

Perbandingan Algoritma Decision Tree dan K-Nearest Neighbor untuk Klasifikasi Penyakit ISPA

Marchell William Putra Pakpahan¹, Bayu Angga Wijaya²,
Marsaulina Lumbantoruan³, Ambarsius Samosir⁴, Mikhael Rafael⁵

¹Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Prima Indonesia, marchellpakpahan@gmail.com

²Teknik Informatika, Fakultas Sains dan Teknologi Universitas Prima Indonesia, bayuanggawijaya@unprimdn.ac.id

³Teknik Informatika, Fakultas Sains dan Teknologi Universitas Prima Indonesia, marsaulinaa7@gmail.com

⁴Teknik Informatika, Fakultas Sains dan Teknologi Universitas Prima Indonesia, ambarsius21@gmail.com

⁵Teknik Informatika, Fakultas Sains dan Teknologi Universitas Prima Indonesia, mikhael.rafael09@gmail.com

Keywords:

ARI,
Decision Tree,
K-Nearest Neighbor,
Classification,
Machine Learning

ABSTRACT

Acute Respiratory Infection (ARI) is one of the most common respiratory diseases with diverse and overlapping clinical symptoms, making initial identification challenging and necessitating a systematic, data-driven classification approach. This study aims to compare the performance of the Decision Tree and K-Nearest Neighbor (KNN) algorithms in classifying ARI-related disease categories. The novelty of this research lies in the specific construction of ARI labels into five distinct categories from the Pediatric Respiratory Infections dataset, coupled with a rigorous feature selection process to handle mixed data types and address class imbalance using weighted evaluation metrics. The dataset consisted of 801 patient records with 91 initial attributes. The classification label was constructed from the Main diagnostic column and grouped into five categories: Asthma/Bronchospasm/Wheezing, Pneumonia/Pneumopathy, Bronchiolitis, Laryngeal/Upper Respiratory, and Other. After feature selection to remove noise and redundancy, 54 features were used, consisting of 24 numerical and 30 categorical features. The research stages included preprocessing, label construction, missing value handling, categorical encoding, KNN normalization, 80:20 train-test splitting, and model evaluation. The results show that Decision Tree achieved higher performance with 67.08% accuracy and 69.12% weighted F1-score, while KNN achieved 65.84% accuracy and 64.18% weighted F1-score. Thus, Decision Tree demonstrates superior performance and interpretability for this specific dataset.

Kata Kunci :

ISPA,
Decision Tree,
K-Nearest Neighbor,
Klasifikasi,
Machine Learning

ABSTRAK

Infeksi Saluran Pernapasan Akut (ISPA) merupakan penyakit pernapasan yang sering ditemukan di masyarakat dengan gejala yang beragam dan sering tumpang tindih, sehingga identifikasi awal secara manual rentan terhadap keterlambatan atau kesalahan. Oleh karena itu, diperlukan pendekatan berbasis data untuk membantu proses klasifikasi secara lebih sistematis dan objektif. Penelitian ini bertujuan membandingkan kinerja algoritma Decision Tree dan K-Nearest Neighbor (KNN) dalam mengklasifikasikan penyakit ISPA. Keterbaruan (novelty) penelitian ini terletak pada pembentukan label ISPA yang spesifik menjadi lima kategori klinis dari dataset Pediatric Respiratory Infections, serta penerapan seleksi fitur dan metrik evaluasi *weighted* untuk mengatasi ketidakseimbangan kelas (*class imbalance*) pada data medis. Dataset yang digunakan terdiri dari 801 data pasien dengan 91 atribut awal. Setelah proses seleksi fitur untuk menghilangkan *noise* dan redundansi, digunakan 54 fitur yang terdiri atas 24 fitur numerik dan 30 fitur kategorikal. Tahapan penelitian meliputi preprocessing, pembentukan label, penanganan nilai kosong, *encoding*, normalisasi untuk KNN, pembagian data latih dan uji rasio 80:20, serta evaluasi menggunakan *confusion matrix*, *accuracy*, *precision weighted*, *recall weighted*, dan *F1-score weighted*. Hasil penelitian menunjukkan Decision Tree memperoleh *accuracy* 67,08% dan *F1-score weighted* 69,12%, sedangkan KNN memperoleh *accuracy* 65,84% dan *F1-score weighted* 64,18%. Dengan demikian, Decision Tree memiliki performa dan interpretabilitas yang lebih baik pada dataset yang digunakan.

