Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

Prediksi Penyakit Kanker Payudara Menggunakan Algoritma Synthetic Minority Oversampling Technique dan Categorical Boosting Classifier

Muhamad Bintang Mandala*, Wina Witanti, Agus Komarudin

Fakultas Sains dan Informatika, Program Studi Teknik Informatika, Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Email: 1,*mbintangmandala237956@gmail.com, 2witanti@gmail.com, 3agus.komarudin@lecture.unjani.ac.id
Email Penulis Korespondensi: mbintangmandala237956@gmail.com
Submitted**: 20/05/2025;** Accepted**: 22/06/2025;** Published**: 25/06/2025**

Abstrak−Kanker payudara merupakan salah satu penyebab kematian tertinggi di dunia dan memiliki prevalensi yang tinggi di kalangan perempuan Indonesia. Model prediksi yang akurat dan efisien sangat dibutuhkan untuk mendukung deteksi dini dan menurunkan angka mortalitas. Penelitian ini bertujuan untuk mengembangkan model klasifikasi kanker payudara menggunakan algoritma CatBoost, yaitu metode boosting berbasis pohon keputusan yang mampu menangani fitur kategorikal secara langsung dan meminimalkan overfitting melalui teknik ordered boosting. Dataset yang digunakan terdiri atas fitur diagnostik tumor payudara, yang dipra-proses dengan pengecekan kelengkapan data serta transformasi atribut numerik menjadi bentuk kategorikal untuk memperjelas distribusi nilai. Ketidakseimbangan kelas antara kasus jinak dan ganas diatasi dengan metode SMOTE (Synthetic Minority Over-sampling Technique) sehingga diperoleh data latih yang seimbang. Optimasi hyperparameter dilakukan dengan Bayesian optimization, menghasilkan konfigurasi terbaik dari parameter depth, learning rate, dan L2 regularization. Model kemudian dilatih dan dievaluasi menggunakan metrik recall, akurasi, dan F1-score, dengan confusion matrix untuk menilai kualitas prediksi. Hasil menunjukkan bahwa CatBoost memiliki performa tinggi dengan recall 1,0, akurasi 98,6% , dan F1-score 0,99, serta mengungguli atau menyamai model pembanding seperti SVM, Neural Network, dan XGBoost. Temuan ini menunjukkan bahwa CatBoost merupakan algoritma yang andal dan efektif dalam mendukung pengambilan keputusan medis untuk diagnosis kanker payudara.

Kata Kunci: CatBoost; Deteksi; Data Mining; Kanker; Machine Learning

Abstract−Breast cancer remains one of the leading causes of mortality worldwide, with high prevalence rates among women in Indonesia. Accurate and efficient diagnostic models are essential to support early detection and reduce mortality. This study aims to develop a predictive model for breast cancer classification using the CatBoost algorithm, a gradient boosting method known for its ability to natively handle categorical features and reduce overfitting through ordered boosting. The dataset used consists of diagnostic features of breast tumors, which were preprocessed by checking completeness and transforming numerical attributes into categorical bins to capture value distribution more effectively. To address class imbalance between benign and malignant cases, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied, resulting in a balanced training set. Optimal hyperparameters for the CatBoost model were obtained using Bayesian optimization, with key parameters including depth, learning rate, and L2 regularization. The model was then trained and evaluated using recall, accuracy, and F1-score metrics, with a confusion matrix used to assess prediction quality. The results demonstrate that CatBoost achieved high performance with a recall of 1,0, accuracy of 98,6%, and F1-score of 0,99, outperforming or matching other benchmark models such as SVM, Neural Network, and XGBoost. These findings highlight the reliability and effectiveness of CatBoost in supporting medical decision-making for breast cancer diagnosis.

Keywords: CatBoost; Detection; Data Mining; Cancer; Machine Learning

1. PENDAHULUAN

Kanker merupakan salah satu penyebab utama kematian di seluruh dunia dan menjadi tantangan serius dalam bidang kesehatan masyarakat [1]. Berdasarkan laporan GLOBOCAN yang dirilis oleh *World Health Organization* (WHO) pada tahun 2019, kanker payudara merupakan jenis kanker yang paling sering terjadi di Indonesia dengan total 58.256 kasus baru. Angka ini menjadikan kanker payudara sebagai penyakit dengan prevalensi tertinggi di antara perempuan Indonesia [2]. Peningkatan angka kejadian setiap tahunnya menandakan perlunya pengembangan metode deteksi dan prediksi kanker yang lebih akurat, cepat, dan dapat diimplementasikan secara luas untuk mendukung upaya pencegahan dan penanganan penyakit ini [3].

Kanker payudara terjadi akibat pertumbuhan sel abnormal pada jaringan payudara yang dapat bersifat ganas dan menyebar ke jaringan tubuh lain [4]. Faktor risiko yang dapat memicu timbulnya kanker payudara mencakup faktor genetik, usia, gaya hidup, paparan hormon, hingga faktor lingkungan [5]. Deteksi dini merupakan langkah krusial dalam menurunkan angka mortalitas akibat kanker payudara, karena intervensi medis pada stadium awal memiliki peluang keberhasilan terapi yang jauh lebih tinggi dibandingkan dengan stadium lanjut [6]. Namun, proses diagnosis kanker payudara secara tradisional, seperti biopsi dan mammografi, dapat memerlukan biaya tinggi, waktu yang lama, serta tergantung pada keahlian tenaga medis [7].

Perkembangan teknologi dalam bidang kecerdasan buatan dan pembelajaran mesin (*machine learning*) memberikan peluang besar untuk meningkatkan akurasi dan efisiensi dalam proses diagnosis penyakit, termasuk kanker payudara [8]. Dengan menggunakan algoritma pembelajaran mesin, sistem dapat belajar dari data historis dan membangun model yang mampu mengklasifikasikan atau memprediksi kondisi pasien secara otomatis [9]. Salah satu bentuk implementasi dari teknologi ini adalah pengembangan sistem pendukung keputusan medis (clinical decision support system) berbasis machine learning yang dapat memberikan rekomendasi diagnosis berdasarkan data input [10].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

Berbagai penelitian terdahulu telah mengevaluasi sejumlah algoritma pembelajaran mesin dalam konteks klasifikasi kanker payudara. Support Vector Machine (SVM) merupakan salah satu algoritma yang banyak digunakan karena kemampuannya dalam menangani klasifikasi non-linear melalui penggunaan kernel [11]. Salah satu studi menunjukkan bahwa penerapan SVM dalam klasifikasi kanker payudara menghasilkan nilai *Area Under Curve* (AUC) sebesar 0,993 serta *F1-score*, *precision*, dan *recall* masing-masing sebesar 0,962 [12]. Dalam studi lain, penerapan SVM dengan optimasi hyperplane mencapai precision sebesar 1,00, recall 0,95, dan F1-score 0,98, yang menunjukkan efektivitas tinggi dari metode tersebut dalam klasifikasi data medis [13].

