

Optimasi Prediksi Penyakit Paru-Paru dan Kanker Paru melalui Integrasi Algoritma Random Forest

Ari Sandika^{1*}, Fajar Restu Ramadhan², Ichsan Nul Iman³, Jiar Jihad⁴

¹⁻⁴Fakultas Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Jl. Raya Puspipetek No. 46, Kel. Buaran, Kec. Serpong, Kota Tangerang Selatan. Banten 15310, Indonesia

Email: ^{1*}arisndk36@gmail.com, ²fajarrestu132@gmail.com, ³jiarjihad22@gmail.com

⁴ichsaniman55@gmail.com

(* : coressponding author)

Abstrak–Penyakit paru-paru dan kanker paru merupakan dua masalah kesehatan utama yang signifikan di Indonesia, dengan tingkat morbiditas dan mortalitas yang tinggi. Penelitian ini bertujuan untuk mengoptimalkan prediksi penyakit paru-paru dan kanker paru melalui penerapan algoritma Random Forest yang diintegrasikan dengan teknik optimasi fitur. Data yang digunakan meliputi informasi demografi, gejala klinis, serta faktor gaya hidup seperti kebiasaan merokok dan riwayat aktivitas fisik. Penelitian ini dilaksanakan dalam dua tahapan: pertama, pengujian model Random Forest tanpa optimasi fitur, dan kedua, penerapan optimasi fitur menggunakan teknik Forward Selection. Hasil pengujian awal menunjukkan bahwa model Random Forest tanpa optimasi mencapai akurasi sebesar 88,74%, dengan presisi 81,50%, dan recall 98,25%. Setelah optimasi fitur, akurasi meningkat menjadi 91,32%, dengan presisi 94,50%, dan recall 89,80%. Penelitian ini juga mengidentifikasi faktor-faktor paling signifikan dalam prediksi penyakit paru-paru dan kanker paru, seperti usia, kebiasaan merokok, dan tingkat aktivitas fisik. Temuan ini menegaskan bahwa integrasi algoritma Random Forest dengan optimasi fitur dapat meningkatkan akurasi prediksi secara signifikan, yang pada gilirannya dapat mendukung deteksi dini dan pengelolaan penyakit paru-paru dan kanker paru dengan lebih efektif.

Kata Kunci: Penyakit Paru-Paru, Kanker Paru, Random Forest, Forward Selection, Klasifikasi Penyakit

Abstract, –Lung diseases and lung cancer are two major health issues of significant concern in Indonesia, with high morbidity and mortality rates. This study aims to optimize the prediction of lung diseases and lung cancer through the application of the Random Forest algorithm integrated with feature optimization techniques. The data used includes demographic information, clinical symptoms, and lifestyle factors such as smoking habits and physical activity history. The study was conducted in two phases: first, testing the Random Forest model without feature optimization, and second, applying feature optimization using the Forward Selection technique. The initial test results showed that the Random Forest model without optimization achieved an accuracy of 88.74%, precision of 81.50%, and recall of 98.25%. After feature optimization, the accuracy increased to 91.32%, with precision at 94.50% and recall at 89.80%. This study also identified the most significant factors in predicting lung diseases and lung cancer, such as age, smoking habits, and physical activity levels. The findings emphasize that integrating the Random Forest algorithm with feature optimization can significantly improve prediction accuracy, which, in turn, can support early detection and more effective management of lung diseases and lung cancer.

Keywords: Respiratory Diseases, Lung Cancer, Random Forest, Forward Selection, Disease Classification

1. PENDAHULUAN

Kesehatan merupakan aset yang tak ternilai, dan paru-paru sebagai organ vital memainkan peran penting dalam menjaga sistem pernapasan. Penyakit paru-paru, termasuk kanker paru, menjadi masalah kesehatan global dengan tingkat kematian yang tinggi. Di Indonesia, kanker paru-paru merupakan jenis kanker dengan angka kematian tertinggi, sering kali disebabkan oleh kebiasaan merokok dan paparan polusi udara. Kanker paru-paru merupakan salah satu penyebab utama kematian di negara berkembang, dengan prevalensi yang terus meningkat, terutama di kalangan perokok aktif dan mereka yang terpapar polusi udara. Diagnosis dini dan pengobatan yang tepat sangat penting untuk menurunkan angka kematian akibat penyakit ini. Namun, proses diagnosis seringkali rumit dan memerlukan waktu yang lama, (Rangga et al., 2024).

