Optimasi Algoritma K-Nearest Neighbor (KNN) dengan Normalisasi dan Seleksi Fitur Dalam Klasifikasi Komplikasi Infark Miokard

Muhammad Rizki Nanda Pratama Fakultas Ilmu Komputer Universitas Sriwijaya Sumatera Selatan, Indonesia

Annisa Darmawahyuni* Intelligent System Research Group Fakultas Ilmu Komputer Universitas Sriwijaya Sumatera Selatan, Indonesia riset.annisadarmawahyuni@gmail.com

Rossi Passarella Fakultas Ilmu Komputer Universitas Sriwijaya Sumatera Selatan, Indonesia

***Abstract*—** Infark miokard merupakan salah satu penyebab utama kematian akibat penyakit kardiovaskular yang sering disertai komplikasi fatal pada fase akut. Penelitian ini mengusulkan optimasi algoritma *K-Nearest Neighbor* (KNN) melalui penerapan normalisasi *Z-Score* dan teknik seleksi fitur untuk mengklasifikasikan jenis komplikasi yang terjadi. Dataset yang digunakan adalah *Myocardial Infarction Complications* dari UCI Machine Learning Repository. Tiga metode seleksi fitur dibandingkan, yaitu *Information Gain*, *Gain Ratio*, dan *Symmetrical Uncertainty*. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* pada beberapa nilai parameter K (3, 5, 7, 9). Hasil menunjukkan bahwa kombinasi KNN dengan seleksi fitur *Information Gain* dan normalisasi menghasilkan performa terbaik dengan akurasi mencapai 93%. Hasil ini membuktikan efektivitas teknik optimasi tersebut dalam meningkatkan akurasi klasifikasi medis multikelas.

Keywords*—K-Nearest Neighbor, Seleksi Fitur, Normalisasi, Infark Miokard, Machine Learning*

I. LATAR BELAKANG

Infark miokard, atau yang lebih dikenal sebagai serangan jantung, terjadi akibat penyumbatan aliran darah ke otot jantung, yang umumnya disebabkan oleh aterosklerosis. Penyumbatan ini dapat menyebabkan kerusakan atau kematian jaringan jantung yang berdampak pada fungsi jantung secara keseluruhan. Komplikasi infark miokard meliputi gagal jantung, aritmia, emboli, dan bahkan kematian mendadak [1]. Secara global, WHO memperkirakan angka kematian akibat penyakit kardiovaskular akan meningkat dari 17 juta pada 2008 menjadi 23,4 juta pada 2030, dengan lebih dari 80% kasus terjadi di negara berkembang [2].

Di Indonesia, penyakit kardiovaskular, termasuk infark miokard, menjadi penyebab kematian utama, dengan prevalensi mencapai 1,5% dari populasi [1]. Faktor risiko utama meliputi hipertensi, diabetes melitus, hiperkolesterolemia, obesitas, dan kebiasaan merokok. Sebuah penelitian menunjukkan bahwa hipertensi ditemukan pada 93,5% penderita infark miokard, obesitas pada 72,6%, kolesterol tinggi pada 69,4%, diabetes melitus pada 56,5%, dan kebiasaan merokok pada 54,8% pasien [1]. Tingginya risiko ini juga dikaitkan dengan gaya hidup yang tidak sehat, seperti pola makan tinggi lemak dan rendah aktivitas fisik [1]. Penyakit ini memiliki tingkat keparahan tinggi karena dapat berkembang menjadi kondisi kritis seperti syok kardiogenik atau peradangan sistemik yang luas (systemic inflammatory response syndrome/SIRS). Pada infark miokard akut dengan elevasi segmen ST (IMA-EST), oklusi total pada arteri koroner dapat menyebabkan iskemia berat dan kerusakan luas pada miokardium [3].

Di Kota Palembang, Sumatera Selatan, angka kejadian penyakit kardiovaskular, termasuk infark miokard, terus menunjukkan peningkatan. Berdasarkan data Riskesdas 2013, Sumatera Selatan termasuk salah satu provinsi dengan prevalensi penyakit jantung koroner (PJK) tertinggi di Indonesia, dengan angka kasus terdiagnosis mencapai 0,7% berdasarkan gejala dan diagnosis dokter [4]. Tingginya angka kejadian penyakit kardiovaskular di kota ini dipengaruhi oleh faktor risiko dominan seperti hipertensi, dislipidemia, diabetes melitus, obesitas, dan kebiasaan merokok. Selain itu, gaya hidup masyarakat yang cenderung rendah aktivitas fisik dan tingginya konsumsi makanan tinggi lemak juga turut memperburuk situasi [4].

Di era modern, perkembangan teknologi berbasis data menawarkan solusi baru untuk mengatasi tantangan ini. Salah satu metode yang menonjol dalam bidang ini adalah algoritma K-Nearest Neighbor (KNN). KNN merupakan algoritma pembelajaran mesin non-parametrik yang bekerja dengan mengidentifikasi sejumlah titik data terdekat berdasarkan metrik jarak tertentu, seperti *Euclidean distance.* Algoritma ini mengklasifikasikan data uji berdasarkan kategori mayoritas dari data tetangga terdekatnya. Sifatnya yang sederhana namun fleksibel membuat KNN sangat cocok untuk digunakan dalam analisis data medis, termasuk deteksi dan prediksi penyakit kardiovaskular. Kelebihan utama KNN adalah kemampuannya menangani data multidimensi tanpa memerlukan asumsi distribusi tertentu. Dalam konteks penyakit jantung, algoritma ini dapat digunakan untuk menganalisis parameter kesehatan seperti tekanan darah, kadar kolesterol, dan indeks massa tubuh untuk mengklasifikasikan risiko pasien, serta memprediksi kemungkinan komplikasi yang dapat terjadi [5].

Dalam penelitian terbaru, KNN telah berhasil diterapkan untuk memetakan tingkat risiko penyebaran COVID-19 di beberapa wilayah di Jawa Tengah, menunjukkan potensi besar algoritma ini dalam menangani data epidemiologi [6]. Seleksi fitur yang tepat dapat berpengaruh signifikan terhadap performa KNN. Seleksi fitur merupakan teknik dalam pembelajaran mesin yang bertujuan untuk mengidentifikasi atribut-atribut yang relevan dan menghilangkan atribut yang tidak relevan, sehingga meningkatkan efisiensi dan akurasi model. Teknik seperti *Information Gain*, *Gain Ratio*, dan *Symmetrical Uncertainty* dapat digunakan untuk mengidentifikasi fitur yang signifikan [7].

