***P-ISSN : 2338-2899***
***E-ISSN: 2807-1271***

# Komparasi Performansi Algoritma Pengklasifikasi *KNN*, *Bagging* Dan *Random Forest* Untuk Prediksi Kanker Payudara

*Agung Mulyo Widodo[1*], Nizirwan Anwar [2], Bambang Irawan[3], Lista Meria [4] , Andika Wisnujati[5]*

*[1,2] Fakultas Ilmu Komputer, Universitas Unggul Esa Unggul, Jakarta Barat, 11510, Indonesia*
*[3] Fakultas Ekonomi dan Bisnis, Universitas Esa Unggul, Jakarta Barat, 11510, Indonesia*
*[4]Program Vokasi, Universitas Muhammadiyah Yogyakarta, Bantul, DIY 55183, Indonesia*

*Email Correspondent Author : agung.mulyo@esaunggul.ac.id*

***Abstract*** **—** Predictive modeling using classification techniques is one of the ways data mining is used to support decision-making systems. Numerous machine learning techniques were incorporated into the development of this predictive classification model. This study compares the accuracy of classification algorithms, specifically the Bagging, KNN, and Random forest algorithms, when used with the same dataset to diagnose breast cancer. According to the comparison, the KNN algorithm has the highest accuracy of the three algorithms, while the random forest algorithm has the lowest**.**

***Keyword*** **—** Bagging, KNN, Random forest, prediksi kanker payudara**.**

*Abstrak* — Pemodelan prediktif menggunakan teknik klasifikasi adalah salah satu cara data mining digunakan untuk mendukung sistem pengambilan keputusan. Banyak teknik pembelajaran mesin dimasukkan ke dalam pengembangan model klasifikasi prediktif ini. Penelitian ini membandingkan akurasi algoritma klasifikasi, khususnya algoritma Bagging, KNN, dan Random forest, ketika digunakan dengan dataset yang sama untuk mendiagnosis kanker payudara. Berdasarkan hasil perbandingan, algoritma KNN memiliki akurasi tertinggi dari ketiga algoritma tersebut, sedangkan algoritma random forest memiliki akurasi yang paling rendah.

*Kata kunci* — Bagging, KNN, Random forest, prediksi kanker payudara.

## I. PENDAHULUAN

Kanker payudara adalah jenis kanker yang berkembang di sel-sel payudara. Kanker payudara dapat mempengaruhi pria dan wanita sama, tetapi jauh lebih umum pada wanita dan sangat jarang pada pria. Kanker payudara merupakan penyakit dengan derajat heterogenitas yang tinggi. Saat ini merupakan penyebab utama kedua kematian akibat kanker pada wanita. Kanker payudara diklasifikasikan sebagai jinak atau ganas. Kanker payudara jinak adalah jenis kanker payudara non-invasif yang jarang menimbulkan ancaman bagi kehidupan pasien. Ini ditemukan di lapisan dalam saluran susu. Jenis kanker ini tidak menyebar ke jaringan yang berdekatan dan tetap berada di dalam saluran. Ini disebut sebagai karsinoma duktal. Karsinogenik Kanker payudara: Karsinoma lobular adalah bentuk agresif dari kanker payudara yang dimulai di lobulus payudara. Ini menyebar dari asalnya di lobulus payudara ke jaringan normal di sekitarnya, menimbulkan ancaman bagi kehidupan. Mereka kadang-kadang tumbuh kembali setelah pengangkatan. Profesional medis menganggap skrining mamografi sebagai metode deteksi dini kanker payudara yang paling dapat diandalkan. Wanita bisa berumur panjang jika kanker payudara terdeteksi dini. Karena kompleksitas kanker payudara, kemungkinan disebabkan oleh kombinasi faktor risiko. Usia, faktor genetik, dan keturunan merupakan faktor risiko kanker payudara yang tidak bisa dihindari. Obesitas, terapi sulih hormon, alkohol, dan merokok merupakan faktor risiko yang dapat dihindari. Paparan radiasi, akhir kehamilan pada usia yang lebih tua, kepadatan tulang yang tinggi, dan terapi hormon pascamenopause merupakan faktor risiko tambahan.

