Vol. 3 No. 1 – Tahun 2024
Indonesian Journal Computer Science
ISSN: 2829-3819

# Deteksi dan Prediksi Cerdas Penyakit Paru-Paru dengan Algoritma Random Fores

**Deny Kurniawan[1], Mochamad Wahyudi[2], Lise Pujiastuti[3], Sumanto[4]**
Universitas Bina Sarana Informatika[1,2,4], STMIK Antar Bangsa[3]
deny.kurniawan@bsi.ac.id[1], wahyudi@bsi.ac.id[2], lise.pujiastuti@gmail.com[3], sumanto@bsi.ac.id[4]

| Diterima | Direvisi | Disetujui |
|---|---|---|
| (27-03-2024) | (12-04-2024) | (29-04-2024) |

**Abstrak** - Penyakit paru-paru, seperti COPD, kanker paru-paru, dan asma, adalah masalah kesehatan global yang menyebabkan lebih dari tujuh juta kematian setiap tahun. Teknologi canggih, termasuk model deep learning dan algoritma Random Forest, telah digunakan secara efektif untuk mendeteksi dan mengklasifikasikan penyakit paru-paru dari data pencitraan dengan akurasi tinggi. Penelitian ini bertujuan menunjukkan efektivitas algoritma Random Forest dalam memprediksi penyakit paru-paru. Dataset yang digunakan terdiri dari 30.000 data dengan 11 atribut, diperoleh dari Kaggle dan diproses menggunakan perangkat lunak Orange versi 3.36.2. Algoritma Random Forest diimplementasikan dengan 10 pohon keputusan dan enam atribut yang dipertimbangkan pada setiap pembagian data. Model ini diuji menggunakan validasi silang dengan 10 lipatan, dan hasil pengujian menunjukkan nilai AUC sebesar 0,993, yang mengindikasikan tingkat akurasi yang sangat tinggi. Matriks kebingungan digunakan untuk mengevaluasi kinerja model, dengan mengukur akurasi, presisi, recall, F1-Score, dan AUC. Model ini menunjukkan akurasi yang tinggi, dengan nilai ROC AUC 0,453 untuk prediksi adanya penyakit paru-paru dan 0,547 untuk prediksi ketiadaan penyakit paru-paru. Hasil ini menunjukkan bahwa algoritma Random Forest dapat menjadi alat yang efektif dalam mengidentifikasi penyakit paru-paru. Penelitian ini berkontribusi pada pengembangan teknik diagnostik yang lebih akurat dan efisien, yang dapat membantu tenaga medis dalam mendiagnosis penyakit paru-paru pada pasien. Dengan pemahaman yang lebih baik tentang penerapan algoritma ini dalam dunia kesehatan, diharapkan dapat meningkatkan kualitas diagnosis dan perawatan pasien secara signifikan..

Kata Kunci : *Random Forest Algorithm, Lung Disease, Orange Software*

***Abstract*** *- Lung diseases, such as COPD, lung cancer, and asthma, are global health issues that cause more than seven million deaths each year. Advanced technologies, including deep learning models and the Random Forest algorithm, have been effectively used to detect and classify lung diseases from imaging data with high accuracy. This study aims to demonstrate the effectiveness of the Random Forest algorithm in predicting lung diseases. The dataset used consists of 30,000 records with 11 attributes, obtained from Kaggle and processed using Orange software version 3.36.2. The Random Forest algorithm was implemented with 10 decision trees and six attributes considered at each data split. The model was tested using cross-validation with 10 folds, and the testing results showed an AUC value of 0.993, indicating a very high accuracy level. A confusion matrix was used to evaluate the model's performance, measuring accuracy, precision, recall, F1-score, and AUC. The model exhibited 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 demonstrate that the Random Forest algorithm can be an effective tool in identifying lung diseases. This study contributes to the development of more accurate and efficient diagnostic techniques, which can assist healthcare professionals in diagnosing lung diseases in patients. With a better understanding of how this algorithm can be applied in the healthcare field, it is expected to significantly improve the quality of patient diagnosis and care..*