Selain SVM, pendekatan berbasis *Neural Network* juga telah diterapkan dan menunjukkan performa yang sangat baik. Model GRU pada studi serupa menghasilkan nilai *F1-score* sebesar 0,982, *precision* 0,980 dan *recall* sebesar 0,981 [14]. Meskipun performanya tinggi, *Neural Network* memerlukan waktu pelatihan yang lama serta sumber daya komputasi yang besar, sehingga dapat menjadi kendala dalam penerapan skala luas di institusi kesehatan yang memiliki keterbatasan infrastruktur [15].

Algoritma boosting seperti XGBoost juga menunjukkan hasil yang menjanjikan. XGBoost merupakan metode ensemble yang menggabungkan banyak pohon keputusan secara iteratif untuk meningkatkan akurasi prediksi [16]. Pada penelitian sebelumnya, XGBoost menghasilkan akurasi sebesar 0,9626, precision 0,9695, recall 0,9358, dan AUC 0,9865 dalam klasifikasi kanker payudara [17]. Hasil ini menunjukkan bahwa metode ensemble memiliki potensi yang sangat tinggi untuk diterapkan dalam diagnosis penyakit kompleks seperti kanker [18]. Namun demikian, XGBoost masih memiliki keterbatasan dalam penanganan fitur kategorikal dan terkadang rentan terhadap *overfitting* jika tidak dilakukan regularisasi yang memadai [19].

Selain itu, dalam pengembangan model prediksi kanker payudara, penting juga untuk mempertimbangkan faktor interpretabilitas dan kemudahan implementasi di lingkungan klinis [20]. Model yang kompleks memang seringkali menghasilkan akurasi tinggi, namun jika sulit untuk dipahami dan dijelaskan kepada tenaga medis, maka nilai praktisnya menjadi terbatas. Oleh karena itu, pemilihan algoritma harus seimbang antara performa prediksi dan kemampuan model untuk memberikan penjelasan yang dapat dipercaya, sehingga dapat membantu proses pengambilan keputusan medis secara efektif.

CatBoost merupakan algoritma boosting berbasis *gradient decision tree* yang dikembangkan untuk mengatasi beberapa kelemahan dari algoritma boosting tradisional [21]. Salah satu keunggulan utama CatBoost adalah kemampuannya dalam menangani fitur kategorikal secara native tanpa memerlukan proses transformasi seperti *one-hot encoding* [22]. Hal ini sangat relevan dalam konteks data medis yang sering kali terdiri atas kombinasi fitur numerik dan kategorikal. Selain itu, CatBoost mengimplementasikan skema regularisasi yang lebih stabil serta memanfaatkan ordered boosting untuk mengurangi kebocoran informasi (*information leakage*), yang umumnya terjadi pada proses pelatihan model boosting konvensional [23].

Di samping keunggulan teknisnya, CatBoost juga dikenal memiliki waktu pelatihan yang lebih cepat dan sensitivitas yang lebih rendah terhadap urutan data pelatihan [24]. Karakteristik ini menjadikannya kandidat yang potensial untuk digunakan dalam pengembangan model prediksi kanker payudara berbasis data diagnosis [25]. Melalui pemanfaatan algoritma ini, prediksi diagnosis dapat dilakukan dengan tingkat akurasi tinggi, efisiensi yang lebih baik, serta mengurangi risiko *overfitting* yang sering menjadi kendala pada model pembelajaran mesin [26].

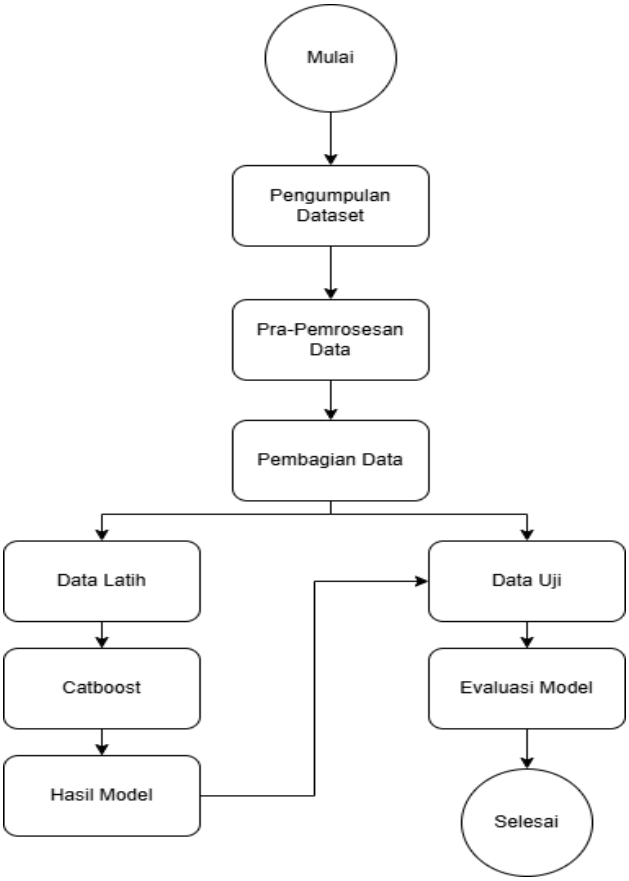
Penelitian ini bertujuan untuk merancang dan mengevaluasi model prediksi penyakit kanker payudara dengan mengimplementasikan algoritma *Synthetic Minority Over-sampling Technique* (SMOTE) sebagai metode penyeimbangan data serta *Categorical Boosting* (CatBoost) sebagai algoritma klasifikasi utama. Proses pengembangan model melibatkan tahapan praproses data, eksplorasi distribusi kelas target, penerapan teknik oversampling, serta pelatihan model klasifikasi dengan pendekatan supervised learning. Pemilihan SMOTE didasarkan pada kemampuannya untuk menghasilkan sampel sintetis yang merepresentasikan kelas minoritas secara lebih realistis dibandingkan metode oversampling konvensional, sehingga membantu mengurangi bias terhadap kelas mayoritas selama pelatihan model. Sementara itu, CatBoost dipilih karena kemampuannya dalam menangani fitur kategorikal secara native serta stabilitas performa yang tinggi pada data tabular, terutama ketika dikombinasikan dengan dataset yang telah seimbang. Evaluasi model dilakukan dengan menggunakan metrik umum klasifikasi seperti akurasi, precision, recall, dan F1-score untuk menilai sejauh mana kombinasi metode ini mampu meningkatkan kemampuan deteksi terhadap kasus kanker payudara, khususnya pada kelas minoritas yang lebih kritis. Dengan rancangan ini, penelitian secara sistematis menelaah efektivitas pendekatan gabungan SMOTE dan CatBoost dalam konteks klasifikasi diagnosis medis berbasis dataset kanker payudara yang bersifat tidak seimbang.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian yang dilakukan dapat dilihat secara jelas pada Gambar 1, yang menggambarkan urutan proses mulai dari pengumpulan data hingga evaluasi model.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403



