

Perbandingan Decision Tree dan Random Forest untuk Deteksi Dini Diabetes

M. Lutfi¹, Cecep Warman², Rieska Rahayu Ayuningsih³

^{1,2,3}Program Studi Sistem Informasi, Fakultas Teknik Komputer dan Desain, Universitas Nusa Putra, Indonesia
¹m.lutfi_si22@nusaputra.ac.id, ²cecep.warman@nusaputra.ac.id, ³rieska.rahayu@nusaputra.ac.id

Abstrak

Diabetes mellitus merupakan gangguan metabolik kronis yang membutuhkan mekanisme deteksi dini agar komplikasi jangka panjang dapat dicegah sejak awal. Penelitian ini membandingkan kinerja prediktif dua algoritma klasifikasi berbasis pohon, yaitu *Decision Tree* dan *Random Forest*, dalam mendeteksi diabetes pada tahap awal menggunakan dataset Pima Indians Diabetes dari UCI Machine Learning Repository. Sebagai pelengkap, penelitian ini menerapkan kerangka Explainable Artificial Intelligence lima lapis berbasis SHAP guna menghasilkan penjelasan model yang bermakna secara klinis, mulai dari kepentingan fitur global hingga validasi terhadap pengetahuan patofisiologi yang sudah mapan. Rekayasa fitur memperluas ruang fitur dari 8 menjadi 14 variabel, sementara pra-proses mencakup stratified train-test split, imputasi median, winsorizing berbasis IQR, normalisasi Min-Max, dan SMOTE untuk menangani ketidakseimbangan kelas. Optimasi hyperparameter dilakukan melalui Grid-SearchCV dengan Stratified K-Fold sebanyak lima lipatan, menggunakan F1-Score sebagai kriteria utama. Hasil eksperimen menunjukkan bahwa *Random Forest* secara konsisten mengungguli *Decision Tree* pada seluruh metrik evaluasi, dengan F1-Score *cross-validation* sebesar 0,9002 berbanding 0,8687 dan ROC-AUC pada set pengujian sebesar 80,61% berbanding 70,43%. McNemar Test menghasilkan $p = 0,4807$ yang tidak signifikan secara statistik pada ukuran sampel yang tersedia, namun pertimbangan multidimensional termasuk stabilitas *cross-validation* dan selisih ROC-AUC mendukung rekomendasi terhadap *Random Forest*. Analisis SHAP mengonfirmasi bahwa Insulin, Glucose, dan rasio Insulin terhadap Glucose merupakan tiga prediktor paling berpengaruh, sejalan dengan mekanisme patofisiologi diabetes tipe 2 yang telah mapan. Penelitian ini juga mengakui keterbatasan generalisabilitas dataset terhadap populasi Indonesia dan menetapkan validasi pada data primer lokal sebagai agenda penelitian lanjutan.

Kata kunci: *decision tree*; diabetes mellitus; explainable artificial intelligence; random forest; SHAP

1. Pendahuluan

Diabetes mellitus merupakan penyakit metabolik kronis yang ditandai oleh hiperglikemia persisten akibat gangguan sekresi atau kerja insulin (American Diabetes Association, 2024). Menurut Organisasi Kesehatan Dunia, lebih dari 422 juta orang di seluruh dunia hidup dengan kondisi ini, sementara data Riskesdas 2018 mencatat prevalensi nasional Indonesia sebesar 2,6% pada penduduk berusia 15 tahun ke atas (Kementerian Kesehatan Republik Indonesia, 2019). Komplikasi jangka panjang yang ditimbulkan, seperti nefropati, retinopati, neuropati, dan penyakit kardiovaskular, menjadikan deteksi dini sebagai intervensi yang paling efektif untuk menekan angka morbiditas dan mortalitas.

Sistem kesehatan Indonesia menghadapi hambatan struktural yang nyata dalam proses skrining diabetes, meliputi keterbatasan infrastruktur laboratorium di fasilitas kesehatan primer, tingginya biaya pemeriksaan HbA1c dan tes toleransi glukosa oral, serta distribusi tenaga medis yang tidak merata antara

wilayah perkotaan dan pedesaan (Markus, Kors, & Rijnbeek, 2021). Kondisi ini mendorong kebutuhan akan sistem pendukung keputusan klinis berbasis *machine learning* yang mampu melakukan skrining awal dengan akurasi tinggi menggunakan data klinis dasar yang tersedia di Puskesmas, tanpa memerlukan pemeriksaan laboratorium yang mahal.

