ISSN: 2477-5126
e-ISSN: 2548-9356

Data Mining berbasis *Nearest Neighbor* dan Seleksi Fitur untuk Deteksi Kanker Payudara

Yohanes Setiawan[1*)],
[1]Jurusan Teknologi Informasi, Fakultas Teknologi Informasi dan Bisnis, Institut Teknologi Telkom Surabaya
[1]Jln. Ketintang no. 156, Kota Surabaya, 60231, Indonesia
[1]email: yohanes@ittelkom-sby.ac.id

***Abstract* – Detecting breast cancer in early stage is not straightforward. This happens because biopsy test requires time to determine whether the type is benign or malignant. Data mining algorithm has been widely used to automate diagnosis of a disease. One of popular algorithms is nearest neighbor based because of its simplicity and low computation. However, too many features can cause low accuracy in nearest neighbor based models. In this research, nearest neighbor based with feature selection is developed to detect breast cancer. Conventional k-Nearest Neighbor (KNN) and Multi Local Means k-Harmonic Nearest Neighbor have been chosen as nearest neighbor based models to experiment. The feature selection method used in this study is filter based, namely Correlation based, Information Gain, and ReliefF. The experimental result shows that the highest recall metric of MLM-KHNN and Correlation based is 92% with 5 features, MLM-KHNN and Information Gain is 94% with 5 features, and MLM-KHNN and ReliefF is 94% with 5 features. Then MLM-KHNN and Information Gain is compared with MLM-KHNN and ReliefF through confusion matrix. MLM-KHNN and Information Gain have higher accuracy and precision than MLM-KHNN and ReliefF. In brief, MLM-KHNN algorithm with Information Gain can increase the recall of the prediction of breast cancer compared with the conventional K-NN algorithm and have been deployed into website using Streamlit such that the model can be used to detect breast cancer from chosen Wisconsin dataset features.**

***Abstrak* – Deteksi kanker payudara pada tahap awal tidak mudah. Hal ini disebabkan oleh tes *biopsy* memerlukan banyak waktu untuk menentukan kankernya berjenis jinak atau ganas. Algoritma *data mining* telah banyak digunakan secara luas untuk melakukan otomatisasi diagnosis penyakit. Salah satu algoritma *data mining* yang populer adalah algoritma berbasis *nearest neighbor* karena kesederhanaan dan komputasinya yang rendah. Namun, terlalu banyak fitur dapat mengakibatkan akurasinya bernilai rendah. Tujuan dari penelitian ini adalah melakukan deteksi kanker payudara berbasis *nearest neighbor* dengan seleksi fitur. Algoritma k-Nearest Neighbor (KNN) konvensional dan *Multi Local Means k-Harmonic Nearest Neighbor* (MLM-KHNN) dipilih sebagai model berbasis *nearest neighbor* yang digunakan dalam penelitian ini. Selanjutnya, metode seleksi fitur yang digunakan adalah *filter based*, yakni *Correlation based, Information Gain,* dan ReliefF. Hasil eksperimen menunjukkan bahwa recall tertinggi dari MLM-KHNN dan *Correlation based* mencapai 92% dengan 5 fitur, MLM-KHNN dan *Information Gain* mencapai 94% dengan 5 fitur, dan MLM-KHNN dan ReliefF mencapai 94% dengan 5 fitur. Selanjutnya dibandingkan *confusion matrix* dari MLM-KHNN dan *Information Gain* dan MLM-KHNN dan ReliefF. Akurasi dan *precision* dari MLM-KHNN dan *Information Gain* lebih tinggi daripada MLM-KHNN dan ReliefF. Dapat disimpulkan bahwa MLM-KHNN dan *Information Gain* dapat meningkatkan recall prediksi dari kanker payudara jika dibandingkan dengan KNN konvensional dan telah melalui proses *deployment* ke dalam website menggunakan Streamlit sehingga model dapat digunakan untuk mendeteksi kanker payudara menggunakan fitur-fitur terpilih dari dataset Wisconsin yang diperoleh.**

***Kata Kunci* – Kanker Payudara, Data Mining, Nearest Neighbor, Seleksi Fitur**

I. PENDAHULUAN

Kanker payudara merupakan jenis kanker yang berasal dari jaringan kelenjar payudara atau juga bisa dari jaringan lemak atau jaringan ikat dalam payudara. Apabila ditemukan benjolan pada payudara ataupun terdapat penemuan yang tidak normal pada *mammogram*, uji *biopsy* diperlukan untuk diagnosis kanker payudara. Di Indonesia, terdapat 42 dari 100.000 penduduk terjangkit penyakit ini dan rata-rata kematiannya mencapai 17 per 100.000 penduduk [1], [2].

Pada umumnya, diagnosis kanker payudara yang terdeteksi awalnya sebagai tumor dapat dikelompokkan menjadi 2 jenis, yakni *benign* (jinak) dan *malignant* (ganas) [2]. Agar bisa menegakkan diagnosis ini, diperlukan tenaga kesehatan dengan kecukupan pengalaman dalam menangani beberapa kasus dalam kanker payudara. Pengalaman tersebut umumnya dapat dicapai oleh para dokter yang telah berada pada pertengahan karirnya atau yang telah berpengalaman mendiagnosis beberapa pasien dengan gejala yang bervariasi [3]. Meski demikian, tidak ada yang menjamin kebergantungan pada akurasinya dan membutuhkan waktu yang lama bagi dokter untuk mengelompokkan kanker ini. Oleh sebab itu, proses otomatisasi deteksi kanker payudara menggunakan komputer semakin meningkat pesat menggunakan teknik *data mining* yang memanfaatkan historikal data pasien [2].