Gambar 1. Tahapan Penelitian

2.2 Pengumpulan Dataset

Data yang digunakan dalam penelitian ini bersumber dari UCI Machine Learning Repository. Dataset yang digunakan dikenal sebagai *Breast Cancer Wisconsin* (Diagnostic) yang terdiri atas 569 entri pasien dan 31 atribut fitur. Setiap entri mewakili hasil ekstraksi fitur numerik dari citra digital massa jaringan payudara yang diperoleh melalui prosedur *Fine Needle Aspiration* (FNA). Atribut dalam dataset ini mencakup informasi statistik seperti rata-rata, deviasi standar, dan nilai ekstrem dari radius, tekstur, perimeter, area, kekompakan, kehalusan, simetri, dan dimensi fraktal. Label target yang tersedia terbagi menjadi dua kelas, yaitu *Malignant* (ganas) dan *Benign* (jinak), yang dikodekan secara biner untuk kebutuhan klasifikasi.

2.3 Pra-Pemrosesan Data

Tahap pra-pemrosesan merupakan langkah krusial dalam menyiapkan data mentah agar sesuai dengan kebutuhan pemodelan berbasis algoritma *machine learning* [27]. Proses ini diawali dengan pemeriksaan terhadap keberadaan nilai kosong atau tidak valid yang berpotensi mengganggu integritas analisis dan pelatihan model. Pemeriksaan dilakukan secara menyeluruh terhadap seluruh atribut dalam dataset untuk memastikan konsistensi dan kelengkapan data. Pada tahapan ini, tidak ditemukan adanya nilai yang hilang, sehingga tidak diperlukan proses imputasi.

Selanjutnya, dilakukan transformasi tipe data guna mengoptimalkan performa algoritma CatBoost yang memiliki keunggulan dalam menangani fitur kategorikal [28]. Karena seluruh atribut dalam dataset merupakan data numerik kontinu, maka diterapkan teknik diskretisasi atau binning untuk mengubah atribut-atribut tersebut ke dalam format kategorikal. Proses ini dilakukan dengan membagi nilai setiap atribut numerik menjadi tiga interval, yaitu rendah (*low*), sedang (*medium*), dan tinggi (*high*), berdasarkan distribusi nilai masing-masing fitur. Hasil dari proses ini menghasilkan atribut tambahan yang merepresentasikan versi kategorikal dari fitur asli. Transformasi ini dimaksudkan untuk memfasilitasi algoritma dalam membentuk pemisahan kelas yang lebih jelas berdasarkan kelompok nilai yang seragam.

2.4 Data Latih

Setelah proses transformasi selesai, data kemudian dibagi menjadi dua subset, yaitu data pelatihan (*training set*) dan data pengujian (*testing set*). Pembagian dilakukan secara proporsional untuk menjaga representasi label kelas pada masing-masing subset. Tahap ini bertujuan agar model dapat dilatih dengan cukup variasi data dan diuji pada data yang belum pernah dilihat sebelumnya guna menilai kemampuan generalisasi model [29].

Sebelum proses pelatihan dimulai, dilakukan analisis distribusi label pada data latih untuk mengidentifikasi adanya ketidakseimbangan jumlah sampel antar kelas. Ketidakseimbangan kelas dalam data pelatihan dapat

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

memengaruhi kinerja model klasifikasi, khususnya dalam mengidentifikasi kelas minoritas [30]. Untuk mengatasi hal tersebut, diterapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE), yaitu pendekatan oversampling yang bekerja dengan membuat sampel sintetis baru dari kelas minoritas melalui interpolasi antara sampel-sampel yang berdekatan dalam ruang fitur [31]. Tujuan dari teknik ini adalah menciptakan representasi kelas yang lebih seimbang agar model yang dibangun tidak bias terhadap kelas mayoritas dan mampu mendeteksi kasus minoritas secara lebih akurat [32].

2.5 Catboost

Pemodelan pada penelitian ini dilakukan menggunakan algoritma CatBoost, salah satu varian dari *gradient boosting decision tree* yang dirancang untuk mengoptimalkan pemrosesan fitur kategorikal [33]. Keunggulan utama CatBoost terletak pada penerapan *ordered boosting* yang efektif dalam mengurangi risiko *overfitting* dan *target leakage*, serta dukungan terhadap *encoding* internal fitur kategorikal tanpa perlu transformasi eksplisit dari pengguna [34].

Proses pelatihan model melibatkan eksplorasi terhadap sejumlah hyperparameter utama, termasuk *learning_rate*, *depth*, *iterations*, *l2_leaf_reg*, dan *loss_function*. Penyesuaian nilai-nilai tersebut dilakukan melalui pendekatan grid search dan validasi silang untuk memperoleh konfigurasi optimal [35]. Selain itu, digunakan pula *Bayesian Optimization* sebagai pendekatan alternatif guna mempercepat eksplorasi ruang hyperparameter secara efisien.

Selama pelatihan, evaluasi kinerja model dilakukan dengan memantau nilai *accuracy* dan *log loss* sebagai indikator utama. Untuk menghindari *overfitting*, diterapkan mekanisme *early stopping* yang secara otomatis menghentikan proses pelatihan apabila tidak terdapat peningkatan performa setelah sejumlah iterasi tertentu.

2.6 Hasil Model

Model CatBoost yang telah dilatih pada data latih selanjutnya digunakan untuk melakukan prediksi terhadap data pengujian. Proses prediksi menghasilkan label klasifikasi yang dibandingkan dengan label sebenarnya dalam data pengujian. Perbandingan ini digunakan sebagai dasar untuk mengukur sejauh mana model mampu mengklasifikasikan data secara akurat. Evaluasi terhadap hasil prediksi dilakukan melalui penghitungan berbagai metrik performa yang mencerminkan kualitas klasifikasi yang dilakukan oleh model.

2.8 Evaluasi Model

Evaluasi model dilakukan dengan pendekatan kuantitatif melalui perhitungan sejumlah metrik evaluasi klasifikasi, yaitu akurasi (*accuracy*), presisi (*precision*), sensitivitas (*recall*), dan *F1-score*. Seluruh metrik ini dihitung berdasarkan nilai-nilai yang diperoleh dari confusion matrix, yang terdiri atas empat komponen utama: *true positive* (TP), yaitu jumlah kasus positif yang berhasil diklasifikasikan dengan benar; *true negative* (TN), yaitu jumlah kasus negatif yang diklasifikasikan dengan benar; *false positive* (FP), yaitu jumlah kasus negatif yang salah diklasifikasikan sebagai positif; dan *false negative* (FN), yaitu jumlah kasus positif yang salah diklasifikasikan sebagai negatif.