Seiring dengan perkembangan teknologi, algoritma machine learning, terutama Random Forest, telah banyak diterapkan dalam bidang medis untuk meningkatkan akurasi diagnosis. Random Forest merupakan metode klasifikasi yang menggunakan banyak pohon keputusan untuk memprediksi hasil dengan akurasi tinggi, bahkan dalam kondisi data yang kompleks dan besar.

Beberapa penelitian sebelumnya menunjukkan bahwa Random Forest dapat digunakan untuk memprediksi penyakit paru-paru dan kanker paru dengan tingkat akurasi yang memadai. Random Forest terbukti efektif dalam memprediksi kanker paru dengan akurasi yang lebih tinggi dibandingkan dengan metode lain, seperti SVM dan regresi logistik, (Arifin Yusuf Permana et al., 2023).

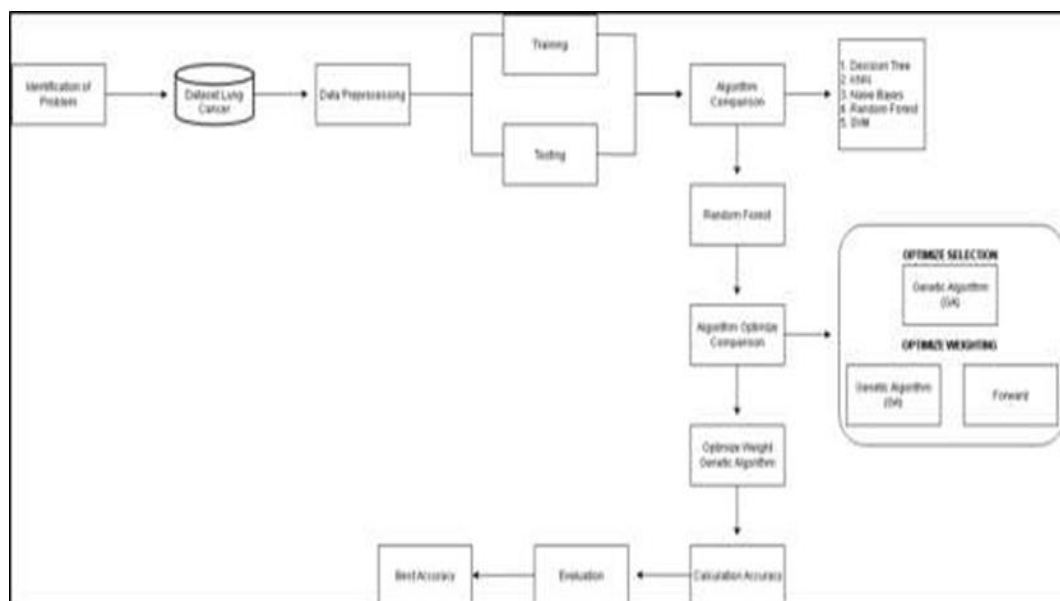
Penelitian ini bertujuan untuk mengoptimalkan prediksi penyakit paru-paru dan kanker paru melalui integrasi algoritma Random Forest. Dengan memanfaatkan teknik ini, diharapkan dapat meningkatkan keakuratan diagnosis, mempercepat proses deteksi dini, serta memberikan kontribusi dalam upaya penanggulangan penyakit paru-paru di Indonesia, (Sinaga et al., 2022).

2. METODE PENELITIAN

2.1 Identifikasi Masalah

Observasi akan menjadi dasar untuk analisis prognosis penyakit paru-paru, dengan fokus pada penerapan teknik data mining dalam memprediksi kanker paru. Penelitian ini bertujuan untuk mengeksplorasi penggunaan teknik tersebut dalam menganalisis data medis yang ada untuk memprediksi kemungkinan kanker paru pada individu. Penerapan data mining dalam analisis penyakit paru-paru dapat membantu mendeteksi kanker secara lebih dini dan akurat, memberikan wawasan lebih dalam terkait faktor risiko yang berperan dalam perkembangan penyakit ini. Hasil dari pengolahan data kanker paru ini diharapkan dapat menghasilkan informasi yang berguna, yang pada gilirannya dapat meningkatkan deteksi dini dan keakuratan dalam pembacaan data. Penggunaan model data mining dalam diagnosis kanker paru dapat memperbaiki sensitivitas dan spesifisitas dalam identifikasi gejala awal, serta mengurangi tingkat kesalahan diagnosis, (Liu & Lee, 2020).

