Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 4 October 2026, pp: 2024-2035
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Early Detection of Pneumonia Based on Clinical Symptoms Using Random Forest

Deteksi Dini Pneumonia Berbasis Gejala Klinis Menggunakan Random Forest

Ridha Mukhtar[1*], Achmad Saleh[2], Rusdah[3]

[1]Program Studi Sistem Informasi, Fakultas Sains dan Teknik, Universitas Sapta Mandiri, Indonesia
[2]Program Studi Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Sapta Mandiri, Indonesia
[3]Fakultas Teknologi Informasi, Universitas Budi Luhur, Indonesia

E-Mail: [1]utar@univsm.ac.id, [2]achmadsaleh209@gmail.com, [3]rusdah@budiluhur.ac.id

Received May 15th 2026; Revised Jun 18th 2026; Accepted Aug 30th 2026; Available Online Oct 30th 2026
Corresponding Author: Ridha Mukhtar
Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Pneumonia is an acute respiratory disease with high morbidity and mortality rates, especially among children and the elderly. Limited access to radiology and laboratory facilities in several regions often hinders early diagnosis of pneumonia. This study aims to develop an early pneumonia detection model based on clinical symptoms using the Random Forest algorithm. The research methodology adopted the CRISP-DM framework, consisting of business understanding, data understanding, data preparation, modeling, evaluation, and deployment stages. The dataset was obtained from medical records at RSUD H. Damanhuri Barabai and included attributes such as gender, age, height, weight, body temperature, cough, sputum, chest pain, shortness of breath, respiratory rate, and vomiting. Model evaluation was conducted using a confusion matrix and 10-fold cross-validation. The results showed that the Random Forest algorithm achieved the best performance with an accuracy of 89.82%, a recall of 95.24%, a precision of 90.53%, and an AUC of 0.9469. These findings indicate that the proposed model is effective for early pneumonia detection and has the potential to support medical decision-making, particularly in healthcare facilities with limited diagnostic resources.

Keywords: Clinical Symptoms, Early Detection, Machine Learning, Pneumonia, Random Forest

Abstrak

Pneumonia merupakan salah satu penyakit pernapasan akut dengan tingkat morbiditas dan mortalitas yang tinggi, terutama pada anak-anak dan lansia. Keterbatasan fasilitas radiologi dan laboratorium di beberapa wilayah menyebabkan proses diagnosis dini pneumonia menjadi kurang optimal. Penelitian ini bertujuan untuk membangun model deteksi dini pneumonia berbasis gejala klinis menggunakan algoritma Random Forest. Metode penelitian mengacu pada kerangka kerja CRISP-DM yang meliputi tahap business understanding, data understanding, data preparation, modelling, evaluation, dan deployment. Dataset penelitian diperoleh dari rekam medis RSUD H. Damanhuri Barabai dengan atribut berupa jenis kelamin, usia, tinggi badan, berat badan, suhu tubuh, batuk, dahak, nyeri dada, sesak napas, laju pernapasan, dan muntah. Evaluasi model dilakukan menggunakan confusion matrix dan 10-fold Cross-validation. Hasil penelitian menunjukkan bahwa algoritma Random Forest menghasilkan performa terbaik dengan nilai akurasi sebesar 89,82%, recall 95,24%, precision 90,53%, dan AUC 0,9469. Hasil ini menunjukkan bahwa model mampu mendeteksi pneumonia secara efektif dan berpotensi menjadi alat bantu pengambilan keputusan medis, khususnya pada fasilitas kesehatan dengan keterbatasan sarana diagnostik.

Kata Kunci: Deteksi Dini, Gejala Klinis, Machine Learning, Pneumonia, Random Forest

1. PENDAHULUAN

Pneumonia merupakan salah satu penyakit infeksi saluran pernapasan akut yang menjadi penyebab utama morbiditas dan mortalitas di dunia, terutama pada kelompok rentan seperti balita dan lansia. Menurut *World Health Organization* (WHO), pneumonia menyebabkan sekitar 15% kematian pada anak di bawah usia lima tahun setiap tahunnya [1]. Selain itu, penyakit ini juga menjadi salah satu penyebab utama rawat inap akibat infeksi saluran pernapasan. Gejala pneumonia seperti demam, batuk, sesak napas, dan nyeri dada sering

kali menyerupai penyakit paru lainnya sehingga dapat menyebabkan keterlambatan diagnosis dan penanganan medis apabila tidak terdeteksi sejak dini.

Dalam praktik medis, diagnosis pneumonia umumnya dilakukan melalui pemeriksaan radiologi seperti *chest* X-ray, CT *scan*, maupun pemeriksaan laboratorium. Namun, fasilitas tersebut tidak selalu tersedia secara memadai, khususnya di fasilitas kesehatan primer dan daerah terpencil yang memiliki keterbatasan infrastruktur kesehatan. Kondisi ini menyebabkan proses diagnosis dini pneumonia menjadi kurang optimal dan berpotensi meningkatkan risiko komplikasi serius seperti gagal napas, sepsis, hingga kematian [2], [3]. Oleh karena itu, diperlukan pendekatan alternatif yang lebih sederhana, cepat, dan terjangkau untuk membantu tenaga medis melakukan deteksi dini pneumonia.

Perkembangan teknologi *machine learning* memberikan peluang baru dalam membantu proses klasifikasi penyakit berdasarkan data klinis pasien. Pendekatan *machine learning* memungkinkan sistem melakukan prediksi otomatis menggunakan pola data gejala klinis tanpa harus bergantung sepenuhnya pada pemeriksaan radiologi [4]. Salah satu algoritma yang banyak digunakan dalam klasifikasi medis adalah Random Forest karena mampu menangani data kompleks, fitur nonlinier, serta menghasilkan performa klasifikasi yang stabil dan akurat [5].

Beberapa penelitian terdahulu menunjukkan bahwa penerapan machine learning dalam deteksi pneumonia memberikan hasil yang menjanjikan dan sedang mendapat perhatian [6]. Penelitian [7] menggunakan algoritma Naïve Bayes untuk diagnosis pneumonia pada balita dan memperoleh tingkat akurasi sebesar 87,57%. Selain itu, penelitian [8] membandingkan beberapa algoritma seperti Logistic Regression, SVM, Naïve Bayes, dan Random Forest pada data rekam medis pasien, di mana Random Forest menunjukkan sensitivitas dan performa terbaik dibandingkan algoritma lainnya. Penelitian [9] juga menunjukkan bahwa model machine learning mampu menghasilkan AUROC rata-rata sebesar 0,88 dalam prediksi pneumonia terkait ventilator.

Pada sebuah Penelitian membuktikan bahwa dalam melakukan deteksi infeksi mycoplasma pneumoniae pneumonia dapat dilakukan dengan menerapkan model *machine learning*. Hasil akhir menunjukan bahwa metode modifikasi *hyper* parameter dengan Hyperband memiliki performa klasifikasi yang sedikit lebih baik jika dibandingkan dengan *Grid Search* dan *Random Search* dengan nilai f-*score* dan akurasi masing-masing yaitu 0.887 dan 0.894 untuk Decision Tree, 0.942 dan 0.947 untuk Logistic Regression, 0.910 dan 0.915 untuk Gradient Boosting Decision Tree serta 0.591 dan 0.715 untuk Support Vector Machine [10] .

