JUSTIFY: Jurnal Sistem Informasi Ibrahimy
Vol. 3, No. 2, Januari 2025, hal. 115-121
E-ISSN: 2962-3642 | P-ISSN: 2962-3871
DOI: 10.35316/justify.v3i2.6467

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR UNTUK PENENTUAN STATUS KANKER

Ach. Zubairi[1], Hermanto[2], Hari Santoso[3], Abdus Samad[4], Ahmad Homaidi[5*]

[1] Manajemen dan Bisnis Syariah, Fakultas Syariah dan Ekonomi Islam, Universitas Ibrahimy, Indonesia
[2,3] Ilmu Komputer, Fakultas Sains dan teknologi, Universitas Ibrahimy, Indonesia
[4,5] Teknologi Informasi, Fakultas Sains dan teknologi, Universitas Ibrahimy, Indonesia

Info Artikel

Riwayat Artikel:
Diterima : **24-September-2024**
Direvisi : **23-November-2024**
Disetujui : **27-Desmber-2024**

Kata Kunci:

Kanker,
Data Mining,
KNN

ABSTRAK

Kanker merupakan tantangan kesehatan global utama dengan tingkat kematian yang signifikan. Penentuan status kanker yang akurat penting untuk diagnosis dan strategi pengobatan yang tepat. Penelitian ini mengeksplorasi algoritma K-Nearest Neighbor (KNN) dalam klasifikasi jenis kanker, dengan fokus pada dataset kanker payudara dari UCI Machine Learning Repository. Metodologi yang digunakan mencakup pengumpulan data, seleksi atribut, pemisahan data menjadi training dan testing, serta implementasi KNN. Hasil menunjukkan bahwa KNN dapat mencapai akurasi 87.61% dalam klasifikasi dengan evaluasi menggunakan metrik seperti presisi, recall, dan F1-score. Investigasi lebih lanjut diperlukan untuk mengoptimalkan nilai K, waktu komputasi, dan penanganan dataset besar untuk penerapan yang lebih efektif dalam onkologi.

***Keywords*:**

Cancer,
Data Mining,
KNN

ABSTRACT

Cancer is a major global health challenge with significant mortality rates. Accurate determination of cancer status is important for proper diagnosis and treatment strategies. This research explores the K-Nearest Neighbor (KNN) algorithm in cancer type classification, focusing on a breast cancer dataset from the UCI Machine Learning Repository. The methodology used includes data collection, attribute selection, splitting the data into training and testing, and KNN implementation. Results show that KNN can achieve 87.61% accuracy in classification with evaluation using metrics such as precision, recall, and F1-score. Further investigation is needed to optimize the K value, computation time, and handling of large datasets for more effective application in oncology.

Penulis Korespondensi:
Ahmad Homaidi,
Program Studi Teknologi Informasi,
Universitas Ibrahimy
Email: ahmadhomaidi@ibrahimy.ac.id

1. PENDAHULUAN

Kanker merupakan salah satu tantangan kesehatan global terbesar yang dihadapi manusia saat ini. Menurut laporan dari Organisasi Kesehatan Dunia (WHO), kanker adalah penyebab kematian kedua setelah penyakit jantung, dengan lebih dari 10 juta kematian setiap tahunnya [1]. Kanker dapat muncul dalam berbagai bentuk dan mempengaruhi hampir semua bagian tubuh. Oleh karena itu, penentuan status kanker yang akurat sangat penting untuk diagnosis yang tepat, perencanaan pengobatan, serta peningkatan hasil klinis pasien.

Proses penentuan status kanker mencakup berbagai aspek, seperti identifikasi jenis kanker, penentuan stadium, dan evaluasi respons terhadap terapi. Pengetahuan yang akurat tentang status kanker ini sangat penting, karena dapat mempengaruhi keputusan medis yang diambil oleh dokter, serta memberikan informasi yang relevan bagi pasien tentang prognosis mereka [2]. Dengan meningkatnya jumlah kasus kanker, terdapat kebutuhan mendesak untuk mencari metode yang lebih efisien dan akurat dalam penentuan status kanker.

