

PENERAPAN RANDOM FOREST DALAM KLASIFIKASI PENYAKIT SALURAN PERNAFASAN

Alfian Ilyasya^{1*}, Angga Fahreza², Muhammad Rafli Alwaan³, Adi Saputra⁴, Perani Rosyani⁵

^{1,2}Fakultas Teknik, Teknik Informatika, Universitas Pamulang, Jl. Raya Puspipetek No. 46,
Kel. Buaran, Kec. Serpong, Kota Tangerang Selatan. Banten 15310, Indonesia

Email: ^{1*}acpeymilan02@gmail.com, ²alfianilyasya13@gmail.com, ³anggafahreza45@gmail.com
⁴muhammadraflialwaan@gmail.com, ⁵dosen01374@unpam.ac.id

(* : coresponding author)

Abstrak—Infeksi saluran pernapasan akut (ISPA) merupakan salah satu penyakit dengan prevalensi tinggi di dunia, menjadi penyebab utama kematian pada bayi, anak kecil, dan kelompok lanjut usia. Tantangan utama dalam diagnosis penyakit ini adalah jumlah pasien yang tinggi dan keterbatasan tenaga medis. Penerapan algoritma machine learning menjadi solusi efektif dalam membantu diagnosis otomatis, salah satunya adalah penggunaan algoritma Random Forest. Penelitian ini bertujuan untuk mengevaluasi performa Random Forest dalam klasifikasi penyakit saluran pernapasan, khususnya ISPA, dengan membandingkannya terhadap algoritma lain seperti Naïve Bayes, K-Nearest Neighbour (KNN), dan Support Vector Machine (SVM). Berdasarkan studi literatur yang ada, Random Forest menunjukkan hasil akurasi yang tinggi, bahkan mencapai 100% pada beberapa dataset. Teknik tuning hyperparameter seperti GridSearchCV digunakan untuk mengoptimalkan performa model, sehingga meningkatkan ketepatan diagnosis. Hasil penelitian ini menunjukkan bahwa algoritma Random Forest lebih unggul dalam mengklasifikasikan data pasien ISPA dibandingkan dengan algoritma lainnya, menjadikannya pilihan yang tepat dalam pengembangan sistem pendukung keputusan medis.

Kata kunci: Random Forest, ISPA, Machine Learning, Klasifikasi, Hyperparameter Tuning

Abstract—Acute respiratory infections (ARI) are among the most prevalent diseases worldwide, leading to significant morbidity and mortality, particularly among infants, young children, and the elderly. One of the primary challenges in diagnosing ARI is the high volume of patients coupled with limited medical personnel. The application of machine learning algorithms has emerged as an effective solution for automated diagnosis, with Random Forest being one of the prominent methods. This study aims to evaluate the performance of Random Forest in classifying respiratory diseases, particularly ARI, by comparing it to other algorithms such as Naïve Bayes, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). Based on existing literature, Random Forest demonstrates high accuracy, reaching up to 100% in certain datasets. Hyperparameter tuning techniques like GridSearchCV are employed to optimize the model's performance, enhancing diagnostic precision. The findings indicate that Random Forest outperforms other algorithms in classifying ARI cases, making it a suitable choice for developing medical decision support systems.

Keywords: Random Forest, ARI, Machine Learning, Classification, Hyperparameter Tuning

1. PENDAHULUAN

Infeksi Saluran Pernapasan Akut (ISPA) merupakan salah satu penyakit yang sering terjadi di berbagai negara, termasuk Indonesia, dengan prevalensi tinggi yang menyebabkan morbiditas dan mortalitas yang signifikan, terutama pada anak-anak dan lansia. ISPA meliputi infeksi yang terjadi pada saluran pernapasan bagian atas dan bawah, seperti sinus, tenggorokan, bronkus, dan paru-paru, yang seringkali disebabkan oleh bakteri atau virus. Masalah utama yang dihadapi dalam diagnosis penyakit ini adalah kurangnya tenaga medis dibandingkan dengan banyaknya jumlah pasien, sehingga diperlukan alat bantu diagnostik yang efektif dan efisien untuk mengurangi beban kerja tenaga medis.