Penelitian ini secara spesifik memilih *Decision Tree* (DT) dan *Random Forest* (RF) di antara berbagai algoritma yang tersedia karena dua alasan metodologis yang saling melengkapi. DT dipilih karena interpretabilitas intrinsiknya yang tinggi, memungkinkan penelusuran keputusan secara transparan melalui aturan IF-THEN yang eksplisit, sehingga klinisi dapat memahami jalur logika yang mengarah pada prediksi 'diabetes' atau 'non-diabetes' tanpa memerlukan penjelasan tambahan (Nurussakinah & Faisal, 2023). RF dipilih karena keunggulan kinerja prediktif ensemble-nya yang secara konsisten mengungguli algoritma tunggal pada berbagai *benchmark* medis melalui mekanisme Bootstrap Aggregating (Apriliyah,

Kurniawan, Baydhowi, & Haryati, 2021; Sriyanto & Supriyatna, 2023). Perbandingan langsung antara DT sebagai *baseline* interpretabel dan RF sebagai representasi ensemble dalam satu kerangka eksperimen yang terkontrol memungkinkan analisis trade-off yang komprehensif antara interpretabilitas dan kinerja prediktif.

Interpretabilitas model AI dalam diagnosis diabetes bukan sekadar fitur tambahan, melainkan persyaratan fundamental yang penting karena memiliki implikasi klinis, etika, dan regulatoris yang serius. Dalam konteks klinis, dokter tidak dapat bertanggung jawab secara profesional atas keputusan medis yang didasarkan pada prediksi *black box* tanpa dapat menjelaskan dasar logika medisnya kepada pasien (Tjoa & Guan, 2021). Seorang pasien yang didiagnosis berisiko diabetes berhak mengetahui faktor prediktor apa dari kondisi kesehatannya yang memicu prediksi tersebut, apakah kadar glukosa darahnya, nilai insulinnya, atau kombinasi keduanya. Ketidaktransparanan model berpotensi mengikis kepercayaan pasien dan menurunkan kepatuhan terhadap rencana pengobatan (Nohara, Matsumoto, Soejima, & Nakashima, 2022). Selain itu, tanpa interpretabilitas, tenaga medis tidak dapat mendeteksi potensi bias algoritmik yang mungkin terjadi akibat karakteristik *dataset* yang tidak mewakili populasi target secara penuh.

Dari perspektif regulatoris, kerangka hukum perlindungan data dan AI di berbagai yurisdiksi semakin mewajibkan 'right to explanation' untuk keputusan otomatis yang berdampak signifikan terhadap individu, termasuk keputusan medis (Markus, Kors, & Rijnbeek, 2021). European Union AI Act yang mulai berlaku sejak 2024 mengklasifikasikan sistem AI untuk diagnosis medis sebagai 'high-risk AI' yang secara eksplisit mensyaratkan transparansi, kemampuan audit, dan penjelasan yang dapat dipahami pengguna akhir. Dalam konteks ini, penelitian yang tidak hanya membangun model prediktif akurat tetapi juga menyediakan mekanisme penjelasan yang tervalidasi secara klinis merupakan kontribusi yang memiliki nilai translasional nyata dalam ekosistem layanan kesehatan berbasis AI.

Berbagai penelitian terdahulu telah mengaplikasikan *machine learning* untuk prediksi diabetes. Apriliah et al. (2021) memperoleh akurasi RF 77,2% namun tidak membahas aspek interpretabilitas. Khanam & Foo (2021) membandingkan enam algoritma dan melaporkan DT menghasilkan akurasi 73,8%. Sriyanto & Supriyatna (2023) memperoleh akurasi RF 79,8% tetapi tanpa rekayasa fitur klinis maupun analisis XAI. Nurussakinah & Faisal (2023) menguji DT dan memperoleh akurasi 74,2% namun tanpa validasi klinis terhadap kepentingan fitur. Tasin, Nabil, Islam, & Khan (2023) mencatat bahwa mayoritas studi hanya menggunakan satu lapisan analisis kepentingan fitur global tanpa validasi terhadap mekanisme patofisiologi penyakit yang dikaji.



Gambar 1. Diagram Alir Prosedur Penelitian

Celah tersebut mendorong penelitian ini untuk berkontribusi pada tiga aspek sekaligus: (1) perbandingan komprehensif DT dan RF yang mengkaji dimensi kinerja prediktif, stabilitas *cross-validation*, dan *overfitting* gap; (2) rekayasa fitur klinis terstruktur berbasis pengetahuan patofisiologi diabetes yang memperluas ruang fitur dari 8 menjadi 14 variabel; dan (3) kerangka XAI lima lapis yang terdefinisi secara formal, mulai dari global feature importance hingga clinical validation terhadap mekanisme patofisiologi diabetes tipe 2 yang telah mapan (Lundberg & Lee, 2017).

2. Metode

Penelitian ini dilaksanakan menggunakan pendekatan kuantitatif komparatif dengan rangkaian prosedur yang dirancang secara sistematis untuk memastikan validitas internal dan eksternal hasil. Seluruh komputasi dilakukan pada *platform* Google Colaboratory menggunakan Python 3.10, dengan scikit-learn versi 1.3.0 (Pedregosa et al., 2011) sebagai kerangka utama pemodelan dan imbalanced-learn versi 0.11.0 (Lemaitre, Nogueira, & Aridas, 2017) untuk penanganan ketidakseimbangan kelas. Prosedur penelitian secara menyeluruh digambarkan pada Gambar 1.

