ANALISIS *MACHINE LEARNING* UNTUK PREDIKSI PENYAKIT PARU-PARU MENGGUNAKAN *RANDOM FOREST*

Ade Christian[1], Hariyanto[2], Ahmad Yani[3], Sumanto[4]
[1] Teknologi Informasi, Universitas Bina Sarana Informatika,
[2] Sistem Informasi, Universitas Bina Sarana Informatika
[3] Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika
[4] Informatika, Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98 RT.2/RW.9, Kwitang, Kec, Senin, Jakarta Pusat, Indonesia
e-mail: [1]**ade.adc@bsi.ac.id** , [2] hariyanto.hro@bsi.ac.id, [3]ahmad.amy@bsi.ac.id, *[4]sumanto@bsi.ac.id

Abstract
Lung diseases, including COPD, lung cancer, and asthma, are serious global health issues, causing over seven million deaths annually. Advanced technologies, such as deep learning and the Random Forest algorithm, have been effectively utilized to detect and classify lung diseases from imaging data with high accuracy. This study aims to demonstrate the effectiveness of Random Forest in predicting lung diseases. The dataset used consists of 30,000 records with 11 attributes, collected from Kaggle and processed using Orange software version 3.36.2. The implementation of the Random Forest algorithm was conducted with 10 decision trees and six attributes considered at each split. The model was tested using Cross Validation with 10 folds. The testing results showed an AUC value of 0.993, indicating a very high level of accuracy. A confusion matrix was used to measure the model's performance through various metrics, including accuracy, precision, recall, F1-score, and AUC. This model achieved high accuracy, with ROC AUC values of 0.453 for predicting the presence of lung disease and 0.547 for predicting its absence. These results confirm that the Random Forest algorithm is an effective predictive tool for identifying lung diseases. This study makes a significant contribution to the development of more accurate and efficient diagnostic techniques, assisting medical professionals in identifying lung diseases in patients. With a deeper understanding of how this algorithm operates in the healthcare domain, it is expected to significantly enhance the quality of patient diagnosis and care.

Keyword: *Lung Disease, Machine learning, Orange Software, RF, Random Forest Algorithm*

PENDAHULUAN

Kesehatan adalah aset berharga, dan paru-paru berperan penting dalam sistem pernapasan. Namun, berbagai penyakit paru-paru, termasuk kanker paru, menjadi ancaman global dengan tingkat kematian yang tinggi. Di Indonesia, kanker paru merupakan penyebab utama kematian akibat kanker, yang sebagian besar dipicu oleh kebiasaan merokok dan paparan polusi udara, menjadikannya tantangan serius dalam upaya kesehatan masyarakat (Sandika et al., 2024) (Putra et al., 2024). Penyakit paru-paru mencakup berbagai gangguan yang menyerang saluran pernapasan dan struktur paru, seperti PPOK, kanker paru, asma, bronkiektasis, penyakit paru interstitial, gangguan akibat paparan lingkungan kerja, serta hipertensi paru. Kondisi-kondisi ini dapat menghambat fungsi pernapasan dan berpotensi menimbulkan dampak serius terhadap kesehatan (Gould et al., 2023). Di seluruh dunia, penyakit paru-paru menjadi isu kesehatan utama, menyebabkan lebih dari tujuh juta kematian setiap tahun akibat kondisi seperti PPOK, infeksi saluran pernapasan bawah, dan kanker paru-paru (Heitlinger, 2023), Oleh karena itu, deteksi dini menjadi langkah krusial dalam upaya pencegahan dan penanganan penyakit paru-paru, guna meningkatkan peluang pengobatan yang lebih efektif dan mengurangi angka kematian.

Teknologi canggih seperti model *machine learning* telah berhasil digunakan untuk mendeteksi dan mengklasifikasikan penyakit paru-paru seperti pneumonia, tuberkulosis, dan kanker paru-paru dari data pencitraan, menunjukkan tingkat akurasi dan efektivitas yang tinggi dalam deteksi penyakit (Priyadarsini et al., 2023) (Swartzendruber et al., 2020). Keberhasilan ini menunjukkan potensi pendekatan berbasis kecerdasan buatan dalam meningkatkan diagnosis penyakit pernapasan. Di negara-negara seperti Indonesia, implementasi jaringan saraf buatan seperti LVQ3 telah diterapkan untuk mendiagnosis berbagai penyakit paru-paru dengan hasil yang menjanjikan (Midyanti et al., 2020). Kemajuan ini menekankan pentingnya inovasi diagnostik dalam meningkatkan manajemen kesehatan paru-paru dan mempercepat deteksi dini penyakit pernapasan. Beberapa penelitian terkait deteksi penyakit paru-paru terangkum tabel 1.