Akurasi (*accuracy*) mengukur proporsi total prediksi yang benar terhadap keseluruhan jumlah data. Metrik ini memberikan gambaran umum mengenai kinerja model, terutama dalam kondisi distribusi kelas yang seimbang. Rumus akurasi dituliskan pada Persamaan (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Keterangan: TP adalah jumlah true positive, TN adalah true negative, FP adalah false positive, dan FN adalah *false negative*.

Presisi (*precision*) mengukur sejauh mana prediksi positif yang dihasilkan oleh model benar-benar merupakan kasus positif aktual. Nilai presisi yang tinggi menunjukkan bahwa kesalahan dalam mengklasifikasikan data negatif sebagai positif (*false positive*) relatif rendah. Rumus presisi diberikan pada Persamaan (2):

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

Keterangan: TP menyatakan jumlah kasus positif yang benar-benar terdeteksi, sedangkan FP menyatakan jumlah kasus negatif yang salah diklasifikasikan sebagai positif.

Recall atau sensitivitas mengukur kemampuan model dalam mendeteksi seluruh kasus positif aktual. Metrik ini sangat penting dalam konteks medis, karena kesalahan dalam mendeteksi kasus positif (*false negative*) dapat berakibat fatal. Nilai *recall* dihitung berdasarkan Persamaan (3):

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

Keterangan: TP menunjukkan jumlah kasus positif yang berhasil dikenali dengan benar, sementara FN menunjukkan jumlah kasus positif yang tidak berhasil dikenali oleh model.

F1-score merupakan rata-rata harmonik dari *precision* dan *recall*. Metrik ini berguna ketika terdapat ketidakseimbangan distribusi kelas, karena *F1-score* memberikan ukuran kinerja model yang mempertimbangkan keseimbangan antara kedua metrik tersebut. Rumus *F1-score* disajikan pada Persamaan (4):

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Keterangan: *F1-score* akan tinggi jika *precision* dan *recall* keduanya tinggi; sebaliknya, jika salah satu rendah, maka *F1-score* juga akan menurun.

Selain metrik-metrik di atas, *confusion matrix* juga digunakan sebagai alat bantu visual untuk menilai performa model klasifikasi. Matriks ini memperlihatkan distribusi jumlah prediksi yang benar dan salah untuk masing-masing kelas, sehingga dapat membantu dalam mengidentifikasi pola kesalahan yang terjadi, seperti jumlah kasus positif yang salah diklasifikasikan sebagai negatif (*false negative*) atau sebaliknya.

Dengan mengombinasikan berbagai metrik evaluasi tersebut, proses evaluasi dapat memberikan pandangan yang lebih menyeluruh terhadap kualitas prediksi model. Hal ini menjadi sangat krusial dalam konteks klasifikasi kanker payudara, di mana keseimbangan antara sensitivitas dan presisi sangat diperlukan untuk mendukung proses diagnosis yang akurat dan minim risiko klinis.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini merupakan data diagnosis kanker payudara yang diperoleh dari repositori UCI Machine Learning Repository, dengan nama resmi *Breast Cancer Wisconsin* (Diagnostic). Dataset ini secara luas digunakan dalam berbagai penelitian klasifikasi medis karena menyediakan informasi diagnostik yang kaya, yang diekstrak dari citra digital jaringan payudara.

Terdapat 569 sampel data, masing-masing dengan 30 fitur numerik dan 1 label target klasifikasi. Label diagnosis terdiri dari dua kelas, yaitu Malignant (ganas) yang direpresentasikan dengan huruf "M", dan Benign (jinak) yang direpresentasikan dengan huruf "B". Setiap baris data mencerminkan hasil diagnosis terhadap satu pasien, dengan berbagai fitur yang mencerminkan karakteristik bentuk, ukuran, dan tekstur jaringan.

Tabel 1 berikut menyajikan cuplikan contoh isi dataset, termasuk sebagian atribut yang digunakan sebagai input dan label diagnosis sebagai target.

Tabel 1. Dataset Kanker Payudara

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	…	symmetry_worst	fractal_dimension
M	17.99	10.38	122.80	1001.0	…	0.4601	0.11890
M	20.57	17.77	132.90	1326.0	…	0.2750	0.08902
M	19.69	21.25	130.00	1203.0	…	0.3613	0.08758
…	…	…	…	…	…	…	…
B	11.42	20.38	77.58	386.1	…	0.2827	0.06771
B	14.91	22.53	97.07	685.0	…	0.2128	0.06115
B	12.45	15.70	82.57	477.1	…	0.2350	0.06592

Karakteristik numerik dari dataset ini memerlukan proses pra-pemrosesan yang cermat, mengingat distribusi dan skala antar fitur dapat sangat bervariasi. Beberapa atribut memiliki rentang nilai yang besar, sementara lainnya cukup sempit. Hal ini penting untuk diperhatikan dalam tahap normalisasi maupun pemilihan algoritma. Karena seluruh fitur bersifat numerik, algoritma CatBoost yang mampu menangani fitur numerik dan kategorikal secara efisien tanpa encoding eksplisit menjadi sangat sesuai untuk digunakan dalam penelitian ini.

3.2 Pra-Pemrosesan Data

Pra-pemrosesan merupakan tahap esensial dalam rangka menyiapkan data agar siap digunakan dalam proses pelatihan model pembelajaran mesin. Proses ini dimulai dengan evaluasi kelengkapan data guna memastikan bahwa seluruh atribut memiliki nilai yang valid dan tidak terdapat nilai kosong (*missing values*) atau kesalahan input. Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung nilai hilang maupun data tidak valid. Oleh karena itu, tidak diperlukan proses imputasi atau pembersihan data lebih lanjut.

Langkah berikutnya dalam tahap pra-pemrosesan adalah melakukan transformasi terhadap atribut numerik menjadi atribut kategorikal melalui metode equal-width binning. Teknik ini bekerja dengan membagi rentang nilai dari setiap atribut numerik menjadi tiga interval yang sama lebar, yang kemudian diberi label sebagai kategori "rendah", "sedang", dan "tinggi". Meskipun algoritma CatBoost secara default mampu mengolah fitur numerik tanpa memerlukan encoding atau transformasi, pendekatan diskretisasi ini dilakukan untuk melihat bagaimana perubahan representasi data dapat memengaruhi performa model. Selain itu, diskretisasi juga membantu mengurangi pengaruh nilai ekstrim (outliers) serta mempermudah interpretasi distribusi nilai atribut dalam konteks klasifikasi.