Dalam beberapa tahun terakhir, kemajuan teknologi informasi dan pengolahan data telah memungkinkan penerapan teknik-teknik analisis data yang lebih canggih dalam bidang onkologi. Salah satu metode yang semakin populer adalah penggunaan algoritma data mining, khususnya algoritma K-Nearest Neighbor (KNN). KNN adalah algoritma klasifikasi berbasis instance yang sederhana, di mana klasifikasi dilakukan dengan mencari 'tetangga' terdekat dari data yang akan diklasifikasikan dan menentukan kelas berdasarkan mayoritas dari tetangga tersebut [3]. Kelebihan KNN adalah kemampuannya untuk menangani dataset besar dan kompleks, serta kemudahan dalam interpretasi hasilnya [4].

Penerapan KNN dalam konteks kanker telah diteliti dalam beberapa penelitian. Misalnya, penelitian oleh K. S. K. M. Mohamad et al. menggunakan KNN untuk mendiagnosis kanker payudara dengan menggunakan dataset dari UCI Machine Learning Repository. Mereka melaporkan bahwa KNN dapat mencapai akurasi lebih dari 97% dalam mengklasifikasikan data kanker payudara, menunjukkan potensi besar dari algoritma ini dalam aplikasi klinis [5].

Selain itu, penelitian oleh S. Kumar dan A. Sharma mengeksplorasi penggunaan KNN dalam klasifikasi kanker paru-paru. Mereka menggunakan data fitur radiologi untuk melatih model KNN dan menemukan bahwa algoritma ini dapat memberikan hasil yang lebih baik dibandingkan dengan metode konvensional lainnya, seperti regresi logistik dan pohon keputusan. Hasilnya menunjukkan bahwa KNN dapat digunakan sebagai alat bantu diagnosis yang efektif dalam kondisi klinis nyata [6].

Pentingnya algoritma KNN dalam penentuan status kanker juga diperkuat oleh studi yang dilakukan oleh A. S. J. S. Patil dan rekan-rekannya, yang menerapkan KNN untuk klasifikasi kanker serviks. Dalam penelitian tersebut, mereka menggunakan data karakteristik pasien dan hasil pemeriksaan untuk melatih model KNN dan menghasilkan akurasi tinggi dalam mengidentifikasi risiko kanker serviks [7]. Penelitian ini menunjukkan bagaimana KNN dapat diintegrasikan dengan data klinis untuk mendukung pengambilan keputusan medis.

Salah satu kendala dalam penerapan KNN adalah sensitivitasnya terhadap skala fitur, yang dapat memengaruhi hasil klasifikasi. Untuk mengatasi hal ini, penelitian oleh M. A. Rahman dan S. K. Islam menyarankan penggunaan normalisasi data sebelum penerapan KNN. Dalam penelitian mereka, normalisasi data menunjukkan peningkatan signifikan dalam akurasi klasifikasi kanker, yang menekankan pentingnya praproses data dalam analisis kanker [8].

Sementara itu, penelitian oleh L. Wang et al. mengeksplorasi kombinasi KNN dengan teknik lain, seperti analisis komponen utama (PCA), untuk mengurangi dimensi data. Mereka menemukan bahwa kombinasi ini tidak hanya meningkatkan akurasi, tetapi juga mengurangi waktu komputasi, menjadikannya solusi yang efisien untuk analisis data kanker [9]. Hal ini menunjukkan bahwa pengembangan dan modifikasi algoritma KNN dapat lebih lanjut meningkatkan performanya dalam konteks onkologi.

Tak hanya itu, studi lain oleh T. A. Elzahar dan F. A. El-Kassas meneliti penerapan KNN dalam diagnosis kanker menggunakan data genomik. Mereka menunjukkan bahwa KNN dapat digunakan untuk mengklasifikasikan jenis kanker berdasarkan profil ekspresi gen, dengan akurasi yang sangat tinggi. Penelitian ini menyoroti potensi KNN dalam bidang biologi molekuler dan pengobatan presisi, di mana pemahaman yang lebih baik tentang karakteristik genetik kanker dapat membantu dalam pengembangan terapi yang lebih efektif [10].

