JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 14 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v14i3.10773

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR UNTUK KLASIFIKASI TINGKAT KEBUGARAN TUBUH BERDASARKAN DATA BIOMETRIK

Christina Nainggolan[1]*, Filipi Siahaan[2], Diva Sirait[3], Johespan Sipayung[4], Indra M Sarkis S[5]
[1, 2, 3, 4, 5] Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Methodist Indonesia, Jl. Hang Tuah No. 8, Medan, Indonesia

Kata Kunci:
K-Nearest Neighbor; klasifikasi; kebugaran tubuh; data biometrik.

Korespondensi Email:
nainggolanchristina21@gmail.com

Abstrak. Tingkat kebugaran fisik adalah parameter penting untuk mengevaluasi kondisi kesehatan tubuh, namun pengukuran manual bersifat subjektif dan memakan waktu. Penelitian ini menerapkan algoritma K-Nearest Neighbor (KNN) untuk mengklasifikasikan tingkat kebugaran tubuh ke dalam empat kategori (A, B, C, D) berdasarkan sepuluh indikator biometrik: usia, tinggi badan, berat badan, persentase lemak tubuh, tekanan darah diastolik dan sistolik, kekuatan genggaman, fleksibilitas, jumlah sit-up, dan hasil lompatan lebar. Dataset berisi 13.393 data, dibagi menjadi data latih (80%) dan data uji (20%) dengan stratified split, dinormalisasi menggunakan Min-Max Scaling, dan dihitung jaraknya dengan Euclidean Distance. Pengujian beberapa nilai K menunjukkan bahwa K=9 memberikan akurasi terbaik sebesar 59,54% pada data uji. Matriks kebingungan menunjukkan model lebih baik mengenali kelas ekstrem (A dan D) dibandingkan kelas menengah (B dan C) yang memiliki kemiripan karakteristik biometrik. Penelitian ini menyimpulkan bahwa KNN dapat diterapkan untuk klasifikasi kebugaran tubuh dengan akurasi yang memadai, serta menjelaskan pengaruh nilai K dan karakteristik data terhadap akurasi klasifikasi.

Abstract.*Physical fitness is an important indicator of health condition, yet traditional manual measurement is subjective and time-consuming. This study applies the K-Nearest Neighbor (KNN) algorithm to classify body fitness levels into four categories (A, B, C, D) based on ten biometric indicators: age, height, weight, body fat percentage, diastolic and systolic blood pressure, grip strength, flexibility, sit-up count, and broad jump distance. The dataset consists of 13,393 records, split into training (80%) and test (20%) sets using stratified sampling, normalized with Min-Max Scaling, with distances computed via Euclidean Distance. Testing several K values showed that K=9 achieved the best accuracy of 59.54% on the test data. The confusion matrix showed the model performs better on extreme classes (A and D) than on middle classes (B and C), which share similar biometric characteristics. This study concludes that KNN can be applied to body fitness classification with adequate accuracy, and explains the effect of K value and data characteristics on classification accuracy.*

1. PENDAHULUAN

Kebugaran jasmani menjadi indikator penting derajat kesehatan seseorang, karena berkaitan dengan kemampuan fungsional tubuh dan pencegahan penyakit tidak menular seperti obesitas, hipertensi, dan gangguan kardiovaskular. Penilaian kebugaran secara konvensional melibatkan banyak parameter biometrik (antropometri, tekanan darah, kekuatan otot, fleksibilitas, daya ledak otot), namun interpretasi manual terhadap parameter tersebut cenderung subjektif, memakan waktu, dan rentan kesalahan tanpa dukungan sistem terkomputerisasi.

Perkembangan machine learning membuka peluang otomatisasi klasifikasi biometrik secara lebih objektif. Berbagai algoritma seperti Naive Bayes, SVM, Decision Tree, Random Forest, dan KNN telah banyak diterapkan di bidang

kesehatan [1], [2], [3]. KNN menjadi salah satu yang paling populer karena sederhana, non-parametrik, dan efektif untuk data numerik multidimensi tanpa asumsi distribusi tertentu [4].

Penelitian terdahulu telah membuktikan keberhasilan KNN pada berbagai kasus, seperti klasifikasi kematangan buah [1], status gizi balita [5], penyakit jantung [6], dan kesegaran daging [7], serta kajian pengaruh jarak Euclidean terhadap akurasi klasifikasi [5], [6]. Namun, penerapan KNN khusus untuk klasifikasi kebugaran tubuh berdasarkan sepuluh atribut biometrik sekaligus—mencakup antropometri, kardiovaskular, kekuatan otot, fleksibilitas, dan daya ledak—masih jarang dikaji, terutama pada dataset berskala besar seperti yang digunakan dalam penelitian ini.

Kesenjangan ini mendasari penelitian yang menerapkan algoritma KNN untuk mengklasifikasikan tingkat kebugaran tubuh ke dalam empat kelas (A, B, C, D) menggunakan dataset biometrik sebanyak 13.393 data. Penelitian ini juga menguji berbagai nilai K untuk memperoleh nilai optimal, serta mengevaluasi model menggunakan confusion matrix dan metrik akurasi.

