JURNAL ILMIAH PERPUSTAKAAN, SISTEM INFORMASI, INFORMATIKA & ELEKTRO
Volume 1, Nomor 1, Juni 2026
.
hlm. 1-10

ANALISIS KOMPARATIF ALGORITMA DECISION TREE DAN RANDOM FOREST DALAM KLASIFIKASI PENYAKIT DIABETES

Andrian Falah Kalyana[1], Feri Sulianta[2]
[1,2]Fakultas Teknik, Universitas Widyatama, Jl. Cikutra no 204 A Bandung 40124
E-mail korespondensi: [1]andrian.falah@widyatama.ac.id, [2]feri.sulianta@widyatama.ac.id

Keywords: *Diabetes, Machine Learning, Decision Tree, Random Forest, Medical Prediction.*

Abstract
Diabetes is a global health issue that requires an accurate early detection system to prevent chronic complications. This study aims to analyze and compare the performance of two *machine learning* algorithms *Decision Tree* and *Random Forest* in predicting the risk of diabetes. The research methodology uses the Pima Indians Diabetes secondary dataset from Kaggle, which was processed through data preprocessing stages, including handling missing values and feature standardization using *StandardScaler*. Model evaluation was conducted by measuring accuracy, precision, recall, and F1-score metrics. The analysis results show that the *Decision Tree* algorithm delivers the most optimal performance with an accuracy rate of 76%. The research findings confirm that glucose and body mass index (BMI) features have the most significant influence on prediction accuracy. It is hoped that the contributions of this research can serve as a reference in the development of an efficient clinical decision support system for early diabetes screening based on computational data.

Kata kunci: *Diabetes, Machine Learning, Decision Tree, Random Forest, Prediksi Medis.*

Abstrak
Diabetes merupakan masalah kesehatan global yang memerlukan sistem deteksi dini yang akurat untuk mencegah komplikasi kronis. Penelitian ini bertujuan untuk menganalisis dan membandingkan performa dua algoritma *Machine Learning*, yaitu *Decision Tree* dan *Random Forest*, dalam memprediksi risiko penyakit diabetes. Metodologi penelitian ini menggunakan dataset sekunder Pima Indians Diabetes dari Kaggle yang diolah melalui tahapan pra-pemrosesan data, termasuk penanganan nilai hilang (missing values) dan standardisasi fitur menggunakan *StandardScaler*. Evaluasi model dilakukan dengan mengukur metrik akurasi, presisi, recall, dan f1-score. Hasil analisis menunjukkan bahwa algoritma *Decision Tree* memberikan performa paling optimal dengan tingkat akurasi sebesar 76%. Temuan penelitian mengonfirmasi bahwa fitur glukosa dan indeks massa tubuh (BMI) memiliki pengaruh paling signifikan terhadap ketepatan prediksi. Kontribusi penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem pendukung keputusan klinis yang efisien untuk skrining awal diabetes berbasis data komputasional.

PENDAHULUAN

Diabetes merupakan kondisi medis kronis yang ditandai dengan tingginya kadar gula darah akibat gangguan produksi atau efektivitas insulin dalam tubuh. Insulin adalah hormon yang diproduksi oleh pankreas, berfungsi mengatur kadar glukosa agar dapat diserap oleh sel sebagai sumber energi (Aditya et al., 2024). Secara global, data *International Diabetes Federation* (IDF) tahun

2021 mencatat lebih dari 537 juta penderita diabetes di seluruh dunia, dengan angka kematian mencapai 6,7 juta jiwa. Angka ini menunjukkan bahwa diabetes merupakan masalah kesehatan global yang memerlukan perhatian serius (Desmita et al., 2025). Tanpa diagnosis dini yang tepat, diabetes dapat memicu komplikasi kesehatan jangka panjang yang serius, yang mengancam kesejahteraan individu maupun masyarakat luas (Ibrahim et al., 2025a).