Beragam algoritma *data mining* dapat digunakan untuk melakukan deteksi kanker payudara, seperti berbasis pohon keputusan (*tree*), maupun jaringan saraf (*neural network*). Namun, algoritma-algoritma tersebut memiliki komputasi yang tinggi [3] sehingga membuat *modelling* dan *deployment* membutuhkan waktu lebih. Kemudian, akurasi yang dihasilkan juga tidak tinggi sehingga mempengaruhi performa deteksi yang membutuhkan sistem yang akurat [1]. Algoritma berbasis *nearest neighbor* dikenal sebagai algoritma dengan

*) **penulis korespondensi**: Yohanes Setiawan
Email: yohanes@ittelkom-sby.ac.id

komputasi yang cepat dan prediksi yang akurat menggunakan kedekatan jarak antara data latih dengan data uji [4].

Penelitian ini menggunakan *data mining* berbasis *Nearest Neighbor* dengan Seleksi Fitur untuk deteksi kanker payudara. Algoritma berbasis *Nearest Neighbor* yang dipilih adalah KNN dan MLM-KHNN. KNN dipilih karena komputasi yang ringan dan akurasi yang baik. Sementara MLM-KHNN merupakan perbaikan dari metode KNN dengan performa yang sama dengan KNN namun dengan akurasi yang lebih baik melalui penggunaan rata-rata harmonik sebagai basis pemilihan *class*. Seleksi fitur digunakan untuk mengatasi dimensi data yang tinggi sehingga akan memperberat komputasi pelatihan dan eksekusi program. Metode seleksi fitur yang digunakan berjenis *filter based* dikarenakan kecepatannya dalam memilih fitur melalui hubungan untuk setiap fitur dengan class-nya. Terdapat 3 jenis metode seleksi fitur *filter based* yang digunakan, yaitu Correlation, Information Gain, dan ReliefF. Ketiga metode ini popular digunakan untuk seleksi fitur dan telah memiliki implementasinya melalui aplikasi data mining yakni WEKA. Tujuan dari penelitian ini adalah meningkatkan keakuratan prediksi jenis kanker payudara melalui komparasi antar metode berbasis *nearest neighbor* disertai seleksi fitur. Rancangan percobaan dilakukan dengan membandingkan antara KNN dengan MLM-KHNN baik tanpa seleksi fitur maupun dengan seleksi fitur untuk setiap metode seleksi fitur berjenis filter based pada tahap *pre-processing* data.

II. PENELITIAN YANG TERKAIT

Metode berbasis *nearest neighbor* pertama kali diperkenalkan melalui KNN. KNN merupakan salah satu algoritma populer yang termasuk dalam daftar 10 algoritma yang sederhana namun efektif dalam berbagai kasus pengenalan pola [5]. Hal ini disebabkan oleh mudahnya perhitungan KNN dan kecepatannya dalam memproses pelatihan (*training*) dan pengujian (*testing*). Studi mengenai KNN telah bervariasi dan diimplementasikan ke dalam beragam permasalahan klasifikasi. Permasalahan yang dimaksud mencakup diagnosis penyakit pada bidang pengolahan citra [6], pengolahan teks untuk analisis sentiment [7], penanggulangan bencana alam [8], dan khususnya pada bidang kesehatan [9]. Selain itu, metode KNN dikembangkan sedemikian rupa sehingga menjadikannya lebih optimal dibandingkan dengan versi orisinalnya [5], [10]–[12]. Kekurangan dari metode KNN adalah pada sensitifitas pemilihan nilai ketetanggaan k yang menjadi hyperparameter-nya. Perbedaan nilai k dapat mempengaruhi akurasi/recall yang cukup signifikan pada model. Salah satu pengembangan KNN yang mengatasi hal tersebut adalah Multi Local Means k-Harmonic Nearest Neighbor (MLM-KHNN) [11]. MLM-KHNN adalah pengembangan KNN atau berbasis *nearest neighbor* yang menggunakan k vektor rata-rata multi lokal dengan implementasi rata-rata harmonik sebagai metrik pengukuran jarak antar sampel. Penggunaan k vektor rata-rata multi local menangkal kekurangan KNN mengenai sensitifitas terhadap pemilihan k. Rata-rata harmonik digunakan untuk mengurangi *error rate* pada klasifikasi yang disebabkan oleh *outlier*. Beberapa penelitian telah menggunakan MLM-KHNN sebagai metode klasifikasi pada kasusnya [13], [14]. Penelitian ini mengusulkan kebaruan ilmiah berupa kerangka kerja dalam pembuatan sistem deteksi kanker payudara dengan metode berbasis *nearest neighbor*, yakni KNN konvensional dan variasi KNN terbaru yakni MLM-KHNN yang dilatih menggunakan fitur-fitur pilihan yang melalui proses seleksi fitur.

III. DATA MINING, KLASIFIKASI BERBASIS *NEAREST NEIGHBOR,* DAN SELEKSI FITUR

Pada bab ini akan dijelaskan landasan teori bagi penelitian ini yang terdiri dari *data mining*, seleksi fitur berjenis *filter based*, klasifikasi berbasis *Nearest Neighbor*, dan metrik evaluasi klasifikasi yang digunakan.

A. Data Mining

Data mining adalah studi mengenai ekstraksi pengetahuan dan pengenalan pola dalam data. Melalui data mining, pengenalan pola diproses melalui pengumpulan dan penggunaan data yang bersifat historikal [15]. Beragam teknik atau algoritma yang digunakan senantiasa bergantung dari penyelesaian masalah secara tepat.

