

Literatur Review: Penerapan Algoritma Random Forest untuk Klasifikasi Penyakit Diabetes

**Fajar Sidiq Wijaya^{1*}, Muhammad Farhan Arotsid², Riedo Adriano³,
Zakia Dwihadi Larasati⁴**

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

Email: ¹fajarsidiq3450@gmail.com, ²mfarotsid15@gmail.com, ³riedoadriano29@gmail.com,
⁴larasatzakiadwihadi@gmail.com,

(* : coresponding author)

Abstract – Penyakit diabetes ini adalah sebuah penyakit kronis yang banyak diderita dan di alami oleh semua kalangan usia. Penyakit diabetes ini salah satu penyebab angka kematian yang tinggi dan jarang terdeteksi secara dini oleh tubuh yang terkena diabetes ini. Oleh karena itu mendiagnosa penyakit diabetes dini sangat krusial untuk menurunkan dari resiko komplikasi dan meningkatkan pemulihannya. Tujuan dari penelitian ini adalah untuk mengembangkan sebuah model dari klasifikasi penyakit diabetes dengan menerapkannya algoritma Random Forest. Model ini juga bertujuan untuk bisa mengklasifikasi beberapa dari gejala awal penyakit diabetes seperti dari garis keturunan, kadar gula darah yang tinggi, hipertensi dan berat badan berlebihan. Hasil dari penelitian ini dapat berkontribusi pada dokter dan tenaga kesehatan serta masyarakat umum untuk mendeteksi penyakit diabetes sejak dini.

Kata Kunci : Penyakit diabetes, Klasifikasi, Deteksi dini, Random Forest

***Abstract** - Diabetes is a chronic disease that affects individuals across all age groups and is a leading cause of mortality, frequently going undetected in its early stages. Early diagnosis of diabetes is thus critical to minimize the risk of complications and enhance recovery potential. This study aims to develop a classification model for diabetes by implementing the Random Forest algorithm. The model is also intended to classify early symptoms associated with diabetes, such as family history, elevated blood glucose levels, hypertension, and obesity. The findings of this study are expected to aid healthcare providers and the general public in detecting diabetes at an early stage.*

Keywords : Diabetes diseases, Classification, Early Detection, Random Forest

1. PENDAHULUAN

Penyakit diabetes menurut Organisasi Kesehatan Dunia (WHO) menjadi salah satu penyakit degeneratif kronis yang disebabkan oleh gangguan dari metabolisme glukosa yang menyebabkan produksi dari insulin yang tidak mencukupi di pankreas dan mengalami kekurangannya produksi insulin baik secara absolut maupun relatif. Penderita yang mengidap penyakit ini mempunyai 2 tipenya, tipe 1 dan tipe 2. namun secara umum gejala awal dari kedua tipe ini seperti mengalami sering merasa kehausan, frekuensi dari buang air kecil meningkat, rasa lelah gangguan penglihatan, keputihan dan luka infeksi yang lama sembuh. Gejala awal dari penyakit diabetes memang sangat bervariasi pada setiap pasien, sehingga sulit untuk dikenali. Menurut dari beberapa pengidap penyakit ini masih belum terdiagnosis dan masih belum menyadari kalau dirinya terkena diabetes, karena gejala awalnya ini memang hanya seperti sakit biasa, Sehingga banyak sekali orang yang telat menyadari bahkan sudah mengarah pada komplikasi.

Pada perkembangan teknologi sekarang, khususnya di bidang Machine Learning sangat membantu dalam dunia kesehatan, yang dimana di kalangan dokter dan tim kesehatan lainnya mendapatkan kesempatan dalam meningkatkan kualitas dari kerja tim tenaga kesehatannya. Dengan mengimplementasikan model dari klasifikasi penyakit diabetes menggunakan algoritma Random Forest, tujuan dari penelitian ini sudah tercapai karena algoritma dari Random Forest memiliki model perkembangan yang lebih konsisten. Sehingga dari tim kesehatan bisa menggunakannya dalam mengenali gejala-gejala awal dari penyakit diabetes ini, supaya nantinya dari pasien yang sudah terkena gejalanya bisa lebih cepat di tangani dan menghindarinya dari komplikasi penyakit lainnya.