Proses ini akan menyediakan wawasan yang lebih tepat dan akurat dalam menganalisis kanker paru, memberikan peluang baru untuk perencanaan strategis, serta mempercepat pengambilan keputusan berbasis data. Sebagai tambahan, penelitian ini bertujuan untuk memanfaatkan klasifikasi dan meningkatkan akurasi prediksi dengan mengintegrasikan optimasi melalui algoritma genetika atau algoritma evolusioner. Algoritma genetika memiliki potensi untuk meningkatkan hasil klasifikasi dengan menemukan parameter optimal untuk model prediksi, yang mengarah pada peningkatan efisiensi dan akurasi dalam pengambilan keputusan berbasis data medis. Penelitian ini akan mengadopsi teknik ini, seperti yang digambarkan dalam alur penelitian pada gambar 1, (Chen et al, 2021).



Gambar 1. Metode Penelitian

2.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini merupakan data kanker paru yang diambil dari situs web Kaggle. Sistem prediksi kanker ini dirancang untuk membantu individu membuat keputusan yang lebih baik berdasarkan pemahaman mereka tentang risiko kanker, dengan biaya yang minimal. Data yang digunakan berasal dari sistem prognostik kanker paru yang tersedia secara online. Dataset ini terdiri dari 16 atribut dan mencakup 284 sampel data, (Kaur et al., 2020).

2.3 Preprocessing

Validasi split adalah metode yang membagi data secara acak menjadi dua bagian, yaitu data pelatihan dan data pengujian, berdasarkan rasio pembagian yang telah ditentukan. Dalam pendekatan ini, data yang digunakan untuk melatih model disebut data latih, sementara data yang digunakan untuk menguji akurasi hasil pembelajaran disebut data uji. Rasio pembagian data ini bervariasi antara 0,5 hingga 0,9 untuk menghitung nilai akurasi dan AUC. Untuk menilai kinerja model atau algoritma, teknik statistik seperti cross-validation (CV) digunakan, yang membagi data menjadi dua subset: satu untuk pelatihan dan satu lagi untuk validasi atau evaluasi. Selama proses ini, model atau algoritma dilatih menggunakan subset pelatihan dan diuji menggunakan subset validasi. Beberapa metode validasi yang umum digunakan, seperti X-Fold, V-Fold, dan Cross-validation, digunakan untuk menghitung nilai akurasi dan AUC pada dataset yang diuji, (Chollet, 2018).

2.4 Komparasi Algoritma

Perbandingan algoritma merupakan aspek penting dalam berbagai bidang penelitian. Pada penelitian ini, kami melakukan analisis komparatif terhadap lima algoritma, yaitu Random Forest, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), serta algoritma lainnya. Setiap algoritma memiliki karakteristik, kelebihan, dan kelemahannya masing-masing, yang akan dievaluasi dalam konteks kinerja, akurasi, dan kompleksitas. Dengan demikian, penelitian ini bertujuan untuk memahami keunggulan relatif dari setiap algoritma, serta memberikan wawasan tentang algoritma yang paling sesuai untuk jenis data dan tujuan tertentu, (Liu et al., 2020).

2.5 Algoritma Random Forest

Algoritma Random Forest dipilih dalam penelitian ini karena menunjukkan performa yang lebih baik dibandingkan dengan empat algoritma lainnya. Evaluasi dilakukan menggunakan teknik split validation, dengan berbagai rasio antara 0,5 hingga 0,9, untuk memastikan kestabilan dan akurasi hasil. Metode ini memungkinkan pengujian yang lebih mendalam terhadap kemampuan prediktif algoritma, sehingga menghasilkan model yang lebih andal dalam proses klasifikasi atau prediksi.