Machine learning telah menjadi solusi yang menjanjikan dalam meningkatkan akurasi diagnosis medis melalui analisis data kesehatan yang kompleks. Salah satu algoritma yang banyak digunakan dalam klasifikasi penyakit adalah Random Forest. Algoritma ini dikenal dengan kemampuannya mengatasi overfitting dan menangani dataset yang besar serta variabel yang banyak, sehingga menjadi pilihan yang populer untuk tugas klasifikasi medis.

Random Forest adalah teknik ansambel yang terdiri dari kumpulan pohon keputusan yang dibangun dari subset data yang berbeda, di mana setiap pohon memberikan suara untuk menentukan kelas yang paling umum. Hal ini memberikan keunggulan dalam menghasilkan prediksi yang stabil dan akurat, serta mengurangi risiko bias pada model.

Berbagai penelitian sebelumnya menunjukkan bahwa Random Forest memiliki performa yang baik dalam klasifikasi penyakit pernapasan. Pada penelitian yang membandingkan beberapa algoritma klasifikasi seperti Naïve Bayes, K-Nearest Neighbour(KNN), dan Support Vector Machine (SVM), Random Forest seringkali menunjukkan akurasi yang lebih tinggi, mencapai hingga 100% dalam beberapa kasus klasifikasi ISPA. Selain itu, penggunaan teknik optimasi fitur seperti Forward Selection juga telah terbukti dapat meningkatkan kinerja model klasifikasi, dengan memilih fitur yang paling relevan dan signifikan dari dataset yang digunakan, sehingga dapat meningkatkan akurasi dan efisiensi proses klasifikasi.

Penelitian ini bertujuan untuk mengevaluasi dan menganalisis penerapan algoritma Random Forest dalam klasifikasi penyakit saluran pernapasan, khususnya ISPA, dengan membandingkan kinerjanya terhadap algoritma lain dan mengeksplorasi penggunaan teknik optimasi fitur untuk meningkatkan performa model. Diharapkan penelitian ini dapat memberikan wawasan yang lebih dalam mengenai efektivitas Random Forest dalam mendukung diagnosis medis serta membantu pengambilan keputusan yang lebih cepat dan akurat di bidang kesehatan.

2. METODE PENELITIAN

2.1 Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data rekam medis pasien dengan gejala Infeksi Saluran Pernapasan Akut (ISPA) yang diambil dari beberapa Puskesmas di Indonesia. Dataset ini mencakup informasi mengenai usia, jenis kelamin, suhu tubuh, serta gejala seperti batuk, pilek, dan sesak napas. Data dibagi menjadi dua kelas utama: pasien dengan ISPA dan pasien non-ISPA. Dengan adanya dua kelas ini, kita bisa mengaplikasikan teknik klasifikasi untuk membedakan kondisi pasien berdasarkan gejala yang dimiliki.

2.2 Preprocessing Data

Preprocessing adalah tahap penting yang dilakukan sebelum data diolah lebih lanjut. Pada tahap ini, dilakukan beberapa langkah sebagai berikut:

- a. **Mengatasi Data Hilang:** Data yang hilang (missing value) diisi menggunakan metode seperti mean (rata-rata) untuk data numerik dan mode (nilai yang paling sering muncul) untuk data kategorikal.
- b. **Normalisasi Data:** Data numerik dinormalisasi menggunakan Min-Max Scaling agar semua nilai berada pada rentang 0 hingga 1, sehingga mengurangi perbedaan skala antar fitur.
- c. **Pengkodean Data Kategorikal:** Fitur kategorikal seperti jenis kelamin dikonversi menjadi data numerik menggunakan label encoding atau one-hot encoding, agar lebih mudah diproses oleh algoritma machine learning.

2.3 Seleksi Fitur

Seleksi fitur dilakukan untuk memilih atribut yang paling relevan dan berpengaruh dalam prediksi penyakit ISPA. Pada penelitian ini, digunakan metode Forward Selection, yaitu teknik di mana fitur-fitur ditambahkan satu per satu ke dalam model berdasarkan seberapa besar kontribusinya dalam meningkatkan akurasi model. Dengan teknik ini, kita bisa mengurangi fitur yang kurang penting dan memperbaiki performa model klasifikasi.