2. METODE PENELITIAN

2.1 Tinjauan Literatur Sistematis (SLR)

Metode *Systematic Literatur Review* (SLR) digunakan untuk pengkajian, penganalisaan dan juga pengumpulan data dari berbagai penelitian terkait klasifikasi pada penyakit diabetes, dengan penekanan pada penggunaan metode dari algoritma Random Forest. Dengan permulaan prosesnya mencari sumber-sumber artikel atau jurnal yang terpercaya, seperti dari database akademik penyedia jurnal-jurnal yang ada. SLR ini bertujuan untuk memberikan sedikit dari keseluruhan terhadap isi-isinya penelitian yang telah dilakukan sebelumnya, dan mengevaluasi dari metode yang telah digunakan dalam mengklasifikasi penyakit diabetes.

2.2 Random Forest

Metode Random Forest merupakan salah satu metode algoritma klasifikasi yang dibuat untuk memprediksi dari sekumpulan prediksi pohon keputusan dan juga termasuk dari metode pembelajaran terarah. Metode ini dipilih karena atas dasar dari kelebihanannya dalam akurasi klasifikasi yang sangat bagus dengan error yang lebih rendah. Pendekatan Random Forest ini digunakan sebagai alat untuk mengklasifikasi penyakit diabetes yang dimana nanti hasil dari pendekatan ini akan bisa mendeteksi dari gejala-gejala awal penyakit diabetes.

2.3 Pohon Keputusan (Decision Tree)

Metode pohon keputusan (decision tree) adalah salah satu algoritma pembelajaran terawasi (supervised learning) yang sering digunakan untuk tugas klasifikasi dan regresi. Pohon keputusan bekerja dengan membagi data menjadi beberapa bagian berdasarkan aturan tertentu, sehingga pada akhirnya dapat mencapai keputusan atau klasifikasi yang diinginkan. Pohon keputusan dapat digunakan untuk memprediksi apakah seseorang menderita diabetes atau tidak berdasarkan beberapa variabel kesehatan. Metode yang intuitif dan cocok digunakan dalam klasifikasi penyakit seperti diabetes karena menghasilkan model yang dapat dieksplorasi dan divalidasi lebih lanjut oleh para ahli medis.

2.4 Analisis Efektivitas Sistem Dalam klasifikasi Penyakit Diabetes

Algoritma *random forest* yang diperkenalkan oleh Leo Breiman pada 2001, seorang profesor ahli dalam bidang machine learning yang berasal dari University of California. Konsep dari random forest adalah menggabungkan banyak pohon keputusan. Algoritma bisa menggabungkan hasil dari beberapa banyak pohon keputusan yang dibangun secara acak, maka model yang dihasilkan bisa lebih akurat. Kelemahan utama dari pohon keputusan adalah kecenderungan overfitting pada data latih, terutama jika pohon terlalu dalam atau kompleks. Di sinilah algoritma seperti random forests (yang menggabungkan beberapa pohon keputusan) bisa membantu mengatasi masalah tersebut dengan meningkatkan akurasi dan mengurangi overfitting. Saat sistem sudah siap, lakukan pengujian dengan data baru yang tidak pernah digunakan sebelumnya untuk melatih model. Hal ini bertujuan untuk memastikan bahwa model dapat diandalkan.

2.5 Research Questions

Research Questions merupakan bagian penting dari setiap penelitian yang membantu menentukan fokus dan arah penelitian. Pertanyaan penelitian ini dirancang untuk mengevaluasi efektivitas algoritma Random Forest dalam klasifikasi penyakit Diabetes, dan untuk memahami faktor-faktor yang mempengaruhi keakuratan model dan perbandingannya dengan metode lain. Pertanyaan untuk penelitian ini adalah sebagai berikut:

Tabel 1. Pertanyaan Penelitian

No	Pertanyaan Penelitian
1	Bagaimana efektivitas algoritma random forest dalam mengklasifikasi risiko penyakit diabetes dibandingkan dengan metode pohon keputusan?

