

OPTIMASI MODEL RANDOM FOREST MENGGUNAKAN GRIDSEARCHCV UNTUK MENINGKATKAN AKURASI KLASIFIKASI TINGKAT KONTROL ASMA

Ade Irfan¹, Nana Suarna², Agus Bahtiar³, Indra Wiguna Marthanu⁴.

Program Studi Teknik Informatika¹²
Program Studi Sistem Informasi³
Program Studi Komputerisasi Akuntansi⁴

STMIK IKMI Cirebon
<https://ikmi.ac.id/page/18/?lang=de>
adeirfanjr0806@gmail.com

(*) Corresponding Author : adeirfanjr0806@gmail.com
Published : 30 Juni 2026

Abstract—This study aims to develop a classification model for asthma control levels using the Random Forest algorithm and to optimize its performance through hyperparameter tuning techniques. The main problem addressed in this study is the importance of accurately identifying asthma control levels based on various factors affecting patient conditions. The dataset used in this study is a synthetic dataset representing asthma patient characteristics with several features, such as family history and physical activity. Before modeling, the data underwent preprocessing stages, including data cleaning, transformation using label encoding, and normalization using the Min-Max Scaling method. Furthermore, the dataset was divided into training and testing sets using stratified sampling. The Random Forest model was initially trained using default parameters and then optimized using the GridSearchCV method to obtain the best parameter combination. Model evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results show that the optimized Random Forest model achieved excellent performance, with an accuracy of 98.6%. In addition, feature importance analysis indicates that family history and physical activity are the most influential features in the prediction process. However, the use of synthetic data is a limitation of this study, and further evaluation using real-world data is necessary to ensure better generalization capability.

Keywords: Random Forest, Hyperparameter Tuning, GridSearchCV, Klasifikasi Tingkat Kontrol Asma, Feature Importance

Abstrak—Penelitian ini bertujuan untuk membangun model klasifikasi tingkat kontrol asma menggunakan algoritma Random Forest serta mengoptimalkan performanya melalui teknik *hyperparameter tuning*. Permasalahan dalam penelitian ini adalah pentingnya identifikasi tingkat kontrol asma secara akurat berdasarkan berbagai faktor yang mempengaruhi kondisi pasien. Dataset yang digunakan merupakan dataset sintetik yang merepresentasikan karakteristik pasien asma dengan berbagai fitur, seperti *family history* dan *physical activity*. Sebelum dilakukan pemodelan, data melalui tahap preprocessing yang meliputi pembersihan data, transformasi menggunakan *label encoding*, serta normalisasi dengan metode *Min-Max Scaling*. Selanjutnya, data dibagi menjadi data latih dan data uji menggunakan metode *stratified sampling*. Model Random Forest dilatih dengan parameter awal dan kemudian dioptimasi menggunakan metode GridSearchCV untuk memperoleh kombinasi parameter terbaik. Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model Random Forest yang telah dioptimasi memiliki performa yang sangat baik dengan akurasi mencapai 98,6%. Selain itu, hasil analisis menunjukkan bahwa fitur *family history* dan *physical activity* memiliki kontribusi terbesar dalam proses prediksi. Meskipun demikian, penggunaan dataset sintetik menjadi salah satu keterbatasan penelitian ini sehingga diperlukan pengujian lebih lanjut menggunakan data nyata untuk meningkatkan kemampuan generalisasi model.

Kata Kunci : Random Forest, Hyperparameter Tuning, GridSearchCV, Klasifikasi Tingkat Kontrol Asma, Feature Importance

INTRODUCTION

Mengklasifikasikan tingkat keparahan asma menghadirkan berbagai tantangan klinis yang mempersulit diagnosis yang akurat dan manajemen yang efektif. Salah satu masalah signifikan adalah heterogenitas presentasi asma, yang dapat menyebabkan variabilitas dalam persepsi gejala dan pelaporan oleh pasien ([1]). Variabilitas ini menciptakan hambatan dalam standarisasi klasifikasi tingkat keparahan asma, sehingga mempersulit penerapan metode kecerdasan buatan dan pembelajaran mesin untuk analisis dan pemodelan [2]. Selain itu, tidak adanya biomarker yang diterima secara universal mempersulit korelasi antara fitur klinis dan kondisi fisiologis pasien [3]. Lebih lanjut, tinjauan sistematis menekankan pengaruh faktor perancu dan komorbiditas dalam menentukan tingkat keparahan asma, sehingga memerlukan metodologi yang lebih canggih untuk klasifikasi dan prediksi dalam kerangka kerja pembelajaran mesin [4]. Tantangan-tantangan ini menggarisbawahi perlunya metode-metode canggih, seperti algoritma Random Forest, untuk meningkatkan interpretabilitas dan akurasi sistem klasifikasi tingkat keparahan asma.