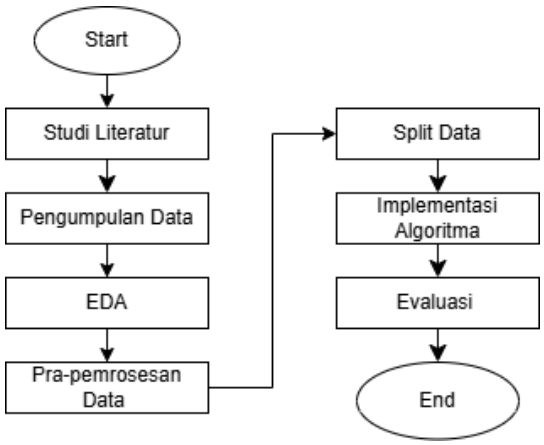
Seiring dengan perkembangan teknologi, pemanfaatan teknologi *Machine Learning* dalam bidang kesehatan, khususnya untuk prediksi penyakit diabetes telah menunjukkan potensi yang sangat besar (Siswoyo & Nurhafidz, 2025). *Machine Learning* adalah cabang kecerdasan buatan yang berfokus pada pengembangan sistem yang mampu belajar dari data untuk meningkatkan akurasi prediksi secara bertahap (Ginting et al., 2022). Model ini memproses parameter klinis seperti kadar glukosa, tekanan darah, BMI, usia, dan riwayat keluarga untuk memprediksi risiko diabetes dengan akurasi tinggi (Handayani et al., 2025).

Di antara berbagai algoritma yang tersedia, *Decision Tree* menawarkan keunggulan dalam hal interpretabilitas, memungkinkan dokter memahami alur logika keputusan, meskipun sering kali rentan terhadap overfitting. Sementara itu, *Random Forest*, sebagai metode *ensemble*, hadir untuk menutupi kelemahan model tunggal dengan meningkatkan akurasi dan stabilitas prediksi melalui agregasi banyak pohon keputusan (Amritha & Dayanti, 2026).

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa data dari dua model *Machine Learning Decision Tree* dan *Random Forest*, dalam prediksi penyakit diabetes. Hasil penelitian ini diharapkan dapat memberikan panduan dalam pemilihan teknik yang optimal untuk mendukung interpretasi hasil prediksi diabetes berbasis *Machine Learning*, serta meningkatkan efektivitas komunikasi hasil prediksi kepada praktisi medis dan pasien (Ibrahim et al., 2025b).

METODE

Metode yang digunakan dalam penelitian ini terdiri dari beberapa tahapan sistematis, mulah dari pengumpulan data hingga evaluasi model. Alur penelitian digambarkan melalui tahapan berikut:



Gambar 1 Flowchart Alur Penelitian

1. Studi Literatur

Literatur dalam studi ini diperoleh dari berbagai macam jurnal medis. Tujuannya adalah untuk memberikan pemahaman mengenai penyakit diabetes dan bagaimana algoritma *Decision Tree* dan *Random Forest* telah digunakan oleh peneliti sebelumnya.

2. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari platform kaggle. Dataset ini dipilih karena memiliki atribut klinis yang relevan untuk klasifikasi diabetes, data tersebut memiliki 9 fitur, dimana diantaranya 8 fitur sebagai input and 1 fitur sebagai output.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) adalah pendekatan yang umum digunakan untuk menganalisis dan menginterpolasi informasi yang berguna melalui grafik dan visualisasi statistik untuk mendapatkan pemahaman yang lebih baik tentang kumpulan data yang sedang dipelajari (Silmina & Perkasa, 2025).

4. Pra-pemrosesan Data

Pra-pemrosesan data dilakukan untuk memastikan dataset yang digunakan memiliki kualitas yang baik dan siap untuk proses pelatihan model Tahapan pra-pemrosesan yaitu pembersihan data, transformasi data, pembagian data (Sidiq et al., 2025).

5. Split Data

Data dibagi menjadi dua yaitu data latih (*training* data) dan data uji (*testing* data). Pembagian data dilakukan untuk menghindari bias dan memastikan evaluasi model yang objektif.