Korespondensi Penulis:

Bayu Angga Wijaya,
Afiliasi: Universitas Prima Indonesia, Medan
Telepon: +62 82165089123
Email: bayuangawijaya@unprimdn.ac.id

**Submitted: 26-05-2026; Accepted: 04-06-2026;
Published: 12-06-2026**

Copyright (c) 2026 The Author (s) This article is distributed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0)

1. PENDAHULUAN

Kemajuan teknologi informasi mendorong pemanfaatan data di bidang kesehatan. Selain menyimpan data, teknologi dapat membantu proses analisis, klasifikasi, prediksi, dan pengambilan keputusan. Pemanfaatan *machine learning* memungkinkan pengolahan data pasien dilakukan secara lebih cepat, sistematis, dan objektif [1], [2]. Sejumlah studi terbaru juga menunjukkan bahwa pendekatan *machine learning* dan AI-IoMT banyak digunakan untuk prediksi penyakit kronis, prognosis berbasis data medis, serta diagnosis dini pada berbagai penyakit [13], [16], [17].

Infeksi Saluran Pernapasan Akut (ISPA) merupakan gangguan pernapasan yang umum terjadi dan menjadi salah satu penyebab utama morbiditas, terutama pada anak-anak. Pengambilan studi kasus ISPA dalam penelitian ini didasari oleh tingginya prevalensi penyakit ini dan tantangan klinis dalam diagnosis awal. Kemiripan gejala (seperti demam, batuk, dan sesak napas) dengan berbagai penyakit pernapasan lain membuat identifikasi awal tidak selalu mudah, terutama ketika jumlah pasien dan variasi kondisi klinis cukup besar. Keterlambatan atau kesalahan klasifikasi dapat berakibat pada penanganan yang tidak tepat. Oleh karena itu, sistem pendukung keputusan berbasis data sangat dibutuhkan untuk membantu tenaga medis dalam triase dan klasifikasi awal.

Klasifikasi dalam *machine learning* dapat digunakan untuk mengelompokkan data pasien berdasarkan pola atribut, seperti gejala, tanda vital, dan indikator klinis. Decision Tree merupakan algoritma berbentuk pohon keputusan yang mudah ditelusuri karena terbentuk dari aturan pada setiap atribut [7], [8]. Sementara itu, K-Nearest Neighbor (KNN) menentukan kelas berdasarkan kedekatan data uji dengan data latih, tetapi kinerjanya dipengaruhi skala data, jumlah fitur, nilai K, dan metode jarak [9], [10]. Penelitian pada penyakit jantung dan kardiovaskular juga menunjukkan pentingnya pemilihan fitur serta model *tree-based* dalam meningkatkan kinerja prediksi [14], [15], [19].

Tujuan utama penelitian ini adalah membandingkan performa algoritma Decision Tree dan KNN dalam mengklasifikasikan penyakit ISPA. Adapun proses seleksi fitur dilakukan sebagai tahapan preprocessing yang krusial, bukan sebagai tujuan utama, dengan alasan untuk mengurangi dimensi data, menghilangkan atribut yang tidak relevan (noise), dan mencegah overfitting, sehingga perbandingan antara kedua algoritma dapat dilakukan pada fitur-fitur yang paling informatif dan representatif. Perbedaan karakteristik Decision Tree dan KNN menjadi dasar perbandingan performa menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix* [11], [12].

2. METODE PENELITIAN

2.1. Jenis Penelitian

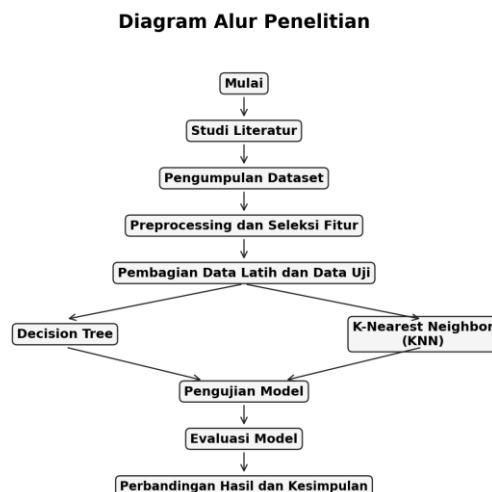
Jenis penelitian ini adalah penelitian eksperimen dengan pendekatan kuantitatif. Eksperimen dilakukan dengan menguji dua algoritma klasifikasi, yaitu Decision Tree dan K-Nearest Neighbor, untuk membandingkan kemampuannya dalam mengklasifikasikan penyakit ISPA. Pendekatan kuantitatif digunakan karena hasil penelitian dianalisis melalui nilai numerik dari metrik evaluasi, seperti *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Penelitian ini berfokus pada proses komputasi sebagai bahan perbandingan algoritma dan tidak dimaksudkan untuk menggantikan diagnosis dokter atau tenaga medis.

