

Perbandingan Matriks jarak pada Algoritma K-NN untuk Prediksi Penyakit Diabetes

Comparison of Distance Matrices in the K-NN Algorithm for Predicting diabetes

Galih Hendro Martono^{1*}, Neny Sulistianingsih²

Universitas Bumigora, Mataram, Indonesia^{1,2}

Article Info

Article history:

Diterima 27 Desember 2023

Direvisi 11 Januari 2024

Disetujui 12 Januari 2024

Kata Kunci:

Cosine Similarity

Euclidean Distance

K-Nearest Neighbor

Manhattan Distance

Minkowski Distance

Pearson Correlation Coefficient

ABSTRAK

Penelitian ini bertujuan untuk mengetahui kinerja berbagai metrik jarak dalam klasifikasi K-Nearest Neighbors (KNN) untuk diagnosis penyakit diabetes. Studi ini menggunakan beberapa metode penghitungan jarak seperti *Euclidean Distance*, *Manhattan Distance*, *Minkowski Distance*, *Cosine Similarity*, dan *Pearson Correlation Coefficient*. Data yang digunakan merupakan data pasien yang menderita penyakit diabetes. Evaluasi dilakukan dengan menggunakan evaluation measure yaitu presisi, sensitivitas, Skor F1, dan akurasi. Hasil penelitian menunjukkan bahwa pengukuran jarak dengan *Manhattan Distance* lebih unggul dibandingkan dengan metode pengukurannya lainnya yaitu memiliki nilai presisi 0,952, nilai sensitivitas sebesar 0,950, nilai skor F1 sebesar 0,939, dan akurasi sebesar 0,950. Di sisi lain, pengukuran jarak dengan *Euclidean Distance* dan *Minkowski Distance* secara konsisten memberikan hasil yang sama yaitu 0,932, 0,940, 0,930 dan 0,940 untuk presisi, sensitivitas, F1-Score dan akurasi berurut-turut. Berbeda dengan, metode *Cosine Similarity*, meskipun akurasinya tinggi yaitu sebesar 0,905 namun nilai evaluation measure yang lainnya seperti presisi, sensitivitas, dan F1-Score menunjukkan nilai yang lebih rendah yaitu 0,865, 0,905, dan 0,881 secara berurut. Sedangkan, pengukuran jarak dengan *Pearson Correlation Coefficient*, meskipun memiliki akurasi yang bagus yaitu 0,900 namun menunjukkan kinerja presisi, sensitivitas, dan skor F1 yang rendah yaitu 0,544, 0,517, dan 0,518 secara berurut.

ABSTRACT

This study aims to determine the performance of various distance metrics in K-Nearest Neighbors (KNN) classification for diabetes diagnosis. This study uses several distance calculation methods such as *Euclidean Distance*, *Manhattan Distance*, *Minkowski Distance*, *Cosine Similarity*, and *Pearson Correlation Coefficient*. The data used is data from patients suffering from diabetes. Evaluation is carried out using evaluation measures, namely precision, sensitivity, F1 score and accuracy. The results of the research show that measuring distance with *Manhattan Distance* is superior to other measurement methods, namely having a precision value of 0.952, a sensitivity value of 0.950, an F1 score value of 0.939, and an accuracy of 0.950. On the other hand, distance measurements with *Euclidean Distance* and *Minkowski Distance* consistently provide the same results, namely 0.932, 0.940, 0.930 and 0.940 for precision, sensitivity, F1-Score and accuracy respectively. In contrast to the *Cosine Similarity* method, although the accuracy is high, namely 0.905, other evaluation measure values such as precision, sensitivity and F1-Score show lower values, namely 0.865, 0.905 and 0.881 respectively. Meanwhile, distance measurement using the *Pearson Correlation Coefficient*, although it has good accuracy, namely 0.900, shows low precision, sensitivity and F1 score performance, namely 0.544, 0.517 and 0.518 respectively.