6. Implementasi Algoritma

- **Algoritma *Decision Tree***

Algoritma *Decision Tree* menggunakan *graph* seperti pohon yang mewakili struktur *root* atau akar dan *leaf* atau daun. Setiap *root* pohon mewakili sebuah atribut. Setiap cabang dari simpul mewakili hasil tes, dan simpul terakhir adalah "daun" yang mewakili label atau kelas. Untuk menentukan *root*, algoritma *Decision Tree* umumnya menggunakan *Information Gain* atau Gini Index (Karo & Hendriyana, 2022).

- **Algoritma *Random Forest***

Random Forest adalah algoritma *Ensemble Learning* yang menerapkan teknik *Bagging* untuk membangun beberapa pohon keputusan untuk menghasilkan sampel acak dan melatih pohon keputusan dari sampel tersebut. Dua aspek utama dalam algoritma *Random Forest* meliputi pembentukan beberapa pohon keputusan selama proses pelatihan dan penggabungan prediksi melalui pemungutan suara mayoritas. *Random Forest* memiliki beberapa keunggulan, antara lain tingkat akurasi yang tinggi, kemampuan untuk menangani data yang mengandung noise, kecepatan kinerja dan kontrol terhadap overfitting (Hanif & Utomo, 2025).

7. Evaluasi Model

Setelah model dilatih, performa keduanya diukur menggunakan data uji. Parameter evaluasi yang digunakan meliputi Akurasi, Presisi, Recall, dan F1-Score. Hasil dari kedua algoritma ini dibandingkan untuk melihat algoritma mana yang memberikan prediksi paling stabil dan akurat dalam mendeteksi penyakit diabetes.

HASIL

1. Statistika Deskriptif

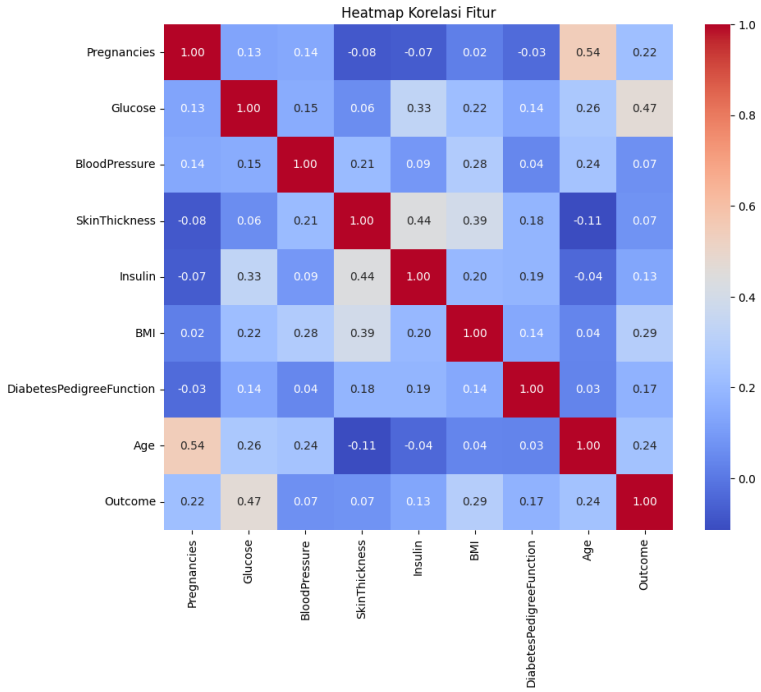
Analisis deskriptif terhadap dataset Kaggale menunjukkan karakteristik klinis respoden yang beragam. Dataset ini terdiri dari 768 baris dengan 8 fitur klinis yaitu *Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction,* dan *Age* dan 1 variabel target (*Outcome*) untuk kelas 0 (tidak diabetes) dan unutk kelas 1 (diabetes). Tabel 1 menyajikan ringkasan statistik untuk variabel utama yang menjadi indikator diabetes.

Tabel 1 Dataset

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

2. Uji Korelasi Dan Asumsi

Uji asumsi dilakukan melalui pemetaan korelasi antar fitur menggunakan *Heatmap*. Hasil menunjukkan bahwa fitur *Glucose* memiliki hubungan linear paling signifikan terhadap *Outcome*. Selain itu, penggunaan *StandatScaler* memastikan asumsi keseragaman skala data terpenuhi sebelum proses pemodelan. Bisa dilihat pada gambar di bawah ini.