Terdapat 2 jenis model dalam data mining, yakni deskriptif dan prediktif [16]. Model deskriptif digunakan untuk menemukan pola yang mendeskripsikan data dan dapat diinterpretasi menjadi sebuah *insight*. Sementara model prediktif digunakan untuk memprediksi suatu atribut spesifik yang tidak diketahui berdasarkan atribut-atribut lainnya, termasuk didalamnya adalah klasifikasi dan regresi. Klasifikasi melakukan prediksi terhadap data kategorikal, dan regresi melakukan prediksi terhadap data kontinu. Pada penelitian ini, deteksi kanker payudara menggunakan model

B. Seleksi Fitur Berjenis Filter Based

Seleksi Fitur merupakan bagian dari tahap *pre-processing* pada *data mining* yang bertujuan untuk memilih sebagian fitur yang spesifik untuk digunakan dalam tahap pemodelan data [17]. Pada umumnya, seleksi fitur memiliki beberapa jenis teknik. Pertama, seleksi fitur yang berjenis *wrapper-based* yang paling populer. Jenis ini melakukan seleksi fitur dikombinasikan dengan algoritma pembelajaran terawasi (supervised learning) dalam prosesnya. Kekurangan utama jenis ini adalah komputasi yang sangat mahal karena harus mencari jumlah fitur yang optimal pada dimensi yang sangat besar. Kedua, seleksi fitur yang berjenis filter-based merupakan jenis seleksi fitur yang sederhana. Jenis ini menggunakan teknik perhitungan matematika dan statistika untuk menentukan fitur-fitur penting yang seharusnya digunakan dalam model machine learning. Terdapat 2 kategori seleksi fitur dengan jenis filter-based, yakni *rank based* dan *subset evaluation based*. *Rank based* didasarkan pada teknik statistika univariat untuk mengevaluasi ranking dari setiap fitur tanpa mempertimbangkan interrelationship antar fitur yang terjadi. Sementara kategori *subset evaluation based* menggunakan teknik statistika multivariat untuk mengevaluasi ranking dari subset keseluruhan fitur yang ada. Pemilihan fitur pada *filter based* didasarkan pada penilaian karakteristik penting pada data [18].

Salah satu metode seleksi fitur filter based adalah Pearson Correlation (PCor). PCor diturunkan dari ilmu statistika dan menjadi pengukuran yang umum dalam mengidentifikasi kebergantungan diantara dua atau lebih variabel acak. Pengukuran kesamaan (*similarity*) yang mengevaluasi penting atau tidaknya suatu atribut dapat dilakukan dengan perhitungan PCor antara atribut independent (fitur) dengan atribut dependennya (kelas) [19]. PCor antara fitur dan kelasnya dapat dilihat pada (1):

$$PCor = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}} \quad (1)$$

dengan X adalah himpunan atribut fitur pada dataset dan Y adalah kelasnya, sedangkan n adalah himpunan keseluruhan atribut fitur dan kelas.

Metode lainnya adalah Information Gain (IG). Pada umumnya, IG digunakan pada data yang berdimensi tinggi untuk mengukur efektifitas fitur dalam penggunaan klasifikasi. IG melakukan seleksi fitur berdasarkan nilai entropi untuk setiap fitur pada dataset [20]. Setiap fitur diperiksa satu per satu dan nilai gain-nya dihitung untuk mengukur seberapa penting fitur tersebut terhadap kelasnya. Fitur-fitur pada dataset di-ranking berdasarkan nilai IGnya. Fitur dengan nilai IG yang rendah berarti tidak terlalu berpengaruh secara signifikan pada data klasifikasi. Sehingga, fitur dengan nilai IG yang rendah dapat dibuang tanpa memiliki dampak yang besar terhadap model klasifikasi. Nilai IG dari kelas Y dan fitur inputan X dikalkulasi melalui (2):

$$IG(X) = H(Y) - H_X(Y) \quad (2)$$

$H(Y)$ menghitung entropi dari kelas Y menggunakan rumus dibawah ini. Entropi merupakan fungsi matematika yang berkorespondensi terhadap kuantitas informasi yang termuat pada sumbernya. Entropi dari kelas Y untuk menentukan relevansi dihitung seperti pada (3):

$$H(Y) = \sum_i -P(v_i)\log_2 P(v_i) \quad (3)$$

dimana untuk setiap i, $P(v_i)$ adalah probabilitas yang memiliki nilai v_i.

Kemudian, ReliefF adalah salah satu metode seleksi fitur filter based dari algoritma Relief yang melakukan pemilihan atribut/fitur berdasarkan kesesuaiannya dengan target/kelas pada kasus klasifikasi [21]. Setiap fitur diberi bobot yang akan digunakan untuk mengenali elemen-elemen yang saling berdekatan. Algoritma Relief mencari elemen-elemen terdekat pada kelas yang sama dan yang berbeda untuk setiap fitur yang ada didalam sebuah dataset. ReliefF merupakan jenis Relief yang dapat mengatasi keberadaan noise dan dataset yang bersifat multiclass, yakni dataset yang memiliki lebih dari dua target melalui perhitungan relevansi bobot untuk setiap fitur. Dari m sampel contoh, dipilih secara acak sebuah sampel contoh (R) berdasarkan selisih antara R dengan contoh terdekat yang sama (H) dan kelas yang berbeda serta melakukan update pada nilai-nilai yang relevan. Kemudian bobot diberikan untuk membedakan contoh tersebut dengan tetangga (neighbor) yang memiliki kelas yang berbeda. Dengan mempertimbangkan kontribusi rata-rata kesalahan terdekat $M(c)$ untuk mengupdate bobotnya, perhitungan probabilitas dari masing-masing kelas juga dilakukan. Bobot X_i dari fitur ke-i di-update melalui persamaan (4):

$$w_i = w_i - \frac{\psi(X_i, R, H)}{m} + \sum_{C \neq C_R} \frac{P(C)\psi(X_i, R, M(c))}{m} \quad (4)$$

dimana fungsi $\Psi(X_i, R, H)$ menghitung jarak antara sampel contoh R dengan contoh terdekat H