2.2. Data Penelitian

Bahan penelitian adalah dataset *Pediatric Respiratory Infections: Epidemiological and Etiological Data from a Cohort of 801 Moroccan Children* yang diperoleh dari Mendeley Data. Dataset ini berisi data infeksi saluran pernapasan pada anak dan digunakan sebagai dasar klasifikasi penyakit ISPA dengan algoritma Decision Tree dan KNN. Dataset terdiri atas 801 data pasien dengan **91 atribut awal (kolom data mentah sebelum proses pembersihan dan seleksi)**. Label klasifikasi dibentuk berdasarkan kolom *Main diagnostic* dan dikelompokkan menjadi lima kategori, yaitu *Asthma/Bronchospasm/Wheezing*, *Pneumonia/Pneumopathy*, *Bronchiolitis*, *Laryngeal/Upper Respiratory*, dan *Other*.

2.3. Prosedur Penelitian

Prosedur penelitian meliputi studi literatur, pengumpulan dataset, pemeriksaan data, preprocessing, pembagian data latih dan data uji, implementasi model, pengujian model, evaluasi, serta perbandingan hasil. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

Pada tahap preprocessing, data dibersihkan dengan merapikan nama kolom dan mengubah nilai kosong seperti N/A, NA, NULL, None, atau data kosong menjadi NaN. Nilai kosong pada fitur numerik diisi dengan median, sedangkan fitur kategorikal diisi dengan modus. Data kategorikal kemudian diubah menjadi numerik menggunakan One-Hot Encoding, dan normalisasi dilakukan pada fitur numerik untuk mendukung KNN. Dalam konteks data kesehatan, penanganan nilai kosong dan reduksi atau seleksi fitur dilaporkan penting untuk menjaga kualitas model, misalnya melalui KNN imputer, PCA, dan ensemble stacking [18], [20].

2.4. Pembagian Data dan Evaluasi

Dataset dibagi menjadi data latih dan data uji dengan rasio 80:20 menggunakan random state 42. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengukur kemampuan model pada data baru. Kedua algoritma diuji dengan data data uji yang sama agar hasil perbandingan lebih objektif.

Tabel 1. Pembagian Data Latih dan Data Uji

Jenis Data	Jumlah
Data latih	640
Data uji	161
Total data	801

Evaluasi model dilakukan menggunakan confusion matrix yang mencakup True Positive, True Negative, False Positive, dan False Negative. Berdasarkan nilai tersebut dihitung accuracy, precision weighted, recall weighted, dan F1-score weighted. Metrik weighted digunakan karena distribusi kelas pada dataset tidak seimbang.

3. HASIL DAN PEMBAHASAN

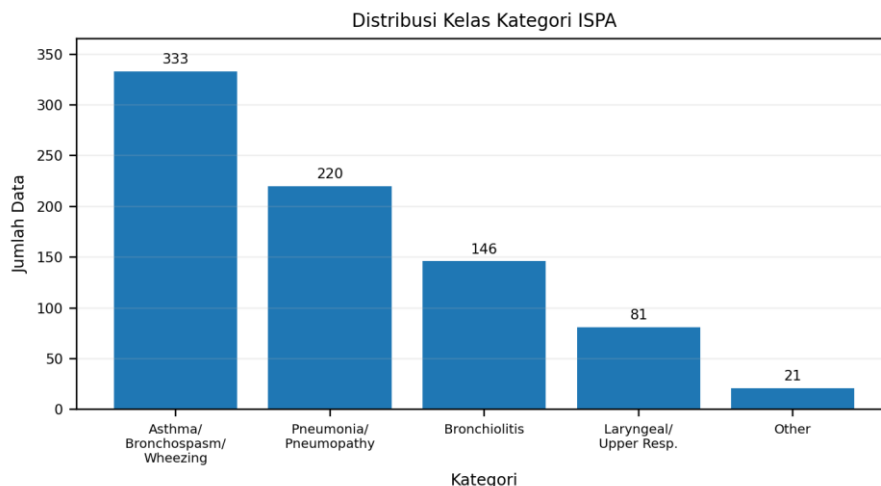
3.1. Gambaran Umum Dataset

Distribusi kategori penyakit ISPA menunjukkan bahwa kelas *Asthma/Bronchospasm/Wheezing* memiliki jumlah data terbanyak (333 data), sedangkan kelas *Other* memiliki jumlah data paling sedikit (21 data). Kondisi ini memperlihatkan adanya ketidakseimbangan kelas (*class imbalance*) yang berpotensi menghasilkan bias jika model dilatih secara naif, di mana model akan cenderung memprediksi kelas mayoritas.

