PERBANDINGAN KINERJA ALGORITMA NAIVE BAYES DAN DECISION TREE DALAM KLASIFIKASI KANKER PARU-PARU

Gifthera Dwilestari [1], Turfa Azmi Afifah [2]
[1] Sistem Informasi, STMIK IKMI Cirebon
[2] Teknik Informatika, STMIK IKMI Cirebon
Jalan Perjuangan 10B – Majasem, Kota Cirebon, Indonesia
ggdwilestari@gmail.com

ABSTRAK

Kanker paru-paru merupakan salah satu penyebab utama kematian di dunia, menjadikan deteksi dini sebagai hal yang sangat penting. Penelitian ini bertujuan untuk membandingkan kinerja algoritma Naive Bayes dan Decision Tree dalam klasifikasi kanker paru-paru. Metode penelitian melibatkan tahapan Knowledge Discovery in Databases (KDD), termasuk pemilihan data, praproses, transformasi, penambangan data, dan interpretasi hasil. Dataset yang digunakan berasal dari repository publik dengan berbagai fitur terkait kondisi pasien. Hasil menunjukkan bahwa Naive Bayes memiliki akurasi lebih tinggi sebesar 92,47% dan unggul dalam mendeteksi kelas positif (recall sebesar 98,77%), tetapi kurang optimal pada kelas negatif. Sementara itu, Decision Tree menunjukkan akurasi 88,17% dengan keseimbangan deteksi antar kelas yang lebih baik (recall kelas negatif sebesar 66,67%). Berdasarkan hasil tersebut, Naive Bayes lebih cocok untuk aplikasi yang membutuhkan deteksi cepat dan efisien, sedangkan Decision Tree lebih sesuai untuk skenario dengan kebutuhan keseimbangan deteksi antar kelas. Penelitian ini memberikan rekomendasi algoritma yang sesuai berdasarkan kebutuhan spesifik aplikasi dan menyarankan pengembangan lebih lanjut untuk optimasi algoritma.

***Kata Kunci**: Kanker Paru-Paru, Naive Bayes, Decision Tree, Klasifikasi, Akurasi, Recall*

1. PENDAHULUAN

Kanker paru-paru adalah salah satu jenis kanker yang paling mematikan di dunia, dengan jumlah kasus yang terus meningkat setiap tahunnya. Menurut data dari World Health Organization (WHO), kanker paru-paru menjadi penyebab utama kematian terkait kanker di seluruh dunia, dengan lebih dari 1,8 juta kematian pada tahun 2020. Tingginya angka kejadian ini disebabkan oleh beberapa faktor, termasuk paparan asap rokok, polusi udara, dan faktor genetik. Selain itu, kanker paru-paru seringkali terdiagnosis pada tahap lanjut karena gejalanya yang tidak spesifik pada tahap awal, sehingga pengobatan menjadi kurang efektif dan tingkat kelangsungan hidup pasien menurun (WHO, 2021).

Klasifikasi kanker paru-paru menjadi tantangan yang signifikan dalam dunia medis karena kompleksitas dan variabilitas dari data pasien. Proses ini tidak hanya melibatkan diagnosis yang tepat, tetapi juga membutuhkan identifikasi karakteristik spesifik dari tumor untuk menentukan subtipe kanker. Kesulitan ini semakin diperparah dengan adanya data yang heterogen dan tidak seimbang, yang dapat mengarah pada kesalahan klasifikasi. Misalnya, beberapa subtipe kanker paru-paru memiliki karakteristik yang tumpang tindih, yang membuat proses klasifikasi menjadi lebih rumit dan membutuhkan pendekatan yang lebih canggih untuk mencapai hasil yang akurat [1]

Penelitian yang dilakukan oleh Perwira Tarigan pada jurnal Multimedia Dan Teknologi Informasi tahun 2024 dengan judul "Analisis Tingkat Akurasi Metode Naïve Bayes Dataset Breast Cancer Wisconsin menjelaskan bahwa Naive Bayes Classifier (NBC) adalah metode penambangan data yang digunakan untuk klasifikasi data dalam suatu kelas. Berdasarkan pandangan literatur dari beberapa penelitian sebelumnya, diketahui bahwa algoritma NBC memiliki tingkat akurasi yang kurang memuaskan. Keakuratan hasil klasifikasi diketahui dipengaruhi oleh fitur-fitur yang digunakan, yang seringkali tidak relevan dan memiliki pengaruh yang rendah terhadap klasifikasi. Dalam penelitian ini, dilakukan analisis tingkat akurasi metode NBC tanpa seleksi fitur dan dengan seleksi fitur pada dataset Wisconsin Breast Cancer. Hasil penelitian menunjukkan bahwa penggunaan metode seleksi fitur dapat meningkatkan akurasi NBC, dengan peningkatan sebesar 0,30% dari akurasi metode NBC tanpa menggunakan seleksi fitur, mencapai tingkat akurasi 97,4% [2]