*Keywords: Random Forest Algorithm, Lung Disease, Orange Software*

## I. PENDAHULUAN

Penyakit paru-paru mencakup berbagai gangguan yang mempengaruhi saluran udara dan struktur paru-paru, termasuk penyakit paru obstruktif kronik (PPOK), kanker paru-paru, asma, bronkiektasis, penyakit paru-paru interstitial, penyakit paru-paru akibat kerja, dan hipertensi paru (Gould et al., 2023). Secara global, penyakit paru-paru merupakan masalah kesehatan yang signifikan, dengan lebih dari tujuh juta kematian setiap tahun dikaitkan dengan kondisi seperti PPOK, infeksi saluran pernapasan bagian bawah, dan kanker paru-paru (Heitlinger, 2023). Teknologi canggih

seperti model pembelajaran mendalam 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 (Jasmine Pemeena Priyadarsini et al., 2023). Memahami mekanisme yang mendasari peradangan paru-paru, seperti peran connexin 43 (Cx43) pada penyakit seperti sindrom gangguan pernapasan akut (ARDS), PPOK, dan asma, sangat penting untuk mengembangkan perawatan yang ditargetkan dan strategi pencegahan (Swartzendruber et al., 2020). Di negara-negara seperti Indonesia, jaringan saraf buatan seperti LVQ3 telah digunakan untuk mendiagnosis berbagai jenis penyakit paru-paru dengan tingkat akurasi yang menjanjikan, menyoroti pentingnya pendekatan diagnostik inovatif dalam mengelola kesehatan paru-paru (Midyanti et al., 2020).

Tabel 1. Penelitian Terdahulu

| Peneliti | Data | Metode | Hasil |
|---|---|---|---|
| Umaidah et al | Penyakit Paru-Paru | C4.5 | Akurasi 89.77% |
| Anshor et al | Kanker paru-paru | *Regresi linier* | Akurasi 90% |
| Alang et al | Citra paru-paru | *SVM* | Akurasi 79%.%. |
| Olha Musa, dan Alang | Penyakit Paru-Paru | *K-Nearest Neighbors* | Akurasi 91.90% |
| Pradana et al | citra paru normal dan citra kanker paru | *Naïve Bayes* | Akurasi 88,33 %. |

Sumber: Penelitian (2024)

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%.

Tujuan inti dari penelitian ini adalah untuk membuktikan penggunaan metode *Random Forest* untuk meningkatkan akurasi dan keandalan memprediksi penyakit paru-paru. *Random Forest* diharapkan dapat mengatasi variasi akurasi yang terlihat pada metode lain, seperti *SVM*, dan menangani data yang kompleks seperti citra paru-paru dengan lebih efektif. Kelebihan metode ini termasuk akurasi yang tinggi, kemampuan mencegah overfitting, fleksibilitas dalam menangani berbagai jenis *variabel*, dan efektivitas dalam menangani *missing values*. Selain itu, *Random Forest* dapat memberikan wawasan penting tentang fitur-fitur individual dalam data medis, meningkatkan efisiensi komputasi, dan toleran terhadap data bising. Dengan tujuan dan kelebihan ini, *Random Forest* diharapkan dapat memberikan kontribusi signifikan dalam peningkatan diagnostik penyakit paru-paru dan kanker paru-paru, serta mendukung penelitian medis dan praktik klinis.

## II. METODOLOGI PENELITIAN

### 1. Pengumpulan Dataset

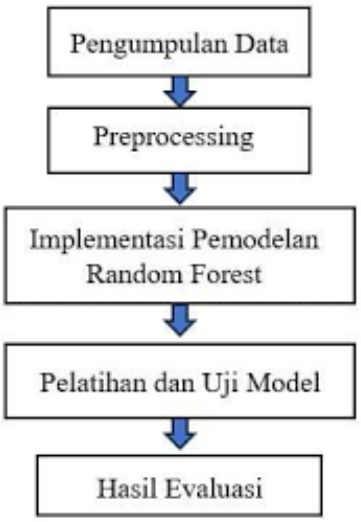
Dataset yang digunakan berasal dari https://www.kaggle.com/ yang bersifat open source. Dataset berjumlah 30000 *record* dan memiliki 11 atribut, yaitu No, Usia, Jenis kelamin, Meroko, Bekerja, Rumah Tangga, Aktivitas Begadang, aktivitas Olahraga, Asuransi, Penyakit Bawaan, dan Hasil (Sriyanto & Supriyatna, 2023).