C. Klasifikasi Berbasis Nearest Neighbor

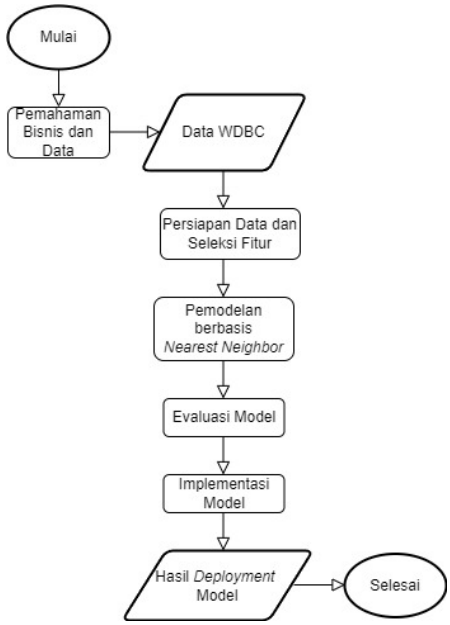
k-Nearest Neighbor, yang biasa disingkat KNN, merupakan salah satu algoritma yang popular karena sederhana namun efektif untuk mengenali pola [5]. Algoritma berbasis *nearest neighbor* berawal dari KNN yang melakukan pembelajaran melalui perbandingan sejumlah data testing dan data training yang memiliki kemiripan. Data training dideskripsikan pada n atribut. Setiap tupel menggambarkan sebuah titik pada ruang dimensi n. Artinya, seluruh data training disimpan dengan pola ruang dimensi n. Ketika diberikan tupel yang tidak diketahui, tetangga terdekat k mencari ruang pola untuk tupel training k yang berjarak dekat dengan tupel tersebut. k tupel training itulah yang disebut sebagai k "tetangga terdekat" dari tupel yang tidak diketahui didalam data testing yang belum terlihat. Namun, penggunaan *majority vote* untuk menentukan kelas dari tupel yang tidak diketahui tersebut kurang sesuai, sehingga [11] mengusulkan Multi Local k-Harmonic Nearest Neighbor (MLM-KHNN).

MLM-KHNN memanfaatkan sebanyak mungkin k multi rata-rata local (*multi local means*) untuk setiap kelas untuk mengurangi sensitifitas pemilihan k. Disamping itu, rata-rata harmonik digunakan untuk pertama kali sebagai metrik kesamaan (*similarity*). Algoritma MLM-KHNN secara lengkap tersaji sebagai berikut:

1. Dapatkan k "tetangga terdekat" untuk setiap kelas ω_j dari sampel training diantara samel testing x dengan persamaan jarak Euclidean.
2. Hitung k vektor multi-local mean untuk setiap kelas ω_j melalui persamaan (5):
$$m_r = \frac{1}{r}\sum_{i=1}^{r} y_{ij} \quad (5)$$
dengan $1 \leq r \leq k$, y_{ij} adalah fitur dari sampel training.
3. Hitung jarak rata-rata harmonik HD antara sampel testing x terhadap k vektor multi-local mean melalui persamaan (6):
$$HD\left(x, m_k(x)\right) = \frac{k}{\sum_{r=1}^{k} \frac{1}{d(x, m_r)}} \quad (6)$$
dengan $d(x, m_r)$ adalah persamaan jarak Euclidean antara sampel testing x terhadap k vektor multi-local mean pada Langkah (2).
4. Berikan label/class pada sampel testing x kepada class ω_j yang memiliki HD minimum.

IV. METODE PENELITIAN

Metode penelitian dapat dilihat pada Gbr. 1 Metode penelitian yang dilakukan berpedoman pada CRISP-DM yang menjadi dasar tahapan proses dari pengolahan *data mining*.



Gbr. 1 Diagram Alur Metode Penelitian

A. Pemahaman Bisnis dan Data

Pemahaman bisnis diperlukan untuk memahami tujuan bisnis dari penelitian. Tujuan bisnis dari penelitian ini adalah menemukan model terbaik berbasis *nearest neighbor* melalui banyaknya fitur yang optimal melalui seleksi fitur yang ada.

Dataset yang digunakan adalah Wisconsin Dataset Breast Cancer (WDBC) yang diperoleh melalui UCI Machine Learning Repository. WDBC merupakan bentuk numerik dari ekstraksi fitur pada citra digital dari *Fine Needle Aspirate* (FNA) massa payudara yang mendeskripsikan karakteristik inti sel. Terdapat 1 kolom yang menjadi *class* dan 31 kolom yang menjadi fitur yang terdiri dari ID Pasien (V1), Diagnosis (V2), Rata-rata radius (V3), rata-rata tekstur (V4), rata-rata perimeter (V5), rata-rata area (V6), rata-rata smoothness (V7), rata-rata compactness (V8), rata-rata concavity (V9), rata-rata concave points (V10), rata-rata simetri (V11), rata-rata dimensi fractal (V12), standard error dari radius (V13), standard error dari tekstur (V14), standard error dari perimeter (V15), standard error dari area (V16), standard error dari smoothness (V17), standard error dari compactness (V18), standard error dari concavity (V19), standard error dari concave points (V20), standard error dari simetri (V21), standard error dari dimensi fraktal (V22), Worst Radius (V23), Worst Tekstur (V24), Worst Perimeter (V25), Worst Area (V26), Worst Smoothness (V27), Worst Compactness (V28), Worst Concativity (V29), rata-rata Worst Concave Points (V30), Worst Simetri (V31), dan Worst Dimensi Fractal (V32).

B. Persiapan Data dan Seleksi Fitur

Persiapan data dilakukan untuk mempersiapkan dataset sebelum masuk ke tahap pemodelan. Berdasarkan dataset yang digunakan, dilakukan beberapa tahapan persiapan data sebagai berikut:

- *Menghapus atribut/fitur yang tidak relevan*

  Penghapusan atribut/fitur yang tidak relevan dilakukan untuk menghindari fitur yang tidak relevan dalam pemodelan data. Dalam konteks penelitian ini, fitur tersebut adalah fitur "ID" yang merupakan nomor kartu identitas pasien.

- *Melakukan Encoding*

  Tahap *encoding* diperlukan agar memastikan bahwa tipe data setiap fitur berjenis numerikal. Pada penelitian ini, *class* data WDBC berjenis *string*, yakni B (*benign)* dan M (*malignant*). Proses *encoding* dilakukan dengan mengubah B menjadi 0 (nol, karena *benign* bukan kanker ganas) dan M menjadi 1 (satu, karena *malignant* merupakan kanker ganas).