Sebagai contoh, Tabel 2 menunjukkan tiga entri data asli untuk tiga atribut numerik, yaitu *radius_mean*, *area_mean*, dan *concavity_worst*. Data ini merepresentasikan nilai kontinu hasil pengukuran karakteristik jaringan payudara.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

Tabel 2. Dataset sebelum pre-procecssing

radius_mean	area_mean	concavity_worst
14.2	660.1	0.234
11.6	330.2	0.145
18.1	1225.4	0.316

Setelah dilakukan diskretisasi menggunakan metode equal-width binning, nilai-nilai tersebut dikonversi menjadi kategori berdasarkan rentang nilai masing-masing atribut. Konversi ini menghasilkan atribut baru yang merepresentasikan versi kategorikal dari fitur-fitur numerik awal, seperti *radius_mean_bin*, *area_mean_bin*, dan *concavity_worst_bin*. Hasil transformasi kategorikal ditampilkan dalam Tabel 3.

Tabel 3. Dataset setelah pre-procecssing

radius_mean	area_mean	concavity_worst
sedang	sedang	tinggi
rendah	rendah	sedang
tinggi	tinggi	tinggi

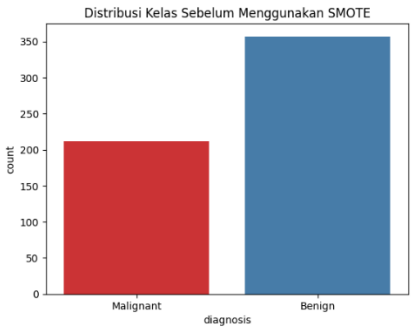
Transformasi ini memberikan informasi tambahan kepada model mengenai posisi relatif nilai fitur dalam distribusi masing-masing atribut. Dengan cara ini, model dapat menangkap perbedaan karakteristik antar entri secara lebih eksplisit melalui batasan-batasan kategori yang telah ditentukan. Representasi kategorikal ini juga dapat meningkatkan stabilitas model dalam menghadapi distribusi data yang tidak seragam atau mengandung nilai ekstrim, sehingga mendukung proses klasifikasi yang lebih robust.

3.3 Balancing Data

Distribusi kelas target dalam dataset menunjukkan ketidakseimbangan yang cukup mencolok, yang dapat memengaruhi performa model klasifikasi secara signifikan. Dari total 569 entri dalam dataset, terdapat 357 sampel yang termasuk dalam kategori jinak (*Benign*) dan 212 sampel dalam kategori ganas (*Malignant*). Ketidakseimbangan ini berpotensi menimbulkan permasalahan bias pada proses pelatihan model, khususnya karena algoritma pembelajaran mesin cenderung mengoptimalkan performa terhadap kelas mayoritas. Akibatnya, model dapat menunjukkan akurasi keseluruhan yang tinggi, namun memiliki performa yang rendah dalam mendeteksi kelas minoritas yang dalam konteks medis justru lebih krusial, seperti kasus kanker ganas.

Untuk mengatasi permasalahan ini, dilakukan proses penyeimbangan data (balancing) melalui pendekatan *oversampling* menggunakan algoritma SMOTE (*Synthetic Minority Over-sampling Technique*). Distribusi kelas sebelum diterapkannya SMOTE ditampilkan dalam Gambar 2. Visualisasi ini menunjukkan bahwa proporsi data kelas *Benign* jauh lebih dominan dibandingkan dengan kelas *Malignant*, dengan selisih jumlah yang cukup besar. Ketimpangan seperti ini berdampak pada ketidakmampuan model dalam mengidentifikasi pola dari kelas minoritas secara akurat, serta menurunkan nilai metrik evaluasi seperti *recall* dan *F1-score* pada kelas kanker ganas.

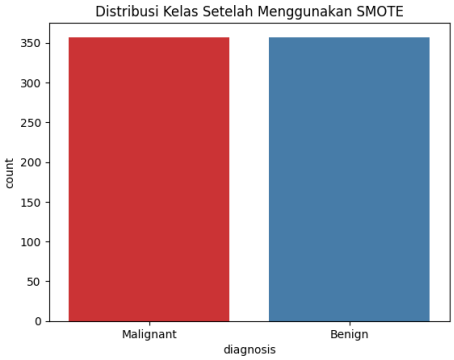
Selain itu, ketimpangan distribusi kelas yang cukup mencolok ini turut mempertegas pentingnya strategi balancing sebelum dilakukan proses pelatihan model secara menyeluruh. Dalam konteks klasifikasi medis, khususnya deteksi kanker, keberpihakan terhadap kelas mayoritas bukan hanya berpotensi menyesatkan hasil evaluasi, tetapi juga berisiko besar terhadap pengambilan keputusan klinis yang bergantung pada keluaran model. Ketidakseimbangan ini dapat menyebabkan model mengabaikan pola-pola penting yang muncul dalam kelas minoritas, sehingga sensitivitas terhadap kondisi kritis seperti kanker ganas menjadi sangat rendah. Oleh karena itu, penerapan teknik seperti SMOTE menjadi sangat relevan untuk menciptakan representasi data yang lebih setara antara kelas *Benign* dan *Malignant*. Dengan penambahan data sintetis yang menyerupai karakteristik kelas minoritas, model diharapkan mampu belajar secara lebih adil terhadap kedua kelas dan meningkatkan kepekaan terhadap kasus-kasus kanker ganas yang cenderung langka namun sangat penting untuk dikenali secara dini. Dengan distribusi data yang telah diseimbangkan, proses pelatihan model juga menjadi lebih stabil dan berpotensi menghasilkan performa evaluasi yang lebih representatif terhadap kondisi nyata di lapangan.



Gambar 2. Distribusi Kelas Sebelum Menggunakan SMOTE

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

Setelah diterapkannya metode SMOTE, jumlah entri pada kelas *Malignant* meningkat hingga setara dengan jumlah entri kelas *Benign*, masing-masing sebanyak 357 data. Proses ini menghasilkan dataset baru yang lebih seimbang dari sisi distribusi kelas target, yang dapat dilihat pada Gambar 3. Penyamaan jumlah sampel antar kelas ini dimaksudkan untuk memastikan bahwa algoritma klasifikasi memperoleh informasi yang cukup dari kedua kelas, serta mengurangi dominasi pola dari kelas mayoritas yang dapat menutupi keberadaan pola minoritas.



Gambar 3. Distribusi Kelas Setelah Menggunakan SMOTE

Dengan struktur data yang telah seimbang, model klasifikasi dapat melakukan pembelajaran terhadap kedua kelas secara lebih proporsional. Hal ini berkontribusi dalam memperbaiki sensitivitas terhadap kelas *Malignant*, yang secara klinis lebih kritis untuk diidentifikasi. Selain itu, penggunaan SMOTE sebagai teknik augmentasi data minoritas telah terbukti efektif dalam berbagai studi untuk meningkatkan ketahanan model terhadap bias distribusi kelas serta memperkuat generalisasi dalam skenario klasifikasi biner yang tidak seimbang.