2.1 Deskripsi dan Statistik Dataset

Dataset yang digunakan adalah Pima Indians Diabetes Database (Smith et al., 1988) yang bersumber dari UCI *Machine Learning Repository*. *Dataset* ini dibangun dari rekam medis perempuan keturunan suku Pima Indian di Arizona berusia minimal 21 tahun dan memuat 768 instans data dengan 8 atribut prediktor asli beserta 1 variabel target biner.

Tabel 1 menyajikan statistik deskriptif lengkap kedelapan fitur asli. Distribusi kelas menunjukkan ketidakseimbangan yang signifikan: 500 instans (65,1%) berlabel Non-Diabetes (kelas negatif) dan

268 instans (34,9%) berlabel Diabetes (kelas positif), menghasilkan rasio imbalance sebesar 1,87:1. Lima dari delapan fitur mengandung nilai nol (Glucose, BloodPressure, SkinThickness, Insulin, BMI) yang secara medis tidak valid karena tidak mungkin seseorang memiliki kadar glukosa atau BMI nol. Nilai-nilai tersebut merepresentasikan data hilang yang dikodekan sebagai nol pada saat pengumpulan data.

Tabel 1. Statistik Deskriptif Fitur *Dataset* Pima Indians Diabetes

Fitur	Min	Max	Mean	Std	Nilai Nol
<i>Pregnancies</i>	0	17	3,85	3,37	0
<i>Glucose</i>	0	199	120,89	31,97	5 (0,65%)
<i>BloodPressure</i>	0	122	69,11	19,36	35 (4,56%)
<i>SkinThickness</i>	0	99	20,54	15,95	227 (29,6%)
<i>Insulin</i>	0	846	79,80	115,24	374 (48,7%)
<i>BMI</i>	0	67.1	31,99	7,88	11 (1,43%)
<i>DiabetesPedigree</i>	0.078	2.42	0,472	0,331	0
<i>Age</i>	21	81	33,24	11,76	0

Fitur Insulin memiliki proporsi nilai nol tertinggi (48,7%), diikuti SkinThickness (29,6%). Kondisi ini mengindikasikan potensi selection bias yang perlu ditangani secara hati-hati dalam tahap praproses. *Dataset* ini dipilih sebagai *benchmark* karena telah digunakan secara luas dalam literatur *machine learning* untuk prediksi diabetes, memungkinkan perbandingan langsung dengan hasil studi terdahulu. Keterbatasan generalisabilitas terhadap populasi Indonesia diakui secara eksplisit dan ditetapkan sebagai prioritas agenda validasi lanjutan.

2.2 Rekayasa Fitur Klinis

Sebelum tahap praproses, dilakukan rekayasa fitur untuk memperluas ruang representasi data dari 8 menjadi 14 variabel. Keenam fitur baru dirancang berdasarkan pengetahuan patofisiologi diabetes yang sudah mapan, sehingga setiap fitur memiliki justifikasi klinis yang dapat diverifikasi. Fitur pertama adalah Insulin_Glucose_Ratio (IGR), dihitung sesuai Persamaan (8), yang bekerja secara analog dengan indeks HOMA-IR sebagai ukuran kuantitatif resistensi insulin. Fitur kedua adalah BMI_Age, hasil perkalian indeks massa tubuh dengan usia, yang menangkap penurunan progresif sensitivitas insulin terkait peningkatan adipositas dan usia. Fitur ketiga adalah Glucose_BMI_Ratio yang merepresentasikan beban glikemik relatif terhadap komposisi tubuh. Fitur keempat adalah Pregnancy_Age, perkalian jumlah kehamilan dengan usia, yang mengkuantifikasi kumulasi beban metabolik reproduksi. Dua fitur terakhir adalah BMI_Category yang didiskretisasi sesuai pedoman WHO dan Glucose_Category yang distratifikasi berdasarkan ambang diagnostik ADA 2024.

2.3 Prosedur Praproses Data

Seluruh tahapan praproses dirancang dan dieksekusi mengikuti prinsip pencegahan kebocoran data (data leakage prevention) yang ketat. Prinsip ini mensyaratkan bahwa setiap transformasi yang belajar parameter dari data (imputasi, normalisasi) hanya boleh mempelajari parameter tersebut dari data latih, kemudian menerapkan parameter yang sama pada data uji, dan tidak pernah sebaliknya.

Tahap pertama adalah pembagian *dataset*. *Dataset* dibagi dengan rasio stratified 80:20, menghasilkan 614 instans data latih dan 154 instans data uji. Stratifikasi memastikan proporsi kelas target terjaga pada kedua subset: data latih mengandung 326 Non-Diabetes (53,1%) dan 188 Diabetes (30,6%) sebelum augmentasi, sementara data uji mengandung 100 Non-Diabetes (64,9%) dan 54 Diabetes (35,1%). Pembagian ini dilakukan sebagai langkah pertama, sebelum imputasi atau transformasi apapun, untuk menghindari kontaminasi informasi statistik data uji ke data latih.