Copyright ©2022 JOMI: Journal of Millennial Informatics. This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

Penulis Korespondensi:

Galih Hendro Martono

Program Studi Ilmu Komputer, Fakultas Teknik

Universitas Bumigora, Mataram, Indonesia.

Email: galih.hendro.m@universitasbumigora.ac.id.

1 PENDAHULUAN

Dalam era medis yang terus berkembang dengan cepat, peningkatan jumlah data medis yang tersedia telah menarik perhatian besar terhadap pengembangan model prediktif yang efisien untuk menganalisis data kesehatan. Beberapa metode seperti *Support Vector Machine* (SVM) [1], *Logistic Regression* [2], *Decision Tree* [3], algoritma *ensemble* [4], [5], algoritma *Boost* [6], dan *Deep Learning* [7]. Salah satu pendekatan yang banyak digunakan dalam dunia kedokteran adalah penerapan algoritma *K-Nearest Neighbors* (KNN), yang sering diterapkan untuk klasifikasi dan regresi berdasarkan data medis.

Beberapa penelitian yang menggunakan KNN seperti yang dilakukan oleh [9]–[14] menggunakan metode KNN pada data medis. Penelitian lain yang dilakukan oleh [14] untuk penyakit Alzheimer, [15] untuk penyakit Parkinson, [12] untuk penyakit ginjal kronis, [13] untuk epilepsi, dan [9] untuk penyakit Tuberculosis (TBC). Berbeda dengan penelitian sebelumnya, penelitian yang dilakukan oleh [10] fokus penelitian adalah untuk variasi dari $n_{neighbor}$ yang digunakan pada penelitian dengan dataset pasien dari Amerika. Dataset pasien penderita penyakit diabetes tersebut hanya mencakup gula darah pasien yang meliputi HbA1c level, gula darah puasa dan gula darah acak. Selain itu, pada penelitian pengukuran jarak juga dibandingkan antara KNN *classic* yang menggunakan *euclidean Distance*, *cosine similarity* dan *cubic Distance*. Sedangkan pada penelitian yang dilakukan oleh [11] dikembangkan metode KNN dengan menggunakan konsep fuzzy yang diterapkan pada pasien wanita hamil yang menderita diabetes. Konsep fuzzy yang diterapkan pada pengembangan KNN ini menggunakan klasifier *uncertainty* yang berkaitan dengan tingkat kepercayaan, menangani fungsi keanggotaan dan hukum operasi (*operation law*) atau aturan yang digunakan dalam pemrosesan.

Penelitian sebelumnya berfokus pada penerapan metode KNN pada berbagai dataset dan penelitian lainnya berfokus pada perhitungan ketetanggaan pada metode KNN, maka pada penelitian ini, fokus penelitian adalah membandingkan berbagai metrik pengukuran jarak yang terdapat pada KNN. Beberapa metrik pengukuran jarak yang dibandingkan dalam penelitian ini antara lain *Euclidean Distance*, *Manhattan Distance*, *Minkowski Distance*, *Cosine Similarity*, dan *Pearson Correlation Coefficient*. Metrik pengukuran jarak memainkan peran krusial dalam menentukan sejauh apa dua data dapat dianggap mirip atau berbeda, sehingga pemilihan metrik yang sesuai dapat memengaruhi hasil akhir dari model KNN. Dalam penelitian digunakan data penyakit diabetes sebagai contoh kasus.

2 METODE PENELITIAN

Penelitian ini dilakukan dalam beberapa tahap. Tahap awal dimulai dengan mengumpulkan data yang digunakan pada penelitian ini. Sumber data yang digunakan adalah informasi diagnosis penyakit diabetes yang diperoleh dari www.kaggle.com. Jumlah data pasien pada data diabetes yang digunakan pada penelitian ini adalah 1000 data pasien. Data ini mencakup berbagai kriteria seperti usia (*age*), tekanan darah (*hypertension*), penyakit jantung (*heart disease*), indeks massa tubuh (BMI), tingkat HbA1c (*HbA1c level*), kadar glukosa darah (*blood glucose level*), dan status diabetes. Selain kriteria tersebut terdapat juga kriteria jenis kelamin (*gender*) dan riwayat merokok (*smoking history*). Akan tetapi kedua kriteria tersebut bersifat kategorikal. Setiap data menyertakan label yang mengindikasikan diagnosis diabetes yang dialami oleh pasien tertentu. Tabel 1 menyajikan beberapa informasi yang dapat diperoleh dari data tersebut.