Pemilihan algoritma Random Forest tidak hanya didasarkan pada hasil penelitian terdahulu, tetapi juga pada karakteristik dataset yang digunakan. Dataset penelitian terdiri atas kombinasi atribut numerik dan kategorikal dengan kemungkinan hubungan nonlinier antarvariabel klinis. Random Forest dipilih karena mampu menangani karakteristik data tersebut secara efektif melalui mekanisme *ensemble learning*, memiliki ketahanan terhadap *overfitting*, mampu melakukan seleksi fitur secara implisit, serta tetap menghasilkan performa yang stabil pada ukuran dataset yang relatif terbatas. Karakteristik ini menjadikan Random Forest sesuai untuk membangun model klasifikasi pneumonia berbasis gejala klinis yang akurat dan mudah diimplementasikan pada lingkungan layanan kesehatan. Namun untuk memperoleh performa yang optimal, diperlukan beberapa strategi seperti melakukan seleksi fitur, mengatur kompleksitas pohon melalui parameter *max_depth*, *min_samples_split*, dan *min_samples_leaf*, serta menerapkan validasi silang (*Cross-validation*). Pendekatan tersebut membantu model memperoleh kemampuan generalisasi yang lebih baik terhadap data baru dan mengurangi kemungkinan model mempelajari pola yang hanya muncul pada data pelatihan [11].

Meskipun penelitian ini melibatkan beberapa algoritma klasifikasi sebagai pembanding, fokus utama penelitian adalah pengembangan model deteksi dini pneumonia menggunakan Random Forest. Decision Tree dan Gradient Boosting digunakan sebagai *baseline* model untuk mengevaluasi keunggulan Random Forest pada dataset yang digunakan

Selain berbasis citra radiologi, penelitian terkait deteksi pneumonia berbasis gejala klinis juga menunjukkan hasil yang cukup efektif. Penggunaan atribut seperti suhu tubuh, batuk, dahak, sesak napas, laju pernapasan, dan nyeri dada terbukti relevan dalam proses klasifikasi pneumonia. Pendekatan berbasis gejala klinis dinilai lebih praktis karena dapat diterapkan pada fasilitas kesehatan dengan keterbatasan alat diagnostik modern. Kajian literatur juga menunjukkan bahwa sistem berbasis *machine learning* dapat membantu tenaga medis dalam pengambilan keputusan awal dengan sensitivitas dan spesifisitas yang tinggi.

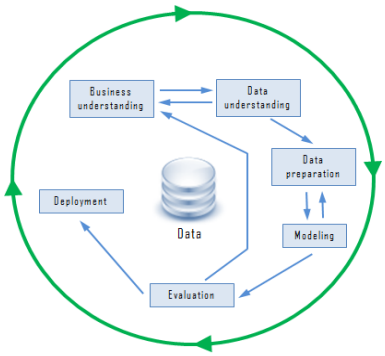
Penelitian ini dilakukan karena masih terbatasnya implementasi sistem deteksi dini pneumonia berbasis gejala klinis menggunakan data rekam medis lokal di Indonesia, khususnya pada rumah sakit daerah. Penelitian ini memanfaatkan data rekam medis dari RSUD H. Damanhuri Barabai dengan atribut klinis seperti jenis kelamin, usia, suhu tubuh, batuk, dahak, nyeri dada, sesak napas, laju pernapasan, dan muntah. Dengan memanfaatkan algoritma Random Forest, penelitian ini mampu menghasilkan model deteksi dini pneumonia yang akurat, murah, dan mudah diimplementasikan sebagai alat bantu pengambilan keputusan medis awal sebelum dilakukan pemeriksaan radiologi atau laboratorium.

Berdasarkan latar belakang dan kajian pustaka tersebut, penelitian ini bertujuan untuk menganalisis performa algoritma Random Forest dalam mendeteksi dini pneumonia berbasis parameter gejala klinis pasien. Penelitian ini juga bertujuan untuk mengetahui apakah model machine learning yang dibangun mampu

memberikan tingkat akurasi, sensitivitas, dan Area Under the ROC Curve (AUC) yang baik dalam membantu proses klasifikasi pneumonia secara otomatis.

2. METODE PENELITIAN

Penelitian ini menerapkan *machine learning* untuk mengenali risiko pneumonia sejak dini berdasarkan parameter klinis pasien. Metodologi yang digunakan mengacu pada siklus CRISP-DM (*Cross Industry Standard Process for Data Mining*) dengan enam tahapan utamanya, mulai dari *Business Understanding* hingga *Deployment*. Dalam prosesnya, tidak ada prasyarat khusus untuk data mentah yang digunakan. Hal ini dikarenakan data tersebut secara otomatis akan disaring dan dibentuk melalui berbagai tahapan pengolahan yang ada di dalam kerangka CRISP-DM [12]. Adapun fase dalam CRISP-DM berjumlah 6 yang dijelaskan pada gambar 1.



Gambar 1. Fase CRISP-DM [13]

2.1. *Business Understanding*

Merujuk pada pedoman klinis, penegakan diagnosis pneumonia tergantung pada pengujian laboratorium dan CT Scan toraks guna mengevaluasi kadar leukosit serta sampel sputum [2]. Kendala utamanya adalah metode konvensional ini tergolong mahal dan sangat bergantung pada kelengkapan infrastruktur kesehatan, sehingga kurang aplikatif di daerah terpencil. Padahal, kecepatan penanganan sangatlah krusial. Penelitian oleh [14] menyatakan keterlambatan dalam menegakkan diagnosis pneumonia dapat menghambat pemberian terapi yang tepat waktu, sehingga berpotensi memperburuk kondisi pasien. Dampak yang ditimbulkan antara lain meningkatnya kebutuhan perawatan di ruang intensif (ICU) serta risiko terjadinya komplikasi yang lebih serius.. Oleh sebab itu, pemanfaatan teknologi *machine learning* untuk deteksi dini dinilai menjadi solusi yang potensial. Selain dapat mempercepat identifikasi penyakit, metode ini juga berpeluang besar untuk meminimalkan beban tenaga medis dan membuka akses kesehatan yang lebih merata.

2.2. Data *Understanding*

Penelitian ini berfokus pada pengembangan model deteksi dini penyakit pneumonia dengan menganalisis parameter gejala klinis pasien. Data yang digunakan bersumber dari rekam medis pasien di RSUD H. Damanhuri, Kabupaten Hulu Sungai Tengah, dengan total keseluruhan mencapai 334 data yang mencakup 11 atribut gejala. Untuk menentukan kesebelas atribut tersebut, peneliti memadukan tinjauan literatur dengan hasil wawancara bersama dokter spesialis penyakit dalam dan kepala bagian rekam medis. Guna mematuhi kode etik kesehatan, seluruh identitas pasien disamarkan dan data diproses secara anonim untuk menjamin kerahasiaan dan privasi pasien.