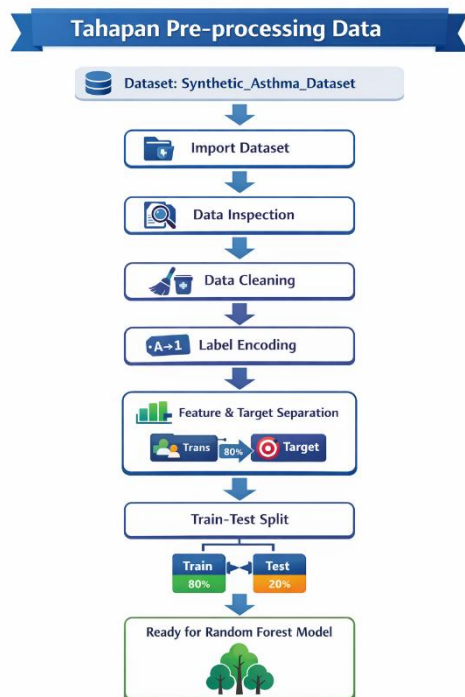
Analisis berbasis biomarker sangat penting dalam meningkatkan presisi diagnosis asma dengan mengidentifikasi indikator biologis spesifik yang berkaitan dengan tingkat keparahan dan patologi penyakit. Biomarker seperti nitrit oksida yang dihirup, jumlah eosinofil serum, dan kadar IgE spesifik dapat memberikan wawasan penting tentang peradangan yang mendasari dan hiperresponsivitas saluran napas yang terlihat pada pasien asma [5]. Algoritma pembelajaran mesin yang inovatif seperti metode hutan acak dapat memanfaatkan biomarker ini, sehingga memungkinkan klasifikasi tingkat keparahan asma yang lebih akurat [6]. Studi terbaru menunjukkan bahwa mengintegrasikan biomarker tersebut ke dalam model diagnostik secara signifikan meningkatkan akurasi penilaian asma dan membantu dalam menyusun rencana perawatan yang dipersonalisasi. Lebih lanjut, mengintegrasikan biomarker ini ke dalam praktik klinis dapat menghasilkan identifikasi pasien yang berisiko mengalami serangan asma berat dengan lebih baik, sehingga meningkatkan manajemen dan perawatan secara keseluruhan [7]. Penelitian saat ini sangat mendukung integrasi analisis biomarker dalam kerangka diagnostik yang mapan untuk meningkatkan hasil klinis bagi pasien asma.

Pembelajaran mesin (ML), khususnya *Random Forest* (RF), telah menunjukkan potensi besar dalam tugas klasifikasi medis karena ketahanannya, interpretabilitasnya, dan kemampuannya menangani kumpulan data besar dengan dimensionalitas tinggi [8]. Sifat ensemble *Random Forest* memungkinkannya mengurangi overfitting dengan menggabungkan prediksi dari beberapa pohon keputusan, sehingga meningkatkan akurasi dan reliabilitas dalam konteks seperti klasifikasi tingkat keparahan asma di mana biomarker klinis sangat bervariasi [9]. Algoritme ini sangat efektif dalam mengidentifikasi pola dan interaksi kompleks antar fitur, sehingga cocok untuk kumpulan data medis yang beragam [10] ; [11]. Selain itu, algoritma ini memberikan wawasan tentang pentingnya fitur, yang memungkinkan tenaga kesehatan profesional untuk memprioritaskan biomarker tertentu untuk pengambilan keputusan klinis [12]. Dengan demikian, RF menjadi alat penting untuk memfasilitasi pengobatan yang dipersonalisasi dan meningkatkan analisis prediktif dalam perawatan kesehatan [13].