Gambar 2 Heatmap Korelasi Fitur

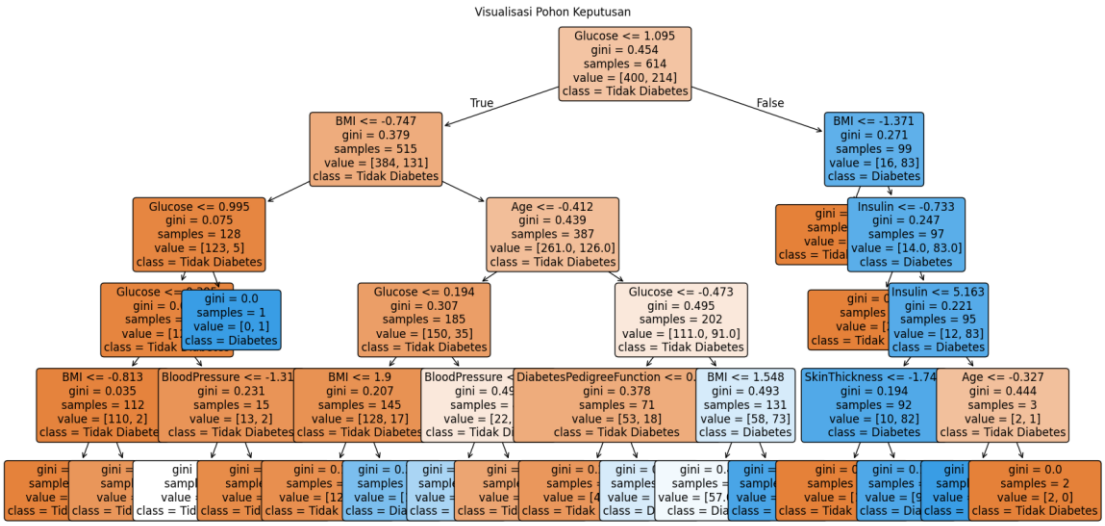
Setelah melakukan pemetaan korelasi antar fitur, tahap selanjutnya pra-pemrosesan data yaitu pembersihan data, handling missing value, dan pembagian data, bisa dilihat pada tabel di bawah ini data yang sudah siap digunakan untuk pemodelan.

Tabel 2 Data Yang Sudah Pra-Pemrosesan

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	125.0	33.6	0.627	50	1
1	1	85.0	66.0	29.0	125.0	26.6	0.351	31	0
2	8	183.0	64.0	29.0	125.0	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

3. Uji Hipotesis Performa Model

Uji Hipotesis dilakukan untuk membandingkan efektivitas kedua model. Data dibagi menjadi 80% data latih dan 20% data uji. Hasil evaluasi berdasarkan *confusion matrix* dan *classification report*. Gambar dibawah ini adalah visualisasi dari pohon keputusan.

