

Perbandingan Algoritma Decision Tree dan Random Forest Dalam Pengklasifikasian Penyakit Tiroid

Muhammad Rizky Ardiansyah Putra¹, Rissa Nurfitriana Handayani²

^{1,2}Program Studi Sistem Informasi, Universitas Adhirajasa Reswara Sanjaya
e-mail: ¹muhammadrizky.ardiansyahputra@gmail.com, ²rissa@ars.ac.id

Abstrak

Penyakit tiroid telah menjadi masalah kesehatan yang semakin meningkat di Indonesia. Salah satu penyebabnya adalah rendahnya kesadaran masyarakat terhadap gejala dan kelainan tiroid, sehingga banyak kasus yang tidak terdiagnosa dan tidak mendapatkan perawatan yang tepat. Terdapat dua jenis tiroid yaitu hipertiroid dan hipotiroid, di antaranya beberapa jenis yang bersifat ganas. Penyakit tiroid ganas dapat mengganggu fungsi tubuh dan bahkan menyebabkan kematian. Dalam dunia kesehatan, melakukan klasifikasi dan prediksi sangat diperlukan, terutama dengan menggunakan teknik *data mining* untuk menganalisis data dalam jumlah besar yang tersimpan pada suatu *database*. *Machine Learning* merupakan bidang ilmu yang berfokus pada pembangunan sistem yang dapat belajar dan meningkatkan performanya berdasarkan data. *Decision Tree* dan *Random Forest* merupakan dua algoritma yang digunakan untuk memprediksi pengobatan penyakit tiroid. Hasil penelitian menunjukkan bahwa algoritma *Random Forest* memiliki akurasi yang lebih tinggi, yaitu 94,81%, dibandingkan dengan *Decision Tree* yang memiliki akurasi 93,51%. Selain itu, nilai AUC (*Area Under Curve*) dari *Random Forest* mencapai 0,983, sedangkan *Decision Tree* hanya 0,922. Kedua nilai ini menunjukkan kategori klasifikasi yang sangat baik. Penelitian ini menyimpulkan bahwa *Random Forest* lebih baik digunakan dalam pengklasifikasian penyakit tiroid karena performanya yang lebih tinggi dan akurasi yang lebih baik.

Kata kunci—Penyakit Tiroid, *Machine Learning*, *Decision Tree*, *Random Forest*

Abstract

Thyroid disease has become an increasingly significant health issue in Indonesia. One of the main causes is the low awareness among the public regarding the symptoms and abnormalities of the thyroid, resulting in many cases going undiagnosed and untreated. There are two types of thyroid conditions: hyperthyroidism and hypothyroidism, with some forms being malignant. Malignant thyroid diseases can disrupt bodily functions and even lead to death. In the healthcare sector, classification and prediction are crucial, especially through data mining techniques to analyze large amounts of data stored in databases. Machine Learning is a field focused on developing systems that can learn and improve their performance based on data. Decision Tree and Random Forest are two algorithms used to predict thyroid disease treatment. The research findings indicate that the Random Forest algorithm has a higher accuracy of 94.81% compared to the Decision Tree algorithm, which has an accuracy of 93.51%. Additionally, the Area Under Curve (AUC) value for Random Forest reaches 0.983, while Decision Tree only achieves 0.922. Both values indicate a very good classification category. This study concludes that Random Forest is more effective for classifying thyroid diseases due to its superior performance and accuracy.

Keywords—*Thyroid Disease, Machine Learning, Decision Tree, Random Forest*

Corresponding Author:

Rissa Nurfitriana Handayani

Email: rissa@ars.ac.id

1. PENDAHULUAN

Penyakit tiroid merupakan salah satu gangguan endokrin yang umum terjadi di seluruh dunia, termasuk di Indonesia, di mana tingkat kesadaran masyarakat terhadap gejala dan dampak penyakit ini masih tergolong rendah [1]. Penyakit tiroid dapat dikategorikan menjadi dua jenis utama, yaitu hipertiroidisme, dimana kelenjar tiroid memproduksi hormon secara berlebihan, dan hipotiroidisme, dimana kelenjar tiroid tidak memproduksi hormon dalam jumlah yang cukup [2]. Meskipun sebagian besar kasus tiroid bersifat jinak, sekitar 10% di antaranya dapat berkembang menjadi kanker tiroid yang bersifat ganas [3]. Penyakit tiroid yang ganas dapat menyebabkan komplikasi serius dan, dalam kasus ekstrem, kematian.

