VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8502

PERBANDINGAN KINERJA ALGORITMA MACHINE LEARNING UNTUK KLASIFIKASI ISPA MENGGUNAKAN DATA KLINIS RUMAH SAKIT

Irvan Lewenusa[1*]; Apriyanto Chandra[2]; Tri Sutrisno[3]

Program Studi Teknik Informatika[1,2]
Universitas Tarumanagara, Jakarta, Indonesia[1,2]
https://untar.ac.id[1,2]
irvanl@fti.untar.ac.id[1*], apriyanto.535210032@stu.untar.ac.id[2]

Program Studi Sistem Informasi[3]
Universitas Tarumanagara, Jakarta, Indonesia[3]
https://untar.ac.id[3]
tris@fti.untar.ac.id[3]

(*) Corresponding Author

Abstract— *Acute Respiratory Infection (ARI) remains one of the leading causes of morbidity and mortality worldwide, particularly among children and elderly populations. The complexity of ARI clinical symptoms necessitates rapid and accurate diagnostic approaches to support healthcare services. This study compares the performance of five machine learning algorithms, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Random Forest, and Gradient Boosting Machine algorithms for ARI classification using clinical hospital data. The study employed a quantitative experimental approach, using 521 outpatient clinical records obtained from XYZ Hospital, Jakarta. The research process included data preprocessing, classification model development, model performance evaluation, and statistical analysis using the Kruskal-Wallis test followed by Dunn's Post Hoc Test with Bonferroni correction. Model performance was assessed using accuracy, precision, recall, and F1-score metrics, which were computed using macro averaging due to the imbalanced class distribution. Statistically significant differences were observed among the algorithms across all evaluation metrics ($p < 0,001$). Effect size analysis using epsilon squared (ε^2) indicated large effects for accuracy ($\varepsilon^2 = 0.828$), precision ($\varepsilon^2 = 0.719$), recall ($\varepsilon^2 = 0.434$), and F1-score ($\varepsilon^2 = 0.654$). The post hoc analysis indicated that Random Forest and Gradient Boosting Machine showed comparable performance and consistently achieved competitive results across evaluation metrics. These findings suggest that ensemble learning methods are better suited to handling the complex clinical data associated with ARI and could help develop decision support systems for early ARI screening. Future studies should incorporate multicenter datasets, hyperparameter optimization, and explainable artificial intelligence techniques to improve model generalizability and interpretability.*

Keywords: *Acute Respiratory Infection, Gradient Boosting Machine, Kruskal-Wallis, Machine Learning, Random Forest.*

Abstrak— Infeksi Saluran Pernapasan Akut (ISPA) masih menjadi salah satu penyebab utama morbiditas dan mortalitas di dunia, khususnya pada kelompok anak-anak dan lansia. Kompleksitas gejala klinis ISPA memerlukan pendekatan diagnosis yang cepat dan akurat untuk mendukung pelayanan kesehatan. Penelitian ini bertujuan untuk membandingkan kinerja lima algoritma *machine learning*, yaitu *Logistic Regression, Naïve Bayes, K-Nearest Neighbor, Random Forest,* dan *Gradient Boosting Machine* dalam klasifikasi ISPA menggunakan data klinis rumah sakit. Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan dataset sebanyak 521 pasien rawat jalan dari Rumah Sakit XYZ, Jakarta. Tahapan penelitian meliputi *data preprocessing*, pembangunan model klasifikasi, evaluasi performa model, serta analisis statistik menggunakan uji *Kruskal-Wallis* yang dilanjutkan dengan *Dunn Post Hoc Test* dan *Bonferroni Correction*. Evaluasi model dilakukan menggunakan *accuracy*, *precision*, *recall*, dan *F1-score* yang dihitung menggunakan pendekatan *macro average* untuk memberikan evaluasi yang lebih seimbang pada

dataset dengan distribusi kelas yang tidak seimbang. Hasil penelitian menunjukkan adanya perbedaan yang signifikan antar-algoritma pada seluruh metrik evaluasi ($p < 0{,}001$). Analisis *effect size* menggunakan *epsilon squared* (ε^2) menunjukkan pengaruh yang besar pada *accuracy* ($\varepsilon^2 = 0.828$), *precision* ($\varepsilon^2 = 0.719$), *recall* ($\varepsilon^2 = 0.434$), dan *F1-score* ($\varepsilon^2 = 0.654$). Hasil uji lanjut menunjukkan bahwa Random Forest dan Gradient Boosting Machine memiliki performa yang tidak berbeda secara signifikan secara statistik dan secara umum menunjukkan kinerja yang lebih baik dibandingkan algoritma lainnya. Temuan ini menunjukkan bahwa metode *ensemble learning* lebih efektif dalam memodelkan hubungan kompleks antarvariabel klinis pada ISPA dan berpotensi mendukung pengembangan sistem pendukung keputusan untuk skrining awal ISPA. Penelitian selanjutnya disarankan untuk menggunakan *dataset multicenter*, optimasi hiperparameter, serta pendekatan *Explainable Artificial Intelligence* untuk meningkatkan kemampuan generalisasi dan interpretabilitas model.

