OPEN ACCESS

Volume 9 (4), October-December 2025, 1562-1571

E-ISSN:2580-1643

Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)

DOI: https://doi.org/10.35870/jtik.v9i4.4236

Analisis Performa Algoritma Random Forest dalam Mengatasi Overfitting pada Model Prediksi

Muhammad Wisnu Nugroho [1*]

[1*] *Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Kota Salatiga, Provinsi Jawa Tengah, Indonesia.*

a r t i c l e i n f o

Article history:
Received 9 May 2025
Received in revised form 20 June 2025
Accepted 1 July 2025
Available online October 2025.

Keywords:
Random Forest; Machine Learning; Prediction Model; Accuracy; Ensemble Learning.

Kata Kunci:
Random Forest; Machine Learning; Model Prediksi; Akurasi; Ensemble Learning.

a b s t r a c t

The Random Forest algorithm is an ensemble-based machine learning method widely used to improve the accuracy of predictive models. This algorithm works by randomly constructing many decision trees and combining the results to produce more accurate predictions and reduce the risk of overfitting. The advantages of Random Forest lie in its ability to handle complex datasets, manage variables with high correlation, and provide stable results in various scenarios. This study aims to analyze the performance of the Random Forest algorithm in overcoming overfitting and improving the accuracy of predictive models in various fields. The method used in this study is a literature study (library research), by collecting and analyzing 40 scientific literature from various sources such as international journals, proceedings, and relevant academic articles. Data were analyzed qualitatively with a comparative-descriptive approach to the advantages and disadvantages of Random Forest compared to other algorithms such as Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, and Neural Networks. The results show that Random Forest excels in handling high-dimensional data, reduces the risk of overfitting, and provides stable prediction results in various applications such as healthcare, finance, manufacturing, and environmental analysis. This research is limited to literature-based analysis without empirical testing using actual datasets. For future research, it is recommended to conduct direct experiments implementing the Random Forest algorithm on real-world datasets, as well as explore combinations of other ensemble algorithms, such as XGBoost or LightGBM, to improve the accuracy and efficiency of predictive models.

a b s t r a k

Algoritma Random Forest merupakan salah satu metode machine learning berbasis ensemble yang banyak digunakan untuk meningkatkan akurasi model prediksi. Algoritma ini bekerja dengan membangun banyak pohon keputusan (Decision Trees) secara acak dan menggabungkan hasilnya untuk menghasilkan prediksi yang lebih akurat serta mengurangi risiko overfitting. Keunggulan Random Forest terletak pada kemampuannya dalam menangani dataset kompleks, mengelola variabel dengan korelasi tinggi, serta memberikan hasil yang stabil dalam berbagai skenario. Penelitian ini bertujuan untuk menganalisis performa algoritma Random Forest dalam mengatasi overfitting serta meningkatkan akurasi model prediksi pada berbagai bidang. Metode yang digunakan dalam penelitian ini adalah studi pustaka (library research), dengan mengumpulkan dan menganalisis 40 literatur ilmiah dari berbagai sumber seperti jurnal internasional, prosiding, dan artikel akademik yang relevan. Data dianalisis secara kualitatif dengan pendekatan komparatif-deskriptif terhadap keunggulan dan kelemahan Random Forest dibandingkan algoritma lain seperti Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, dan Neural Networks. Hasil penelitian menunjukkan bahwa Random Forest unggul dalam menangani data berdimensi tinggi, mengurangi risiko overfitting, dan memberikan hasil prediksi yang stabil di berbagai aplikasi seperti kesehatan, keuangan, manufaktur, dan analisis lingkungan. Penelitian ini dibatasi pada analisis berbasis literatur tanpa pengujian empiris menggunakan dataset aktual. Untuk penelitian selanjutnya, disarankan agar dilakukan eksperimen langsung dengan implementasi algoritma Random Forest pada dataset nyata serta eksplorasi kombinasi algoritma ensemble lainnya seperti XGBoost atau LightGBM untuk meningkatkan akurasi dan efisiensi model prediksi.

Corresponding Author. Email: wisnu.nugrojo@gmail.com [1].

ACM Computing Classification System (CCS)
EBSCOhost
Communication and Mass Media Complete (CMMC)

1. Pendahuluan

Perkembangan teknologi digital telah mendorong peningkatan penerapan *Artificial Intelligence (AI)* dan *Machine Learning* dalam mendukung proses pengambilan keputusan berbasis data di berbagai sektor. *Machine learning* memungkinkan sistem untuk mempelajari pola dari data historis, kemudian menggunakannya dalam membuat prediksi dengan tingkat akurasi yang semakin tinggi. Model prediktif berbasis *machine learning* telah banyak dimanfaatkan dalam bidang kesehatan, keuangan, manufaktur, pertanian, serta teknologi informasi. Salah satu tantangan dalam pengembangan model semacam ini adalah akurasi yang tidak konsisten, yang sering kali disebabkan oleh fenomena *overfitting*, *underfitting*, serta distribusi data yang tidak seimbang. Hal ini mendorong kebutuhan akan algoritma yang mampu meningkatkan ketepatan prediksi tanpa mengurangi kapasitas model dalam mengenali pola dari data yang belum pernah dilihat sebelumnya (Sobari *et al*., 2025). Algoritma *Random Forest* merupakan salah satu pendekatan populer yang dirancang untuk meningkatkan performa model prediksi. Algoritma ini merupakan pengembangan dari *Decision Tree* yang bekerja dengan cara membangun sejumlah pohon keputusan secara paralel, lalu menggabungkan hasilnya untuk menghasilkan prediksi akhir.