Rumusan masalah penelitian ini: (1) bagaimana implementasi KNN untuk klasifikasi kebugaran tubuh berdasarkan data biometrik; (2) bagaimana pengaruh nilai K terhadap akurasi klasifikasi; dan (3) seberapa besar akurasi yang dicapai model KNN pada dataset ini. Tujuan penelitian ini adalah membangun model klasifikasi kebugaran tubuh menggunakan KNN, menganalisis pengaruh nilai K terhadap akurasi, serta mengevaluasi performa model melalui confusion matrix dan accuracy.

2. TINJAUAN PUSTAKA

2.1. Kebugaran Jasmani dan Data Biometrik

Kebugaran jasmani (*physical fitness*) didefinisikan sebagai kemampuan tubuh untuk melakukan aktivitas fisik sehari-hari tanpa mengalami kelelahan yang berarti dan masih memiliki cadangan energi untuk menghadapi keadaan darurat. Penilaian kebugaran umumnya melibatkan data biometrik seperti antropometri (tinggi dan berat badan), komposisi tubuh (persentase lemak tubuh), kardiovaskular (tekanan darah sistolik dan diastolik), kekuatan otot (*grip force*), fleksibilitas (*sit and bend forward*), daya tahan otot (*sit-up*), dan daya ledak otot (*broad jump*) [7]. Kombinasi atribut-atribut tersebut memberikan representasi multidimensi terhadap kondisi fisik seseorang yang sesuai digunakan sebagai fitur dalam model klasifikasi berbasis machine learning.

2.2. Algoritma K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) merupakan algoritma klasifikasi berbasis instance-based learning yang mengklasifikasikan data baru berdasarkan mayoritas kelas dari sejumlah K data tetangga terdekat pada data latih [4]. KNN termasuk algoritma non-parametrik karena tidak membangun model pembelajaran secara eksplisit (*lazy learner*), melainkan menyimpan seluruh data latih dan melakukan perhitungan jarak saat proses prediksi. Kedekatan antar data pada KNN umumnya diukur menggunakan metrik Euclidean Distance [5], [6].

Setelah jarak antara data uji dan seluruh data latih dihitung, dipilih sejumlah **K** data dengan jarak terkecil sebagai tetangga terdekat. Selanjutnya, kelas data uji ditentukan berdasarkan kelas yang paling banyak muncul (*suara mayoritas*). KNN dipilih karena sederhana, mudah diimplementasikan, tidak memerlukan asumsi distribusi data, serta memberikan performa yang baik pada berbagai kasus klasifikasi data biometrik dan kesehatan [8], [9], [10].

2.3. Euclidean Distance

Euclidean Distance merupakan metrik jarak yang mengukur jarak garis lurus antara dua titik pada ruang berdimensi *n*. Metrik ini banyak digunakan dalam algoritma KNN karena sederhana dan sesuai untuk data numerik kontinu [5]. Perhitungan jarak Euclidean dinyatakan dengan persamaan berikut.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Dengan:

- $d(x, y)$= jarak antara data uji dan data latih;
- x_i= nilai atribut ke-i pada data uji;
- y_i= nilai atribut ke-i pada data latih;
- n= jumlah atribut yang digunakan.

Semakin kecil nilai jarak yang diperoleh, semakin tinggi tingkat kemiripan antara data uji dan data latih sehingga data tersebut berpeluang menjadi salah satu **K** tetangga terdekat dalam proses klasifikasi.

2.4. Normalisasi Data (Min-Max Scaling)

Karena atribut biometrik pada dataset memiliki satuan dan rentang nilai yang berbeda-beda (misalnya usia dalam tahun dan tekanan darah dalam mmHg), maka diperlukan normalisasi data agar setiap atribut memiliki kontribusi yang setara terhadap perhitungan jarak. Penelitian ini menggunakan metode **Min-Max Scaling** yang mentransformasikan nilai setiap atribut ke dalam rentang [11]. Proses normalisasi dilakukan menggunakan persamaan berikut.

$$x' = \frac{x - x_{menit}}{x_{maks} - x_{menit}}$$

dengan keterangan sebagai berikut:

- x'= nilai hasil normalisasi.
- x= nilai asli suatu atribut.
- x_{menit}= nilai minimum atribut.
- x_{maks}= nilai maksimum atribut.

Melalui proses normalisasi ini, seluruh atribut memiliki skala yang sama sehingga tidak ada atribut tertentu yang mendominasi perhitungan jarak pada algoritma KNN.

2.5. Confusion Matrix dan Accuracy

Evaluasi performa model klasifikasi pada penelitian ini menggunakan confusion matrix, yaitu tabel yang merangkum jumlah prediksi benar dan salah untuk setiap kelas, serta metrik *accuracy* yang menyatakan proporsi prediksi benar terhadap keseluruhan data uji [12], [10]. Nilai *accuracy* dihitung menggunakan persamaan berikut.

$$Accuracy = \frac{\sum Prediksi\ Benar}{Total\ Data\ Uji} \times 100\%$$

dengan keterangan sebagai berikut:

- Akurasi = tingkat akurasi model klasifikasi.
- Σ Prediksi Benar = jumlah data uji yang berhasil diklasifikasikan dengan benar.
- Total Data Uji = jumlah seluruh data yang digunakan pada proses pengujian.