2	Faktor-faktor apa saja yang paling memengaruhi akurasi algoritma randomforest dalam klasifikasi penyakit diabetes?
3	Bagaimana performa algoritma random forest dalam klasifikasi penyakitdiabetes ketika diterapkan pada dataset dengan variasi jumlah data latih?

2.6 Proses Penelitian

Proses penelitian dilakukan dengan mengakses berbagai website dan jurnal-jurnal akademik, seperti ResearchGate, IEEE Xplore, ScienceDirect dan Google Scholar, dengan kata kunci yang relevan seperti “Random Forest dalam Diagnosis Medis Diabetes”, “RandomForest for Diabetes Classification” dan “Machine Learning untuk Klasifikasi Diabetes dengan Random Forest”. Random Forest dalam klasifikasi penyakit Diabetes dan implementasi sistem. Artikel-artikel yang ditemukan kemudian dianalisis untuk memperoleh informasi yang relevan dan saling melengkapi penelitian ini.

2.7 Pengumpulan Data

Pengumpulan data dilakukan dengan tujuan untuk mendapatkan dataset yang relevan dan berkualitas dalam klasifikasi penyakit diabetes. Data dikumpulkan melalui dua sumber utama: data primer dari dataset kesehatan terbuka serta literatur pendukung untuk mendukung analisis algoritma, mengutamakan karya ilmiah terbaru dan memiliki ISSN.

2.7.1 Data Primer

Data utama untuk penelitian ini adalah dataset kesehatan yang berisi catatan pasien yang telah diidentifikasi berdasarkan kriteria tertentu yang relevan dengan klasifikasi diabetes. Dataset ini diambil dari sumber dataset kesehatan terbuka, seperti Pima Indians Diabetes Database, yang tersedia di UCI Machine Learning Repository atau Kaggle. Dataset ini berisi beberapa fitur yang relevan dengan diagnosis diabetes, termasuk:

- Kadar glukosa darah
- Indeks Massa Tubuh (BMI)
- Tekanan darah
- Usia

Riwayat keluarga terkait diabetes, dan faktor lainnya.

Dataset ini dipilih karena memiliki variabel yang umum digunakan dalam diagnosis diabetes, sehingga cocok untuk analisis klasifikasi menggunakan algoritma random forest dan pohon keputusan.

2.7.2 Data Sekunder

Data sekunder mencakup literatur ilmiah dari jurnal, buku, dan artikel yang terkait dengan penerapan algoritma random forest dan pohon keputusan dalam klasifikasi penyakit. Pencarian literatur dilakukan di beberapa basis data akademik seperti ResearchGate, IEEE Xplore, ScienceDirect, dan Google Scholar dengan kata kunci “random forest untuk klasifikasi diabetes,” “pohon keputusan dalam klasifikasi penyakit,” “prediksi diabetes menggunakan algoritma Machine Learning,” dan “Machine Learning untuk Klasifikasi Diabetes dengan Random Forest”.

Literatur ini digunakan untuk mendukung analisis algoritma yang diterapkan, serta membandingkan hasil penelitian ini dengan penelitian sejenis. Selain itu, literatur ini juga membantu dalam proses benchmarking model yang digunakan.

2.7.3 Kriteria Inklusi dan Eksklusi

Kriteria inklusi:

- Dataset yang berfokus pada klasifikasi penyakit diabetes dan memiliki variabel yang relevan.

- Artikel yang mengkaji penerapan algoritma random forest dan pohon keputusan untuk klasifikasi penyakit.
- Penelitian yang menggunakan teknik validasi model seperti cross-validation.

Kriteria eksklusi:

- Artikel yang tidak mencakup penerapan random forest atau pohon keputusan dalam penyakit diabetes.
- Dataset dengan fitur yang tidak relevan dengan diagnosis diabetes.