Tabel 1. Penelitian Terdahulu

Peneliti	Data	Algoritma	Akurasi
(Sofyan et al., 2023)	Penyakit Paru-Paru	C4.5	89.77%
(Wahid et al., 2023)	Kanker paru-paru	*Regresi linier*	90%
(Prasetyo et al., 2022)	Citra paru-paru	*SVM*	79%.%.
(Musa & Alang, 2017)	Penyakit Paru-Paru	*K-NEAREST NEIGHBORS*	91.90%
(Yunianto et al., 2021)	Citra paru normal dan citra kanker paru	*NAÏVE BAYES*	88,33 %.

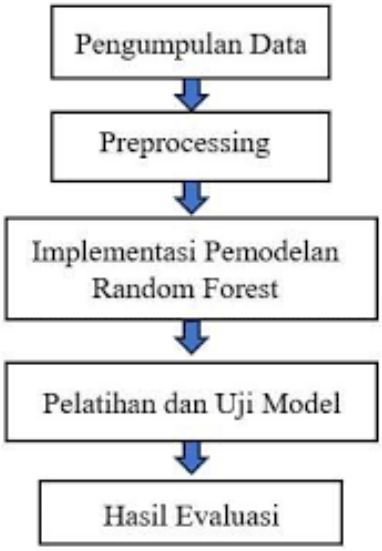
Tabel 1 menyajikan hasil penelitian tentang penyakit paru-paru menggunakan berbagai metode analisis data. (Sofyan et al., 2023) menggunakan metode C4.5 untuk menganalisis penyakit paru-paru dengan akurasi sebesar 89,77%. (Wahid et al., 2023) menerapkan *regresi linier* untuk kanker paru-paru dan mencapai akurasi 90%. (Prasetyo et al., 2022) menggunakan *Support Vector Machine (SVM)* untuk menganalisis citra paru-paru dan memperoleh akurasi 79%. (Musa & Alang, 2017) menggunakan metode *K-Nearest Neighbors* untuk penyakit paru-paru dan mendapatkan akurasi tertinggi, yaitu 91,90%. (Yunianto et al., 2021) menggunakan *Naïve Bayes* untuk menganalisis citra paru normal dan citra kanker paru dengan akurasi 88,33%.

Berdasarkan berbagai penelitian sebelumnya, metode seperti C4.5, regresi linier, SVM, K-Nearest Neighbors (KNN), dan Naïve Bayes telah digunakan untuk menganalisis dan memprediksi penyakit paru-paru, masing-masing dengan tingkat akurasi yang bervariasi. Meskipun KNN mencatat akurasi tertinggi (91,90%), metode ini memiliki keterbatasan dalam skalabilitas dan sensitivitas terhadap data berdimensi tinggi. Sementara itu, SVM hanya mencapai akurasi 79%, menunjukkan keterbatasannya dalam menangani data citra medis yang kompleks. Regresi linier meskipun cukup akurat (90%), kurang optimal dalam menangani hubungan non-linear dalam data kesehatan. *Random Forest* (RF) hadir sebagai solusi unggul, karena mampu mengatasi *overfitting* yang sering terjadi pada C4.5, menangani data berdimensi tinggi lebih baik dibandingkan KNN, serta memiliki fleksibilitas lebih tinggi dibandingkan *Naïve Bayes* dalam menangani variabel yang saling bergantung. Selain itu, RF dapat menangani missing values dengan lebih efektif, meningkatkan akurasi prediksi dan keandalan dalam menganalisis data medis. Dengan keunggulannya dalam mengidentifikasi fitur penting dalam data kesehatan dan memberikan prediksi yang lebih stabil, *Random Forest* menjadi metode yang sangat cocok untuk deteksi dini penyakit paru-paru, membantu dalam pengambilan keputusan medis yang lebih akurat dan efisien. Penelitian ini bertujuan untuk menganalisis efektivitas algoritma Random Forest (RF) dalam mendeteksi penyakit paru-paru dan membandingkannya dengan metode lain seperti C4.5, regresi linier, SVM, KNN, dan Naïve Bayes. Dengan keunggulannya dalam mengatasi *overfitting*, menangani data berdimensi tinggi, serta mengelola *missing values* secara lebih efektif, RF diharapkan dapat memberikan akurasi prediksi yang lebih tinggi dan stabil. Selain itu, penelitian ini bertujuan untuk mengidentifikasi fitur penting dalam data kesehatan guna mendukung diagnosis yang lebih akurat dan efisien, serta menjadi dasar bagi pengembangan model *machine learning* untuk aplikasi medis.

METODE PENELITIAN

Penelitian ini dilakukan melalui proses yang sistematis dan terarah guna menyelesaikan permasalahan yang dikaji. Langkah-langkah penelitian diawali dengan pengumpulan data, kemudian dilanjutkan dengan tahap *preprocessing*, yang bertujuan untuk membersihkan serta mempersiapkan data sebelum dilakukan analisis lebih lanjut. Selanjutnya, dilakukan

implementasi pemodelan *Random Forest* , di mana algoritma diterapkan untuk membangun model prediksi. Setelah itu, model akan melalui tahap pelatihan dan pengujian guna mengevaluasi kinerjanya. Tahap terakhir adalah analisis hasil evaluasi, di mana performa model diukur berdasarkan akurasi dan efektivitasnya dalam mendeteksi penyakit paru-paru. Tahapan penelitian diilustrasikan dalam Gambar 1.