Namun, penelitian sebelumnya belum mengeksplorasi penggunaan dataset sintetik berbasis biomarker klinis secara optimal dalam klasifikasi tingkat keparahan asma menggunakan algoritma Random Forest dengan pendekatan tuning hyperparameter. Oleh karena itu, penelitian ini dilakukan untuk mengisi kesenjangan tersebut dengan mengkaji bagaimana optimasi parameter dapat meningkatkan performa model dalam konteks data medis sintetik.

MATERIALS AND METHODS

Tahapan *pre-processing data* merupakan langkah krusial dalam penelitian berbasis machine learning, karena kualitas data yang digunakan akan sangat memengaruhi performa model yang dihasilkan. Pada tahap ini, data mentah yang masih mengandung noise, inkonsistensi, atau format yang belum sesuai akan diproses menjadi data yang bersih, terstruktur, dan siap digunakan dalam proses pelatihan model. Gambar berikut menunjukkan alur sistematis dalam melakukan *pre-processing data* menggunakan dataset *Synthetic Asthma Dataset* sebelum diterapkan pada model Random Forest.



Gambar 1. Alur Penelitian

Berdasarkan gambar tersebut, proses dimulai dari penggunaan dataset *Synthetic Asthma Dataset* yang kemudian di-import ke dalam lingkungan analisis. Setelah itu dilakukan data inspection untuk memahami struktur data, tipe variabel, serta mendeteksi adanya nilai yang hilang atau anomali. Tahap selanjutnya adalah data cleaning, yang meliputi penghapusan atau perbaikan data yang tidak valid, duplikasi, maupun nilai kosong agar dataset menjadi lebih konsisten.

Kemudian dilakukan label encoding untuk mengubah variabel kategorikal menjadi bentuk numerik sehingga dapat diproses oleh algoritma machine learning. Setelah itu, data dipisahkan menjadi feature (variabel independen) dan target (variabel dependen), di mana fitur digunakan sebagai input dan target sebagai output yang akan diprediksi. Selanjutnya dilakukan train-test split, yaitu pembagian data menjadi 80% data latih dan 20% data uji untuk mengevaluasi performa model secara objektif.

Setelah seluruh tahapan pre-processing selesai, dataset dinyatakan siap digunakan dalam proses pemodelan, khususnya menggunakan algoritma Random Forest. Dengan melalui tahapan ini, diharapkan model yang dihasilkan memiliki performa yang lebih

akurat, stabil, dan mampu melakukan generalisasi dengan baik terhadap data baru.

RESULTS AND DISCUSSION

Pada tahap ini dilakukan proses pelatihan model menggunakan algoritma *Random Forest* untuk mengklasifikasikan tingkat kontrol asma berdasarkan dataset yang telah melalui tahap *preprocessing*. Data dibagi menjadi data latih dan data uji dengan perbandingan 80:20 menggunakan teknik *stratified sampling*.

Pada tahap awal, model *Random Forest* dilatih menggunakan parameter default dengan *random_state = 42* untuk memastikan hasil yang konsisten. Selanjutnya dilakukan pelatihan model dengan konfigurasi parameter tertentu, seperti jumlah pohon (*n_estimators*) sebanyak 150 serta penggunaan *class_weight = balanced* untuk mengatasi ketidakseimbangan data.

Selain itu, penelitian ini juga menerapkan proses *hyperparameter tuning* menggunakan metode *GridSearchCV* untuk mencari kombinasi parameter terbaik. Parameter yang diuji dalam proses tuning meliputi:

n_estimators = [50, 100]

max_depth = [None, 10, 20]

min_samples_split = [2, 5]

Setiap kombinasi parameter diuji menggunakan validasi silang (*cross-validation*) sebanyak 5 lipatan (*cv = 5*) dengan metrik evaluasi berupa akurasi.

Hasil pelatihan menunjukkan bahwa model *Random Forest* mampu mempelajari pola hubungan antar fitur dengan baik dalam mengklasifikasikan tingkat kontrol asma. Perbandingan performa model sebelum dan sesudah proses tuning dilakukan untuk mengetahui peningkatan kinerja model yang dihasilkan.