Namun, penelitian ini tetap dapat dilanjutkan ke tahapan berikutnya dengan menerapkan strategi khusus untuk mengatasi ketidakseimbangan tersebut, tanpa harus melakukan oversampling sintesis (seperti SMOTE) yang berisiko menimbulkan noise pada data medis. Strategi yang diterapkan adalah: (1) Penggunaan parameter `class_weight='balanced'` pada algoritma Decision Tree untuk memberikan bobot lebih besar pada kelas minoritas selama pelatihan; (2) Penggunaan parameter `weights='distance'` pada KNN; dan (3) Penggunaan metrik evaluasi *weighted* (precision, recall, dan F1-score) yang memperhitungkan proporsi setiap kelas, sehingga performa model dinilai secara adil dan tidak bias terhadap kelas mayoritas.

Tabel 2. Distribusi Kategori Penyakit ISPA

Kategori ISPA	Jumlah Data
Asthma/Bronchospasm/Wheezing	333
Pneumonia/Pneumopathy	220
Bronchiolitis	146
Laryngeal/Upper Respiratory	81
Other	21



Gambar 2. Grafik Distribusi Kelas Kategori ISPA

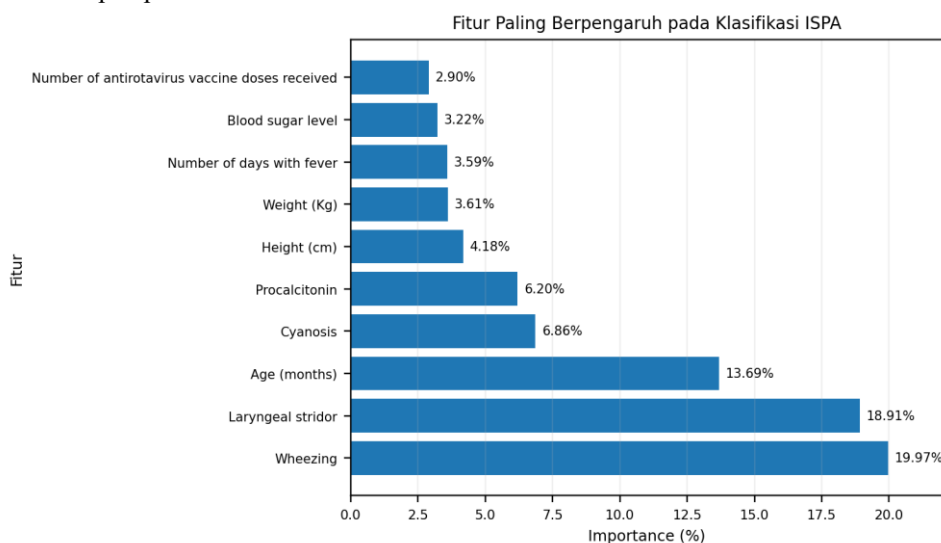
3.2. Hasil Preprocessing dan Seleksi Fitur

Dataset awal memiliki 91 atribut. Setelah proses pembersihan data dan seleksi fitur, jumlah fitur dikurangi menjadi 54 fitur (terdiri dari 24 fitur numerik dan 30 fitur kategorikal). Alasan penggunaan seleksi fitur dalam penelitian ini adalah untuk mengurangi dimensi data (*dimensionality reduction*), menghilangkan fitur yang tidak relevan atau redundan (seperti ID pasien atau data yang memiliki nilai kosong terlalu banyak), serta mencegah *overfitting*. Dengan mempertahankan hanya 54 fitur yang paling informatif, komputasi menjadi lebih efisien dan model dapat belajar pola yang lebih general, bukan menghafal noise pada data. Fitur yang digunakan dipilih dari atribut yang berkaitan dengan kondisi awal pasien, meliputi data demografi, riwayat kesehatan, gejala, tanda vital, dan tanda klinis. Kolom *Main diagnostic* hanya digunakan untuk membentuk label dan tidak dipakai sebagai fitur input agar tidak terjadi *data leakage*.