Penelitian yang dilakukan oleh Tarigan (2024) memberikan wawasan penting tentang peran seleksi fitur dalam meningkatkan akurasi Naive Bayes Classifier. Namun, penelitian ini berfokus pada dataset Wisconsin Breast Cancer, yang berbeda dari konteks penelitian kanker paru-paru yang saya lakukan. GAP yang terlihat adalah kurangnya penerapan seleksi fitur pada dataset kanker paru-paru yang lebih kompleks dan variabel. Sementara Tarigan menunjukkan peningkatan akurasi melalui seleksi fitur, penelitian saya akan mengeksplorasi efektivitas NBC dalam klasifikasi kanker paru-paru dan membandingkannya dengan algoritma lain, yaitu Decision Tree, untuk mengetahui algoritma mana yang memberikan performa terbaik dalam konteks ini. [3]

Untuk menyelesaikan masalah klasifikasi kanker paru-paru, penelitian ini akan menggunakan dua algoritma, yaitu Decision Tree dan Naive Bayes,

dengan pendekatan Knowledge Discovery in Databases (KDD). Proses KDD akan melibatkan tahapan seleksi fitur, praproses data, dan evaluasi model untuk menentukan algoritma yang paling akurat dalam mengklasifikasikan kanker paru-paru. Dengan menggunakan kedua algoritma ini, penelitian ini bertujuan untuk mengidentifikasi pola-pola yang signifikan dalam dataset kanker paru-paru dan menilai efektivitas metode tersebut dalam mengatasi tantangan klasifikasi yang telah diidentifikasi sebelumnya.

Tujuan dari penelitian ini adalah untuk membandingkan kinerja algoritma Naive Bayes dan Decision Tree dalam klasifikasi kanker paru-paru. Melalui analisis ini, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam bidang klasifikasi data medis, khususnya dalam meningkatkan akurasi deteksi dan klasifikasi kanker paru-paru. Hasil penelitian ini juga diharapkan dapat memberikan rekomendasi yang tepat terkait penggunaan algoritma dalam proses diagnosis klinis yang lebih efektif dan efisien.

Implikasi dari penelitian ini diharapkan dapat membantu para praktisi medis dalam memilih algoritma yang paling efektif untuk klasifikasi kanker paru-paru, yang pada gilirannya dapat meningkatkan kualitas diagnosis dan pengobatan bagi pasien. Selain itu, penelitian ini juga dapat menjadi referensi bagi pengembangan sistem pendukung keputusan berbasis AI dalam bidang kesehatan, khususnya untuk penyakit kanker paru-paru, serta menjadi dasar bagi penelitian lanjutan yang bertujuan untuk mengoptimalkan algoritma klasifikasi dalam skenario medis yang lebih kompleks [4].

2. TINJAUAN PUSTAKA

2.1. Metode Literatur Review

Metode literature review yang digunakan dalam penelitian ini adalah dengan cara mereview pada artikel-artikel yang terkait dengan Perbandingan kinerja algoritma naive bayes dan decision tree dalam klasifikasi kanker paru-paru. Langkah-langkah dari literature review meliputi 4 tahapan, yaitu (1) formulasi permasalahan, (2) pencarian literature, (3) evaluasi data, (4) analisis dan interpretasi.

Hasil literature review yang telah dilakukan pada jurnal-jurnal penelitian terkait topik Perbandingan kinerja algoritma naive bayes dan decision tree dalam klasifikasi kanker paru-paru dapat dijabarkan sebagai berikut :

Penelitian pertama yang dilakukan oleh Fika Afiani Ri'fati Rizki, Budhi Irawan, Anggunmeka Luhur Prasasti pada jurnal Seminar Nasional Teknologi Komputer & Sains tahun 2019 dengan judul Deteksi Hand, Foot, and Mouth Disease Menggunakan Metode Klasifikasi Naïve Bayes Berbasis Android menjelaskan bahwa Hand, Foot and Mouth Disease (HMFD) adalah penyakit menular yang disebabkan sekelompok virus dari Enterovirus. Meskipun tergolong penyakit ringan, HMFD juga dapat menyebabkan komplikasi yang berujung kematian jika disebabkan oleh virus enterovirus 71 (EV71). Berdasarkan masalah tersebut dibuat aplikasi system pakar berbasis android dengan metode sistem pakar Naïve Bayes yang dapat mendeteksi gejala HFMD pada citra telapak tangan pengguna serta mengolah gejala yang dirasakan. Aplikasi ini dapat memberikan keluaran berupa informasi terdeteksi dini atau tidaknya penyakit HFMD. Dari hasil pengujian yang telah dilakukan, sistem pakar ini memiliki tingkat akurasi sebesar 95,58 % pada pengujian seluruh dataset, dan 100% pada pengujian partisi data dengan perbandingan data training: data testing sebesar 70%:30%.