Gambar 1. Tahapan Penelitian (Siregar et al., 2023)

1. Pengumpulan Dataset

Dataset yang digunakan diperoleh dari Kaggle (https://www.kaggle.com/) sebagai sumber data terbuka yang dapat diakses secara publik. Dataset ini mencakup 30.000 sampel dengan 11 variabel, yang mencerminkan berbagai faktor kesehatan dan gaya hidup individu. Variabel-variabel tersebut meliputi identifikasi unik, rentang usia, jenis kelamin, kebiasaan merokok, status pekerjaan, struktur keluarga, pola tidur (terutama kebiasaan begadang), tingkat aktivitas fisik, kepemilikan asuransi kesehatan, riwayat penyakit bawaan, serta hasil akhir pemeriksaan kesehatan (Sriyanto & Supriyatna, 2023).

2. Praprosess Dataset

Analisis data dalam penelitian ini dilakukan menggunakan software Orange versi 3.36.2. Sebelum proses analisis, dilakukan tahap prapemrosesan data untuk memastikan bahwa dataset tersedia dalam format yang sesuai dan dapat diinterpretasikan dengan baik oleh Orange. Tahapan ini mencakup pembersihan data, normalisasi, serta pengolahan atribut agar hasil analisis lebih optimal dan akurat (Sriyanto & Supriyatna, 2023).

3. Implementasi *Random Forest*

Data yang diperoleh dari langkah sebelumnya akan diterapkan ke alat data mining, yaitu menggunakan *Orange* Data Mining. Pada tahap implementasi, langkah-langkah pengisian data ke dalam alat ini dilakukan, dan hasilnya akan menampilkan prediksi terkait penyakit paru-paru. Berikut adalah langkah-langkah yang dijalankan dalam implementasi algoritma *Random Forest* (Utami & Saptiari, 2020).

4. Pelatihan dan Uji Model

Tahap pelatihan dan uji model bertujuan untuk mengevaluasi performa *Random Forest* dalam memprediksi penyakit paru-paru. Pada proses ini, data yang telah mengalami tahap pra-pemrosesan dipisahkan menjadi dua bagian utama, yaitu data pelatihan (*training set*) yang digunakan untuk membangun model dan data pengujian (*testing set*) yang berfungsi untuk mengevaluasi kinerja model. Data pelatihan model menggunakan algoritma *Random Forest* membangun sekumpulan pohon keputusan berdasarkan pola yang ditemukan dalam data. Setelah model dilatih, dilakukan pengujian menggunakan data testing untuk menilai akurasi, presisi,

recall, dan nilai F1-*score*. Selain itu, digunakan metode validasi silang (*cross-validation*) untuk menghindari bias dan memastikan bahwa model memiliki generalisasi yang baik terhadap data baru. Hasil dari tahap ini akan memberikan gambaran tentang efektivitas *Random Forest* dalam menangani data medis, serta menunjukkan sejauh mana model dapat digunakan untuk deteksi dini penyakit paru-paru secara andal.

5. Hasil Evaluasi

Proses penilaian model bertujuan untuk mengukur efektivitas RF dalam mengidentifikasi serta mengelompokkan kasus penyakit paru-paru berdasarkan data uji. Kinerja model dievaluasi menggunakan sejumlah metrik, termasuk akurasi, presisi, *recall*, dan F1-score, yang membantu menilai sejauh mana prediksi yang dihasilkan mendekati kondisi sebenarnya. Selain itu, *confusion matrix* digunakan untuk mengidentifikasi kesalahan klasifikasi dan memahami bagaimana model membagi data ke dalam kategori yang benar. Untuk meningkatkan keandalan serta menghindari *overfitting*, dilakukan validasi silang (*cross-validation*) guna memastikan bahwa model tetap stabil saat diterapkan pada data baru. Hasil dari tahapan ini akan menjadi indikator utama dalam menilai sejauh mana RF dapat dijadikan pendekatan yang efektif dalam deteksi dini penyakit paru-paru, sekaligus memberikan dasar bagi peningkatan model di masa mendatang.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dan pembahasan dari penelitian yang telah dilakukan, mencakup analisis data, penerapan algoritma, evaluasi performa model, serta interpretasi temuan yang diperoleh. Hasil yang diperoleh akan dibandingkan dengan penelitian sebelumnya guna menilai efektivitas metode yang digunakan, sementara pembahasan akan menjelaskan makna dari hasil tersebut serta implikasinya dalam konteks deteksi penyakit paru-paru.