### 2. Praprosess Dataset

Kegiatan analisis data menggunakan *software orange* versi 3.36.2. Praproses data dilakukan untuk mengolah dataset menjadi bentuk data yang dimengerti oleh *tools orange* (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) (Siregar et al., 2023)



Sumber: Hasil Penelitian (2024)
Gambar 1. Tahapan Penelitian

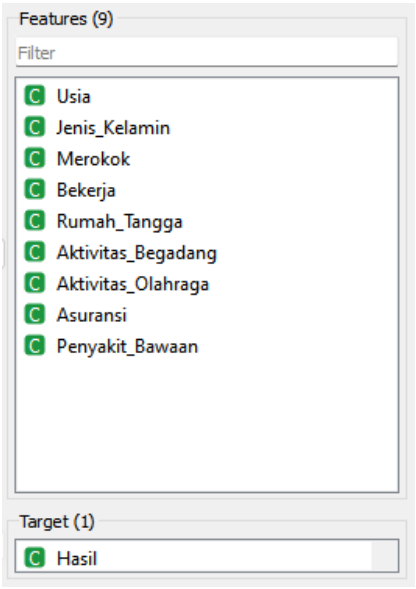
Gambar 1 menggambarkan alur tahapan penelitian dalam penerapan algoritma Random Forest. Tahapan dimulai dengan Pengumpulan Data, di mana data yang relevan untuk penelitian dikumpulkan. Setelah itu, dilakukan Preprocessing, yakni tahap persiapan data agar siap digunakan dalam pemodelan, seperti pembersihan dan transformasi data. Selanjutnya adalah Implementasi Pemodelan Random Forest**,** di mana algoritma Random Forest diterapkan pada dataset untuk membangun model prediksi. Setelah model dibangun, dilakukan Pelatihan dan Uji Model, di mana model dilatih menggunakan data training dan diuji untuk mengevaluasi performanya. Akhirnya, Hasil Evaluasi digunakan untuk menilai kinerja model berdasarkan metrik yang relevan, seperti akurasi, presisi, dan recall.

## III. HASIL DAN PEMBAHASAN

Hasil dan pembahasan penelitian yang dilakukan adalah sebagai berikut.

### 1. Praprosess Dataset

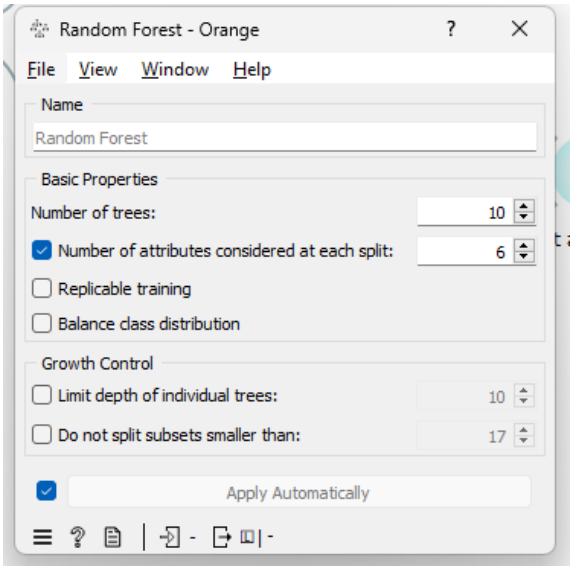
Dalam penelitian ini, dataset telah dianggap baik dan tidak mengandung nilai yang kosong (*missing value*). Tahap pra-pemrosesan data melibatkan pemilihan atribut yang akan digunakan sebagai atribut fitur dan atribut target. Dalam kasus ini, atribut yang diambil sebagai target adalah atribut "Hasil", sementara atribut lainnya akan dijadikan atribut fitur. Pada gambar 3 disajikan rincian dataset yang digunakan.




