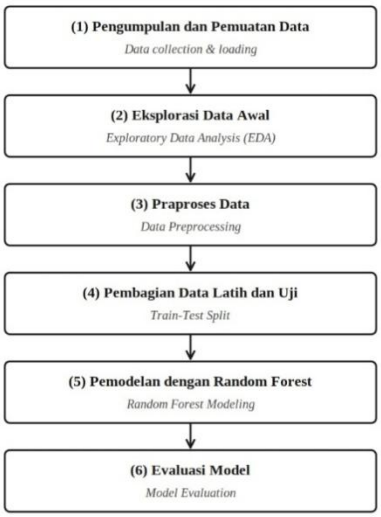
Meskipun berbagai pendekatan ML telah terbukti efektif, terdapat beberapa celah penelitian yang masih perlu dijawab: (1) perbandingan komprehensif performa algoritma ML dengan metrik evaluasi standar; (2) analisis feature importance untuk mendukung interpretabilitas klinis; serta (3) integrasi pendekatan Explainable AI (XAI) agar keputusan model dapat dipahami oleh praktisi medis [10]. Hal ini mendorong perlunya penelitian yang tidak hanya berfokus pada akurasi, tetapi juga pada keterbacaan dan kepercayaan sistem AI di lingkungan klinis.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis machine learning dengan algoritma Random Forest untuk melakukan klasifikasi diagnosis kanker payudara. Data bersumber dari dataset Wisconsin Breast Cancer Diagnostic (WBCD) dalam format Excel. Dataset ini terdiri dari 569 sampel dengan 30 fitur numerik yang diekstraksi dari citra digital aspirasi jarum halus (FNA), mencakup karakteristik radius, tekstur, perimeter, luas, kehalusan, kekompakan, cekungan, titik cekungan, simetri, dan dimensi fraktal masing-masing dalam representasi rata-rata, standar kesalahan, dan nilai terburuk.

Alur penelitian ini dilakukan melalui tahapan: (1) pengumpulan dan pemuatan data, (2) eksplorasi data awal, (3) praproses data, (4) pembagian data latih dan uji, (5) pemodelan dengan Random Forest, serta (6) evaluasi model.

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 64-69
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :



A. Eksplorasi dan Praproses Data

Tahap eksplorasi data dilakukan untuk memahami struktur dataset, termasuk dimensi data, nilai yang hilang (missing values), dan distribusi kelas target. Kolom id dihapus karena tidak memiliki nilai prediktif. Variabel target (diagnosis) dikonversi dari format kategorikal (M = Malignant/Ganas, B = Benign/Jinak) menjadi format numerik biner (M = 1, B = 0). Penanganan outlier dilakukan menggunakan metode IQR Clipping dengan membatasi nilai ekstrem pada batas bawah (Q1 - 1.5 x IQR) dan batas atas (Q3 + 1.5 x IQR), sehingga seluruh 569 sampel tetap dipertahankan. Standardisasi fitur menggunakan StandardScaler juga diterapkan untuk menjaga konsistensi alur praproses, meskipun Random Forest secara teknis tidak sensitif terhadap skala fitur.

B. Pembagian Data dan Pemodelan

Dataset dibagi menjadi data latih (80%) dan data uji (20%) menggunakan stratified splitting untuk memastikan proporsi kelas yang seimbang. Model yang digunakan adalah Random Forest Classifier dari pustaka scikit-learn dengan konfigurasi: jumlah pohon (n_estimators) = 100, class_weight='balanced' untuk menangani ketidakseimbangan kelas, dan random_state=42 untuk reprodusibilitas.

C. Evaluasi Model

Evaluasi model menggunakan metrik akurasi, presisi, recall, dan F1-score dalam classification report. Analisis kesalahan dilakukan melalui confusion matrix yang menguraikan nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Validasi model menggunakan 5-Fold Stratified Cross Validation untuk mengukur generalisasi model. Analisis feature importance juga dilakukan untuk mengidentifikasi variabel paling berkontribusi. Seluruh implementasi menggunakan Python dengan pustaka pandas, numpy, scikit-learn, matplotlib, dan seaborn.