3.4 Implementasi Catboost

Pelatihan model CatBoost dalam penelitian ini dimulai dengan proses pencarian kombinasi parameter yang paling optimal menggunakan pendekatan *Bayesian Optimization*. Metode ini dipilih karena lebih efisien dibandingkan pencarian acak (random search) maupun pencarian menyeluruh (grid search), terutama saat bekerja dengan ruang parameter yang cukup luas. Tiga parameter utama yang dicoba adalah kedalaman pohon (*depth*), kecepatan pembelajaran (*learning_rate*), dan nilai regularisasi L2 pada daun pohon (*l2_leaf_reg*). Rentang nilai awal untuk setiap parameter ditentukan berdasarkan pengalaman umum dalam pemodelan berbasis pohon, seperti yang ditunjukkan pada Tabel 3.

Tabel 3. Inisiasi Parameter BayesianOptimization

Parameter	Nilai Bayesian Optimization
depth	(4, 10)
learning_rate	(0.01, 0.3)
12_leaf_reg	(1, 10)

Proses *Bayesian Optimization* dijalankan sebanyak 20 kali (iterasi) untuk mencoba berbagai kombinasi parameter dan melihat pengaruhnya terhadap akurasi model. Hasil dari setiap iterasi ini dapat dilihat pada Tabel 4. Dari hasil tersebut, kombinasi parameter terbaik ditemukan pada iterasi ke-2, yaitu *depth* sebesar 7, nilai *learning_rate* sebesar 0.0552, dan nilai *l2_leaf_reg* sebesar 2.404, dengan akurasi validasi mencapai 0.979.

Tabel 4. Hasil Hyperparameter Tuning BayesianOptimization

iter	target	depth	12_leaf_reg	learning_rate
1	0.9755	6.247	9.556	0.2223
2	0.979	7.592	2.404	0.05524
3	0.9685	4.349	8.796	0.1843
4	0.9702	8.248	1.185	0.2913
5	0.9737	8.995	2.911	0.06273
6	0.9737	7.615	2.467	0.07781
7	0.972	7.571	2.39	0.1197
8	0.9755	5.951	5.312	0.05141
9	0.9702	9.658	4.311	0.07418
10	0.9702	9.365	8.754	0.1298
11	0.9702	5.994	7.119	0.2893
12	0.9772	6.207	9.554	0.1897
13	0.9755	6.188	9.554	0.1838
14	0.9772	7.662	2.993	0.2931

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

iter	target	depth	12_leaf_reg	learning_rate
15	0.9772	4.245	5.4	0.1816
16	0.9737	5.85	8.75	0.03721
17	0.9737	6.231	9.545	0.1668
18	0.9772	7.652	2.39	0.02023
19	0.9772	7.674	1.554	0.1332
20	0.9755	7.554	7.477	0.04712

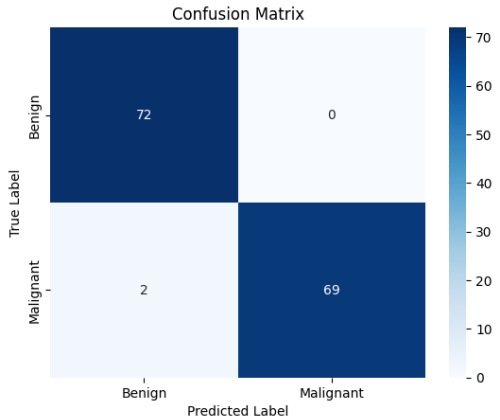
Pemilihan *depth* 7 menunjukkan bahwa model memiliki cukup fleksibilitas untuk mengenali pola yang kompleks dalam data, tapi tidak terlalu dalam sehingga masih dapat menghindari risiko *overfitting*. Nilai *learning_rate* yang kecil membuat model belajar secara bertahap, yang biasanya membantu proses pelatihan menjadi lebih stabil dan hasil akhirnya lebih baik. Sementara itu, regularisasi L2 yang sedang membantu mengontrol kompleksitas model, agar model tidak terlalu menyesuaikan diri dengan data latih.

Setelah mendapatkan parameter terbaik, model dilatih ulang pada data latih dengan jumlah iterasi sebanyak 500. Selain itu, nilai random seed juga diatur agar hasil pelatihan bisa diulang dengan konsisten. Dengan pengaturan tersebut, model CatBoost dapat menyesuaikan diri dengan karakteristik data kanker payudara secara efektif dan menghasilkan kinerja yang sangat baik.

3.5 Evaluasi Model

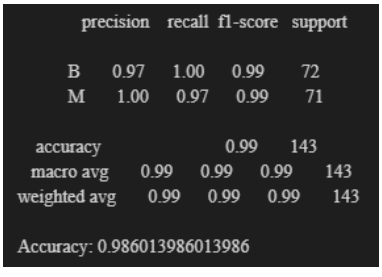
Evaluasi kinerja model CatBoost dilakukan pada data uji yang telah diseimbangkan menggunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE). Penyeimbangan ini bertujuan untuk mengatasi ketimpangan distribusi kelas, yang umum terjadi pada kasus diagnosis kanker, agar model tidak bias terhadap kelas mayoritas. Proses evaluasi dilakukan dengan mengacu pada sejumlah metrik umum dalam klasifikasi biner, yaitu akurasi, presisi, recall (sensitivitas), dan F1-score. Nilai-nilai metrik ini diperoleh dari hasil prediksi model yang kemudian diringkas dalam bentuk *confusion matrix* dan laporan klasifikasi, seperti disajikan pada Gambar 3 dan Gambar 4.

Berdasarkan Gambar 3, model menunjukkan kinerja yang sangat baik dalam membedakan antara kelas *Benign* (jinak) dan *Malignant* (ganas). Dari total 143 data uji, model berhasil mengklasifikasikan seluruh kasus *Benign* dengan benar (72/72), serta mengenali 69 dari 71 kasus *Malignant* secara akurat. Hanya dua sampel *Malignant* yang diklasifikasikan secara keliru sebagai *Benign* (*false negative*). Dengan demikian, tingkat kesalahan model, khususnya pada kasus yang berpotensi serius, tergolong sangat rendah.



Gambar 3. Confusion Matrix

Gambar 4 menampilkaan hasil evaluasi numerik yang mendetail. Model mencatatkan akurasi keseluruhan sebesar 98,6%, nilai presisi sebesar 1,00 untuk kelas Malignant, dan recall sebesar 0,97. Nilai F1-score pada kedua kelas juga sangat tinggi, yaitu 0,99, yang menunjukkan konsistensi kinerja model dalam menangani ketidakseimbangan kelas sekaligus menjaga sensitivitas dan spesifisitas. Nilai *weighted average* dan *macro average* pada seluruh metrik sama-sama sebesar 0,99, mencerminkan distribusi performa yang stabil antar kelas.