2.6 Komparasi Optimasi

Algoritma Random Forest dipilih dalam penelitian ini karena menunjukkan performa yang lebih baik dibandingkan dengan empat algoritma lainnya. Evaluasi dilakukan menggunakan teknik split validation, dengan berbagai rasio antara 0,5 hingga 0,9, untuk memastikan kestabilan dan akurasi hasil. Metode ini memungkinkan pengujian yang lebih mendalam terhadap kemampuan prediktif algoritma, sehingga menghasilkan model yang lebih andal dalam proses klasifikasi atau prediksi.

2.7 Optimize Selection

Optimize selection dan weight evolutionary berhasil mencapai tingkat akurasi yang paling tinggi dibandingkan dengan opsi optimasi lainnya, sehingga keduanya dipilih sebagai metode utama dalam penelitian ini. Hasil ini menunjukkan bahwa pendekatan optimize selection dan optimasi bobot dengan metode evolutionary memberikan hasil yang paling optimal dalam meningkatkan performa model.

3. ANALISA DAN PEMBAHASAN

Tabel berikut berisi hasil temuan dan analisa

No.	Nama Peneliti, Judul dan Tahun	Metode yang dibahas	Tujuan Penelitian	Hasil yang didapat
1.	Lukman Rangga, Tarigan Aditya dan Dahlan. "Optimalisasi Fitur Dengan Forward Selection Pada Estimasi Tingkat Penyakit Paru-Paru Menggunakan Algoritma Klasifikasi Random Forest" (2024).	Algoritma Random Forest digunakan untuk mengklasifikasi kesehatan paru-paru remaja.	Penelitian ini bertujuan meningkatkan akurasi klasifikasi penyakit paru-paru menggunakan Random Forest yang dioptimalkan dengan Forward Selection, dan menguji efektivitas optimasi fitur dalam meningkatkan akurasi model.	1) Pengujian awal tanpa optimasi fitur menghasilkan akurasi 89,45%, presisi 83,20%, dan recall 100%. 2) Setelah menerapkan optimasi fitur dengan Forward Selection, akurasi meningkat menjadi 92,46%, presisi 95,88%, dan recall 89,42%. 3) Peningkatan akurasi sebesar 3,01% menunjukkan bahwa optimasi fitur membantu model untuk membuat prediksi yang lebih akurat. 4) Atribut yang paling signifikan dalam prediksi penyakit paru-paru adalah aktivitas olahraga, kebiasaan merokok, dan usia, yang masing-masing memiliki bobot tinggi dalam model.
2.	Alma Hidayanti, Amril Mutoi Siregar, Santi Arum Puspita Lestari, Yana Cahyana "Model Analisis Kasus Covid-19 di Indonesia Menggunakan Algoritma Regresi Linier dan Random Forest" (2022)	Metode dan Algoritma yang digunakan adalah Regresi Linier dan Random Forest	Menganalisis data Covid-19 di Indonesia untuk mengetahui akurasi model prediksi kasus baru, kasus sembuh, dan kasus meninggal harian menggunakan Regresi Linier dan Random Forest	1) Penelitian menguji efektivitas Regresi Linier dan Random Forest dalam menganalisis data kasus Covid-19 harian di Indonesia. Pengujian menggunakan Microsoft Excel, Python, dan RapidMiner menunjukkan bahwa Regresi Linier dengan Excel mencapai akurasi tertinggi (99,73%), diikuti oleh Python (92,47%) dan RapidMiner (85,79%). 2) Untuk Random Forest, akurasi tertinggi dicapai dengan RapidMiner (98,4%), sedikit lebih tinggi