2.3. Data *Preparation*

Dalam menyusun pemodelan ini, proses seleksi atribut diterapkan untuk memastikan hanya data yang benar-benar relevan yang dipakai untuk mendiagnosis pneumonia pada pasien anak usia 0–10 tahun di RSUD H. Damanhuri Barabai. Fitur-fitur medis seperti usia, jenis kelamin, tinggi dan berat badan, indikator vital (nadi, suhu, laju napas), hingga keluhan pernapasan (batuk, berdahak, sesak, dan nyeri dada) dipertahankan karena terbukti berkontribusi langsung pada penegakan diagnosis secara klinis.

Di sisi lain, variabel administratif seperti ID pasien, lokasi bangsal, tanggal masuk, hingga kategori perawatan harus dibuang guna menghindari bias dalam proses klasifikasi. Guna memperkuat keputusan seleksi ini, peneliti memanfaatkan analisis *feature importance* bawaan algoritma Random Forest. Pendekatan ini menjamin bahwa hanya atribut dengan bobot pengaruh terkuatlah yang dimasukkan ke dalam pemodelan, yang pada akhirnya akan mendongkrak tingkat akurasi dan meminimalkan potensi *overfitting*.

2.4. Modeling

Menurut Wisdatuti [15], Tingkat keberhasilan sebuah model klasifikasi sangat dipengaruhi oleh proses seleksi fitur (*feature selection*), kualitas data yang digunakan, serta ketepatan memilih algoritma yang selaras dengan karakteristik data itu sendiri. Setiap algoritma tentu memiliki keunggulannya masing-masing; ada yang sangat optimal ketika memproses data berbentuk angka (numerik), namun ada pula yang lebih tangguh dalam menangani data kategorikal maupun data dengan distribusi yang tidak seimbang (*imbalanced* data). Lebih dari sekadar memilih algoritma yang tepat, tahapan validasi juga memegang peranan krusial dalam memastikan keandalan hasil penelitian yang dilakukan. Teknik seperti k-fold cross-validation digunakan untuk memastikan bahwa model tidak hanya unggul di data latih, tetapi juga memiliki kemampuan generalisasi terhadap data uji. Evaluasi model klasifikasi dilakukan dengan metrik seperti akurasi, presisi, *recall*, dan AUC, yang masing-masing memberikan gambaran berbeda tentang performa model, terutama dalam konteks data tidak seimbang seperti kasus diagnosis penyakit [16]. Random Forest ditetapkan sebagai model utama penelitian, sedangkan Decision Tree dan Gradient Boosting digunakan sebagai model pembanding untuk mengevaluasi performa klasifikasi.

2.4.1. Decision Tree

Dijelaskan pada [17] persamaan yang menjabarkan tentang Decision Tree mulai dari klasifikasi sederhana dan lainnya. Persamaan matematika untuk Decision Tree pada klasifikasi sederhana dengan dua kelas (positif dan negatif) dapat dilihat pada Persamaan 1.

1. Hitung entropy (E) dari dataset:

$$E = -p_pos\ log2\ p_pos - p_neg\ log2\ p_neg \quad (1)$$

Dengan: p_pos adalah proporsi data yang termasuk kelas positif dan p_neg adalah proporsi data yang termasuk kelas negatif.

2. Hitung gain (G) dari setiap atribut (A) pada dataset, dapat dilihat pada persamaan (2).

$$G(A) = E(D) - \Sigma(Di/D) * E(Di) \quad (2)$$

Dengan: E(D) adalah entropy awal dataset; Di adalah subset data yang memiliki nilai atribut A = i; D adalah jumlah total data pada dataset; E(Di) adalah entropy subset data dengan nilai atribut A = i.

3. Pilih atribut dengan gain tertinggi sebagai split node pada decision tree.

Ulangi langkah b dan c untuk setiap subset data pada setiap split node hingga mencapai kondisi terminasi, seperti mencapai kedalaman maksimal atau seluruh data pada subset memiliki kelas yang sama.

2.4.2. Gradient Boosting

Algoritma ini seringkali menghasilkan akurasi tinggi, tetapi lebih sensitif terhadap parameter *tuning* dan lebih berisiko *overfitting* jika tidak dikontrol dengan baik [16].

Gradient Boosting bekerja dengan cara menyusun model prediksi secara bertahap. Prinsip utamanya adalah terus memperbaiki kelemahan dari model-model terdahulu dengan meminimalkan nilai *loss function*. Secara teknis, setiap *weak learner* yang ditambahkan akan dilatih secara spesifik untuk menangani sisa eror atau residual (*negative gradient*) dari hasil prediksi sebelumnya. Proses perbaikan yang dilakukan secara berulang (iteratif) inilah yang membuat model mampu mendongkrak tingkat akurasinya [18]. Model awal dinyatakan pada persamaan 3.

$$F_0(x) = arg\ \min_{\gamma} \sum_{i=1}^{n} (y_i - \gamma)^2 \quad (3)$$

Pada setiap iterasi, residual dihitung menggunakan Persamaan 4.

$$r_i^{(m)} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad (4)$$

Model kemudian diperbarui dengan menambahkan *weak learner* baru menggunakan Persamaan 5.

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \quad (5)$$

dengan parameter *learning rate* ν berfungsi sebagai regularisasi untuk mengurangi risiko *overfitting*. Setelah seluruh iterasi selesai, prediksi akhir model diperoleh melalui Persamaan 6.

$$\hat{y}(x) = F_M(x) = \sum_{m=1}^{M} \nu \cdot h_m(x) \quad (6)$$

2.4.3. Random Forest

Random Forest bekerja dengan membangun sekumpulan decision tree acak dan menentukan hasil akhir melalui *majority* voting. Pendekatan ini ampuh meningkatkan akurasi dan menghindari *overfitting* yang kerap menimpa Decision Tree biasa. Kelebihan lainnya, algoritma ini mampu mengidentifikasi tingkat kontribusi tiap fitur (*feature importance*) terhadap hasil prediksi. [19]. Contoh persamaan dalam persamaan algoritma pada machine learning seperti random forest ditunjukkan pada Persamaan 7.

$$\text{Entropy(Y)} \ = \ -\sum_{1} p\,(\text{c|Y})\text{log2p(c} \ \vee \ \text{Y)} \quad (7)$$

Information Gain, Y merupakan kumpulan kasus dan dan p (c|Y) meupakan rasio nilai Y terhadap kelas c.

$$\text{InformationGain(y, a)} \ = \text{Entropy(Y)} \ - \sum_{veValues} \frac{|Yv|}{|Ya|} Entropy\ (Yv) \quad (8)$$