Penelitian kedua yang dilakukan oleh Yuce Yuliani dalam Journal of Multidisciplinary Inquiry in Science Technology and Educational Research tahun 2024, dengan judul "Perbandingan Algoritma Klasifikasi untuk Deteksi Intrusi pada Jaringan Komputer (Literature Review)", menyoroti pentingnya Sistem Deteksi Intrusi (IDS) dalam menjaga keamanan jaringan komputer di era digital. Studi ini membandingkan algoritma klasifikasi seperti Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, dan Neural Networks. Hasilnya menunjukkan bahwa Neural Networks dan Random Forest memiliki akurasi tinggi namun membutuhkan sumber daya komputasi besar, sementara Naive Bayes dan Decision Tree lebih unggul dalam hal kecepatan dan efisiensi komputasi. Penelitian ini memberikan panduan bagi peneliti dan praktisi untuk memilih algoritma yang sesuai berdasarkan kebutuhan aplikasi dan karakteristik data. [5]

Penelitian ketiga yang dilakukan oleh Eri Mardiani dalam Journal Digital Transformation Technology (Digitech) tahun 2023, dengan judul "Analisis Kompleksitas Password Dengan Metode KNN, Naïve Bayes, Decision Tree, Ensemble Methods Dan Linear Regression", menyoroti pentingnya keamanan kata sandi dalam melindungi informasi sensitif di era digital. Penelitian ini menggunakan dataset dari Kaggle.com dan aplikasi Orange untuk memvisualisasikan data serta menganalisis kompleksitas password melalui lima metode algoritma: K-Nearest Neighbor (k-NN), Naïve Bayes, Decision Tree, Ensemble Methods, dan Linear Regression. Hasil analisis menunjukkan bahwa Random Forest merupakan algoritma dengan akurasi tertinggi, diikuti oleh Naïve Bayes, ada Boost, Decision Tree, dan k-NN. Namun, model Linear Regression tidak cocok untuk dataset ini. Kesimpulan penelitian ini menggarisbawahi keunggulan Random Forest sebagai algoritma paling akurat untuk memprediksi kompleksitas password. [6]

Penelitian keempat yang dilakukan oleh Azminuddin I. S. Azis dalam Jurnal Rekayasa Sistem dan Teknologi Informasi tahun 2019, berjudul "Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara", menyoroti pentingnya pra-pengolahan data untuk meningkatkan akurasi

prediksi kanker payudara. Dengan menerapkan teknik seperti Missing Value Replacement, Data Transformation, Smoothing Noisy Data, Feature Selection, Data Validation, dan Unbalanced Class Reduction, penelitian ini berhasil mengusulkan pendekatan machine learning yang lebih efisien. Hasilnya meliputi model C4.5 – Z-Score – Genetic Algorithm untuk Breast Cancer Dataset dengan akurasi 77,27%, 7-Nearest Neighbor – Min-Max Normalization – Particle Swarm Optimization untuk Wisconsin Breast Cancer Dataset - Original dengan akurasi 97,85%, Artificial Neural Network – Z-Score – Forward Selection untuk Wisconsin Breast Cancer Dataset – Diagnostic dengan akurasi 98,24%, serta 11-Nearest Neighbor – Min-Max Normalization – Particle Swarm Optimization untuk Wisconsin Breast Cancer Dataset – Prognostic dengan akurasi 83,33%. Pendekatan ini terbukti lebih unggul dibandingkan metode machine learning standar maupun metode terkait sebelumnya, menjadikannya solusi yang efisien untuk prediksi kanker payudara. [7]

Penelitian kelima yang dilakukan oleh Riska Chairunisa pada jurnal Rekayasa Sistem dan Teknologi Informasi tahun 2020 dengan judul Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray menjelaskan bahwa Kanker merupakan salah satu penyakit yang mematikan di dunia dengan tingkat kematian 57,3% pada tahun 2018 di benua Asia. Maka dari itu, diperlukannya diagnosis dini untuk menghindari peningkatan angka kematian yang disebabkan oleh penyakit kanker. Seiring berkembangnya pembelajaran mesin, data gen kanker dapat diolah menggunakan microarray untuk deteksi terjangkitnya penyakit kanker sejak dini. Namun permasalahan yang dimiliki microarray adalah jumlah atribut yang sangat banyak sehingga perlu dilakukan reduksi dimensi. Untuk mengatasi permasalahan tersebut, dalam makalah ini digunakan reduksi dimensi Discrete Wavelet Transform (DWT). Selanjutnya digunakan Classification and Regression Tree (CART) dan Random Forest (RF) sebagai metode klasifikasinya. Tujuan penggunaan kedua metode klasifikasi tersebut untuk mengetahui metode klasifikasi mana yang menghasilkan performa paling baik. Pada penelitian ini digunakan lima data microarray yaitu Colon Tumor, Breast Cancer, Lung Cancer, Prostate Tumor dan Ovarian Cancer dari Kent-Ridge Biomedical Dataset. Akurasi terbaik yang didapat pada penelitian ini untuk data breast cancer sebesar 76,92% dengan CART-DWT, Colon Tumor sebesar 90,1% dengan RF-DWT, lung cancer sebesar 100% dengan RF-DWT, prostate tumor sebesar 95,49% dengan RF-DWT, dan ovarian cancer sebesar 100% dengan RF-DWT. Dari hasil tersebut maka dapat disimpulkan bahwa RF-DWT lebih baik dibandingkan CART-DWT [8]