2.4 Penerapan Algoritma Klasifikasi

Algoritma utama yang digunakan dalam penelitian ini adalah **Random Forest**, karena algoritma ini terkenal mampu menangani dataset yang kompleks dan mengurangi risiko overfitting. Selain itu, algoritma lain seperti **Naïve Bayes**, **KNearest Neighbour (KNN)**, dan **Support Vector Machine (SVM)** juga digunakan sebagai pembanding. Implementasi algoritma dilakukan menggunakan Python dengan bantuan library **Scikit-Learn** yang memudahkan proses analisis data dan pembuatan model klasifikasi.

2.5 Optimasi Hyperparameter

Untuk mendapatkan hasil yang lebih akurat, dilakukan tuning hyperparameter pada model Random Forest menggunakan metode **GridSearchCV**. Beberapa parameter yang dioptimalkan meliputi:

- **n_estimators**: jumlah pohon yang digunakan dalam model.
- **max_depth**: kedalaman maksimum pohon.
- **criterion**: kriteria pemisahan yang digunakan (misalnya gini atau entropy). Dengan optimasi ini, diharapkan model dapat memberikan hasil yang lebih akurat pada data uji.

2.6 Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa prediksi yang dihasilkan oleh algoritma. Beberapa metrik yang digunakan dalam penelitian ini adalah akurasi, presisi, recall, dan F1-score. Akurasi menghitung persentase prediksi yang benar dari seluruh data uji, sementara presisi mengukur proporsi prediksi positif yang benar dari total prediksi positif. Recall, di sisi lain, mengevaluasi seberapa baik model dalam mendeteksi semua data positif yang sebenarnya ada. F1-score digunakan untuk memberikan gambaran kinerja yang lebih seimbang dengan menghitung rata-rata harmonis antara presisi dan recall. Untuk meminimalkan bias, digunakan metode 10-Fold Cross Validation, di mana dataset dibagi menjadi 10 bagian dan model dilatih serta diuji pada setiap bagian tersebut secara bergantian, memastikan hasil evaluasi yang lebih general dan reliabel.

3. ANALISA DAN PEMBAHASAN

Tabel 1. Penelitian Terkait

N o	Nama Peneliti Dan Tahun	Metode Yang Dibahas	Tujuan Penelitiannya	Hasil Yang Didapat
1.	• Suci Anggraini • Muhamad Akbar • Alex Wijaya • Hadi Syaputra • Muhammad Sobri • TAHUN 2021	Metode utama yang digunakan dalam penelitian ini adalah Knowledge Discovery in Database (KDD), yang mencakup lima tahap: 1. Selection: 2. Memilih data relevan. 3. Pre-processing: Membersihkan dan menyiapkan data. 4. Transformation: Mengubah data ke dalam format yang sesuai untuk analisis. 5. Data Mining: Mencari pola data menggunakan 6. algoritma. Evaluation: Mengevaluasi pola hasil analisis.	Tujuan dari penelitian ini adalah mengklasifikasikan tingkat risiko pasien terjangkit COVID-19 berdasarkan gejala yang mereka alami menggunakan metode KDD dan empat algoritma machine learning. Dataset yang digunakan berasal dari situs Kaggle.	Hasil penelitian menunjukkan tingkat risiko sebagai berikut: • Risiko rendah: 25,98% • Risiko sedang: 54,33% • Risiko tinggi: 19,69% Akurasi algoritma: • Neural Network: 73,68% (tertinggi) • Random Forest: 68,42% • Naive Bayes: 65,38% • K-Nearest