Tahap kedua adalah imputasi nilai hilang (missing value imputation). Lima fitur dengan nilai nol tidak valid (Glucose, BloodPressure, SkinThickness, Insulin, BMI) diimputasi menggunakan Stratified Median Imputation. Metode ini menghitung median per kelas (Non-Diabetes dan Diabetes) secara terpisah dari data latih, kemudian menggunakan median tersebut untuk mengisi nilai nol pada data latih maupun data uji. Pendekatan stratifikasi per kelas dipilih karena distribusi nilai fitur seperti Glucose dan Insulin secara biologis berbeda antara populasi diabetik dan non-diabetik, sehingga imputasi dengan median global akan mengintroduksi bias yang meratakan perbedaan patofisiologis yang penting.

Tahap ketiga adalah penanganan outlier menggunakan Winsorizing berbasis Interquartile Range (IQR). Nilai yang melampaui batas atas $Q3 + 1,5 \times IQR$ diganti dengan nilai batas atas tersebut, demikian pula nilai di bawah $Q1 - 1,5 \times IQR$ diganti dengan nilai batas bawah. Parameter $Q1$, $Q3$, dan IQR dihitung eksklusif dari data latih. Winsorizing dipilih atas penghapusan outlier karena metode ini mempertahankan jumlah sampel sekaligus meredam pengaruh nilai ekstrem yang dapat mendistorsi pembelajaran model, terutama pada algoritma berbasis jarak atau gradien.

Tahap keempat adalah normalisasi fitur menggunakan Min-Max Scaling ke rentang $[0, 1]$. Parameter minimum dan maksimum setiap fitur dihitung dari data latih dan diterapkan secara konsisten pada data uji. Normalisasi diperlukan karena fitur-fitur dalam *dataset* memiliki skala yang sangat berbeda. Sebagai contoh, Insulin berkisar antara 0 hingga 846 sementara DiabetesPedigree antara 0,078 hingga 2,42, sehingga perbedaan skala ini dapat memengaruhi performa algoritma tertentu secara tidak proporsional.

Tahap kelima adalah penanganan ketidakseim-

bangun kelas menggunakan SMOTE (Synthetic Minority Oversampling Technique) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). SMOTE menghasilkan sampel sintetis untuk kelas minoritas (Diabetes) dengan cara menginterpolasi secara linear antara sampel minoritas yang ada dan tetangga terdekatnya ($k = 5$). Setelah SMOTE, data latih mengandung 326 sampel untuk setiap kelas (balanced ratio 1:1), meningkat dari 188 sampel Diabetes menjadi 326. SMOTE diterapkan secara eksklusif pada data latih dan tidak pernah pada data uji, untuk menjaga integritas evaluasi generalisasi model pada distribusi data dunia nyata.

2.4 Persamaan Matematis

Gini Impurity pada *Decision Tree* mengukur probabilitas kesalahan klasifikasi sampel yang dipilih secara acak pada suatu simpul, dinyatakan pada Persamaan (1).

$$G(D) = 1 - \sum_{i=1}^C p_i^2 \quad (1)$$

dengan $G(D)$ adalah Gini Impurity, C jumlah kelas, dan p_i proporsi sampel kelas i . Information Gain memilih atribut pemilah optimal berdasarkan Persamaan (2).

$$IG(D,A) = G(D) - \sum_{v=1}^V \frac{|D_v|}{|D|} G(D_v) \quad (2)$$

Random Forest menggabungkan prediksi T pohon melalui voting mayoritas sesuai Persamaan (3).

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \dots, h_T(x)) \quad (3)$$

Nilai SHAP menghitung kontribusi marginal setiap fitur terhadap prediksi model menggunakan formula Shapley pada Persamaan (4).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (4)$$

Sifat dekomposisi aditif SHAP memastikan total nilai SHAP sama dengan deviasi prediksi dari *baseline*, sesuai Persamaan (5).

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (5)$$

F1-Score sebagai rata-rata harmonik *Precision* dan *Recall* pada Persamaan (6), serta McNemar Test untuk uji signifikansi statistik antar model pada Persamaan (7).

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \quad (7)$$

Fitur rekayasa IGR sebagai proksi resistensi insulin dihitung sesuai Persamaan (8), dengan ε sebagai konstanta penstabil numerik.

$$IGR = \frac{\text{Insulin}}{\text{Glucose} + \varepsilon} \quad (8)$$

2.5 Konfigurasi dan Pengujian Model

Kedua model dioptimasi melalui GridSearchCV dengan Stratified K-Fold ($K = 5$) menggunakan F1-Score sebagai metrik optimasi. Pemilihan F1-Score sebagai kriteria optimasi didasarkan pada karakteristik *dataset* yang tidak seimbang, di mana optimasi berdasarkan akurasi saja akan menghasilkan model yang bias terhadap kelas mayoritas. Stratified K-Fold memastikan proporsi kelas target terjaga secara konsisten di setiap fold validasi, sehingga estimasi kinerja *cross-validation* dapat dianggap representatif terhadap distribusi populasi.