- *Membagi Data Latih dan Data Uji*

  Pembagian data latih (*training*) dan data uji (*testing*) perlu dilakukan sebelum tahapan selanjutnya untuk menghindari *data leakage*, yakni kebocoran data akibat telah diketahui/dilihat dari proses pelatihan. Penelitian ini membagi dataset menjadi 70% data latih dan 30% data uji dengan memastikan bahwa setiap *class* memiliki proporsi yang sama (tidak ada yang lebih banyak ataupun lebih sedikit antara data latih dan data uji)

Selanjutnya, proses seleksi fitur dilakukan menggunakan data latih yang telah siap diolah. Metode seleksi fitur yang digunakan adalah Correlation, Information Gain, dan ReliefF. Eksperimen pemilihan fitur dilakukan dengan memanfaatkan perangkat lunak *Waikato Environment for Knowledge Analysis* (WEKA) yang dikembangkan oleh *University of Waikato* [22]. WEKA mengimplementasikan Persamaan (1) untuk menghitung korelasi antar atribut pada metode seleksi fitur Correlation, kemudian persamaan (2) dan (3) untuk menghitung *information gain* (IG) pada metode seleksi fitur Information Gain, dan persamaan (4) untuk perhitungan bobot pada metode seleksi fitur ReliefF. Jumlah fitur yang dipilih adalah 5 fitur, 10 fitur, dan 15 fitur (maksimal top 50% dari 30 fitur yang ada di data WDBC) untuk setiap metode seleksi fitur dikarenakan semakin tinggi nilai dari Correlation/Information Gain/ReliefF maka semakin tinggi pula *importance* (kepentingan) antara fitur tersebut dengan *class*-nya.

C. Pemodelan berbasis Nearest Neighbor

Tahap pemodelan data merupakan tahap membuat model *machine learning* yang dilakukan berbasis *nearest neighbor*. Algoritma berbasis *nearest neighbor* yang dipilih adalah KNN dan MLM-KHNN. Metode KNN dan MLM-KHNN diimplementasikan menggunakan Python. Metode KNN konvensional diimplementasikan menggunakan *Scikit-Learn* [23] pada library Python yang mengimplementasikan pemilihan kelas berdasarkan tetangga berjarak terdekat. Metode MLM-KHNN dibangun secara *scratch* menggunakan bahasa pemrograman Python berdasarkan 4 langkah algoritma pada bab sebelumnya yang memuat persamaan (5) dan (6). Penelitian ini akan membandingkan KNN dan MLM-KHNN untuk 3 jenis skenario percobaan, yakni menggunakan seleksi fitur Correlation, Information Gain, dan ReliefF untuk 5 fitur, 10 fitur, dan 15 fitur). Pemilihan *hyperparameter* pada KNN dan MLM-KHNN adalah nilai k yang berkisar dari 1-10 dan akan diambil k dengan metrik recall terbaik dalam setiap skenario-nya.

D. Evaluasi Model

Pada umumnya, klasifikasi menggunakan confusion matrix untuk menentukan performa algoritmanya [24].

ISSN: 2477-5126
e-ISSN: 2548-9356

Confusion matrix merupakan tabel dua dimensi yang menggambarkan "aktual" dan "prediksi", dan kedua dimensi memiliki True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Pemahaman terhadap keempat nilai tersebut dapat dilihat pada Tabel I.

TABEL I.
CONFUSION MATRIX

Kelas	Positif Prediksi	Negatif Prediksi
Positif Aktual	TP	FN
Negatif Aktual	FP	TN

Evaluasi model dilakukan untuk memilih model terbaik yang siap di-*deploy* dari tahap pemodelan sebelumnya. Metrik evaluasi yang digunakan adalah recall (*sensitivity*) dari *class* "1 (*malignant*)" karena bertujuan untuk meminimalisir *False Negative* (sistem yang mendeteksi pasien sebagai *benign*/jinak meskipun seharusnya yang benar adalah *malignant*/ganas) untuk menghindari kesalahan penanganan pada pasien yang seharusnya ditangani secara kanker ganas. Untuk setiap model KNN dan MLM-KHNN, model dengan recall tertinggi dipilih sebagai model yang terbaik.

E. Implementasi Model

Implementasi model merupakan tahap produksi (*deployment*) dari model terbaik yang terpilih dari metrik recall terbaik. Proses *deployment* dilakukan berbasis website secara *localhost* menggunakan Streamlit.

V. HASIL DAN PEMBAHASAN

Berdasarkan metode penelitian yang telah dijelaskan, penelitian ini memiliki tujuan bisnis yakni membangun model optimal berbasis *nearest neighbor* untuk deteksi kanker payudara. Kemudian, dataset WDBC dilakukan tahap persiapan data melalui penghapusan fitur yang tidak sesuai, yakni ID Pasien, transformasi *encoding* pada *class* yang masih belum bertipe data numerik, dan pembagian data menjadi 70% data latih dan 30% data uji.

Tahapan seleksi fitur dilakukan pada data latih menggunakan metode Correlation, Information Gain, dan ReliefF. Dataset yang telah bersih menjadi *input* ke dalam WEKA untuk dilakukan *ranking* fitur berdasarkan nilai yang tertinggi untuk masing-masing metode. Contoh tampilan output WEKA untuk seleksi fitur dapat dilihat pada Gbr. 2.

```
Attribute selection output

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 31 diagnose):
        Information Gain Ranking Filter

Ranked attributes:
 0.671   23 perimeter_worst
 0.657   21 radius_worst
 0.65    24 area_worst
 0.644    8 concave points_mean
 0.616   28 concave points_worst
 0.564    3 perimeter_mean
 0.524    4 area_mean
 0.518    1 radius_mean
 0.493   14 area_se
 0.471    7 concavity_mean
 0.454   27 concavity_worst
 0.365   13 perimeter_se
 0.359   11 radius_se
 0.309    6 compactness_mean
 0.288   26 compactness_worst

Selected attributes: 23,21,24,8,28,3,4,1,14,7,27,13,11,6,26 : 15
```

Gbr. 2 Contoh Tampilan *Output* WEKA pada Seleksi Fitur

Dari 30 fitur yang ada, dipilih 5, 10, dan 15 fitur dengan *ranking* tertinggi dari masing-masing metode seleksi fitur. Fitur-fitur tersebut menjadi masukan kedalam model berbasis *nearest neighbor* untuk dibandingkan performansinya. Fitur-fitur yang terpilih dari proses seleksi fitur dapat dilihat pada Tabel II.