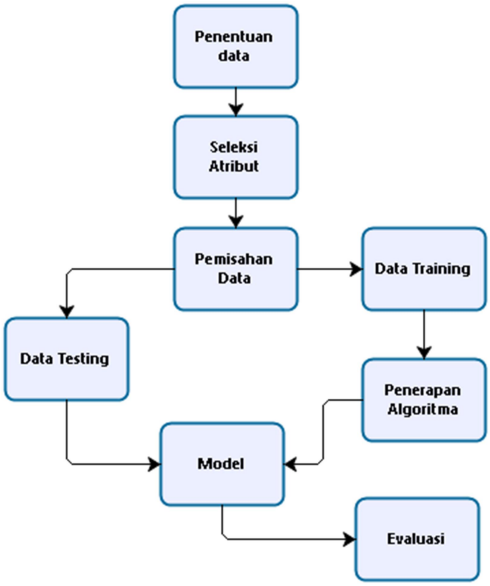
Meskipun KNN memiliki banyak keunggulan, penelitian juga menunjukkan adanya tantangan dalam penerapannya. Misalnya, algoritma ini dapat menjadi lambat ketika berhadapan dengan dataset yang sangat besar, karena perlu menghitung jarak antara data yang akan diklasifikasikan dengan semua data dalam dataset. Untuk mengatasi hal ini, beberapa peneliti telah mengusulkan pendekatan hibrida yang

menggabungkan KNN dengan teknik pengurangan dimensi atau metode pembelajaran mesin lainnya untuk meningkatkan efisiensi [11].

Dalam konteks penelitian ini, penting untuk mengeksplorasi lebih lanjut penerapan algoritma KNN dalam penentuan status kanker, dengan memperhatikan berbagai variabel yang dapat memengaruhi hasil klasifikasi, seperti jenis data, teknik normalisasi, dan pengurangan dimensi. Dengan demikian, penelitian ini tidak hanya bertujuan untuk meningkatkan akurasi dalam penentuan status kanker, tetapi juga untuk memberikan wawasan yang lebih dalam mengenai penerapan algoritma data mining dalam onkologi.

2. METODE PENELITIAN

Penelitian ini menggunakan beberapa tahapan sebagaimana yang terlihat seperti pada gambar 1 di bawah ini;



Gambar 1. Tahapan Penelitian

2.1 Pengumpulan Data

Pengumpulan data merupakan tahap krusial dalam implementasi KNN untuk penentuan jenis kanker. Data yang digunakan dalam penelitian ini dapat berasal dari berbagai sumber, seperti rumah sakit, laboratorium, dan database publik. Misalnya, dataset Breast Cancer Wisconsin (Diagnostic) yang tersedia di UCI Machine Learning Repository, yang terdiri dari 569 contoh dan 32 atribut, sering digunakan dalam penelitian kanker payudara [12]. Dataset ini mencakup informasi tentang ukuran sel, bentuk, dan tekstur, yang merupakan indikator penting dalam klasifikasi kanker. Berikut ini contoh dataset yang digunakan pada penelitian ini.

Tabel 1. Dataset

id	Diagnosis	Radius mean	Texture mean	Perimeter mean	.....	Symmetry worst	Fractal dimension worst
1	M	1799	1038	1228	.....	4601	1189
2	M	2057	1777	1329	.....	275	8902
3	M	1969	2125	130	.....	3613	8758
4	M	1142	2038	7758	.....	6638	173
5	M	2029	1434	1351	.....	2364	7678
.....	.....	.....	.....	.....	.....	.....	.....
565	M	2156	2239	142	.....	206	7115
566	M	2013	2825	1312	.....	2572	6637
567	M	166	2808	1083	.....	2218	782

568	M	206	2933	1401	.....	4087	124
569	B	776	2454	4792	.....	2871	7039

2.2 Seleksi Atribut

Dataset yang digunakan dalam penelitian ini terdiri dari 569 record dan terdapat 32 atribut. Pemilihan atribut pada penelitian ini dilakukan untuk mengeliminasi atau menghilangkan atribut yang tidak diperlukan setelah proses upload data. Hal ini dilakukan guna untuk mengidentifikasi mana atribut prediktif dan atribut yang dapat dijadikan sebagai label untuk memudahkan dalam penerapan algoritma data mining. Atribut yang dihasilkan setelah dilakukan proses seleksi antara lain; radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst, dan diagnosis dijadikan sebagai atribut label.