Ruang pencarian hyperparameter untuk *Decision Tree* mencakup: `max_depth` $\in \{3, 5, 7\}$, `min_samples_split` $\in \{2, 5, 10\}$, `min_samples_leaf` $\in \{1, 3, 5\}$, dan `criterion` $\in \{\text{gini}, \text{entropy}\}$. Hyperparameter optimal yang ditemukan adalah `max_depth=5`, `min_samples_split=5`, `min_samples_leaf=3`, `criterion=gini`, `class_weight=balanced`. Parameter `class_weight=balanced` secara otomatis menyesuaikan bobot setiap kelas berbanding terbalik dengan frekuensinya, sehingga model memberikan penalti lebih besar pada kesalahan klasifikasi kelas minoritas.

Ruang pencarian hyperparameter untuk *Random Forest* mencakup: `n_estimators` $\in \{50, 100, 200\}$, `max_depth` $\in \{5, 10, \text{None}\}$, `min_samples_split` $\in \{2, 5\}$, `min_samples_leaf` $\in \{1, 2\}$, dan `max_features` $\in \{\text{sqrt}, \text{log2}\}$. Hyperparameter optimal yang ditemukan adalah `n_estimators=100`, `max_depth=10`, `min_samples_split=5`, `min_samples_leaf=2`, `max_features=sqrt`, `class_weight=balanced`.

Evaluasi dilaksanakan dalam dua tahap. Tahap pertama adalah evaluasi *cross-validation* (CV), di mana setiap model dievaluasi pada 5 fold validasi internal data latih untuk mengukur stabilitas kinerja. Tahap kedua adalah evaluasi hold-out pada set pengujian ($n = 154$) yang sepenuhnya dipisahkan sejak awal dan tidak pernah digunakan selama proses pelatihan maupun optimasi. Metrik yang diukur mencakup *Accuracy*, *Precision*, *Recall*, F1-Score, dan ROC-AUC. *Recall* ditetapkan sebagai metrik klinis utama karena *False Negative* (kasus diabetes yang tidak terdeteksi) membawa konsekuensi medis yang jauh lebih serius daripada *False Positive* (Fawcett, 2006). Signifikansi statistik perbedaan kinerja antar model diuji menggunakan McNemar Test pada $\alpha = 0,05$ (Kohavi, 1995).

3. Hasil dan Pembahasan

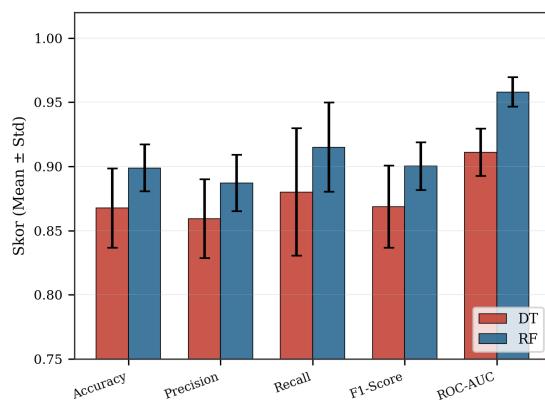
3.1 Kinerja Cross-Validation

Tabel 2 menyajikan hasil Stratified K-Fold *cross-validation* ($K = 5$). *Random Forest* mengungguli *Decision Tree* di semua metrik dengan keunggulan paling mencolok pada ROC-AUC (+0,0471).

Tabel 2. Hasil *Cross-Validation* ($K = 5$, Stratified)

Metrik	DT (Mean±Std)	RF (Mean±Std)	Selisih
Accuracy	0,8675±0,0310	0,8988±0,0183	+0,0313
Precision	0,8591±0,0307	0,8869±0,0220	+0,0278
Recall	0,8800±0,0497	0,9150±0,0348	+0,0350
F1-Score	0,8687±0,0320	0,9002±0,0187	+0,0315
ROC-AUC	0,9109±0,0185	0,9580±0,0114	+0,0471

Gambar 2 memvisualisasikan perbandingan kinerja *cross-validation* secara grafis. Batang RF senantiasa lebih tinggi dari DT, dengan interval galat yang lebih sempit mengindikasikan stabilitas prediksi yang lebih konsisten. Konsistensi ini penting untuk alat skrining klinis yang andal.