Penelitian keenam yang dilakukan oleh Anang Susilo dalam Jurnal Sains dan Teknologi tahun 2023, berjudul "Perbandingan Kinerja K-Nearest Neighbors dan Naive Bayes untuk Klasifikasi Perilaku Nasabah pada Pembayaran Kredit Bank", membahas analisis klasifikasi nasabah kredit bank untuk mengidentifikasi potensi pembayaran kredit, baik lancar maupun bermasalah. Penelitian ini menggunakan metode Naive Bayes dan K-Nearest Neighbors (k-NN) untuk membangun model prediktif dari data nasabah yang mencakup kategori kolektibilitas seperti lancar, DPK (Dalam Perhatian Khusus), kurang lancar, diragukan, dan macet. Naive Bayes dipilih karena menghasilkan akurasi tinggi dengan sedikit data pelatihan, sementara k-NN dipilih karena ketahanannya terhadap data noise. Hasil penelitian menunjukkan bahwa Naive Bayes memiliki kinerja lebih baik dengan tingkat akurasi 70%, dibandingkan dengan k-NN yang hanya mencapai akurasi 40%. Kesimpulan ini memberikan wawasan penting bagi bank untuk memilih metode klasifikasi yang lebih efektif dalam memprediksi perilaku pembayaran nasabah. [9]

Penelitian ketujuh yang dilakukan oleh Scholastica Larissa Zefira Lewoema pada jurnal Information Technology Ampera tahun 2024 dengan judul Implementasi Data Mining Pada Klasifikasi Status Gizi Bayi Dengan Metode Decision TreeCHAID menjelaskan bahwa Penelitian ini mengukur akurasi metode Decision Tree CHAID dalam mengklasifikasikan status gizi bayi dengan menambahkan atribut jenis kelamin dan lokasi desa posyandu. Hasil penelitian ini adalah situs web berbasis server lokal untuk menguji sistem klasifikasi tersebut. Prosesnya meliputi impor data, pembagian data latih dan uji, pelatihan model, pemilihan algoritma, dan pengujian matriks. Dari 3106 data antara Januari hingga Februari 2024, akurasi pada data uji mencapai 0,90, pada data latih 0,99, dan akurasi algoritma CHAID 0,84. Variabel yang digunakan meliputi usia, desa, posyandu, tinggi badan, berat badan, dan jenis kelamin. Kelas status gizi meliputi gizi baik, gizi buruk, gizi kurang, gizi berlebih, obesitas, dan risiko gizi berlebih[10]

Penelitian kedelapan yang dilakukan oleh Adiwijaya pada jurnal MEDIA INFORMATIKA BUDIDARM tahun 2018 dengan judul Deteksi Kanker Berdasarkan Klasifikasi Microarray Data menjelaskan bahwa Kanker merupakan salah satu penyakit yang dapat menyebabkan kematian manusia didunia dan menjadi penyebab kematian terbesar setelah penyakit jantung. Karena itu diperlukan suatu teknologi DNA microarray yang digunakan untuk memeriksa bagaimana pola ekspresi gen berubah dalam kondisi yang berbeda, sehingga teknologi tersebut mampu men-deteksi seseorang terkena kanker atau tidak dengan analisis yang akurat. Besarnya

dimensi pada microarray data dapat berpengaruh terhadap analisis ekspresi gen yang digunakan untuk mencari gen informatif, untuk itu diperlukan suatu metode reduksi dimensi dan klasifikasi yang baik sehingga mampu mendapatkan hasil maupun akurasi yang terbaik. Banyak teknik yang dapat diterapkan dalam DNA microarray, salah satunya BPNN Back

Propagation Neural Network sebagai klasifikasi dan PCA sebagai reduksi dimensi, dimana keduanya telah teruji pada beberapa penelitian sebelumnya. Dengan menerapkan BPNN dan PCA pada beberapa jenis data kanker, didapatkan bahwa BPNN dan PCA mendapatkan hasil akurasi lebih dari 80% dengan waktu training time 0-4 detik. [11]

Penelitian kesembilan yang dilakukan oleh Perwira Tarigan pada jurnal Multimedia Dan Teknologi Informasi pada tahun 2024 dengan judul Analisis Tingkat Akurasi Metode Naïve Bayes Dataset Breast Cancer Wisconsin menjelaskan bahwa Naive Bayes Classifier (NBC) adalah metode penambangan data yang digunakan untuk klasifikasi data dalam suatu kelas. Berdasarkan pandangan literatur dari beberapa penelitian sebelumnya, diketahui bahwa algoritma NBC memiliki tingkat akurasi yang buruk dari segi akurasi. Keakuratan hasil klasifikasi diketahui dipengaruhi oleh fiturfitur yang digunakan seringkali tidak relevan dan memilki pengaruh yang rendah terhadap klasifikasi. Pada penelitian ini dilakukan analisis tingkat akurasi metode NBC tanpa seleksi fitur dan dengan seleksi fitur pada dataset Wisconsin Breast Cancer. Dalam penelitian ini memperoleh hasil pengukuran tingkat akurasi metode NBC tanpa seleksi fitur dan dengan gain rasio serta seleksi fitur F-relief pada dataset Breast Cancer Wisconsin sehingga bisa disimpulkan bahwa metode seleksi fitur bisa meningkatkan akurasi pada NBC. Hal ini didukung juga dengan penbisa penelitian sebelumnya. Peningkatan akurasi pada metode klasifikasi NBC dengan menggunakan seleksi fitur Relief-F meningkat sebesar 0,30% dari akurasi metode NBC tanpa menggunakan seleksi fitur dengan tingkat akurasi 97,4%[3]