Tabel 1. Informasi data

Kriteria	Mean	Standar Deviasi	Median	Modus	Min	Max
Age	41.274	22.910	42.000	80.000	0.080	80.000
Hypertension	0.068	0.252	0.000	0.000	0.000	1.000
Heart_disease	0.032	0.176	0.000	0.000	0.000	1.000
BMI	26.959	6.786	27.320	27.320	10.300	69.370
HbA1c level	5.560	1.098	5.800	6.600	3.500	9.000
Blood Glucose Level	138.419	38.704	145.000	130.000	80.000	300.000
Diabetes	0.082	0.275	0.000	0.000	0.000	1.000

Lebih lanjut, berdasarkan data tersebut dilakukan proses pengecekan data kosong dan data duplikat. Diketahui bahwa pada data diabetes yang digunakan tidak terdapat data kosong dan data duplikat sehingga *preprocessing* data tidak perlu dilakukan dan dapat digunakan untuk tahap selanjutnya.

Tahap penelitian dilanjutkan dengan proses prediksi diagnosis penyakit diabetes dengan menggunakan metode *K-Nearest Neighbor* (KNN). Metode KNN merupakan metode klasifikasi nonparametrik yang sederhana namun efektif dalam banyak kasus akan tetapi sangat bergantung pada nilai parameter $n_{neighbor}$ yang digunakan pada saat penggunaan metode KNN tersebut [13]. Pada metode KNN, pengukuran jarak yang biasa digunakan adalah *Euclidean Distance*. Formula yang digunakan untuk mengukur jarak dengan *Euclidean Distance* dapat dilihat pada formula berikut.

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

dengan,

d : Jarak
 x : Titik x
 y : Titik y
 n : Jumlah kriteria yang digunakan
 x_i : Nilai kriteria ke- i dari data pertama yang dibandingkan
 y_i : Nilai kriteria ke- i dari data kedua yang dibandingkan

Pada penelitian ini beberapa metode pengukuran jarak lainnya digunakan selain dengan *Euclidean Distance*. Metode pengukuran jarak tersebut antara lain; *Manhattan Distance*, *Minkowski Distance*, *Cosine Similarity* dan *Pearson Correlation Coefficient*. Pengukuran jarak dengan menggunakan *Manhattan Distance* tidak jauh berbeda berbeda jika dibandingkan dengan *euclidean Distance*. Pada pengukuran jarak dengan *Manhattan Distance*, jarak antara dua titik pada sebuah vektor dimensi N dihitung dengan menjumlahkan selisih mutlak antara ukuran-ukuran pada seluruh dimensi dua titik yang dibandingkan. Formula untuk *Manhattan Distance* adalah sebagai berikut.

$$d(x, y) = \sum_{i=1}^n |x_{1i} - x_{2i}| \quad (2)$$

dengan,

d : Jarak
 x : Titik x
 y : Titik y
 n : Jumlah kriteria yang digunakan
 x_{1i}, x_{2i} : nilai kriteria pada dimensi ke- i dari dua titik data yang dibandingkan

Perhitungan jarak dengan metode *Minkowski Distance* merupakan pengembangan dari metode *Euclidean Distance* dan *Manhattan Distance*. Sehingga formula untuk pengukuran jarak dengan *Minkowski Distance* adalah sebagai berikut.

$$d(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^r \right)^{\frac{1}{r}} \quad (3)$$

dengan,

d : Jarak
 x : Titik x
 y : Titik y
 n : Jumlah kriteria yang digunakan
 x_i, y_i : nilai kriteria pada dimensi ke- i dari titik x dan y
 r : parameter yang dapat disesuaikan

Pada perhitungan jarak dengan *Minkowski Distance*, nilai parameter $r = 1$ maka formula perhitungan *Minkowski Distance* dapat menjadi perhitungan jarak *Manhattan Distance*. Sedangkan jika nilai parameter $r = 2$ maka perhitungan *Minkowski Distance* menjadi perhitungan jarak *Euclidean Distance*.