Kata kunci: *Infeksi Saluran Pernapasan Akut, Gradient Boosting Machine, Kruskal-Wallis, Pembelajaran Mesin, Random Forest.*

PENDAHULUAN

Infeksi Saluran Pernapasan Akut (ISPA) masih menjadi salah satu penyebab utama morbiditas dan mortalitas di dunia, terutama pada kelompok rentan seperti anak-anak dan lansia. Kompleksitas gejala klinis ISPA menyebabkan proses diagnosis memerlukan ketelitian yang tinggi, sehingga keterlambatan diagnosis dapat memengaruhi penanganan pasien. Perkembangan *machine learning* memberikan peluang untuk mendukung proses diagnosis berbasis data klinis melalui identifikasi pola yang sulit dikenali secara manual. Berbagai penelitian melaporkan bahwa pendekatan ini mampu meningkatkan akurasi klasifikasi penyakit respirasi sekaligus mendukung pengembangan sistem pendukung keputusan klinis (*clinical decision support system*) (Guan et al., 2026, Zhou et al., 2024).

Berbagai algoritma *machine learning*, seperti *Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbor (KNN), Random Forest*, dan *Gradient Boosting Machine (GBM),* telah banyak diterapkan pada klasifikasi penyakit respirasi dengan tingkat performa yang bervariasi. Algoritma berbasis *ensemble learning,* khususnya *Random Forest* dan GBM, dilaporkan memiliki kemampuan yang baik dalam memodelkan hubungan nonlinier antarvariabel klinis, *sedangkan Logistic Regression, Naïve Bayes,* dan *KNN* masih banyak digunakan sebagai model pembanding (*baseline*) dalam penelitian bidang kesehatan (Hasan et al., 2024, Zhou et al., 2024; Bose & Bose, 2025; Warburton, 2025).

Meskipun penelitian mengenai klasifikasi penyakit menggunakan *machine learning* telah banyak dilakukan, masih terdapat beberapa *research gap*. Sebagian besar penelitian sebelumnya menggunakan dataset publik atau berukuran terbatas, sehingga kurang merepresentasikan kondisi klinis di rumah sakit di Indonesia. Selain itu, penelitian yang membandingkan beberapa paradigma *machine learning* menggunakan data klinis lokal masih relatif terbatas. Penelitian terdahulu juga umumnya hanya melaporkan metrik performa klasifikasi tanpa disertai pengujian signifikansi statistik maupun *effect size*, sehingga belum dapat memastikan apakah perbedaan performa antaralgoritma benar-benar bermakna.

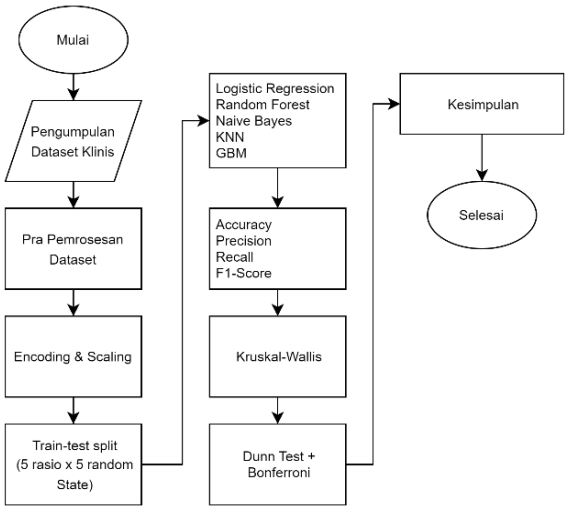
Berdasarkan *research gap* tersebut, penelitian ini bertujuan untuk membandingkan performa *Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbor, Random Forest*, dan *Gradient Boosting Machine* dalam mengklasifikasikan pasien ISPA menggunakan data klinis pasien rawat jalan dari Rumah Sakit XYZ. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, yang dilengkapi dengan uji *Kruskal–Wallis, Dunn Post Hoc Test* dengan koreksi *Bonferroni*, serta analisis *effect size* menggunakan *epsilon squared* (ε^2). Penelitian ini berkontribusi dengan menyajikan studi komparatif lima algoritma menggunakan data klinis nyata dari rumah sakit di Indonesia serta memberikan evaluasi statistik yang lebih komprehensif untuk mendukung pengembangan sistem pendukung keputusan pada skrining awal ISPA.

BAHAN DAN METODE

Penelitian ini menggunakan pendekatan kuantitatif eksperimental untuk membandingkan kinerja lima algoritma *machine learning*, yaitu *Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbor (KNN), Random Forest,* dan *Gradient Boosting Machine (GBM)*, dalam klasifikasi ISPA menggunakan data klinis pasien. Implementasi dilakukan menggunakan Python 3.12.4, sedangkan tahapan penelitian meliputi pengumpulan data, data *preprocessing*, pembangunan model, evaluasi performa, dan analisis statistik sebagaimana ditunjukkan pada Gambar 1.

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

INTI NUSA MANDIRI
DOI: https://doi.org/10.33480/inti.v21i1.8502



Sumber: (Hasil Penelitian, 2025)

Gambar 1. Alur Penelitian

Karakteristik Dataset

Dataset terdiri atas 521 data pasien rawat jalan yang diperoleh dari Rumah Sakit XYZ Jakarta pada periode Agustus 2024–Februari 2025. Dataset memuat atribut usia, jenis kelamin, suhu tubuh, *respiratory rate*, *heart rate*, SpO_2, batuk, pilek, mual, dan status ISPA sebagai variabel target. Seluruh data telah dianonimkan sesuai dengan prosedur etika penelitian. Distribusi kelas terdiri atas 499 pasien ISPA (95,78%) dan 22 pasien Non-ISPA (4,22%), sehingga termasuk kategori *imbalanced* dataset dapat dilihat pada Tabel 1. Statistik deskriptif disajikan pada Tabel 2 dan Tabel 3.