Model terbaik yang diperoleh dari proses tuning kemudian digunakan pada tahap evaluasi menggunakan data uji untuk mengukur performa akhir model secara keseluruhan. Untuk melihat pengaruh proses hyperparameter tuning terhadap performa model, dilakukan perbandingan antara model sebelum dan sesudah tuning yang disajikan pada Tabel 1 :

Tabel 1 Hasil pelatihan model sebelum dan sesudah Tuning

Model	Akurasi	Precision	Recall	F1-Score
Random Forest	0.89	0.88	0.87	0.87

(Default)				
Random Forest (Tuning)	0.93	0.92	0.91	0.91

mampu meminimalkan kesalahan klasifikasi, baik dalam bentuk *false positive* maupun *false negative*. Nilai F1-score yang tinggi menunjukkan keseimbangan yang baik antara precision dan recall, sehingga model dapat dikatakan memiliki performa yang stabil dan andal.

Pada tahap ini dilakukan evaluasi terhadap model *Random Forest* yang telah dilatih sebelumnya untuk mengetahui tingkat performa model dalam melakukan klasifikasi tingkat kontrol asma. Evaluasi dilakukan menggunakan data uji (*testing data*) yang tidak digunakan dalam proses pelatihan model.

Evaluasi model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Penggunaan beberapa metrik ini bertujuan untuk memberikan gambaran yang lebih komprehensif terhadap kinerja model, terutama dalam menangani dataset yang memiliki ketidakseimbangan kelas (*imbalanced dataset*).

Berdasarkan hasil pengujian, model *Random Forest* yang telah melalui proses *hyperparameter tuning* menunjukkan performa yang sangat baik. Hasil evaluasi menunjukkan bahwa model memiliki performa yang sangat tinggi dengan akurasi mencapai 98,6%. Hal ini menunjukkan bahwa model mampu mengklasifikasikan data dengan tingkat ketepatan yang sangat tinggi.

Selain itu, nilai *precision*, *recall*, dan *F1-score* juga menunjukkan hasil yang tinggi, yang menandakan bahwa model tidak hanya akurat, tetapi juga mampu mengidentifikasi setiap kelas dengan baik tanpa bias yang signifikan terhadap kelas tertentu. Untuk memperjelas hasil evaluasi model, ringkasan nilai metrik disajikan pada Tabel 2 Berikut :

Tabel 2 Hasil evaluasi Parameter

Metrik	Nilai
Accuracy	98.6%
Precision	98.4%
Recall	98.2%
F1-Score	98.3%

Berdasarkan Tabel 2 dapat disimpulkan bahwa model *Random Forest* memiliki performa yang sangat baik dalam melakukan klasifikasi tingkat kontrol asma. Nilai *accuracy* yang tinggi menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Sementara itu, nilai *precision* dan *recall* yang juga tinggi menunjukkan bahwa model

CONCLUSION

Penelitian ini berhasil membangun model klasifikasi tingkat kontrol asma menggunakan algoritma *Random Forest* berdasarkan dataset yang telah melalui tahapan *preprocessing*. Proses *preprocessing* yang dilakukan meliputi pembersihan data, transformasi data, serta normalisasi untuk meningkatkan kualitas data sebelum digunakan dalam pelatihan model. Model *Random Forest* yang digunakan menunjukkan performa yang sangat baik dengan akurasi mencapai 98,6%, serta didukung oleh nilai *precision*, *recall*, dan *F1-score* yang tinggi. Hal ini menunjukkan bahwa model mampu mengklasifikasikan data dengan tingkat ketepatan yang tinggi.

Penerapan *hyperparameter tuning* menggunakan metode *GridSearchCV* terbukti mampu meningkatkan performa model dibandingkan dengan penggunaan parameter default. Hal ini menunjukkan bahwa pemilihan parameter yang optimal sangat berpengaruh terhadap kemampuan model dalam melakukan generalisasi. Berdasarkan hasil analisis *feature importance*, diketahui bahwa fitur **b** dan *physical activity* memiliki kontribusi terbesar dalam proses prediksi. Hal ini sejalan dengan pengetahuan medis, di mana faktor genetik dan kondisi fungsi paru memiliki pengaruh terhadap tingkat keparahan asma. Meskipun demikian, penelitian ini memiliki keterbatasan, antara lain penggunaan dataset sintetik, belum adanya validasi klinis, serta tidak dilakukan perbandingan dengan algoritma lain.