Kombinasi KNN dengan seleksi fitur memungkinkan algoritma untuk fokus pada atribut yang penting, mengurangi *overfitting*, dan mempercepat proses klasifikasi. Dalam konteks klasifikasi komplikasi infark miokard, optimasi ini

memungkinkan deteksi dini berdasarkan faktor risiko utama seperti tekanan darah, kadar kolesterol, dan riwayat medis.

II. METODOLOGI PENELITIAN

A. Dataset dan Sumber Data

Dataset yang digunakan dalam penelitian ini adalah *Myocardial Infarction Complications Dataset dari UCI Machine Learning Repository.* Dataset ini berisi rekam medis pasien infark miokard dan dikumpulkan dari Rumah Sakit Kardiologi di Rusia. Dataset ini memiliki 1700 sampel dengn fitur klinis yang mencakup karakteristik pasien, faktor risiko, dan komplikasi yang muncul setelah serangan jantung.

Secara keseluruhan, dataset ini terdiri dari 111 variabel independen (features) yang mencakup berbagai aspek karakteristik pasien serta 12 variabel dependen (target) yang dapat digunakan dalam model prediktif berbasis machine learning. Variabel independen mencakup 111 fitur, yaitu diantaranya :

- Data Demografis: Usia, jenis kelamin, ID pasien
- Faktor Risiko: Riwayat hipertensi, diabetes, merokok, dan kadar kolesterol.
- Parameter Hemodinamik: Tekanan darah, denyut jantung, fraksi ejeksi jantung.
- Parameter Biokimia Darah : Kadar natrium, kalium, ALT, leukosit, dan kreatinin
- Komplikasi Infark Miokard: Gagal jantung, aritmia, emboli, dan kematian pasien.

Variabel target yang digunakan dalam analisis ini adalah lethal outcome akibat komplikasi infark miokard, yang dikategorikan ke dalam delapan kelas, yaitu:

- Kategori 0: Status pasien tidak diketahui (pasien masih hidup).
- Kategori 1: Kematian akibat syok kardiogenik.
- Kategori 2: Kematian akibat edema paru.
- Kategori 3: Kematian akibat ruptur miokard.
- Kategori 4: Kematian akibat perkembangan gagal jantung kongestif.
- Kategori 5: Kematian akibat tromboemboli.
- Kategori 6: Kematian akibat asistol.
- Kategori 7: Kematian akibat fibrilasi ventrikel.

B. K-Nearest Neighbors

K-Nearest Neighbors (KNN) adalah metode klasifikasi yang menggunakan "suara mayoritas" dari k tetangga terdekat untuk menentukan kelas suatu objek. Kedekatan antar data diukur berdasarkan jarak tertentu. Algoritma KNN populer karena kesederhanaannya, fleksibilitasnya, dan efektivitasnya dalam berbagai penerapan. KNN dapat digunakan baik variabel diskrit maupun kontinu. Namun kelemahan dari metode ini yaitu perhitungan dan pengurutan jarak secara iteratif yang sangat besar secara komputasi, sehingga tidak sesuai untuk kumpulan data yang besar [8].

Overfitting dan sensitivitas tinggi terhadap noise dapat terjadi apabila nilai k terlalu kecil. Namun nilai k yang terlalu besar juga dapat menyebabkan bias yang lebih tinggi dan akurasi yang lebih rendah karena mungkin termasuk sampel yang bukan tetangga sebenarnya [9]. Dalam menghitung jarak antara objek dengan data pada metode KNN, dapat digunakan rumus jarak Euclidean pada Persamaan (1) berikut.

$$d = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

dengan :

d = jarak antara data dengan objek pelatihan
p_i = fitur ke-i dari objek
q_i = fitur ke-i data
n = jumlah fitur yang digunakan

Nilai parameter K yang digunakan dalam penelitian ini ditentukan melalui pendekatan empiris berdasarkan studi terdahulu [5][7] yang menunjukkan bahwa nilai kecil K (3–9) memberikan keseimbangan terbaik antara bias dan variansi pada dataset medis berukuran menengah. Oleh karena itu, penelitian ini menguji empat nilai K, yaitu **K = 3, 5, 7, dan 9**, untuk mengidentifikasi titik optimal akurasi klasifikasi komplikasi.

C. Seleksi Fitur

Seleksi fitur adalah proses penting dalam *machine learning* yang bertujuan untuk mengurangi jumlah fitur yang tidak relevan dari dataset. Seleksi fitur dapat meningkatkan akurasi model dan efisiensi pemrosesan data dengan mengidentifikasi dan memilih fitur-fitur yang paling signifikan. *Information Gain* merupakan teknik *unsupervised learning* dengan metode filter. Seleksi fitur ini biasanya digunakan untuk meranking atribut mana yang dianggap paling berpengaruh terhadap kelasnya. *Information Gain* menghitung seberapa banyak informasi relevan yang disediakan oleh fitur mengenai kelas atau label pada suatu dataset. Entropi sendiri ialah sebuah metrik pengukuran yang menentukan ketidakteraturan dalam suatu data dengan mengukur ketidakmurnian pada fitur tertentu [10]. Rumus perhitungan *Information Gain* ditujukan pada Persamaan (2) sebagai berikut [11].

$$Entropy(S) = \sum_{i=1}^{c} -p_i \cdot log_2 p_i$$
$$Gain(S,A) = Entropy(S) - \sum_{(A)} \frac{|s_v|}{|s|} Entropy(S_v) \quad (2)$$

dengan :

c = Akumulasi nilai dari kelas klasifikasi
p_i = Probabilitas kemunculan kelas ke-i
v = kemungkinan nilai fitur A
A = fitur
$Gain(S,A)$ = Nilai gain dari fitur
(A) = kemungkinan nilai himpunan A
S_v = jumlah contoh nilai dari v
S = jumlah keseluruhan sampel data

Gain Ratio merupakan pengembangan dari *Information Gain* yang mengurangi tingkat kebiasan. Metode ini mempertimbangkan jumlah dan ukuran cabang saat memilih parameter, digunakan untuk meningkatkan akurasi prediksi. Metode ini meningkatkan akses ke informasi dengan memasukkan informasi internal (distribusi entropi), dalam perhitungannya [12]. Rumus yang digunakan ditujukan pada

Persamaan (3) sebagai berikut [13].

$$SplitInfo(S,A) = -\sum_{j=1}^{k} \frac{s_j}{s} \times log_2 \frac{s_j}{s} \quad (3)$$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}$$

dengan :