HASIL DAN PEMBAHASAN

A. Distribusi Data dan Karakteristik Dataset

Dataset Wisconsin Breast Cancer Diagnostic (WBCD) terdiri dari 569 sampel dengan 30 fitur numerik dan satu variabel target. Tidak ditemukan missing value pada seluruh kolom. Distribusi kelas target menunjukkan 357 sampel jinak (Benign/B) dan 212 sampel ganas (Malignant/M), dengan rasio sekitar 63:37. Ketidakseimbangan kelas yang moderat ini diatasi menggunakan class_weight='balanced'. Setelah pemisahan 80:20, diperoleh 455 data latih dan 114 data uji.

B. Hasil Evaluasi Model

Setelah model dilatih dan diuji pada 114 data uji. Akurasi yang diperoleh sebesar 97,37%, artinya dari 114 prediksi hanya 3 yang meleset. Gambar 1 menampilkan rinciannya.

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 64-69
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :

Gambar 1. Classification Report Model Random Forest

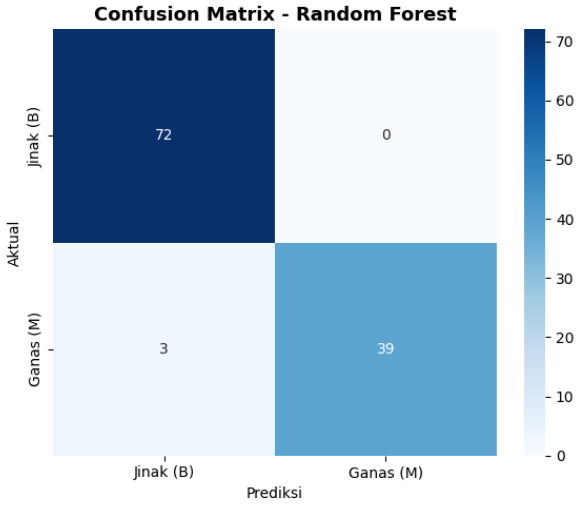
```
Classification Report:
              precision    recall  f1-score   support

   Jinak (B)       0.96      1.00      0.98        72
   Ganas (M)       1.00      0.93      0.96        42

    accuracy                           0.97       114
   macro avg       0.98      0.96      0.97       114
weighted avg       0.97      0.97      0.97       114
```

Dari Gambar 1 terlihat dua hal yang menarik. Pertama, recall kelas Jinak (B) mencapai 1,00 semua 72 sampel jinak terdeteksi tanpa ada yang keliru. Kedua, presisi kelas Ganas (M) juga sempurna di angka 1,00, yang berarti setiap kali model menyatakan "ini ganas", prediksi itu selalu benar. Hanya saja recall kelas Ganas sebesar 0,93 menunjukkan 3 dari 42 kasus ganas luput terdeteksi ini yang perlu dicermati.

C. Analisis Confusion Matrix

Gambar 2 memperlihatkan secara langsung di mana saja model yang benar dan di mana model yang salah.

Gambar 2. Confusion Matrix Model Random Forest



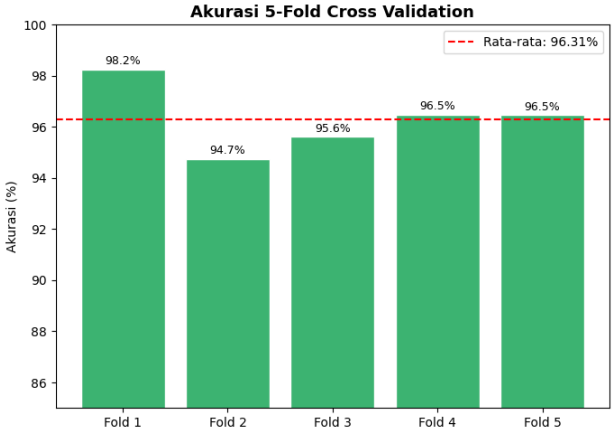
Hasilnya menunjukkan 72 sampel jinak benar diprediksi jinak (TN), dan 39 sampel ganas benar diprediksi ganas (TP). Yang menarik adalah FP = 0 tidak ada satu pun pasien jinak yang salah dicap ganas. Ini penting karena diagnosis palsu positif bisa memicu prosedur medis yang tidak perlu. Di sisi lain, ada 3 kasus False Negative yaitu pasien yang sebenarnya menderita kanker ganas tapi model tidak dapat mendeteksinya. Secara persentase memang kecil (7,14%), tetapi dalam praktik klinis 3 kasus yang terlewat itu bukanlah hal yang sepele.