Penelitian kesepuluh yang dilakukan oleh Agung Mulyo Widodo pada jurnal prosiding sisfotek tahun 2021 dengan judul Performansi K-NN, J48, Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes menejelaskan bahwa Diabetes adalah penyakit kronis yang ditandai dengan ciri-ciri berupa tingginya kadar gula (glukosa) darah. Penyakit ini sering kali ditemukan pada orang dewasa yang sudah lanjut usia, namun penyakit ini juga dapat menyerang orang yang tergolong masih muda. Seiring dengan kemajuan teknologi machine learning sebagai pendukung pembuat keputusan, banyak dibuat model prediksi apakah seseorang dapat diklasifikasi sebagai penderita diabetes atau tidak menderita diabetes dengan menggunakan algoritma-algoritma tertentu. Pada penelitian dilakukan pembuatan model prediksi apakah seseorang terklasifikasi sebagai penderita diabetes atau tidak, berdasarkan paramater/ variabel yaitu berat badan, tinggi badan, kadar kolesterol, gula saat puasa, gula saat tidak puasa, kadar asam urat dan juga jenis kelamin. Model prediksi dibuat dengan menggunakan algoritma-algoritma pengklaisifkasi K-NN, J48 (dengan dasar decision tree), naive bayes dan regresi logistic. Kemudia dilakukan analisis performansi terhadap hasil-hasil testing dari masing-masing algoritma-algoritma tersebut, dan diperoleh bahwa algoritma K-NN menghasilkan model prediksi dengan akurasi yang tertinggi dibandingkan ketiga algoritma yang digunakan dalam penelitian ini. [12]

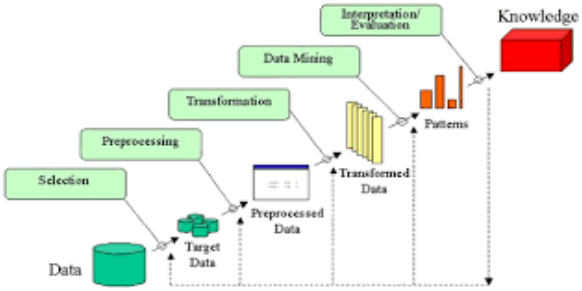
Tabel 1. Ringkasan Hasil Literature View

Judul Penelitian	Nama Penelitian	Masalah	Metode
Deteksi Hand, Foot, and Mouth Disease Menggunakan Metode Klasifikasi Naïve Bayes Berbasis Android	Fika Afiani Ri'fati Rizki, Budhi Irawan, Anggunmeka Luhur Prasasti	Deteksi gejala Hand, Foot, and Mouth Disease (HFMD) pada citra telapak tangan pengguna serta mengolah gejala yang dirasakan.	Naïve Bayes berbasis Android
Perbandingan Algoritma Klasifikasi untuk Deteksi Intrusi pada Jaringan Komputer (Literature Review)	Yuce Yuliani	Keamanan jaringan komputer dan perbandingan algoritma klasifikasi untuk deteksi intrusi.	Perbandingan berbagai algoritma klasifikasi seperti Decision Tree, Random Forest, SVM, KNN, Naive Bayes, dan Neural Networks
Analisis Kompleksitas Password Dengan Metode KNN, Naïve Bayes, Decision Tree, Ensemble Methods Dan Linear Regression	Eri Mardiani	Analisis prediksi kompleksitas password dengan metode-metode algoritma yang berbeda.	K-Nearest Neighbor, Naïve Bayes, Decision Tree, Ensemble Methods, Linear Regression
Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara	Azminuddin I. S. Azis	Prediksi Kanker Payudara menggunakan pendekatan machine learning yang efisien.	C4.5 – Z-Score – Genetic Algorithm, 7-Nearest Neighbor – Min-Max Normalization – Particle Swarm Optimization, Artificial Neural Network – Z-Score – Forward Selection, 11-Nearest Neighbor – Min-Max Normalization – Particle Swarm Optimization
Perbandingan CART dan Random Forest untuk	Riska Chairunisa	Diagnosis dini kanker menggunakan data gen	Classification and Regression Tree (CART) dan Random Forest (RF)