1. Pengumpulan Dataset

Dataset yang digunakan diperoleh dari platform Kaggle (https://www.kaggle.com/) dan ditampilkan dalam bentuk cuplikan layar pada Gambar 2 untuk memberikan gambaran mengenai struktur data yang digunakan dalam penelitian ini.

Data Table (1) - Orange

	Hasil	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begadang	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan
1	Ya	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak
2	Tidak	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada
3	Tidak	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
4	Tidak	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada
5	Ya	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
6	Tidak	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
7	Ya	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak
8	Tidak	Muda	Pria	Aktif	Tidak	Ya	Ya	Sering	Tidak	Tidak
9	Ya	Tua	Wanita	Aktif	Ya	Ya	Ya	Jarang	Ada	Ada
10	Ya	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada
11	Ya	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada
12	Tidak	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak
13	Tidak	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
14	Tidak	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada
15	Ya	Muda	Wanita	Pasif	Ya	Tidak	Ya	Sering	Tidak	Ada
16	Ya	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada
17	Ya	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada
18	Tidak	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak
19	Tidak	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
20	Tidak	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada
21	Ya	Muda	Wanita	Pasif	Ya	Tidak	Ya	Sering	Tidak	Ada
22	Ya	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak
23	Tidak	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada
24	Tidak	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
25	Tidak	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada
26	Ya	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
27	Ya	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada
28	Ya	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada
29	Tidak	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak
30	Tidak	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak
31	Tidak	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada
32	Tidak	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada
33	Ya	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak
34	Tidak	Muda	Pria	Aktif	Tidak	Ya	Ya	Sering	Tidak	Tidak

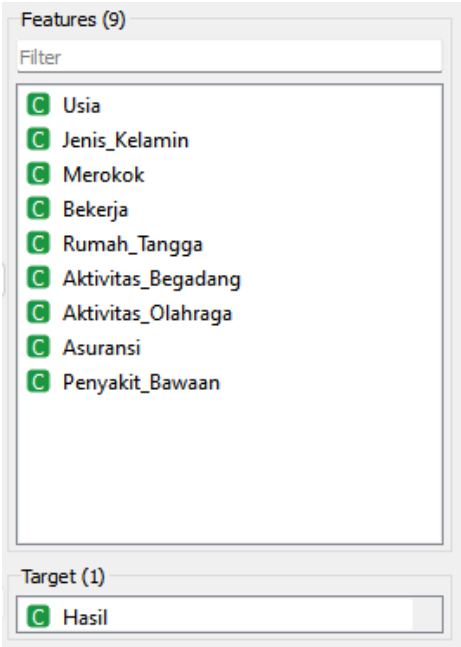
30k | 30k | 30k

Gambar 2. Tangkapan Layar Dataset Paru-Paru

2. Praproses Dataset

Dalam penelitian ini, dataset telah dianggap baik dan tidak mengandung nilai yang kosong (*missing value*). Tahap pra-pemrosesan data melibatkan pemilihan atribut yang akan digunakan sebagai atribut fitur dan atribut target. Dalam kasus ini, atribut yang diambil sebagai target adalah

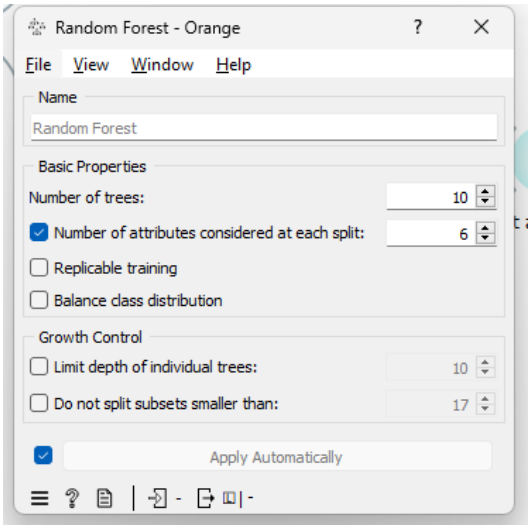
atribut "Hasil", sementara atribut lainnya akan dijadikan atribut fitur. Pada gambar 3 disajikan rincian dataset yang digunakan.