Gambar 4. Hasil Evaluasi Model

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

Tingginya nilai recall pada kelas Malignant menjadi sorotan utama dalam evaluasi ini, mengingat pentingnya deteksi dini dalam konteks diagnosis kanker. Kesalahan dalam mengklasifikasikan kasus ganas sebagai jinak (*false negative*) dapat berdampak fatal terhadap pasien, baik dari sisi keterlambatan penanganan medis maupun risiko lanjutan terhadap kesehatan. Oleh karena itu, performa recall yang tinggi menunjukkan bahwa model memiliki sensitivitas yang cukup baik dalam mengenali kasus-kasus yang bersifat kritis.

Di sisi lain, presisi yang tinggi juga menjadi nilai tambah, karena dapat meminimalkan kemungkinan prediksi positif palsu (false positive), yang berisiko menimbulkan kecemasan dan prosedur medis yang tidak perlu. Kombinasi antara recall dan presisi yang tinggi kemudian tercermin dalam nilai F1-score yang optimal, yang menunjukkan keseimbangan performa model dalam menangani dua sisi permasalahan klasifikasi.

Keberhasilan CatBoost dalam mencapai hasil ini tidak terlepas dari keunggulan algoritmik yang dimilikinya. CatBoost dirancang untuk menangani data numerik dan kategorikal secara efisien, serta memiliki teknik *ordered boosting* yang mampu mengurangi risiko *overfitting*. Model ini juga mampu mengelola data dengan korelasi tinggi antar fitur, seperti yang terdapat pada dataset ini, tanpa memerlukan praproses yang kompleks.

Selain itu, penerapan SMOTE terbukti efektif dalam mengatasi ketimpangan distribusi kelas. Ketika data pelatihan didominasi oleh kelas mayoritas, model cenderung mempelajari pola yang tidak representatif bagi kelas minoritas. Dengan penambahan sampel sintetis melalui SMOTE, model memiliki eksposur yang lebih baik terhadap karakteristik kasus Malignant, yang kemudian tercermin pada peningkatan performa evaluasi.

4. KESIMPULAN

Penelitian ini berhasil membangun model prediksi kanker payudara menggunakan algoritma CatBoost yang dioptimalkan melalui pendekatan Bayesian optimization, serta didukung oleh proses pra-pemrosesan data dan penyeimbangan kelas menggunakan SMOTE. Model yang dikembangkan menunjukkan performa tinggi dengan akurasi sebesar 98,6%, recall 1,00 untuk kelas jinak, 0,97 untuk kelas ganas, serta F1-score mencapai 0,99 pada kedua kelas, yang mengindikasikan kemampuan model dalam memberikan prediksi yang sangat akurat dan sensitif dalam membedakan kondisi kanker payudara, sehingga memiliki potensi besar dalam mendukung deteksi dini secara klinis. Untuk pengembangan lebih lanjut, disarankan agar model diuji pada dataset yang lebih besar dan beragam guna memastikan kemampuan generalisasi, serta dilakukan validasi menggunakan data klinis nyata untuk meningkatkan reliabilitas penerapan di dunia medis. Di samping itu, penerapan teknik validasi silang (cross-validation) dan evaluasi terhadap data eksternal juga penting dilakukan agar tingkat kepercayaan terhadap performa model semakin kuat, sehingga hasil prediksi dapat digunakan secara andal dalam pengambilan keputusan klinis yang bersifat kritis.

REFERENCES

[1] F. Guida *et al.*, “Global and regional estimates of orphans attributed to maternal cancer mortality in 2020,” *Nat Med*, vol. 28, no. 12, pp. 2563–2572, Dec. 2022, doi: 10.1038/s41591-022-02109-2.

[2] S. N. Rasman and H. Trustisari, “Deskriptif Literatur Review: Pendampingan Pasien Kanker Payudara pada Perawatan Paliatif,” *Advances in Cancer Science*, vol. 1, no. 1, p. 9, Jun. 2024, doi: 10.47134/acsc.v1i1.6.

[3] E. Lestari and W. Isti Rahayu, “Prediksi Keganasan Kanker Payudara Dengan Pendekatan Machine Learning,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 3, pp. 1966–1971, Nov. 2023, doi: 10.36040/jati.v7i3.6963.

[4] G. A. Marhaeni, N. N. Suindri, N. P. G. Arneni, N. Habibah, and N. N. A. Dewi, “Edukasi Tentang Kanker Payudara Meningkatkan Perilaku Pemeriksaan Payudara Sendiri (SADARI) Pada Remaja Putri,” *Jurnal Pengabdian Masyarakat Sasambo*, vol. 5, no. 2, p. 136, May 2024, doi: 10.32807/jpms.v5i2.1438.

[5] N. M. Mochtar, L. R. Aisy, D. N. Irawati, and Y. W. Finansah, “Hubungan Faktor Genetik dan Faktor Usia terhadap Kejadian Kanker Payudara pada Wanita di RSUD Dr.Soedomo Trenggalek Periode 2020-2021,” *JurnalMU: Jurnal Medis Umum*, vol. 1, no. 3, pp. 175–184, Dec. 2024, doi: 10.30651/jmu.v1i3.24772.

[6] J. Cathryne, J. E. Siahaan, M. P. Soumokil, M. T. Cengga, and M. Sampepadang, “Gambaran Faktor-Faktor Pemeriksaan Payudara Sendiri Sebagai Deteksi Dini Kanker Payudara,” *Nursing Current: Jurnal Keperawatan*, vol. 12, no. 2, pp. 214–226, Dec. 2024, doi: 10.19166/nc.v12i2.8999.

[7] I. G. B. Setiawan and P. A. T. Adiputra, “Analisis Biaya dan Manfaat Pemeriksaan CA 15-3 dalam Diagnostik dan Pemantauan Kanker Payudara di Era BPJS,” *JBN (Jurnal Bedah Nasional)*, vol. 7, no. 1, p. 23, Jan. 2023, doi: 10.24843/JBN.2023.v07.i01.p04.

[8] A. M. Majid and I. Nawangsih, “Perbandingan Metode Ensemble Untuk Meningkatkan Akurasi Algoritm Machine Learning Dalam Memprediksi Penyakit Breast Cancer (Kanker Payudara),” *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)*, vol. 23, no. 1, p. 97, Feb. 2024, doi: 10.53513/jis.v23i1.9563.

[9] J. Badriyah, Nilam Ramadhani, Agung Muliawan, Khanun Roisatul Ummah, and Ata Amrullah, “Penerapan Dimensi Reduksi Pada Machine Learning Dalam Klasifikasi Kanker Payudara Berdasarkan Parameter Medis,” *Jurnal RESTIKOM : Riset Teknik Informatika dan Komputer*, vol. 6, no. 3, pp. 526–533, Dec. 2024, doi: 10.52005/restikom.v6i3.379.