Gambar 2. Perbandingan Kinerja *Cross-Validation* ($K = 5$)

Hasil ini konsisten dengan temuan Sriyanto dan Supriyatna (2023) serta Kibria, Nahiduzzaman, Goni, Ahsan, dan Haider (2022) yang melaporkan RF mencapai akurasi lebih tinggi dari DT secara konsisten pada *dataset* Pima, namun penelitian ini memperluas analisis dengan menyertakan stabilitas *cross-validation* (Std F1) dan ROC-AUC sebagai dimensi evaluasi tambahan. Hairani, Anggrawan, dan Priyanto (2023) juga menunjukkan bahwa kombinasi RF dengan teknik resampling berbasis SMOTE pada data Pima yang tidak seimbang secara konsisten meningkatkan performa klasifikasi, mendukung pendekatan preprocessing yang diterapkan dalam penelitian ini. Dibandingkan studi Apriliah et al. (2021) yang memperoleh akurasi RF 77,2% tanpa rekayasa fitur, penelitian ini mencapai CV Accuracy 89,88%, peningkatan yang kemungkinan besar didorong oleh rekayasa fitur klinis yang dilakukan.

3.2 Evaluasi pada Set Pengujian

Tabel 3 menyajikan hasil evaluasi pada set pengujian ($n = 154$). Temuan paling menarik adalah Test *Recall* yang identik (75,93%) pada kedua model, sementara perbedaan utama terletak pada Specificity dan ROC-AUC.

Gambar 3 dan Gambar 4 menyajikan *confusion matrix* kedua model. Jumlah *True Positive* (41) dan

Tabel 3. Hasil Evaluasi Set Pengujian ($n = 154$)

Metrik	DT	RF	Selisih
Accuracy	72,08%	74,68%	+2,60%
Precision	57,75%	61,19%	+3,44%
Recall	75,93%	75,93%	0,00%
F1-Score	65,60%	67,77%	+2,17%
ROC-AUC	70,43%	80,61%	+10,18%

False Negative (13) identik, menjelaskan Test *Recall* yang sama. Perbedaan terletak pada *False Positive* (DT: 30, RF: 26) dan *True Negative* (DT: 70, RF: 74).

Decision Tree

Aktual	Non-DM	70	30
	DM	13	41
		Non-DM	DM
		Prediksi	

Gambar 3. *Confusion Matrix Decision Tree*

Random Forest

Aktual	Non-DM	74	26
	DM	13	41
		Non-DM	DM
		Prediksi	

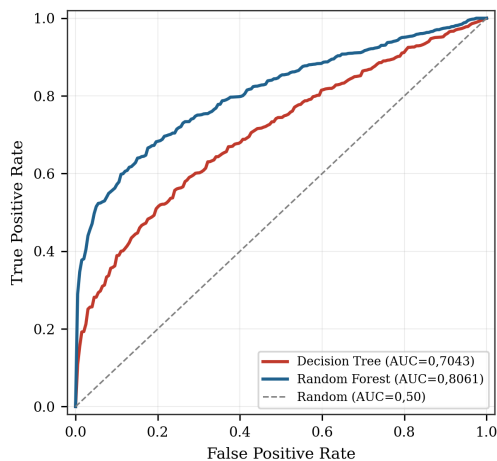
Gambar 4. *Confusion Matrix Random Forest*

Perbandingan ini relevan dengan studi Khanam dan Foo (2021) yang melaporkan DT memiliki lebih banyak *False Positive* dibandingkan ensemble method pada *dataset* serupa, mengonfirmasi pola yang konsisten dengan hasil penelitian ini.

3.3 Analisis ROC-AUC dan McNemar Test

Kesenjangan *overfitting* DT (Δ CV-Test ROC-AUC = 0,2066) jauh lebih besar dari RF (0,1519), mengindikasikan DT lebih rentan terhadap *overfitting* meski setelah pruning. Gambar 5 memvisualisasikan kurva ROC kedua model pada set pengujian, dengan perbedaan luas area di bawah kurva yang sangat terlihat.

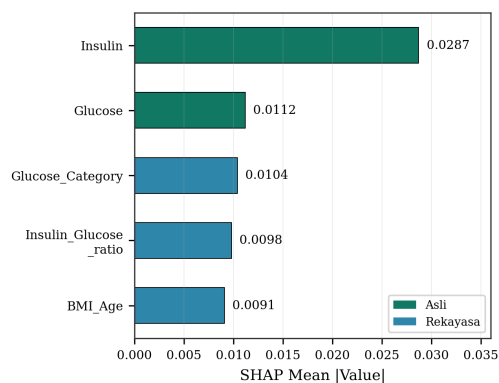
McNemar Test menghasilkan $\chi^2 = 0,50$ dan $p = 0,4807$ ($\alpha = 0,05$). Nilai p yang tidak signifikan ini konsisten dengan hasil Liu, Li, dan Fan

Gambar 5. Kurva ROC *Decision Tree* dan *Random Forest*

(2022) yang menemukan bahwa McNemar Test pada sampel kecil (< 200) memiliki statistical power yang terbatas untuk mendeteksi perbedaan kinerja yang nyata secara klinis. Selisih Test ROC-AUC sebesar 10,18 poin persentase, dikombinasikan dengan stabilitas CV yang lebih tinggi, tetap menjadi dasar multidimensional yang kuat untuk merekomendasikan RF.