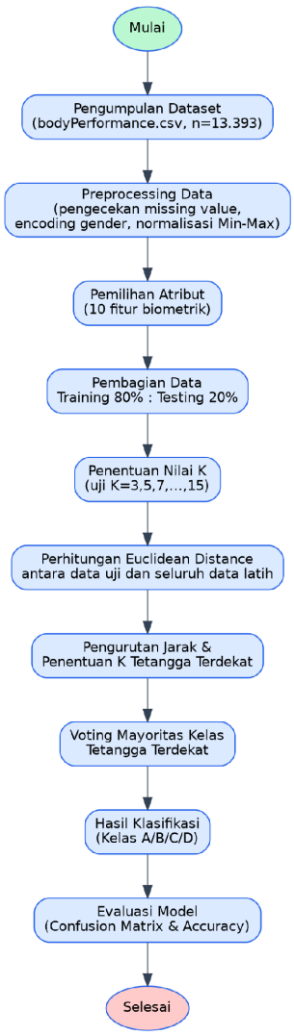
Semakin tinggi nilai *accuracy* yang diperoleh, semakin baik kemampuan model K-Nearest Neighbor dalam mengklasifikasikan tingkat kebugaran tubuh berdasarkan data biometrik. Penelitian ini menggunakan nilai *accuracy* sebagai metrik utama untuk mengevaluasi kinerja model klasifikasi.

2.6. Penelitian Terdahulu

Beberapa penelitian terdahulu yang relevan dengan penelitian ini antara lain penerapan KNN untuk klasifikasi kematangan buah mangga menggunakan fitur citra HSV yang dipublikasikan pada JITET [1], penerapan KNN untuk klasifikasi status gizi balita [13], klasifikasi penyakit jantung menggunakan KNN [8], serta kajian mengenai pengaruh metrik jarak Euclidean dan Manhattan terhadap hasil klasifikasi [5], [6]. Penelitian-penelitian tersebut secara konsisten menunjukkan bahwa performa KNN sangat dipengaruhi oleh pemilihan nilai K, metrik jarak, serta kualitas praproses data, yang menjadi dasar pertimbangan metodologi pada penelitian ini.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk menerapkan dan mengevaluasi algoritma K-Nearest Neighbor (KNN) dalam mengklasifikasikan tingkat kebugaran tubuh. Tahapan penelitian secara umum digambarkan pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

3.1. Sumber dan Deskripsi Data

Data yang digunakan pada penelitian ini adalah dataset *"bodyPerformance"* yang berisi 13393 data hasil pengukuran biometrik dan tes kebugaran fisik. Dataset terdiri atas 10 atribut numerik sebagai fitur, 1 atribut kategorikal (jenis kelamin), dan 1 atribut kelas target berupa tingkat kebugaran (A, B, C, D). Deskripsi statistik setiap atribut numerik ditunjukkan pada Tabel 1, sedangkan distribusi jumlah data pada setiap kelas ditunjukkan pada Tabel 2 dan Gambar 2.

Tabel 1. Deskripsi Statistik Atribut Biometrik

Atribut	Min	Maks	Rata-rata	Std. Dev.
Usia (tahun)	21.0	64.0	36.78	13.63
Tinggi badan (cm)	125.0	193.8	168.56	8.43
Berat badan (kg)	26.3	138.1	67.45	11.95
Lemak tubuh (%)	3.0	78.4	23.24	7.26
Tekanan darah diastolik	0.0	156.2	78.80	10.74
Tekanan darah sistolik	0.0	201.0	130.23	14.71
Kekuatan genggaman (kgf)	0.0	70.5	36.96	10.62
Sit and bend forward (cm)	-25.0	213.0	15.21	8.46
Jumlah sit-up (kali)	0.0	80.0	39.77	14.28
Broad jump (cm)	0.0	303.0	190.13	39.87

Tabel 2. Distribusi Kelas pada Dataset

Kelas	Jumlah Data	Persentase
A	3348	25.00%
B	3347	24.99%
C	3349	25.01%
D	3349	25.01%

3.2. Preprocessing Data

Tahap praproses data meliputi: (1) pemeriksaan missing value, di mana hasil pemeriksaan menunjukkan tidak terdapat nilai yang hilang pada seluruh atribut dataset; (2) encoding atribut kategorikal jenis kelamin (*gender*) menjadi nilai numerik (Laki-laki=1, Perempuan=0) sebagai informasi deskriptif tambahan pada dataset, namun atribut ini tidak diikutsertakan dalam perhitungan jarak Euclidean maupun proses klasifikasi; dan (3) normalisasi seluruh atribut numerik menggunakan metode Min-Max Scaling ke dalam rentang [0,1], sehingga atribut dengan skala nilai besar seperti *broad jump* (cm) tidak mendominasi perhitungan jarak dibandingkan atribut dengan skala nilai kecil seperti *sit and bend forward* (cm).