Sumber: Hasil Penelitian (2024)
Gambar 2**.** Features dan Target

### 2. Implementasi Algoritma *Random Forest*

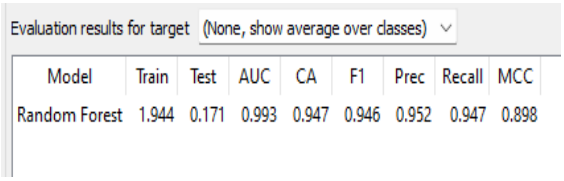
Parameter pengujian yang digunakan dalam penelitian ini meliputi *Number of trees*: 10. *Number of atrributess considered at each split*: 6. Detail dari *parameter* yang digunakan dalam algoritma *Random Forest* dapat dilihat pada Gambar 4.




Sumber: Hasil Penelitian (2024)
**Gambar 3.** Parameter *Random Forest*

### 3. 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.




Sumber: Hasil Penelitian (2024)
Gambar 4. Pengujian Model

### 4. Matrik Konfusi

Matrik konfusi adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi atau prediksi pada suatu set data *testing* yang nilai- nilai sebenarnya sudah diketahui. Matrik konfusi merupakan tabulasi silang antara data kelas *positif* dan kelas *negatif* yang masuk dalam kelas prediksi dan kelas aktual (Sriyanto & Supriyatna, 2023). Matrik konfusi terdiri dari *True positive* (TP), *False Positive* (FP), *False Negative* (FN), dan *True Negative* (TN) [22]. Tabel matrik konfusi disajikan pada Tabel 2.

Tabel 2. Matrik Konfusi

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

Sumber: Hasil Penelitian (2024)

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

a. Akurasi (*Accuracy*)

Akurasi adalah persentase yang menunjukkan seberapa tepat model dapat mengklasifikasikan data secara keseluruhan. Nilai akurasi mengukur proporsi data yang terklasifikasi dengan benar dibagi dengan total keseluruhan data yang ada (Nugroho, 2019 dalam SENDY, 2023). Rumus yang digunakan untuk menghitung tingkat akurasi adalah sebagai berikut.

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

b. Presisi (*Precision*)

*(Positive Predictive Value)* Presisi 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). Rumus untuk menghitung nilai presisi adalah sebagai berikut.

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

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).

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

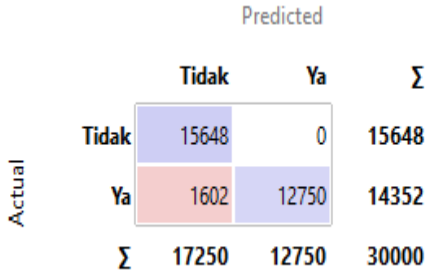
d. F1-Score

*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).

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

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

Merupakan daerah di bawah kurva *Receiver Operating Characteristic* (*ROC*). Nilai *AUC* memiliki rentang antara 0,5 sampai dengan 1. Interpretasi nilai *AUC* dapat diklasifikasikan menjadi lima bagian yang berbeda, yaitu: 0,5 – 0,6 (akurasi salah), 0,6 – 0,7 (tingkat akurasi lemah), 0,7 – 0,8 (tingkat akurasi sedang), 0,8 – 0,9 (tingkat akurasi tinggi), dan 0,9 – 1 (tingkat akurasi sangat tinggi) [(Sriyanto & Supriyatna, 2023)].



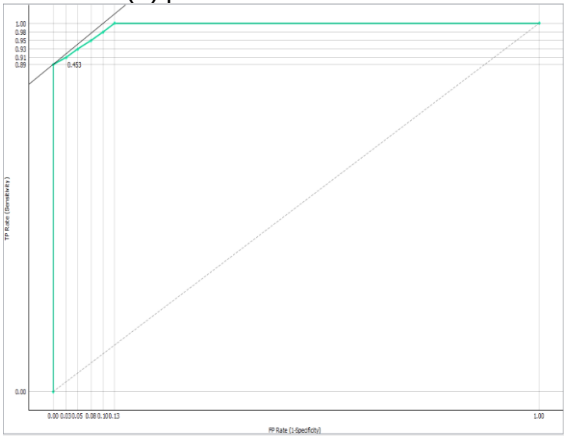

Sumber: Hasil Penelitian (2024)