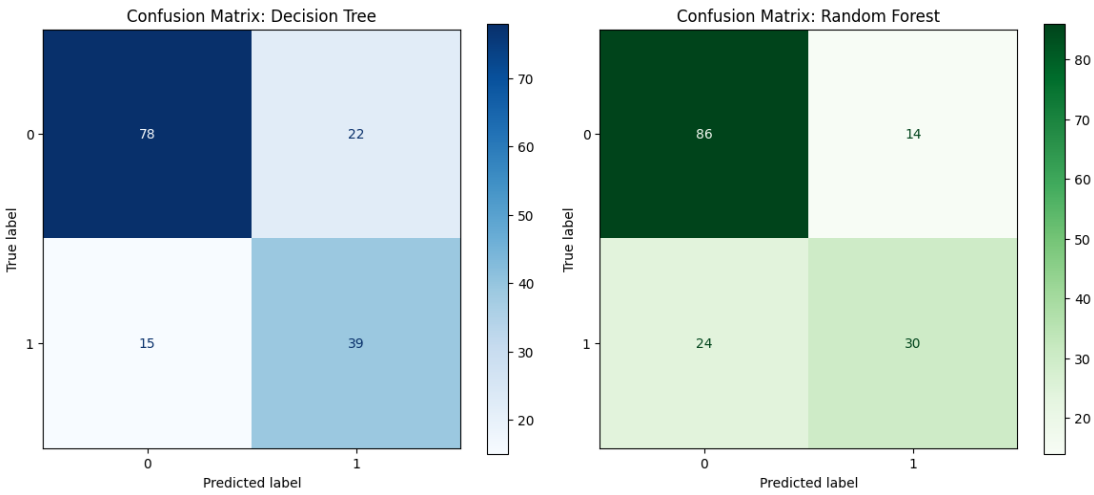


Gambar 3 Visualisasi Pohon Keputusan

Tabel 3 Hasil Evaluasi Perbandingan Kedua Model

	Model	Accuracy	Precision	Recall	F1-Score
0	Decision Tree	0.760	0.639	0.722	0.678
1	Random Forest	0.753	0.682	0.556	0.612

Berdasarkaan tabel 3, *Decision Tree* menunjukkan performa akurasi sebesar 76%, sedikit unggul dibandingkan *Random Forest* yang mencapai 75%. Namun, perlu dicatat bahwa *Random Forest* memiliki nilai *Precision* yang lebih baik (0,682), yang mengindikasikan kemampuan model dalam meminimalkan kesalahan prediksi positif palsu (*False Positive*).



Gambar 4 *Confusion Matrix Decision Tree & Random Forest*

Berikut hasil analisis menggunakan *confusion matrix*. Akurasi tinggi pada kelas 0 (tidak diabetes) dan untuk kelas 1 (diabetes). Kedua model ini memiliki performa yang sangat baik dalam mengklasifikasikan tidak diabetes dan diabetes.

Confusion matrix pada Algoritma *Decision Tree*, model ini berhasil mengklasifikasikan 78% tidak diabetes (kelas 0) dengan benar, dari total 100 pasien. Untuk pasien yang terkena diabetes (kelas 1), model berhasil mengidentifikasi 39 kasus yang benar dati total 54 pasien.

Confusion matrix pada Algoritma *Random Forest*, model ini berhasil mengklasifikasikan 86% tidak diabetes (kelas 0) dengan benar, dari total 100 pasien. Untuk pasien yang terkena diabetes (kelas 1), model berhasil mengidentifikasi 30 kasus yang benar dati total 54 pasien.

Penulisan Rumus

1. Algoritma Decision Tree

Gini Impurity

Gini Impurity digunakan untuk mengukur seberapa sering elemen yang dipilih secara acak dari set data akan salah diklasifikasikan.

$$Gini(S) = 1 - \sum_{i=1}^{c} (p_i)^2 \qquad (1)$$

Keterangan:

S: Himpunan data (dataset).

Σ: Jumlah kelas (dalam kasus kamu ada 2: Diabetes atau Tidak).

p_i: Probabilitas atau proporsi sampel yang termasuk dalam kelas i.

Entropy

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2(p_i) \qquad (2)$$

Serupa dengan Gini Impurity hanya saja memiliki perbedaan yaitu *entropy* mengizinkan penggunanya untuk memilih pemisahan yang meminimalisir ketidakpastian di dalam klasifikasi, sedangkan Gini impurity akan langsung meminimalisasi kemungkinan kesalahan dalam melakukan klasifikas

2. Algoritma Random Forest

algoritma ini membuat banyak pohon (default 100 pohon). Rumusnya bukan lagi satu pohon, melainkan Voting Mayoritas:

$$y = mode\{T_1(x), T_2(x), \dots, T_n(x)\} \qquad (3)$$

Artinya, hasil akhir y adalah kelas yang paling banyak dipilih oleh seluruh pohon T yang ada di dalam forest tersebut.