Judul Penelitian	Nama Penelitian	Masalah	Metode
Deteksi Kanker berbasis Klasifikasi Data Microarray		kanker yang diolah dengan microarray.	dengan reduksi dimensi Discrete Wavelet Transform (DWT)
Perbandingan Kinerja K-Nearest Neighbors dan Naive Bayes Untuk Klasifikasi Perilaku Nasabah Pada Pembayaran Kredit Bank	Anang Susilo	Klasifikasi perilaku nasabah pada pembayaran kredit bank.	K-Nearest Neighbors dan Naive Bayes
Implementasi Data Mining Pada Klasifikasi Status Gizi Bayi Dengan Metode Decision TreeCHAID	Scholastica Larissa Zefira Lewoema	Klasifikasi status gizi bayi berdasarkan variabel seperti usia, tinggi badan, berat badan, jenis kelamin, dan lainnya.	Decision Tree CHAID
Deteksi Kanker Berdasarkan Klasifikasi Microarray Data	Adiwijaya	Deteksi kanker berdasarkan klasifikasi data microarray.	Back Propagation Neural Network (BPNN) dan PCA
Analisis Tingkat Akurasi Metode Naïve Bayes Dataset Breast Cancer Wisconsin	Perwira Tarigan	Analisis tingkat akurasi metode Naïve Bayes dengan dan tanpa seleksi fitur pada dataset Breast Cancer Wisconsin.	Naïve Bayes dengan dan tanpa seleksi fitur
Performansi K-NN, J48, Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes	Agung Mulyo Widodo	Prediksi klasifikasi diabetes menggunakan algoritma K-NN, J48, Naive Bayes, dan regresi logistik.	K-NN, J48, Naive Bayes, Regresi Logistik

3. METODE PENELITIAN

Metode penelitian yang digunakan dalam studi ini dirancang untuk secara sistematis mengeksplorasi dan menganalisis kinerja dua algoritma klasifikasi, yaitu Naive Bayes dan Decision Tree, dalam konteks klasifikasi kanker paru-paru. Penelitian ini mengikuti pendekatan Knowledge Discovery in Databases (KDD), yang mencakup tahapan-tahapan kritis mulai dari pemilihan data, praproses data, transformasi, penambangan data, hingga interpretasi hasil. Berikut ini alur penelitian termuat dalam gambar dibawah ini :



Gambar 1 Tahapan Metode penelitian

Tabel 2 Deskripsi Aktivitas Metode Penelitian

Tahapan	Aktivitas	Deskripsi Aktivitas
Selection	Pemilihan data dari sumber	Pada tahap ini, data yang relevan untuk penelitian dipilih dari sumber yang tersedia. Dalam konteks penelitian ini, data yang digunakan adalah dataset kanker paru-paru. Tahap ini juga melibatkan pemilihan fitur atau atribut yang dianggap penting untuk analisis lebih lanjut, yang akan digunakan dalam proses seleksi fitur dan pengembangan model.
Transformation	Melakukan transformasi data	Pada tahap ini, data yang telah dipraproses akan diubah atau ditransformasikan ke dalam format yang lebih sesuai untuk analisis. Transformasi bisa melibatkan reduksi dimensi atau seleksi fitur untuk mengidentifikasi atribut yang paling berpengaruh terhadap klasifikasi. Dalam penelitian ini, fitur yang relevan akan dipilih untuk meningkatkan akurasi model klasifikasi.
Data Mining	Mmebuat model decision tree dan naïve bayes	Tahap ini adalah inti dari KDD, di mana algoritma data mining diterapkan untuk menemukan pola atau model yang berguna. Dalam penelitian ini, dua algoritma klasifikasi, yaitu Naive Bayes dan Decision Tree, akan diterapkan pada data yang telah diproses dan ditransformasikan. Model-model ini akan dilatih dan diuji untuk menghasilkan prediksi klasifikasi kanker paru-paru.
Interpretation	Melakukan analisa hasil	Pada tahap terakhir ini, hasil dari model yang telah dikembangkan dievaluasi dan diinterpretasikan. Evaluasi dilakukan dengan menggunakan metrik-metrik seperti akurasi, precision, recall, dan F1-score. Peneliti kemudian akan menafsirkan hasil ini untuk menentukan keefektifan masing-masing algoritma dalam klasifikasi kanker paru-paru. Hasil ini juga akan dibandingkan dengan temuan dari penelitian lain untuk mendapatkan wawasan lebih dalam tentang performa algoritma yang diuji.

Sumber data yang digunakan dalam penelitian ini berasal dari dataset kanker paru-paru yang diperoleh dari repository publik atau institusi medis yang terpercaya. Dataset ini berisi berbagai informasi terkait pasien kanker paru-paru, termasuk karakteristik demografis, hasil pemeriksaan klinis, serta detail mengenai kondisi tumor. Data ini mencakup berbagai fitur yang relevan, seperti usia pasien, riwayat merokok, hasil radiologi, dan hasil biopsi. Adapun link sumber data yaitu https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer?resource=download.

Data yang digunakan dalam penelitian ini diperoleh dari dataset sekunder yang sudah tersedia di repository publik atau dari institusi medis terpercaya yang memiliki koleksi data terkait kanker paru-paru. Dataset ini biasanya sudah dipublikasikan oleh lembaga penelitian atau rumah sakit dan digunakan secara luas dalam studi-studi medis. Pengumpulan data dilakukan dengan mengunduh dataset dari sumber yang telah diverifikasi keasliannya dan diakui dalam komunitas akademik.

Kinerja model dievaluasi menggunakan metrik-metrik seperti akurasi (accuracy), presisi (precision), recall, dan F1-score. Metrik-metrik ini membantu dalam memahami seberapa baik model dalam mengklasifikasikan data dengan benar.