2.8 Analisis Data

Analisis data dapat dilakukan dengan beberapa langkah utama guna mengevaluasi performa algoritma random forest dan pohon keputusan dalam klasifikasi diabetes. Langkah-langkah analisis ini meliputi:

1. Pra-pengolahan Data

- **Data Cleaning:** Menangani data yang hilang dan menghapus data yang tidak relevan.
- **Normalisasi:** Melakukan normalisasi fitur agar nilai fitur berada dalam skala yang seragam, sehingga algoritma bekerja lebih optimal.
- **Split Data:** Dataset dibagi menjadi data latih dan data uji dengan perbandingan 80:20.

2. Pelatihan Model

Algoritma random forest dan pohon keputusan dilatih menggunakan data latih. Parameter utama seperti jumlah pohon dalam random forest (`n_estimators`) dan kedalaman pohon (`max_depth`) disesuaikan untuk mengoptimalkan akurasi model. Penggunaan cross-validation diterapkan untuk mengurangi kemungkinan overfitting dan meningkatkan generalisasi model.

3. Evaluasi Model

Setelah model dilatih, kinerja model diuji pada data uji menggunakan metrik-metrik berikut:

- **Akurasi:** Mengukur seberapa banyak prediksi yang benar secara keseluruhan.
- **Presisi dan Recall:** Mengukur kinerja model dalam mengidentifikasi kasus diabetes (positif) dan bukan diabetes (negatif).
- **F1 Score:** Menggabungkan presisi dan recall untuk memberi gambaran keseimbangan antara keduanya.
- **AUC-ROC Curve:** Digunakan untuk mengevaluasi performa model pada berbagai threshold.

4. Analisis Perbandingan

Model random forest dan pohon keputusan dibandingkan untuk menentukan model mana yang lebih efektif dalam klasifikasi penyakit diabetes. Analisis ini melibatkan perbandingan kinerja pada metrik-metrik yang telah disebutkan, serta perbandingan feature importance dari kedua algoritma, untuk menentukan variabel mana yang paling signifikan dalam menentukan risiko diabetes.

5. Interpretasi dan Diskusi Hasil

Hasil dari evaluasi model ini diinterpretasikan untuk memahami kekuatan dan kelemahan setiap algoritma dalam mendeteksi diabetes. Analisis ini mencakup interpretasi feature importance, di mana fitur seperti kadar glukosa atau BMI dapat dilihat sebagai faktor risiko utama berdasarkan hasil random forest. Diskusi juga mencakup perbandingan hasil dengan studi literatur yang relevan.

3. ANALISA DAN PEMBAHASAN

Bagian ini berisi kesimpulan, hasil dan pembahasan dari topik penelitian

Tabel 1. Penelitian Terkait

NO	Author/ Tahun	Metode Penelitian	Kelebihan Random Forest	Kekurangan Random Forest	Faktor Pengaruh Keberhasilan	Manfaat Sistem
1.	(Ajeng Citra Mawani , Rusdah, Law Li Hin , Dian Anubhakti,2023)	Deteksi dini gejala awal penyakit diabetes menggunakan algoritma Random Forest	Algoritma Random Forest memberikan hasil terbaik setelah dilakukan klasifikasi dan data uji	Penggunaan number of trees dan maximal depth yang terbatas dapat menyebabkan kehilangan potensi akurasi yang lebih tinggi	Penggunaan algoritma Random Forest pada metode splitting data komposisi 90:10 dengan dataset asli	Sistem mampu memprediksi gejala awal diabetes sejak dini dengan akurasi 90,38%
2.	(Sriyanto, Agiskaria Supriyana, 2023)	Prediksi penyakit diabetes menggunakan algoritma Random Forest	Random forest dapat menggabungkan decision trees sehingga menghasilkan model yang lebih akurat	Ukuran dataset yang kecil dan tidak seimbang, serta tidak adanya perbandingan dengan algoritma lainnya	Pada perhitungan AUC (area under curve) memperoleh nilai yang sangat tinggi sehingga klasifikasi memiliki tingkat akurasi yang tinggi	Hasil yang akurat, sistem dapat diimplementasikan pada bentuk perangkat lunak untuk prediksi penyakit diabetes
3.	(Anggita Ghazali , Hasih Pratiwi, Sri Sulistijowati Handayani, 2023)	Implementasi data mining menggunakan metode Random Forest dan Support Vector Machine dalam klasifikasi penyakit diabetes	Algoritma Random Forest dapat menghasilkan error yang relatif rendah, performa yang baik dalam klasifikasi dan cocok untuk jumlah data yang besar	Kurangnya validasi dan analisis terhadap variabel pada dataset yang menjelaskan apakah variabel tersebut memiliki pengaruh terhadap deteksi	Penggunaan metode Random Forest dengan split data 80:20 serta variabel menghasilkan akurasi yang tinggi	Dengan metode dan variabel yang digunakan, akurasi yang didapatkan dalam mendeteksi diabetes adalah