Data Mining (DM) adalah teknik yang digunakan untuk menemukan, pola pengetahuan baru yang tersembunyi dan berguna dari database besar. Dari statistik, kecerdasan buatan, dan gudang data, sangat mudah untuk merancang metode dan prosedur untuk mengklasifikasikan data untuk penggunaan aplikasi dunia nyata. Konsep DM sebenarnya merupakan pemisahan dari proses penemuan pengetahuan. DM telah menjadi teknologi terkini dalam penelitian yang ada dan untuk aplikasi bidang medis. Sejumlah teknik data mining telah dilakukan pada pembelajaran data mining untuk meningkatkan kinerja seperti Regresi, Algoritma Genetika, klasifikasi Bays, pengelompokan k-means, aturan asosiasi, prediksi dll. Klasifikasi adalah salah satu yang paling umum digunakan. Klasifikasi adalah salah satu masalah data mining yang paling penting. Inputnya adalah dataset dari training record, dimana setiap record memiliki beberapa atribut. Atribut numerik adalah atribut dengan domain numerik dan atribut kategoris adalah atribut dengan domain non-numerik. Selain itu, ada juga atribut khusus yang disebut label Kelas. Klasifikasi ini dimaksudkan untuk membangun model konsol yang dapat digunakan untuk memprediksi label kelas masa depan, catatan yang tidak berlabel. Ada banyak model klasifikasi yang digunakan dan kami menggunakan algoritma yang disebut Bagging, *K*-Nearest Neighbors (KNN) dan *Random Forest*. Tujuan dari studi ini adalah untuk mengkaji perbandingan penggunaan algoritma-algoritma tersebut sebagai teknik pengklasifkasi untuk memprediksi kanker payudara.

Makalah ini dibagi menjadi tujuh bagian. Bagian 1, berkaitan dengan konsep keseluruhan topik. Bagian 2, review terhadap literatur, membahas tentang algoritma-algortma yang dipakai pada penelitian dengan mengacu pada penulis yang telah melakukan penelitian tentang topik ini. Bagian 3, berkaitan dengan metode yang ada di bidang teknis. Bagian 4, berfokus pada analisis terhadap hasil penerapan algoritma yang digunakan . Bagian 5, menyimpulkan tentang penelitian yang dilakukan.

## II. TINJAUAN PUSTAKA

Sejumlah penelitian tentang penerapan teknik data mining untuk mendeteksi kanker telah dilakukan. Gökbay [1] meneliti perilaku filter Median di berbagai skala dan mengusulkan sistem yang mampu mendeteksi massa ganas di mammogram dan membantu ahli radiologi dengan skrining mamografi. Delen [2] memprediksi model survivabilitas kanker prostat menggunakan tiga teknik data mining yang terkenal, yaitu pohon keputusan, jaringan saraf tiruan, dan mesin vektor pendukung, serta satu model statistik tradisional, regresi logistik. Metode validasi silang 10 kali lipat digunakan untuk mengevaluasi prediksi selama konstruksi model. Model mesin vektor dukungan ditemukan sebagai pengklasifikasi paling sukses. Setelah memberikan informasi singkat tentang kanker payudara di bagian Pendahuluan, Danac et al. [3] memperoleh data umum tentang jaringan menggunakan program pengenalan pola Xcyt. Sel kanker payudara diidentifikasi menggunakan salah satu algoritme pohon keputusan yang tersedia di *tool* Waikato Environment for Knowledge Analysis (WEKA), khususnya algoritme C4.5. Gupta dkk. [4] melakukan penelitian kanker yang berkaitan dengan penggunaan teknik klasifikasi data mining untuk diagnosis kanker payudara. Mereka menyimpulkan bahwa penggunaan metode klasifikasi data mining dapat diterima dan dapat membantu dalam deteksi dini kanker dan menghindari biopsi. Güllüolu [5] melakukan penelitian pendahuluan menggunakan data mining untuk mendiagnosis kanker. Tujuannya adalah untuk memberikan informasi tentang bagaimana para ahli kesehatan harus melakukan pendekatan data mining di sektor kesehatan untuk mendapatkan perspektif yang berbeda tentang proses pengambilan keputusan. Kharya [6] meninjau literatur tentang diagnosis kanker payudara menggunakan teknik data mining seperti pohon keputusan, penambangan aturan asosiasi, jaringan saraf tiruan, naive bayes, jaringan bayesian, dan pengklasifikasi mesin vektor pendukung.