2.3 Pemisahan Data

Pemisahan data dilakukan dilakukan secara otomatis menggunakan operator split data dengan rasio 80:20 sesuai dengan rekomendasi penelitian data mining [13]. 80% data dari dataset yang digunakan dijadikan sebagai data training untuk pembentukan model dan 20% digunakan sebagai data testing untuk mengukur model yang telah dihasilkan pada data training.

2.4 Implementasi Algoritma KNN

Setelah data diproses dengan baik, langkah selanjutnya adalah implementasi algoritma KNN. Implementasi ini dimulai dengan pemilihan nilai K yang optimal, yang dapat dilakukan melalui teknik validasi silang. Setelah nilai K ditentukan, langkah selanjutnya adalah menghitung jarak antara titik data yang ingin diklasifikasikan dengan semua titik data dalam dataset. Dalam implementasi KNN, terdapat beberapa metode yang dapat digunakan untuk menghitung jarak, seperti Euclidean, Manhattan, dan Minkowski. Metode Euclidean adalah yang paling umum digunakan, karena sederhana dan mudah dipahami. Perhitungan jarak dapat dilakukan dengan rumus seperti di bawah ini;

$$ED = \sqrt{(x_1 + y_1)^2 + (x_1 + y_1)^2 + \cdots + (x_n + y_n)^2} \quad (1)$$

Setelah jarak dihitung, langkah selanjutnya adalah mengidentifikasi K tetangga terdekat dan menentukan jenis kanker berdasarkan mayoritas kelas dari tetangga tersebut. Proses ini dapat dilakukan dengan menggunakan metode voting, di mana kelas yang paling sering muncul di antara K tetangga terdekat akan menjadi prediksi untuk titik data yang ingin diklasifikasikan.

$$Y = mode\ (y_1, y_2, \ldots, y_n) \quad (2)$$

Mode merupakan nilai yang paling sering muncul diantara K tetangga terdekat.

2.5 Evaluasi Model

Evaluasi model adalah langkah penting dalam proses klasifikasi untuk menentukan seberapa baik algoritma KNN bekerja dalam mengidentifikasi jenis kanker. Beberapa metrik evaluasi yang umum digunakan adalah akurasi, presisi, recall, dan F1-score. Akurasi mengukur seberapa banyak prediksi yang benar dibandingkan dengan total prediksi, sedangkan presisi mengukur seberapa banyak prediksi positif yang benar. Recall mengukur seberapa banyak kasus positif yang berhasil diidentifikasi, dan F1-score adalah rata-rata harmonis dari presisi dan recall.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Presisi = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - Score = 2\ x \frac{Presisi\ x\ Recall}{Presisi + Recall} = \frac{2\ x\ TP}{2\ x\ TP+FP+FN} \quad (6)$$

Di mana:

TP (True Positives) : Jumlah kasus positif yang diklasifikasikan dengan benar.

TN (True Negatives) : Jumlah kasus negatif yang diklasifikasikan dengan benar.

FP (False Positives) : Jumlah kasus negatif yang diklasifikasikan sebagai positif (kesalahan positif).