Rumus normalisasi Min-Max Scaling yang digunakan adalah sebagai berikut.

$$x' = \frac{x - x_{menit}}{x_{maks} - x_{menit}}$$

Keterangan:

- x'= nilai hasil normalisasi.
- x= nilai data asli.
- x_{menit}= nilai minimum pada suatu atribut.
- x_{maks}= nilai maksimum pada suatu atribut.

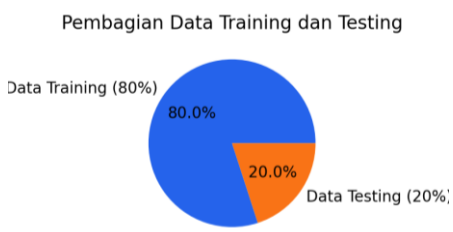
dengan x adalah nilai atribut asli, x_min dan x_max masing-masing adalah nilai minimum dan maksimum atribut tersebut pada data latih, dan x' adalah nilai atribut setelah dinormalisasi.

3.3. Pemilihan Atribut

Seluruh 10 atribut numerik biometrik (usia, tinggi badan, berat badan, persentase lemak tubuh, tekanan darah diastolik, tekanan darah sistolik, kekuatan genggaman, *sit and bend forward*, jumlah *sit-up*, dan *broad jump*) digunakan sebagai fitur dalam proses klasifikasi karena masing-masing merepresentasikan aspek kebugaran yang berbeda (antropometri, kardiovaskular, kekuatan, fleksibilitas, dan daya ledak otot) sehingga tidak dilakukan seleksi/reduksi atribut lebih lanjut.

3.4. Pembagian Data Training dan Testing

Dataset dibagi menjadi data latih (training) dan data uji (*testing*) dengan proporsi 80:20 menggunakan teknik *stratified split* agar proporsi masing-masing kelas pada data latih dan data uji tetap seimbang. Hasil pembagian data menghasilkan 10714 data latih dan 2679 data uji, sebagaimana ditunjukkan pada Gambar 3.



Gambar 3. Proporsi Pembagian Data Training dan Testing

3.5. Penentuan Nilai K

Pemilihan nilai K yang tepat sangat memengaruhi performa klasifikasi KNN. Nilai K yang terlalu kecil menyebabkan model sensitif terhadap noise (*overfitting*), sedangkan nilai K yang terlalu besar dapat menyebabkan batas antar kelas menjadi kurang tegas (*underfitting*) [4], [14], [15]. Penelitian ini menguji nilai K ganjil dari 3 hingga 15 untuk menghindari hasil voting seri (*tie*) pada klasifikasi 4 kelas, dengan hasil ditunjukkan pada Tabel 3 dan Gambar 4.

Tabel 3. Akurasi Model pada Berbagai Nilai K

Nilai K	Akurasi
3	52.82%
5	55.17%
7	56.96%
9	59.54%
11	58.79%
13	58.31%
15	58.12%

Berdasarkan Tabel 3 dan Gambar 4, akurasi tertinggi diperoleh pada K=9 sebesar 59.54%. Penelitian ini memilih K=9 sebagai nilai K final karena selain memberikan akurasi tertinggi, nilai K=9 juga memberikan kompleksitas komputasi yang relatif rendah serta kurva akurasi yang telah mulai stabil (konvergen) mulai dari K=9, sementara pada K=11 hingga K=15 akurasi justru sedikit menurun.

3.6. Perhitungan Euclidean Distance

Setelah data dinormalisasi, jarak antara data uji dan setiap data latih dihitung menggunakan rumus Euclidean Distance sebagai berikut.

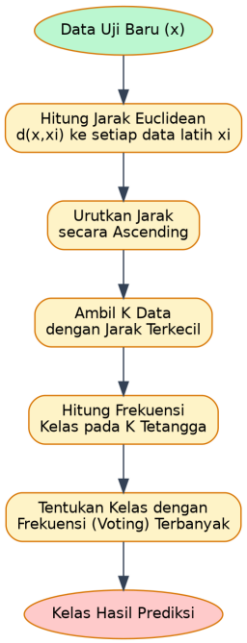
$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

dengan d(x,y) adalah jarak Euclidean antara data uji x dan data latih y, n adalah jumlah atribut/fitur yang digunakan (n=10), serta xi dan yi adalah nilai atribut ke-i dari data x dan y setelah dinormalisasi.

3.7. Proses Klasifikasi KNN

Tahapan klasifikasi menggunakan algoritma KNN pada penelitian ini digambarkan pada Gambar 5, dengan langkah-

langkah sebagai berikut: (1) menghitung jarak Euclidean antara data uji dan seluruh data latih; (2) mengurutkan jarak tersebut secara ascending (dari yang terkecil); (3) memilih K data latih dengan jarak terkecil sebagai tetangga terdekat; (4) menghitung frekuensi kemunculan masing-masing kelas pada K tetangga terdekat tersebut; dan (5) menetapkan kelas dengan frekuensi (*voting*) terbanyak sebagai hasil prediksi kelas data uji.