$Gain(S,A)$ = Nilai *gain* dari fitur

$SplitInfo(S,A)$ = Informasi intrinsik setelah dibagi oleh atribut A

s_j = jumlah sampel pada cabang ke-j

s = jumlah total sampel dalam himpunan data S

Symmetrical Uncertainty (SU) digunakan untuk mengukur korelasi antar fitur, dengan nilai berkisar antara 0 sampai 1. SU melihat seberapa berpengaruh suatu variabel terhadap kelas label. Semakin tinggi nilai SU suatu variabel, maka semakin berpengaruh variabel tersebut terhadap kelas label, dan sebaliknya. Rumus yang digunakan ditujukan pada Persamaan (4) sebagai berikut [14].

$$SU = 2 \times \frac{Gain(S,A)}{Entropi(S) + Entropi(A)} \quad (4)$$

dengan :

$Gain(S,A)$ = Nilai *gain* dari fitur

$Entropi(S)$ = Entropi dari kelas target S

$Entropi(A)$ = Entropi dari kelas target A

D. Confusion Matrix dan Metrik Evaluasi

Confusion matrix adalah metode yang lebih efisien dalam menilai kinerja suatu model klasifikasi. Matriks ini berbentuk tabel yang menunjukkan jumlah data uji yang dikategorikan dengan benar maupun salah, sehingga mempermudah proses evaluasi akurasi sistem klasifikasi [15]. Dalam evaluasi kinerja model klasifikasi menggunakan *confusion matrix*, terdapat empat istilah utama yang merepresentasikan hasil klasifikasi, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).

True Negative mengacu pada jumlah data negatif yang diklasifikasikan dengan benar, sedangkan *False Positive* terjadi ketika data negatif salah diklasifikasikan sebagai positif. Sementara itu, *True Positive* menunjukkan jumlah data positif yang teridentifikasi dengan benar, dan *False Negative* merupakan kebalikan dari *True Positive*, yaitu kondisi di mana data positif diklasifikasikan secara keliru sebagai negatif. Berdasarkan nilai-nilai tersebut, maka dapat dilakukan perhitungan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* dalam model klasifikasi [16]. Metrik evaluasi yang digunakan dalam menentukan kebaikan model adalah akurasi, presisi, *recall,* dan F1-*score* yang masing – masing persamaannya ditujukan dari Persamaan (5) sampai Persamaan (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

E. Normalisasi Z-Score

Normalisasi *Z-Score* adalah teknik transformasi data numerik yang digunakan untuk menyetarakan skala fitur sebelum digunakan dalam algoritma berbasis jarak, seperti K-Nearest Neighbor (KNN). Metode ini mengonversi setiap nilai data menjadi skor standar dengan cara mengurangi rata- rata [17]. Persamaan (9) menunjukkan perhitungan normalisasi dengan metode *Z-score*.

$$Z - score = \frac{x - \mu}{\sigma} \quad (9)$$

dengan :

x = Nilai dari fitur

μ = Nilai rata – rata dari fitur

σ = Simpangan baku (standar deviasi) dari fitur

F. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE merupakan teknik *oversampling* berbasis interpolasi yang dirancang untuk mengatasi ketidakseimbangan kelas dalam dataset, terutama pada klasifikasi multikelas atau biner dengan distribusi label yang timpang. Metode ini bekerja dengan membuat data sintetis baru dari kelas minoritas, bukan sekadar menduplikasi data yang ada. SMOTE memilih titik acak dalam kelas minoritas dan menciptakan sampel baru dengan cara menginterpolasi antara titik tersebut dan tetangganya dalam ruang fitur, sehingga menghasilkan distribusi yang lebih seimbang dan mengurangi *overfitting* [18]. Persamaan SMOTE ditujukan pada Persamaan (10) sebagai berikut.

$$x_s = x_i + \delta \cdot (x_{zi} - x_i) \quad (10)$$

dengan :

x_i = Sampel dari kelas minoritas

x_{zi} = Tetangga terdekat dari x_i dalam kelas minoritas

δ = Bilangan acak dari 0 sampai 1.

III. HASIL ANALISIS

A. Tahap Pre-Processing

Tujuan utama dari proses ini adalah untuk memastikan bahwa data yang dimasukkan ke dalam model telah bersih, representatif, dan dalam skala yang seragam. Dalam penelitian ini, proses *pre-processing* terdiri dari tiga tahapan utama, yaitu imputasi nilai hilang (*missing values*), penyeimbangan distribusi kelas target dengan algoritma SMOTE, dan normalisasi data menggunakan pendekatan Z- *Score*.

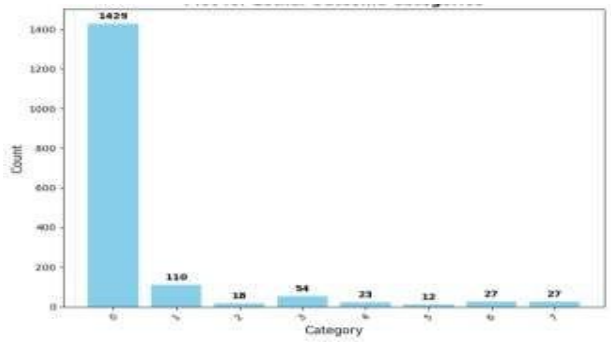
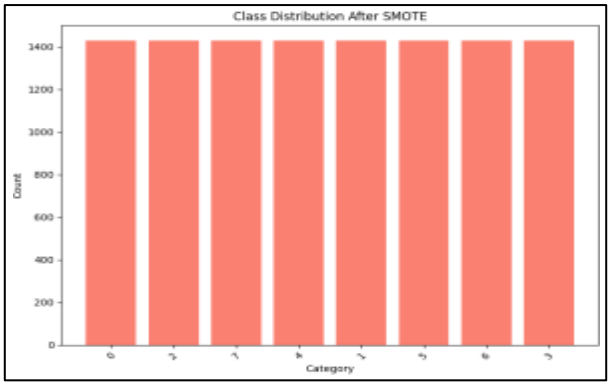


Fig. 1. Distribusi Kelas pada *Outcome* Infark Miokard

Pada Fig. 1. terlihat bahwa adanya ketidakseimbangan dalam kelas *outcome* infark miokard. Distribusi label menunjukkan bahwa mayoritas pasien tidak mengalami kematian (kelas 0 mencakup sekitar 84,06% kasus) dan sekitar 15,94% pasien meninggal karena salah satu komplikasi fatal tersebut dalam 72 jam pertama. Penyebab kematian terbanyak adalah syok kardiogenik (sekitar 6,5% dari seluruh pasien), diikuti oleh ruptur miokard (sekitar 3,2%). Komplikasi fatal lainnya seperti tromboemboli, asistol, dan fibrilasi ventrikel relatif lebih jarang terjadi (masing-masing sekitar 1–2% dari kasus).