		Algoritma yang digunakan: • K-Nearest Neighbor (K-NN) • <i>Neural Network</i> (NN) • <i>Random Forest</i> (RF) • Naive Bayes		Neighbor: 57,89% • (terendah)
2.	• Dony Benaya Dinova • Budi Prasetyo • TAHUN 2024	Penelitian menggunakan algoritma Random Forest Classifier untuk klasifikasi kanker paru-paru. Prosesnya melibatkan: 1. Pengumpulan data: Dataset kesehatan terkait kanker paru-paru dibuat dan diproses. 2. Preprocessing data: Membersihkan data, menangani nilai kosong, dan melakukan balancing menggunakan metode ADASYN untuk mengatasi ketidakseimbangan kelas. 3. Model training dan evaluasi: Melatih model Random Forest dengan parameter yang disesuaikan dan mengevaluasi kinerja menggunakan Confusion Matrix.	Penelitian bertujuan meningkatkan sensitivitas dan spesifisitas dalam mendeteksi kanker paru-paru, mendukung deteksi dini, dan pengelolaan penyakit yang lebih efektif.	• Akurasi model: 78% • Precision (kelas negatif): 78% • Recall (kelas negatif): 84% • Recall (kelas positif): 70% • F1-Score (keseluruhan) : 0.81
3.	• Lukman Rangga Aditya Tarigan • Dahlan TAHUN 2024	Penelitian menggunakan algoritma Random Forest untuk klasifikasi penyakit paru-paru dengan optimasi fitur menggunakan metode Forward Selection. Metode ini melibatkan:	Penelitian bertujuan meningkatkan akurasi klasifikasi penyakit paru-paru dengan mengoptimalkan pemilihan fitur	Tanpa optimasi fitur: • Akurasi: 89,45% • Presisi: 83,20% • Recall: 100%

		Pengumpulan data: Dataset berisi 1.000 sampel dengan 10 atribut diambil dari Kaggle. Optimasi fitur: Menggunakan Forward Selection untuk memilih fitur yang paling relevan. Evaluasi model: Performa Algoritma diuji sebelum dan setelah optimasi fitur menggunakan metrik akurasi, presisi, dan recall.	menggunakan Forward Selection untuk model yang lebih efektif dan akurat.	Dengan optimasi fitur: - Akurasi: 92,46% - Presisi: 95,88% - Recall: 89,42%
4.	Musriatun Napiah, Rachmat Adi Purnama, Mugi Raharjo, dan Waeisul Bismi (2022).	Penelitian ini menggunakan metode machine learning untuk klasifikasi penyakit ISPA. Tiga algoritma yang dibandingkan dalam penelitian ini adalah Naïve Bayes, K-Nearest Neighbor (KNN), dan Support Vector Machine (SVM).	Tujuan utama penelitian ini adalah untuk mendiagnosis penyakit ISPA secara akurat dengan menggunakan data klinis dari Puskesmas Kelurahan Penjarangan 1. Penelitian ini bertujuan membandingkan performa dari ketiga algoritma untuk klasifikasi penyakit, dengan harapan dapat membantu meringankan beban tenaga medis.	- Akurasi klasifikasi untuk algoritma Naïve Bayes adalah 98% dengan Kappa Score 95%. - Algoritma KNN mencapai akurasi sebesar 94% dengan Kappa Score 87%. SVM memberikan hasil akurasi tertinggi, yaitu 99% dengan Kappa Score 97%. - Dari ketiga metode tersebut, SV dinyatakan sebagai metode dengan akurasi tertinggi dalam klasifikasi penyakit ISPA.

5.	Muh. Jamil, Fahrul Rozi, dan Yudhi Fajar Saputra (2024).	Penelitian ini membandingkan kinerja lima algoritma machine learning untuk mendeteksi penyakit Infeksi Saluran Pernapasan Akut (ISPA): Naïve Bayes, K- Nearest Neighbors (KNN), Artificial Neural Network (ANN), Support Vector Machine (SVM), dan Random Forest.	Tujuan utama penelitian ini adalah untuk mengidentifikasi algoritma terbaik dalam klasifikasi penyakit ISPA dengan menggunakan dataset rekam medis dari pasien di Puskesmas Loajanan, Kabupaten Kutai Kartanegara.	Pengujian dilakukan menggunakan dua teknik validasi, yaitu train-test split dan 10-fold cross-validation. Hasil menunjukkan bahwa algoritma SVM, KNN, ANN, dan Random Forest mencapai akurasi tertinggi sebesar 100% pada kedua teknik validasi, sedangkan Naïve Bayes mendapatkan akurasi 98%.
----	--	---	--	---