Validasi Silang (Cross-Validation) Validasi silang dapat digunakan untuk memastikan bahwa model tidak overfitting dan mampu memberikan hasil yang konsisten ketika diterapkan pada data baru. Proses ini melibatkan pembagian dataset menjadi beberapa subset dan melatih serta menguji model pada subset yang berbeda-beda.

4. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk membandingkan kinerja algoritma Naive Bayes dan Decision Tree dalam klasifikasi penyakit kanker paru-paru. Proses klasifikasi dilakukan dengan menggunakan dataset kanker paru-paru yang berisi berbagai fitur seperti usia, kebiasaan merokok, konsumsi alkohol, dan gejala medis lainnya. Metode evaluasi yang digunakan meliputi akurasi, precision, recall, dan F1-score, untuk menentukan algoritma mana yang lebih efektif dalam melakukan klasifikasi berdasarkan hasil prediksi dan hasil sebenarnya.

Dari hasil evaluasi, algoritma Naive Bayes menunjukkan keunggulan dalam hal akurasi, dengan nilai 92,47%. Algoritma ini bekerja dengan menghitung probabilitas bersyarat dari setiap fitur dan mengasumsikan bahwa setiap fitur independen satu sama lain. Keunggulan utama Naive Bayes adalah kemampuannya untuk bekerja cepat dan efisien pada data besar, serta memberikan hasil prediksi yang baik pada kelas YES. Precision untuk kelas YES mencapai 93,02% dan recall mencapai 98,77%, yang menunjukkan bahwa model ini sangat andal dalam mendeteksi kasus positif (YES). Namun, kinerja Naive Bayes pada kelas NO tidak sebaik pada kelas YES. Recall pada kelas NO hanya 50%, menandakan bahwa model ini gagal mengidentifikasi separuh dari data NO dengan benar, sehingga meningkatkan risiko kesalahan dalam mendeteksi kasus negatif.

Sementara itu, algoritma Decision Tree menunjukkan performa yang sedikit lebih rendah dalam hal akurasi, yaitu 88,17%. Meskipun demikian, Decision Tree memiliki kekuatan dalam mendeteksi pola kompleks antar fitur, membuatnya cocok untuk data yang tidak sepenuhnya independen. Precision pada kelas YES mencapai 94,87%, sedikit lebih tinggi dibandingkan Naive Bayes. Selain itu, recall pada kelas NO sebesar 66,67% menunjukkan bahwa Decision Tree lebih baik dalam mendeteksi data negatif (NO) dibandingkan Naive Bayes. Struktur pohon yang dihasilkan juga memberikan interpretasi yang mudah dipahami oleh pengguna, sehingga algoritma ini lebih unggul dalam hal transparansi dan penjelasan hasil.

Dari hasil perbandingan, terlihat bahwa Naive Bayes memiliki akurasi lebih tinggi secara keseluruhan, namun kinerjanya pada kelas NO tidak sebaik Decision Tree. Decision Tree memberikan keseimbangan yang lebih baik antara precision dan recall, terutama pada kelas NO, meskipun akurasinya sedikit lebih rendah. Dalam konteks klasifikasi kanker paru-paru, yang mengharuskan deteksi akurat baik pada kasus positif maupun negatif, Decision Tree mungkin menjadi pilihan yang lebih baik. Hal ini disebabkan oleh kemampuannya untuk menangani interaksi kompleks antar fitur dan memberikan hasil yang lebih konsisten pada kedua kelas.

Namun, jika tujuan utama penelitian adalah memastikan deteksi cepat dan akurat terhadap kasus positif (YES), Naive Bayes dapat menjadi pilihan yang lebih tepat. Naive Bayes juga lebih unggul dalam efisiensi komputasi, sehingga dapat diterapkan dengan baik pada data skala besar atau dalam sistem yang membutuhkan prediksi waktu nyata (real-time).

Secara keseluruhan, Decision Tree lebih baik dalam menangani imbalance data dan mendeteksi kelas NO, sedangkan Naive Bayes lebih unggul dalam efisiensi dan akurasi keseluruhan, terutama pada data dengan fitur yang bersifat independen. Pemilihan algoritma terbaik sangat bergantung pada konteks dan kebutuhan spesifik aplikasi. Dalam situasi di mana kesalahan deteksi negatif dapat berdampak serius, seperti pada diagnosis awal penyakit, Decision Tree bisa menjadi solusi yang lebih tepat. Namun, untuk aplikasi yang mengutamakan kecepatan dan akurasi umum, Naive Bayes mungkin menjadi pilihan yang lebih efisien.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, perbandingan kinerja antara algoritma Naive Bayes dan Decision Tree dalam klasifikasi kanker paru-paru memberikan wawasan penting mengenai kelebihan dan kekurangan masing-masing metode. Naive Bayes menunjukkan akurasi keseluruhan sebesar 92,47%, dengan

keunggulan dalam mendeteksi kelas YES, dibuktikan dengan precision sebesar 93,02% dan recall mencapai 98,77%, sehingga andal dalam mendeteksi pasien yang benar-benar terdiagnosis kanker paru-paru. Namun, algoritma ini memiliki kelemahan dalam mendeteksi kelas NO, dengan recall hanya sebesar 50%, yang berarti separuh pasien non-kanker tidak terdeteksi dengan benar. Sebaliknya, Decision Tree, meskipun akurasinya sedikit lebih rendah sebesar 88,17%, lebih baik dalam mendeteksi kelas NO, dengan recall mencapai 66,67%, serta precision untuk kelas YES yang tinggi, yaitu 94,87%, memberikan keseimbangan lebih baik antara deteksi kelas positif dan negatif.