Random Forest memanfaatkan metode *ensemble learning*, yaitu dengan mengintegrasikan prediksi dari beberapa model dasar guna memperoleh hasil yang lebih stabil. Dengan membangun pohon keputusan secara independen dan menggabungkan prediksi melalui proses *voting* (untuk *klasifikasi*) atau perataan (*averaging*, untuk *regresi*), algoritma ini secara signifikan mampu menurunkan risiko *overfitting* yang umum terjadi pada model berbasis pohon tunggal. Selain itu, *Random Forest* memiliki keunggulan dalam mengelola data berdimensi tinggi dan tetap berfungsi baik ketika data mengandung nilai hilang maupun pencilan (*outlier*). Kinerja algoritma ini telah dibuktikan dalam berbagai implementasi. Di bidang kesehatan, misalnya, *Random Forest* digunakan untuk mendukung proses diagnosis melalui analisis data rekam medis pasien, seperti tekanan darah, kadar kolesterol, serta riwayat penyakit. Model ini terbukti dapat mengenali pola-pola kompleks yang tidak dapat ditangkap oleh algoritma lain. Dalam sektor keuangan, *Random Forest* digunakan dalam evaluasi kelayakan kredit dan deteksi anomali transaksi keuangan. Bank serta institusi keuangan mengandalkan model ini untuk meminimalkan risiko gagal bayar serta mengidentifikasi potensi penyimpangan dalam transaksi (Saputra *et al*., 2024). Di sektor industri, penerapan *Random Forest* dimanfaatkan dalam prediksi kerusakan mesin berbasis data sensor, yang mendukung pengembangan sistem perawatan prediktif (*predictive maintenance*) guna menekan biaya operasional dan mengurangi waktu henti produksi. Kendati demikian, penggunaan *Random Forest* tidak lepas dari tantangan teknis. Salah satu kendala utama adalah kebutuhan komputasi yang relatif besar dibandingkan model pohon tunggal, khususnya saat jumlah pohon dalam model meningkat secara signifikan. Selain itu, performa *Random Forest* sangat dipengaruhi oleh pengaturan parameter seperti jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), dan jumlah fitur yang digunakan dalam pemisahan (*split*). Jika parameter ini tidak diatur dengan tepat, kinerja model dapat mengalami penurunan (Wahyuni & Irawan, 2025). Tantangan lain muncul pada skenario *big data*, di mana jumlah fitur yang sangat besar dapat menyebabkan hambatan pada efisiensi proses pelatihan. Untuk itu, diperlukan strategi optimasi, seperti penggunaan *parallel computing* atau seleksi fitur yang efektif, guna menjaga efisiensi model (Aulia *et al*., 2024).

Sejumlah studi telah menunjukkan bahwa *Random Forest* merupakan salah satu algoritma yang unggul dalam mengurangi *overfitting*. Dengan membangun banyak pohon secara acak dan menggabungkan prediksinya, algoritma ini dapat menurunkan varians model serta meningkatkan kapasitas generalisasi. Misalnya, Alhabib (2022) melaporkan bahwa *Random Forest* menunjukkan performa lebih baik dibandingkan *Naïve Bayes* dan *Decision Tree* dalam prediksi penyakit jantung. Temuan serupa dilaporkan oleh Irfannandhy *et al*. (2024), yang menunjukkan bahwa kombinasi *Random Forest* dengan metode penyeimbangan data seperti *SMOTE* dapat meningkatkan akurasi dan mengurangi *overfitting* dalam prediksi risiko diabetes. Rayadin *et al*. (2024) juga menunjukkan bahwa algoritma ini tetap andal dalam menangani data yang mengandung *noise* dan *outlier*. Namun, kajian lebih lanjut masih diperlukan untuk memahami dampak pengaturan parameter model terhadap

kemampuannya dalam melakukan generalisasi, khususnya dalam perbandingan langsung dengan algoritma *ensemble* lain seperti *XGBoost* atau *Gradient Boosting*. Sejumlah publikasi masih terbatas pada studi kasus sektoral dan belum menyelidiki pengaruh interaksi antarparameter terhadap variasi performa model. Dengan dasar tersebut, studi ini bertujuan untuk menelaah secara sistematis efektivitas *Random Forest* dalam menekan risiko *overfitting* pada berbagai jenis aplikasi. Pembahasan meliputi karakteristik algoritma, faktor teknis yang memengaruhi kinerja, serta strategi optimasi yang digunakan dalam penelitian terdahulu. Fokus diberikan pada pemahaman menyeluruh terhadap peran *Random Forest* dalam menghasilkan model prediktif yang akurat dan tahan terhadap gangguan distribusi data. Pemahaman yang diperoleh dari kajian ini diharapkan dapat menjadi landasan dalam pengembangan model prediksi berbasis *Random Forest* secara lebih efisien dan presisi. Selain itu, pembahasan ini juga memberikan arah bagi penelitian lanjutan yang menggabungkan *Random Forest* dengan pendekatan lain seperti *deep learning* atau *boosting*, untuk menghasilkan sistem prediktif yang lebih adaptif terhadap tantangan data modern.