Dalam konteks kesehatan modern, pentingnya diagnosis dini dan tepat sangat ditekankan untuk mencegah perkembangan penyakit yang lebih lanjut. Data dari Kementerian Kesehatan Indonesia menunjukkan peningkatan signifikan jumlah penderita gangguan tiroid dalam beberapa tahun terakhir, dengan lebih dari 17 juta kasus tercatat pada tahun 2017 [1]. Salah satu tantangan utama dalam menangani penyakit ini adalah banyaknya kasus yang tidak terdiagnosis akibat gejala yang sering tidak disadari oleh pasien.

Perkembangan teknologi informasi, terutama dalam bidang *machine learning* dan *data mining*, telah membuka peluang baru dalam mendukung diagnosis dan pengelolaan penyakit, termasuk penyakit tiroid. *Machine learning* menawarkan kemampuan untuk memproses data dalam jumlah besar dan menemukan pola-pola yang tidak mudah dikenali oleh manusia [4]. Dalam hal ini, algoritma *Decision Tree* dan *Random Forest* telah banyak digunakan untuk keperluan klasifikasi dan prediksi dalam bidang kesehatan [5].

Algoritma *Decision Tree* bekerja dengan memecah dataset menjadi beberapa subset berdasarkan atribut yang paling signifikan, menghasilkan pohon keputusan yang dapat digunakan untuk memprediksi klasifikasi suatu dataset [6]. Di sisi lain, *Random Forest* adalah algoritma *ensemble* yang menggabungkan beberapa pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil [7]. Penelitian sebelumnya telah menunjukkan bahwa *Random Forest* sering kali mengungguli *Decision Tree* dalam hal akurasi, terutama dalam kasus klasifikasi medis yang melibatkan data *kompleks* dan *variabel* yang banyak [1].

Penelitian ini bertujuan untuk membandingkan kinerja kedua algoritma tersebut dalam klasifikasi penyakit tiroid. Berdasarkan *dataset* yang diambil dari *Kaggle*, penelitian ini mengevaluasi akurasi masing-masing algoritma dalam memprediksi keberhasilan diagnosis penyakit tiroid. Dengan semakin pentingnya diagnosa berbasis teknologi, hasil dari penelitian ini diharapkan dapat memberikan wawasan baru mengenai metode mana yang lebih efektif digunakan dalam pengklasifikasian penyakit tiroid, serta memberikan kontribusi dalam pengembangan sistem diagnosa otomatis berbasis *machine learning*.

Selain itu, pemilihan algoritma yang tepat tidak hanya berdampak pada akurasi, tetapi juga pada efisiensi dalam pengolahan data, terutama ketika data yang digunakan sangat besar dan kompleks. Oleh karena itu, penelitian ini akan mengevaluasi tidak hanya dari sisi akurasi, tetapi juga dari sisi keandalan model dalam mengelola dataset yang besar.

2. METODE PENELITIAN

Penelitian ini dilakukan menggunakan *dataset "Thyroid Disease"* dari *kaggle.com*, dengan total 17 atribut dan 383 data. Beberapa atribut yang digunakan antara lain usia, jenis kelamin, riwayat merokok, dan status fungsi tiroid. Pada penelitian ini dilakukan proses *split data* yang dibagi menjadi 80% untuk pelatihan (*data training*) dan 20% untuk pengujian (*data testing*).

2.1. Algoritma yang digunakan

1. *Decision Tree*: Algoritma ini membentuk struktur pohon dimana setiap simpul (*node*) pohon merepresentasikan atribut yang telah diuji. Setiap cabang merupakan pembagian hasil uji dan *node* daun (*leaf*) merepresentasikan kelompok kelas tertentu. Level *node*

teratas dari sebuah decision tree merupakan akar (*root*) berupa atribut yang memiliki pengaruh terbesar pada suatu kelas tertentu [8].