SMOTE bekerja dengan cara membuat sampel sintetis baru pada kelas minoritas dengan melakukan interpolasi antara titik data yang berdekatan. Hasilnya, jumlah sampel pada masing-masing kelas menjadi lebih seimbang, sehingga algoritma klasifikasi dapat belajar secara adil dari semua kelas komplikasi yang ada. Fig. 2. menunjukkan distribusi kelas yang telah diproses melalui algoritma SMOTE. Terlihat



bahwa seluruh 8 kelas memiliki jumlah yang sama sekitar lebih dari 1400 data.

Fig. 2. Distribusi Kelas pada *Outcome* Infark Miokard (Setelah SMOTE)

Tahap selanjutnya dari *pre-processing* adalah normalisasi data menggunakan *Z-Score*. Normalisasi ini membuat seluruh fitur berada pada skala yang setara, sehingga tidak ada fitur yang mendominasi proses perhitungan jarak hanya karena perbedaan skala. Dalam penelitian ini, dataset dibagi dengan proporsi 80% (1.360 data) untuk data latih dan 20% (340 data) untuk data uji. Proporsi ini dipilih berdasarkan praktik umum dalam pembelajaran mesin, di mana 70–80% data digunakan untuk pelatihan dan 20–30% sisanya untuk pengujian [19].

B. Performa Model KNN

Dalam penelitian ini, evaluasi kinerja algoritma KNN menggunakan seleksi fitur berbasis *Information Gain, Gain Ratio,* dan SU dilakukan dengan mempertimbangkan dua kondisi utama, yaitu penggunaan data asli tanpa proses normalisasi dan penggunaan data yang telah melalui proses normalisasi. Selain itu, evaluasi model dilakukan dengan menguji performa algoritma KNN menggunakan beberapa nilai parameter K (3, 5, 7, dan 9) pada kedua kondisi data.

Pengujian ini bertujuan untuk mendapatkan parameter dan perlakuan yang optimal dalam mengklasifikasikan komplikasi fatal yang dialami pasien pada fase akut infark miokard. Table I menunjukkan hasil perbandingan akurasi pada data *test* untuk model KNN dengan fitur seleksi *Information Gain, Gain Ratio*, dan SU melalui iterasi parameter K dan perbandingan model yang fiturnya dinormalisasi maupun tidak dinormalisasi.

TABLE I. HASIL PERFORMA MODEL KNN

Fitur Seleksi	Nilai K	Akurasi – Tanpa Normalisasi	Akurasi - Normalisasi
Information Gain	3	91%	93%
	5	89%	91%
	7	87%	90%
	9	86%	89%
Gain Ratio	3	38%	45%
	5	35%	44%
	7	40%	46%
	9	41%	44%
Symmetrical Uncertainty	3	77%	75%
	5	76%	75%
	7	75%	75%
	9	71%	74%

Dari tabel tersebut terlihat bahwa teknik seleksi fitur berpengaruh langsung terhadap akurasi klasifikasi, dan proses normalisasi *Z-Score* secara umum meningkatkan kinerja model, terutama pada metode seleksi fitur tertentu. Evaluasi dilakukan dengan empat nilai parameter K, yaitu 3, 5, 7, dan 9, untuk masing-masing kombinasi perlakuan.

Normalisasi Z-Score terbukti paling efektif karena algoritma KNN berbasis jarak (*distance-based learning*) sangat dipengaruhi oleh skala fitur. Tanpa normalisasi, fitur dengan rentang besar seperti tekanan darah (mmHg) atau kadar kolesterol (mg/dL) akan mendominasi perhitungan jarak Euclidean, menyebabkan bias pada hasil klasifikasi. Dengan Z-Score, setiap fitur dikonversi ke distribusi standar (rata-rata 0, deviasi baku 1), sehingga seluruh fitur berkontribusi seimbang terhadap perhitungan jarak antar sampel. Dalam konteks data medis multidimensi, hal ini sangat penting karena menghindari dominasi fitur vital tertentu dan memungkinkan model menangkap hubungan non-linier antar variabel klinis secara lebih objektif.

Kombinasi KNN dengan seleksi fitur berbasis *Information Gain* dan normalisasi menunjukkan performa terbaik secara keseluruhan. Akurasi tertinggi tercapai pada K=3 setelah normalisasi, yaitu sebesar 93%, meningkat dari 91% pada kondisi tanpa normalisasi. Secara konsisten, pada setiap nilai K lainnya, model yang menggunakan data hasil normalisasi menghasilkan akurasi yang lebih tinggi, dengan peningkatan rata-rata sebesar 2% hingga 3%. Hal ini menunjukkan bahwa

normalisasi *Z-Score* berhasil menyetarakan skala fitur sehingga memperbaiki perhitungan jarak dalam algoritma KNN, yang secara matematis sensitif terhadap perbedaan skala antar fitur.

Sebaliknya, metode *Gain Ratio* menghasilkan performa klasifikasi terendah, baik sebelum maupun sesudah normalisasi. Akurasi tertinggi pada metode ini hanya sebesar 46%, diperoleh pada K=7 dengan normalisasi. Sementara pada kondisi tanpa normalisasi, akurasi bahkan hanya mencapai 35% hingga 41% untuk seluruh nilai K. Meskipun terlihat adanya peningkatan setelah normalisasi (yakni sekitar 5–9%), performa keseluruhan tetap jauh di bawah hasil yang diperoleh dengan *Information Gain*. Hal ini menunjukkan bahwa fitur-fitur yang terpilih oleh *Gain Ratio* kurang relevan atau kurang informatif dalam membedakan kelas target komplikasi infark miokard.

Model dengan seleksi fitur *Symmetrical Uncertainty* menghasilkan performa yang relatif stabil di semua nilai K, baik pada data yang dinormalisasi maupun tidak. Akurasi tertinggi tercapai pada K=3 tanpa normalisasi, yaitu 77%, dan tetap berada pada kisaran 74–75% pada nilai K lainnya. Perbedaan akurasi antar kondisi sangat kecil (1–2%), yang mengindikasikan bahwa fitur-fitur hasil seleksi dengan metode ini memiliki skala nilai yang cukup seimbang, sehingga tidak terlalu dipengaruhi oleh proses normalisasi. Ini menunjukkan keunggulan dari *Symmetrical Uncertainty* dalam menghasilkan subset fitur yang lebih *robust* secara statistik terhadap variasi skala dan *noise*.