Gambar 3. Features dan Target

3. Implementasi Algoritma Random Forest

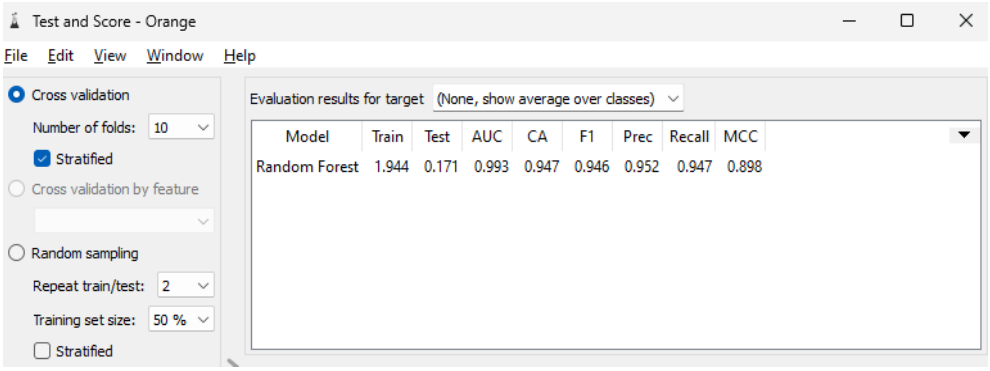
Parameter pengujian yang digunakan dalam penelitian ini meliputi *Number of trees*: 10. *Number of atrributess considered at each split*: 6. Detail dari *parameter* yang digunakan dalam algoritma *Random Forest* dapat dilihat pada Gambar 4.



Gambar 4. Parameter *Random Forest*

4. Pengujian Model

Model diuji menggunakan *Cross Validation* dengan *Number Fold* 10. Berdasarkan hasil pengujian model, diperoleh nilai *AUC* sebesar 0,993. Untuk detail hasil pengujian lengkap, dapat dilihat pada Gambar 4.



Gambar 4. Pengujian Model

5. Matrik Konfusi

Confusion matrix/matrik konfusi yang digunakan dalam penelitian ini terdiri dari empat elemen utama, yaitu True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN). Tabel ini digunakan untuk mengevaluasi hasil prediksi model dengan membandingkan kelas aktual dan kelas yang diprediksi (Sriyanto & Supriyatna, 2023). Tabel 2 menyajikan hasil perhitungan *confusion matrix* yang digunakan dalam penelitian ini.

Tabel 2. Confusion matrix/matrik konfusi

Fakta	Prediksi	
	Negatif	*Positif*
Negatif	TN *(True Negative)*	FP *(False Positive)*
Positif	FN (False Negative)	TP (True Positive)

Terdapat beberapa metrik kinerja yang umumnya digunakan, di antaranya adalah sebagai berikut:

a. Akurasi (*Accuracy*)

Akurasi mengindikasikan seberapa efektif model dalam mengelompokkan data dengan benar secara keseluruhan. Nilai ini diperoleh dengan menghitung rasio antara jumlah prediksi yang tepat terhadap total data yang digunakan, sehingga mencerminkan kinerja model dalam melakukan klasifikasi secara akurat (Nugroho, 2019 dalam sendy, 2023). Persentase akurasi dalam penelitian ini dihitung menggunakan rumus berikut.

$$Accurary = \frac{TP+TN}{TP+TN+FP+FN} \quad [1]$$

b. Presisi (*Precision*)

Presisi *(Positive Predictive Value)* digunakan untuk menilai kinerja sistem dengan menghitung data yang diklasifikasikan dengan benar dan data yang salah diklasifikasikan. Data yang terklasifikasikan dengan benar menjadi acuan untuk memperoleh nilai presisi, dengan membaginya dengan hasil prediksi *False Positive*, yaitu data yang terprediksi tidak tepat. Hal ini memungkinkan penentuan sejauh mana kesesuaian antara data acuan dengan data prediksi. Dengan demikian, semakin banyak data acuan yang terprediksi tidak tepat dalam proses klasifikasi, nilai presisi akan semakin kecil (Nugroho, 2019 dalam (Sendy, 2023). Persentase presisi dalam penelitian ini dihitung menggunakan rumus berikut.

$$Precision = \frac{TP}{TP+FP} \quad [2]$$

c. *Recall* atau Sensitivity

Recall adalah metode untuk mengevaluasi kinerja suatu sistem dalam mengidentifikasi kembali informasi. Ini membandingkan rasio data yang diprediksi dengan benar terhadap keseluruhan data yang sebenarnya *positif* (Nugroho, 2019 dalam sendy, 2023).

$$Recall = \frac{TP}{TP+FN} \quad [3]$$

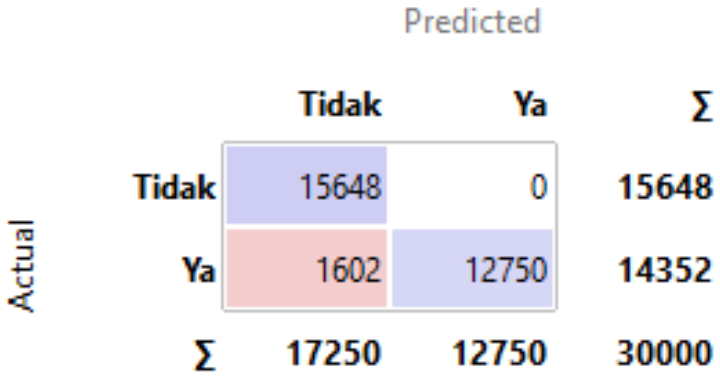
d. F1-Score

F1-Score adalah nilai yang membandingkan nilai rata-rata antara presisi dan *Recall*. Rumus yang digunakan untuk menghitung nilai *F1-Score* adalah sebagai berikut (Saputro & Sari, 2019 dalam (Sendy, 2023).

