https://jurnal.pustakagalerimandiri.co.id/index.php/pustakaai
DOI : https://doi.org/10.55382/jurnalpustakaai.v6i1.1751

JURNAL PUSTAKA AI
JURNAL PUSAT AKSES KAJIAN TEKNOLOGI ARTIFICIAL INTELLIGENCE
Pustaka Galeri Mandiri
Vol. 6 No. 1 (2026) 12 – 24
E ISSN: 2809-4069

Evaluasi Model *Machine Learning* Prediksi Diabetes Menggunakan *Nested Cross-Validation* dan SHAP Terintegrasi

Errie Tri Armawan[1], Riana Safitri[2], Lutvi Riyandari[3]
[1,2,3]Teknik Informatika, STMIK Widya Utama Purwokerto
[1]errietri415@gmail.com . [2]rianasafitri@swu.ac.id . [3]lutvi@swu.ac.id

Abstract

This study evaluates and compares the performance of three machine learning algorithms Logistic Regression, Random Forest and XGBoost for diabetes prediction using nested cross-validation (5-fold outer, 3-fold inner) with an integrated preprocessing pipeline to prevent data leakage. The Pima Indians Diabetes Dataset (n = 768) was used. The Friedman test (p = 0.819) confirmed no statistically significant difference among the three models. Thus, Logistic Regression was selected based on the parsimony principle to achieve the highest stability (AUC-ROC 72.3% ± 1.6% in nested cross-validation) and precision of 74.8% ± 6.8%. On the independent test set, the model achieved accuracy of 69.5%, AUC-ROC of 81.4% and PR-AUC of 65.9%. Threshold analysis revealed that lowering the decision threshold to 0.30 increased recall to 83.3%, with a 95% bootstrap confidence interval for AUC-ROC of [0.737; 0.876] on the test set. SHAP analysis identified Glucose, BMI and DiabetesPedigreeFunction as the top three predictors, consistent with diagnostic criteria and risk factors in the American Diabetes Association and World Health Organization guidelines. This alignment demonstrates that the model learns clinically meaningful patterns rather than mere statistical correlations.

Keywords: machine learning, diabetes prediction, nested cross-validation, SHAP, interpretability

Abstrak

Penelitian ini mengevaluasi dan membandingkan kinerja tiga algoritma *machine learning Logistic Regression*, *Random Forest*, dan *XGBoost* untuk prediksi diabetes menggunakan *nested cross-validation* (5-*fold outer*, 3-*fold inner*) dengan *pipeline* preprocessing terintegrasi untuk mencegah *data leakage*. Dataset yang digunakan adalah *Pima Indians Diabetes Dataset* (n = 768). Uji Friedman (p = 0,819) mengonfirmasi bahwa ketiga model tidak berbeda secara statistik, sehingga *Logistic Regression* dipilih berdasarkan prinsip parsimoni dengan stabilitas tertinggi (*AUC-ROC* 72,3% ± 1,6% pada *nested cross-validation*) dan *precision* 74,8% ± 6,8%. Pada data uji independen, model mencapai akurasi 69,5%, *AUC-ROC* 81,4%, dan *PR-AUC* 65,9%. Analisis threshold menunjukkan bahwa penurunan threshold ke 0,30 meningkatkan recall menjadi 83,3%, dengan 95% confidence interval AUC-ROC = [0,737; 0,876] pada data uji independen. Analisis SHAP mengidentifikasi *Glucose*, BMI, dan *DiabetesPedigreeFunction* sebagai tiga prediktor teratas, yang selaras dengan kriteria diagnostik dan faktor risiko dalam pedoman American Diabetes Association dan World Health Organization. Kesesuaian ini membuktikan bahwa model mempelajari pola yang bermakna secara klinis, bukan sekadar korelasi statistik.