Pengklasifikasi yang paling sukses ditemukan sebagai pohon keputusan. Dalam sebuah penelitian yang dilakukan oleh Poyraz [7], *tool* Weka digunakan untuk memeriksa dataset kanker payudara Wisconsin dari UCI yang dikaitkan dengan kanker payudara. Algoritma data mining J48, KStar (K*), regresi logistik, dan Naive-Bayes dibandingkan untuk tingkat keberhasilannya. Dengan presisi 96,92 persen, algoritma regresi logistik ditentukan sebagai yang paling akurat. Majali dkk. [8] menggunakan teknik data mining untuk mengembangkan sistem diagnosis dan prognosis kanker. Algoritme FP Growth, sejenis algoritme penambangan aturan asosiasi, dan algoritme ID3, sejenis algoritme pohon keputusan, digunakan dalam artikel mereka untuk mendeteksi kanker pada tahap awal. Amutha dan Savithri [9] menggunakan teknik data mining untuk melakukan penelitian tentang deteksi dini kanker payudara. Penelitian mereka berpusat pada algoritma seperti Decision Trees, IBk, Support Vector Machine, Sequential Minimal Optimization (SMO), dan Neural Networks. Amutha dan Savithri menegaskan bahwa berbagai teknik klasifikasi dapat digunakan untuk mendiagnosis kanker dini secara akurat. Kaur dan Singh [10] memasukkan studi rinci tentang penyebab kanker payudara dan review penelitian tentang penggunaan teknik data mining untuk deteksi kanker payudara dalam makalah mereka. Kim dkk. [11] menggunakan mesin vektor pendukung untuk mengembangkan model prediksi baru untuk kekambuhan kanker payudara (SVM). SVM dibandingkan dengan jaringan saraf tiruan dan model regresi bahaya Coxproportional. Hasil penelitian menunjukkan bahwa model yang diusulkan untuk prediksi kekambuhan kanker payudara berdasarkan SVM (BCRSVM) efektif dalam memprediksi kekambuhan kanker payudara dengan sensitivitas tinggi (0,89), spesifisitas (0,73), nilai prediksi positif (0,75), dan nilai prediksi negatif ( -0,75). (0,89). Avşar Aydn dan Kaya Keleş [12] menggunakan dataset antena untuk mendeteksi kanker payudara menggunakan Algoritma K-Nearest Neighbor dengan validasi silang sepuluh kali lipat. Metode KNN, sebuah algoritma data mining, ditemukan untuk menghasilkan 90,0 persen hasil yang akurat.

## III. METODE

Pada penelitian ini, kami menggunakan *tool* WEKA. *Tool* ini dipilih terutama karena penggunaannya yang luas dalam literatur. Karena hal tersebut di atas, penggunaan perangkat lunak Weka dianggap lebih tepat daripada penggunaan alat penambangan data lainnya dalam pekerjaan saat ini. Selama tahap pra-pemrosesan data, kumpulan data dari berbagai sumber digabungkan dan diubah menjadi format arff, yang merupakan format file asli Weka. Mengikuti langkah-langkah ini, perangkat lunak Weka digunakan untuk menjalankan algoritme klasifikasi untuk menentukan apakah pasien menderita kanker/tumor payudara.

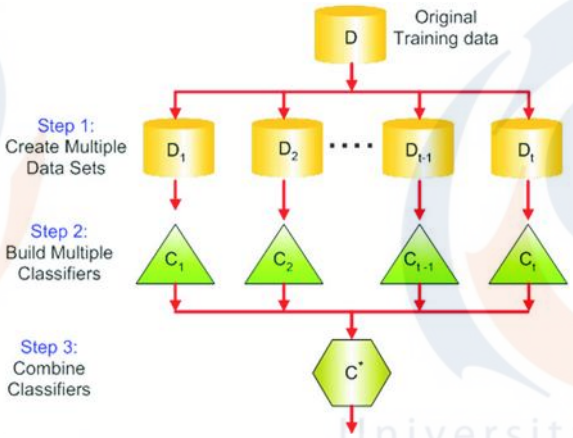
Algoritma-algoritma yang digunakan pada penelitian ini, dapat dijelaskan sebagai berikut.

### *A. Algoritma Bagging*

Breiman mengusulkan Algoritma Bagging [16] pada tahun 1994 sebagai algoritma ensemble pembelajaran mesin yang meningkatkan akurasi metode klasifikasi statistik. Algoritma Bagging berasal dari teknik "Bootstrap Aggregating" dan telah diterapkan secara luas di bidang-bidang seperti biostatistik dan penginderaan jauh. Ini digunakan dalam teknik kecerdasan buatan untuk akuisisi

informasi. Dengan mengurangi varians, algoritma ini berkontribusi untuk mencegah over-learning. Intinya, ini menghasilkan beberapa sampel pelatihan dari berbagai kombinasi data pelatihan. Algoritme Bagging dioptimalkan untuk kumpulan data pelatihan dengan dimensi yang relatif kecil. Operasi dasar algoritma digambarkan pada Gambar 1.