Gambar 5. Diagram Alur Algoritma Klasifikasi KNN

4. HASIL DAN PEMBAHASAN

4.1. Contoh Perhitungan Manual

Sebagai ilustrasi proses klasifikasi, diambil satu data uji dengan indeks ke-13088 pada dataset, dengan atribut asli sebagaimana ditunjukkan pada Tabel 4.

Tabel 4. Data Uji untuk Contoh Perhitungan Manual

Atribut	Nilai
Usia (tahun)	64
Jenis kelamin	F
Tinggi badan (cm)	148.6
Berat badan (kg)	66.46
Lemak tubuh (%)	41.7
Tekanan darah diastolik	80
Tekanan darah sistolik	120
Kekuatan genggaman (kgf)	16.2
Sit and bend forward (cm)	0.9
Jumlah sit-up (kali)	8
Broad jump (cm)	144
Kelas aktual	**D**

Setelah dilakukan normalisasi Min-Max terhadap 10 atribut numerik data uji tersebut menggunakan parameter min dan maks dari data latih, diperoleh vektor fitur ternormalisasi. Selanjutnya, jarak Euclidean dihitung antara data uji tersebut dengan seluruh data latih menggunakan Persamaan (2). Sebagai contoh perhitungan, jarak antara data uji dengan data latih terdekat pertama (indeks data latih 8429) dihitung sebagai berikut.i

$$d(Uji, Latih) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_{10} - y_{10})^2} = 0.1568$$

Perhitungan yang sama diulang terhadap seluruh 10714 data latih. Hasil jarak kemudian diurutkan secara ascending, dan diambil 9 data dengan jarak terkecil sebagai tetangga terdekat (K=9), sebagaimana ditunjukkan pada Tabel 5.

Tabel 5. Daftar 9 Tetangga Terdekat (K=9)

Peringkat	Indeks Data Latih	Jarak Euclidean	Kelas
1	8429	0.1568	D
2	6486	0.1968	D
3	12357	0.2022	C
4	9898	0.2066	B
5	11128	0.2076	D
6	1557	0.2084	D
7	3478	0.2100	B
8	11279	0.2105	D
9	12166	0.2141	D

Berdasarkan Tabel 5, dari 9 tetangga terdekat diperoleh komposisi kelas sebagai berikut: kelas D sebanyak 6 data, kelas C sebanyak 1 data, kelas B sebanyak 2 data. Karena kelas D memiliki frekuensi (*voting*)

terbanyak, maka data uji tersebut diklasifikasikan ke dalam kelas D. Hasil ini sesuai dengan kelas aktual data uji, yaitu kelas D, sehingga prediksi untuk data uji ini dinyatakan benar.

4.2. Hasil Klasifikasi pada Keseluruhan Data Uji

Proses klasifikasi yang sama diterapkan pada seluruh 2679 data uji menggunakan model KNN dengan K=9. Ringkasan hasil klasifikasi ditunjukkan pada Tabel 6, sedangkan hasil evaluasi menggunakan confusion matrix ditunjukkan pada Tabel 7 dan Gambar 6.

Tabel 6. Rekapitulasi Hasil Klasifikasi pada Data Uji

Kelas	Jumlah Data Uji	Prediksi Benar	Prediksi Salah
A	670	538	132
B	669	257	412
C	670	373	297
D	670	427	243

4.3. Confusion Matrix

Tabel 7. Confusion Matrix Hasil Klasifikasi

Aktual \ Prediksi	A	B	C	D
A	538	91	38	3
B	249	257	155	8
C	81	175	373	41
D	27	56	160	427

Rumus akurasi (*accuracy*) yang digunakan untuk mengevaluasi model adalah sebagai berikut.

$$Accuracy = \frac{\sum Prediksi\ Benar}{Total\ Data\ Uji} \times 100\% \quad (3)$$

Berdasarkan Tabel 7, total prediksi benar diperoleh dengan menjumlahkan nilai diagonal confusion matrix, yaitu 538 + 257 + 373 + 427 = 1595 dari total 2679 data uji. Dengan demikian, akurasi model dihitung sebagai berikut.

$$Accuracy = \frac{1595}{2679} \times 100\% = 59.54\%$$

4.4. Evaluasi per Kelas

Tabel 8. Precision, Recall, dan F1-Score per Kelas

Kelas	Precision	Recall	F1-Score
A	60.11%	80.30%	68.75%
B	44.39%	38.42%	41.19%
C	51.38%	55.67%	53.44%
D	89.14%	63.73%	74.32%

Model KNN dengan K=9 menghasilkan akurasi sebesar 59.54% pada data uji. Berdasarkan Tabel 8, kelas D memiliki precision tertinggi (89.14%), yang berarti ketika model memprediksi kelas D, prediksi tersebut sangat mungkin benar. Sebaliknya, kelas B memiliki precision dan recall terendah (44.39% dan 38.42%), menunjukkan bahwa model paling sering keliru dalam mengenali data pada kelas B.