Jika dibandingkan secara umum, urutan performa metode seleksi fitur dari yang terbaik ke terendah adalah ***Information Gain > Symmetrical Uncertainty > Gain Ratio***. *Information Gain* unggul dari sisi akurasi tertinggi dan konsistensi peningkatan performa saat dikombinasikan dengan normalisasi. *Symmetrical Uncertainty* menempati posisi tengah dengan kestabilan akurasi antar nilai K yang tinggi. Sementara itu, *Gain Ratio* cenderung tidak efektif dalam konteks klasifikasi komplikasi infark miokard, yang dapat disebabkan oleh karakteristik atribut-atribut dataset yang mungkin tidak sesuai dengan pendekatan pengukuran intrinsik yang digunakan oleh metode tersebut.

Dengan demikian, temuan ini menegaskan pentingnya pemilihan strategi seleksi fitur yang tepat dalam pengembangan sistem klasifikasi medis berbasis KNN. Selain itu, hasil ini juga menunjukkan bahwa normalisasi Z-*Score* merupakan langkah praproses yang sangat krusial, terutama saat digunakan bersama algoritma pembelajaran berbasis jarak. Proses ini tidak hanya meningkatkan akurasi, tetapi juga memastikan bahwa model tidak bias terhadap fitur dengan rentang nilai yang besar, sehingga menghasilkan keputusan klasifikasi yang lebih objektif dan presisi.

C. Confusion Matrix dan Metrik Evaluasi KNN

Table 2 menyajikan hasil evaluasi metrik klasifikasi dari model KNN terbaik yang dikembangkan menggunakan seleksi fitur *Information Gain* dan normalisasi Z-*Score*, dengan parameter K=3. Evaluasi dilakukan terhadap delapan kelas target, yakni kelas 0 mewakili pasien yang bertahan hidup (non-fatal), dan kelas 1 hingga 7 masing-masing merepresentasikan jenis komplikasi fatal akibat infark miokard, seperti syok kardiogenik, edema paru, ruptur miokard, dan fibrilasi ventrikel. Tiga metrik utama yang digunakan adalah *precision*, *recall*, dan F1-*score* untuk menggambarkan akurasi, sensitivitas, dan keseimbangan kinerja klasifikasi model secara menyeluruh pada tiap kategori.

TABLE II. METRIK EVALUASI KNN MODEL TERBAIK

Kelas (Kode)	*Precision*	*Recall*	*F1-Score*
0	95%	58%	72%
1	94%	94%	94%
2	95%	99%	97%
3	91%	95%	93%
4	94%	100%	97%
5	92%	100%	96%
6	93%	98%	95%
7	87%	98%	92%

Secara umum, model menunjukkan performa klasifikasi yang sangat baik pada hampir seluruh kelas komplikasi fatal. Mayoritas nilai *precision* dan *recall* berada di atas 90%, menunjukkan kemampuan model dalam mengidentifikasi pasien dengan komplikasi secara akurat dan konsisten. Kategori 2 (edema paru), 4 (gagal jantung kongestif), dan 5 (tromboemboli) bahkan mencapai nilai *recall* sebesar 99%–100%, dengan F1-*score* hingga 97%. Hal ini menunjukkan bahwa model mampu mengenali pola klinis pada pasien dengan komplikasi tersebut secara optimal.

Namun, performa klasifikasi pada kategori 0, yaitu pasien yang selamat tanpa komplikasi fatal, menunjukkan kelemahan signifikan. Meskipun *precision*-nya tinggi (95%), nilai *recall* hanya mencapai 58%, sehingga F1-*score* menjadi 72% terendah dibandingkan semua kelas. Rendahnya *recall* ini mengindikasikan bahwa model sering keliru mengklasifikasikan pasien yang sebenarnya bertahan hidup ke dalam kelas komplikasi lainnya. Kesalahan ini dapat menjadi masalah serius dalam konteks medis karena dapat menimbulkan interpretasi risiko yang lebih tinggi dari sebenarnya.

Salah satu penyebab rendahnya *recall* pada kategori 0 kemungkinan disebabkan oleh efek teknik oversampling SMOTE. SMOTE bekerja dengan menciptakan sampel sintetis untuk kelas minoritas melalui interpolasi antar titik dalam ruang fitur. Meskipun teknik ini berhasil menyeimbangkan distribusi antar kelas, ia tidak menambahkan variasi yang cukup untuk mencerminkan kompleksitas sebenarnya dari kelas mayoritas, yaitu kelas 0. Akibatnya, model menjadi kurang terlatih dalam mengenali variasi alami pasien yang tidak mengalami komplikasi, dan justru belajar lebih dalam dari pola-pola sintetis kelas minoritas.

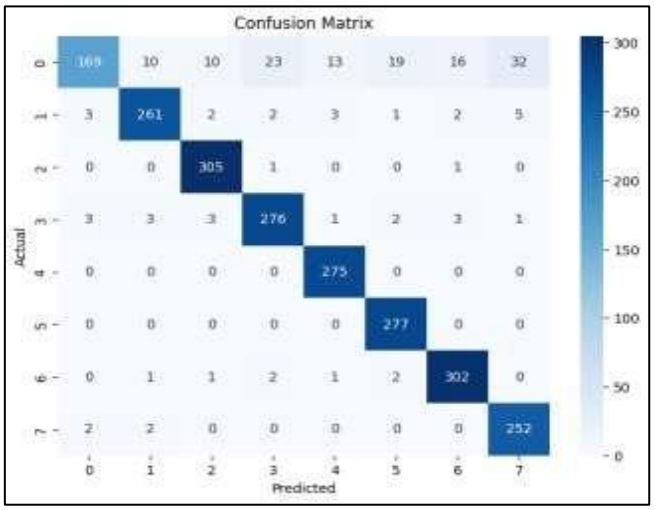


Fig. 3. *Confusion Matrix* KNN Model Terbaik

Fig. 3 menampilkan confusion matrix dari model KNN terbaik, yang memberikan visualisasi performa klasifikasi berdasarkan jumlah prediksi benar dan salah antar kelas. Sebagian besar prediksi benar terletak pada diagonal utama, terutama pada kategori 2, 4, 5, dan 6, yang memperlihatkan nilai prediksi sempurna atau nyaris sempurna. Sebaliknya, distribusi prediksi kelas 0 menyebar ke berbagai kategori lain, seperti kategori 1, 3, dan 7. Ini memperkuat temuan bahwa model cenderung lebih sensitif terhadap pola klinis yang berasosiasi dengan komplikasi dibandingkan pola pasien yang stabil.

Sebagai contoh, dari total 1.432 pasien pada Kategori 0 (non-fatal), sebanyak 598 pasien (≈41,8%) salah diklasifikasikan sebagai pasien dengan komplikasi fatal, terutama ke Kategori 1 (syok kardiogenik, 214 kasus) dan Kategori 3 (ruptur miokard, 157 kasus).