D. Hasil Cross Validation

Untuk memastikan model tidak hanya "hafal" data latih, dilakukan uji validasi silang. Hasilnya ada pada Gambar 3.

JISKA: Jurnal Sistem Informasi Dan Informatika
Vol 4 No 1 (2026): Januari 2026 Hal. 64-69
http://jurnal.unidha.ac.id/index.php/jiska
E-ISSN : 2985-9735
P-ISSN :

Gambar 3. Hasil 5-Fold Stratified Cross Validation

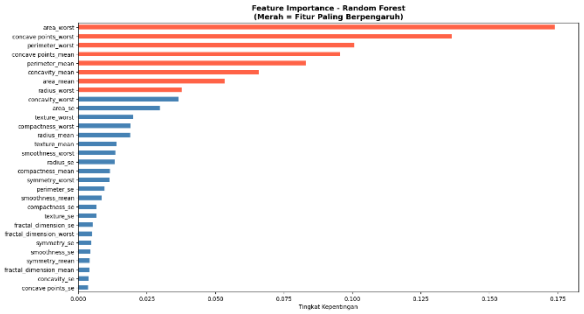


Lima fold menghasilkan akurasi 98,2%, 94,7%, 95,6%, 96,5%, dan 96,5%. Fold 2 yang paling rendah (94,7%) dan Fold 1 yang paling tinggi (98,2%), tapi selisihnya tidak terlalu jauh. Rata-rata 96,31% ini hanya berbeda sekitar 1% dari akurasi data uji. Artinya performa model tidak turun drastis ketika dihadapkan pada data yang berbeda-beda sebagai tanda yang cukup baik bahwa model tidak sekadar menghafal.

E. Analisis Feature Importance

Salah satu kelebihan Random Forest adalah bisa menunjukkan fitur mana yang paling mempengaruhi keputusannya. Gambar 4 menampilkan urutan fitur berdasarkan tingkat kepentingannya.

Gambar 4. Feature Importance



area_worst menjadi fitur paling berpengaruh dengan skor sekitar 0,175. Posisi kedua dan ketiga ditempati concave points_worst (~0,135) dan perimeter_worst (~0,100). Kalau diperhatikan, delapan fitur teratas semu

anya berkaitan dengan ukuran sel dan tingkat kecekungkannya bukan tekstur, bukan simetri. Ini tidak mengejutkan. Sel kanker ganas memang cenderung lebih besar dan bentuknya lebih tidak beraturan, jadi wajar kalau fitur-fitur ukuran dan bentuk itulah yang paling "diandalkan" model.

F. Pembahasan

Dari keseluruhan hasil, ada dua hal yang paling menonjol. Pertama, FP = 0. Tidak ada pasien jinak yang salah divonis ganas, ini penting karena kesalahan semacam itu bisa memicu tindakan medis yang tidak perlu dan membebani pasien secara psikologis maupun finansial. Kedua, selisih antara akurasi data uji (97,37%) dan rata-rata cross validation (96,31%) hanya sekitar 1%. Kalau selisihnya besar, itu tanda model cuma bagus di data latih saja. Tapi dengan selisih sekecil ini, model tampaknya cukup konsisten.

Dari sisi praproses, IQR Clipping dipilih supaya data tidak terbuang. Hasilnya semua 569 sampel tetap bisa dipakai. Parameter class_weight=balanced juga terbukti membantu: meski kelas jinak lebih banyak (63%), model tidak jadi malas mengenali kelas ganas. Terbukti dari recall jinak 1,00 dan presisi ganas 1,00 yang sama-sama tinggi. Untuk feature importance, temuan bahwa area_worst dan concave points_worst mendominasi sebenarnya cukup masuk akal secara biologis, karena sel ganas itu memang lebih besar dan lebih tidak beraturan bentuknya. Jadi model seperti "belajar" hal yang sama dengan yang dokter perhatikan.