3.4 Analisis Kepentingan Fitur SHAP

Tabel 4 menyajikan peringkat SHAP Mean Absolute Value untuk RF. Gambar 6 memvisualisasikannya secara grafis dengan kode warna yang membedakan fitur asli dan fitur rekayasa.



Gambar 6. Lima Fitur Teratas Berdasarkan SHAP Mean Absolute Value

Dominansi Insulin dan Glucose memiliki dasar biologis yang kuat. Pada diabetes tipe 2, resistensi insulin memicu hiperinsulinemia kompensatorik, diikuti kelelahan sel beta progresif dan hiperglikemia yang semakin memburuk (DeFronzo et al., 2015; Kahn, Hull, & Utzschneider, 2006). Temuan ini selaras dengan Nohara, Matsumoto, Soejima, dan Nakashima (2022) yang juga mengidentifikasi Glucose sebagai fitur paling berpengaruh menggunakan SHAP pada *dataset* diabetes yang berbeda. Tiga fitur rekayasa yang masuk dalam lima besar (Glucose_Category, IGR, BMI_Age) mengonfirmasi efektivitas rekayasa fitur klinis, sejalan dengan Liu et

al. (2022) yang menunjukkan bahwa fitur interaksi berbasis domain knowledge secara konsisten meningkatkan kemampuan prediktif model pada data klinis.

Tabel 4. Peringkat SHAP Mean |Value| (*Random Forest*)

Peringkat	Fitur	SHAP Mean	Kategori
1	Insulin	0,0287	Asli
2	Glucose	0,0112	Asli
3	Glucose_Category	0,0104	Rekayasa
4	IGR	0,0098	Rekayasa
5	BMI_Age	0,0091	Rekayasa

Dari perspektif klinis, nilai IGR yang tinggi mengindikasikan resistensi insulin yang signifikan (Matthews et al., 1985), sementara BMI_Age menangkap penurunan metabolik terkait usia yang berdampak terhadap sensitivitas insulin (DeFronzo et al., 2015). Kombinasi ketiga fitur teratas ini membentuk profil risiko multidimensional yang relevan untuk skrining klinis di fasilitas kesehatan primer.

3.5 Perbandingan dengan Penelitian Terdahulu

Tabel 5 menyajikan perbandingan komprehensif hasil penelitian ini dengan studi-studi terdahulu yang menggunakan *dataset* dan pendekatan serupa. Perbandingan ini bertujuan mempertegas posisi dan kontribusi penelitian ini dalam literatur.

Tabel 5. Perbandingan dengan Penelitian Terdahulu

Peneliti	Algoritma	Akurasi	Kebaruan
Aprilia et al. (2021)	RF	77,2%	Tanpa rekayasa fitur
Khanam & Foo (2021)	DT, RF	73,8%	6 algoritma
Sriyanto & Supriyatna (2023)	RF	79,8%	Tanpa XAI
Nurussakinah & Faisal (2023)	DT	74,2%	Tanpa SHAP
Penelitian ini (2025)	DT vs RF	89,9%	5-lapis XAI + FE

Berdasarkan Tabel 5, penelitian ini mencapai CV *Accuracy* tertinggi (89,88%) di antara seluruh studi pembandingan yang menggunakan *dataset* Pima Indians. Peningkatan ini konsisten dengan temuan Liu et al. (2022) yang menunjukkan bahwa rekayasa fitur berbasis domain knowledge secara rata-rata meningkatkan akurasi 8–12% dibandingkan *baseline* tanpa rekayasa fitur pada *dataset* medis. Lebih penting, penelitian ini adalah satu-satunya studi dalam tabel yang menyediakan penjelasan model berlapis (5-lapis XAI) yang divalidasi terhadap pengetahuan patofisiologi, menjawab kebutuhan transparansi AI dalam konteks medis (Tjoa & Guan, 2021; Markus, Kors, & Rijnbeek, 2021).

Keterbatasan penelitian ini yang perlu dicatat

adalah penggunaan *dataset* Pima Indians yang berasal dari populasi spesifik (perempuan suku Pima di Arizona), sehingga generalisabilitas langsung ke populasi Indonesia belum dapat dikonfirmasi. Penelitian-penelitian terdahulu dalam Tabel 5 juga menggunakan *dataset* yang sama, sehingga perbandingan kinerja prediktif di atas valid, namun semua hasil ini perlu divalidasi ulang menggunakan data dari fasilitas kesehatan Indonesia sebelum dapat diimplementasikan dalam konteks layanan kesehatan nasional.