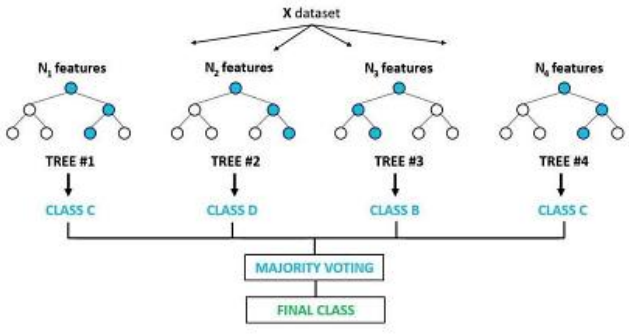
Sedangkan Values (a) yaitu seluruh nilai yang memungkinkan pada suatu kumpulan kasus: a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a. Atribut yang dipilih sebagai simpul, baik akar (root) maupun simpul internal, akan disesuaikan dengan *information gain* yang diperoleh melalui hasil tertinggi pada atribut-atribut yang tersedia. Penentuannya akan menggunakan rumus gain ratio, sedangkan gain ratio diperoleh dengan menentukan nilai split information, seperti dijelaskan pada persamaan (9) dan (10).

$$SplitInformation\ (\text{S, A}) = \sum_{i=1}^{c} \left(\frac{|Si|}{|S|}\right) \log_2 \left(\frac{|Si|}{|S|}\right) \quad (9)$$

Split Information (S,A) merupakan nilai untuk estimasi *entropy* pada variable *input* S dengan kelas c dan |*Si*|/|*S*| yaitu probabilitas kelas i pada atribut.

$$GainRation\ (\text{S, A}) = \frac{InformationGain\ (S, A)}{SplitInformation\ (S, A)} \quad (10)$$

Pendekatan ini tidak hanya meningkatkan akurasi dan kestabilan model, tetapi juga efektif mencegah *overfitting*, sehingga sangat ideal digunakan untuk mendeteksi penyakit dari data klinis medis.



Gambar 2 Pohon Keputusan Random Forest [20]

2.5. *Evaluation*

Evaluasi performa penting dilakukan guna menguji tingkat akurasi model klasifikasi. Dalam konteks medis seperti deteksi pneumonia, penentuan metrik evaluasi yang tepat mutlak diperlukan karena kesalahan prediksi berdampak serius pada keputusan penanganan pasien. Untuk mengukur keandalan model tersebut, parameter yang biasa digunakan adalah *Accuracy*, *Precision*, dan *Recall*, yang keseluruhannya dirumuskan berdasarkan nilai *confusion matrix* (TP, TN, FP, dan FN) [21], [22]. *Accuracy*, *Precision*, dan *Recall* adalah dapat ditunjukkan pada persamaan 11-13.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

Untuk meningkatkan keandalan evaluasi model, digunakan metode k-fold cross-validation dengan membagi dataset menjadi beberapa bagian *(fold).* Metode ini merotasi pembagian data latih dan uji secara bergantian, sehingga seluruh dataset dimanfaatkan secara seimbang untuk menyeleksi algoritma deteksi pneumonia berdasarkan parameter gejala klinis yang paling optimal.

2.6. *Deployment*

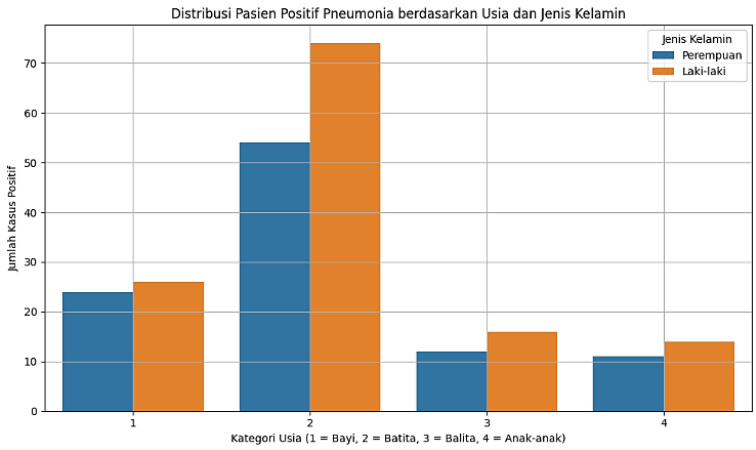
Setelah model terbaik diperoleh, model tersebut kemudian diterapkan pada aplikasi web sederhana berbasis Streamlit. Sistem ini memungkinkan pengguna memasukkan data gejala pasien secara manual atau melalui unggahan dataset, lalu memproses data tersebut secara otomatis. Selain itu, sistem juga menampilkan hasil preprocessing, jumlah data yang diproses, serta visualisasi hasil prediksi untuk menunjukkan distribusi pasien yang terindikasi dan tidak terindikasi pneumonia.

3. HASIL DAN PEMBAHASAN

Bagian ini menjelaskan hasil penerapan *machine learning* untuk deteksi dini pneumonia berdasarkan gejala klinis pasien. Penelitian dilakukan menggunakan tahapan CRISP-DM mulai dari pemahaman masalah, pengolahan data, pemodelan, hingga evaluasi. Beberapa algoritma seperti Random Forest, Decision Tree, dan Gradient Boosting dibandingkan menggunakan metrik *accuracy*, *precision*, *recall*, dan AUC untuk menentukan model terbaik. Selain itu, model dengan performa terbaik juga diimplementasikan ke dalam sistem berbasis web agar proses deteksi dini pneumonia dapat dilakukan dengan lebih cepat dan efisien.

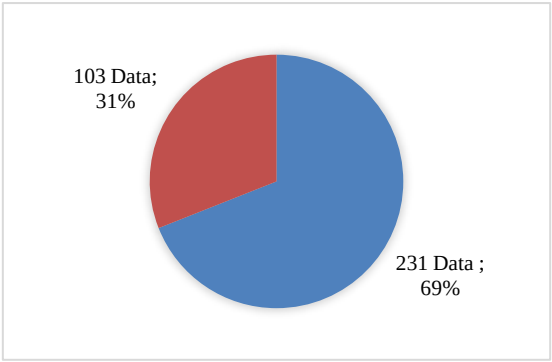
3.1. Dataset

Penelitian ini bertujuan untuk mengembangkan model deteksi dini pneumonia berdasarkan gejala klinis pasien menggunakan data rekam medis dari RSUD H. Damanhuri Kabupaten Hulu Sungai Tengah. Dataset yang digunakan terdiri dari 334 data pasien anak usia 0–10 tahun dengan 11 atribut gejala. Pemilihan atribut dilakukan untuk memastikan hanya data yang relevan dan berpengaruh terhadap deteksi pneumonia yang digunakan dalam proses pemodelan.



Gambar 3 Visualisasi Dataset Awal

Visualisasi pada Gambar 3 menunjukkan distribusi pasien positif pneumonia berdasarkan kategori usia dan jenis kelamin. Kelompok batita laki-laki (kategori usia 2) memiliki jumlah kasus tertinggi dengan lebih dari 70 kasus, diikuti batita perempuan dengan lebih dari 50 kasus. Sementara itu, pada kelompok balita dan anak-anak jumlah kasus terlihat menurun, dengan distribusi yang relatif seimbang antara pasien laki-laki dan perempuan.