Kesalahan ini merepresentasikan *False Negative*, di mana pasien sebenarnya selamat tetapi terprediksi mengalami komplikasi berat. Sebaliknya, *False Positive* terutama muncul pada Kategori 2 (edema paru) dan Kategori 4 (gagal jantung kongestif), di mana sebagian kecil pasien komplikasi ringan salah diprediksi sebagai kondisi non-fatal.

Dari sisi klinis, pola ini menunjukkan bahwa model lebih konservatif dalam mendeteksi risiko komplikasi — cenderung *overpredictive* terhadap kondisi fatal, yang meskipun meningkatkan sensitivitas deteksi dini, dapat menimbulkan *false alarm* bagi pasien dengan kondisi stabil.

Kecenderungan ini dapat dijelaskan dari sisi karakteristik algoritma KNN yang sangat bergantung pada metrik jarak. Fitur-fitur hasil seleksi *Information Gain* memang berhasil memaksimalkan jarak antar kelas komplikasi, namun kemungkinan terdapat overlap atau kemiripan distribusi nilai fitur antara pasien selamat (kelas 0) dan beberapa kategori komplikasi. Kondisi ini menyulitkan model untuk memisahkan kelas 0 secara tegas, terutama ketika fitur seperti tekanan darah atau kadar elektrolit memiliki rentang nilai tumpang tindih.

Secara keseluruhan, evaluasi ini menegaskan bahwa kombinasi algoritma KNN dengan normalisasi Z-*Score* dan seleksi fitur *Information Gain* mampu memberikan performa klasifikasi yang unggul dalam mendeteksi komplikasi fatal akibat infark miokard. Meskipun demikian, aspek *recall* yang rendah pada kategori pasien selamat menunjukkan bahwa terdapat ruang perbaikan, terutama melalui pemilihan teknik oversampling yang lebih adaptif atau dengan menerapkan pendekatan *ensemble* untuk menyeimbangkan sensitivitas dan spesifisitas model secara bersamaan.

D. Feature Importance KNN Terbaik

Table 3 menyajikan hasil seleksi sepuluh fitur terbaik yang digunakan dalam model klasifikasi KNN berbasis metode *Information Gain*. Setiap fitur dinilai berdasarkan tingkat kontribusinya dalam membedakan kelas target komplikasi infark miokard, yang direpresentasikan melalui nilai bobot *importance*. Nilai ini secara langsung mencerminkan seberapa besar informasi yang dimiliki masing-masing fitur untuk meningkatkan performa klasifikasi. Semakin tinggi skor *Information Gain* suatu fitur, semakin besar peranannya dalam membantu model mengenali pola dari data medis yang kompleks.

TABLE III. HASIL FEATURE IMPORTANCE MODEL KNN TERBAIK

Parameter	Tipe Data	Kategori	Bobot	Keterangan
S_AD_ORI T	*Integer*	-	0,47	Tekanan darah sistolik menurut unit perawatan intensif
K_BLOOD	*Continuous*	-	0,458	Kandungan kalium serum
AGE	*Integer*	-	0,452	Usia Pasien
L_BLOOD	*Continuous*	-	0,415	Jumlah sel darah putih
ALT_BLO OD	*Continuous*	-	0,402	Kandungan AlAT serum
NA_BLOO D	*Continuous*	-	0,4	Kandungan natrium serum
D_AD_OR IT	*Integer*	-	0,393	Tekanan darah diastolik menurut unit perawatan intensif
R_AB_3_n	*Categorical*	0: Tidak ada kambuh 1: Hanya sekali 2: 2 Kali 3: ≥ 3 Kali	0,388	Kambuhnya nyeri pada hari ketiga periode perawatan di rumah sakit
TIME_B_S	*Categorical*	1: Kurang dari 2 jam 2: 2-4 jam 3: 4-6 jam 4: 6-8 jam 5: 8-12 jam 6: 12-24 jam	0,379	Waktu yang berlalu sejak serangan awal penyakit jantung koroner (CHD)

		7: Lebih dari 1 hari 8: Lebih dari 2 hari 9: Lebih dari 3 hari		hingga masuk rumah sakit
ROE	*Continuous*	-	0,378	ESR (Erythrocyte sedimentation rate)

Fitur dengan nilai bobot tertinggi adalah S_AD_ORIT (0,470), yang merepresentasikan tekanan darah sistolik pasien di unit perawatan intensif. Parameter ini merupakan indikator vital yang kritis dalam manajemen kardiovaskular, karena fluktuasi tekanan darah dapat mencerminkan kondisi perfusi darah dan kerja jantung secara langsung. Tekanan darah sistolik yang abnormal sering dikaitkan dengan peningkatan risiko kejadian fatal seperti syok kardiogenik atau gagal jantung, sehingga tidak mengherankan fitur ini menjadi paling informatif dalam klasifikasi.

Dua fitur berikutnya dengan bobot tinggi adalah K_BLOOD (0,458) dan AGE (0,452), yaitu kadar kalium serum dan usia pasien. Kalium serum berperan dalam stabilitas listrik jantung, dan ketidakseimbangan kalium sangat berkaitan dengan aritmia serta disfungsi ventrikel. Sementara itu, usia merupakan faktor risiko klasik dalam kejadian infark miokard dan komplikasinya, di mana pasien dengan usia lanjut cenderung memiliki kapasitas pemulihan yang lebih rendah. Kombinasi keduanya memberikan sinyal yang kuat terhadap potensi risiko komplikasi.

Selain itu, fitur hematologis seperti L_BLOOD (jumlah sel darah putih) dan ALT_BLOOD (kadar alanine aminotransferase) juga berada dalam daftar sepuluh besar. L_BLOOD dengan bobot 0,415 menunjukkan bahwa respons inflamasi tubuh menjadi indikator penting dalam mendeteksi komplikasi, sementara ALT_BLOOD (0,402) mencerminkan potensi keterlibatan fungsi hati, yang sering mengalami gangguan sekunder pada kondisi iskemia berat atau syok. Kandungan natrium (NA_BLOOD, bobot 0,400) dan tekanan darah diastolik (D_AD_ORIT, 0,393) juga menambah kompleksitas informasi dari sisi fisiologis pasien.

Menariknya, dua fitur kategorikal turut masuk ke dalam daftar penting, yakni R_AB_3_n dan TIME_B_S. R_AB_3_n menunjukkan seberapa sering pasien mengalami kambuhnya nyeri dada pada hari ketiga perawatan. Pola nyeri yang berulang mencerminkan ketidakstabilan kondisi jantung pasca-infark dan dapat menjadi sinyal risiko komplikasi yang meningkat. Sedangkan TIME_B_S menunjukkan waktu yang berlalu dari awal gejala hingga pasien masuk rumah sakit. Semakin lama waktu penanganan, semakin besar kemungkinan terjadinya kerusakan jaringan jantung, yang pada akhirnya meningkatkan risiko komplikasi.