				diabetes		98%.
4.	(Bagus Rizki Prasetyo, Eka Dyar Wahyu ni,Prisa Marga Kusumantara, 2024	Komparasi performa model berbasis algoritma Random Forest dan LightGBM dalam melakukan klasifikasi penyakit diabetes melitus gestasional	Random Forest memiliki keunggulan pada penggunaan cpu yang rendah dengan kisaran antara 1% hingga 3,9%. Dibandingkan LightGBM	Random Forest membutuhkan waktu pemrosesan 20 mili seconds yang dimana lebih lama dari LightGBM yaitu 3 mili seconds	Penggunaan ADASYN pada Model 2 dan 3 yang berbasis Random Forest mampu membuat model menjadi lebih stabil dengan meningkatkan jumlah kelas minoritasnya	Sistem dapat melakukan prediksi digunakan untuk melakukan deteksi dini diabetes melitus gestasional pada ibu hamil dengan nama CheckDMG
5.	(Resa Budi Prasetyo, 2024)	Prediksi dini penyakit diabetes pada ibu hamil dengan algoritma Random Forest	Random Forest memiliki proses seleksi fitur yang memungkinkan fitur terbaik, meningkatkan kinerja terhadap model klasifikasi	pada model Random Forest memiliki kompleksitas yang tinggi dan keterbatasan dalam interpretabilitas dan ketidakseimbangan data	Menggunakan metrik utama untuk menilai kinerja adalah akurasi, mengukur model dalam prediksi yang benar dibandingkan dengan total prediksi	Random Forest mampu memberikan hasil prediksi yang handal dan konsisten, dengan akurasi akhir 98%

4. KESIMPULAN

Penelitian Ini mengevaluasi efektivitas algoritma Random Forest dan pohon keputusan dalam klasifikasi penyakit diabetes. Berdasarkan analisis dan pengujian yang dilakukan, beberapa kesimpulan utama yang dapat diambil:

1. **Efektivitas Algoritma** : Random forest terbukti memiliki performa yang lebih tinggi dibandingkan pohon keputusan dalam klasifikasi penyakit diabetes. Hal ini disebabkan oleh sifat random forest yang menggabungkan prediksi dari beberapa pohon keputusan, yang meningkatkan akurasi dan mengurangi kemungkinan overfitting.
2. **Pengaruh Faktor-faktor Kesehatan** : Variabel seperti kadar glukosa darah, indeks massa tubuh (BMI), dan tekanan darah teridentifikasi sebagai faktor yang paling signifikan dalam model klasifikasi diabetes. Random forest, dengan fitur feature importance, memberikan wawasan yang mendalam mengenai kontribusi setiap variabel dalam menentukan risiko diabetes.
3. **Penerapan pada Pengujian data dan Pelatihan** : Pengujian data latih dan data uji menunjukkan bahwa Random Forest lebih stabil dan generalis dalam memprediksi kelas, sehingga memberikan hasil yang konsisten dan dapat diandalkan dalam berbagai skenario. Sementara itu, pohon keputusan cenderung overfitting pada data latih jika tidak dikontrol kedalamannya.
4. **Evaluasi Kinerja Model** : Berdasarkan matrik evaluasi diantaranya akurasi, presisi, recall, dan F1 score, algoritma random forest menghasilkan yang sangat baik, terutama dalam mendeteksi kasus positif diabetes. Hal ini menunjukkan bahwa algoritma ini cocok untuk aplikasi medis yang memerlukan akurasi yang tinggi.