TABEL II.
DAFTAR FITUR YANG TERPILIH

Metode	Banyaknya Fitur	Fitur-fitur yang Terpilih
Correlation	5	V25, V30, V23, V10, V26
	10	V25, V30, V23, V10, V26, V5, V3, V6, V9, V29
	15	V25, V30, V23, V10, V26, V5, V3, V6, V9, V29, V8, V28, V13, V16, V15
Information Gain	5	V25, V23, V26, V10, V30
	10	V25, V23, V26, V10, V30, V5, V6, V3, V16, V9
	15	V25, V23, V26, V10, V30, V5, V6, V3, V16, V9, V29, V15, V13, V8, V28
ReliefF	5	V23, V24, V25, V30, V26
	10	V23, V24, V25, V30, V26, V10, V3, V5, V6, V4
	15	V23, V24, V25, V30, V26, V10, V3, V5, V6, V4, V9, V27, V29, V13, V7

Setelah memperoleh informasi terkait fitur-fitur yang terpilih dari ketiga metode seleksi fitur, metode berbasis *nearest neighbor* diimplementasikan ke dalam bahasa pemrograman Python melalui Jupyter Notebook untuk mencari model terbaik. KNN dan MLM-KHNN memiliki *hyperparameter* yang sama, yakni nilai k. Data latih akan dilakukan proses pelatihan menggunakan nilai k dengan range 1-10 untuk setiap metode. Nilai k dari model yang memiliki recall terbaik akan dipilih sebagai *hyperparameter* model optimalnya

ISSN: 2477-5126
e-ISSN: 2548-9356

Uji coba pertama dilakukan melalui proses seleksi fitur dengan metode Correlation. Performansi model dapat dilihat pada Tabel III. Model terbaik tetap dipegang oleh MLM-KHNN dengan recall sebesar 92% dengan besaran *hyperparameter* k optimal yang mendominasi sebesar 3. k optimal diperoleh melalui nilai recall terbesar dari iterasi nilai k dari k = 1 hingga k = 10. Hanya ada satu model, yakni model KNN konvensional dengan banyaknya fitur sebanyak 10 memiliki k bernilai besar yakni k = 9. Artinya, dibutuhkan 9 *neighbors* untuk bisa menemukan recall terbaik dari model dengan performansi yang terbaik. Dalam hal ini, MLM-KHNN menunjukkan konsistensi metrik yang tidak berubah pada jumlah fitur yang mengalami perubahan signifikan. Sementara itu, KNN menunjukkan perbedaan metrik ketika jumlah fitur ikut berubah.

TABEL III.
RECALL PADA UJI COBA CORRELATION

Banyaknya Fitur (n)	KNN	MLM-KHNN
$n = 5$	86% ($k_{optimal} = 3$)	92% ($k_{optimal} = 4$)
$n = 10$	84% ($k_{optimal} = 9$)	92% ($k_{optimal} = 3$)
$n = 15$	83% ($k_{optimal} = 3$)	92% ($k_{optimal} = 3$)

Uji coba kedua dilakukan melalui proses seleksi fitur dengan metode Information Gain. Performansi model dapat dilihat pada Tabel IV. Berbeda dengan Correlation, seleksi fitur dengan Information Gain menunjukkan perbedaan metrik recall yang signifikan pada MLM-KHNN, yakni sebesar 94%. Sementara KNN memiliki hasil metrik yang nyaris serupa dengan seleksi fitur Correlation, dengan terdapat perbedaan pada stagnansi metrik pada saat banyaknya fitur sebesar 10 dan 15 fitur. Melalui perhitungan *ranking* berbasis entropinya, Information Gain mampu menaikkan akurasi sebesar 2% dibandingkan dengan Correlation. Sementara, *hyperparameter* k pada MLM-KHNN menunjukkan konsistensi sebesar 3.

TABEL IV.
RECALL PADA UJI COBA INFORMATION GAIN

Banyaknya Fitur (n)	KNN	MLM-KHNN
$n = 5$	86% ($k_{optimal} = 3$)	94% ($k_{optimal} = 3$)
$n = 10$	84% ($k_{optimal} = 9$)	92% ($k_{optimal} = 3$)
$n = 15$	84% ($k_{optimal} = 9$)	92% ($k_{optimal} = 3$)

Uji coba ketiga dilakukan melalui proses seleksi fitur dengan metode ReliefF. Performansi model dapat dilihat pada Tabel V. Model seleksi fitur ReliefF menghasilkan nilai recall yang sama dengan model Information Gain untuk setiap skenario banyaknya fitur. Hal ini dapat disebabkan oleh komputasi Information Gain dan ReliefF memiliki kemiripan, yakni menghitung probabilitas antara *class* dengan masing-masing fitur. Perbedaannya terletak pada perhitungan probabilitas untuk entropi pada Information Gain dan perhitungan probabilitas untuk perhitungan bobot yang mengukur kedekatan antar elemen. Nilai k pada metode MLM-KHNN menunjukkan konsistensi seperti model sebelumnya saat menggunakan seleksi fitur Information Gain. Artinya, pemilihan fitur mempengaruhi nilai k terbaik dalam menentukan recall yang tertinggi. Sehingga dapat terlihat bahwa sensitifitas MLM-KHNN dalam memilih *hyperparameter*-nya sangat kecil jika dibandingkan dengan KNN konvensional yang berubah ketika pemilihan k-nya berubah. Terlihat pula bahwa model menggunakan seleksi fitur dengan ReliefF memiliki hasil yang sama dengan model yang menggunakan seleksi fitur Information Gain. Oleh karena itu, performansi keduanya perlu dibandingkan dengan metrik-metrik lainnya seperti *precision*, akurasi, dan *f1-score* untuk banyaknya fitur $n = 5$ dengan metode MLM-KHNN (model dengan recall tertinggi) agar dapat dipilih model yang terbaik untuk menjadi model yang siap untuk dilanjutkan pada proses *deployment*.