Tabel 1. Distribusi Kelas

	Jumlah	Persentase
ISPA	499	95,78%
Non-ISPA	22	4,22%
Total	521	100%

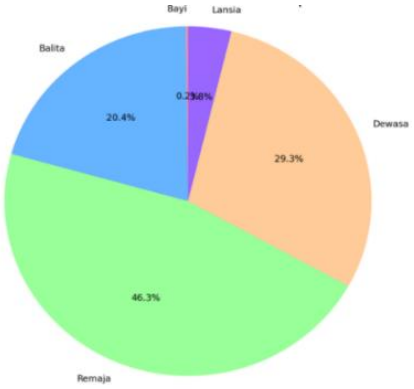
Sumber : (Hasil Penelitian, 2025)

Tabel 2. Karakteristik Dataset

Variabel	Tipe	Satuan
Usia	Numerik	Tahun
Jenis Kelamin	Kategorikal	
Suhu Tubuh	Numerik	^oC
Frekuensi Pernapasan	Numerik	Kali/menit
Denyut Jantung	Numerik	Bpm
SpO2	Numerik	%
Batuk	Kategorikal	
Pilek	Kategorikal	
Demam	Kategorikal	
Mual	Kategorikal	
ISPA	Target	

Sumber : (Hasil Penelitian, 2025)

Informasi ini penting untuk memahami karakteristik data dan menginterpretasikan hasil klasifikasi. Sebaran data dapat dilihat pada Gambar 2.



Sumber: (Hasil Penelitian, 2025)

Gambar 2. Sebaran Data Berdasarkan Kelompok Usia.

Tabel 3. Data Statistik Deskriptif

	Min	Max	Mean	std
Usia	0	86	18.8	18.8
Suhu tubuh	36	39.8	36.5	0.6
Denyut Jantung	60	152	104.3	19.6
Frekuensi Pernapasan	18	48	22.4	3.5
SpO₂	95	100	97.7	2.9

Sumber : (Hasil Penelitian, 2025)

Data Preprocessing

Data *preprocessing* meliputi pemeriksaan data duplikat, analisis distribusi *missing value*, mean *imputation*, label *encoding*, dan standardisasi fitur menggunakan *StandardScaler*. Tidak ditemukan data duplikat, sedangkan sekitar 8% nilai atribut mengalami *missing value*. Distribusi *missing value* berdasarkan kelas ditunjukkan pada Tabel 4 dan menunjukkan bahwa sebagian besar nilai yang hilang berada pada kelas ISPA. Seluruh nilai yang hilang kemudian ditangani menggunakan *mean imputation* untuk mempertahankan jumlah observasi. Ringkasan *preprocessing* disajikan pada Tabel 5.

Tabel 4. Distribusi *Missing Value* Berdasarkan Kelas.

Variabel	ISPA	Non-ISPA	Total
Sesak Napas	117	5	122
Suhu Tubuh	21	0	21
Denyut Jantung	57	1	58
Mual	40	5	45
SpO_2	87	0	87
Pilek	2	0	2
Frekuensi Pernapasan	119	5	124

Sumber : (Hasil Penelitian, 2025)

VOL. 21. NO. 1 AGUSTUS 2026
P-ISSN: 0216-6933 | E-ISSN: 2685-807X
Diterbitkan Oleh: LPPM Universitas Nusa Mandiri

Berdasarkan Tabel 4, jumlah *missing value* pada kelas Non-ISPA relatif lebih kecil dibandingkan kelas ISPA. Hal ini menunjukkan bahwa sebagian besar nilai yang hilang berasal dari kelas mayoritas sehingga penerapan *mean imputation* diperkirakan tidak mengubah karakteristik utama kelas minoritas secara signifikan. Pendekatan ini dipilih untuk mempertahankan seluruh observasi selama proses pelatihan model tanpa mengurangi jumlah sampel yang tersedia.

Tabel 5. Ringkasan Hasil *Data Preprocessing*.

Tahapan	Jumlah data
Sebelum Preprocessing	521
Setelah Preprocessing	521
Data dihapus	0

Sumber : (Hasil Penelitian, 2025)

Berdasarkan hasil *preprocessing* pada Tabel 5, jumlah data sebelum dan sesudah *preprocessing* tetap sebanyak 521 sampel karena tidak ditemukan data duplikat. Oleh karena itu, proses *preprocessing* difokuskan pada penanganan *missing value*, transformasi atribut kategorikal, dan normalisasi fitur numerik.

Pembagian Dataset

Dataset dibagi menggunakan lima rasio *train-test split* (50:50, 60:40, 70:30, 80:20, dan 90:10) dengan lima nilai *random state* sehingga diperoleh 25 skenario eksperimen untuk setiap algoritma. Proses pembagian menggunakan *stratified train-test split (stratify)* agar proporsi kelas ISPA dan Non-ISPA tetap terjaga pada data latih maupun data uji.

Pembangunan Model

Pembangunan model pada setiap algoritma menggunakan parameter yang ditampilkan pada tabel 6 sebagai berikut:

Tabel 6. Parameter Algoritma *Machine Learning*.