Set pelatihan awalnya dibagi menjadi N sub-cluster. Masing-masing sub-cluster ini berfungsi sebagai set pelatihan untuk generasi classifier. Pengklasifikasi yang menyatukan pengklasifikasi ini digunakan untuk menggabungkannya. Akibatnya, prosedur ini disebut sebagai Bagging. Untuk menyederhanakan, jika data latih terdiri dari N item, kumpulan data latih dengan N sampel disubstitusikan untuk pemilihan data acak dari kumpulan latih. Dalam hal ini, beberapa sampel dihilangkan dari set data pelatihan tertentu, sementara yang lain disertakan dalam beberapa set data pelatihan. Setiap pohon keputusan dilatih menggunakan data pelatihan yang terdiri dari sampel yang dihasilkan secara acak, dan hasilnya ditentukan oleh suara mayoritas.




Gambar 1. Prinsip kerja Algoritma Bagging [17]

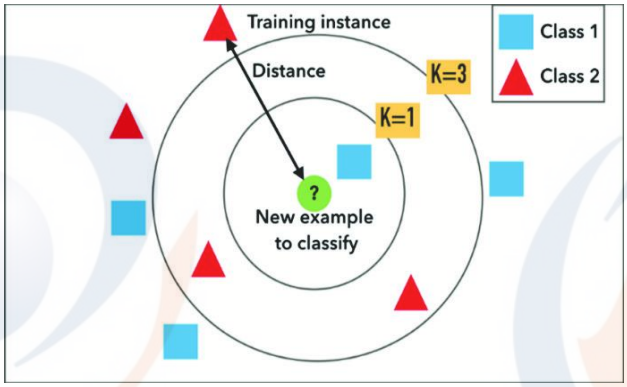
## B. Algoritma KNN (Algoritma IBK)

Algoritma IBk adalah nama Weka untuk Algoritma K-Nearest Neighbors (KNN). IBk adalah akronim dari pembelajaran berbasis *instance* dengan parameter *k*. Ini adalah algoritma klasifikasi. Dengan memanfaatkan teknik klasifikasi lazy, maka dapat menyelesaikan masalah klasifikasi. Jarak antara dua pengamatan dihitung menggunakan perhitungan rata-rata lokal pada algoritma ini [14]. Algoritma KNN ini merupakan teknik pembelajaran terawasi yang memanfaatkan teknik pembelajaran berbasis kesamaan. Algoritma non-parametrik ini adalah algoritma pembelajaran berbasis instance yang banyak digunakan yang secara akurat memprediksi kasus uji. Algoritma K-Nearest Neighbor menetapkan kelas sampel baru dengan menghitung jarak antara sampel dalam sampel saat ini (k-tetangga terdekat) dan nilai k sampel yang diberikan. Ia menemukan k sampel "dekat" dengan memanfaatkan ukuran jarak seperti Euclidean, Manhattan, Chebyshev, dan Levenshtein (Edit Jarak). Jarak Euclidian digunakan dalam penelitian ini bersama dengan Algoritma IBk di Weka karena efisiensi dan produktivitasnya, seperti yang ditunjukkan pada Persamaan (1).

$$distance\ (X_1, X_2\ ) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2)} \quad (1)$$

Dalam klasifikasi menggunakan Algoritma K-Nearest Neighbor, parameter KNN menentukan jumlah tetangga terdekat yang akan digunakan ketika mengklasifikasikan sampel uji, dan hasilnya ditentukan oleh suara mayoritas, seperti yang diilustrasikan pada Gambar 2. Weka secara otomatis memilih nilai terbaik menggunakan opsi "*cross-validation*".

Jika opsi *cross- validation* tidak digunakan, parameter *k* yang dipilih mungkin terlalu kecil atau terlalu besar. Karena itu, parameter *k* penting untuk menjaga algoritma KNN terhadap kumpulan data *noise*.