Pada pengukuran jarak dengan menggunakan *Cosine Similarity*, yang diukur adalah kemiripan antara dua vektor dalam ruang berdimensi banyak yang mengukur kosinus sudut antara keduanya. Sehingga formula yang digunakan untuk menghitung jarak dengan *Cosine Similarity* adalah sebagai berikut.

$$d(X, Y) = \frac{X \cdot Y}{\|X\| \times \|Y\|} \quad (4)$$

dengan,

d : Jarak
 X : Vektor X
 Y : Vektor Y

Sedangkan pada pengukuran dengan menggunakan *Pearson Correlation Coefficient*, jarak pada dua titik diukur dengan metrik statistik yang digunakan untuk mengukur sejauh mana dua titik tersebut berkorelasi linear satu sama lain. Berdasarkan hal tersebut, fomula dari perhitungan jarak dengan menggunakan *Pearson Correlation Coefficient* adalah sebagai berikut.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (5)$$

dengan,

r : Nilai *Pearson Correlation Coefficient*
 x_i, y_i : nilai-nilai pada variabel x dan y untuk observasi ke- i .
 $\bar{x}, \bar{y}$: rata-rata dari variabel x dan y secara berturut-turut

Nilai dari r akan berkisar dari -1 hingga 1 dengan nilai 1 yang menggambarkan linieritas yang positif antara kedua variabel yang dibandingkan, sedangkan kebalikannya jika bernilai -1 maka linieritas tersebut bersifat negatif dan 0 berarti tidak ada hubungan antara kedua variabel yang dibandingkan.

Setelah analisis metode KNN dengan berbagai matrik jarak dilakukan, evaluasi dilakukan terhadap luaran yang dihasilkan masing-masing metode KNN dengan berbagai matrik tersebut. Evaluasi yang dilakukan dengan mengukur presisi, sensitivitas, F1-Score dan akurasi dari hasil deteksi metode KNN tersebut jika dibandingkan total diagnosis yang ada pada data yang digunakan. Formula presisi, sensitivitas, F1-Score dan akurasi yang digunakan pada penelitian ini adalah sebagai berikut.

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

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

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

$$\text{Akurasi} = \frac{TP + TN + FP + FN}{TP + TN + FP + FN} \quad (9)$$

dengan,

- TP : Prediksi benar pada kelas positif (*True Positive*)
- TN : Prediksi benar kelas negatif (*True Negative*).
- FP : Prediksi salah kelas positif (*False Positive*)
- FN : prediksi salah kelas Negatif (*False Negative*)

3 HASIL DAN ANALISIS

Implementasi dari metode KNN untuk prediksi penyakit diabetes pada pasien dilakukan dengan beberapa parameter metode yang ditentukan sebelumnya. Pada implementasi KNN ini nilai $n_{neighbor}$ yang digunakan adalah 3. Sehingga saat pengujian dengan matrik pengukuran jarak yang berbeda, nilai $n_{neighbor} = 3$ untuk semua matrik pengukuran jarak yang digunakan.