Kata kunci: machine learning, prediksi diabetes, nested cross-validation, SHAP, interpretabilitas

Submitted : 15-03-2026 | Reviewed : 25-03-2026 | Accepted : 31-03-2026

1. Pendahuluan

Penerapan *machine learning* untuk prediksi diabetes sudah cukup luas dilakukan [1]-[4], namun banyak penelitian justru melaporkan hasil yang terkesan terlalu bagus. Hal ini bukan karena modelnya benar-benar sempurna, melainkan karena ada celah metodologi yang sering terlewatkan terutama masalah kebocoran data (*data leakage*) dan skema evaluasi yang kurang ketat. Kondisi ini menjadi hambatan nyata ketika model ingin diterapkan di lingkungan klinis, mengingat kejelasan proses dan kemampuan interpretasi model sama pentingnya dengan angka akurasi yang dihasilkan [5]. Dengan jumlah penderita diabetes yang sudah melewati 589 juta jiwa di seluruh dunia dan sekitar 250 juta kasus yang bahkan belum terdiagnosis [6], kebutuhan akan model prediksi yang benar-benar andal dan dapat dipercaya menjadi semakin mendesak terutama untuk mendukung deteksi dini dan mencegah komplikasi seperti kerusakan penglihatan, gangguan fungsi ginjal, hingga penyakit jantung [7].

Bila ditelaah lebih dalam, penelitian-penelitian yang sudah ada [1]-[4] menunjukkan tiga pola kelemahan yang berulang dan secara konsisten membuat angka performa tampak lebih tinggi dari yang sebenarnya. Kelemahan pertama adalah penggunaan *k-fold cross-validation* biasa tanpa pemisahan bertingkat untuk proses pemilihan *hyperparameter* [1]-[4], [8]. Pendekatan ini menciptakan situasi di mana data yang sama dipakai sekaligus untuk menyetel parameter model sekaligus menilai performanya. Akibatnya, angka akurasi yang dilaporkan cenderung lebih tinggi dari yang seharusnya karena ada kebocoran di level metodologi [9]. Studi sebelumnya juga menunjukkan bahwa tidak adanya pemisahan tegas antara tahap pemilihan model dan tahap penilaian akhir meningkatkan risiko model yang terlalu disesuaikan dengan data latih [10].

Kelemahan kedua berkaitan dengan urutan tahapan pra-pemrosesan. Beberapa penelitian melakukan imputasi dan normalisasi data sebelum proses pembagian data untuk *cross-validation* [1]-[3]. Ini secara teknis menyebabkan kebocoran informasi, di mana statistik dari data validasi sudah ikut memengaruhi proses pelatihan model sejak awal [9]. Hasilnya adalah estimasi performa yang tidak mencerminkan kondisi data baru yang sesungguhnya, sehingga model sulit digeneralisasi ke situasi klinis yang sebenarnya.

Kelemahan ketiga ada pada sisi interpretabilitas. Sejumlah penelitian memang sudah menggunakan berbagai pendekatan interpretabilitas untuk memahami kontribusi fitur dalam prediksi diabetes, mulai dari SHAP [11], berbagai metode XAI termasuk knowledge graph [12], LIME dan analisis fitur lanjutan [13]-[15], serta teknik lanjutan seperti counterfactual analysis dan integrated gradients [11]. Pada studi yang menggunakan dataset serupa, analisis SHAP menunjukkan bahwa glukosa dan BMI merupakan prediktor dominan [11], sementara studi pada dataset lain mengidentifikasi faktor risiko yang berbeda sesuai karakteristik datanya [13], [15]. Meskipun demikian, evaluasi formal dan sistematis untuk memverifikasi apakah peringkat *feature importance* tersebut sesuai dengan hierarki faktor risiko dalam pedoman klinis resmi WHO dan ADA belum dilakukan. Celah inilah yang pada akhirnya membuat kepercayaan terhadap model sulit dibangun di lingkungan klinis, karena klinisi membutuhkan penjelasan yang tidak hanya akurat secara statistik, tetapi juga bermakna secara medis dan terverifikasi terhadap standar yang telah mapan.