2. *Random Forest* Algoritma dalam *machine learning* yang digunakan untuk pengklasifikasian *dataset* dalam jumlah besar. Karena fungsinya bisa dipakai dengan banyak dimensi dan dengan berbagai skala dan performa yang tinggi. Cara kerja algoritma *random forest* dengan membuat beberapa pohon keputusan dan menyatukannya untuk memperoleh prediksi yang lebih stabil serta akurat yang biasanya dilatih 10 dengan metode bagging yang diharapkan menghasilkan *model machine learning* yang lebih baik [9].

3.3. Evaluasi Hasil

Hasil klasifikasi dari kedua algoritma dievaluasi menggunakan *confusion matrix* untuk menghitung akurasi, presisi, dan *recall*. *Confusion matrix* merupakan teknik analisa yang efektif dalam pengukuran kinerja sistem klasifikasi menggunakan bentuk tabel yang akan diklasifikasikan dengan nilai benar dan salah, sehingga mempermudah proses evaluasi hasil akurasi sistem klasifikasi dan dapat digunakan untuk menentukan model terbaik dengan mengukur kinerja dari model dan membandingkan dengan satu sama lain [10]. Selain itu dalam mengevaluasi akurasi dari klasifikasi secara visual menggunakan kurva ROC (*Receiver Operating Characteristic*) dan AUC (*Area Under Curve*). Evaluasi menggunakan kurva ROC dan AUC dapat memberikan informasi yang berguna tentang kinerja *model*, terutama dalam kasus di mana kelas-kelas yang diamati tidak seimbang [11] yang menunjukkan performa klasifikasi secara keseluruhan.

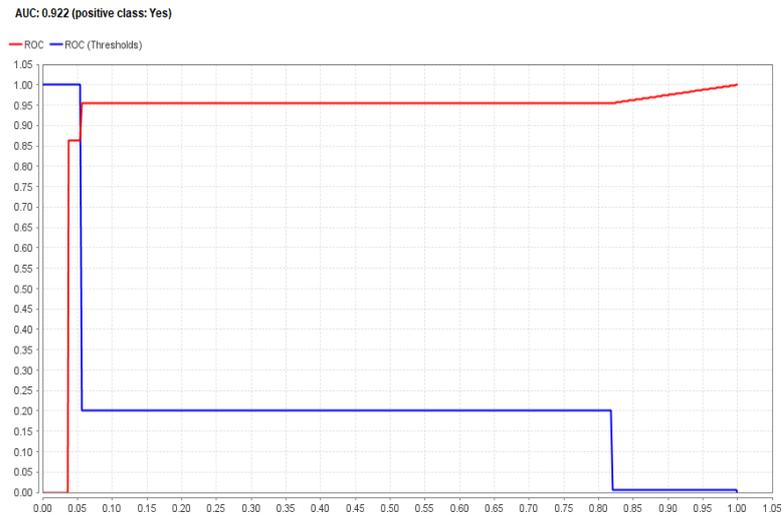
3. HASIL DAN PEMBAHASAN

Tahapan ini merupakan proses pengujian menggunakan software *RapidMiner* terhadap algoritma *Decison Tree* dan *Random Forest* menggunakan 77 *data testing* hasil dari *split data* yang dilakukan dari total 383 data yang tersedia pada *kaggle.com* dengan nama *dataset* “*Thyroid Disease*” dengan *recurred* menjadi label data.

3.1 Hasil Eksperimen Algoritma Decision Tree

Gambar 1. Hasil Akurasi *Decision Tree*

Pada gambar 1 menunjukkan hasil *accuracy* pengujian menggunakan algoritma *decision tree* memperoleh hasil 93.51% dengan *class precision* prediksi *yes* 90.48% sedangkan untuk prediksi *no* memperoleh hasil sebesar 94.64% dan *class recall true yes* memperoleh hasil 86.36% sedangkan untuk *true no* memperoleh hasil sebesar 96.36%



Gambar 2. AUC Algoritma *Decision Tree*

Pada gambar 2 merupakan kurva ROC algoritma *decision tree* menunjukkan nilai AUC (Area Under Curve) sebesar 0.922 yang memiliki hasil diagnose *excellent classification*.