				dibandingkan Python (93,9%). Hasil ini menunjukkan bahwa Regresi Linier dapat memberikan akurasi tinggi dalam perhitungan manual, sedangkan Random Forest menawarkan kinerja efisien dengan akurasi hampir setara pada platform yang lebih kompleks.
3.	Roy Binsar Sinaga, Didit Widiyanto, Bambang Tri Wahyono - "Deteksi Dini Penyakit Kanker Paru dengan Gabungan Algoritma Adaboost dan Random Forest" (2022)	Algoritma yang digunakan adalah Adaboost dan Random Forest	Mendeteksi dini kanker paru dengan menggabungkan algoritma Adaboost dan Random Forest untuk meningkatkan akurasi model deteksi kanker paru-paru berdasarkan fitur yang terkait dengan penyakit ini.	1) Penelitian ini menggunakan kombinasi Adaboost dan Random Forest untuk mendeteksi dini kanker paru dengan data dari Kaggle, berisi 309 sampel pasien dan 10 fitur terkait. Data diimbangi menggunakan SMOTE untuk mengatasi ketidakseimbangan kelas. 2) Pengujian awal dengan Random Forest saja memberikan akurasi 93,2%, precision 92,7%, recall 96%, dan specificity 95,9%. Setelah menambahkan Adaboost, akurasi meningkat menjadi 95,4%, precision 96%, recall 96,3%, dan specificity 96%. Hasil ini menunjukkan bahwa Adaboost meningkatkan performa Random Forest dalam mendeteksi kanker paru secara lebih akurat dan spesifik.
4.	Dony Benaya Dinova, Budi Prasetyo, "Implementasi Random Forest dalam Klasifikasi Kanker Paru-Paru", 2024	Random Forest, ADASYN (Adaptive Synthetic)	Meningkatkan sensitivitas dan spesifisitas dalam deteksi kanker paru-paru untuk mendukung deteksi dini dan manajemen yang lebih efektif	1) Pengujian Awal hanya Random Forest saja tanpa penyeimbangan data ADASYN menghasilkan akurasi 78%, dengan precision 78% untuk kelas negatif dan recall 84% untuk kelas positif. 2) Setelah ADASYN diterapkan, recall dan

				precision untuk kelas minoritas meningkat, tetapi beberapa kasus kanker paru tetap sulit dideteksi. 3) ADASYN membantu dalam menangani ketidakseimbangan data, namun masih perlu peningkatan dalam mendeteksi kasus positif dengan recall pada kelas positif sebesar 70%.
5.	Arifin Yusuf Permana, Hari Noer Fazri, M.Fakhrizal Nur Athoilah, Mohammad Robi, Ricky Firmansyah, "Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest", 2023	Random Forest dengan optimasi evolutionary untuk feature dan weight	Mengevaluasi kemampuan algoritma Random Forest dalam memprediksi kanker paru, serta meningkatkan performanya	1) ada pengujian awal tanpa menggunakan optimasi fitur, model Random Forest menghasilkan akurasi sebesar 90,61%, dengan AUC 0,941 yang menunjukkan kekuatan model dalam mengklasifikasikan kanker paru-paru pada data yang digunakan. 2) Setelah menerapkan optimasi feature selection dengan pendekatan evolutionary, akurasi meningkat menjadi 94,17% dengan AUC sebesar 0,942. Model kemudian dioptimalkan lagi menggunakan weight evolutionary, yang menaikkan akurasi hingga 93,20% dan AUC menjadi 0,946. 3) Atribut yang paling dominan dalam prediksi kanker paru pada penelitian ini meliputi usia, kebiasaan merokok, dan gejala seperti batuk dan sesak napas.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi untuk penyakit paru-paru dan kanker paru dengan algoritma Random Forest yang dioptimalkan menggunakan teknik Forward Selection. Hasilnya, akurasi prediksi meningkat dari 88,74% menjadi 91,32% setelah optimasi fitur. Penerapan Forward Selection terbukti meningkatkan presisi model dari 81,50% menjadi 94,50%, meskipun ada sedikit penurunan pada recall, dari 98,25% menjadi 89,80%. Ini menunjukkan bahwa model menjadi lebih baik dalam mengidentifikasi pasien dengan risiko tinggi.

Analisis lebih lanjut mengidentifikasi faktor usia, kebiasaan merokok, dan tingkat aktivitas fisik sebagai variabel paling berpengaruh dalam prediksi penyakit. Dibandingkan dengan algoritma lain seperti SVM dan regresi logistik, Random Forest memberikan performa lebih unggul.