Selain itu, Decision Tree menawarkan keunggulan dalam interpretasi melalui struktur pohon keputusan yang lebih mudah dipahami dan digunakan. Hasil penelitian ini menunjukkan bahwa pemilihan algoritma terbaik bergantung pada tujuan dan konteks penerapannya. Naive Bayes lebih cocok untuk efisiensi komputasi dan deteksi cepat, seperti dalam sistem real-time, sedangkan Decision Tree lebih sesuai untuk skenario yang memerlukan ketelitian tinggi dalam mendeteksi kasus negatif, seperti diagnosis awal penyakit. Kombinasi atau optimasi kedua algoritma ini juga dapat dipertimbangkan untuk menghasilkan model klasifikasi yang lebih akurat dan andal dalam aplikasi medis yang kompleks seperti deteksi kanker paru-paru. Untuk pengembangan lebih lanjut, tuning parameter pada kedua algoritma dapat dilakukan, seperti pemilihan distribusi yang tepat pada Naive Bayes atau penerapan metode seperti pruning atau ensemble methods pada Decision Tree, guna meningkatkan akurasi dan mengurangi kesalahan klasifikasi.

DAFTAR PUSTAKA

[1] "Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, S., Chairunisa, R., & Astuti, W. (2017a). Terakreditasi SINTA Peringkat 2 Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray. Masa Berlaku Mulai, 1(3), 805–812."

[2] "Tarigan, P., & Prabowo, A. (2024a). Analisis Tingkat Akurasi Metode Naïve Bayes Dataset Breast Cancer Wisconsin-Perwira Analisis Tingkat Akurasi Metode Naïve Bayes Dataset Breast Cancer Wisconsin. 06(02). https://doi.org/10.54209/jatilima.v6i02.505."

[3] "Tarigan, P., & Prabowo, A. (2024b). Analisis Tingkat Akurasi Metode Naïve Bayes Dataset Breast Cancer Wisconsin-Perwira Analisis Tingkat Akurasi Metode Naïve Bayes Dataset Breast Cancer Wisconsin. 06(02). https://doi.org/10.54209/jatilima.v6i02.505."

[4] "Mulyo Widodo, A., Salsabila Anggraeni, Y., Anwar, N., Ichwani, A., & Anggara Sekti, B. (n.d.-a). Performansi K-NN, J48, Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes."

[5] "Susilo Yuda Irawan, A., Heryana, N., Siti Hopipah, H., Rahma Putri, D., & Hs Ronggo Waluyo Puseurjaya Telukjambe Timur Karawang Jawa Barat, J. (2021). Identifikasi Website Phishing dengan Perbandingan Algoritma Klasifikasi. In Syntax: Jurnal Informatika (Vol. 10, Issue 01). www.phishtank.com."

[6] "Mardiani, E., Rahmansyah, N., Wijaya, Y. F., Fitri, A. A., Mustafa, R., Rizki, M. R., & Pramesti, K. M. (2024). Analisis Kompleksitas Password Dengan Metode KNN, Naïve Bayes, Decision Tree, Ensemble Methods Dan Linear Regression. Digital Transformation Technology, 3(2), 955–966. https://doi.org/10.47709/digitech.v3i2.3513."

[7] "Dirjen, S. K., Riset, P., Pengembangan, D., Dikti, R., Azis, A. I. S., Surya, I., Idris, K., Santoso, B., Mustofa, Y. A., & Informatika, J. T. (2017). Terakreditasi SINTA Peringkat 2 Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara. Masa Berlaku Mulai, 1(3), 458–469."

[8] "Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, S., Chairunisa, R., & Astuti, W. (2017b). Terakreditasi SINTA Peringkat 2 Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray. Masa Berlaku Mulai, 1(3), 805–812."

[9] "Susilo, A. (n.d.). Perbandingan Kinerja K-Nearest Neighbors dan Naive Bayes Untuk Klasifikasi Perilaku Nasabah Pada Pembayaran Kredit Bank. 3(2), 2023."

[10] "Setyo, W. N., & Wardhana, S. (2019). Implementasi Data Mining Pada Penjualan Produk Di Cv Cahaya Setya Menggunakan Algoritma Fp-Growth. Petir, 12(1), 54–63. https://doi.org/10.33322/petir.v12i1.416."

[11] "Basuki, T. L., Jondri, M. S., & Wisesty, U. N. (2020). Deteksi Polycystic Ovarian Syndrome (PCOS) Menggunakan Klasifikasi Microarray Data dengan Algoritma Artificial Neural Network (ANN) Backpropagation dan Principal Component Analysis. E-Proceeding of Engineering."

[12] "Mulyo Widodo, A., Salsabila Anggraeni, Y., Anwar, N., Ichwani, A., & Anggara Sekti, B. (n.d.-b). Performansi K-NN, J48, Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes."