JTS : Jurnal Teknik
P-ISSN: 2302-8734
E-ISSN: 2581-0006

FT UMT

Vol. 15 No. 2 Th. 2026
Halaman : 23 - 33
Agustus 2026

Ensemble Learning untuk Prediksi Risiko Penyakit Paru-Paru

Fathimah Nabilah[1], Ilamsyah[2], Muhammad Arba Adnandi[3], Putri Dhamma Yanthi[4], Rizki Sulaiman[5]

[1,2,3,4,5]*Program Studi Ilmu Komputer Jurusan Fakultas Teknologi dan Bisnis Universitas yatsi Madani*
[1,2,3,4,5]Jl. Aria Santika - Karawaci Kota Tangerang
E-mail: [1]fathimah@uym.ac.id, [2]**ilamsyah@raharja.info**,
[3]arba@uym.ac.id, [4]putridhamma@uym.ac.id, [5]rizki@uym.ac.id

Receive: 26-05-2026 Accepted: 27-07-2026

Abstract

Smoking is a major risk factor for lung diseases such as COPD and lung cancer. Using the Ensemble algorithm, this study aims to predict respiratory organ damage caused by smoking. The Ensemble method was chosen for its ability to improve prediction accuracy by combining models such as XGBoost, Random Forest, and Logistic Regression through the Voting Classifier. Results showed that the Voting Classifier model achieved 87% accuracy, higher than individual models such as XGBoost (81%), Logistic Regression (77%), and Random Forest (74%). Important factors in the prediction included smoking duration and the number of cigarettes consumed daily. Although the Medium class has a higher error rate, the ensemble model remains a reliable method for early detection and screening, but it is effective in classifying the extent of lung damage, especially in the High class and medical data-based screening.

Keywords *: Lungs, Ensemble, XGBoost, Random Forest, Logistic Regression.*

Abstrak

Merokok merupakan faktor risiko utama penyakit paru-paru seperti PPOK dan kanker paru-paru. Dengan menggunakan algoritma Ensemble, penelitian ini bertujuan untuk memprediksi kerusakan organ pernapasan yang disebabkan oleh merokok. Metode Ensemble dipilih karena kemampuannya untuk meningkatkan akurasi prediksi dengan menggabungkan model seperti XGBoost, Random Forest, dan Regresi Logistik melalui Voting Classifier. Hasil menunjukkan bahwa model Voting Classifier mencapai akurasi 87%, lebih tinggi daripada model individual seperti XGBoost (81%), Regresi Logistik (77%), dan Random Forest (74%). Faktor penting dalam prediksi meliputi durasi merokok dan jumlah rokok yang dikonsumsi setiap hari. Meskipun kelas Sedang memiliki tingkat kesalahan yang lebih tinggi, model ensemble tetap merupakan metode yang andal untuk deteksi dini dan skrining, tetapi efektif dalam mengklasifikasikan tingkat kerusakan paru-paru, terutama pada kelas Tinggi dan skrining berbasis data medis.

Kata kunci : Paru-paru, Ensemble, XGBoost, Random Forest, Regresi Logistik

PENDAHULUAN

Merokok merupakan kebiasaan berisiko tinggi yang sangat merugikan kesehatan, terutama bagi sistem pernapasan. Dalam asap rokok terdapat lebih dari 7.000 senyawa kimia, di antaranya sekitar 69 zat diketahui bersifat karsinogenik—artinya dapat memicu kanker (Djajalaksana, 2025). Senyawa beracun seperti nikotin, karbon monoksida, tar, formaldehida, benzena, dan arsenik langsung masuk ke dalam tubuh melalui saluran pernapasan ketika seseorang menghirup asap rokok.

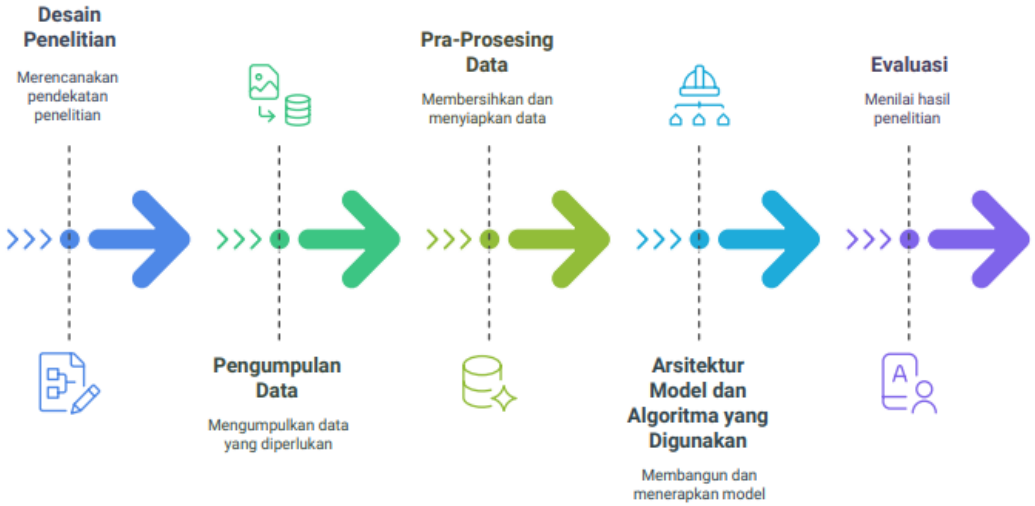
Setiap tahun, penggunaan tembakau menyebabkan lebih dari 8 juta kematian di seluruh dunia, dengan 1,3 juta di antaranya terjadi pada perokok pasif—individu yang tidak merokok namun terpapar asap rokok dari orang lain (WHO, 2023). Hal ini menunjukkan bahwa risiko kesehatan akibat merokok tidak hanya mengancam perokok aktif, tetapi juga membahayakan orang-orang di sekitarnya.

Di Indonesia, jumlah perokok masih tinggi dan cenderung naik, terutama di kalangan remaja dan usia produktif. Prevalensi perokok remaja usia 13–15 tahun naik dari 18,3% menjadi 19,2%. Tren ini menunjukkan makin banyak remaja mulai merokok sejak dini — kondisi yang memicu ketergantungan lebih cepat dan membuat mereka lebih sulit berhenti di kemudian hari (Prabawati & Nurhidayah, 2024).

Salah satu algoritma yang sering digunakan dan efektif dalam pengolahan data medis adalah Ensemble, yaitu menggabungkan prediksi dari beberapa model machine learning untuk meningkatkan akurasi dan stabilitas prediksi (Wibisono dkk, 2024). Kelebihannya terletak pada kemampuan mengurangi kelemahan model tunggal dengan memadukan kekuatan berbagai model sekaligus. Dalam bidang medis, hal ini sangat berguna karena membantu dokter dan peneliti mengidentifikasi faktor-faktor utama yang memengaruhi risiko penyakit, misalnya dalam kasus penyakit pernapasan.

METODE PENELITIAN

Penelitian ini bertujuan membangun model prediksi tingkat kerusakan organ pernapasan akibat kebiasaan merokok dan faktor risiko lainnya. Penelitian menggunakan basis data klinis sebagai dasar pembangunan dan pengujian model. Tahapan penelitian meliputi: tinjauan literatur, pengumpulan dan pra-pemrosesan data, perancangan model, implementasi algoritma, dan evaluasi performa. Alur penelitian yang di gunakan dapat dilihat pada gambar berikut ini:



Dari gambar diatas dapat dijelaskan alur penelitian sebagi berikut :

2.1. Desain Penelitian

Penelitian menggunakan pendekatan kuantitatif dan metode *data mining* untuk memprediksi tingkat kerusakan paru berdasarkan data pasien. Model *ensemble* dibangun dengan Voting Classifier, yang menggabungkan tiga algoritma dasar: Random Forest, Logistic Regression, dan XGBoost. Algoritma tersebut telah terbukti efektif dalam

membangun model prediksi risiko penyakit, dengan akurasi yang lebih tinggi melalui pemilihan fitur yang optimal pada dataset (Wardani dan Akbar, 2022).

2.2. Pengumpulan Data

Data diambil dari Kaggle.com (open source), terdiri dari 1.000 record dengan 26 atribut. Data mencakup femografi: usia, jenis kelamin. Faktor risiko: merokok, perokok pasif, obesitas. Gejala klinis: batuk darah, sesak napas, nyeri dada. ariabel target adalah Level, yang menunjukkan tingkat keparahan kerusakan paru, terbagi dalam tiga kategori yaitu Low: gejala ringan atau tahap awal. Medium: gejala signifikan tapi belum berat. High: kondisi berat, berisiko tinggi terhadap penyakit kronis atau kanker paru.

2.3. Pra-Prosesing Data

Pemrosesan memiliki beberapa tahapan yaitu Label Encoding: mengubah nilai kategori Level menjadi angka agar bisa diproses oleh model, Hapus Duplikat: menghilangkan data yang sama persis untuk menjaga keunikan sampel. Serta deteksi & Hapus Outlier: outlier diidentifikasi dengan Z-score > 3 dan dihapus agar data lebih bersih dan representatif.

- Standarisasi Data: Fitur dinormalisasi menggunakan StandardScaler agar memiliki distribusi yang seragam
- Penanganan Ketidakseimbangan Data: Diterapkan teknik oversampling menggunakan SMOTE (Synthetic Minority Oversampling Technique) untuk menyeimbangkan jumlah data antar kelas.
- Pembagian Data: Setelah tahap pra-pemrosesan selesai, data dibagi menjadi dua bagian: 80% untuk data latih (*training set*) dan 20% untuk data uji (*testing set*) menggunakan metode *train-test split*. Untuk menjaga keseimbangan distribusi kelas target di kedua subset, diterapkan teknik *stratified sampling*. Pendekatan ini memastikan bahwa model dilatih secara optimal dan dievaluasi secara adil terhadap data baru yang belum pernah dilihat selama pelatihan.Model Arsitektur dan Algoritma

Model yang digunakan dalam penelitian ini adalah *Ensemble Voting Classifier* dengan pendekatan *soft voting*, di mana probabilitas prediksi dari setiap model dikombinasikan untuk menghasilkan keputusan akhir.

Algoritma yang digunakan dalam ensemble ini terdiri dari:

- XGBoost

Merupakan kumpulan decision tree di mana pembentukan pohon berikutnya bergantung pada pohon sebelumnya (Sykuron, 2020). Metode ini dikenal efisien dan sering digunakan dalam tugas klasifikasi karena kemampuannya mengelola data yang kompleks serta secara bertahap mengurangi kesalahan (error) melalui pendekatan iteratif.

- Random Forest

Algoritma ini beroperasi dengan membangun sejumlah besar pohon keputusan (decision trees) secara acak, lalu menggabungkan prediksi dari seluruh pohon tersebut untuk menghasilkan keputusan akhir yang lebih akurat dan mengurangi risiko overfitting (Nugroho, 2025).

- Logistic Regression

Adalah algoritma yang sederhana namun efektif untuk klasifikas dan dapat digunakan untuk memprediksi probabilitas variabel dependen kategori (Setyawan, 2025).

Seluruh model dikemas dalam satu pipeline terintegrasi yang mencakup penanganan ketidakseimbangan kelas menggunakan SMOTE dan *scaling* fitur dengan *Standard Scaler*. Dengan pendekatan ini, proses mulai dari pra-pemrosesan hingga pelatihan model berjalan secara konsisten dan efisien. Untuk mendapatkan performa terbaik, hyperparameter disesuaikan secara otomatis menggunakan GridSearchCV.

Setelah proses tuning, diperoleh kombinasi parameter optimal sebagai berikut:

- Random Forest
  - Jumlah pohon (n estimators) bisa diatur, misalnya 100 atau 200 pohon.
  - Kedalaman maksimum dari setiap pohon keputusan (max depth) dapat ditentukan, seperti tanpa batas (None), atau dibatasi hingga kedalaman 10 atau 15.
- XGBoost
  - Laju pembelajaran (learning rate) digunakan untuk mengatur seberapa besar langkah pembelajaran model, contohnya 0.01 atau 0.1.
  - Kedalaman maksimum pohon (max depth) juga bisa diatur, seperti 3, 6, atau 10.
- Parameter Umum
  - Minimum sampel yang dibutuhkan untuk membagi sebuah node bisa ditetapkan (min samples split), contohnya 5 sampel.
  - Nilai acak (random state) seperti 42 digunakan untuk memastikan hasil eksperimen bisa diulang secara konsisten.

Karena berfungsi sebagai baseline sederhana, model Logistic Regression digunakan sebagai salah satu komponen dalam ensemble Voting Classifier tanpa melakukan proses tuning parameter. Fokus tuning diberikan pada dua model utama, XGBoost dan Random Forest, yang memiliki ruang parameter dan kompleksitas yang lebih besar untuk dioptimalkan.

2.4. Evaluasi

Model dievaluasi menggunakan Stratified 5-Fold Cross Validation, dengan metrik: Precision, Recall, F1-Score, dan Confusion Matrix. Teknik ini membagi data menjadi 5 bagian, di mana setiap bagian memiliki proporsi kelas yang seimbang. Proses pelatihan dan pengujian diulang 5 kali: setiap kali, satu bagian dipakai sebagai data uji, dan empat sisanya untuk pelatihan.

Pemilihan 5-*Fold Cross* dipilih karena karena menyeimbangkan akurasi, efisiensi, dan stabilitas evaluasi sehingga hasilnya lebih akurat dan mencerminkan kemampuan model dalam menghadapi data baru (Bakheet. 2021).

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan algoritma Random Forest, mencakup tahapan: pra-pemrosesan data, penyeimbangan kelas, pengaturan hyperparameter, dan evaluasi model menggunakan validasi silang dan confusion matrix.

3.1. Dataset

Dataset berisi data klinis dan perilaku pasien terkait kebiasaan merokok dan gangguan pernapasan. Terdapat 20 fitur, seperti: merokok, gangguan tidur, kelelahan, batuk, dan sesak napas. Variabel target adalah Level, yang menunjukkan tingkat keparahan kondisi pasien.

3.2. Pra-Processing Data

Tujuannya adalah membersihkan dan mempersiapkan data teks agar siap untuk dianalisis (Limbong, 2024). Analisis awal menunjukkan bahwa distribusi kelas Level tidak seimbang—jumlah data pada kategori High, Medium, dan Low tidak proporsional. Untuk

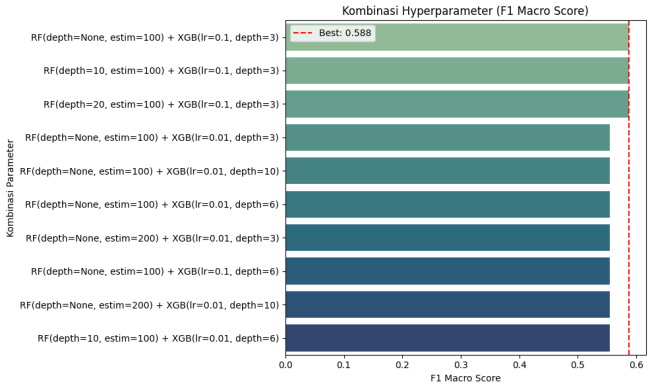
mengatasi ketidakseimbangan tersebut, diterapkan teknik SMOTE (Synthetic Minority Oversampling Technique), yang menghasilkan data sintetis untuk kelas minoritas melalui interpolasi antar titik data yang mirip.

3.3. Evaluasi Prediksi

Model dievaluasi menggunakan F1-Score, yaitu metrik yang menggabungkan presisi dan recall untuk memberikan ukuran kinerja yang lebih seimbang, terutama pada dataset yang tidak seimbang (Syahputra, 2025). Hasil pencarian hyperparameter terbaik (via Grid Search) menunjukkan:

- Random Forest: max_depth = None, n_estimators = 100
- XGBoost: learning_rate = 0.1, max_depth = 3

Kombinasi ini memberikan performa terbaik dalam validasi silang, dan digunakan sebagai konfigurasi akhir untuk pelatihan model.



Gambar 2. Evaluasi Kombinasi Hyperparameter RF + XGBoost (F1 Macro)

Gambar 2 menyajikan visualisasi sepuluh kombinasi parameter terbaik berdasarkan nilai rata-rata skor F1 Macro yang diperoleh dari proses validasi silang menggunakan Stratified 5-Fold Cross Validation. Setiap bar pada grafik mewakili satu kombinasi parameter dengan nilai mean test score yang dihasilkan. Kombinasi terbaik ditandai dengan garis putus-putus berwarna merah.

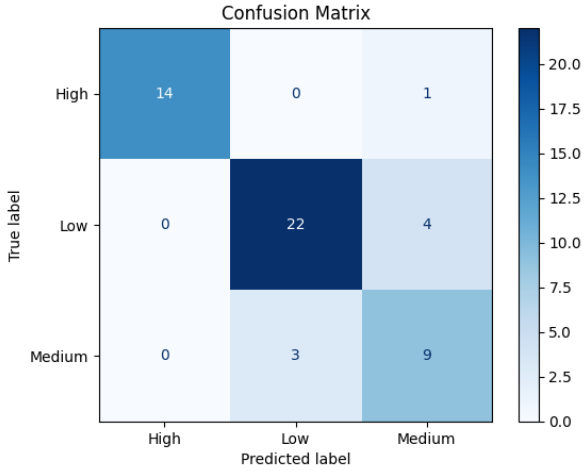
Dari hasil visualisasi tersebut, dapat dilihat bahwa kombinasi terbaik diperoleh pada konfigurasi Random Forest dengan n_estimators=100 dan max_depth=None, serta XGBoost dengan learning rate = 0.1 dan max depth = 3, yang menghasilkan skor F1 Macro tertinggi sebesar 0.588. Kombinasi ini menunjukkan bahwa Random Forest bekerja optimal dengan jumlah pohon sedang tanpa pembatasan kedalaman, sedangkan XGBoost lebih efisien dengan kedalaman pohon yang dangkal dan laju pembelajaran yang moderat.

Setelah dilakukan training dan validasi silang, model ensemble memberikan hasil sebagai berikut:

Table 2. Hasil Evaluasi Model Voting Classifier

	Precision	Recall	F1-Score	Support
High	1.00	0.93	0.97	15
Low	0.92	0.85	0.88	26
Medium	0.67	0.83	0.74	12

accuracy			0.87	53
Macro avg	0.86	0.87	0.86	53
Weighted avg	0.88	0.87	0.87	53



Gambar 3. Confusion Matrix Hasil Evaluasi Voting Classifier

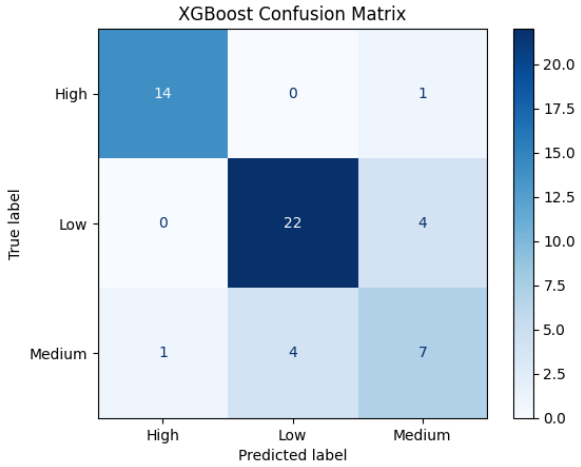
Berdasarkan classification report Voting Classifier pada Table 2, kelas High memiliki performa prediksi yang sangat tinggi, dengan nilai precision sebesar 1.00, recall 0.93, dan f1-score mencapai 0.97. Ini menunjukkan bahwa model mampu mengidentifikasi kelas High hampir tanpa kesalahan. Pada kelas Low, precision dan recall masing-masing sebesar 0.92 dan 0.85 dengan f1-score 0.88, yang juga menandakan performa klasifikasi yang baik dan seimbang. Sementara itu, kelas Medium menunjukkan hasil yang relatif lebih rendah, dengan precision sebesar 0.67, recall 0.83, dan f1-score 0.74. Meskipun demikian, hal ini masih dianggap wajar mengingat jumlah data kelas Medium yang lebih sedikit dan kemungkinan kemiripan pola fitur dengan kelas lainnya.

Secara keseluruhan, model memperoleh nilai accuracy sebesar 87%, dengan macro average f1-score sebesar 0.86 dan weighted average sebesar 0.87, yang mencerminkan bahwa model memiliki kinerja klasifikasi yang baik dan relatif stabil di antara semua kelas.

Selain itu, Gambar 3 menunjukkan bahwa dari 15 data dalam kelas *High*, sebanyak 14 berhasil diklasifikasikan dengan benar, sedangkan 1 data salah diprediksi sebagai kelas *Medium*. Pada kelas *Low*, 22 data diklasifikasikan secara akurat, namun 4 data lainnya keliru diprediksi sebagai *Medium*. Sementara itu, untuk kelas *Medium*, 9 data berhasil dikenali dengan tepat, tetapi 3 data salah diklasifikasikan sebagai *Low*. Temuan ini mengindikasikan bahwa kesalahan klasifikasi paling sering terjadi antara kelas *Low* dan *Medium*, kemungkinan besar disebabkan oleh kemiripan dalam karakteristik fitur kedua kelas tersebut.

Table 3. Hasil Evaluasi Model XGBoost

	precision	recall	f1-score	support
High	0.93	0.93	0.93	15
Low	0.85	0.85	0.85	26
Medium	0.58	0.58	0.58	12
accuracy			0.81	53
macro avg	0.79	0.79	0.79	53
Weighted avg	0.81	0.81	0.81	53

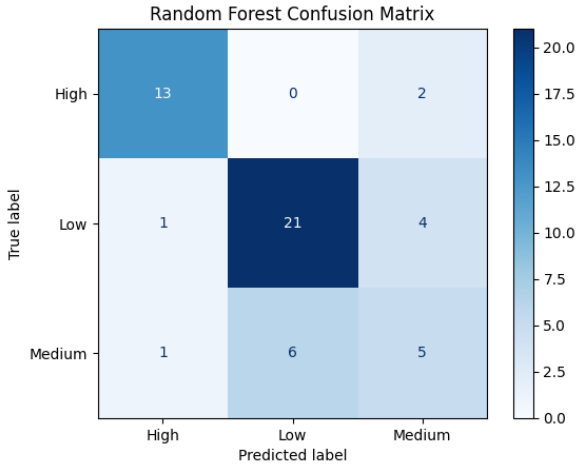


Gambar 4. Confusion Matrix Hasil Evaluasi Model XGBoost

Gambar 4 memberikan gambaran lebih rinci tentang pola kesalahan klasifikasi. Dari 12 data yang seharusnya termasuk dalam kelas Medium, hanya 7 yang berhasil diprediksi dengan benar, sementara 4 lainnya salah diklasifikasikan sebagai Low dan 1 sebagai High. Di sisi lain, dari kelas Low, terdapat 4 data yang salah masuk ke kelas Medium. Pola kesalahan ini menunjukkan adanya tumpang tindih yang signifikan dalam distribusi fitur antara kelas Medium dan Low, yang menyulitkan model dalam membedakan batas antar kategori tersebut. Secara keseluruhan, XGBoost menunjukkan performa yang sangat solid, terutama dalam mengklasifikasikan kelas High dan Low, namun masih menyisakan ruang perbaikan untuk meningkatkan akurasi pada kelas Medium. Rata-rata skor *macro* untuk *precision*, *recall*, dan *f1-score* berada di angka 0,79, yang menunjukkan bahwa model memiliki keseimbangan performa di seluruh kelas, meskipun belum optimal untuk kelas dengan karakteristik ambigu seperti Medium. Hasil ini semakin memperkuat posisi XGBoost sebagai model yang andal dalam konteks klasifikasi kesehatan pernapasan, terutama jika nantinya dikombinasikan dengan model lain dalam pendekatan *ensemble* guna meningkatkan stabilitas, generalisasi, dan kemampuan prediksi secara menyeluruh.

Table 4. Hasil Evaluasi Model Random Forest

	precision	recall	f1-score	support
High	0.87	0.87	0.87	15
Low	0.78	0.79	0.79	26
Medium	0.45	0.43	0.43	12
accuracy			0.74	53
macro avg	0.70	0.70	0.70	53
Weighted avg	0.73	0.73	0.73	53



Gambar 5. Confusion matrix Hasil Evaluasi Model Random Forest

Berdasarkan *classification report* Random Forest (Tabel 4), model ini memiliki akurasi keseluruhan 74% — cukup baik, tapi masih di bawah XGBoost dan VotingClassifier.

- Kelas High: performa terbaik dengan *precision, recall,* dan *f1-score* masing-masing 0.87 yang artinya model sangat akurat mengenali pasien dengan kerusakan paru berat.
- Kelas Low: *precision 0.78, recall 0.81, f1-score 0.79* — cukup baik, tapi masih ada kesalahan.
- Kelas Medium: performa paling rendah — *precision 0.45, recall 0.42, f1-score 0.43* — menunjukkan model kesulitan membedakan kelas ini.

Gambar 5 memperjelas:

- Dari 15 data High, 13 benar, 2 salah (terprediksi sebagai Medium).
- Dari 26 data Low, 21 benar, 4 salah ke Medium, 1 ke High.
- Dari 12 data Medium, hanya 5 benar — 6 salah ke Low, 1 ke High.

Table 5. Hasil Evaluasi Model Logistic Regression

	precision	recall	f1-score	support
High	1.00	0.93	0.97	15
Low	0.92	0.85	0.88	26
Medium	0.67	0.83	0.74	12
accuracy			0.87	53
macro avg	0.86	0.87	0.86	53
Weighted avg	0.88	0.87	0.87	53

Evaluasi model Logistic Regression (Tabel 5) menunjukkan performa cukup baik secara umum, dengan akurasi 77%. Performa terbaik terlihat pada kelas High dan Low, masing-masing dengan *precision, recall,* dan *f1-score* sebesar 0.87 dan 0.85 — artinya model cukup akurat mengenali pola di kedua kelas ini. Namun, pada kelas Medium, performanya jauh lebih rendah: *precision, recall,* dan *f1-score* hanya 0.50. Ini menunjukkan model kesulitan membedakan kelas Medium, kemungkinan karena fitur yang mirip dengan kelas lain atau jumlah data yang tidak seimbang.

Table 6. Perbandingan Performa Model Klasifikasi

Model	Accuracy	Precision	Recall	F1-Score
Voting Classifier	87%	0.88	0.87	0.87
XGBoost	81%	0.81	0.81	0.81
Logistic Regression	77%	0.77	0.77	0.77
Random Forest	74%	0.73	0.74	0.73

Berdasarkan hasil evaluasi pada Tabel 6, *Voting Classifier Ensemble* merupakan model dengan performa paling optimal dalam memprediksi tingkat kerusakan organ pernapasan. Model ini meraih akurasi tertinggi sebesar 87%, mengungguli XGBoost (81%), Logistic Regression (77%), dan Random Forest (74%). Selain akurasi, metrik rata-rata *precision*, *recall*, dan F1-*score* dari *Voting Classifier* juga menunjukkan hasil terbaik, masing-masing sebesar 0,88; 0,87; dan 0,87. Temuan ini mengindikasikan bahwa pendekatan ensemble berhasil mengintegrasikan kekuatan beberapa algoritma—yaitu XGBoost, Random Forest, dan Logistic Regression—sehingga menghasilkan prediksi yang lebih stabil, seimbang, dan akurat.

SIMPULAN DAN SARAN

a. Simpulan

Merokok adalah faktor risiko utama yang menyebabkan kerusakan serius pada sistem pernapasan. Untuk mencegah penyakit kronis seperti PPOK dan kanker paru, diperlukan deteksi dini melalui model prediktif. Penelitian ini menggunakan pendekatan *ensemble machine learning* berbasis Voting Classifier, yang menggabungkan tiga algoritma: XGBoost, Random Forest, dan Logistic Regression. Proses pengembangan model mencakup pengumpulan data, pra-pemrosesan, penyetelan hyperparameter dengan GridSearchCV, dan validasi menggunakan Stratified 5-Fold Cross Validation.

Hasil evaluasi menunjukkan bahwa Voting Classifier memberikan performa terbaik dibanding model tunggal, dengan akurasi 87%, presisi 0,88, recall 0,87, dan f1-score 0,87.