3.2 Hasil Eksperimen Algoritma *Random Forest*

Gambar 3. Hasil Akurasi *Random Forest*

Pada gambar 3 menunjukkan hasil *accuracy* pengujian menggunakan algoritma *random forest* memperoleh hasil 94.81% dengan *class precision* prediksi *yes* 90.91% sedangkan untuk prediksi *no* memperoleh hasil sebesar 96.36% dan *class recall true yes* memperoleh hasil 90.91% sedangkan untuk *true no* memperoleh hasil sebesar 96.36%

Gambar 4. AUC Algoritma *Random Forest*

Pada gambar 4 merupakan kurva ROC algoritma *decision tree* menunjukkan nilai AUC (*Area Under Curve*) sebesar 0.983 yang memiliki hasil diagnose *excellent classification*.

Hasil pengujian menunjukkan bahwa algoritma *Random Forest* memiliki akurasi dan AUC yang lebih tinggi dibandingkan *Decision Tree*.

3.3 Hasil Akurasi

Algoritma *Random Forest* memiliki akurasi sebesar 94.81% dan nilai AUC sebesar 0.983, yang lebih unggul dibandingkan *Decision Tree* dengan akurasi 93.51% dan AUC 0.922. Hal ini menunjukkan bahwa *Random Forest* lebih baik dalam menangani variabilitas data dan memiliki kemampuan prediksi yang lebih akurat seperti yang ditunjukkan pada gambar 5.

Decision Tree			
Accuracy: 93.51%			AUC = 0.922
	True No	True Yes	Class Precision
Pred. No	53	3	94.64%
Pred. Yess	2	19	90.48%
Class Recall	96.36%	86.36%	
Random Forest			
Accuracy: 94.81%			AUC = 0.983
	True No	True Yes	Class Precision
Pred. No	53	2	96.36%
Pred. Yess	2	20	90.91%
Class Recall	96.36%	90.91%	

Gambar 5. Perbandingan Hasil Penelitian

Keunggulan *Random Forest* terletak pada kemampuannya melakukan ensemble learning dengan menggabungkan beberapa pohon keputusan untuk mendapatkan prediksi yang lebih stabil. Sementara itu, *Decision Tree* cenderung lebih rawan terhadap overfitting pada dataset dengan variabilitas yang tinggi.

4. KESIMPULAN

Penelitian ini berhasil membandingkan kinerja algoritma *Decision Tree* dan *Random Forest* dalam pengklasifikasian penyakit tiroid menggunakan dataset "*Thyroid Disease*" yang tersedia di *kaggle.com*. Berdasarkan hasil yang diperoleh, algoritma *Random Forest* menunjukkan performa yang lebih unggul dibandingkan dengan *Decision Tree*, baik dalam hal akurasi maupun nilai AUC. *Random Forest* mencapai akurasi sebesar 94.81% dengan AUC sebesar 0.983, sedangkan *Decision Tree* hanya menghasilkan akurasi 93.51% dengan AUC 0.922. Temuan ini menegaskan bahwa *Random Forest* lebih efektif dalam mendeteksi kasus positif dan mengurangi jumlah prediksi salah positif, yang sangat penting dalam konteks diagnosis penyakit.

Keunggulan *Random Forest* dapat diatributkan pada kemampuannya dalam melakukan ensemble learning, di mana beberapa pohon keputusan digabungkan untuk meningkatkan stabilitas dan akurasi prediksi. Selain itu, *Random Forest* juga lebih mampu menangani variabilitas dalam data dan meminimalkan risiko overfitting dibandingkan dengan *Decision Tree*. Hasil penelitian ini memberikan wawasan penting bagi praktisi kesehatan dan peneliti dalam memilih algoritma yang tepat untuk pengklasifikasian penyakit, khususnya dalam kasus-kasus di mana akurasi diagnosis sangat krusial.

Mengingat tingginya prevalensi penyakit tiroid dan dampak yang signifikan terhadap kesehatan masyarakat, penggunaan algoritma machine learning seperti *Random Forest* dalam diagnosis penyakit tiroid dapat memberikan kontribusi yang berarti terhadap deteksi dini dan pengobatan yang lebih efektif. Penelitian ini juga menggarisbawahi pentingnya pemanfaatan

teknik analisis data yang canggih dalam bidang kesehatan untuk meningkatkan hasil klinis. Untuk penelitian selanjutnya, disarankan untuk melakukan pengujian dengan *dataset* yang lebih besar dan beragam serta menerapkan algoritma lain yang mungkin dapat lebih meningkatkan akurasi dan keandalan dalam klasifikasi penyakit tiroid..