3. Metriks Evaluasi

Presentasi total prediksi yang benar

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

PEMBAHASAN

Hasil analisis menunjukkan bahwa algoritma *Decision Tree* memiliki performa yang lebih baik dibandingkan dengan *Random Forest* dalam mengklasifikasikan pasien diabetes. *Decision Tree* berhasil mencapai akurasi rata-rata sebesar 76%, sedangkan *Random Forest* memiliki akurasi rata-rata 75%. Hal ini menunjukkan bahwa *Decision Tree* tidak hanya memiliki akurasi lebih tinggi, tetapi juga lebih stabil dan andal secara statistik.

Hasil confusion matrix memperlihatkan karakteristik yang berbeda antara kedua model dalam mendeteksi pasien diabetes. Decision Tree menunjukkan kemampuan yang lebih baik dalam menangkap kasus positif dengan berhasil mengklasifikasikan 39 dari 54 pasien positif diabetes (True Positive), sementara Random Forest hanya berhasil mengidentifikasi 30 dari 54 pasien positif. Hal ini membuat nilai recall Decision Tree lebih unggul dalam mendeteksi penyakit. Namun demikian, Random Forest memiliki keunggulan signifikan dalam memprediksi pasien non-diabetes. Random Forest hanya menghasilkan 14 kesalahan prediksi pada orang sehat (False Positive), lebih rendah dibandingkan Decision Tree yang mencapai 22 kesalahan. Dengan tingkat False Positive yang lebih rendah ini, Random Forest memiliki nilai precision yang lebih baik, yang berarti model ini lebih dipercaya saat memberikan vonis positif.

Interpretasi terhadap efektivitas model tidak dapat dilepaskan dari peran tahap pra-pemrosesan data. Penanganan nilai nol (*zero values*) pada fitur klinis seperti *Glucose* dan *BMI* melalui teknik imputasi median telah terbukti meningkatkan akurasi model secara signifikan. Data medis sering kali mengandung *noise* atau data hilang yang jika tidak ditangani dengan tepat, akan menyebabkan bias pada fase *training*. Perbedaan mendasar dalam penelitian ini adalah penggunaan *StandardScaler* yang disesuaikan secara spesifik untuk fitur dengan rentang nilai luas seperti *Insulin*, sehingga mencegah satu fitur mendominasi proses pembelajaran model dibandingkan fitur lainnya.

KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa penggunaan algoritma *Machine Learning* memberikan kontribusi signifikan dalam memvalidasi parameter klinis sebagai prediktor risiko diabetes. Melalui analisis komparatif, ditemukan bahwa meskipun Random Forest memiliki keunggulan dalam stabilitas melalui mekanisme *ensemble*, pada dataset dengan karakteristik

tertentu, *Decision Tree* mampu memberikan performa akurasi dan *recall* yang bersaing. Temuan utama penelitian ini menegaskan bahwa tingkat glukosa darah dan indeks massa tubuh (BMI) merupakan faktor determinan paling konsisten dalam klasifikasi risiko medis. Secara substansial, penelitian ini memberikan kontribusi pada pengembangan ilmu pengetahuan dengan menunjukkan bahwa efektivitas model prediksi tidak hanya bergantung pada kompleksitas algoritma, tetapi juga pada optimalisasi pra-pemrosesan data klinis. Hal ini membuka wawasan baru bahwa model yang lebih sederhana dan dapat diinterpretasikan secara visual (interpretable AI) memiliki nilai guna yang tinggi dalam mendukung pengambilan keputusan klinis yang cepat dan akurat di fasilitas kesehatan dengan sumber daya terbatas.

Berdasarkan temuan penelitian ini, disarankan bagi peneliti selanjutnya untuk melakukan eksplorasi lebih mendalam dengan mengintegrasikan teknik penanganan ketidakseimbangan data (class imbalance) seperti *SMOTE* guna meningkatkan sensitivitas model terhadap kelompok pasien berisiko tinggi. Selain itu, pengembangan penelitian ke depan dapat diarahkan pada pengujian model menggunakan dataset primer yang lebih luas dan variatif secara demografis untuk memperkuat generalisasi hasil. Dari sisi perkembangan ilmu pengetahuan, disarankan adanya integrasi antara pendekatan komputasional dengan analisis psikologi kesehatan guna memahami bagaimana hasil prediksi otomatis ini dapat memengaruhi persepsi risiko dan kepatuhan pasien dalam menjalani gaya hidup sehat. Kontribusi baru ini diharapkan dapat menciptakan sistem deteksi dini yang tidak hanya akurat secara teknis, tetapi juga adaptif terhadap aspek perilaku manusia dalam manajemen kesehatan kronis.