2. Metodologi Penelitian

Penelitian ini menggunakan pendekatan *library research* atau studi kepustakaan, yang dilaksanakan melalui tahapan pengumpulan, seleksi, analisis, dan sintesis literatur ilmiah secara sistematis. Seluruh literatur yang dianalisis memiliki relevansi dengan topik algoritma *Random Forest* dalam upaya peningkatan akurasi model prediksi (*Zed*, 2008). Kriteria inklusi yang digunakan dalam proses seleksi literatur meliputi: (1) artikel jurnal atau prosiding ilmiah yang dipublikasikan antara tahun 2018 hingga 2025 guna menjamin aktualitas informasi; (2) publikasi yang secara eksplisit membahas penerapan algoritma *Random Forest* dalam permasalahan *klasifikasi* maupun *regresi*; (3) dokumen yang tersedia dalam bahasa Indonesia atau bahasa Inggris; serta (4) literatur yang menyajikan data kuantitatif atau komparatif mengenai performa *Random Forest*. Di sisi lain, kriteria eksklusi meliputi: (1) tulisan yang bersifat opini tanpa metodologi ilmiah yang terverifikasi; (2) artikel populer non-akademik; dan (3) publikasi duplikat atau berulang. Dari total 95 dokumen awal yang diperoleh melalui basis data seperti *IEEE Xplore*, *ScienceDirect*, *SpringerLink*, *Google Scholar*, dan *ResearchGate*, sebanyak 42 literatur memenuhi seluruh kriteria inklusi dan digunakan sebagai sumber utama dalam analisis. Prosedur analisis dilakukan melalui tahap pengkodean tematik (*thematic coding*) dan klasifikasi literatur berdasarkan beberapa dimensi, yaitu: (a) bidang penerapan (misalnya kesehatan, keuangan, manufaktur, dan lingkungan); (b) algoritma pembanding yang digunakan dalam studi (seperti *Decision Tree*, *Support Vector Machine (SVM)*, *K-Nearest Neighbors (KNN)*, dan lainnya); (c) ukuran performa model yang dilaporkan, seperti akurasi, *F1-score*, sensitivitas, dan sebagainya; serta (d) pengaturan parameter *Random Forest* yang dianalisis, termasuk *n_estimators*, *max_depth*, dan teknik *feature selection*.

Proses sintesis dilakukan dengan membandingkan hasil temuan dari masing-masing publikasi dalam satu kerangka evaluasi terstruktur. Literatur dikelompokkan berdasarkan tipe penelitian, apakah bersifat eksperimental, *review*, atau studi kasus. Selanjutnya, dilakukan identifikasi atas persamaan dan perbedaan dalam hal performa algoritma, karakteristik penggunaannya, serta hambatan teknis yang dilaporkan dalam implementasi. Analisis juga mencakup evaluasi terhadap metode validasi model seperti *cross-validation* dan teknik *parameter tuning* yang digunakan dalam studi-studi tersebut. Dengan mempertimbangkan periode penerbitan, kriteria seleksi yang terdefinisi, serta sistematika dalam klasifikasi dan penelaahan isi, pendekatan ini diharapkan mampu memberikan representasi yang sahih mengenai efektivitas algoritma *Random Forest* dalam mengurangi risiko *overfitting* dan meningkatkan akurasi prediksi pada berbagai bidang penerapan.

3. Hasil dan Pembahasan

Hasil

Konsep Dasar *Machine Learning* dan Model Prediksi

Machine learning merupakan cabang dari *Artificial Intelligence (AI)* yang memungkinkan sistem komputer untuk mempelajari pola dari data dan menghasilkan keputusan atau prediksi tanpa pemrograman eksplisit. Perkembangan teknologi ini didorong oleh

ketersediaan data dalam jumlah besar serta kemajuan dalam kemampuan komputasi. Secara umum, *machine learning* terbagi menjadi tiga kategori utama, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Pada *supervised learning*, algoritma dilatih menggunakan *dataset* yang telah diberi label, dengan tujuan untuk memetakan input ke output tertentu. Contoh aplikasinya antara lain adalah klasifikasi email sebagai *spam* atau bukan, serta prediksi harga rumah berdasarkan sejumlah fitur. Sebaliknya, *unsupervised learning* digunakan ketika *dataset* tidak memiliki label, sehingga algoritma dituntut untuk mengenali pola tersembunyi secara mandiri. Pendekatan ini sering digunakan dalam analisis klaster, misalnya dalam segmentasi pelanggan. Adapun *reinforcement learning* berfokus pada pembelajaran melalui interaksi agen dengan lingkungan, di mana algoritma belajar berdasarkan sistem *reward* dan *punishment* untuk mencapai tujuan, seperti dalam pengembangan sistem permainan dan robotika (Depari *et al*., 2022). Salah satu penggunaan utama dari *machine learning* adalah pengembangan model prediktif yang digunakan untuk memperkirakan nilai atau hasil tertentu berdasarkan pola historis. Model ini dapat diterapkan dalam berbagai sektor, seperti prediksi harga saham di bidang keuangan, deteksi penyakit pada sektor kesehatan, dan proyeksi permintaan pasar di bidang bisnis.

Model prediksi bekerja dengan membangun relasi antara variabel independen (*features*) dan variabel dependen (*target*), sehingga dapat digunakan untuk menghasilkan output yang akurat pada data baru. Dalam kerangka *supervised learning*, model prediksi diklasifikasikan menjadi dua tipe, yaitu *regression* dan *classification*. *Regression* digunakan untuk memperkirakan nilai kontinu, seperti suhu udara, sedangkan *classification* bertujuan untuk mengelompokkan ke dalam kategori tertentu, misalnya menentukan apakah pelanggan akan melakukan pembelian. Namun, membangun model prediksi yang akurat memerlukan perhatian khusus terhadap fenomena *overfitting* dan *underfitting*. *Overfitting* terjadi saat model belajar terlalu rinci dari data latih hingga kehilangan kemampuan generalisasi, sementara *underfitting* muncul ketika model terlalu sederhana sehingga tidak mampu menangkap pola penting. Untuk mengatasi hal ini, berbagai teknik telah digunakan, termasuk *cross-validation*, *regularization*, dan *ensemble learning*. Salah satu metode *ensemble learning* yang banyak digunakan adalah *Random Forest*, yakni pengembangan dari algoritma *Decision Tree* yang dapat meningkatkan akurasi model dengan menggabungkan hasil dari sejumlah pohon keputusan (Apriliah *et al*., 2021). Selain struktur model, kualitas data juga memiliki pengaruh signifikan terhadap akurasi prediksi. Data dengan banyak nilai hilang (*missing values*), pencilan (*outliers*), atau distribusi tidak seimbang cenderung menghasilkan model yang tidak stabil. Oleh karena itu, tahap *data preprocessing* menjadi sangat penting. Proses ini meliputi pembersihan data, normalisasi, transformasi fitur, dan penggunaan teknik *sampling*. Selain itu, pemilihan fitur atau *feature selection* juga diperlukan agar model tidak dibebani oleh fitur yang tidak relevan, yang dapat meningkatkan beban komputasi dan risiko *overfitting*.