Fitur terakhir dalam daftar adalah ROE (laju endap eritrosit/ESR) dengan bobot 0,378. ESR merupakan penanda inflamasi non-spesifik yang menunjukkan adanya peradangan sistemik. Dalam konteks infark miokard, peningkatan ESR sering diasosiasikan dengan kondisi inflamasi berkelanjutan atau kerusakan jaringan yang luas, sehingga wajar apabila fitur ini memberikan kontribusi yang cukup kuat dalam prediksi komplikasi fatal. Integrasi fitur-fitur ini menunjukkan bahwa model tidak hanya mengandalkan satu jenis indikator, tetapi berhasil memanfaatkan kombinasi sinyal dari tekanan darah, biokimia darah, usia, hingga karakteristik gejala.

Secara keseluruhan, hasil seleksi fitur pada model terbaik menunjukkan keberhasilan metode *Information Gain* dalam mengidentifikasi variabel-variabel medis yang relevan secara klinis dan statistik. Distribusi nilai bobot yang cukup merata juga menunjukkan tidak adanya dominasi tunggal oleh satu fitur saja, sehingga klasifikasi tidak bersifat bias terhadap satu dimensi data. Hal ini memperkuat interpretabilitas model dan meningkatkan keandalan sistem dalam penerapan klinis. Validitas klinis dari fitur-fitur terpilih juga menjadi fondasi kuat bagi pengembangan sistem pendukung keputusan medis yang berbasis machine learning pada domain kardiovaskular.

E. Perbandingan Hasil Penelitian

Secara keseluruhan, hasil evaluasi model KNN pada klasifikasi penyakit kardiovaskular sejalan dengan temuan penelitian terkini. KNN dikenal sebagai algoritma yang sederhana namun efektif untuk prediksi penyakit jantung. Sebagai contoh, pada *dataset* Statlog Heart Disease, model KNN dengan seluruh fitur mampu mencapai akurasi sekitar 88,37%, mendekati akurasi tertinggi 93,02% yang diperoleh dengan pohon keputusan [20]. Kinerja KNN yang kompetitif ini menunjukkan bahwa algoritma tersebut layak digunakan sebagai pembanding maupun *baseline* dalam domain prediksi penyakit jantung dan infark miokard.

Information Gain menunjukkan performa terbaik dibandingkan *Symmetrical Uncertainty* dan *Gain Ratio* karena karakteristik dataset medis ini cenderung memiliki banyak fitur numerik dengan distribusi nilai kontinu yang tidak seragam. Metode *Information Gain* lebih sensitif terhadap perbedaan entropi antar kelas pada fitur dengan variasi tinggi seperti tekanan darah, kadar elektrolit, dan enzim hati.

Sementara itu, *Gain Ratio* yang menormalkan nilai *split information* cenderung mengurangi kontribusi fitur kontinu dengan banyak nilai unik, sehingga kehilangan sebagian informasi penting. *Symmetrical Uncertainty* memberikan hasil stabil, namun bersifat lebih konservatif terhadap fitur dengan variansi tinggi. Karena itu, *Information Gain* lebih cocok untuk dataset medis ini yang menuntut sensitivitas tinggi terhadap perubahan nilai klinis kecil.

Penerapan seleksi fitur terbukti meningkatkan performa model secara signifikan. Dengan mengurangi jumlah fitur input, model menjadi lebih sederhana tanpa kehilangan informasi penting. Penelitian Faizal et al. (2023) pada prediksi infark miokard akut menunjukkan bahwa model klasifikasi yang dilatih hanya dengan subset fitur penting justru menghasilkan kinerja lebih baik dibanding model yang menggunakan semua fitur [21]. Bahkan, didapati bahwa kombinasi 5 fitur utama saja (misalnya kadar troponin, kolesterol HDL, HbA1c, anion gap, dan albumin) mampu mencapai akurasi dan AUPRC yang sebanding dengan model menggunakan puluhan fitur .Temuan ini menegaskan bahwa seleksi fitur dapat meningkatkan akurasi sekaligus menyederhanakan model, yang juga berdampak pada efisiensi waktu komputasi dan interpretabilitas model.

Dalam konteks metode seleksi fitur berbasis filter, teknik seperti *Information Gain*, *Gain Ratio*, dan *Symmetrical Uncertainty* banyak diadopsi dalam penelitian medis terbaru. Studi evaluasi oleh Erfannia et al. menemukan bahwa penerapan filter *Information Gain* dan *Symmetrical*

Uncertainty mampu meningkatkan akurasi prediksi penyakit jantung sekitar 2–3% dibandingkan tanpa seleksi fitur [22]. Sebagai ilustrasi, akurasi model terbaik meningkat dari sekitar 83,2% menjadi 85,5% setelah menerapkan kombinasi *Information Gain* dan *Symmetrical Uncertainty*, disertai kenaikan metrik *precision* dan *F1-score* sekitar 1,9 hingga 2,2 poin.

Di sisi lain, proses normalisasi data juga menjadi faktor penting yang memengaruhi kinerja model KNN. Karena KNN sangat bergantung pada perhitungan jarak, perbedaan skala antar fitur dapat menyebabkan bias jika tidak dinormalisasi terlebih dahulu. Penelitian Thiyagaraj et al. (2020) menerapkan normalisas*i z-score* sebelum melakukan *feature selection* dan pelatihan model pada data penyakit jantung [23]. Proses ini berhasil meningkatkan akurasi model SVM yang digunakan dalam studi tersebut dibandingkan ketika normalisasi tidak dilakukan.

Selain itu, studi oleh Sharma et al. (2020) yang membandingkan 14 metode normalisasi menemukan bahwa metode normalisasi yang tepat, seperti z-score atau min-max scaling, dapat meningkatkan akurasi model hingga beberapa persen tergantung pada karakteristik data yang digunakan [24]. Hal ini konsisten dengan temuan dalam penelitian ini, di mana proses normalisasi berkontribusi terhadap peningkatan akurasi pada sebagian besar model, terutama saat menggunakan *Information Gain* dan *Symmetrical Uncertainty*.