Tabel 3. Jumlah Fitur yang Digunakan

Jenis Fitur	Jumlah
Fitur numerik	24
Fitur kategorikal	30
Total fitur	54

Contoh fitur yang digunakan meliputi Age, Gender, History of fever, History of cough, History of rhinorrhea, Oxygen saturation, Axillary temperature, Respiratory rate, Heart rate, Nasal flaring, Rhonchi, Crackles, Wheezing, Hypoventilation, dan Labored breathing. Analisis feature importance pada Decision Tree digunakan untuk mengetahui kontribusi fitur terhadap keputusan klasifikasi.



Gambar 3. Grafik Fitur Paling Berpengaruh pada Klasifikasi ISPA

Tabel 4. Sepuluh Fitur dengan Pengaruh Terbesar

Fitur	Feature Importance
-------	--------------------

Wheezing	19,97%
Laryngeal stridor	18,91%
Age (months)	13,69%
Cyanosis	6,86%
Procalcitonin	6,20%
Height (cm)	4,18%
Weight (Kg)	3,61%
Number of days with fever	3,59%
Blood sugar level	3,22%
Number of antirotavirus vaccine doses received	2,90%

Berdasarkan *feature importance*, fitur *Wheezing*, *Laryngeal stridor*, dan *Age (months)* memiliki pengaruh terbesar. Hal ini sejalan dengan literatur medis yang menyatakan bahwa mengi (*wheezing*) dan stridor adalah indikator klinis utama dalam membedakan obstructive airway disease (seperti asma/bronkiolitis) dengan infeksi saluran pernapasan atas pada anak-anak, sementara usia merupakan faktor risiko demografis yang signifikan [3], [4].

3.3. Implementasi Algoritma

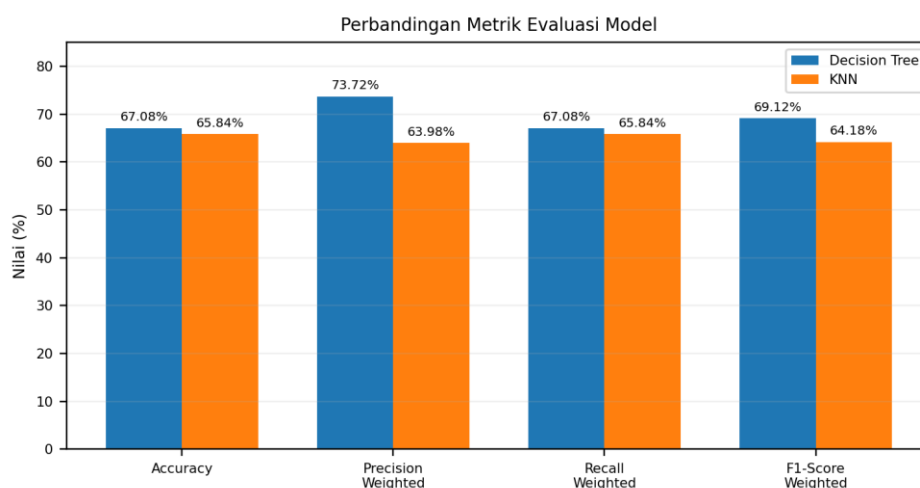
Decision Tree diterapkan dengan parameter *criterion* Gini, *max depth* 7, *min samples leaf* 5, *class weight balanced*, dan *random state* 42. Parameter *class weight balanced* digunakan karena dataset memiliki distribusi kelas tidak seimbang, sehingga model dapat memberikan perhatian lebih pada kelas dengan jumlah data lebih kecil. KNN diterapkan dengan nilai $K = 7$, *weights distance*, dan *metric* Euclidean. Nilai $K = 7$ dipilih berdasarkan hasil uji coba (hyperparameter tuning) pada rentang nilai ganjil ($K = 3, 5, 7, 9$). Nilai ganjil dipilih untuk menghindari hasil voting yang seri (tie), dan hasil eksperimen menunjukkan bahwa $K = 7$ menghasilkan F1-score weighted yang paling optimal dibandingkan nilai K lainnya. Pada KNN, normalisasi fitur numerik dilakukan terlebih dahulu menggunakan Standard Scaler (Z-score normalization). Normalisasi ini sangat penting karena KNN menghitung jarak Euclidean antar data; tanpa normalisasi, fitur dengan skala nilai besar (misalnya, jumlah sel darah) akan mendominasi perhitungan jarak dibandingkan fitur dengan skala kecil (misalnya, suhu tubuh), yang akan merusak akurasi model.