Dalam beberapa tahun terakhir, perkembangan *machine learning* juga didorong oleh munculnya pendekatan *deep learning*, khususnya *Neural Networks*, yang telah menunjukkan keunggulan dalam bidang seperti pengenalan wajah, pemrosesan bahasa alami (*Natural Language Processing*/*NLP*), dan sistem rekomendasi. Meskipun *deep learning* menawarkan akurasi yang tinggi dalam sejumlah aplikasi, algoritma klasik seperti *Random Forest* masih banyak digunakan karena kemudahan interpretasi, waktu pelatihan yang relatif singkat, dan performa yang baik pada *dataset* berdimensi tinggi (Imaniar Ikko Mulya Rizky *et al*., 2023). Secara keseluruhan, *machine learning* telah menjadi pilar penting dalam pengembangan sistem berbasis data. Dengan pemilihan algoritma yang tepat, pemrosesan data yang efektif, serta penyetelan parameter yang optimal, model prediktif dapat memberikan hasil yang akurat dan mendukung pengambilan keputusan yang lebih informatif. Seiring dengan bertambahnya data yang tersedia dan kemajuan teknologi komputasi, penerapan *machine learning* diperkirakan akan terus berkembang di berbagai sektor.

Algoritma *Random Forest*

Algoritma *Random Forest* merupakan salah satu pendekatan dalam *machine learning* yang mengadopsi prinsip *ensemble learning*. Teknik ini menggabungkan sejumlah model prediksi untuk menghasilkan akurasi dan kestabilan yang lebih baik dibandingkan dengan model tunggal. *Random Forest* dikembangkan sebagai

perbaikan atas algoritma *Decision Tree*, yang meskipun sederhana dan mudah dipahami, cenderung mengalami *overfitting* ketika diterapkan pada data kompleks. Untuk mengatasi hal tersebut, *Random Forest* membangun banyak pohon keputusan (*decision trees*) dan menggabungkan prediksinya melalui mekanisme *voting* (untuk klasifikasi) atau perataan (*averaging*) (untuk regresi), yang menghasilkan model lebih tangguh terhadap *outlier* dan variasi data (*Alhabib*, 2022). Proses kerja algoritma ini dimulai dengan pembentukan sejumlah pohon berdasarkan *subset* acak dari *dataset*, baik dari segi baris maupun fitur. Teknik ini dikenal sebagai *bootstrap aggregating* atau *bagging*, di mana setiap pohon dilatih menggunakan sampel acak dari data latih. Selain itu, hanya sebagian fitur yang dipertimbangkan dalam setiap pemisahan (*split*) pada pohon keputusan. Tujuan utamanya adalah untuk mengurangi korelasi antar pohon, sehingga hasil akhir lebih stabil dan tidak terlalu dipengaruhi oleh pola spesifik pada data latih. Setelah seluruh pohon selesai dibangun, hasil prediksi dari masing-masing pohon digabungkan melalui mekanisme *majority voting* untuk klasifikasi atau perataan untuk regresi (*Breiman*, 2001).

Salah satu kelebihan utama *Random Forest* adalah kemampuannya dalam menangani *dataset* berdimensi besar serta nilai yang hilang. Berbeda dengan beberapa algoritma lain yang memerlukan imputasi, *Random Forest* dapat beroperasi langsung pada data dengan *missing values* melalui pemanfaatan fitur yang tersedia di masing-masing pohon. Selain itu, algoritma ini dapat mengukur pentingnya setiap fitur melalui komponen *feature importance*, sehingga pengguna dapat mengidentifikasi variabel yang paling berpengaruh terhadap prediksi (Salsabil *et al*., 2024). Jika dibandingkan dengan algoritma lain seperti *Support Vector Machine (SVM)* atau *Neural Networks*, *Random Forest* menawarkan keunggulan dalam berbagai aspek. Pertama, algoritma ini relatif tidak sensitif terhadap *outliers* dan tidak membutuhkan proses normalisasi data yang kompleks. Kedua, kemampuan *Random Forest* dalam menangani *dataset* yang tidak seimbang lebih baik dibandingkan banyak metode lain, karena mekanisme *bagging* memungkinkan setiap pohon untuk belajar dari variasi sampel yang berbeda, termasuk kelas minoritas. Ketiga, dalam hal konfigurasi, *Random Forest* tergolong sederhana karena hanya memerlukan pengaturan beberapa parameter utama seperti jumlah pohon (*n_estimators*) dan kedalaman maksimum pohon (*max_depth*). Meskipun demikian, algoritma ini juga memiliki keterbatasan. Salah satu kendala yang sering ditemui adalah kebutuhan sumber daya komputasi yang tinggi, terutama jika jumlah pohon yang dibangun sangat banyak. Hal ini dapat memperlambat proses pelatihan maupun prediksi. Selain itu, pemilihan parameter yang tidak tepat tetap dapat menurunkan performa, meskipun secara umum *Random Forest* lebih tahan terhadap *overfitting* dibandingkan *Decision Tree* tunggal (Rayadin *et al*., 2024). Dalam praktiknya, *Random Forest* telah diterapkan secara luas dalam berbagai bidang. Di sektor kesehatan, algoritma ini digunakan untuk mendeteksi penyakit berdasarkan data klinis, seperti risiko diabetes atau penyakit jantung. Di sektor keuangan, digunakan dalam evaluasi kredit dan deteksi penipuan transaksi. Dalam industri manufaktur, algoritma ini membantu memprediksi potensi kegagalan mesin berdasarkan data sensor, sehingga mendukung strategi *predictive maintenance* guna mengurangi waktu henti produksi. Meskipun telah muncul algoritma yang lebih kompleks seperti *Gradient Boosting Machines (GBM)* dan *Deep Learning*, *Random Forest* tetap menjadi pilihan yang kompetitif karena stabilitas, kecepatan, dan kemudahan interpretasinya. Dengan penyetelan parameter yang tepat serta pemahaman terhadap karakteristik data, algoritma ini dapat digunakan secara efektif dalam membangun model prediktif pada berbagai bidang.