Gambar 4 Distribusi Label pada dataset

Gambar 4 menunjukkan distribusi label pada dataset deteksi dini pneumonia yang terdiri dari 231 data terindikasi pneumonia (69%) dan 103 data tidak terindikasi pneumonia (31%). Meskipun jumlah data antar kelas belum sepenuhnya seimbang, komposisi tersebut masih cukup baik untuk proses klasifikasi pada kasus medis. Kondisi ini umum terjadi karena sebagian besar data berasal dari pasien dengan gejala yang mengarah pada pneumonia. Analisis distribusi label dilakukan agar model machine learning tetap mampu mengenali kedua kelas dengan baik dan tidak hanya fokus pada kelas mayoritas. Setelah proses *filterisasi* awal dilakukan, diperoleh dataset awal seperti yang ditunjukkan pada Tabel 1.

Tabel 1 Dataset Awal Hasil Pemilihan Atribut

Usia (Tahun)	JK	Tinggi Badan	Berat Badan	Nadi	Suhu Badan	Laju Nafas	Nyeri Dada	Sesak Nafas	Batuk	Dahak	Muntah	Label
0	L	43	3	139	40	22	No	Yes	No	Yes	Yes	Terindikasi Pneumonia
0	P	43	3,3	110	38	35	Yes	Yes	No	Yes	Yes	Terindikasi Pneumonia
0	P	43	3,3	139	40	22	No	No	No	No	No	Tidak Terindikasi Pneumonia
0	L	45	3	136	38	22	Yes	Yes	No	Yes	Yes	Terindikasi Pneumonia
0	L	45	3,2	139	39	30	No	Yes	No	No	No	Tidak Terindikasi Pneumonia

3.2. Data *Preparation*

Data hasil *query* rekam medis terlebih dahulu dilakukan proses perbaikan dan penyesuaian sebelum digunakan dalam penelitian. Data deskriptif dari rekam medis diubah ke dalam format Excel agar lebih mudah dianalisis dan diolah. Tahap selanjutnya adalah *data cleansing*, yaitu pengecekan struktur, format, dan konsistensi data sesuai kebutuhan penelitian. Selain itu, dilakukan penanganan *missing value* pada data yang kosong atau tidak terisi. Nilai suhu tubuh yang kosong diisi menggunakan rata-rata suhu tubuh pasien, sedangkan data batuk yang kosong diisi dengan nilai "Ya" berdasarkan pertimbangan penelitian dari [23] yang menyatakan gejala batuk tidak selalu muncul pada awal penyakit, tetapi dapat berkembang dan menjadi lebih nyata setelah beberapa hari perjalanan infeksi seiring dengan keterlibatan jaringan paru yang lebih luas. Selanjutnya melakukan transformasi data karena sebagian besar algoritma seperti Decision Tree, Random Forest membutuhkan input berupa angka, mengurangi resiko error saat training model, serta maka dilakukan transformasi data kategorikal menjadi *numerical*.

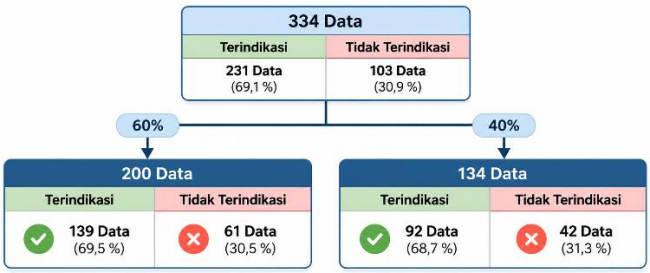
Hasil akhir dari tahap data *preparation* adalah dataset terstruktur dan bersih yang siap untuk diterapkan pada algoritma machine learning guna mendeteksi pneumonia secara dini. Proses ini memastikan bahwa data yang digunakan valid, konsisten, dan berkualitas tinggi, sehingga hasil analisis dapat diandalkan. seperti tabel dataset training yang ditunjukkan pada Tabel 2.

Tabel 2. Dataset *Training*

Jenis Kelamin	Kategori Usia	Tinggi Badan	Berat Badan	Nadi	Suhu Badan	Laju Pernafasan	Nyeri dada	Sesak nafas	Batuk	Dahak	Muntah	Label
0	1	45	3	115	38	35	1	0	1	1	1	1
Jenis Kelamin	Kategori Usia	Tinggi Badan	Berat Badan	Nadi	Suhu Badan	Laju Pernafasan	Nyeri dada	Sesak nafas	Batuk	Dahak	Muntah	Label

Jenis Kelamin	Kategori Usia	Tinggi Badan	Berat Badan	Nadi	Suhu Badan	Laju Pernafasan	Nyeri dada	Sesak nafas	Batuk	Dahak	Muntah	Label
0	1	45	3	133	37	30	1	1	0	0	1	1
0	1	45	3,2	140	40	30	0	1	1	1	1	0
1	1	47	3,2	136	36	22	1	1	0	0	1	0
0	1	47,4	3,5	110	38	30	1	1	1	1	1	1

Sebelum memasuki tahap pemodelan, langkah awal yang dilakukan adalah memisahkan data fitur dan target dalam dataset. Fitur terdiri dari seluruh kolom kecuali kolom 'Label', sedangkan kolom 'Label' berfungsi sebagai target.



Gambar 5. Penentuan data latih dan data uji menggunakan stratified random sampling pada Dataset

Setelah proses pemisahan data dilakukan, dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 60:40. Sebanyak 60% data digunakan untuk melatih model, sedangkan 40% sisanya digunakan untuk menguji performa model pada data baru. Pembagian ini bertujuan agar model dapat mempelajari pola data dengan baik sekaligus mengurangi risiko *overfitting*, sehingga hasil evaluasi lebih mampu menggambarkan performa model pada kondisi nyata.

3.3. *Modeling*

Proses pemodelan dilakukan menggunakan tiga algoritma klasifikasi, yaitu Random Forest, Decision Tree, dan Gradient Boosting. Evaluasi dilakukan melalui dua pendekatan utama, yaitu split data (80:20, 70:30, 60:40) dan 10-Fold cross-validation, serta dilengkapi dengan proses tuning hyperparameter menggunakan metode manual dan GridSearchCV.

3.3.1. *Perbandingan Model Machine Learning*

Tabel 3 menyajikan hasil evaluasi model untuk tiga algoritma *machine learning* Random Forest, Decision Tree, dan Gradient Boosting, pada tiga skenario pembagian data (*split data*): 80:20, 70:30, dan 60:40. Evaluasi dilakukan menggunakan metrik *Accuracy* dan AUC.