4.5. Pembahasan

Untuk mengetahui pengaruh jumlah data latih terhadap kinerja model, dilakukan pengujian tambahan dengan mengubah-ubah persentase data latih yang digunakan, yaitu 10%, 25%, 50%, 75%, dan 100% dari total 10.714 data latih yang ada.Sementara itu, jumlah data uji tetap tidak berubah, yakni sebanyak 2.679 data dalam setiap pengujian. Pengambilan data latih dalam bentuk subset dilakukan secara stratified agar setiap kelas (A, B, C, D) memiliki proporsi yang seimbang.Setiap persentase (selain 100%) diuji lima kali dengan membagi data secara acak yang berbeda, sehingga dapat diperoleh nilai akurasi yang lebih stabil secara rata-rata.Hasil pengujian disajikan pada Tabel 9.

Tabel 9. Hasil Klasifikasi KNN (K=9) dengan Variasi Proporsi Data Latih

Persentase Data Latih	Jumlah Data Latih	Jumlah Data Uji	Akurasi (%)
10%	1.071	2.679	54,87
25%	2.678	2.679	55,89

Persentase Data Latih	Jumlah Data Latih	Jumlah Data Uji	Akurasi (%)
50%	5.357	2.679	56,02
75%	8.035	2.679	57,48
100%	10.714	2.679	**58,38**

Hasil pengujian menunjukkan bahwa secara konsisten seising bertambahnya proporsi data latih terpercaya 54,87% pada penggunaan 10% data latih hingga mencapai 58,38% pada penggunaan 100% data latih. Peningkatan yang cukup besar terjadi dari 75% ke 100%, yaitu dari 57,48% menjadi 58,38%.Hal ini menunjukkan bahwa model KNN di dataset ini masih bisa mendapat manfaat dari penambahan data latih dan belum mencapai titik di mana performa tidak lagi meningkat (saturasi) pada rentang proporsi yang diuji.

Untuk melihat pola kesalahan klasifikasi pada setiap proporsi data latih secara lebih jelas, confusion matrix hasil pengujian (rata-rata dari lima kali percobaan, dibulatkan menjadi bilangan bulat terdekat) dapat dilihat pada Tabel 10 hingga Tabel 14.

Tabel 10. Proporsi Data Latih 10%

Akt.\Pred	A	B	C	D
A	**516**	136	17	1
B	259	**304**	89	17
C	115	254	**270**	31
D	28	80	182	**380**

Tabel 11. Proporsi Data Latih 25%

Akt.\Pred	A	B	C	D
A	**531**	116	21	2
B	260	**288**	107	14
C	121	221	**289**	39
D	32	88	160	**390**

Tabel 12. Proporsi Data Latih 50%

Akt.\Pred	A	B	C	D
A	**538**	111	20	1
B	258	**277**	119	14
C	128	198	**301**	43
D	36	87	162	**385**

Tabel 13. Proporsi Data Latih 75%

Akt.\Pred	A	B	C	D
A	**548**	107	15	1
B	254	**291**	112	12
C	124	194	**306**	46
D	35	88	153	**394**

Tabel 14. Proporsi Data Latih 100%

Akt.\Pred	A	B	C	D
A	**550**	106	13	1
B	257	**285**	116	11
C	122	189	**324**	35
D	23	93	149	**405**

Berdasarkan Tabel 10 hingga Tabel 14, tampak bahwa nilai diagonal (prediksi benar) untuk kelas A dan D terus-menerus lebih besar dibandingkan kelas B dan C pada setiap proporsi data latih. Misalnya pada proporsi 100%, kelas A dan D masing-masing memiliki 550 dan 405 prediksi yang benar dari total 670 data uji, sedangkan kelas B dan C hanya mendapatkan 285 dan 324 prediksi yang benar. Kesalahan dalam mengklasifikasikan kelas B dan C juga terasa saling bertukar: sebagian besar prediksi yang salah untuk kelas B jatuh ke kelas C (dan sebaliknya), bukan ke kelas A atau D.Hal ini menunjukkan bahwa kedua kelas ini paling sulit dibedakan oleh model dibandingkan dengan pasangan kelas lainnya. Pola ini terus

muncul di semua bagian data latih yang diuji, artinya menambah data latih tidak mengubah pola kesalahan tersebut, tetapi hanya membuat jumlah prediksi yang benar bertambah secara perlahan.

Temuan ini memperkuat hasil penilaian per kelas pada Tabel 8, yang menunjukkan bahwa model KNN mampu mengenali kelas ekstrem, yakni kelas A (kebugaran terbaik) dan kelas D (kebugaran terendah), dengan tingkat recall yang lebih tinggi dibandingkan kelas B dan C yang berada di tengah. Hal ini bisa dijelaskan karena data di kelas A dan D biasanya memiliki ciri-ciri biometrik yang lebih ekstrem dan berbeda secara nyata, seperti genggaman yang sangat kuat atau sangat lemah, sehingga jarak Euclidean antara data tersebut dengan data dari kelas lain biasanya lebih jauh dan lebih mudah dibedakan. Sebaliknya, kelas B dan C memiliki rentang nilai atribut biometrik yang tumpang tindih, sehingga data dari kelas B seringkali memiliki tetangga terdekat yang berasal dari kelas C dan sebaliknya, yang menyebabkan kesalahan klasifikasi lebih tinggi pada kedua kelas tersebut. Fenomena serupa mengenai kesulitan mengelompokkan data ke dalam kelas yang saling dekat atau tumpang tindih juga pernah ditemukan dalam penelitian sebelumnya tentang klasifikasi berbasis KNN.