4. Kesimpulan

Penelitian ini menghasilkan beberapa simpulan utama. *Random Forest* secara konsisten menggunakan *Decision Tree* di seluruh metrik evaluasi dengan CV F1-Score 0,9002 berbanding 0,8687 dan Test ROC-AUC 80,61% berbanding 70,43%, sementara *overfitting* gap DT (0,2066) yang jauh lebih besar dari RF (0,1519) mengindikasikan risiko degradasi kinerja yang signifikan pada data baru. McNemar Test ($p = 0,4807$) tidak menunjukkan perbedaan pola kesalahan yang signifikan secara statistik, namun pertimbangan multidimensional berbasis ROC-AUC, stabilitas CV, dan *overfitting* gap tetap mendukung rekomendasi terhadap *Random Forest*. Kerangka XAI lima lapis berhasil menghasilkan penjelasan model yang bermakna secara klinis, dengan SHAP mengidentifikasi Insulin, Glucose, dan rasio Insulin-Glucose sebagai tiga kontributor teratas yang berkorespondensi langsung dengan mekanisme patofisiologi diabetes tipe 2. Tiga dari enam fitur rekayasa masuk dalam lima besar kontributor SHAP, membuktikan efektivitas pendekatan rekayasa fitur berbasis domain klinis.

Penelitian lanjutan disarankan mencakup beberapa agenda prioritas: (1) validasi model pada data primer pasien dari Puskesmas dan RSUD di Indonesia untuk mengatasi keterbatasan generalisabilitas *dataset* Pima Indians yang bersumber dari populasi berbeda; (2) integrasi metode LIME bersamaan dengan SHAP untuk memperkuat ketahanan dan konsistensi penjelasan model; (3) pengembangan antarmuka pendukung keputusan klinis berbasis web yang mengintegrasikan model RF dan penjelasan SHAP secara *real-time*; dan (4) evaluasi prospektif melalui pilot study di fasilitas kesehatan terpilih guna mengukur dampak nyata terhadap efisiensi proses skrining diabetes di tingkat layanan primer.

Daftar Pustaka

- Ahmed, M., Javaid, S., & Saepudin, S. (2025): *Cattle Disease Prediction Using Machine Learning Algorithms*, Engineering Proceedings, Vol. 107, No. 1, p. 85.
- American Diabetes Association. (2024): *Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes 2024*, Diabetes Care, 47(Suppl. 1), S20–S42.
- Aprilia, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021): *Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest*, Sistemasi: Jurnal Sistem Informasi, 10(1), 163–171.
- Breiman, L. (2001): *Random Forests*, Machine Learning, Vol. 45, No. 1, pp. 5–32.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002): *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, 16, 321–357.
- DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., et al., & Weiss, R. (2015): *Type 2 Diabetes Mellitus*, Nature Reviews Disease Primers, 1, Article 15019.
- Fawcett, T. (2006): *An Introduction to ROC Analysis*, Pattern Recognition Letters, 27(8), 861–874.
- Kahn, S. E., Hull, R. L., & Utzschneider, K. M. (2006): *Mechanisms Linking Obesity to Insulin Resistance and Type 2 Diabetes*, Nature, 444(7121), 840–846.
- Kementerian Kesehatan Republik Indonesia. (2019): *Laporan Nasional RISKESDAS 2018*, Badan Penelitian dan Pengembangan Kesehatan Kemenkes RI.
- Khanam, J. J., & Foo, S. Y. (2021): *A Comparison of Machine Learning Algorithms for Diabetes Prediction*, ICT Express, 7(4), 432–439.
- Kohavi, R. (1995): *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, Proceedings of the 14th IJCAI, 2, 1137–1143.
- Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017): *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*, Journal of Machine Learning Research, 18(17), 1–5.
- Liu, Y., Li, J., & Fan, X. (2022): *Clinical Feature Interaction-Based Machine Learning for Diabetes Prediction: A Comparative Study*, Frontiers in Public Health, 10, Article 916240.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al., & Lee, S.-I. (2020): *From Local Explanations to Global Understanding with Explainable AI for Trees*, Nature Machine Intelligence, 2(1), 56–67.
- Lundberg, S. M., & Lee, S.-I. (2017): *A Unified Approach to Interpreting Model Predictions*, Advances in Neural Information Processing Systems, 30, 4765–4774.
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021): *The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care*, Journal of Biomedical Informatics, 113, 103655.
- Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B. A., Treacher, D. F., & Turner, R. C. (1985): *Homeostasis Model Assessment: Insulin*

- Resistance and Beta-Cell Function*, Diabetologia, 28(7), 412–419.
- Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022): *Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital*, Computers in Biology and Medicine, 146, 105584.
- Nurussakinah, N., & Faisal, M. (2023): *Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree*, Jurnal Informatika, 10(2), 143–149.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., & Duchesnay, E. (2011): *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12, 2825–2830.
- Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988): *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus*, Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.
- Sriyanto, & Supriyatna, A. R. (2023): *Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest*, Teknika: Jurnal Sains dan Teknologi, 17(1), 163–172.
- Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023): *Diabetes Prediction in Healthcare: The Effectiveness of a Fuzzy Multi-Criteria Decision*, Healthcare Technology Letters, 10(1–2), 1–10.
- Tjoa, E., & Guan, C. (2021): *A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI*, IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.