Tabel 3 Hasil Teknik Split Data

Komposisi Dataset	*Random Forest*		*Decision Tree*		*Gradient Boosting*	
	Accuracy	AUC	*Accuracy*	AUC	*Accuracy*	AUC
80:20:00	86.9	0.91	85.1	0.9	85.1	0.9
70:30:00	89	0.9	85.9	0.79	88.4	0.9
60:40:00	89	0.94	86.5	0.85	88.5	0.9

Hasil pemodelan menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dibandingkan dengan Decision Tree dan Gradient Boosting pada seluruh skenario pembagian data. Pada pembagian data 80:20, Random Forest memperoleh akurasi 86.9% dengan AUC 0.91. Pada *split* 70:30, akurasi meningkat menjadi 89.0% dengan AUC 0.90, sedangkan pada split 60:40 model kembali mencapai akurasi 89.0% dengan AUC tertinggi sebesar 0.94. Hasil ini menunjukkan bahwa Random Forest mampu memberikan prediksi yang akurat dan stabil dalam mendeteksi dini pneumonia.

Sementara itu, Decision Tree menghasilkan akurasi antara 85.1%–86.5% dengan AUC 0.79–0.90, namun performanya cenderung kurang stabil terutama pada split 70:30. Gradient Boosting memberikan hasil yang cukup baik dengan akurasi 85.1%–88.5% dan AUC sekitar 0.90, tetapi masih berada di bawah Random Forest. Secara keseluruhan, evaluasi menggunakan metrik *Accuracy* dan AUC menunjukkan bahwa Random Forest merupakan algoritma paling optimal dan konsisten untuk deteksi dini pneumonia karena memiliki tingkat akurasi tinggi serta kemampuan yang baik dalam membedakan kasus positif dan negatif.

3.3.2. 10 Fold Cross-validation

Kemudian melakukan model dengan pendekatan 10 fold Cross-validation digunakan untuk mendapatkan estimasi performa model yang lebih stabil dan tidak tergantung pada satu komposisi dataset tertentu di jabarkan pada Table 4.

Tabel 4 Hasil Teknik 10 Fold Cross-validation

Model	Accuracy	Recall (Sensitivitas)	Precision	AUC
Random Forest	0.8982	0.9524	0.9053	0.9469
Decision Tree	0.8413	0.8831	0.887	0.8153
Gradient Boosting	0.8842	0.9481	0.8963	0.9403

Random Forest memiliki akurasi lebih baik pada 0.89 diikuti oleh gradient boosting dengan akurasi 0.88. Pada sensitivitas random forest kebali terlihat lebih baik dengan 0.95 dan 0.94 pada gradient boosting yang mana ini menjelaskan dalam konteks deteksi penyakit sangat penting karena dapat mengenali lebih banyak kasus positif. Performa dalam menganalisa *false posistif* juga ditunjukkan oleh random forest dengan hasil *precision* 0.90 dan di ikuti gradient boosting dengan 0.89.

Secara keseluruhan, hasil menunjukkan bahwa model Random Forest layak untuk dipilih sebagai algoritma utama dalam sistem deteksi dini pneumonia yang dikembangkan, karena memberikan keseimbangan yang optimal antara akurasi dan kemampuan diskriminasi model yakni pada AUC juga ditunjukkan random forest.

Berdasarkan hasil evaluasi performa model, algoritma Random Forest dipilih sebagai metode terbaik untuk deteksi dini pneumonia karena memiliki akurasi tertinggi yang menunjukkan kemampuan model dalam mengenali kasus pneumonia secara akurat dan konsisten. Pada tahap pemodelan, algoritma random forest dipilih sebagai pendekatan klasifikasi utama untuk menilai kesejahteraan desa. Algoritma ini dipilih karena kemampuannya dalam menangani kompleksitas dataset, menangani fitur nonlinier, dan memberikan estimasi probabilitas kelas. Selanjutnya, proses tuning hyperparameter dilakukan untuk mengoptimalkan performa model.

3.3.3. *Tuning* Manual dan GridSearch CV

Proses *tuning* dilakukan untuk mendapatkan kombinasi parameter terbaik pada algoritma Random Forest. Pengujian dilakukan menggunakan dua pendekatan manual dan satu pendekatan GridSearchCV dengan evaluasi berdasarkan nilai *Mean Accuracy* dan Mean AUC hasil cross-validation.

Tabel 5 Hasil Tuning Manual dan Gridsearch CV

Metode *Tuning*	*max_depth*	*min_samples_leaf*	*n_estimators*	*max_features*	*Accuracy* CV	AUC CV
Manual-1	5	2	100	sqrt	0.863	0.9155
Manual-2	10	2	200	sqrt	0.8692	0.9136
GridSearchCV	*None*	4	200	sqrt	0.8842	0.9238

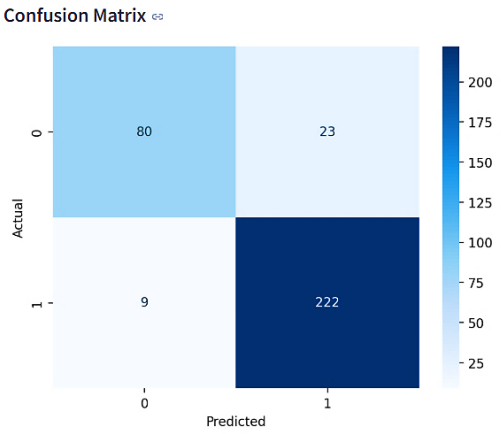
Pada tuning manual pertama dengan parameter *max_depth* = 5, min_*samples_leaf* = 2, n_*estimators* = 100, dan max_*features* = sqrt, model menghasilkan *Mean Accuracy* 0.8630 dan *Mean* AUC 0.9155. *Tuning* manual kedua dengan *max_depth* = 10 dan n_*estimators* = 200 meningkatkan akurasi menjadi 0.8692 dengan *Mean* AUC 0.9136. Sementara itu, pendekatan GridSearchCV memberikan hasil terbaik dengan parameter *max_depth* = None, *min_samples_leaf* = 4, n_estimators = 200, dan *max_features* = sqrt. Model ini memperoleh *Mean Accuracy* sebesar 0.8842 dan *Mean* AUC 0.9238.

Berdasarkan hasil tersebut, GridSearchCV dipilih sebagai pendekatan terbaik karena mampu menghasilkan performa yang lebih tinggi melalui proses pencarian parameter yang lebih sistematis. Secara keseluruhan, Random Forest dengan 10-Fold Cross-validation menjadi model terbaik dengan *accuracy* 89.82%, *recall* 95.24%, *precision* 90.53%, dan AUC 0.9469 sehingga dinilai efektif untuk mendukung deteksi dini pneumonia.

3.4. Evaluasi Performa

Untuk menilai kinerja model secara menyeluruh, dilakukan analisis menggunakan *confusion matrix* sebagai salah satu alat evaluasi utama. *Confusion matrix* memberikan gambaran detail mengenai jumlah prediksi yang benar dan salah dari model terhadap masing-masing kelas (positif dan negatif). Dengan melihat metrik seperti *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN), kita dapat menghitung berbagai indikator performa seperti akurasi, sensitivitas (*recall*), dan presisi. Hal ini sangat krusial dalam konteks deteksi dini pneumonia, di mana kesalahan klasifikasi — terutama kesalahan dalam mengenali

kasus positif — dapat berdampak serius pada keputusan klinis. Gambar 3.5 berikut menampilkan *confusion matrix* hasil prediksi model pada data uji.