Algoritma	Parameter
Logistic Regression	Solver=liblinear, max_iter=1000, class_weight=balanced
Random Forest	n_estimators=100, max_depth=5, class_weight=balanced
KNN	K=3, metric=euclidean
Gaussian Naïve Bayes	Gaussian distribution
GBM	n_estimators=100 learning_rate=0.01 max_depth=3 random_state=42

Sumber : (Hasil Penelitian, 2025)

Penelitian ini tidak menerapkan *hyperparameter tuning* karena bertujuan membandingkan performa algoritma menggunakan konfigurasi *baseline* yang umum digunakan pada penelitian sebelumnya. Pada GBM digunakan *learning_rate* = 0.01 dan *n_estimators* = 100 untuk memperoleh proses pembelajaran yang stabil dengan kompleksitas komputasi tetap rendah. Optimasi hiperparameter menjadi bagian dari penelitian selanjutnya. *Logistic Regression* dan *Random Forest* menggunakan *class_weight='balanced'*, sedangkan KNN, *Gaussian Naïve Bayes*, dan GBM menggunakan konfigurasi standar dari *scikit-learn*.

Evaluasi Model

Performa setiap algoritma dievaluasi menggunakan *accuracy, precision, recall,* dan *F1-score.* Mengingat dataset memiliki distribusi kelas yang tidak seimbang, nilai *precision, recall*, dan *F1-score* dihitung menggunakan pendekatan *macro average* sehingga setiap kelas memperoleh bobot yang sama dalam proses evaluasi. (Stopa et al., 2024).

Analisis Statistik

Hipotesis H0: tidak terdapat perbedaan performa yang signifikan antar algoritma *machine learning*. H1: Terdapat setidaknya satu algoritma *machine learning* yang memiliki performa berbeda secara signifikan. Uji *Kruskal-Wallis* digunakan karena data performa model tidak diasumsikan berdistribusi normal. Pengujian dilakukan terhadap nilai *accuracy*, *precision*, *recall*, dan *F1-score* yang diperoleh dari 25 skenario eksperimen pada masing-masing algoritma. Penggunaan lima rasio *train-test split* dengan lima *random state* bertujuan untuk mengevaluasi konsistensi performa algoritma pada berbagai proporsi data latih dan data uji. Setiap skenario menggunakan *stratified train-test split* sehingga distribusi kelas tetap terjaga pada setiap iterasi. Pendekatan ini dipilih untuk mengevaluasi konsistensi performa algoritma pada berbagai konfigurasi pembagian data sehingga hasil evaluasi tidak bergantung pada satu konfigurasi *train-test split* tertentu. Meskipun demikian, penelitian ini mengakui bahwa penggunaan *Stratified K-Fold Cross Validation* dapat memberikan estimasi performa yang lebih stabil dan direkomendasikan untuk penelitian selanjutnya. Rumus statistika *Kruskal-Wallis* ditunjukkan pada persamaan (1) sebagai berikut:

$$k = (N-1)\frac{\sum_{i=1}^{g} n_i(r_i - r)^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - r)^2} \quad (1)$$

Jika hasil pengujian menunjukkan p-value < 0,05 maka dilakukan *Dunn Post Hoc Test* dengan *Bonferroni Correction* untuk mengidentifikasi

pasangan algoritma yang berbeda performa secara signifikan. Selanjutnya, analisis *effect size* menggunakan *epsilon squared* (ε^2) dilakukan untuk mengukur besarnya pengaruh perbedaan performa antaralgoritma. Seluruh proses analisis data dilakukan menggunakan *Python* 3.12.4 pada perangkat komputer dengan spesifikasi Apple M2 Chip, RAM 8 GB, dan sistem operasi macOS. Dalam penelitian ini, setiap skenario eksperimen diperlakukan sebagai satu unit observasi independen yang dihasilkan dari konfigurasi pembagian data yang berbeda. Oleh karena itu, analisis statistik difokuskan pada evaluasi konsistensi performa algoritma di berbagai kondisi eksperimen, bukan pada perbandingan antarrasio *train-test split.*

Analisis *Feature Importance*

Untuk memberikan interpretasi awal terhadap kontribusi masing-masing fitur dalam proses klasifikasi, penelitian ini melakukan analisis *feature importance* menggunakan algoritma Random Forest. Nilai *feature importance* dihitung berdasarkan rata-rata penurunan *impurity (Mean Decrease in Impurity)* yang dihasilkan selama proses pembentukan pohon keputusan. Semakin tinggi nilai *feature importance*, semakin besar kontribusi relatif suatu fitur terhadap proses klasifikasi pada model *Random Forest.*

HASIL DAN PEMBAHASAN

Hasil Evaluasi Model Klasifikasi

Penelitian ini melakukan evaluasi terhadap lima algoritma *machine learning*, yaitu *Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbors (KNN), dan Gradient Boosting Machine (GBM)* untuk klasifikasi pasien Infeksi Saluran Pernapasan Akut (ISPA) menggunakan data klinis pasien rawat jalan Rumah Sakit XYZ.

Tabel 7. Perbandingan Performa Terbaik Model Klasifikasi ISPA

Algoritma	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.81	0.56	0.73	0.56
Random Forest	0.96	0.83	0.72	0.75
Naïve Bayes	0.90	0.61	0.74	0.64
KNN	0.95	0.47	0.49	0.48
GBM	0.97	0.87	0.67	0.72

*Nilai *precision, recall, dan F1-score* dihitung menggunakan pendekatan *macro average.*

Sumber: (Hasil Penelitian, 2025).

Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Untuk memastikan bahwa perbedaan performa antarmodel tidak hanya bersifat numerik tetapi juga signifikan secara statistik, penelitian ini menerapkan pengujian *Kruskal-Wallis* yang dilanjutkan dengan uji *post-hoc Dunn* menggunakan koreksi *Bonferroni* serta analisis *effect size* menggunakan *epsilon squared* (ε^2). Ringkasan performa terbaik untuk masing-masing algoritma ditunjukkan pada Tabel 7. Mengingat distribusi kelas pada dataset sangat tidak seimbang (95,78% ISPA dan 4,22% Non-ISPA), dilakukan evaluasi tambahan terhadap performa masing-masing algoritma pada kelas minoritas (Non-ISPA). Hasil evaluasi tersebut disajikan pada Tabel 8.

Tabel 8. Perbandingan Performa Kelas Minoritas Non-ISPA

Algoritma	Precision	Recall	F1-Score
Logistic Regression	0.14	0.64	0.23
Random Forest	0.70	0.46	0.53
Naïve Bayes	0.25	0.56	0.35
KNN	0.00	0.00	0.00
GBM	0.77	0.36	0.47

*Nilai *precision, recall, dan F1-score* dihitung menggunakan pendekatan *macro average.*

Sumber: (Hasil Penelitian, 2025).

Berdasarkan hasil evaluasi pada kelas minoritas (Non-ISPA), terdapat perbedaan kemampuan antaralgoritma dalam mendeteksi sampel kelas minoritas. *Random Forest* memperoleh nilai *F1-score* tertinggi (0,53), yang menunjukkan keseimbangan terbaik antara *precision* dan *recall* dalam mengidentifikasi kasus Non-ISPA. Sementara itu, *Gradient Boosting Machine* (GBM) menghasilkan nilai *precision* tertinggi (0,77), yang mengindikasikan bahwa sebagian besar prediksi Non-ISPA yang dihasilkan model merupakan prediksi yang benar. Namun demikian, nilai *recall* GBM yang relatif lebih rendah (0,36) menunjukkan bahwa masih terdapat sejumlah kasus Non-ISPA yang tidak terdeteksi.

Logistic Regression memperoleh nilai *recall* yang cukup tinggi (0,64), tetapi memiliki *precision* yang rendah (0,14), sehingga menghasilkan cukup banyak prediksi positif yang tidak sesuai *(false positive*). Di sisi lain, algoritma *K-Nearest Neighbor* (KNN) tidak mampu mengidentifikasi kelas Non-ISPA pada skenario pengujian, yang ditunjukkan oleh nilai *precision, recall,* dan *F1-score* sebesar 0. Hasil ini menunjukkan bahwa nilai *accuracy* yang tinggi tidak selalu mencerminkan kemampuan model dalam mengenali kelas minoritas pada dataset dengan distribusi kelas yang tidak

seimbang. Oleh karena itu, evaluasi menggunakan metrik per kelas, seperti *precision, recall,* dan *F1-score* untuk kelas minoritas, diperlukan agar performa model dapat dinilai secara lebih komprehensif.

Analisis Performa Model

Berdasarkan hasil rata-rata pengujian pada berbagai skenario pembagian data dan *random state*, algoritma berbasis *ensemble*, yaitu *Gradient Boosting Machine* dan *Random Forest*, menunjukkan performa yang lebih stabil dibandingkan model lainnya. Secara umum, model *ensemble*, yaitu *Random Forest* dan *Gradient Boosting Machine*, menunjukkan performa yang kompetitif pada sebagian besar metrik evaluasi. Hasil pengujian statistik menunjukkan bahwa performa GBM tidak selalu berbeda secara signifikan dibandingkan dengan *Random Forest* dan *Naïve Bayes* pada seluruh metrik evaluasi. Oleh karena itu, interpretasi keunggulan model dilakukan berdasarkan kombinasi hasil evaluasi numerik dan hasil pengujian statistik. Tingginya *accuracy* pada KNN kemungkinan dipengaruhi oleh dominasi kelas mayoritas sehingga tidak sepenuhnya merefleksikan kemampuan model dalam mendeteksi seluruh kasus secara seimbang.

Analisis *Feature Importance*

Berdasarkan analisis *feature importance* menggunakan algoritma *Random Forest*, variabel usia (*age*) memiliki nilai kepentingan tertinggi (0,189), diikuti oleh suhu tubuh (0,160), *heart rate* (0,151), batuk (0,137), dan pilek (0,119). Hasil ini menunjukkan bahwa model lebih banyak memanfaatkan kelima variabel tersebut dalam membedakan pasien ISPA dan Non-ISPA. Tabel 9 menampilkan *feature importance Random Forest*. Nilai tersebut menunjukkan kontribusi relatif terhadap keputusan model *Random Forest* dan tidak menunjukkan hubungan kausal antarvariabel.