$$\text{F1-Score} = 2\ x\ \frac{Recall \text{ x } Precision}{Recall+Precision} \quad [4]$$

e. Nilai *Area Under Curve* (*AUC*)

UC (*Area Under the* Curve) mengukur luas di bawah kurva *Receiver Operating Characteristic* (ROC) dan memiliki nilai yang berkisar antara 0,5 hingga 1. Interpretasi nilai AUC terbagi dalam lima kategori: 0,5–0,6 menunjukkan akurasi rendah, 0,6–0,7 mengindikasikan akurasi lemah, 0,7–0,8 menandakan akurasi sedang, 0,8–0,9 mencerminkan akurasi tinggi, dan 0,9–1 menggambarkan tingkat akurasi yang sangat baik (Sriyanto & Supriyatna, 2023).

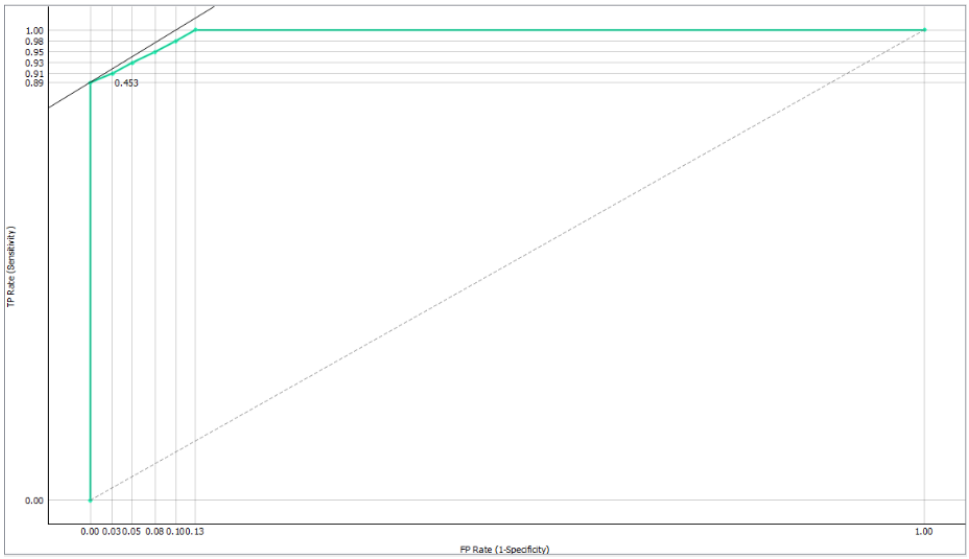


Gambar 6. Matrik konfusi

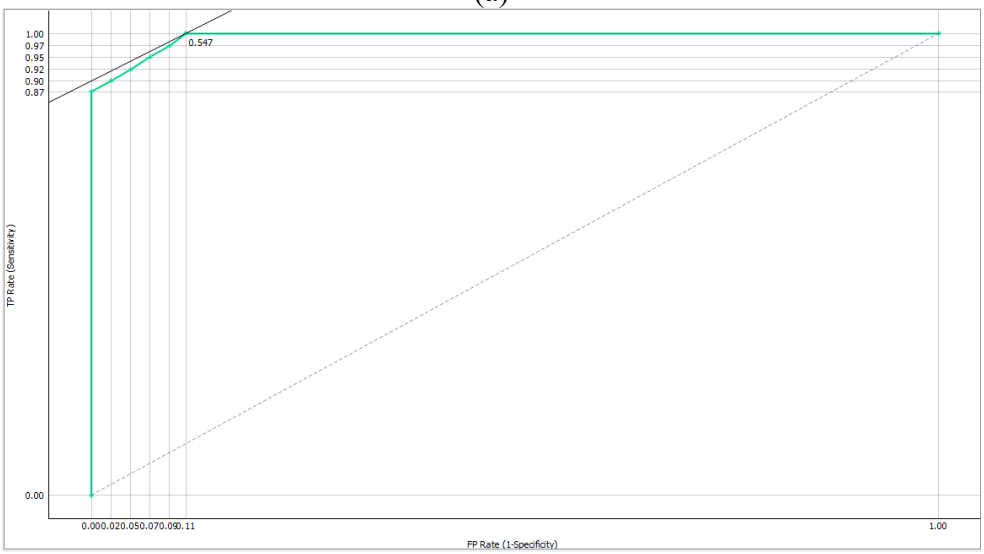
Dengan menggunakan 30.000 data latih dan menjalankan proses perulangan sebanyak 100 kali, model menghasilkan akurasi sebesar 12.750 untuk kategori "Ya" dan 15.648 untuk kategori "Tidak". Hasil perhitungan lebih lanjut yang diperoleh dari *confusion matrix* dapat dilihat pada Gambar 6 di atas.