Dengan demikian, menambah jumlah data latih masih memberikan manfaat positif bagi akurasi model, tetapi masih ada masalah utama yaitu kesamaan karakteristik atribut antara kelas B dan C.Hal ini tetap menjadi hambatan utama karena kesalahan klasifikasi pada kedua kelas tersebut lebih disebabkan oleh kemiripan nilai atribut biometriknya, bukan hanya karena jumlah data latih yang terbatas.

4.6. Keterbatasan Penelitian

Meskipun model KNN yang dibangun menunjukkan performa cukup baik, penelitian ini memiliki beberapa keterbatasan:

1. Tidak ada validasi silang. Evaluasi hanya menggunakan single hold-out split 80:20, sehingga akurasi 59.54% berisiko bervariasi tergantung komposisi data yang terbentuk secara kebetulan. Penggunaan k-fold cross-validation akan memberikan estimasi performa yang lebih robust.
2. Tidak ada perbandingan algoritma lain. KNN tidak dibandingkan dengan Random Forest, SVM, atau Decision Tree pada dataset yang sama, sehingga klaim performa "cukup baik" masih bersifat relatif.
3. Anomali data tekanan darah. Ditemukan nilai minimum 0 pada tekanan darah sistolik/diastolik (Tabel 1) yang secara fisiologis tidak valid, kemungkinan missing value ter-encode sebagai nol. Nilai ini lolos dari pemeriksaan missing value standar, ikut ternormalisasi Min-Max, dan berpotensi mendistorsi perhitungan jarak Euclidean.
4. Tidak ada pembobotan atribut. Seluruh atribut biometrik diberi bobot sama dalam perhitungan jarak tanpa mempertimbangkan relevansinya terhadap klasifikasi, meski pendekatan ini sudah disebutkan sebagai arah pengembangan pada bagian Pembahasan namun belum diuji.
5. Analisis *overlap* kelas B-C masih kualitatif. Belum ada bukti kuantitatif (misalnya boxplot distribusi per-atribut per-kelas) yang menunjukkan atribut mana yang paling tumpang tindih antara kedua kelas.
6. Atribut gender tidak dimanfaatkan sebagai fitur. Jenis kelamin memang di-encode secara numerik pada tahap praproses, namun tidak diikutsertakan dalam perhitungan jarak Euclidean maupun proses klasifikasi. Padahal, secara fisiologis capaian grip force, sit-up, dan broad jump dapat berbeda signifikan antara laki-laki dan perempuan, sehingga pengabaian atribut ini berpotensi mengurangi kemampuan model dalam membedakan kelas kebugaran yang saling tumpang tindih.
7. Hanya satu metrik jarak. Penelitian hanya menggunakan Euclidean Distance tanpa pembanding seperti Manhattan Distance, meski literatur rujukan [5], [6] menunjukkan pemilihan metrik jarak dapat memengaruhi hasil klasifikasi KNN secara signifikan.

Keterbatasan ini membuka peluang pengembangan lanjutan, terutama validasi model yang lebih robust, pembersihan data yang lebih menyeluruh, kejelasan perlakuan atribut kategorikal, serta eksplorasi metode dan metrik jarak lain untuk meningkatkan akurasi klasifikasi.

5. KESIMPULAN

1. Algoritma K-Nearest Neighbor (KNN) berhasil diimplementasikan untuk mengklasifikasikan tingkat kebugaran tubuh ke dalam empat kelas (A, B, C, D) berdasarkan sepuluh atribut biometrik pada dataset sebanyak 13393 data, melalui tahapan praproses (normalisasi Min-Max), pembagian data (80:20), perhitungan jarak Euclidean, dan voting mayoritas kelas tetangga terdekat.
2. Nilai K memberikan pengaruh signifikan terhadap akurasi klasifikasi. Akurasi meningkat seiring bertambahnya K dan mencapai puncaknya pada K=9 dengan akurasi tertinggi 59.54%, sebelum cenderung menurun pada K=11 hingga K=15. Penelitian ini memilih K=9 sebagai nilai optimal karena memberikan akurasi tertinggi sekaligus kompleksitas komputasi yang relatif rendah.
3. Model KNN dengan K=9 menghasilkan akurasi sebesar 59.54% pada data uji. Model menunjukkan performa lebih baik dalam mengenali kelas ekstrem (A dan D) dibandingkan kelas menengah (B dan C) yang memiliki karakteristik biometrik yang saling tumpang tindih.
4. Kelebihan algoritma KNN pada penelitian ini adalah kesederhanaan dan interpretabilitas hasil, sedangkan kekurangannya terletak pada sensitivitas terhadap nilai K dan penurunan performa pada kelas yang saling tumpang tindih. Pengembangan lebih lanjut dapat dilakukan melalui penerapan pembobotan atribut atau varian KNN lainnya untuk meningkatkan akurasi klasifikasi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan, arahan, dan fasilitas selama proses penelitian dan penulisan artikel ini.