Tabel 9. *Feature Importance Random Forest*

Peringkat	Fitur	*Importance*
1	Usia	0.189
2	Suhu Tubuh	0.160
3	Denyut Jantung *(Heart Rate)*	0.151
4	Batuk	0.137
5	Pilek	0.119
6	SpO2	0.091
7	Demam	0.048
8	Mual	0.036
9	Frekuensi Pernapasan (*Respiratory Rate)*	0.029
10	Sesak Napas	0.022
11	Jenis Kelamin	0.017

Sumber : (Hasil Penelitian, 2025)

Di sisi lain, SpO_2 memperoleh nilai f*eature importance* sebesar 0,091, sedangkan *respiratory rate* (RR) sebesar 0,029. Hal ini menunjukkan bahwa kedua fitur tersebut tetap memberikan kontribusi terhadap proses klasifikasi, namun kontribusinya relatif lebih rendah dibandingkan dengan beberapa variabel klinis lainnya pada dataset penelitian ini. Perbedaan tingkat kepentingan antarfitur dipengaruhi oleh karakteristik dataset dan mekanisme pembentukan pohon keputusan pada *Random Forest*. Oleh karena itu, nilai *feature importance* merepresentasikan kontribusi relatif fitur terhadap model yang dibangun dan tidak dapat diinterpretasikan sebagai hubungan sebab-akibat.

Analisis *Confusion Matrix*

Selain evaluasi numerik, penelitian ini juga menganalisis *confusion matrix* untuk mengevaluasi pola kesalahan klasifikasi pada setiap algoritma. Hasil *confusion matrix* menunjukkan bahwa *Random Forest* dan GBM menghasilkan jumlah *false positive* yang sama rendahnya dibandingkan dengan algoritma lainnya. Dalam konteks klinis, tingginya *false positive* dapat meningkatkan risiko pemeriksaan medis tambahan yang tidak diperlukan. Oleh karena itu, keseimbangan antara *precision* dan *recall* menjadi faktor penting dalam evaluasi model klasifikasi medis. Tabel 10 menyajikan *confusion matrix* dari masing-masing algoritma pada skenario dengan performa terbaik, serta melengkapi metrik evaluasi seperti *accuracy, precision, recall*, dan *F1-score*.

Tabel 10. *Confusion Matrix Pada Skenario Pengujian Terbaik.*

Model	TN	FP	FN	TP
Logistic Regression	1	1	7	44
Random Forest	1	1	1	50
Naïve Bayes	1	1	4	47
KNN	0	2	1	50
GBM	1	1	1	50

Sumber : (Hasil Penelitian, 2025)

Berdasarkan Tabel 10, *Random Forest* dan GBM menghasilkan *false positive* yang sama rendahnya, yaitu 1 sampel, serta *false negative* yang juga rendah, yaitu 1 sampel. Sebaliknya, *Logistic Regression* menghasilkan jumlah *false negative* yang lebih tinggi dibandingkan algoritma lainnya, sedangkan KNN menghasilkan jumlah *false positive* tertinggi. Hasil ini menunjukkan bahwa *Random Forest* dan GBM memiliki kemampuan yang lebih baik dalam menyeimbangkan kesalahan klasifikasi dibandingkan dengan algoritma pembanding pada skenario pengujian terbaik.

Uji *Kruskal-Wallis* dan *Effect Size*

Untuk mengevaluasi signifikansi perbedaan performa antaralgoritma, penelitian ini menerapkan pengujian statistik menggunakan metode *Kruskal-Wallis*. Tabel 11 menunjukkan hasil pengujian Kruskal-Wallis pada setiap evaluasi performa.

Tabel 11. Hasil Uji Statistik Kruskal-Wallis.

Metric	H Value	p-value	ε^2
Accuracy	103.315	<0,001	0.828
Precision	90.258	<0,001	0.719
Recall	56.109	<0,001	0.434
F1-score	82.518	<0,001	0.654

Sumber : (Hasil Penelitian, 2025)

Hasil uji Kruskal-Wallis menunjukkan bahwa terdapat perbedaan performa yang signifikan antaralgoritma klasifikasi pada seluruh metrik evaluasi. Nilai H sebesar 103.315 ($p < 0,001$) untuk *accuracy*, 90.258 ($p < 0,001$) untuk *precision*, 56.109 ($p < 0,001$) untuk *recall*, dan 82.518 ($p < 0,001$) untuk *F1-score* menunjukkan bahwa minimal terdapat satu algoritma yang memiliki performa berbeda secara signifikan dibandingkan algoritma lainnya. Analisis *effect size* menggunakan *epsilon squared* (ε^2) menunjukkan nilai 0.828 untuk *accuracy*, 0.719 untuk *precision*, 0.434 untuk *recall*, dan 0.654 untuk *F1-score*. Seluruh nilai tersebut termasuk dalam kategori large *effect size*, yang menunjukkan bahwa perbedaan performa antar-algoritma tidak hanya signifikan secara statistik, tetapi juga memiliki makna praktis yang kuat.

Uji *Post Hoc Dunn* Koreksi *Bonferroni*

Karena hasil uji *Kruskal-Wallis* menunjukkan perbedaan yang signifikan pada seluruh metrik evaluasi ($p < 0,05$), maka dilakukan analisis lanjutan menggunakan uji *post hoc Dunn* dengan koreksi *Bonferroni* untuk mengidentifikasi pasangan algoritma yang memiliki perbedaan performa secara signifikan. Hasil uji *Dunn* pada metrik *F1-score* yang disajikan pada Tabel 8 menunjukkan adanya perbedaan signifikan antara GBM dan KNN ($p < 0,001$), GBM dan *Logistic Regression* ($p < 0,001$), serta *Random Forest* dan KNN ($p < 0,001$). Namun demikian, tidak ditemukan perbedaan signifikan antara GBM dan *Random Forest* ($p = 1,000$) maupun antara GBM dan *Naïve Bayes* ($p = 0,593$). Selain itu, tidak ditemukan perbedaan signifikan antara *Random Forest* dan *Naïve Bayes* ($p=0,100$) maupun antara *Naïve Bayes* dan Logistic Regression ($p=0,160$). Hasil ini menunjukkan bahwa meskipun GBM menghasilkan performa rata-rata yang tinggi, keunggulannya tidak selalu berbeda secara signifikan dibandingkan *Random Forest* dan *Naïve Bayes* pada metrik *F1-score.*