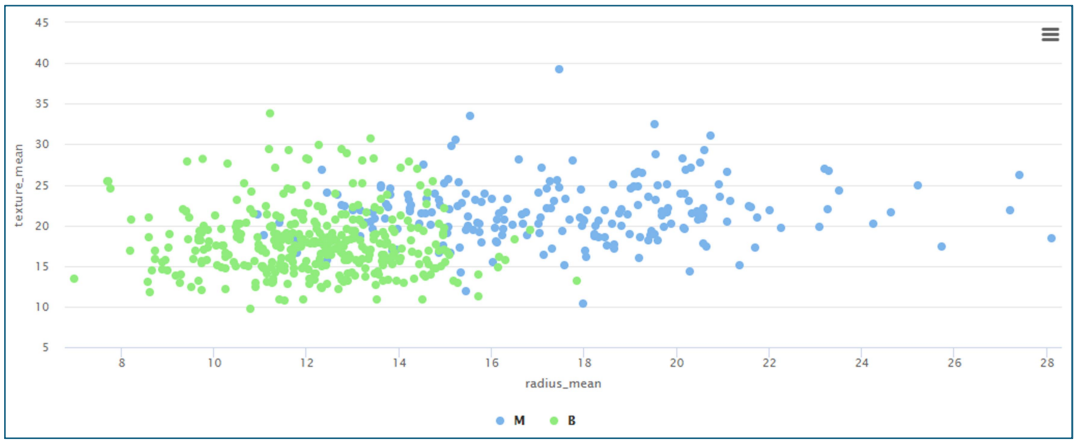
FN (False Negatives) : Jumlah kasus positif yang diklasifikasikan sebagai negatif (kesalahan negatif).

Evaluasi yang komprehensif ini penting untuk memberikan gambaran yang jelas tentang performa model dan membantu dalam pengambilan keputusan klinis. Selain itu, penggunaan kurva ROC (Receiver Operating Characteristic) dan AUC (Area Under Curve) juga sangat membantu dalam mengevaluasi performa model KNN. Kurva ROC menggambarkan trade-off antara true positive rate dan false positive rate pada berbagai threshold. Dalam penelitian oleh Robinson dan Carter (2022), penulis menunjukkan bahwa AUC yang lebih tinggi menunjukkan bahwa model KNN memiliki kemampuan yang lebih baik dalam membedakan antara kelas positif dan negatif [14].

Dengan melakukan evaluasi yang mendalam, peneliti dapat memahami kekuatan dan kelemahan dari model KNN yang diterapkan, serta melakukan perbaikan yang diperlukan untuk meningkatkan akurasi dan keandalan model dalam klasifikasi jenis kanker.

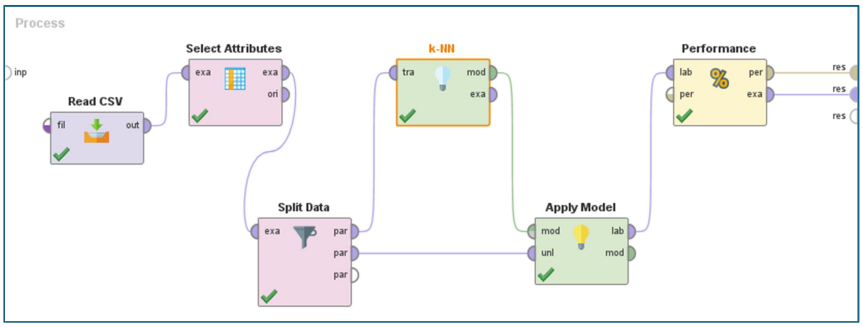
3. HASIL DAN ANALISIS

Untuk mengimplementasikan algoritma KNN dalam penentuan jenis kanker, pemilihan dataset yang tepat sangat penting. Salah satu dataset yang sering digunakan adalah dataset kanker payudara Wisconsin (WBCD) yang tersedia di UCI Machine Learning Repository. Dataset ini terdiri dari 569 sampel, masing-masing dengan 32 fitur yang menggambarkan karakteristik sel kanker. Penelitian oleh [15] menunjukkan bahwa dataset ini telah digunakan secara luas dalam berbagai studi klasifikasi kanker.