Perbandingan Akurasi dengan Algoritma Lain

Dalam ranah *machine learning*, pemilihan algoritma sangat dipengaruhi oleh karakteristik *dataset*, tujuan analisis, serta kebutuhan akan efisiensi komputasi dan kemudahan interpretasi. Algoritma *Random Forest* kerap dibandingkan dengan pendekatan lain seperti *Decision Tree*, *Support Vector Machine (SVM)*, *K-Nearest Neighbors (KNN)*, *Naïve Bayes*, dan *Neural Networks* dalam hal ketepatan prediksi, kemampuan generalisasi, serta kompleksitas operasional. Meskipun *Random Forest* dikenal memiliki performa yang andal di banyak skenario, terdapat kondisi tertentu di mana algoritma lain menunjukkan hasil yang lebih optimal. Oleh sebab itu, penting untuk mengevaluasi kinerja *Random Forest* secara komparatif terhadap algoritma lain berdasarkan jenis data dan jenis masalah yang dihadapi (*Rafrastara et al*., 2023).

Salah satu pembanding utama *Random Forest* adalah *Decision Tree*, mengingat *Random Forest* merupakan gabungan dari sejumlah pohon keputusan. Keunggulan *Decision Tree* terletak pada kesederhanaan dan kecepatan pelatihan, namun model ini rentan mengalami *overfitting*, terutama jika pohon dibangun terlalu dalam atau jika terdapat fitur yang saling berkorelasi. *Random Forest* mengurangi risiko tersebut dengan menerapkan teknik *bagging*, yang menurunkan varians model dan meningkatkan daya generalisasi. Penelitian menunjukkan bahwa akurasi *Random Forest* umumnya melampaui *Decision Tree* tunggal, khususnya pada data yang kompleks dan memiliki banyak variabel. Meskipun demikian, untuk *dataset* kecil yang tidak terlalu rumit, *Decision Tree* tetap dapat menjadi pilihan efisien karena waktu pelatihan yang singkat dan penggunaan sumber daya yang rendah.
Dibandingkan dengan *Support Vector Machine (SVM)*, *Random Forest* unggul dalam hal kecepatan dan skalabilitas, terutama pada *dataset* berskala besar. *SVM* dikenal efektif dalam menangani data berdimensi tinggi serta kasus dengan batas keputusan non-linier yang kompleks. Namun, kelemahan *SVM* terletak pada kebutuhan komputasi yang tinggi, terutama saat menggunakan kernel kompleks seperti *Radial Basis Function (RBF)*.

Sementara itu, *Random Forest* dapat dilatih secara paralel sehingga lebih efisien dalam memproses data berukuran besar. Dari sisi akurasi, *SVM* cenderung unggul jika data memiliki batas keputusan yang tegas dan bersih dari *noise*, namun *Random Forest* sering menunjukkan hasil lebih baik ketika data mengandung fitur yang tidak terstruktur atau redundan (Manullang *et al.*, 2023). Dalam perbandingan dengan *K-Nearest Neighbors (KNN)*, *Random Forest* menunjukkan keunggulan dalam menangani *outliers* dan dalam efisiensi komputasi saat bekerja dengan *dataset* besar. *KNN* menghitung jarak antara setiap titik uji dan seluruh data latih untuk menentukan prediksi, yang menjadi hambatan utama dalam skala besar. Meskipun *KNN* cukup akurat pada data kecil, waktu komputasi yang diperlukan meningkat secara eksponensial ketika jumlah sampel membesar. Sebaliknya, *Random Forest*, setelah proses pelatihan selesai, dapat memberikan prediksi lebih cepat karena hanya melalui jalur pohon keputusan. Dibandingkan dengan *Naïve Bayes*, *Random Forest* memberikan hasil yang lebih akurat pada data dengan hubungan antar fitur yang kompleks. *Naïve Bayes* bekerja dengan asumsi independensi antar fitur, yang jarang ditemukan pada data nyata. Meskipun demikian, *Naïve Bayes* memiliki keunggulan dalam hal kecepatan dan kinerja yang baik pada data teks, seperti klasifikasi email atau analisis sentimen. Pada kasus dengan korelasi fitur yang signifikan, *Random Forest* lebih dapat diandalkan karena mampu memodelkan interaksi antar variabel secara lebih efektif (Jan Melvin Ayu Soraya Dachi & Sitompul, 2023). Perbandingan dengan *Neural Networks*, khususnya dalam bentuk *Deep Learning* seperti *Convolutional Neural Networks (CNN)* dan *Recurrent Neural Networks (RNN)*, menunjukkan perbedaan yang mencolok dalam hal skala dan kompleksitas. *Neural Networks* sangat efektif dalam menangani data tidak terstruktur, seperti gambar, teks, dan suara, serta memiliki potensi akurasi yang tinggi dalam tugas-tugas tersebut. Namun, algoritma ini membutuhkan pelatihan yang lama, konsumsi sumber daya komputasi besar, serta sulit untuk diinterpretasikan. *Random Forest*, dengan struktur yang lebih sederhana dan hasil yang dapat dijelaskan, lebih unggul dalam aplikasi yang membutuhkan efisiensi dan transparansi. Secara umum, kinerja *Random Forest* relatif konsisten pada berbagai jenis data, terutama yang memiliki banyak fitur dengan interaksi yang rumit. Untuk *dataset* yang lebih bersih dan linier, algoritma seperti *SVM* atau *Neural Networks* bisa lebih unggul. Namun, dalam banyak aplikasi industri dan akademik, *Random Forest* sering kali dipilih karena memberikan keseimbangan antara akurasi, kecepatan, dan interpretabilitas. Dalam praktiknya, banyak pendekatan *hybrid* dikembangkan dengan menggabungkan *Random Forest* dan algoritma lain, misalnya *Gradient Boosting Machines (GBM)* atau *XGBoost*, untuk meningkatkan performa prediksi. Pendekatan ini memperkuat posisi *Random Forest* sebagai komponen penting dalam sistem klasifikasi dan regresi modern, baik di lingkungan penelitian maupun dunia industri.