3.4. Hasil Evaluasi Model

Hasil evaluasi memperlihatkan bahwa Decision Tree memperoleh nilai yang lebih tinggi pada seluruh metrik evaluasi utama dibandingkan KNN. Rincian hasil evaluasi ditunjukkan pada Tabel 5 dan divisualisasikan pada Gambar 4.

Tabel 5. Hasil Evaluasi Decision Tree dan KNN

Model	Accuracy	Precision Weighted	Recall Weighted	F1-Score Weighted
Decision Tree	67,08%	73,72%	67,08%	69,12%
KNN	65,84%	63,98%	65,84%	64,18%



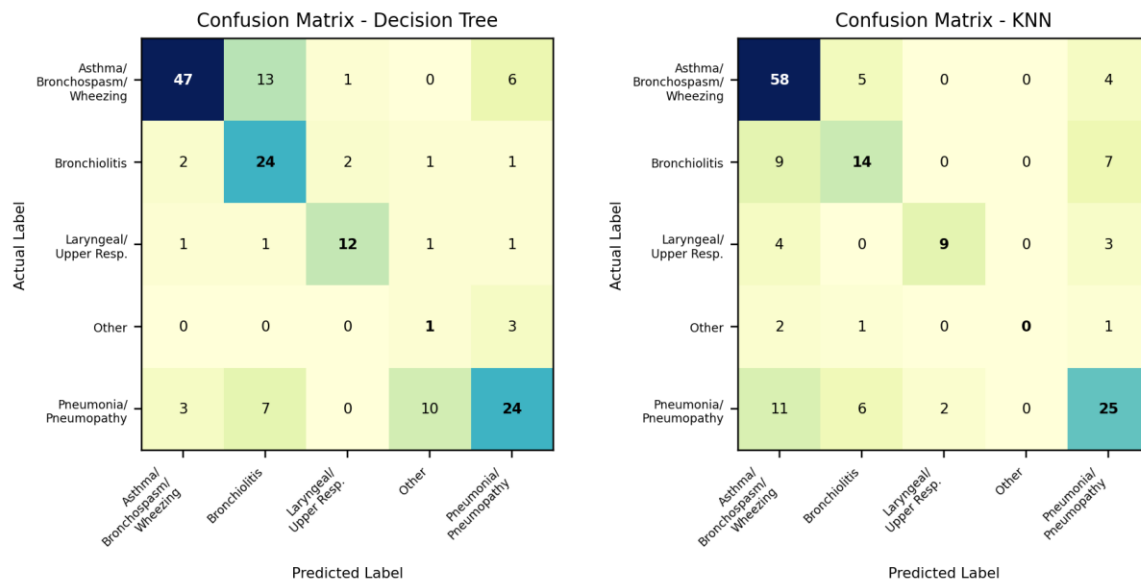
Gambar 4. Grafik Perbandingan Metrik Evaluasi Model

Decision Tree memperoleh *accuracy* sebesar 67,08%, sedangkan KNN memperoleh *accuracy* sebesar 65,84%. Dari sisi *F1-score weighted*, Decision Tree memperoleh 69,12%, sedangkan KNN memperoleh 64,18%. Keunggulan Decision Tree paling terlihat pada *precision weighted*, yaitu 73,72% dibandingkan 63,98% pada KNN. Pembahasan dari nilai-nilai tersebut menunjukkan bahwa Decision Tree lebih unggul dalam menangani data tabular dengan campuran fitur numerik dan kategorikal. Sebagaimana didukung oleh penelitian [6] dan [11], algoritma berbasis pohon keputusan (tree-based) cenderung lebih robust pada data medis karena mampu membentuk aturan klasifikasi (if-then rules) yang secara implisit menangani interaksi antar fitur kategorikal hasil encoding. Sebaliknya, performa KNN yang lebih rendah dapat diatribusikan pada "curse of dimensionality". Meskipun telah dinormalisasi,

keberadaan 30 fitur kategorikal hasil One-Hot Encoding membuat ruang fitur menjadi sangat jarang (sparse), sehingga perhitungan jarak Euclidean menjadi kurang representatif dalam menentukan tetangga terdekat [9], [10].