SIMPULAN

Penelitian ini menerapkan Random Forest pada dataset WBCD untuk klasifikasi kanker payudara dan mendapat akurasi 97,37% pada data uji. Dari 114 data yang diuji, hanya 3 yang salah prediksi, semuanya False Negative, tidak ada False Positive sama sekali. Hasil cross validation dengan rata-rata 96,31% menunjukkan

E-ISSN : 2985-9735
P-ISSN :

model cukup stabil saat dihadapkan pada data berbeda. Fitur paling berpengaruh adalah area_worst, concave points_worst, dan perimeter_worst, ketiganya mencerminkan ukuran dan bentuk sel yang memang berbeda antara tumor jinak dan ganas.

Tentu ada keterbatasan yang perlu diakui. Dataset WBCD sudah cukup tua dan bersifat statis yang tidak mencerminkan variasi pasien di dunia nyata, apalagi populasi Indonesia. Model ini juga belum pernah dicoba di lingkungan klinis nyata. Ke depannya, perlu dicoba pada dataset yang lebih besar dan beragam. Eksplorasi hyperparameter tuning seperti Grid Search juga bisa membuka peluang peningkatan performa. Kalau datanya berupa citra medis langsung, pendekatan deep learning seperti CNN mungkin lebih cocok.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah mendukung pelaksanaan penelitian dan penulisan artikel ini, termasuk lembaga afiliasi penulis dan seluruh rekan yang telah memberikan masukan konstruktif.

DAFTAR PUSTAKA

[1] S. M. Khoirunnisa, D. Setiawan, M. J. Postma, and L. A. De Jong, “Cancer Treatment and Research Communications Trends in breast cancer in Indonesia from 2017 to 2020 : A national-level analysis by age and disease severity,” vol. 45, no. September, 2025.
[2] S. Zakareya, H. Izadkhah, and J. Karimpour, “A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images,” pp. 1–23, 2023.
[3] S. Devi, R. K. Ghanekar, J. A. Pande, D. Dumbre, R. Chavan, and H. Gupta, “Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models,” vol. 25, pp. 1077–1085, 2024, doi: 10.31557/APJCP.2024.25.3.1077.
[4] T. Islam et al., “Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI,” Sci. Rep., pp. 1–17, 2024, doi: 10.1038/s41598-024-57740-5.
[5] A. A. Balasubramanian et al., “Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology,” 2024.
[6] A. Jafari, “Computer Methods in Biomechanics and Biomedical Engineering : Imaging & Visualization Machine-learning methods in detecting breast cancer and related therapeutic issues : a review a review,” Comput. Methods Biomech. Biomed. Eng. Imaging Vis., vol. 12, no. 1, 2024, doi: 10.1080/21681163.2023.2299093.
[7] C. Chandra, D. P. Mulya, and Faradika, “Deteksi Serangan Siber Menggunakan Machine Learning : Studi Pada Sistem Informasi Akademik,” vol. 3, no. 2, pp. 106–110, 2025.
[8] M. D. Desriansyah, I. U. Sari, and Zulfahmi, “Analisis Efektivitas Algoritma Machine Learning dalam Deteksi Hoaks : Pada Berita Digital Berbahasa Indonesia,” vol. 3, no. 2, pp. 63–69, 2025.
[9] A. Yaqoob, N. K. Verma, M. A. Mir, G. G. Tejani, H. M. H. O. Nashwa Hassan Babiker Eisa, and M. A. Shah, “SGA-Driven feature selection and random forest classification for enhanced breast cancer diagnosis : A comparative study,” pp. 1–23, 2025.
[10] J. Ganesan et al., “Enhancing breast cancer detection accuracy through machine learning , deep learning and transfer learning techniques for clinical practice,” vol. 3, 2026.