3.1. Hasil

Diagnosis penyakit diabetes dilakukan dengan metode KNN dengan berbagai matrik pengukuran jarak yang digunakan. Evaluasi yang dilakukan dengan menggunakan ukuran presisi, sensitivitas, F1-score dan akurasi menunjukkan bahwa walaupun pada metode K-NN, *euclidean Distance* merupakan metode pengukuran jarak yang standar digunakan pada metode K-NN akan tetapi jika dibandingkan dengan metode lainnya, pengukuran jarak dengan menggunakan *Manhattan Distance* menghasilkan hasil evaluasi yang lebih dominan jika dibandingkan dengan *Euclidean Distance*. Kesimpulan ini ditunjukkan dengan hasil implementasi metode *Manhattan Distance* yang lebih tinggi nilai evaluasi dengan menggunakan ukuran presisi, sensitivitas, F1-score dan akurasi yaitu sebesar 0.952, 0.950, 0.939 dan 0.950, berturut-turut. Nilai ini merupakan yang paling tinggi untuk matrik pengukuran jarak yang digunakan pada penelitian ini. Diikuti oleh *Minkowski Distance* dan *Euclidean Distance* dengan nilai ukuran evaluasi yang sama, yaitu 0.932, 0.940, 0.930 dan 0.940 untuk presisi, sensitivitas, F1-Score dan akurasi berturut-turut. Diikuti oleh *Cosine similarity* dan terakhir oleh *Pearson Correlation Coefficient*. Hasil implementasi metode K-NN dengan berbagai matrik pengukuran jarak dapat dilihat pada Tabel 2.

Tabel 2. Hasil KNN dengan berbagai matrik pengukuran jarak

Matrik Pengukuran Jarak	Presisi	Sensitivitas	F1-Score	Akurasi
<i>Euclidean Distance</i>	0.932	0.940	0.930	0.940
<i>Manhattan Distance</i>	0.952	0.950	0.939	0.950
<i>Minkowski Distance</i>	0.932	0.940	0.930	0.940
<i>Cosine Similarity</i>	0.865	0.905	0.881	0.905
<i>Pearson Correlation Coefficient</i>	0.544	0.517	0.518	0.900

3.2. Analisis

Analisis performa metrik pengukuran jarak menunjukkan hasil yang menarik. *Euclidean Distance* menunjukkan kinerja yang baik dengan presisi, sensitivitas, dan F1-score sekitar 0.930 hingga 0.940, serta akurasi mencapai 0.940. Di sisi lain, *Manhattan Distance* memberikan hasil yang sangat baik dengan presisi, sensitivitas, dan F1-score sekitar 0.95, dan akurasi mencapai 0.950. *Minkowski Distance* menunjukkan hasil serupa dengan *Euclidean Distance*, menegaskan konsistensinya dengan presisi, sensitivitas, dan F1-score sekitar 0.930 hingga 0.940, dan akurasi mencapai 0.940. Namun, *Cosine Similarity*, meskipun memiliki akurasi tinggi sekitar 0.905, menunjukkan presisi, sensitivitas, dan F1-score yang lebih rendah dibandingkan metrik lainnya, berkisar antara 0.865 hingga 0.881. Selain itu, *Pearson Correlation Coefficient*, meskipun memiliki akurasi tinggi sekitar 0.900, menunjukkan kinerja yang rendah dalam presisi, sensitivitas, dan F1-score, terutama presisi yang hanya sekitar 0.544.

Dalam memilih metrik pengukuran jarak, dapat disimpulkan bahwa *Manhattan Distance* menonjol sebagai pilihan yang optimal dengan kinerja yang tinggi di semua metrik evaluasi. Sementara *Euclidean Distance* dan *Minkowski Distance* memberikan hasil yang konsisten dan dapat dianggap baik. Meskipun *Cosine Similarity* menunjukkan akurasi yang tinggi, perlu diperhatikan bahwa metrik ini kurang unggul dalam presisi, sensitivitas, dan F1-score. Oleh karena itu, pemilihan *Pearson Correlation Coefficient* sebaiknya diperhitungkan kembali, mengingat presisinya yang rendah, meskipun memiliki akurasi yang tinggi. Kesimpulan ini menyoroti pentingnya mempertimbangkan berbagai metrik evaluasi dan tujuan spesifik dalam memilih metrik pengukuran jarak yang sesuai dengan karakteristik data yang dimiliki. Metode pengukuran jarak dengan Manhattan lebih baik jika dibandingkan dengan pengukuran jarak lainnya karena pengukuran dengan Manhattan cenderung lebih tahan terhadap outlier. Hal ini karena dalam pengukuran jarak dengan Manhattan, jarak dihitung sebagai jumlah perbedaan absolut antara koordinat titik-titik yang bersangkutan. Analisis ini menyoroti pentingnya mempertimbangkan seluruh metrik evaluasi untuk mendapatkan gambaran komprehensif tentang kinerja metrik pengukuran jarak yang digunakan. Namun demikian, pemilihan metode pengukuran jarak haruslah disesuaikan dengan karakteristik data dan masalah yang dihadapi.