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada dosen pembimbing beserta dosen-dosen pengajar yang sudah banyak membantu dan membimbing dalam setiap perkuliahan. Ucapan terima kasih juga kepada rekan-rekan kelas yang banyak memberikan bantuan. Tanpa dukungan dan bantuannya penelitian ini mungkin tidak dapat terselesaikan dengan baik.

DAFTAR PUSTAKA

- [1] R. A. Astuti and F. N. Irfani, "Pemeriksaan Imunologi Terhadap Kadar Hormon Thyroid Stimulating Hormone (TSH) Pada Pasien Gangguan Tiroid Di RSUD Panembahan Senopati Periode 2020-2022," *Jurnal Ilmiah Kedokteran Dan Kesehatan*, vol. 3, no. 1, Article 1, 2024. [Online]. Available: <https://doi.org/10.55606/klinik.v3i1.2631>
- [2] L. Safitri, K. C. Murtiwiyati, S. Chodidjah, and D. Indayanti, "Perbandingan Metode Algoritma Decision Tree C4.5 Dan Naïve Bayes Untuk Memprediksi Penyakit Tiroid," *Journals of Ners Community*, vol. 13, no. 5, Article 5, 2022. [Online]. Available: <https://doi.org/10.55129/jnerscommunity.v13i5.2121>
- [3] Y. Saputri and M. M. Damayanti, "Karakteristik Pasien dengan Nodul Tiroid di Rumah Sakit X Bandung," *Jurnal Riset Kedokteran*, vol. 1, no. 2, pp. 71–79, 2021. [Online]. Available: <https://doi.org/10.29313/jrk.v1i2.438>
- [4] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes terhadap Analisis Sentimen Pengguna Layanan BPJS," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, pp. 646–654, 2020. [Online]. Available: <https://doi.org/10.32493/informatika.v5i4.7622>
- [5] F. Y. Pamuji and V. P. Ramadhan, "Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotherapy," *Jurnal Teknologi Dan Manajemen Informatika*, vol. 7, no. 1, Article 1, 2021. [Online]. Available: <https://doi.org/10.26905/jtmi.v7i1.5982>
- [6] W. A. Rahmat, S. M. Ladjamuddin, and D. T. Awaludin, "Perbandingan Algoritma Decision Tree, Random Forest Dan Naïve Bayes Pada Prediksi Penilaian Kepuasan Penumpang Maskapai Pesawat Menggunakan Dataset Kaggle," 2023. [Online]. Available: <https://ejournal.istn.ac.id/index.php/rekayasainformasi/article/view/1726>
- [7] I. Kurniawan, D. C. P. Buani, A. Abdussomad, W. Apriliah, and R. A. Saputra, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, Article 2, 2023. [Online]. Available: <https://doi.org/10.25126/jtiik.20231026225>
- [8] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris," *Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar*, vol. 7, no. 2, Article 2, 2021. [Online]. Available: <https://doi.org/10.35329/jiik.v7i2.203>
- [9] K. A. Khalim, U. Hayati, and A. Bahtiar, "Perbandingan Prediksi Penyakit Hipertensi Menggunakan Metode Random Forest Dan Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, Article 1, 2023. [Online]. Available: <https://doi.org/10.36040/jati.v7i1.6376>
- [10] A. M. Majid and I. Nawangsih, "Perbandingan Metode Ensemble Untuk Meningkatkan Akurasi Algoritma Machine Learning Dalam Memprediksi Penyakit Breast Cancer (Kanker

- Payudara)," *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika Dan Komputer)*, vol. 23, no. 1, Article 1, 2024. [Online]. Available: <https://doi.org/10.53513/jis.v23i1.9563>
- [11] A. H. Permana, F. R. Umbara, and F. Kasyidi, "Klasifikasi Penyakit Jantung Tipe Kardiovaskular Menggunakan Adaptive Synthetic Sampling dan Algoritma Extreme Gradient Boosting," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 1, Article 1, 2024. [Online]. Available: <https://doi.org/10.47065/bits.v6i1.5421>