Gambar 6. Confusion Matrix

Gambar 6 menunjukkan hasil evaluasi model Random Forest menggunakan *confusion matrix* pada data pengujian. Berdasarkan matriks tersebut, model berhasil mengklasifikasikan 222 data pneumonia sebagai pneumonia (*True Positive*) dan 80 data non-pneumonia sebagai non-pneumonia (*True Negative*). Namun, masih terdapat 23 data non-pneumonia yang diprediksi sebagai pneumonia (*False Positive*) serta 9 data pneumonia yang diprediksi sebagai non-pneumonia (*False Negative*). Secara keseluruhan, model mampu mengklasifikasikan 302 dari 334 data dengan benar, yang menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan kasus pneumonia dan non-pneumonia.

Berdasarkan hasil confusion matrix, model memperoleh nilai accuracy sebesar 90,42%, yang menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan tepat. Selain itu, nilai precision sebesar 90,61% mengindikasikan bahwa mayoritas prediksi pneumonia yang dihasilkan model memang merupakan kasus pneumonia yang sebenarnya. Nilai F1-Score sebesar 93,27% juga menunjukkan keseimbangan yang baik antara precision dan recall, sehingga model memiliki performa yang stabil dalam melakukan klasifikasi.

Nilai *recall* (*sensitivity*) sebesar 96,10% menjadi salah satu keunggulan utama model karena menunjukkan kemampuan yang sangat baik dalam mendeteksi kasus pneumonia. Tingginya recall mengindikasikan bahwa hanya sedikit kasus pneumonia yang tidak teridentifikasi oleh model, yaitu sebanyak 9 kasus (*False Negative*). Dalam konteks layanan kesehatan, kondisi ini sangat penting karena kesalahan dalam mendeteksi pasien pneumonia dapat berdampak pada keterlambatan penanganan medis. Meskipun nilai *specificity* sebesar 77,67% menunjukkan masih adanya sejumlah kasus non-pneumonia yang terdeteksi sebagai pneumonia, kesalahan tersebut relatif lebih dapat ditoleransi karena pasien masih dapat menjalani pemeriksaan lanjutan untuk memastikan diagnosis. Oleh karena itu, model Random Forest dapat dikatakan efektif untuk digunakan sebagai alat bantu deteksi dini pneumonia berdasarkan data gejala klinis.

3.5. Diskusi

Hasil penelitian menunjukkan bahwa algoritma Random Forest mampu memberikan performa yang baik dalam deteksi dini pneumonia berbasis gejala klinis dengan akurasi sebesar 89,82%, *recall* 95,24%, *precision* 90,53%, dan AUC 0,9469. Tingginya nilai *recall* menunjukkan bahwa model mampu mengidentifikasi sebagian besar kasus pneumonia dengan benar, sehingga berpotensi mengurangi risiko keterlambatan diagnosis akibat kasus positif yang tidak terdeteksi. Performa tersebut diduga diperoleh karena Random Forest menggunakan mekanisme *ensemble learning* yang menggabungkan banyak pohon keputusan sehingga mampu menangkap hubungan nonlinier antarvariabel klinis, mengurangi varians model, dan meningkatkan kemampuan generalisasi pada dataset berukuran terbatas.

Penelitian ini juga konsisten dengan beberapa penelitian sebelumnya yang menunjukkan keunggulan Random Forest dalam klasifikasi pneumonia. Effah et al. (2022) [6] melaporkan bahwa Random Forest mampu memprediksi pneumonia berdasarkan parameter non-invasif dengan tingkat akurasi yang tinggi serta mampu mengidentifikasi fitur-fitur klinis yang paling berpengaruh terhadap diagnosis pneumonia. Selain itu, Liang et al. (2022) [24] menemukan bahwa Random Forest menjadi salah satu algoritma terbaik dalam prediksi *ventilator-associated* pneumonia karena mampu mengolah data klinis yang kompleks dan heterogen. Hasil penelitian ini juga mendukung temuan Liu et al. (2022) [25] yang menunjukkan bahwa *pendekatan machine learning* berbasis Random Forest efektif digunakan untuk mendukung pengambilan keputusan klinis pada kasus pneumonia anak.

Meskipun menunjukkan performa yang baik, penelitian ini masih memiliki beberapa keterbatasan. Dataset yang digunakan hanya berasal dari satu rumah sakit sehingga karakteristik data mungkin belum sepenuhnya mewakili populasi yang lebih luas. Selain itu, jumlah data yang relatif terbatas dapat memengaruhi kemampuan generalisasi model ketika diterapkan pada lingkungan klinis yang berbeda. Namun demikian, model yang dihasilkan memiliki implikasi praktis yang penting karena dapat digunakan sebagai alat bantu skrining awal pneumonia pada fasilitas kesehatan yang memiliki keterbatasan akses terhadap pemeriksaan radiologi atau laboratorium. Penelitian selanjutnya disarankan menggunakan dataset *multisenter* dengan jumlah data yang lebih besar serta membandingkan Random Forest dengan metode yang lebih mutakhir seperti XGBoost, LightGBM, atau CatBoost untuk memperoleh performa prediksi yang lebih optimal.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, algoritma Random Forest terbukti menjadi metode terbaik untuk mendeteksi dini pneumonia berdasarkan parameter gejala klinis pasien. Pengujian menggunakan metode *10-Fold Cross-validation* menunjukkan bahwa model mampu memperoleh *accuracy* sebesar 89.82%, *recall* positif 95.24%, *precision* 90.53%, dan AUC 0.9469. Nilai tersebut menunjukkan bahwa Random Forest memiliki kemampuan yang sangat baik dalam mengenali pasien yang terindikasi pneumonia dengan tingkat keakuratan yang tinggi. Tingginya nilai sensitivitas juga menjadi poin penting karena dapat mengurangi risiko pasien pneumonia tidak terdeteksi (*false negative*).

Meskipun demikian, hasil evaluasi menunjukkan bahwa *recall* pada kelas negatif masih berada pada angka 77.67%, sehingga model masih berpotensi menghasilkan *false positive*, yaitu kondisi ketika pasien sehat diprediksi positif pneumonia. Namun, secara keseluruhan, performa Random Forest tetap lebih unggul dibandingkan algoritma Decision Tree dan Gradient Boosting, baik dari sisi akurasi maupun kemampuan model dalam membedakan kelas berdasarkan nilai AUC.

Dengan demikian, hipotesis penelitian ini dapat dinyatakan terbukti, yaitu algoritma Random Forest berbasis gejala klinis mampu digunakan secara efektif sebagai alat bantu deteksi dini pneumonia. Penerapan model ini diharapkan dapat membantu tenaga medis dalam mengambil keputusan secara lebih cepat dan tepat, terutama di daerah yang memiliki keterbatasan fasilitas diagnostik seperti laboratorium dan radiografi dada, sehingga pasien dengan risiko tinggi dapat segera diprioritaskan untuk mendapatkan penanganan lebih lanjut.