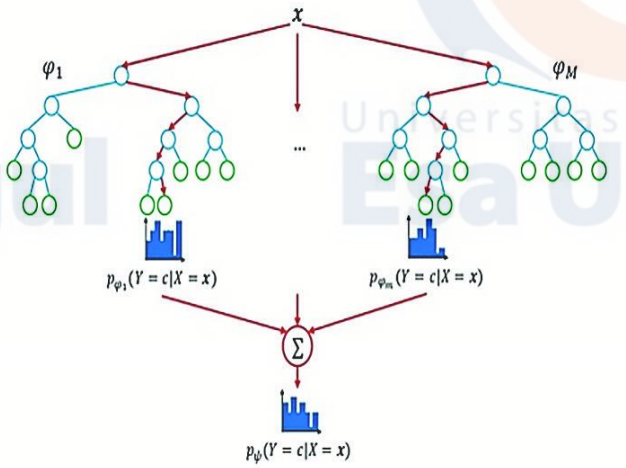

Gambar 2. Prinsip kerja Algoritma IBk (*KNN*) [17]

## C. Algoritma Random Forest (RF)

Algoritma *Random Forest* (*RF*) adalah kumpulan pengklasifikasi untuk berbagai jenis pohon. Leo Breiman [15] mempelopori penggunaan hutan acak dalam studi formal. Dia menjelaskan metode untuk menghasilkan hutan pohon yang tidak terkait dalam karyanya dengan menggabungkan prosedur Classification and Regression Treelike (CART-like) dengan optimasi node acak dan bagging. Algoritma Breiman bertujuan untuk menggabungkan keputusan dari beberapa pohon multivariat, yang masing-masing telah dilatih dalam cluster pelatihan yang berbeda, daripada menghasilkan pohon keputusan tunggal.

Algoritma *Random Forest* (*RF*) menggunakan beberapa pohon keputusan untuk menentukan nilai klasifikasi selama proses klasifikasi. Daripada membagi setiap simpul menjadi cabang berdasarkan cabang terbaik di antara semua variabel, Algoritma Hutan Acak membagi setiap simpul menjadi variabel berdasarkan variabel terbaik di antara variabel yang dipilih secara acak di setiap simpul. Setiap dataset dihasilkan menggunakan data yang telah dipindahkan dari aslinya. Kemudian, dengan menggunakan fitur seleksi acak, pohon dikembangkan tanpa dipangkas [15]. Strategi ini membedakan dan meningkatkan Algoritma

RF. Selain itu, Algoritma RF cepat, kuat untuk kemampuan beradaptasi yang ekstrim, dan mampu bekerja dengan jumlah pohon yang tidak terbatas.




Gambar 3. Prinsip kerja Algoritma *Random Forest* (*RF)* [13]

Akurasi klasifikasi Algoritma RF tergantung pada parameter yang ditentukan pengguna seperti N (jumlah pohon) dan m (jumlah variabel / parameter yang digunakan di setiap node). Akibatnya, memilih parameter terbaik untuk data meningkatkan akurasi klasifikasi. Breiman [15] menegaskan bahwa pengaturan jumlah variabel m sama dengan akar kuadrat dari jumlah total variabel M (jumlah variabel keseluruhan) umumnya menghasilkan solusi optimal. Algoritma RF menghasilkan pohon melalui penggunaan algoritma Classification and Regression Tree (CART). Percabangan dilakukan pada setiap node sesuai dengan kriteria algoritma CART (misalnya, indeks GINI). Untuk menentukan kriteria percabang an digunakan metode Koefisien GINI. Algoritma Random Forest diilustrasikan pada Gambar 3.

## IV. HASIL DAN DISKUSI

Data kanker payudara diambil dari situs web (http://tunedit.org/rsist of epo/UCI/breast-cancer.arff). Dataset ini berisi 286 instance dan sepuluh atribut, sembilan di antaranya unik dan satu di antaranya khusus untuk kelas. Di bawah ini adalah dataset dan daftar atributnya.

Tabel 1. Atribut Dataset Kanker Payudara

| No | Attribut | Data Type | Data Missing |
|---|---|---|---|
| 1 | Age | Nominal | 0 |
| 2 | Menopause | Nominal | 0 |
| 3 | Tumor-Size | Nominal | 0 |
| 4 | Inv-Nodes | Nominal | 0 |
| 5 | Node-Caps | Nominal | 8 |
| 6 | Deg-Malig | Nominal | 0 |
| 7 | Breast | Nominal | 0 |
| 8 | Breast-Quad | Nominal | 1 |
| 9 | Irradiat | Nominal | 0 |
| 10 | Class | Nominal | 0 |

Setelah itu, dataset dikonversi ke format arff, yang kompatibel dengan *tool* Weka dilakukan pra-pemrosesan memiliki peran penting karena sebaik apapun teknik yang digunakan jika data masukan mengandung kesalahan maka kesimpulan yang dihasilkan juga akan salah. Tahapan pra-pemrosesan meliputi pembersihan data, integrasi, transformasi hingga reduksi data. Tahap preprocessing juga akan membantu data analyst memahami data. Semakin banyak data yang dipahami, semakin baik model yang dihasilkan.