TABEL V.
RECALL PADA UJI COBA RELIEFF

Banyaknya Fitur (n)	KNN	MLM-KHNN
$n = 5$	86% ($k_{optimal} = 3$)	94% ($k_{optimal} = 3$)
$n = 10$	84% ($k_{optimal} = 9$)	92% ($k_{optimal} = 3$)
$n = 15$	84% ($k_{optimal} = 9$)	92% ($k_{optimal} = 3$)

Metrik-metrik tambahan untuk penggunaan metode Information Gain dan ReliefF pada 5 fitur ditunjukkan pada Gbr. 3 dan Gbr. 4 Hasil tersebut menunjukkan bahwa meskipun keduanya memiliki hasil recall yang sama, namun metrik yang lain berbeda. Akurasi pada seleksi fitur Information Gain lebih tinggi dibandingkan Relief-F, serta recall pada class Benign dan precision pada class Malignant berturut-turut hanya menyentuh 74% dan 68%. Oleh karena itu, dipilih MLM-KHNN dengan $k = 3$ pada 5 fitur pilihan dari hasil metode seleksi fitur Information Gain sebagai model yang terbaik.

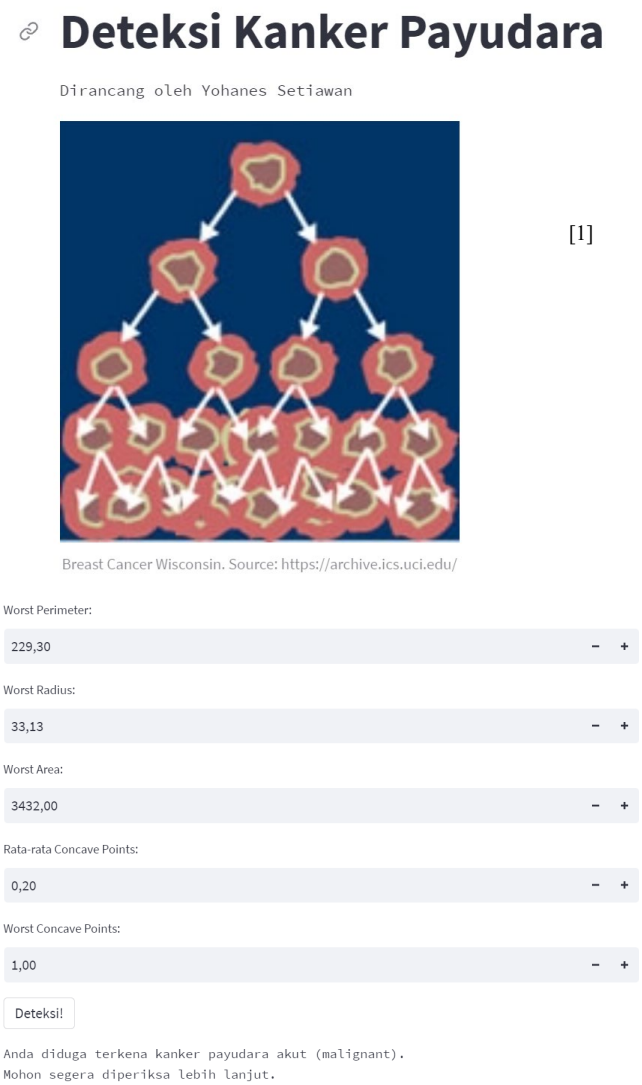
	precision	recall	f1-score
0	0.96	0.80	0.87
1	0.74	0.94	0.83
accuracy			0.85

Gbr. 3 Metrik Tambahan pada MLM-KHNN menggunakan Metode Seleksi Fitur Information Gain

	precision	recall	f1-score
0	0.95	0.74	0.83
1	0.68	0.94	0.79
accuracy			0.81

Gbr. 4 Metrik Tambahan pada MLM-KHNN menggunakan Metode Seleksi Fitur ReliefF

Selanjutnya, model di-*deploy* ke dalam website berbasis *localhost* menggunakan Streamlit. Contoh tampilan *website* dapat dilihat pada Gbr. 5 Implementasi model melalui *deployment* pada website perlu dilakukan agar *user* dapat memanfaatkan model yang telah dirancang sebagai deteksi awal kanker payudara dari data WDBC yang telah diperoleh. Penggunaan seleksi fitur mempermudah proses implementasi model sehingga *user* tidak perlu memberikan banyak *input* agar bisa melakukan deteksi kanker payudara melainkan cukup menggunakan 5 *input* fitur saja.



Gbr. 5 Contoh Tampilan *Deployment* Model

VI. KESIMPULAN

Penelitian ini mengkaji deteksi kanker payudara melalui teknik data mining berbasis *nearest neighbor* dengan seleksi fitur. Metode *nearest neighbor* yang digunakan adalah k-Nearest Neighbor (KNN) dan Multi Local Mean k-Harmonic Nearest Neighbor (MLM-KHNN) yang diimplementasikan pada Python dengan 3 metode seleksi fitur berjenis filter based, yaitu Correlation, Information Gain, dan ReliefF pada skenario 5 fitur, 10 fitur, dan 15 fitur. Hasil uji coba menunjukkan bahwa MLM-KHNN pada $k = 3$ dengan metode seleksi fitur Information Gain memiliki nilai recall terbaik sebesar 94% disertai metrik-metrik lainnya yang berada diatas 70%. Seleksi fitur yang dilakukan membantu proses *deployment* model sehingga hanya 5 fitur saja yang diaplikasikan ke dalam website. Proses *deployment* model dilakukan dengan Streamlit sehingga dapat diaplikasikan didalam keseharian. Saran untuk penelitian selanjutnya dapat dikembangkan melalui uji coba dengan metode-metode *machine learning* terbaru lainnya sehingga dapat memiliki hasil dengan metrik yang lebih baik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Institut Teknologi Telkom Surabaya yang telah memberikan dukungan kepada penelitian ini.