6. Hasil Nilai AUC

Hasil prediksi penyakit paru-paru divisualisasikan melalui kurva ROC, yang menunjukkan performa model dalam membedakan kategori. Gambar 7 menyajikan Kurva ROC untuk kategori "Ya" (a) dengan nilai 0.453 dan kategori "Tidak" (b) dengan nilai 0.547, yang merepresentasikan tingkat keakuratan model dalam klasifikasi data.



(a)



(b)

Gambar 7. Kurva *ROC* (a) Ya dan *ROC* (b) Tidak

Gambar 7.a menampilkan kurva ROC yang menunjukkan hubungan antara *False Positive Rate* (sumbu x) dan *True Positive Rate* (sumbu y), dengan nilai AUC sebesar 0.453 sebagai indikator prediksi keberadaan penyakit paru-paru. Hasil ini menunjukkan bahwa model memiliki akurasi tinggi, karena mendekati titik 0.1. Sementara itu, Gambar 7.b juga memperlihatkan kurva ROC dengan pola serupa, di mana nilai AUC sebesar 0.547 digunakan untuk memprediksi ketidakhadiran penyakit paru-paru. Hasil ini menunjukkan bahwa model tetap memiliki akurasi yang baik, dengan performa klasifikasi yang stabil.

Hasil perbandingan dengan penelitian sebelumnya:

Hasil penelitian pada Tabel 3 menunjukkan bahwa *Random Forest* unggul dalam akurasi dibandingkan dengan metode lainnya, mencapai 94.7%, yang lebih tinggi dibandingkan *K-Nearest Neighbors* (91.90%), C4.5 (89.77%), dan *Naïve Bayes* (88.33%). Keunggulan ini menegaskan bahwa *Random Forest* memiliki performa lebih baik dalam menangani data medis, terutama dalam prediksi penyakit paru-paru, dengan stabilitas dan keandalan yang lebih tinggi dibandingkan metode sebelumnya.

Tabel 3. Perbandingan hasil penelitian penyakit paru-paru

Algoritma	Dataset	Akurasi
C4.5	Penyakit Paru-Paru	89.77%
Regresi Linier	Kanker Paru-Paru	90%
SVM	Citra Paru-Paru	79%
K-Nearest Neighbors	Penyakit Paru-Paru	91.90%
Naïve Bayes	Citra Paru Normal dan Kanker Paru	88.33%
Random Forest	Penyakit Paru-Paru	**94.7%**

KESIMPULAN

Hasil implementasi algoritma *Random Forest* dalam prediksi penyakit paru-paru menunjukkan kinerja yang sangat baik dengan tingkat akurasi 94,7%, F1-score 0.946, Precision 0.952, dan *Recall* 0.947. Pengujian dilakukan dengan parameter *Number of Trees*: 10 dan *Number of Attributes Considered at Each Split*: 6 menggunakan *tools Orange Data Mining*. Hasil evaluasi berdasarkan *confusion matrix* menunjukkan bahwa model memiliki tingkat kesalahan klasifikasi yang rendah, sementara analisis kurva evaluasi yang mendekati titik 0.1 mengkonfirmasi stabilitas model dalam melakukan klasifikasi data. Hal ini membuktikan bahwa *Random Forest* mampu menggeneralisasi data dengan baik dan memberikan prediksi yang andal, menjadikannya metode yang layak digunakan dalam pengembangan model prediksi penyakit paru-paru atau kasus serupa.

Meskipun hasil penelitian ini telah menunjukkan tingkat akurasi yang tinggi, beberapa aspek masih dapat ditingkatkan dalam penelitian lanjutan. Salah satu perbaikan yang dapat dilakukan adalah optimasi parameter model melalui *hyperparameter tuning*, seperti *Grid Search* atau *Random Search*, guna mencari kombinasi jumlah pohon keputusan dan atribut terbaik untuk meningkatkan performa prediksi. Selain itu, pengujian dengan dataset yang lebih besar dan beragam diperlukan untuk memastikan bahwa model dapat beradaptasi dengan berbagai kondisi medis dan karakteristik pasien yang lebih luas. Untuk mendapatkan perbandingan yang lebih komprehensif, penelitian selanjutnya juga dapat mempertimbangkan algoritma lain, seperti XGBoost, LightGBM, atau *Deep Learning*, guna mengevaluasi metode mana yang lebih optimal dalam prediksi penyakit paru-paru. Lebih lanjut, implementasi model dalam sistem berbasis web atau aplikasi berbasis AI dapat dikembangkan agar dapat digunakan secara *real-time* oleh tenaga medis, sehingga meningkatkan efektivitas dalam diagnosis dini penyakit paru-paru. Dengan hasil yang menjanjikan ini, penelitian ini diharapkan dapat menjadi rujukan bagi pengembangan model prediksi berbasis kecerdasan buatan, serta berkontribusi dalam kemajuan teknologi medis untuk mendukung deteksi penyakit secara lebih akurat dan efisien.