Pada dataset asli terdapat 1 (satu) data yang hilang pada atribut breast-quad dan 8 data yang hilang pada atribut node-caps. Data blank bisa disebabkan oleh banyak hal seperti kesalahan operator, kesalahan sensor, atau memang datanya tidak ada. Pada dataset asli terdapat 1 (satu) data yang hilang pada atribut breast-quad dan 8 data yang hilang pada atribut node-caps, seperti tampak pada gambar 4. Data blank bisa disebabkan oleh banyak hal seperti kesalahan operator, kesalahan sensor, atau memang datanya tidak ada. Hal ini perlu diatasi karena beberapa teknik pembelajaran mesin tidak ingin memproses data kosong atau dapat menyebabkan model yang dihasilkan menjadi tidak akurat. Salah satu cara untuk menangani data yang hilang adalah dengan menghapus baris, hal ini dapat dilakukan jika jumlahnya tidak terlalu banyak. Di bagian ini digunakan filter, yag terdapat pada *tool* WEKA.

Kemudian, dengan menggunakan alat perangkat lunak data mining *Knowledge Extraction based on Evolutionary Learning* (*Keel*), teknik *cross-validation* 10 *fold* digunakan untuk mendapatkan hasil yang paling akurat. Untuk mendapatkan ukuran kinerja yang paling akurat, digunakan metode *cross-validation k*-fold. Setelah pengacakan data, kumpulan data asli yang diacak dibagi menjadi *k* bagian yang sama dengan menggunakan metode validasi silang *k*-fold.

Dalam makalah kami, kami memilih nilai $k = 10$, yang berarti bahwa satu lipatan telah dihapus per iterasi dan setiap lipatan digunakan sekali untuk pengujian dan sembilan kali untuk pelatihan. Hal ini karena dalam *k*-fold cross validation, ketika satu bagian disisihkan, *k*-1 bagian yang tersisa menjalani pelatihan dan kinerjanya diukur. Prosedur ini diulang k kali dengan bagian yang berbeda dihapus setiap kali. Jadi, pada akhir proses, skor sepuluh hasil – yang diperoleh dari sepuluh pelatihan dan sepuluh set tes – (nilai *k*) dirata-ratakan untuk menghasilkan prediksi tunggal, dan hasil nyata diperoleh dengan mengambil rata-rata dari ini sepuluh hasil.

Model dapat mengevaluasi seberapa baik prediksi kelas instance yang merupakan% tertentu dari semua data (hold-out). Besar % menurut masukan pengguna (default: 66). Data latih dibagi menjadi dua menurut persentase, bagian pertama adalah data uji, bagian kedua adalah data latih. Dalam makalah ini kami memilih sesuai default *tool*

WEKA, yakni 66% untuk persentase pemisahan (*percentage split*).

Kami menjalankan semua pengklasifikasi yang digunakan pada makalah ini di Weka di setiap kelas. Karena fakta bahwa akurasi adalah metrik kinerja yang disukai, hasil akurasi untuk algoritme ini ditampilkan di Tabel. 1 dan pengukuran akurasi ditunjukkan pada Persamaan. (2) [14]. True Positive, True Negative, False Positive, dan False Negative adalah singkatan dari True Positive, True Negative, False Positive, dan False Negative, masing-masing. Hasil pengujian tipe TP menunjukkan bahwa model benar mengidentifikasi kelas positif, sedangkan hasil pengujian tipe TN menunjukkan model benar mengidentifikasi kelas negatif. FP menunjukkan hasil tes di mana model mengidentifikasi kelas positif salah, sedangkan FN menunjukkan hasil tes di mana model mengidentifikasi kelas negatif salah. Dengan demikian, akurasi didefinisikan sebagai proporsi prediksi yang benar yang dibuat oleh model dan dihitung dengan membagi jumlah prediksi yang benar dengan jumlah total prediksi.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Dari dataset dimana dengan total 286 rekor dalam dataset kanker payudara, Setelah pra-pemrosesan, jumlah kasus akan menjadi 277 yang digunakan dalam proses klasifikasi. Di antara 277 kejadian tersebut, 196 kejadian tergolong kejadian tidak berulang (*No-recurrence-events*) dan 81 kejadian tergolong kejadian berulang (*Recurrence-events***)**.