Keberhasilan evaluasi model tidak hanya bergantung pada akurasi semata, melainkan juga memerlukan penilaian yang menyeluruh, termasuk presisi, sensitivitas, dan F1-score. Dalam mengoptimalkan performa, perlu diingat bahwa terdapat kompromi atau keseimbangan yang harus diperhitungkan antara berbagai metrik evaluasi. Sehingga, dalam pemilihan metrik pengukuran jarak, pertimbangan harus disesuaikan dengan tujuan khusus dan karakteristik data yang sedang dihadapi.

4 KESIMPULAN

Dari penelitian yang dilakukan terkait dengan analisis performa metrik pengukuran jarak pada metode K-NN untuk prediksi penyakit diabetes menunjukkan bahwa pilihan optimal bervariasi tergantung pada kebutuhan spesifik dan karakteristik data. Meskipun *Manhattan Distance* menonjol dengan kinerja tinggi di semua metrik evaluasi, *Euclidean Distance* dan *Minkowski Distance* juga memberikan hasil yang konsisten. *Cosine Similarity*, meskipun memiliki akurasi tinggi, menunjukkan keterbatasan dalam presisi, sensitivitas, dan F1-score. *Pearson Correlation Coefficient*, meskipun memiliki akurasi yang tinggi, menunjukkan kinerja yang kurang memuaskan, terutama dalam presisi.

Penelitian sebelumnya berfokus pada implementasi KNN untuk berbagai kasus dan parameter ketetanggaan pada KNN, sedangkan penelitian ini berfokus pada perbandingan metode pengukuran jarak yang terdapat pada KNN. Penelitian tentang pengukuran jarak pada KNN memiliki potensi kontribusi yang besar dalam pengembangan *machine learning* dan *data mining*. Pengembangan selanjutnya dari metode pengukuran jarak pada KNN dapat dilakukan untuk penanganan data yang tidak terstruktur. Selain itu, penelitian lebih lanjut juga dapat mempertimbangkan penggunaan metrik pengukuran jarak lainnya yang belum dieksplorasi dalam penelitian ini. Dengan mencakup variasi metrik pengukuran jarak, penelitian masa depan dapat memberikan pemahaman yang lebih komprehensif tentang bagaimana berbagai metode pengukuran jarak memengaruhi kinerja model. Lebih lanjut untuk penelitian di masa depan, disarankan untuk mengeksplorasi variasi parameter $n_{neighbor}$ pada model K-NN. Dengan menguji berbagai nilai $n_{neighbor}$, penelitian dapat memberikan wawasan lebih lanjut tentang bagaimana performa model KNN bereaksi terhadap perubahan jumlah tetangga yang dipertimbangkan. Langkah ini dapat membantu peneliti dan praktisi untuk menyesuaikan model KNN dengan lebih baik sesuai dengan kebutuhan spesifik dataset.