Pembahasan

Algoritma *Random Forest* telah diterapkan secara luas dalam berbagai sektor karena kemampuannya dalam menangani *dataset* berskala besar, mengurangi risiko *overfitting*, serta menghasilkan prediksi yang akurat. Keandalannya dalam tugas *klasifikasi* dan *regresi* menjadikannya salah satu pilihan utama dalam aplikasi praktis, mulai dari bidang kesehatan, keuangan,

manufaktur, hingga lingkungan. Berbagai studi menunjukkan penerapan *Random Forest* dalam menyelesaikan persoalan dunia nyata melalui pendekatan berbasis data yang valid dan efisien (Iskandar, 2023). Dalam bidang kesehatan, *Random Forest* banyak digunakan untuk mendukung proses diagnosis penyakit dan prediksi risiko kesehatan. Salah satu contoh implementasinya adalah deteksi kanker payudara dengan memanfaatkan data medis seperti *mammogram* atau hasil biopsi. Berbagai fitur seperti ukuran tumor, tekstur sel, dan kepadatan jaringan digunakan dalam model untuk mengklasifikasikan apakah suatu sel bersifat jinak atau ganas. Dalam prediksi diabetes, algoritma ini digunakan untuk mengevaluasi faktor risiko seperti kadar glukosa darah, tekanan darah, indeks massa tubuh (*Body Mass Index/BMI*), serta riwayat keluarga. Beberapa studi menunjukkan bahwa penerapan teknik *ensemble learning* berbasis *Random Forest* mampu meningkatkan ketepatan prediksi dibandingkan model konvensional seperti *logistic regression* atau *Decision Tree* tunggal. Selain itu, algoritma ini juga digunakan dalam analisis data genomik untuk mengidentifikasi gen terkait penyakit tertentu, yang mendukung pengembangan di bidang bioinformatika dan terapi presisi.

Pada sektor keuangan, *Random Forest* digunakan dalam deteksi penipuan transaksi (*fraud detection*) dan analisis risiko kredit. Seiring meningkatnya volume transaksi digital, kebutuhan untuk mendeteksi pola anomali secara otomatis menjadi krusial. *Random Forest* digunakan untuk mengklasifikasikan transaksi berdasarkan variabel seperti lokasi, nominal, serta riwayat pembelian. Algoritma ini terbukti lebih efektif dibandingkan pendekatan berbasis aturan karena kemampuannya dalam mengelola variabilitas pola perilaku. Selain itu, dalam evaluasi kelayakan kredit, bank dan perusahaan *fintech* memanfaatkan *Random Forest* untuk memprediksi kemungkinan gagal bayar berdasarkan riwayat transaksi, pendapatan, dan kebiasaan keuangan. Keunggulannya terletak pada kemampuannya menangani *dataset* dengan fitur yang kompleks dan distribusi kelas yang tidak seimbang (Fatunnisa & Marcos, 2024). Di sektor manufaktur, algoritma ini digunakan dalam implementasi *predictive maintenance* dengan tujuan memprediksi kerusakan mesin sebelum terjadi kegagalan. Data sensor seperti suhu, getaran, dan tekanan dianalisis untuk mengidentifikasi pola yang menunjukkan potensi kerusakan. Dengan model prediktif berbasis *Random Forest*, perusahaan dapat menerapkan strategi pemeliharaan preventif dan meminimalkan *downtime* produksi. Contohnya, perusahaan otomotif besar seperti Tesla dan General Motors telah mengintegrasikan *machine learning*, termasuk *Random Forest*, dalam sistem kendaraan untuk memantau keausan komponen secara *real-time* dan merancang jadwal perawatan yang lebih efisien. Pada bidang lingkungan dan pertanian, *Random Forest* digunakan dalam berbagai studi, seperti prediksi cuaca, pemantauan perubahan iklim, serta optimalisasi hasil pertanian. Dalam prediksi cuaca, algoritma ini digunakan untuk menganalisis data historis seperti suhu, kelembapan, tekanan udara, dan kecepatan angin guna memperkirakan kondisi atmosfer secara lebih presisi. Selain itu, dalam pengawasan deforestasi, *Random Forest* digunakan untuk memproses citra satelit dan mengklasifikasikan tutupan lahan, serta mendeteksi perubahan lingkungan secara *real-time*. Organisasi seperti NASA telah mengimplementasikan teknologi ini untuk memantau deforestasi di kawasan hutan tropis seperti Amazon dan Asia Tenggara.