Temuan tersebut mengindikasikan bahwa model *ensemble*, khususnya *Gradient Boosting Machine* dan *Random Forest*, memiliki kemampuan klasifikasi yang relatif sebanding dalam mendeteksi kasus ISPA berdasarkan data klinis pasien. Oleh karena itu, pemilihan model tidak hanya perlu mempertimbangkan nilai performa rata-rata, tetapi juga hasil pengujian statistik yang menunjukkan signifikansi perbedaan antarmodel. Tabel 12 menunjukkan perbandingan antara algoritma.

Tabel 12. Hasil Uji Post Hoc Dunn.

Perbandingan	Adjusted p-value	Keterangan
GBM vs KNN	<0,001	Signifikan
GBM vs LR	<0,001	Signifikan
GBM vs RF	1.000	Tidak Signifikan
GBM vs NB	0.593	Tidak Signifikan
RF vs KNN	<0,001	Signifikan
RF vs LR	<0,001	Signifikan
RF vs NB	0.100	Tidak Signifikan
NB vs KNN	<0,001	Signifikan
NB vs LR	0.160	Tidak Signifikan
LR vs KNN	0.049	Signifikan

Sumber : (Hasil Penelitian, 2025)

Implikasi Klinis dan Keterbatasan Penelitian

Hasil penelitian menunjukkan bahwa algoritma berbasis ensemble learning, khususnya *Random Forest* dan *Gradient Boosting Machine*, berpotensi mendukung proses skrining awal ISPA berdasarkan data klinis pasien. Pemanfaatan model klasifikasi berbasis *machine learning* dapat membantu tenaga kesehatan dalam memberikan penilaian awal secara lebih cepat dan objektif sehingga berpotensi mendukung pengembangan *clinical decision support system* pada pelayanan kesehatan.

Penelitian ini memiliki beberapa keterbatasan. Dataset yang digunakan hanya berasal dari satu rumah sakit dan memiliki distribusi kelas yang tidak seimbang, sehingga dapat membatasi kemampuan model untuk menggeneralisasi. Meskipun pembagian data telah dilakukan menggunakan *stratified repeated train-test split* untuk menjaga proporsi setiap kelas pada seluruh skenario pengujian, penelitian selanjutnya disarankan untuk menggunakan dataset *multicenter*, menerapkan teknik penyeimbangan data seperti SMOTE atau ADASYN, serta menggunakan *Stratified K-Fold Cross Validation* untuk memperoleh estimasi performa yang lebih stabil dan representatif. Selain itu, penerapan metode *Explainable Artificial Intelligence (XAI),* seperti *SHAP* atau *LIME*, dapat meningkatkan interpretabilitas model sehingga lebih mudah diimplementasikan dalam pengambilan keputusan klinis.

KESIMPULAN

Penelitian ini berhasil melakukan analisis komparatif terhadap lima algoritma *machine learning*, yaitu *Logistic Regression* (LR), *Random Forest* (RF), *Naïve Bayes* (NB), *K-Nearest Neighbor* (KNN), dan *Gradient Boosting Machine* (GBM) untuk klasifikasi pasien Infeksi Saluran Pernapasan Akut (ISPA) menggunakan data klinis pasien rawat jalan Rumah Sakit XYZ. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* pada berbagai skenario pembagian data dan *random state* untuk memperoleh hasil yang lebih robust. Hasil penelitian menunjukkan bahwa terdapat perbedaan performa yang signifikan antara algoritma klasifikasi pada seluruh metrik evaluasi. Uji *Kruskal-Wallis* menghasilkan nilai H sebesar 103,315 untuk *accuracy*, 90,258 untuk precision, 56,109 untuk *recall*, dan 82,518 untuk *F1-score* dengan nilai $p < 0,001$ pada seluruh metrik. Analisis *effect size* menggunakan *epsilon squared* (ε^2) menunjukkan kategori *large effect size* pada seluruh metrik evaluasi, yang mengindikasikan bahwa perbedaan performa antar-algoritma tidak hanya signifikan secara statistik tetapi juga memiliki makna praktis yang kuat.

Berdasarkan hasil evaluasi, model berbasis *ensemble*, khususnya *Random Forest dan Gradient Boosting Machine,* menunjukkan performa yang kompetitif dan tidak berbeda secara signifikan secara statistik pada sebagian besar metrik evaluasi berdasarkan uji *Dunn* dengan koreksi *Bonferroni* dalam klasifikasi pasien ISPA. Temuan penelitian ini menunjukkan bahwa model berbasis *ensemble learning*, khususnya *Random Forest* dan *Gradient Boosting Machine*, berpotensi mendukung *pengembangan clinical decision support system* untuk skrining awal ISPA. Selain itu, penelitian ini menegaskan pentingnya penggunaan metrik *macro average* serta evaluasi performa pada kelas minoritas dalam menilai model klasifikasi pada dataset dengan distribusi kelas yang tidak seimbang. Pendekatan tersebut memberikan gambaran performa model yang lebih representatif dibandingkan dengan hanya mengandalkan nilai *accuracy*. Penggunaan data klinis nyata dari rumah sakit juga berkontribusi terhadap pengembangan penelitian *healthcare informatics* berbasis populasi lokal yang masih relatif terbatas di Indonesia.