Studi-studi yang ada umumnya menerapkan pendekatan-pendekatan tersebut secara terpisah ada yang sudah memperhatikan interpretabilitas dengan berbagai metode XAI [11]-[15], ada yang menyoroti pentingnya evaluasi yang lebih ketat [9], [10] namun belum ada yang mengintegrasikan *nested cross-validation*, *pipeline* anti-kebocoran data, dan validasi interpretabilitas SHAP terhadap pedoman klinis WHO dan ADA dalam satu kerangka yang utuh. Khususnya pada dataset Pima Indians, meskipun interpretabilitas telah diterapkan menggunakan SHAP [11] maupun knowledge graph [12], aspek validasi metodologis yang ketat dan validasi klinis formal terhadap pedoman WHO dan ADA belum diintegrasikan secara bersamaan.

Penelitian ini bertujuan mengevaluasi dan membandingkan kinerja tiga algoritma *machine learning Logistic Regression*, *Random Forest*, dan *XGBoost* dalam memprediksi diabetes menggunakan kerangka validasi yang lebih ketat, sekaligus menganalisis kesesuaian interpretabilitas model terhadap standar klinis WHO dan ADA. Untuk menjawab tujuan tersebut, penelitian ini dirancang dengan tiga pendekatan utama: (1) menerapkan *nested cross-validation* (5-*fold outer*, 3-*fold inner*) yang secara tegas memisahkan proses penyetelan *hyperparameter* dari tahap penilaian performa; (2) mengintegrasikan seluruh tahap pra-pemrosesan ke dalam *pipeline* yang berjalan secara independen di setiap *fold*, sehingga data validasi tidak terlibat dalam proses pelatihan; dan (3) memvalidasi hasil interpretabilitas model secara sistematis terhadap faktor risiko klinis yang tercantum dalam pedoman WHO dan ADA menggunakan analisis SHAP [16].

Dengan pendekatan ini, penelitian berharap dapat menghasilkan estimasi performa yang lebih konservatif namun lebih dapat diandalkan, dibandingkan studi-studi sebelumnya yang menggunakan skema validasi sederhana [1]-[4].

Berdasarkan uraian di atas, kontribusi utama penelitian ini dapat dirumuskan secara eksplisit sebagai berikut. Pertama, penelitian ini mengintegrasikan tiga komponen metodologis yang selama ini diterapkan secara terpisah dalam literatur prediksi diabetes berbasis dataset Pima Indians, yaitu: (1) *nested cross-validation* dengan *pipeline*

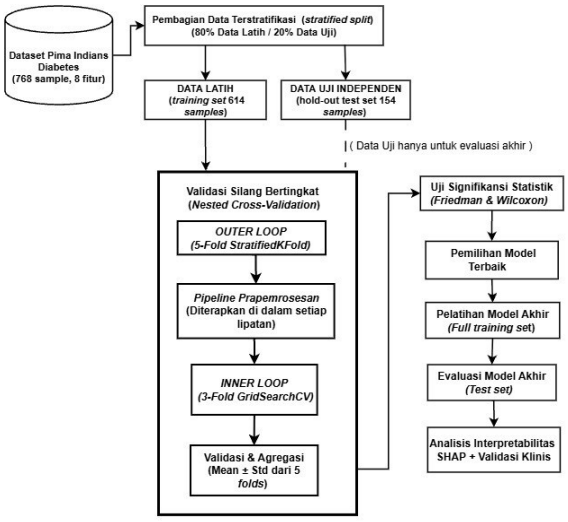
Submitted : 15-03-2026 | Reviewed : 25-03-2026 | Accepted : 31-03-2026