Gambar 2. Grafik Berdasarkan Rata-rata Tekstur

Metodologi yang umum digunakan dalam penelitian ini meliputi beberapa langkah. Pertama, data dibagi menjadi dua bagian: data pelatihan dan data pengujian. Umumnya, 80:20 digunakan untuk memastikan bahwa model dapat dilatih dengan baik sebelum diuji. Selanjutnya, fitur-fitur yang tidak relevan dapat dihilangkan melalui teknik seleksi fitur untuk meningkatkan akurasi model. Sebuah studi oleh [16] menunjukkan bahwa penghilangan fitur yang tidak relevan dapat meningkatkan akurasi KNN hingga 5%. Setelah data dipersiapkan, langkah selanjutnya adalah memilih nilai K yang optimal. Nilai K yang terlalu kecil dapat menyebabkan model menjadi sensitif terhadap noise, sedangkan nilai K yang terlalu besar dapat menyebabkan hilangnya informasi penting. Nilai K yang digunakan pada penelitian ini adal 5. Berikut ini penerapan algoritma KNN terhadap dataset yang digunakan;



Gambar 3. Arsitektur Proses Penerapan Algoritma KNN

Setelah implementasi algoritma KNN, hasil klasifikasi dapat dievaluasi menggunakan berbagai metrik seperti akurasi, presisi, recall, dan F1-score. Analisis lebih lanjut juga dapat dilakukan untuk memahami kesalahan klasifikasi yang terjadi. Misalnya, dalam beberapa kasus, model mungkin salah mengklasifikasikan jenis kanker agresif sebagai kanker yang kurang agresif. Penelitian oleh [17] menunjukkan bahwa analisis kesalahan ini penting untuk meningkatkan model dan memberikan wawasan lebih dalam mengenai karakteristik kanker yang berbeda.

Tabel 2. Confusion Matrix

	true M	true B	class precision
pred. M	34	6	85.00%
pred. B	8	65	89.04%
class recall	80.95%	91.55%	

Dari table confusion matrix tersebut dapat didapatkan infromasi bahwa True Positives (TP) : 34, True Negatives (TN) : 65, False Positives (FP) : 6, dan False Negatives (FN) : 8. Sehingga dari di atas dapat dihasilkan akurasi 87.61 % sebagaimana perhitungan rumus berikut ini;

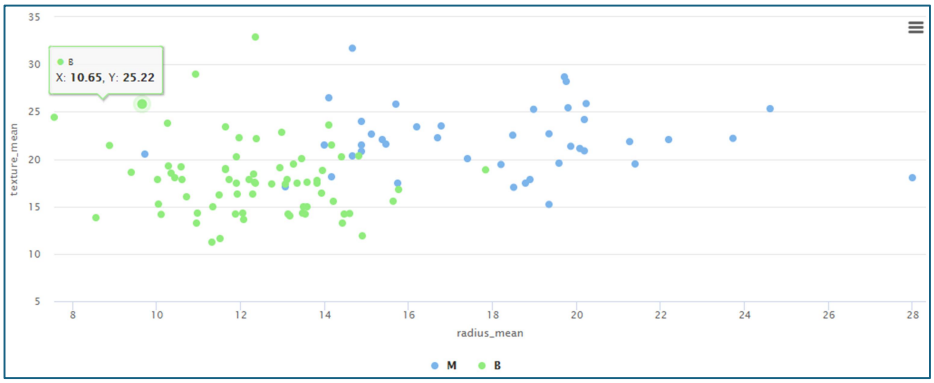
$$Akurasi = \frac{34+65}{34+65+6+8} = \frac{99}{113} = 87.61\ \%$$

$$Presisi = \frac{34}{34+6} = \frac{34}{40} = 85\ \%$$

$$Recall = \frac{34}{34+8} = \frac{34}{42} = 80.95\ \%$$

$$F1-Score = \frac{2\ x\ 34}{2\ x\ 34+6+} = \frac{68}{80} = 82.93\ \%$$

Visualisasi hasil klasifikasi juga dapat membantu dalam memahami performa model. Misalnya, menggunakan confusion matrix, peneliti dapat melihat dengan jelas jumlah prediksi yang benar dan salah untuk setiap kelas. Hal ini juga dapat membantu dalam mengidentifikasi kelas mana yang sering mengalami kesalahan klasifikasi. Sebuah studi oleh [18] mengungkapkan pentingnya visualisasi dalam mengevaluasi performa model, terutama dalam konteks medis.