3.5. Analisis Confusion Matrix

Confusion matrix digunakan untuk melihat kesesuaian antara label aktual dan label hasil prediksi. Target klasifikasi berbentuk multiclass karena terdiri atas lima kategori penyakit ISPA. Hasil confusion matrix Decision Tree dan KNN ditampilkan pada Gambar 5.



Gambar 5. Confusion Matrix Decision Tree dan KNN

Pada Decision Tree, model mampu mengklasifikasikan 47 data Asthma/Bronchospasm/Wheezing, 24 data Bronchiolitis, 12 data Laryngeal/Upper Respiratory, 1 data Other, dan 24 data Pneumonia/Pneumopathy dengan benar. Kesalahan klasifikasi masih terjadi terutama pada kelas Pneumonia/Pneumopathy yang sebagian diprediksi sebagai Other dan Bronchiolitis. Pada KNN, model mampu mengklasifikasikan 58 data Asthma/Bronchospasm/Wheezing, 14 data Bronchiolitis, 9 data Laryngeal/Upper Respiratory, 0 data Other, dan 25 data Pneumonia/Pneumopathy dengan benar. KNN tidak mampu mengenali kelas Other secara tepat karena jumlah data pada kelas tersebut sangat sedikit.

3.6. Analisis Perbandingan Model

Berdasarkan hasil evaluasi, Decision Tree memiliki performa lebih baik dibandingkan KNN. Perbedaan tersebut terlihat pada *accuracy*, *precision weighted*, *recall weighted*, dan *F1-score weighted*. Meskipun selisih *accuracy* tidak terlalu besar, nilai *F1-score weighted* menunjukkan bahwa Decision Tree memiliki keseimbangan *precision* dan *recall* yang lebih baik pada dataset tidak seimbang. Decision Tree lebih unggul karena mampu membentuk aturan klasifikasi berdasarkan fitur yang tersedia dan tidak terlalu sensitif terhadap skala data. Sebaliknya, KNN sangat bergantung pada perhitungan jarak. Pada dataset dengan banyak fitur kategorikal hasil *encoding* dan distribusi kelas tidak seimbang, jarak antar data dapat menjadi kurang representatif sehingga performa KNN menurun.

Penting untuk dicatat bahwa generalisasi hasil penelitian ini memiliki batasan. Kesimpulan bahwa Decision Tree lebih baik berlaku spesifik untuk karakteristik dataset yang digunakan (kohort anak di Maroko). Kondisi demografis, genetik, lingkungan, dan sistem kesehatan yang berbeda di wilayah lain (misalnya di Indonesia) dapat mempengaruhi pola gejala dan distribusi penyakit ISPA. Oleh karena itu, hasil klasifikasi ini dimaksudkan sebagai bahan analisis pendukung dalam konteks penelitian informatika kesehatan, bukan untuk menggantikan diagnosis klinis dokter secara universal.

4. KESIMPULAN

Penelitian ini membandingkan algoritma Decision Tree dan K-Nearest Neighbor dalam klasifikasi penyakit ISPA menggunakan dataset *Pediatric Respiratory Infections*. Keterbaruan (novelty) dari penelitian ini adalah penerapan skema klasifikasi lima kategori spesifik pada dataset tersebut, dikombinasikan dengan seleksi fitur yang ketat dan penggunaan metrik evaluasi *weighted* untuk mengatasi bias akibat ketidakseimbangan kelas data medis.

Hasil evaluasi menunjukkan bahwa Decision Tree memiliki performa lebih baik dibandingkan KNN dengan *accuracy* 67,08%, *precision weighted* 73,72%, *recall weighted* 67,08%, dan *F1-score weighted* 69,12%. Sementara itu, KNN memperoleh *accuracy* 65,84%, *precision weighted* 63,98%, *recall weighted* 65,84%, dan *F1-score weighted* 64,18%. Berdasarkan analisis *confusion matrix* dan *feature importance*, Decision Tree lebih mampu mengenali beberapa kelas utama serta memberikan hasil yang lebih mudah diinterpretasikan. Fitur yang paling berpengaruh adalah *Wheezing*, *Laryngeal stridor*, dan *Age (months)*. Penelitian selanjutnya disarankan menggunakan dataset yang

lebih seimbang atau metode penyeimbangan seperti SMOTE, melakukan optimasi parameter lebih lanjut, serta membandingkan dengan algoritma lain seperti *Random Forest*, *Support Vector Machine*, *Naive Bayes*, atau *Gradient Boosting*.