Penelitian ini memiliki beberapa keterbatasan. Dataset yang digunakan berasal dari satu rumah sakit dan memiliki distribusi kelas yang tidak seimbang, sehingga dapat membatasi generalisasi hasil penelitian. Meskipun pembagian data telah dilakukan *menggunakan stratified train-test split* untuk menjaga proporsi setiap kelas, penelitian selanjutnya disarankan untuk menerapkan teknik penyeimbangan data, seperti SMOTE atau ADASYN, serta menggunakan *Stratified K-Fold Cross Validation* untuk memperoleh estimasi performa model yang lebih andal. Selain itu, penerapan metode *Explainable Artificial Intelligence (XAI),* seperti *SHAP* atau *LIME*, dapat dipertimbangkan untuk meningkatkan interpretabilitas model guna mendukung pengambilan keputusan klinis.

REFERENSI

Bose, S., & Bose, S. (2025). *Random Forests: The Wisdom of Crowds in Action*. https://doi.org/10.65525/jetcsa.v1i1.5.

Cortez, L. F., Luis, B. A., Christian, S. C., Henning, V. M., & Livia, K. (2024). *Additional file 2 of Comparative microbiome analysis in cystic fibrosis and non-cystic fibrosis bronchiectasis.* https://doi.org/10.6084/m9.figshare.25854671.v1.

Frasca, M., La Torre, D., Pravettoni, G., & Cutica, I. (2024). Explainable and interpretable artificial intelligence in medicine: A systematic bibliometric review. *Discover Artificial Intelligence*, *4*(1), 15. https://doi.org/10.1007/s44163-024-00114-7

Guan, C., Chen, F., Song, Y., Huang, Y., Zhou, Y., Wang, Z., & Cheng, J. (2026). A machine learning-based model for assessing community-acquired pneumonia severity using routine blood tests. Frontiers in Cellular and Infection Microbiology, 15, 1605502. https://doi.org/10.3389/fcimb.2025.1605502.

Hasan, M. N., Prodhan, M. E., Alam, A., Sultana, S., Rahman, M., & Parvin, H. (2024). A comparative study of machine learning algorithms in predicting acute respiratory infection of under-five children in Bangladesh using imbalanced data. In *2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)* (pp. 274–278). IEEE. https://doi.org/10.1109/BECITHCON64160.2024.10962602.

Kassaw, A. K., Bekele, G., Kassaw, A. K., & Yimer, A. (2024). Prediction of acute respiratory infections using machine learning techniques in Amhara Region, Ethiopia. Scientific Reports, 14(1), 27968. https://doi.org/10.1038/s41598-024-76847-3.

Khairunissa, N., Gunawan, P. H., Rohmawati, A. A., & Fikriansyah, M. (2024). Waiting Period Prediction of Telkom University Student Alumni Using K-Nearest Neighbor and Naïve Bayes. 22–27. https://doi.org/10.1109/icodsa62899.2024.10652228.

Nasution, I. S., Anggraini, F. A., Hsb, M. F. R., Tyas, D. A., Pinasti, N. E., Andriyani, R., & Nasution, E. Y. (2025). *Penyakit Saluran Nafas Atas: Epidemiologi, Patogenesis, Pengobatan, dan Pencegahan*. *2*(1), 411–420. https://doi.org/10.57235/helium.v2i1.5163.

Nita, Y., Rokayah, R., & Alfian, R. (2026). Clinical Decision Support Systems to Improve Antibiotic Prescribing in Community and Clinical Pharmacy: A Systematic Review. Indonesian Journal of Pharmacy. https://doi.org/10.22146/ijp.25049.

Shah, R., Pawar, A., & Kumar, M. (2024). *Enhancing Machine Learning Model Using Explainable AI* (pp. 287–297). Springer International Publishing. https://doi.org/10.1007/978-981-99-6906-7_25.

Shankhdhar, D., & Agarwal, D. V. (2025). A Review of Machine Learning Techniques for Disease Classification in Healthcare Data. 8(3). https://doi.org/10.5281/zenodo.15699589.

Stopa, Ł., Stopa, W., & Stopa, Z. (2024). Correlation between Tomography Scan Findings and Clinical Presentation and Treatment Outcomes in Patients with Orbital Floor Fractures. Diagnostics, 14(3), 245. https://doi.org/10.3390/diagnostics14030245.

Warburton, D. M. (2025). *Predictive analytics for healthcare insurance risk assessment using ensemble learning models*. https://doi.org/10.5281/zenodo.14598758

Zhao, X. Y., & Nie, X. (2022). Status Forecasting Based on the Baseline Information Using Logistic Regresssion. *Entropy*, *24*(10), 1481. https://doi.org/10.3390/e24101481.

Zhou, L.-x., Zhou, Q., Gao, T.-m., Xiang, X.-x., Zhou, Y., Jin, S.-j., Qian, J.-j., Zhou, B.-h., Bai, D.-s., & Jiang, G.-q. (2024). Machine learning predicts acute respiratory failure in pancreatitis patients: A retrospective study. International Journal of Medical Informatics, 192, 105629. https://doi.org/10.1016/j.ijmedinf.2024.105629