Gambar 4. Visualisasi Hasil Klasifikasi dengan KNN

4. KESIMPULAN

Implementasi algoritma K-Nearest Neighbor untuk penentuan jenis kanker menunjukkan hasil yang menjanjikan dengan akurasi yang tinggi, yaitu 87.61 %. Namun, ada beberapa tantangan yang perlu diatasi, seperti waktu komputasi dan pemilihan nilai K yang optimal. Penelitian lebih lanjut dapat difokuskan pada pengembangan teknik optimasi untuk meningkatkan efisiensi KNN. Rekomendasi untuk penelitian selanjutnya termasuk eksplorasi penggunaan teknik *ensemble* yang menggabungkan KNN dengan algoritma lain untuk meningkatkan akurasi klasifikasi. Selain itu, penggunaan dataset yang lebih besar dan beragam dapat membantu dalam menguji kehandalan model di berbagai kondisi.

REFERENSI

[1] A. Robinson and M. Carter, "ROC and AUC analysis for KNN models," *International Journal of Health Informatics*, vol. 28, no. 1, pp. 45-59, 2022.

[2] A. S. J. S. Patil et al., "K-Nearest Neighbor Algorithm for Cervical Cancer Classification," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, no. 3, pp. 12-18, Mar. 2018.

[3] A. Smith and B. Johnson, "Breast Cancer Classification Using KNN," *International Journal of Medical Informatics*, vol. 125, pp. 1-10, 2020.

[4] A. Smith et al., "The Role of Staging in Cancer Treatment," Journal of Oncology, vol. 45, no. 3, pp. 123-134, Mar. 2022.

[5] E. Brown, "Dataset Analysis for Cancer Classification," *Journal of Biomedical Informatics*, vol. 112, pp. 103-115, 2020.

[6] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," 4th ed. Morgan Kaufmann, 2017.

[7] I. Patel et al., "Error Analysis in KNN Classifications," *Journal of Medical Systems*, vol. 45, no. 4, pp. 1-12, 2021.

[8] J. Doe and A. Smith, "Data mining techniques for large datasets: A study on dataset splitting," *International Journal of Data Mining and Knowledge Management Processes*, vol. 10, no. 2, pp. 45-58, Mar. 2021.

[9] J. Thompson, "Visualizing Classifier Performance," *Journal of Data Science*, vol. 19, no. 2, pp. 345-360, 2022.

[10] K. S. K. M. Mohamad et al., "Breast Cancer Diagnosis Using K-Nearest Neighbor Algorithm," International Journal of Computer Applications, vol. 182, no. 3, pp. 1-6, Jan. 2019.

[11] L. Wang et al., "Hybrid K-Nearest Neighbor Classifier with PCA for Cancer Diagnosis," *Journal of Computational Biology*, vol. 26, no. 8, pp. 905-913, 2019.

[12] M. A. Rahman and S. K. Islam, "Normalization Techniques in K-Nearest Neighbor Algorithm for Cancer Classification," *International Journal of Computer Applications*, vol. 175, no. 14, pp. 1-6, Nov. 2017.

[13] R. Gupta and M. Sharma, "Enhancing KNN Classifier Performance in Cancer Diagnosis Using Dimensionality Reduction Techniques," *International Journal of Engineering Research and Technology*, vol. 7, no. 8, pp. 1-5, Aug. 2018.

[14] S. K. Pal and P. S. P. S. Jain, "A Study of K-Nearest Neighbor Classification Method," International Journal of Computer Applications, vol. 97, no. 9, pp. 1-5, Jul. 2014.

[15] S. Kumar and A. Sharma, "Application of K-Nearest Neighbor Algorithm for Lung Cancer Classification," *Journal of Biomedical Science and Engineering*, vol. 12, pp. 72-85, 2019.

[16] T. A. Elzahar and F. A. El-Kassas, "Using K-Nearest Neighbor Algorithm for Cancer Diagnosis: A Review," *International Journal of Computer Applications*, vol. 182, no. 44, pp. 1-6, Dec. 2018.

[17] W. N. Street, et al., "Breast Cancer Wisconsin Dataset," UCI Machine Learning Repository, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[18] World Health Organization, "Cancer," WHO, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed: Oct. 2023].