Dalam bidang pertanian, algoritma ini digunakan untuk menganalisis faktor seperti kualitas tanah, curah hujan, serta jenis pupuk guna menentukan strategi tanam yang paling efisien (Herjanto & Carudin, 2024). Selain itu, *Random Forest* juga diterapkan dalam sistem rekomendasi di sektor *e-commerce* dan hiburan digital. Platform seperti Netflix, Spotify, dan Amazon menggunakan algoritma ini untuk menyarankan produk, film, atau lagu berdasarkan histori interaksi pengguna dan atribut demografis. Dengan mengelompokkan pengguna berdasarkan preferensi serupa, model mampu memberikan rekomendasi yang lebih personal dan relevan. Dari berbagai studi tersebut, dapat disimpulkan bahwa *Random Forest* adalah algoritma yang fleksibel, mampu beradaptasi pada berbagai jenis data, serta menunjukkan performa yang tinggi pada beragam aplikasi. Kemampuannya dalam mengurangi *overfitting*, memilih fitur yang paling relevan, dan bekerja dengan baik meski pada *dataset* yang bising, menjadikannya salah satu algoritma yang paling banyak digunakan dalam pengembangan sistem prediktif berbasis data. Efektivitas *Random Forest* dalam mengendalikan *overfitting* juga terbukti pada studi kasus lintas sektor. Di bidang kesehatan, prediksi

kanker payudara dan risiko diabetes menunjukkan bahwa *Random Forest* lebih akurat dibandingkan pendekatan konvensional seperti *logistic regression* dan *Decision Tree*. Di sektor keuangan, algoritma ini berhasil mengidentifikasi pola anomali transaksi serta menjaga akurasi prediksi pada *dataset* yang sangat tidak seimbang. Dalam industri manufaktur, kemampuan model untuk mengenali pola dari data sensor secara *real-time* memungkinkan pemeliharaan berbasis prediksi yang lebih efisien dan akurat. Demikian pula, di bidang lingkungan dan pertanian, *Random Forest* mampu mengelola variabel dalam jumlah besar dan mengatasi *noise* tanpa kehilangan stabilitas model (Herjanto & Carudin, 2024). Penerapannya dalam sistem rekomendasi juga membuktikan bahwa *Random Forest* mampu menyesuaikan prediksi berdasarkan preferensi pengguna tanpa terlalu bergantung pada interaksi terbatas, yang berisiko menimbulkan *overfitting*. Keberhasilan ini menunjukkan bahwa stabilitas dan kemampuan generalisasi *Random Forest* sangat dipengaruhi oleh karakteristik data, mekanisme pemilihan fitur secara acak, serta strategi agregasi dari hasil pohon keputusan yang dibangun. Berdasarkan berbagai kasus nyata di berbagai sektor, dapat ditegaskan bahwa *Random Forest* merupakan algoritma yang efektif dan fleksibel dalam mengembangkan model prediksi yang akurat dan tahan terhadap *overfitting*. Keandalannya dalam berbagai domain aplikasi, dari diagnosis medis hingga pemodelan spasial, menjadikannya salah satu pendekatan yang relevan di era *big data* dan transformasi digital saat ini.

4. Kesimpulan dan Saran

Algoritma *Random Forest* merupakan salah satu metode dalam *machine learning* yang terbukti efektif dalam meningkatkan ketepatan model prediktif serta mengurangi risiko *overfitting*. Melalui pendekatan *ensemble learning*, algoritma ini membangun dan menggabungkan sejumlah pohon keputusan (*decision trees*) untuk memperkuat kemampuan generalisasi terhadap data yang belum pernah dilihat. Keunggulan struktural ini menjadikan *Random Forest* sebagai algoritma yang dapat diandalkan dalam berbagai sektor, termasuk kesehatan, keuangan, manufaktur, lingkungan, dan sistem rekomendasi.

Jika dibandingkan dengan algoritma lain seperti *Decision Tree*, *Support Vector Machine (SVM)*, *K-Nearest Neighbors (KNN)*, *Naïve Bayes*, dan *Neural Networks*, *Random Forest* kerap menunjukkan performa yang lebih stabil dan akurat, terutama pada *dataset* dengan dimensi tinggi, jumlah fitur yang banyak, serta hubungan antar variabel yang kompleks. Walaupun algoritma lain dapat menunjukkan keunggulan pada skenario tertentu misalnya *SVM* untuk data berdimensi tinggi atau *Neural Networks* untuk data tidak terstruktur seperti gambar dan teks *Random Forest* tetap menjadi pilihan strategis karena kemudahan dalam interpretasi hasil, efisiensi pelatihan, serta daya tahannya terhadap variasi data. Seiring dengan meningkatnya kebutuhan akan model prediksi yang andal dan dapat dijelaskan (*explainable*), penggunaan *Random Forest* diperkirakan akan terus berkembang, baik sebagai model utama maupun sebagai bagian dari pendekatan gabungan dengan teknik lain seperti *boosting* dan *deep learning*. Oleh karena itu, pemahaman yang mendalam mengenai kekuatan, batasan, serta parameter-parameter penting dalam konfigurasi *Random Forest* sangat diperlukan bagi peneliti dan praktisi *data science* dalam mengembangkan solusi prediktif yang akurat, efisien, dan relevan dengan kebutuhan dunia nyata.