Secara keseluruhan, Random Forest merupakan algoritma yang kuat dan efektif dalam mendeteksi diabetes dibandingkan dengan algoritma yang lainnya. Temuan ini dapat menjadi acuan untuk pengembangan sistem pendukung keputusan medis yang lebih akurat dalam diagnosis diabetes.

REFERENCES

- R. Andanika Siallagan, (2021). "PREDIKSI PENYAKIT DIABETES MELLITUS MENGGUNAKAN ALGORITMA C4.5," J.RESPONSIF, vol. 3, no. 1, pp. 44 – 52.
- N. M. Putry, (2022). "Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus," EVOLUSI J. Sains dan Manaj., vol. 10, no. 1.
- E. C. Johns, F. C. Denison, J. E. Norman, dan R. M. Reynolds, (2018). "Gestational Diabetes Mellitus: Mechanisms, Treatment, and Complications," Trends in Endocrinology & Metabolism.
- C. C. Olisah, L. Smith, dan M. Smith, (2022). "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," Comput Methods Programs Biomed.
- C. Z. V. Junus, T. Tarno, dan P. Kartikasari, (2023). "Klasifikasi Menggunakan Metode Support Vector Machine Dan Random Forest Untuk Deteksi Awal Risiko Diabetes Mellitus," Jurnal Gaussian, vol. 11, no. 3, hlm. 386 – 396.
- Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4. 5. Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen).
- Nugraha, W., & Sabaruddin, R. (2021). Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4. 5, Random Forest, dan SVM.
- A. Primajaya dan B. N. Sari, (2018). "Random Forest Algorithm for Prediction of Precipitation," Indones. J. Artif. Intell. Data Min.
- A. Pinandito, S. A. Wicaksono, dan S. H. Wijoyo, (2023). "Implementasi Machine Learning Dalam Deteksi Risiko Tinggi Diabetes Melitus Pada Kehamilan".
- Rianti Nurpalah, Meti Kusmiati, Meri Meri, Hendro Kasmanto, dan Dina Ferdiani, (2023). "Deteksi Dini Diabetes Melitus Gestasional (Dmg) Melalui Pemeriksaan Glukosa Darah Sebagai Upaya Pencegahan Komplikasi Pada Ibu Hamil,."
- Rosyani Perani, dkk, (2021). "Klasifikasi Citra menggunakan Metode Random Forest dan Sequential Minimal Optimization (SMO)" : JUSTIN (Jurnal Sistem dan Teknologi Informasi),
- Rosyani Perani, dkk, (2023). "Literature Review : Implementasi Sistem Pakar untuk Diagnosa Penyakit Diabetes menggunakan Metode Fuzzy", BIKER : Jurnal Ilmu Komputer, Teknik dan Multimedia, (Juni, 2023).

- Wasis Haryono, Nida Fitrillah. (2022). “Sistem Informasi Perhitungan Kebutuhan Gizi Ibu Hamil menggunakan Metode Harris Benedict”, OKTAL : Jurnal Ilmu Komputer dan Sains, (November, 2022).
- Hadi Zakaria, Triani Krismonica Ningsih, (2023). “ Implementasi Algoritma K- NearestNeighbor pada Sistem Deteksi Penyakit Jantung : Studi kasus: Klinik Makmur Jaya”, LOGIC : Jurnal Ilmu Komputer dan Pendidikan, (Desember, 2023).
- Fajar Agung Nugroho, Wahyu Santoso, (2024). “ Penerapan Sistem Penjualan Makanan berbasis Android Kotlin dengan Metode Prototype”, OKTAL: Jurnal Ilmu Komputer dan Sains, (September, 2024).