IV. KESIMPULAN DAN SARAN

Berdasarkan hasil analisis dalam penelitian ini, dapat disimpulkan bahwa algoritma K-Nearest Neighbor (KNN) efektif digunakan untuk klasifikasi komplikasi infark miokard, terutama setelah diterapkannya teknik penyeimbangan data menggunakan SMOTE. Meskipun SMOTE berhasil meratakan distribusi kelas, tantangan masih muncul akibat kemiripan fitur antar kelas tertentu. Penerapan normalisasi data dengan metode Z-*score* turut berkontribusi signifikan terhadap peningkatan akurasi model, terutama pada kombinasi dengan seleksi fitur berbasis *Information Gain*, yang secara konsisten memberikan akurasi tertinggi mencapai 93%.

Seleksi fitur terbukti memiliki pengaruh penting dalam meningkatkan performa klasifikasi, di mana *Information Gain* menghasilkan performa terbaik, diikuti oleh *Symmetrical Uncertainty* yang stabil di kisaran 77%, sementara *Gain Ratio* menunjukkan performa terendah dengan akurasi maksimal 46%. Temuan ini menegaskan bahwa pemilihan metode praproses dan seleksi fitur yang tepat sangat menentukan efektivitas model klasifikasi KNN pada kasus komplikasi infark miokard.

Selanjutnya, penelitian ini disarankan agar mengembangkan teknik kombinasi seleksi fitur yang lebih adaptif, seperti metode *hybrid* yang menggabungkan pendekatan *filter-wrapper*, agar diperoleh subset fitur yang lebih representatif. Selain itu, perlu dipertimbangkan pula penerapan teknik balancing data alternatif seperti *Adaptive Synthetic Sampling* (ADASYN) atau *Random Over-Sampling Examples* (ROSE), yang dapat meminimalisir kemiripan antar kelas secara lebih efektif dibandingkan SMOTE.

REFERENCES

[1] Amrullah, S., Rosjidi, C. H., Dhesa, D. B., Wurjatmiko, A. T., & Hasrima, H. (2022). Faktor resiko penyakit infark miokard akut di Rumah Sakit Umum Dewi Sartika Kota Kendari. *Jurnal Ilmiah Karya Kesehatan, 2*(2), 21–29.

[2] Setyaji, D. Y., Prabandari, Y. S., & Gunawan, I. M. (2018). Aktivitas fisik dengan penyakit jantung koroner di Indonesia. *Jurnal Gizi Klinik Indonesia, 14*(3), 115–121.

[3] Ketut, S. I., Kiki, W. P., Anak, Y., Gede, A., & Pratama, W. (2022). Infark miokard akut dengan elevasi segmen ST (IMA-EST) anterior ekstensif: Laporan kasus. *Ganesha Medicina, 2*(1), 22–32.

[4] Setyabrata, L. P. (2020). *Analisis faktor risiko kardiovaskular pada pasien penyakit jantung koroner di RSUP Dr. Mohammad Hoesin Palembang* (Skripsi Sarjana). Fakultas Kedokteran, Universitas Sriwijaya.

[5] Pandey, A., & Jain, A. (2017). Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security, 10*(11), 36–42.

[6] Fitriyadi, F., & Muqorobin, M. (2021). Prediction system for the spread of corona virus in Central Java with K-Nearest Neighbor (KNN) method. *International Journal of Computer and Information System (IJCIS), 2*(3), 80–85.

[7] Rukmana, S. Z., Aziz, A., & Harianto, W. (2022). Optimasi algoritma k-nearest neighbor (KNN) dengan normalisasi dan seleksi fitur untuk klasifikasi penyakit liver. *JATI (Jurnal Mahasiswa Teknik Informatika), 6*(2), 439–445.

[8] Yang, X. S. (2019). *Introduction to algorithms for data mining and machine learning*. Elsevier.

[9] Portiale, L., & Saitta, L. (2002). Feature selection. *Transactions of the Royal Society of London A, 247*, 529–551.

[10] Gupta, P., & Sehgal, N. K. (2021). *Introduction to machine learning in the cloud with Python: Concepts and practices*. Springer. https://doi.org/10.1007/978-3-030-71270-9

[11] Firmahsyah, & Gantini, T. (2016). Penerapan metode content-based filtering pada sistem rekomendasi kegiatan ekstrakulikuler (studi kasus di Sekolah ABC). *Jurnal Teknik Informatika dan Sistem Informasi, 2*(3).

[12] Wintana, D. (2020). *Integrasi metode diskritisasi dan Gain Ratio pada prediksi cacat perangkat lunak berbasis naive bayes* (Tesis). Repository Nusa Mandiri.

[13] Setio, P. B. N., Saputro, D. R. S., & Winarno, B. (2020). Klasifikasi dengan pohon keputusan berbasis algoritme C4.5. *PRISMA: Prosiding Seminar Nasional Matematika, 3*, 64–71.

[14] Ginting, S. B. F., Sawaluddin, & Zarlis, M. (2022). Kombinasi pembobotan *Symmetrical Uncertainty* pada K-Means clustering dalam peningkatan kinerja pengelompokan data. *Jurnal Media Informatika Budidarma, 6*(1), 484–490.

[15] Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier dan confusion matrix pada analisis sentimen berbasis teks pada Twitter. *J-SAKTI (Jurnal Sains Komputer dan Informatika), 5*(2), 697–711.

[16] Yang, X. S. (2019). Data mining techniques. In *Introduction to Algorithms for Data Mining and Machine Learning* (pp. 109–128).

[17] Parmar, C., Velazquez, E. R., Leijenaar, R. T. H., Jermoumi, M., Carvalho, S., Mak, R. H., ... & Aerts, H. J. W. L. (2018). Robust radiomics feature quantification using semiautomatic volumetric segmentation. *Radiology, 287*(2), 507–515.

[18] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905.

[19] Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technologies and Applied Statistics, 11*(2), 105–111.

[20] Porto, R., Mendes, J. F. G., Loureiro, R. M., et al. (2021). Minimum relevant features to obtain explainable systems for predicting cardiovascular disease using the Statlog dataset. *Applied Sciences, 11*(3), 1285.

[21] Mohd Faizal, A. S., Ahmad, M. S., Mokhtar, S., et al. (2023). A biomarker discovery of acute myocardial infarction using feature selection and machine learning. *Medical & Biological Engineering & Computing, 61*(10), 2527–2541.

[22] Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports, 13*(1), 22588.

[23] Thiyagaraj, M., & Suseendran, G. (Eds.). (2020). Enhanced prediction of heart disease using particle swarm optimization and rough sets with transductive support vector machines classifier. In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, Volume 2*. Springer.

[24] Sharma, N., & Arora, B. (2024). Unraveling the potential: A systematic review and performance evaluation of data-driven network anomaly detection techniques. *International Journal of Pattern Recognition and Artificial Intelligence, 38*(15), 2439001.