REFERENSI

Aditya, M. F., Pramuntadi, A., Wijaya, D. P., & Wicaksono, Y. (2024). Implementasi Metode Decision Tree pada Prediksi Penyakit Diabetes Melitus Tipe 2: Implementation of Decision Tree Method for Diabetes Mellitus Type 2 Prediction. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 1104–1110. https://doi.org/10.57152/malcom.v4i3.1284

Amritha, Y. D., & Dayanti, A. R. (2026). *Model Machine Learning yang Dioptimalkan untuk Prediksi Penyakit Jantung Menggunakan R Shiny*. *8*.

Desmita, N. L., Lonang, S., & Kumoro, D. T. (2025). COMPARATIVE ANALYSIS OF DECISION TREE AND RANDOM FOREST ALGORITHMS FOR PREDICTING DIABETES MELLITUS. *SainsTech Innovation Journal*, *8*(1), 507–518. https://doi.org/10.37824/sij.v8i1.2025.783

Ginting, R. G., Girsang, E., Ginting, J. B., & Hartono, H. (2022). ANALISIS DETERMINAN DAN PREDIKSI PENYAKIT DIABETES MELITUS TIPE 2 MENGGUNAKAN METODE MACHINE LEARNING: SCOPING REVIEW. *Jurnal Maternitas Kebidanan*, *7*(1), 58–72. https://doi.org/10.34012/jumkep.v7i1.2538

Handayani, O. P., Ashari, I. A., & Ardianto, R. (2025). *Systematic Literature Review: Penerapan Machine Learning dalam Diagnosis dan Prediksi Penyakit Diabetes*. *14*(2).

Hanif, H., & Utomo, D. W. (2025). Prediksi Diabetes menggunakan Metode Ensemble Learning dengan Teknik Soft Voting. *Infotekmesin*, *16*(1), 127–134. https://doi.org/10.35970/infotekmesin.v16i1.2534

Ibrahim, M. C., Fachruddin, F., & Nurhadi, N. (2025a). Perbandingan Data Prediksi Diabetes Menggunakan Machine Learning: Comparison of Diabetes Prediction Data Using Machine Learning. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(4), 1423–1436. https://doi.org/10.57152/malcom.v5i4.2301

Ibrahim, M. C., Fachruddin, F., & Nurhadi, N. (2025b). Perbandingan Data Prediksi Diabetes Menggunakan Machine Learning: Comparison of Diabetes Prediction Data Using Machine Learning. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(4), 1423–1436. https://doi.org/10.57152/malcom.v5i4.2301

Karo, I. M. K., & Hendriyana, H. (2022). Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score. *Jurnal Teknologi Terpadu*, *8*(2), 94–99. https://doi.org/10.54914/jtt.v8i2.564

Sidiq, S., Alfian, A., & Mabrur, N. S. (2025). Pengembangan Model Prediksi Risiko Diabetes Menggunakan Pendekatan AdaBoost dan Teknik Oversampling SMOTE. *Jurnal Ilmiah Informatika Dan Ilmu Komputer (JIMA-ILKOM)*, *4*(1), 13–23. https://doi.org/10.58602/jima-ilkom.v4i1.41

Silmina, E. P., & Perkasa, L. (2025). EDA and Tableau Analysis for Identification of Heart Disease Risk Factors. *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, *5*(1), 79. https://doi.org/10.30811/jaise.v5i1.6389

Siswoyo, B., & Nurhafidz, M. I. (2025). Penerapan Algoritma Random Forest Untuk Prediksi Risiko Diabetes Berdasarkan Data Kesehatan Pasien. *Jurnal Teknologi Informasi Digital*, *1*(1), 35–38.