5. Daftar Pustaka

Alhabib, I. (2022). Komparasi metode deep learning, naïve Bayes dan random forest untuk prediksi penyakit jantung. *Informatics for Educators and Professional: Journal of Informatics, 6*(2), 176. https://doi.org/10.51211/itbi.v6i2.1881.

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi random forest. *Sistemasi, 10*(1), 163. https://doi.org/10.32520/stmsi.v10i1.1129.

Aulia, Y., Andriyansyah, A., Suharjito, S., & Nensi, S. W. (2024). Analisis prediksi stroke dengan membandingkan tiga metode klasifikasi decision tree, naïve Bayes, dan random forest. *Jurnal Ilmu Komputer dan Informatika, 3*(2), 89–98. https://doi.org/10.54082/jiki.90.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Depari, D. H., Widiastiwi, Y., & Santoni, M. M. (2022). Perbandingan model decision tree, naïve Bayes dan random forest untuk prediksi klasifikasi penyakit jantung. *Informatik: Jurnal Ilmu Komputer, 18*(3), 239. https://doi.org/10.52958/iftk.v18i3.4694.

Fatunnisa, A., & Marcos, H. (2024). Prediksi kelulusan tepat waktu siswa SMK teknik komputer menggunakan algoritma random forest. *Jurnal Manajemen Informatika (JAMIKA), 14*(1), 101–111. https://doi.org/10.34010/jamika.v14i1.12114

Herjanto, M. F. Y., & Carudin, C. (2024). Analisis sentimen ulasan pengguna aplikasi Sirekap pada Play Store menggunakan algoritma random forest classifier. *Jurnal Informatika dan Teknik Elektro Terapan, 12*(2), 1204–1210. https://doi.org/10.23960/jitet.v12i2.4192.

Imaniar Ikko Mulya Rizky, Suhendro Yusuf Irianto, & Sriyanto. (2023). Perbandingan kinerja algoritma naïve Bayes, support vector machine dan random forest untuk prediksi penyakit ginjal kronis. *Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat, 18*, 139–151.

Irfannandhy, R., Handoko, L. B., & Ariyanto, N. (2024). Analisis performa model random forest dan CatBoost dengan teknik SMOTE dalam prediksi risiko diabetes. *Edumatic: Jurnal Pendidikan Informatika, 8*(2), 714–723. https://doi.org/10.29408/edumatic.v8i2.27990.

Iskandar, D. (2023). Optimasi parameter random forest menggunakan grid search untuk analisis time series. *Petir, 16*(2), 267–277. https://doi.org/10.33322/petir.v16i2.2084.

Jan Melvin Ayu Soraya Dachi, & Pardomuan Sitompul. (2023). Analisis perbandingan algoritma XGBoost dan algoritma random forest ensemble learning pada klasifikasi keputusan kredit. *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam, 2*(2), 87–103. https://doi.org/10.55606/jurrimipa.v2i2.1470

Manullang, O., Prianto, C., & Harani, N. H. (2023). Analisis sentimen untuk memprediksi hasil calon pemilu presiden menggunakan lexicon based dan random forest. *Jurnal Ilmiah Informatika, 11*(2), 159–169. https://doi.org/10.33884/jif.v11i02.7987.

Prasojo, B., & Haryatmi, E. (2021). Analisa prediksi kelayakan pemberian kredit pinjaman dengan metode random forest. *Jurnal Nasional Teknologi dan Sistem Informasi, 7*(2), 79–89. https://doi.org/10.25077/teknosi.v7i2.2021.79-89.

Rafrastara, F. A., Supriyanto, C., Paramita, C., & Astuti, Y. P. (2023). Deteksi malware menggunakan metode stacking berbasis ensemble. *Jurnal Informatika: Jurnal Pengembangan IT, 8*(1), 11–16. https://doi.org/10.30591/jpit.v8i1.4606.

Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementasi ensemble learning metode XGBoost dan random forest untuk prediksi waktu penggantian baterai aki. *BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer, 5*(2), 111–119.

Ristianto, A., & Sanjaya, F. I. (2025). Klasifikasi kualitas air sungai Daerah Istimewa Yogyakarta (DIY) menggunakan algoritma random forest. *Jurnal Sistem Informasi dan Sistem Komputer, 10*(1), 168–181.

Salsabil, M., Lutvi, N., & Eviyanti, A. (2024). Implementasi data mining dalam melakukan prediksi penyakit diabetes menggunakan metode random forest dan XGBoost. *Jurnal Ilmiah Komputasi, 23*(1), 51–58. https://doi.org/10.32409/jikstik.23.1.3507.

Saputra, D. B., Atina, V., & Nastiti, F. E. (2024). Penerapan model CRISP-DM pada prediksi nasabah kredit menggunakan algoritma random forest. *Idealis: Indonesia Journal Information System,*

7, 240–247.

Sobari, S., Purnamasari, A. I., Bahtiar, A., & Kaslani, K. (2025). Meningkatkan model prediksi kelulusan santri tahfidz di Pondok Pesantren Al-Kautsar menggunakan algoritma random forest. *Jurnal Informatika dan Teknik Elektro Terapan, 13*(1).

Tarigan, L. R. A., & Dahlan, D. (2024). Optimalisasi fitur dengan forward selection pada estimasi tingkat penyakit paru-paru menggunakan algoritma klasifikasi random forest. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(5), 10341–10348.

Wahyuni, R., & Irawan, Y. (2025). Model prediksi risiko kebakaran hutan menggunakan algoritma random forest dengan seleksi fitur lasso regression. *JEKIN - Jurnal Teknik Informatika, 5*(1), 91–100.

Zed, M. (2008). *Metode penelitian kepustakaan*. Yayasan Pustaka Obor Indonesia.