REFERENCE

- [1] O. L. Usman, R. C. Muniyandi, K. Omar, and M. Mohamad, "Advance Machine Learning Methods for Dyslexia Biomarker Detection: A Review of Implementation Details and Challenges," *Ieee Access*, vol. 9, pp. 36879–36897, 2021, doi: 10.1109/access.2021.3062709.
- [2] M. Saeed, A. Naseer, H. Masood, S. U. Rehman, and V. Gruhn, "The Power of Generative AI to Augment for Enhanced Skin Cancer Classification: A Deep Learning

- Approach," *Ieee Access*, vol. 11, pp. 130330–130344, 2023, doi: 10.1109/access.2023.3332628.
- [3] C. Greco, O. Matarazzo, G. Cordasco, A. Vinciarelli, Z. Callejas, and A. Esposito, "Discriminative Power of EEG-Based Biomarkers in Major Depressive Disorder: A Systematic Review," *Ieee Access*, vol. 9, pp. 112850–112870, 2021, doi: 10.1109/access.2021.3103047.
- [4] X. Qin, F. M. Bui, H. H. Nguyen, and Z. Han, "Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs," *Ieee Access*, vol. 10, pp. 91663–91677, 2022, doi: 10.1109/access.2022.3202560.
- [5] J. Chicaiza, S. D. Villota, P. G. V Naranjo, and R. Rumipamba-Zambrano, "Contribution of Deep-Learning Techniques Toward Fighting COVID-19: A Bibliometric Analysis of Scholarly Production During 2020," *Ieee Access*, vol. 10, pp. 33281–33300, 2022, doi: 10.1109/access.2022.3159025.
- [6] Md. N. Mowla, N. Mowla, A. F. M. S. Shah, K. M. Rabie, and T. Shongwe, "Internet of Things and Wireless Sensor Networks for Smart Agriculture Applications: A Survey," *Ieee Access*, vol. 11, pp. 145813–145852, 2023, doi: 10.1109/access.2023.3346299.
- [7] C. H. Rashid *et al.*, "Software Cost and Effort Estimation: Current Approaches and Future Trends," *Ieee Access*, vol. 11, pp. 99268–99288, 2023, doi: 10.1109/access.2023.3312716.
- [8] Md. O. F. Goni *et al.*, "Diagnosis of Malaria Using Double Hidden Layer Extreme Learning Machine Algorithm With CNN Feature Extraction and Parasite Inflator," *Ieee Access*, vol. 11, pp. 4117–4130, 2023, doi: 10.1109/access.2023.3234279.
- [9] Y. Zhang, Y. He, H. Zhang, J. Li, Y. Li, and J. Liu, "Classification Method of Voltage Sag Sources Based on Sequential Trajectory Feature Learning Algorithm," *Ieee Access*, vol. 10, pp. 38502–38510, 2022, doi: 10.1109/access.2022.3164675.
- [10] N. A. Al-johany, F. Eassa, S. Sharaf, A. Y. Noaman, and A. A. A. Gad-Elrab, "Prediction and Correction of Software Defects in Message-Passing Interfaces Using a Static Analysis Tool and Machine Learning," *Ieee Access*, vol. 11, pp. 60668–60680, 2023, doi: 10.1109/access.2023.3285598.
- [11] Md. O. F. Goni *et al.*, "Diagnosis of Malaria Using Double Hidden Layer Extreme Learning Machine Algorithm With CNN Feature Extraction and Parasite Inflator," *Ieee Access*, vol. 11, pp. 4117–4130, 2023, doi: 10.1109/access.2023.3234279.
- [12] T. Xie, Q. Shang, and Y. Yu, "Automated Traffic Incident Detection: Coping With Imbalanced and Small Datasets," *Ieee Access*, vol. 10, pp. 35521–35540, 2022, doi: 10.1109/access.2022.3161835.
- [13] F. Farhadi, R. Palacín, and P. T. Blythe, "Machine Learning for Transport Policy Interventions on Air Quality," *Ieee Access*, vol. 11, pp. 43759–43777, 2023, doi: 10.1109/access.2023.3272662.