REFERENSI

[1] WHO, “Pneumonia in children,” Nov. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] Kemenkes, “Pedoman Nasional Pelayanan Kedokteran Tata Laksana Pneumonia Pada Dewasa,” *Kementeri. Kesehat. RI*, pp. 1–85, 2023, [Online]. Available: jdih.kemkes.go.id

[3] World Health Organization, *Guideline on Management of Pneumonia and Diarrhoea in Children up to 10 Years of Age*. 2024. [Online]. Available: https://iris.who.int/handle/10665/381761

[4] F. Faisal, Irwandi, R. Aprilia, Suharni, and Efriza, “Tinjauan Literatur: Faktor Risiko dan Epidemiologi Pneumonia pada Balita,” *Sci. J.*, vol. 3, no. 3, pp. 166–173, 2024, doi: 10.56260/sciena.v3i3.144.

[5] S. Usabili and U. Indahyanti, “Pemodelan Deteksi Dini Diabetes Mellitus menggunakan Pendekatan Ensemble Learning,” *Manaj. Pelayanan Kesehat.*, vol. 1, no. 1, p. 12, Jan. 2024, doi: 10.47134/mpk.v1i1.2924.

[6] C. Y. Effah *et al.*, “Machine learning-assisted prediction of pneumonia based on non-invasive measures,” *Front. Public Heal.*, vol. 10, 2022, doi: 10.3389/fpubh.2022.938801.

[7] T. Z. Yehuala, B. M. Fente, S. M. Wubante, and N. M. Derseh, “Exploring machine learning algorithms to predict acute respiratory tract infection and identify its determinants among children under five in Sub-Saharan Africa,” *Front. Pediatr.*, vol. 12, no. November, 2024, doi: 10.3389/fped.2024.1388820.

[8] A. W. Demsash *et al.*, “Data-driven machine learning Data-driven machine learning algorithm model for pneumonia prediction and determinant factor stratification among children aged 6–23 months in Ethiopia (BMC Infectious Diseases, (2025), 25, 1, (647), 10.1186/s12879-025-10916-4),” *BMC Infect. Dis.*, vol. 25, no. 1, 2025, doi: 10.1186/s12879-025-11137-5.

[9] T. Frondelius *et al.*, “Early prediction of ventilator-associated pneumonia with machine learning models: A systematic review and meta-analysis of prediction model performance☆,” *Eur. J. Intern. Med.*, vol. 121, pp. 76–87, Mar. 2024, doi: 10.1016/j.ejim.2023.11.009.

[10] I. A. Firdaus, “Deteksi Infeksi Mycoplasma Pneumoniae Pneumonia Menggunakan Komparasi Algoritma Klasifikasi Machine Learning,” 2022. doi: 10.31328/jointecs.v7i1.3242.

[11] S. Mahmuda, “Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube,” *J. Jendela Mat.*, vol. 2, no. 01, pp. 21–31, 2024, doi: 10.57008/jjm.v2i01.633.

[12] M. L. Amodio, “Effects of seasonal variation on MAP response of fresh-cut artichokes,” in *Acta Horticulturae*, 2024, pp. 175–176. doi: 10.17660/ActaHortic.2024.1386.24.

[13] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, “The evolution of CRISP-DM for Data Science:

Methods, Processes and Frameworks," *SBC Rev. Comput. Sci.*, vol. 4, no. 1, pp. 28–43, 2024, doi: 10.5753/reviews.2024.3757.

[14] M. Bouam *et al.*, "Delayed diagnosis of pneumonia in the emergency department: factors associated and prognosis," *Front. Med.*, vol. 10, no. May, 2023, doi: 10.3389/fmed.2023.1042704.

[15] S. H. Widiastuti, Purwanto, A. Wibowo, and R. W. Wisnuwardani, "Impact of Filter-Based Feature Selection on Prediction Model Performance for Early Detection of Stunting," *2024 9th Int. Conf. Informatics Comput. ICIC 2024*, pp. 1–7, 2024, doi: 10.1109/ICIC64337.2024.10956483.

[16] D. Anggreani, Hamdani, and Nurmisba, "Grid Search Hyperparameter Analysis in Optimizing The Decision Tree Method for Diabetes Prediction," *Indones. J. Data Sci.*, vol. 5, no. 3, pp. 190–197, 2024, doi: 10.56705/ijodas.v5i3.190.

[17] A. Hamid, L. Harmaji, and D. Mahdiana, "Deteksi Dini Hipertensi Pada Perusahaan Jasa Keamanan Menggunakan Metode Decision Tree," *J. Media Inform. Budidarma*, vol. 8, no. 2, p. 900, 2024, doi: 10.30865/mib.v8i2.7513.

[18] A. Saleh *et al.*, "Early Detection of Dengue Hemorrhagic Fever Using Patient Medical Data with Ensemble Learning Methods," vol. 8, no. 3, pp. 635–645, 2025, doi: http://dx.doi.org/10.24014/ijaidm.v8i3.38088.

[19] A. C. Mawarni, R. Rusdah, L. L. Hin, and D. Anubhakti, "Deteksi Dini Gejala Awal Penyakit Diabetes Menggunakan Algoritma Random Forest," *IDEALIS Indones. J. Inf. Syst.*, vol. 6, no. 2, pp. 165–171, Jul. 2023, doi: 10.36080/idealis.v6i2.3018.

[20] Ary Prandika Siregar, Dwi Priyadi Purba, Jojor Putri Pasaribu, and Khairul Reza Bakara, "Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke," *J. Penelit. Rumpun Ilmu Tek.*, vol. 2, no. 4, pp. 155–164, 2023, doi: 10.55606/juprit.v2i4.3039.

[21] A. Wibowo, A. F. N. Masruriyah, and S. Rahmawati, "Refining Diabetes Diagnosis Models: The Impact of SMOTE on SVM, Logistic Regression, and Naïve Bayes for Imbalanced Datasets," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 197–207, 2025, doi: 10.35882/jeeemi.v7i1.596.

[22] N. Elsayed, S. A. Elaleem, and M. Marie, "Improving Prediction Accuracy using Random Forest Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, pp. 436–441, 2024, doi: 10.14569/IJACSA.2024.0150445.

[23] D. Org Nur-Uz-Zaman, R. Ara, and M. Hossain, "A review on pneumonia in children: Clear insights," *Mediterr. J. Pharm. Pharm. Sci.*, vol. 4, no. 4, p. 57, 2024, doi: https://doi.org/10.5281/zenodo.14279043.

[24] Y. Liang *et al.*, "Early prediction of ventilator-associated pneumonia in critical care patients: a machine learning model," *BMC Pulm. Med.*, vol. 22, no. 1, pp. 1–9, 2022, doi: 10.1186/s12890-022-02031-w.

[25] Y. C. Liu *et al.*, "Evaluation of the Need for Intensive Care in Children with Pneumonia: Machine Learning Approach," *JMIR Med. Informatics*, vol. 10, no. 1, pp. 1–11, 2022, doi: 10.2196/28934.