Gambar 6. Matrik konfusi

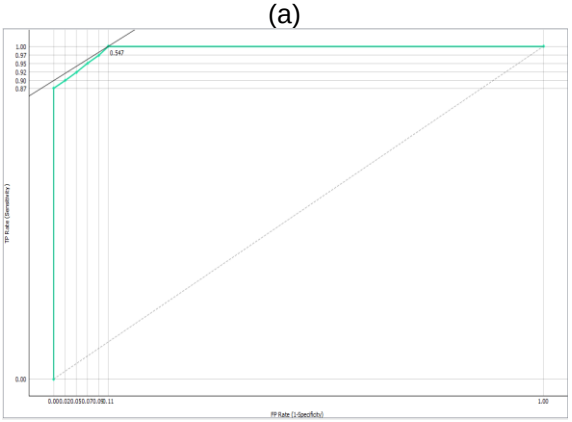
Dari 30000 data latih yang digunakan dengan perulangan sebanyak 100 kali, maka diperoleh tingkat akurasi 12750 untuk Ya dan 15648 Tidak. Untuk hasil dari *confusion matrix* tampak pada gambar 6 diatas.

**5. Hasil Nilai *AUC***

Hasil prediksi penyakit paru - paru dapat dilihat dalam kurva *ROC*. Pada Gambar 7 disajikan Kurva *ROC* Ya (a) pada titik 0.453 dan Tidak (b) pada titik 0.547.



(a)



(b)

Sumber: Hasil Penelitian (2024)

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

Gambar 7.a menampilkan kurva *ROC* yang menggambarkan hubungan antara tingkat *False Positive rate* (sumbu x) dengan *true positive rate* (sumbu y). Nilai *AUC* yang diperoleh adalah 0,453, yang digunakan sebagai acuan untuk memprediksi keberadaan penyakit paru-paru. Dapat disimpulkan bahwa perfoma model yang dihasilkan mendekati titik 0.1 yang artinya hasil klafisikasi memiliki tingkat akurasi yang tinggi.

Gambar 7.b menampilkan kurva *ROC* yang menggambarkan hubungan antara tingkat *False Positive rate* (sumbu x) dengan *true positive rate* (sumbu y). Nilai *AUC* yang diperoleh adalah 0,547, yang digunakan untuk memprediksi ketiadaan penyakit paru-paru. Dapat disimpulkan bahwa perfoma model yang dihasilkan mendekati titik 0.1 yang artinya hasil klasifikasi memiliki tingkat akurasi yang tinggi.

## IV. KESIMPULAN

Hasil implementasi algoritma Random Forest pada penelitian ini menunjukkan performa yang sangat baik dalam memprediksi penyakit paru-paru. Dengan parameter pengujian yang meliputi jumlah pohon keputusan (Number of Trees) sebanyak 10 dan jumlah atribut yang dipertimbangkan di setiap split sebanyak 6, model yang diuji menggunakan tools Orange Data Mining menghasilkan tingkat akurasi yang tinggi. Hasil uji confusion matrix menunjukkan akurasi sebesar 0.947, F1 Score 0.946, Precision 0.952, dan Recall 0.947, yang menjadikan model ini sangat andal untuk digunakan sebagai rujukan dalam pengembangan model prediksi serupa. Kurva performa yang mendekati titik 0.1 lebih lanjut menunjukkan kemampuan model dalam memberikan prediksi yang presisi. Untuk pengembangan di masa depan, disarankan untuk mengeksplorasi jumlah pohon keputusan yang lebih bervariasi serta atribut lain yang relevan guna meningkatkan kemampuan model dalam berbagai skenario data. Selain itu, pengujian model pada dataset yang lebih besar dan lebih kompleks, serta penerapan teknik ensemble atau hybrid, dapat memberikan hasil yang lebih optimal. Penggunaan metode validasi silang yang lebih dalam juga diharapkan dapat meningkatkan akurasi dan generalisasi model di berbagai kondisi dan domain aplikasi lainnya.

## V. REFERENSI

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.

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.