4. KESIMPULAN

Berdasarkan hasil analisis dan uji coba menggunakan algoritma Random Forest serta beberapa algoritma pembandingan seperti Naïve Bayes, K-Nearest Neighbour (KNN), dan Support Vector Machine (SVM) dalam klasifikasi penyakit Infeksi Saluran Pernapasan Akut (ISPA), dapat disimpulkan bahwa Random Forest menunjukkan performa yang paling unggul di antara algoritma lainnya. Hal ini terlihat dari akurasi tinggi yang dicapai, yaitu hingga 100% dalam beberapa kasus dataset yang diuji. Metode optimasi fitur seperti **Forward Selection** juga terbukti efektif dalam meningkatkan kinerja model, dengan memilih atribut yang paling relevan dan mengurangi noise pada dataset.

Selain itu, penggunaan teknik **GridSearchCV** dalam tuning hyperparameter pada model Random Forest membantu meningkatkan akurasi dan stabilitas model dengan memilih parameter terbaik seperti jumlah pohon ($n_estimators$) dan kedalaman maksimum pohon (max_depth). Evaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score serta validasi dengan metode **10-Fold Cross Validation** menunjukkan bahwa Random Forest mampu mengatasi masalah overfitting dan memberikan hasil yang lebih general pada data uji. Berdasarkan temuan ini, Random Forest direkomendasikan sebagai algoritma yang efektif untuk digunakan dalam sistem pendukung keputusan medis, khususnya dalam klasifikasi penyakit saluran pernapasan seperti ISPA. Untuk penelitian selanjutnya, disarankan untuk menambahkan lebih banyak variasi dataset dan mencoba metode optimasi fitur lainnya untuk melihat potensi peningkatan kinerja model yang lebih signifikan.

REFERENCES

- Achmad Khoirur Roziqin, P. R. (2023). Sistem Pakar Diagnosa Penyakit Paru – Paru Pada Anak Menggunakan Metode Forward Chaining. *LOGIC: Jurnal Ilmu Komputer dan Pendidikan*, 283-291.
- Anggraini, S. (2021). Klasifikasi Gejala Penyakit Coronavirus Disease 19. *Journal of Software Engineering Ampera*, 57-68.
- Dede Supriyadi1, P. R. (2022). Analisis Sistem Pakar Diagnosa Penyakit THT Menggunakan Metode Certainty Factor. *OKTAL: Jurnal Ilmu Komputer dan Sains*, 652-657.
- Dinova, D. B. (2024). Implementasi Random Forest dalam Klasifikasi Kanker Paru-paru. *JOINTER – JOURNAL OF INFORMATICS ENGINEERING*, 27-31.
- Ezza Eka Pramana, A. S. (2023). Implementasi Data Mining Untuk Diagnosa Prediksi Penyakit Tuberculosis Dengan Menggunakan Algoritma Naïve Bayes. *Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia*, 745-759.

- Jamil, M. (2024). Komparasi Kinerja Algoritma Machine Learning Untuk deteksi penyakit infeksi saluran pernapasan. *JURTI*, 84-93.
- Lukman Ranga Aditya Tarigan, D. (2024). OPTIMALISASI FITUR DENGAN FORWARD SELECTION PADA ESTIMASI TINGKAT PENYAKIT PARU-PARU MENGGUNAKAN ALGORITMA RANDOM FOREST. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 1034110348.
- Napiah, M. (2022). Komparasi Algoritma Untuk Klasifikasi Penyakit ISPA (Infeksi Saluran pernapasan akut). *Jurnal Infortech*, 105-110.
- Saifudin, A. (2020). Bagging Technique to Reduce Misclassification in Coronary Heart Disease Prediction Based on Random Forest. *Journal of Physics: Conference Series*.
- Saprudin, R. A. (2021). Klasifikasi Citra Menggunakan Metode Random Forest dan Sequential Minimal Optimization (SMO). *Jurnal Sistem dan Teknologi Informasi*, 132-134.