Matriks konfusi (*Confusion Matrix*) dari masing-masing algoritma, dijelaskan di bawah ini.

Tabel 2. Matriks Konfusi Algoitma Bagging

| Class Target | No-recurrence-events | Recurrence-events |
|---|---|---|
| No-recurrence-events | 183 | 13 |
| Recurrence-events | 61 | 20 |

Statistik kinerja yang menunjukkan seberapa akurat model pengklasifikasi dapat memprediksi kelas sebenarnya dari setiap instance menurut opsi pengujian. Untuk opsi uji *cross validation* 10 kali lipat yang dipilih, ada 277 instans yang diuji dalam 10 iterasi. Pada hasil yang ditunjukkan, terdapat 203 kejadian (73,29%) yang diprediksi dengan benar oleh kelas, dan 74 kejadian lainnya (26,72%) yang salah diprediksi.. Matriks konfusi yang menunjukkan berapa banyak instance yang diprediksi untuk setiap kelas (per kolom) dan jumlah instance yang sesuai dengan label aktual (per baris). Dalam matriks tersebut ada 244 kejadian yang diprediksi atau diklasifikasikan sebagai 'Ya'. Dari 244 contoh ini, 183 contoh diberi label 'Ya' (juga disebut Positif Benar), dan 61 contoh diberi label 'Tidak' (juga disebut Positif Palsu). Nilai Instans yang diklasifikasikan dengan Benar dalam Ringkasan 213 (73,29%) adalah nilai (183 + 20) dari matriks konfusi ini. Sebaliknya, nilai *Instances Classified Instances* diklasifikasikan dengan Salah pada Ringkasan 74 (26,72%) adalah nilai (61 + 13) dari matriks konfusi ini. Akurasi dari algoritma ini berdasrkan Persamaan.(2) adalah 73,29%.

Selanjutnya untuk algoritma IBk (KNN), ditampilkan pada table 3. Untuk opsi uji *cross validation* 10 kali lipat yang dipilih, ada 277 instans yang diuji dalam 10 iterasi. Pada hasil yang ditunjukkan, terdapat 206 kejadian (74,37%) yang diprediksi dengan benar oleh kelas, dan 71 kejadian lainnya (25,63%) salah diprediksi.

Tabel 3. Matriks Konfusi Algoitma IBk

| Class Target | No-recurrence-events | Recurrence-events |
|---|---|---|
| No-recurrence-events | 176 | 20 |
| Recurrence-events | 51 | 30 |

Dari matriks yang menunjukkan berapa banyak instance yang diprediksi untuk setiap kelas (per kolom) dan jumlah instance yang sesuai dengan label aktual (per baris). Dalam matriks ini ada 206 kejadian yang diprediksi atau diklasifikasikan sebagai "Ya". Dari 206 contoh ini, 176 contoh diberi label "Ya" (juga disebut Positif Benar), dan 51 contoh diberi label "Tidak" (juga disebut Positif Salah). *Incorrectly Classified Instances* dalam ringkasan 206 (73,29%) adalah nilai (176 + 30) dari matriks konfusi ini. Sebaliknya, nilai *Incorrectly Classified Instances* pada Ringkasan 71 (25,63%) adalah nilai (20 + 51) dari matriks konfusi ini. Akurasi dari algoritma ini adalah 74,37%.

Untuk algoritma Random Forest, ditampilkan pada tabel 4.

Tabel 4. Matriks Konfusi Algoitma Random Forest

| Class Target | No-recurrence-events | Recurrence-events |
|---|---|---|
| No-recurrence-events | 174 | 22 |
| Recurrence-events | 53 | 28 |

Untuk opsi uji *cross validation* 10 *fold* yang dipilih, ada 277 instans yang diuji dalam 10 iterasi. Pada hasil yang ditunjukkan, terdapat 202 kejadian (72,92%) yang diprediksi dengan benar oleh kelas, dan 75 kejadian lainnya (27,08%) yang salah diprediksi. Matriks konfusi yang menunjukkan berapa banyak instance yang diprediksi untuk setiap kelas (per kolom) dan jumlah instance yang sesuai dengan label aktual (per baris). Dalam contoh ini ada 202 kejadian yang diprediksi atau diklasifikasikan sebagai "Ya". Dari 202 contoh ini, 174 contoh diberi label "Ya" (juga disebut Positif Benar), dan 28 contoh diberi label "Tidak"

(juga disebut Positif Salah). Nilai Instance yang *Correctly Classified Instances* dalam Ringkasan 202 (72,92%) adalah nilai (174 + 28) dari matriks konfusi ini. Sebaliknya, nilai Instance yang *Incorrectly Classified Instances* dalam ringkasan 75 (27,08%) adalah nilai (22 + 53) dari matrks konfusi ini.