DAFTAR PUSTAKA

[1] M. Muchtar and R. A. Muchtar, “Perbandingan Metode KNN dan SVM dalam Klasifikasi Kematangan Buah Mangga Berdasarkan Citra HSV dan Fitur Statistik,” Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 2, pp. 876–884, 2024, doi: 10.23960/jitet.v12i2.4010.

[2] T. Aprilia, “Klasifikasi Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor dan Metode Naive Bayes,” SATESI: Jurnal Sains Teknologi dan Sistem Informasi, vol. 4, no. 2, pp. 156–163, 2024, doi: 10.54259/satesi.v4i2.3167.

[3] N. T. Ujianto, Gunawan, H. Fadillah, A. P. Fanti, A. D. Saputra, and I. G. Ramadhan, “Penerapan Algoritma K-Nearest Neighbors (KNN) untuk Klasifikasi Citra Medis,” IT-Explore: Jurnal Penerapan Teknologi Informasi dan Komunikasi, vol. 4, no. 1, pp. 33–43, 2025, doi: 10.24246/itexplore.v4i1.2025.pp33-43.

[4] N. Khairunisa et al., “Analisis Perbandingan Algoritma CNN dan YOLO dalam Mengidentifikasi Kerusakan Jalan,” vol. 12, no. 3, 2024.

[5] S. Buana Prameswary, D. Agustianingsih, and A. G. Nugraheni, “Perbandingan Implementasi Metode K-Nearest Neighbor Menggunakan Jarak Euclidean dan Manhattan pada Analisa Klasifikasi Penyakit Anemia,” Technology and Informatics Insight Journal, vol. 4, no. 1, pp. 7–13, 2025.

[6] P. P. Allorerung, A. Erna, M. Bagussahrir, and S. Alam, “Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit,” JISKA (Jurnal Informatika Sunan Kalijaga), vol. 9, no. 3, pp. 178–191, 2024, doi: 10.14421/jiska.2024.9.3.178-191.

[7] F. Putra, H. F. Tahiyat, and R. M. Ihsan, “Application of K-Nearest Neighbor Algorithm Using Wrapper as Preprocessing for Determination of Human Weight Information,” MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 1, pp. 273–281, 2024.

[8] A. Yogianto, A. Homaidi, and Z. Fatah, “Implementasi Metode K-Nearest Neighbors (KNN) untuk Klasifikasi Penyakit Jantung,” G-Tech: Jurnal Teknologi Terapan, vol. 8, no. 3, pp. 1720–1728, 2024, doi: 10.33379/gtech.v8i3.4495.

[9] S. A. Bugis, C. Cakra, A. M. Islah, M. S. Said, D. Suarna, and M. S. Said, “Implementasi

Algoritma K-Nearest Neighbor (K-NN) dalam Perancangan Alat Pendeteksi Tingkat Kesegaran Daging," Simtek: Jurnal Sistem Informasi dan Teknik Komputer, vol. 9, no. 1, 2024.

[10] J. Lemantara and T. Lusiani, "Analisis Prediksi Penyakit Diabetes pada Wanita Menggunakan Metode Naïve Bayes dan K-Nearest Neighbor," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4911.

[11] W. Aprilliandhika and F. F. Abdulloh, "Comparison of K-Nearest Neighbor and Support Vector Machine Algorithm Optimization with Grid Search CV on Stroke Prediction," Jurnal Teknik Informatika (JUTIF), vol. 5, no. 4, pp. 991–1000, 2024, doi: 10.52436/1.JUTIF.2024.5.4.1951.

[12] A. Andrian and K. Harefa, "Implementasi Algoritma K-Nearest Neighbor untuk Klasifikasi Risiko Stunting Balita Berbasis Web di Posyandu Belimbing," Jurnal Komputer Teknologi Informasi Sistem Komputer (JUKTISI), vol. 5, no. 1, pp. 1091–1099, 2026, doi: 10.62712/juktisi.v5i1.1247.

[13] D. Heksaputra, "Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 3, pp. 812–818, 2024, doi: 10.57152/malcom.v4i3.1268.

[14] I. M. Anggita, M. Naufal, and F. Alzami, "Implementasi Grid Search CV KNN dengan Preprocessing Z-Score Outlier Removal untuk Sistem Prediksi Risiko Kehamilan," Building of Informatics, Technology and Science (BITS), vol. 7, no. 2, pp. 1202–1213, 2025, doi: 10.47065/bits.v7i2.8206.

[15] A. Yaqin, D. Kurniawan, and J. Zeniarja, "Optimasi Algoritma K-Nearest Neighbors Menggunakan GridSearchCV untuk Klasifikasi Penyakit Diabetes," Infotekmesin, vol. 16, no. 1, pp. 75–84, 2025, doi: 10.35970/INFOTEKMESIN.V16I1.2557.