[10] K. H. Lee *et al.*, “Machine learning-based clinical decision support system for treatment recommendation and overall survival prediction of hepatocellular carcinoma: a multi-center study,” *NPJ Digit Med*, vol. 7, no. 1, p. 2, Jan. 2024, doi: 10.1038/s41746-023-00976-8.

[11] N. Arya, A. Mathur, S. Saha, and S. Saha, “Proposal of SVM Utility Kernel for Breast Cancer Survival Estimation,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 20, no. 2, pp. 1372–1383, Mar. 2023, doi: 10.1109/TCBB.2022.3198879.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 616−625
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7403

[12] J. Kusuma, B. H. Hayadi, W. Wanayumini, And R. Rosnelly, "Komparasi Metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk Klasifikasi Kanker Payudara," *MIND Journal*, vol. 7, no. 1, pp. 51–60, Jun. 2022, doi: 10.26760/mindjournal.v7i1.51-60.
[13] N. Tri, R. Adiningrum, R. Rianti, and C. Prianto, "Rancang Bangun Aplikasi Prediksi Kanker Payudara Dengan Pendekatan Machine Learning," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 11, no. 3, pp. 2830–7062, doi: 10.23960/jitet.v11i3%20s1.3351.
[14] D. R. I. M. Setiadi *et al.*, "Integrating Hybrid Statistical and Unsupervised LSTM-Guided Feature Extraction for Breast Cancer Detection," *Journal of Computing Theories and Applications*, vol. 2, no. 4, pp. 536–552, May 2025, doi: 10.62411/jcta.12698.
[15] L. J. H. Leow *et al.*, "A Convolutional Neural Network-Based Auto-Segmentation Pipeline for Breast Cancer Imaging," *Mathematics*, vol. 12, no. 4, p. 616, Feb. 2024, doi: 10.3390/math12040616.
[16] Rahmanul Hoque, Suman Das, Mahmudul Hoque, and Mahmudul Hoque, "Breast Cancer Classification using XGBoost," *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1985–1994, Feb. 2024, doi: 10.30574/wjarr.2024.21.2.0625.
[17] M. Ravly Andryan *et al.*, "Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (SVM) Untuk Diagnosa Penyakit Kanker Payudara," *Jurnal Informatika dan Komputer*, vol. 6, no. 1, pp. 1–5, 2022.
[18] A. Pandey, V. Ramesh, R. Mohan, S. Kaliappan, S. R. A, and K. Gurunathan, "Image Processing Based Early Breast Cancer Detection in Mammography Images Using GRU and XGBoost Approach," in *2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC)*, IEEE, May 2024, pp. 1–6. doi: 10.1109/ICECCC61767.2024.10593949.
[19] F. Lin *et al.*, "Identification of lysine lactylation (kla)-related lncRNA signatures using XGBoost to predict prognosis and immune microenvironment in breast cancer patients," *Sci Rep*, vol. 14, no. 1, p. 20432, Sep. 2024, doi: 10.1038/s41598-024-71482-4.
[20] E. R. Webb *et al.*, "Kindlin-1 regulates IL-6 secretion and modulates the immune environment in breast cancer models," *Elife*, vol. 12, Mar. 2023, doi: 10.7554/eLife.85739.
[21] L. Zhang and D. Jánošík, "Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches," *Expert Syst Appl*, vol. 241, p. 122686, May 2024, doi: 10.1016/j.eswa.2023.122686.
[22] M. Katlav and F. Ergen, "Improved forecasting of the compressive strength of ultra-high-performance concrete (UHPC) via the CatBoost model optimized with different algorithms," *Structural Concrete*, vol. 26, no. 1, pp. 212–235, Feb. 2025, doi: 10.1002/suco.202400163.
[23] H. Zhao, Z. Ma, and Y. Sun, "Predict Onset Age of Hypertension Using CatBoost and Medical Big Data," in *2020 International Conference on Networking and Network Applications (NaNA)*, IEEE, Dec. 2020, pp. 405–409. doi: 10.1109/NaNA51271.2020.00075.
[24] M. Mao *et al.*, "Application of FCEEMD-TSMFDE and adaptive CatBoost in fault diagnosis of complex variable condition bearings," *Sci Rep*, vol. 14, no. 1, p. 30448, Dec. 2024, doi: 10.1038/s41598-024-78845-x.
[25] A. Alobaid and T. Bonny, "A Comparative Analysis of Machine and Deep Learning Models in the Early Detection of Breast Cancer," in *2024 Advances in Science and Engineering Technology International Conferences (ASET)*, IEEE, Jun. 2024, pp. 1–9. doi: 10.1109/ASET60340.2024.10708703.
[26] S. Jin, Q. Li, S. Liu, O. K. Joel, and X. Ge, "Short - Term Temperature Forecasting Using LSTM -CatBoost Combination Method," in *2024 16th International Conference on Wireless Communications and Signal Processing (WCSP)*, IEEE, Oct. 2024, pp. 1217–1222. doi: 10.1109/WCSP62071.2024.10826699.
[27] A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," *BMC Med Res Methodol*, vol. 22, no. 1, p. 282, Nov. 2022, doi: 10.1186/s12874-022-01758-8.
[28] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: 10.1186/s40537-020-00369-8.
[29] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.
[30] P. Billion Polak, J. D. Prusa, and T. M. Khoshgoftaar, "Low-shot learning and class imbalance: a survey," *J Big Data*, vol. 11, no. 1, p. 1, Jan. 2024, doi: 10.1186/s40537-023-00851-z.
[31] S. Yan, Z. Zhao, S. Liu, and M. Zhou, "BO-SMOTE: A Novel Bayesian-Optimization-Based Synthetic Minority Oversampling Technique," *IEEE Trans Syst Man Cybern Syst*, vol. 54, no. 4, pp. 2079–2091, Apr. 2024, doi: 10.1109/TSMC.2023.3335241.
[32] W. Rahayu *et al.*, "Synthetic Minority Oversampling Technique (SMOTE) for Boosting the Accuracy of C4.5 Algorithm Model," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 3, no. 3, pp. 624–630, Jun. 2024, doi: 10.59934/jaiea.v3i3.469.
[33] M. Luo *et al.*, "Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass," *Forests*, vol. 12, no. 2, p. 216, Feb. 2021, doi: 10.3390/f12020216.
[34] S. Zhang *et al.*, "Prediction of traffic accident impact range based on CatBoost ensemble algorithm," in *Second International Conference on Algorithms, Microchips, and Network Applications (AMNA 2023*, May 2023, p. 54. doi: 10.1117/12.2679147.
[35] Y. F. Zamzam, T. H. Saragih, R. Herteno, Muliadi, D. T. Nugrahadi, and P.-H. Huynh, "Comparison of CatBoost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 2, pp. 125–136, Mar. 2024, doi: 10.35882/jeeemi.v6i2.382.