DAFTAR PUSTAKA

Gould, G. S., Hurst, J. R., Trofor, A., Alison, J. A., Fox, G., Kulkarni, M. M., Wheelock, C. E., Clarke, M., & Kumar, R. (2023). Recognising the importance of chronic lung disease: a consensus statement from the Global Alliance for Chronic Diseases (Lung Diseases group). *Respiratory Research*, *24*(1), 15.

Heitlinger, E. (2023). Globale Belastung durch Lungenkrankheiten bekämpfen. *Healthbook TIMES Das Schweizer Ärztejournal Journal Des Médecins Suisses*, *7*(5–6), 4–5.

Jasmine Pemeena Priyadarsini, M., Kotecha, K., Rajini, G. K., Hariharan, K., Utkarsh Raj, K., Bhargav Ram, K., Indragandhi, V., Subramaniyaswamy, V., & Pandya, S. (2023). Lung diseases detection using various deep learning algorithms. *Journal of Healthcare Engineering*, *2023*(1), 3563696.

Midyanti, D. M., Bahri, S., & Hidayati, R. (2020). Diagnosis of lung disease using Learning Vector Quantization 3 (LVQ3). *Scientific Journal of Informatics*, *7*(2), 174.

Musa, O. R., & Alang, A. (2017). ANALISIS Penyakit Paru-Paru Menggunakan Algoritma K-Nearest Neighbors Pada Rumah Sakit Aloei Saboe Kota Gorontalo. *ILKOM Jurnal Ilmiah*, *9*(3), 348–352.

Prasetyo, T. M., Amrullah, A., Syahrir, S., & Sari, B. N. (2022). Implementasi Algoritma SVM (Support Vector Machine) Dalam Klasifikasi Penyakit Paru-Paru Berdasarkan Fitur Pola Bentuk. *Jurnal Teknologi Informasi*, *6*(1), 1–6.

Putra, B. S. C., Tahyudin, I., Kusuma, B. A., & Isnaini, K. N. (2024). Efektivitas Algoritma *Random Forest* , XGBoost , dan Logistic Regression dalam Prediksi Penyakit Paru-paru. *Techno.Com*, *23*(4), 909–922.

Sandika, A., Ramadhan, F. R., Iman, I. N., & Jihad, J. (2024). Optimasi Prediksi Penyakit Paru-Paru dan Kanker Paru melalui Integrasi Algoritma *Random Forest* . *BIIKMA : Buletin Ilmiah Ilmu Komputer Dan Multimedia*, *2*(3), 585–591.

SENDY, H. P. (2023). *Evaluasi Kinerja Metode Support Vector Machine (Svm), Naive Bayes Dan Decision Tree Untuk Diagnosa Penyakit Jantung*.

Siregar, A. P., Purba, D. P., Pasaribu, J. P., & Bakara, K. R. (2023). Implementasi Algoritma *Random Forest* Dalam Klasifikasi Diagnosis Penyakit Stroke. *Jurnal Penelitian Rumpun Ilmu Teknik*, *2*(4), 155–164.

Sofyan, F. M. A., Voutama, A., & Umaidah, Y. (2023). Penerapan Algoritma C4. 5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(2), 1409–1415.

Sriyanto, S., & Supriyatna, A. R. (2023). Prediksi Penyakit Diabetes Menggunakan Algoritma *Random Forest* . *TEKNIKA*, *17*(1), 163–172.

Swartzendruber, J. A., Nicholson, B. J., & Murthy, A. K. (2020). The role of connexin 43 in lung disease. *Life*, *10*(12), 1–11. https://doi.org/10.3390/life10120363

Utami, N. W., & Saptiari, N. N. (2020). Penerapan Data Mining Untuk Klasifikasi Penyebab Kematian Menggunakan Algoritma Support Vector Machine. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi| JIITUJ|*, *4*(2), 234–240.

Wahid, M. A. R., Nugroho, A., & Anshor, A. H. (2023). Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier. *Bulletin of Information Technology (BIT)*, *4*(1), 63–74.

Yunianto, M., Anwar, F., Septianingsih, D. N., Ardyanto, T. D., & Pradana, R. F. (2021). Klasifikasi Kanker Paru Paru Menggunakan Naïve Bayes Dengan Variasi Filter Dan Ekstraksi Ciri Gray Level Co-Occurance Matrix (GLCM). *Indonesian Journal of Applied Physics*, *11*(2), 256–268.