REFERENSI

- [1] A. Y. Saleh, C. K. Chin, V. Penshie, and H. R. H. Al-Absi, "Lung cancer medical images classification using hybrid CNN-SVM," *Int. J. Adv. Intell. Informatics*, vol. 7, no. 2, pp. 151–162, 2021, doi: 10.26555/ijain.v7i2.317.
- [2] C. M. Bertonecelli, P. Altamura, E. R. Vieira, S. S. Iyengar, F. Solla, and D. Bertonecelli, "PredictMed: A logistic regression-based model to predict health conditions in cerebral palsy," *Health Informatics J.*, vol. 26, no. 3, pp. 2105–2118, 2020, doi: 10.1177/1460458219898568.
- [3] A. S. Abdullah, S. Selvakumar, P. Karthikeyan, and M. Venkatesh, "Comparing the Efficacy of Decision Tree and its Variants using Medical Data," *Indian J. Sci. Technol.*, vol. 10, no. 18, pp. 1–8, 2017, doi: 10.17485/ijst/2017/v10i18/111768.
- [4] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, vol. 22, no. 19, pp. 1–37, 2022, doi: 10.3390/s22197268.
- [5] P. Jiang, H. Suzuki, and T. Obi, "XAI-based cross-ensemble feature ranking methodology for machine learning models," *Int. J. Inf. Eng.*, vol. 15, no. 4, pp. 1759–1768, 2023.
- [6] M. J. Raihan, M. A. M. Khan, S. H. Kee, and A. Al Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-33525-0.
- [7] H. Sun, Z. Liu, G. Wang, W. Lian, and J. Ma, "Intelligent analysis of medical big data based on deep learning," *IEEE*

- Access, vol. 7, pp. 142022–142037, 2019, doi: 10.1109/ACCESS.2019.2942937.
- [8] D. Putra, *Pengolahan Citra Digital*. Penerbit ANDI.
 - [9] M. Muthahir, T. T. S. Sibarani, and Al-Khowarizmi, “Analysis K-Nearest Neighbors (KNN) in Identifying Tuberculosis Disease (Tb) By Utilizing Hog Feature Extraction,” *Int. Comput. Sci. Inf. Technol. J.*, vol. 1, no. 1, pp. 33–38, 2020.
 - [10] A. Ali, M. Alrubei, La. F. M. Hassan, A.-J. Mohannad, and S. Abdulwahed, “Diabetes Classification Based on KNN,” *IIUM Eng. J.*, vol. 21, no. 1, pp. 173–181, 2020.
 - [11] H. Salem, M. Y. Shams, O. M. Elzeki, M. A. Elfattah, J. F. Al-amri, and S. Elnazer, “Fine-Tuning Fuzzy KNN Classifier Based on Uncertainty Membership for the Medical Diagnosis of Diabetes,” *Appl. Sci.*, vol. 12, no. 3, pp. 1–26, 2022, doi: 10.3390/app12030950.
 - [12] C. Priyadharshini, K. Sanjeev, M. Vignesh, N. Saravanan, and M. Somu, “KNN based Detection and Diagnosis of Chronic Kidney Disease,” vol. 25, no. 4, pp. 2870–2877, 2021.
 - [13] Z. Wang, J. Na, and B. Zheng, “An Improved kNN Classifier for Epilepsy Diagnosis,” *IEEE Access*, vol. 8, pp. 100022–100030, 2020, doi: 10.1109/ACCESS.2020.2996946.
 - [14] M. Balamurugan, A. Nancy, and S. Vijaykumar, “Alzheimer’s disease diagnosis by using dimensionality reduction based on KNN Classifier,” *Biomed. Pharmacol. J.*, vol. 10, no. 4, pp. 1823–1830, 2017, doi: 10.13005/bpj/1299.
 - [15] Z. Cai *et al.*, “An Intelligent Parkinson’s Disease Diagnostic System Based on a Chaotic Bacterial Foraging Optimization Enhanced Fuzzy KNN Approach,” *Comput. Math. Methods Med.*, vol. 2018, 2018, doi: 10.1155/2018/2396952.
 - [16] A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja’afari, and S. Abdulwahed, “Diabetes classification based on KNN,” *IIUM Eng. J.*, vol. 21, no. 1, pp. 175–181, 2020, doi: 10.31436/iiumej.v21i1.1206.