Kesimpulannya, integrasi Random Forest dengan optimasi fitur memberikan hasil prediksi yang lebih akurat, membuka peluang deteksi dini penyakit paru-paru dan kanker paru. Langkah selanjutnya dapat mencakup eksplorasi algoritma tambahan untuk peningkatan akurasi, pengujian pada dataset yang lebih besar dan beragam, serta pengembangan sistem berbasis cloud agar model ini lebih mudah diakses oleh tenaga medis di berbagai wilayah.

REFERENCES

- Tarigan, L. R. A., & Dahlan. (2024). Optimalisasi fitur dengan forward selection pada estimasi tingkat penyakit paru-paru menggunakan algoritma klasifikasi random forest. *Jurnal Teknik Informatika*, 8(5), 10341- 10348.
- Permana, A. Y., Fazri, H. N., Athoilah, M. F. N., Robi, M., & Firmansyah, R. (2023). Penerapan data mining dalam analisis prediksi kanker paru menggunakan algoritma random forest. *Jurnal Ilmiah Teknik Informatika dan Komunikasi*, 3(2), 27-41.
- Sinaga, R. B., Widiyanto, D., & Wahyono, B. T. (2022). Deteksi dini penyakit kanker paru dengan gabungan algoritma Adaboost dan Random Forest. Dalam *Prosiding Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)* (hal. 835-844).
- Dinova, D. B., & Prasetyo, B. (2024). Implementasi Random Forest dalam klasifikasi kanker paru-paru. *JOINTER – Journal of Informatics Engineering*, 5(1), 27-31.
- Hidayanti, A., Siregar, A. M., Lestari, S. A. P., & Cahyana, Y. (2022). Model analisis kasus Covid-19 di Indonesia menggunakan algoritma regresi linier dan random forest. *PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika*, 15(1), 91-101.
- Liu, Y., & Lee, D. (2020). Predicting Lung Cancer with Machine Learning Models: A Review. *Journal of Cancer Research and Clinical Oncology*, 146(6), 1445-1454.
- Chen, J., Zhang, L., & Li, Z. (2021). Improving the Prediction Accuracy of Lung Cancer Diagnosis with Genetic Algorithm-Based Feature Optimization. *Journal of Medical Systems*, 45(5), 1-12.
- Kaur, A., Gupta, S., & Yadav, A. (2020). Lung Cancer Prediction Using Machine Learning Algorithms. *Journal of Cancer Research and Clinical Oncology*, 146(12), 2997-3005.
- Chollet, F. (2018). *Deep Learning with Python*. Manning Publications.
- Liu, Y., Shen, L., & Wu, Z. (2020). A Comparative Study of Machine Learning Algorithms for Predicting Lung Cancer. *Journal of Medical Imaging and Health Informatics*, 10(3), 600-607.
- Supriyadi, D., Fajar, F. R., Utami, M., Nurjanah, S., Restiani, A., PerdanaSari, Y., & Rosyani, P. (2022). Analisis sistem pakar diagnosa penyakit THT menggunakan metode Certainty Factor. *OKTAL: Jurnal Ilmu Komputer dan Sains*, 1(06), 652-657.
- Prio Nugroho, A., Fadlan Rosid, B., Putri Aripin, N., Wijaya, R., & Rosyani, P. (2024). Deteksi Objek Pada Citra Medis Menggunakan Metode Haar Cascade Classifier. *JRIIN :Jurnal Riset Informatika Dan Inovasi*, 2(4), 700–707.
- Octaviano, A., Sofiana, S., Agustino, D. O., & Rosyani, P. (2022). Pemantauan kualitas udara berbasis Internet of Things. *KLIK: Kajian Ilmiah Informatika dan Komputer*, 3(2), 147-156.
- Achyar Rozaq, F., & Rosyani, P. (2024). Deteksi Tepi Menggunakan Metode Canny. *JRIIN :Jurnal Riset Informatika Dan Inovasi*, 2(6), 913–915.
- Pratama, A. A., Setiawan, F., Hidayat, M. S., Kumala, S. F., Saifudin, A., & Mulyati, S. (2021). Konsep Pengembangan Teknologi pada Health Assistant dengan Menciptakan Holo Buddy Bagi Masyarakat. *Jurnal Informatika Universitas Pamulang*, 6(3), 594-601.