DAFTAR PUSTAKA

[1] 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.

[2] A. S. Elkorany, M. Marey, K. M. Almustafa, and Z. F. Elsharkawy, "Breast Cancer Diagnosis Using Support Vector Machines Optimized by Whale Optimization and Dragonfly Algorithms," *IEEE Access*, vol. 10, pp. 69688–69699, 2022, doi: 10.1109/ACCESS.2022.3186021.

[3] M. Monirujjaman Khan *et al.*, "Machine Learning Based Comparative Analysis for Breast Cancer Prediction," *J Healthc Eng*, vol. 2022, 2022, doi: 10.1155/2022/4365855.

[4] A. W. Satria Bahari Johan, S. W. Putri, G. Hajar, and A. Y. Wicaksono, "Modified KNN-LVQ for Stairs Down Detection Based on Digital Image," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 12, no. 3, p. 141, Nov. 2021, doi: 10.24843/lkjiti.2021.v12.i03.p02.

[5] J. Gou *et al.*, "A representation coefficient-based k-nearest centroid neighbor classifier," *Expert Syst Appl*, vol. 194, May 2022, doi: 10.1016/j.eswa.2022.116529.

[6] C. Paramita, E. Hari Rachmawanto, C. Atika Sari, and D. R. Ignatius Moses Setiadi, "Klasifikasi Jeruk Nipis Terhadap Tingkat Kematangan Buah Berdasarkan Fitur Warna Menggunakan K-Nearest Neighbor," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 4, no. 1, pp. 1–6, Jan. 2019, doi: 10.30591/jpit.v4i1.1267.

[7] D. Apriliani, A. Susanto, M. Fikri Hidayattullah, and G. Wiro Sasmito, "Sentimen Analisis Pandangan Masyarakat Terhadap Vaksinasi Covid 19 Menggunakan K-Nearest Neighbors," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 1, 2023.

[8] A. A. A'Ziziyyah, I. I. Nugroho, R. Sabillillah, B. A. S. Aji, and K. Amiroh, "Perbandingan Sistem Deteksi Banjir Menggunakan Algoritma Naive Bayes Dan K-NN Berbasis IOT," *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 7, no. 1, 2022.

[9] T. A. Assegie, "An optimized K-Nearest neighbor based breast cancer detection," *Journal of Robotics and Control (JRC)*, vol. 2, no. 3, pp. 115–118, May 2021, doi: 10.18196/jrc.2363.

[10] M. J. Vikri *et al.*, "Penerapan Fungsi Exponential Pada Pembobotan Fungsi Jarak Euclidean Algoritma K-Nearest Neighbor," *Generation Journal*, vol. 6, no. 2

[11] Z. Pan, Y. Wang, and W. Ku, "A new k-harmonic nearest neighbor classifier based on the multi-local means," *Expert Syst Appl*, vol. 67, pp. 115–125, Jan. 2017, doi: 10.1016/j.eswa.2016.09.031.

[12] Z. Pan, Y. Pan, Y. Wang, and W. Wang, "A new globally adaptive k-nearest neighbor classifier based on local mean optimization," *Soft computing*, vol. 25, no. 3, pp. 2417–2431, Feb. 2021, doi: 10.1007/s00500-020-05311-x.

[13] T. Widiharih and M. A. Mukid, "Credit Scoring Menggunakan Metode Local Means Based K Harmonic Nearest Neighbor (MLMKHNN)," *MEDIA STATISTIKA*, vol. 11, no. 2, pp. 107–117, Dec. 2018, doi: 10.14710/medstat.11.2.107-117.

[14] A. Assegaf, M. A. Mukid, and A. Hoyyi, "Analisis Kesehatan Bank Menggunakan Local Mean K-Nearest Neighbor dan Multi Local Means K-Harmonic Nearest Neighbor," vol. 8, no. 3, pp. 343–355, 2019, [Online]. Available: http://ejournal3.undip.ac.id/index.php/gaussian

[15] M. Siahaan, "Data Mining Strategi Pembangunan Infrastruktur Menggunakan Algoritma K-Means," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 11, no. 3, pp. 316–324, Dec. 2022, doi: 10.32736/sisfokom.v11i3.1453.

[16] W. T. Wu *et al.*, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Military Medical Research*, vol. 8, no. 1. BioMed Central Ltd, Dec. 01, 2021. doi: 10.1186/s40779-021-00338-z.

[17] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," *Clinical eHealth*, vol. 4, pp. 12–23, 2021, doi: 10.1016/j.ceh.2020.11.001.

[18] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4. King Saud bin Abdulaziz University, pp. 1060–1073, Apr. 01, 2022. doi: 10.1016/j.jksuci.2019.06.012.

[19] H. Huang, R. Jia, X. Shi, J. Liang, and J. Dang, "Feature selection and hyper parameters optimization for short-term wind power forecast," *Applied Intelligence*, vol. 51, no. 10, pp. 6752–6770, Oct. 2021, doi: 10.1007/s10489-021-02191-y.

[20] A. Thakkar and R. Lohiya, "Attack classification using feature selection techniques: a comparative study," *J Ambient Intell Humaniz Comput*, vol. 12, no. 1, pp. 1249–1266, Jan. 2021, doi: 10.1007/s12652-020-02167-9.

[21] C. Eiras-Franco, B. Guijarro-Berdiñas, A. Alonso-Betanzos, and A. Bahamonde, "Scalable feature selection using ReliefF aided by locality-sensitive hashing," *International Journal of Intelligent Systems*, vol. 36, no. 11, pp. 6161–6179, Nov. 2021, doi: 10.1002/int.22546.

[22] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Burlington: Morgan Kaufmann, 2016.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and others. "Scikit-learn: Machine learning in Python. Journal of Machine Learning Research", pp. 2825–2830, 2011

[24] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis," in *Procedia Computer Science*, 2021, vol. 191, pp. 487–492. doi: 10.1016/j.procs.2021.07.062.