REFERENSI

- [1] N. Nigar, A. Jaleel, S. Islam, M. K. Shahzad, and E. A. Affum, "IoMT Meets Machine Learning: From Edge to Cloud Chronic Diseases Diagnosis System," *Journal of Healthcare Engineering*, vol. 2023, 2023.
- [2] A. S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily, "An explainable machine learning approach for Alzheimer's disease classification," *Scientific Reports*, vol. 14, no. 1, pp. 1-18, 2024.
- [3] W. P. Hanage and W. Schaffner, "Burden of Acute Respiratory Infections Caused by Influenza Virus, Respiratory Syncytial Virus, and SARS-CoV-2 with Consideration of Older Adults: A Narrative Review," *Infectious Diseases and Therapy*, vol. 14, pp. 5-37, 2025.
- [4] J. M. Caldwell, C. M. Espinosa, R. Banerjee, and J. B. Domachowske, "Rapid diagnosis of acute pediatric respiratory infections with Point-of-Care and multiplex molecular testing," *Infection*, vol. 53, Suppl. 1, pp. 1-15, 2025.
- [5] S. Gongora Alonso, G. Marques, D. Agarwal, I. De la Torre Diez, and M. Franco-Martin, "Comparison of Machine Learning Algorithms in the Prediction of Hospitalized Patients with Schizophrenia," *Sensors*, vol. 22, no. 7, 2022.
- [6] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, 2022.
- [7] A. Y. Yildiz and A. Kalayci, "Gradient Boosting Decision Trees on Medical Diagnosis Over Tabular Data," in *2025 IEEE Conference on AI and Data Analytics (ICAD)*, 2025.
- [8] T. T. Tuyen et al., "Prediction of white spot disease susceptibility in shrimps using decision trees based machine learning models," *Applied Water Science*, vol. 14, no. 1, pp. 1-15, 2024.
- [9] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, no. 1, pp. 1-11, 2022.
- [10] B. Sun and H. Chen, "A Survey of k Nearest Neighbor Algorithms for Solving the Class Imbalanced Problem," *Wireless Communications and Mobile Computing*, vol. 2021, 2021.
- [11] B. J. Chadha and R. Singh, "A Comparative Analysis of Decision Tree, K-Nearest Neighbors and Naive Bayes Algorithms for Classification Tasks," vol. 14, no. 04, pp. 2-5, 2025.
- [12] J. E. Simarmata, G. W. Weber, and D. Chrisinta, "Performance Evaluation of Classification Methods on Big Data: Decision Trees, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines," *Journal of Mathematics, Statistics and Computing*, vol. 20, no. 3, pp. 623-638, 2024.
- [13] R. Islam, A. Sultana, and M. R. Islam, "A comprehensive review for chronic disease prediction using machine learning algorithms," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 27, 2024, doi: 10.1186/s43067-024-00150-4.
- [14] M. A. Khan et al., "Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN)," *Scientific Reports*, vol. 14, no. 26241, 2024, doi: 10.1038/s41598-024-78021-1.
- [15] F. Asadi, R. Homayounfar, Y. Mehrali, C. Masci, S. Talebi, and F. Zayeri, "Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms," *Scientific Reports*, vol. 14, no. 22230, 2024, doi: 10.1038/s41598-024-72819-9.
- [16] S. Srinivasan et al., "Detection of Parkinson disease using multiclass machine learning approach," *Scientific Reports*, vol. 14, no. 13813, 2024, doi: 10.1038/s41598-024-64004-9.
- [17] M. Nandagopal et al., "A Deep Auto-Optimized Collaborative Learning (DACL) model for disease prognosis using AI-IoMT systems," *Scientific Reports*, vol. 14, no. 10280, 2024, doi: 10.1038/s41598-024-59846-2.
- [18] A. Altamimi et al., "An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques," *BMC Medical Research Methodology*, vol. 24, no. 221, 2024, doi: 10.1186/s12874-024-02324-0.
- [19] S. Hossain et al., "Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023," *BMC Cardiovascular Disorders*, vol. 24, no. 214, 2024, doi: 10.1186/s12872-024-03883-2.
- [20] P. Chakraborty et al., "Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing," *BMC Bioinformatics*, vol. 25, no. 329, 2024, doi: 10.1186/s12859-024-05866-8.