Model ini sangat akurat dalam mengklasifikasikan pasien ke dalam kategori kerusakan High dan Low, dengan tingkat kesalahan prediksi yang rendah. Namun, untuk kelas Medium, hasilnya masih kurang optimal karena jumlah datanya lebih sedikit dan pola fiturnya mirip dengan kelas lain.

b. Saran

Berdasarkan hasil penelitian, berikut beberapa saran untuk pengembangan di masa mendatang Untuk meningkatkan performa model prediksi, langkah pertama yang dapat dilakukan adalah menyeimbangkan data antar kelas, terutama untuk kelas Medium yang menunjukkan performa lebih rendah. Hal ini bisa diatasi dengan menambah jumlah data pada kelas minoritas atau menerapkan teknik lainnya. Selanjutnya, eksplorasi penggunaan algoritma lain seperti LightGBM, CatBoost, atau bahkan model deep learning dapat dilakukan untuk mengevaluasi apakah pendekatan tersebut memberikan hasil prediksi yang lebih baik, khususnya pada kelas-kelas yang sulit diprediksi. Selain itu, penerapan metode interpretasi model seperti feature importance untuk membantu mengidentifikasi fitur-fitur paling berpengaruh dalam proses prediksi, yang sangat berguna bagi tenaga kesehatan dalam mendukung keputusan klinis. Terakhir, pengembangan model menjadi aplikasi web atau mobile akan memungkinkan pemanfaatannya sebagai alat skrining awal, baik oleh dokter maupun masyarakat umum, sehingga meningkatkan aksesibilitas dan dampak praktis dari solusi ini.

DAFTAR PUSTAKA

Bakheet dan A. Al-Hamadi, (2021). Automatic detection of COVID-19 using pruned GLCM-Based texture features and LDCRF classification, Comput. Biol. Med., vol. 137, no. June, hal. 104781, 2021

Djajalaksana, Susanthy. (2023). Wujudkan Generasi Bebas Tembakau. Jurnal Klinik dan Riset Kesehatan. vol 4 No 3

World Health Organization (WHO). Tobacco Fact Sheet, https://www.who.int/news-room/fact-sheets/detail/tobacco, diakses tgl 11 Juni 2025

Limbong, Hendra Halomoan, Norhikmah. (2024). Optimasi Analisis Sentimen Ulasan Aplikasi Amikom One Menggunakan SMOTE pada Algoritma Artificial Neural Network. Sistemasi:Jurnal Sistem Informasi Volume 13, Nomor 5

Lina Putri Prabawati, Siti Nurhidayah. (2024). Problematika Rokok di Indonesia: Pemetaan Masalah dan Prediksi Kebijakan Pengendalian Konsumsi Rokok Kalangan Remaja, Vol 1, No5

Nugroho, Muhammad Wisnu. (2025). Analisis Performa Algoritma Random Forest dalam Mengatasi Overfitting pada Model Prediksi. Jurnal Teknologi Informasi dan Komunikasi Volume 9 No 4

Setyawan, Nanda Hendra, Nur Wakhidah. (2025). Analisis Perbandingan Metode Logistic Regression, Random Forest, Gradient Boosting Untuk Prediksi Diabetes. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika). Vol. 10, No. 1

Syahputra, Firman, Dahri Yani Hakim Tanjung, Wiwi Verina, Muhammad Ihsan, Rofiqoh Dewi, Andi Sanjaya. (2025). Analisis Kinerja Algoritma Klasifikasi terhadap

Dataset Penerimaan Pegawai Outsourcing. Jurnal Minfo Polgan Volume 14, Nomor 1.

Syukron, Muhamad , Rukun Santoso, Tatik Widiharih, (2020). Perbandingan Metode Smote Random Forest Dan Smote XGBOOST Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data, Jurnal Gaussian, Vol 9 , No 3

Wardani, Kartina Diah Kusuma, Akbar. Memen, (2022). Diabetes Risk Prediction using Feature Importance Extreme Gradient Boosting (XGBoost), J. RESTI (Rekayasa Sist. dan Teknologi Informasi), vol. 7, no. 4

Wibisono, Setyawan, Endang Lestariningsih, Wiwien Hadikurniawati,Heribertus Yulianton, Taufiq Dwi Cahyono. (2024). Optimalisasi Model Klasifikasi Diabetes Menggunakan Ensemble Learning Adaboost, Gradient Boosting, dan XGBoost. Jurnal Teknik Informatika Unika ST. Thomas (JTIUST), Volume 09 Nomor 02.