## V. KESIMPULAN

Diagnosis penyakit adalah tugas yang sangat menantang di bidang perawatan kesehatan. Banyak teknik data mining yang digunakan dalam proses pengambilan keputusan. Dalam percobaan kami, kami telah menerapkan teknik klasifikasi data mining Bagging, IBk (K-Means) dan RandomForest yang digunakan untuk mengklasifikasikan penyakit kanker payudara berulang dan tidak berulang. Kinerja pengklasifikasi dievaluasi melalui matriks kebingungan dalam hal akurasi. Algoritma IBk memberikan 74,37 % yang memberikan akurasi yang lebih baik daripada Algoritma lainnya. Sebagai pekerjaan masa depan, teknik yang sama digunakan untuk menerapkan dataset penyakit lain seperti Diabetes, Penyakit jantung, Kanker paru-paru, dan lain sebagainya.

## DAFTAR ACUAN


[1] İ. Z. Gökbay, *Machine learning techniques in breast cancer detection*. MSc, Bahçeşehir University Institute of Science and Technology, İstanbul, Turkey, 2007

[2] D. Delen, ”Analysis of cancer data: a data mining approach,” *Expert Syst*, 26, 100-112. 2009 https://doi.org/10.1111/j.1468-0394.2008.00480.x

[3] M. Danacı, M.Çelik, & A. E. Akkaya, “Prediction and diagnosis of breast cancer cells using data mining methods,” *ASYU'2010*, Kayseri, Turkey, 9-12, 2010.

[4] S. Gupta, K. Dharminder, & S. Anand, (),”Data mining classificati on techniques applied for breast cancer diagnosis and prognosis,” *Indian J ComputSci and Engineer*, 2, 188-195, 2011.

[5] S. S. Güllüoğlu, “Data Mining Studies in Medical and Healthcare: A Preliminary Study for Cancer Diagnosis,” *AJIT-e: Online Academic Journal of Information Technology*, 2, 355-360, 2011

[6] S. Kharya, “Using data mining techniques for diagnosis and prognosis of cancer disease,” *Int J ComputSci, Engineer and Inf Tech*, 2, 55-66, 2012. https://doi.org/10.5121/ijcseit.2012.2206.

[7] O. Poyraz, *Data mining aplications in medicine: Breast cancer data set analysis*, MSc, Trakya University Institute of Science and Technology, Edirne, Turkey, 2012

[9] J. Majali, R. Niranjan, V. Phatak, & O. Tadakhe, “ Data Mining Techniques for Diagnosis and Prognosis of Breast Cancer,” *Int J ComputSci and Inf Tech*, 5, 6487-6490, 2014.

[10] R. Amutha, & M. Savithri, “Diagnosis and Prognosis of Breast Cancer Using Data Mining Techniques,” *Paripex Indian J Res*, 4, 6-8, 2015.

[11] S. Kaur, & R. Singh,”A Review of Data Mining Based Breast Cancer Detection and Risk Assessment Techniques,” *Int J ComputSci and Inform Secur*, 14, 241-251. 2016.

[12] W. Kim, K. S. Kim, J. E. Lee, D. Y. Noh, S. W Kim, Y. S. Jung, et al.,” Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine. *J Breast Canc*,”15,230-238,2012. https://doi.org/10.4048/jbc.2012.15.2.230

[13] E. A. Aydın, & M. K. Keles,” Breast cancer detection using K-nearest neighbors data mining method obtained from the bow-tie antenna dataset.” International Journal of RF and Microwave Computer-Aided Engineering, 27(6). 2109. https://doi.org/10.1002/mmce.

[14] N. S. Altman, ”An Introduction to Kernel and Nearest Neighbor Non-parametric Regression,” *The American Statistician,46*(3), 175. 1992. https://doi.org/10.2307/2685209

[15] L. Breiman, *Random forests*, Mach Learn, 45, 5-32. 2001. https://doi.org/10.1023/A:1010933404324.

[16] L Breiman, *Bagging predictors*, *Machine Learning*, *24*(2), 123, 1996. https://doi.org/10.1007/bf00058655

[17] M. Kaya Keleş,” Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study,” Technical Gazette 26, 1, 149-155, 2019.