anti-kebocoran data, (2) uji signifikansi statistik antar model menggunakan Friedman dan Wilcoxon dengan koreksi Bonferroni, dan (3) validasi *feature importance* SHAP secara sistematis terhadap pedoman klinis ADA dan WHO semuanya dalam satu kerangka metodologis yang utuh. Kedua, penelitian ini menghasilkan estimasi performa yang konservatif sekaligus membuktikan bahwa model yang dipilih tidak hanya valid secara statistik tetapi juga bermakna secara medis. Integrasi ketiga aspek tersebut secara simultan, khususnya pada dataset Pima Indians, merupakan kontribusi yang belum dijumpai pada penelitian-penelitian sebelumnya.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan eksperimental untuk mengevaluasi kinerja dan interpretabilitas tiga algoritma *machine learning*: *Logistic Regression*, *Random Forest*, dan *XGBoost* dalam prediksi diabetes melitus. Implementasi dilakukan menggunakan Python 3.10.12 dengan pustaka scikit-learn 1.3.2 [17], *XGBoost* 2.0.1, dan SHAP 0.43.0 [16] pada platform Google Colaboratory. Rancangan penelitian menerapkan *nested cross-validation* untuk memperoleh estimasi performa yang tidak bias [10] serta analisis SHAP yang divalidasi terhadap pedoman klinis WHO [18], sehingga langsung mengatasi masalah optimistic bias dan *data leakage* [9].

Gambar 1 mengilustrasikan alur metodologi penelitian yang dirancang untuk mengatasi kelemahan metodologis pada penelitian sebelumnya. Tahapannya meliputi pembagian data secara terstratifikasi dengan rasio 80:20 (n = 614 data latih dan n = 154 data uji) untuk memastikan representasi kelas diabetes yang proporsional. *Nested cross-validation* dengan *preprocessing pipeline* terintegrasi diterapkan untuk mencegah *data leakage*. Pengujian signifikansi statistik menggunakan metode *Friedman* dan Wilcoxon dilakukan untuk membandingkan performa model secara objektif. Model terbaik dipilih berdasarkan prinsip parsimoni dan interpretabilitas, lalu dievaluasi pada data uji independen untuk memvalidasi kemampuan generalisasinya. Terakhir, analisis interpretabilitas menggunakan SHAP divalidasi terhadap pedoman klinis WHO dan ADA untuk memastikan model mempelajari pola yang bermakna secara klinis. Setiap tahapan dirancang untuk menghasilkan estimasi performa yang konservatif namun dapat dipercaya.



Gambar 1. Kerangka Kerja Metodologi Penelitian: Pembagian Data Terstratifikasi, Validasi Silang Bertingkat, Uji Signifikansi Statistik, dan Analisis Interpretabilitas

2.1. Dataset dan Pembagian Data

Penelitian ini menggunakan *benchmark* dataset Pima Indians Diabetes yang diperoleh dari UCI *Machine Learning* Repository [19], yang telah banyak digunakan untuk mengevaluasi metode prediksi diabetes. Dataset ini dikumpulkan oleh National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dan mencakup data dari 768 pasien perempuan keturunan Pima Indian berusia ≥ 21 tahun. Dataset ini telah digunakan dalam sejumlah penelitian prediksi diabetes, baik pada skala internasional [8] maupun nasional [1]-[4], sehingga memungkinkan perbandingan hasil penelitian secara langsung. Selain itu, dataset ini menyediakan dokumentasi fitur klinis yang lengkap dan terstandarisasi serta tersedia secara publik, yang mendukung transparansi dan keterbukaan proses penelitian.

Pengukuran glukosa dalam dataset ini didasarkan pada nilai glukosa plasma dua jam pasca *oral glucose tolerance test* (2-hour post-OGTT), yang sesuai dengan kriteria diagnosis diabetes menurut American Diabetes Association

(ADA), yaitu ≥ 200 mg/dL. Dataset terdiri dari delapan fitur prediktor klinis dan satu variabel target biner (*Outcome*), dengan nilai 0 untuk individu non-diabetes dan nilai 1 untuk individu dengan diabetes. Distribusi kelas menunjukkan 268 kasus diabetes (34,9%) dan 500 kasus non-diabetes (65,1%). Rincian masing-masing variabel penelitian disajikan pada Tabel 1.

Tabel 1. Deskripsi Variabel dan Statistik Deskriptif Dataset

No	Variabel	Satuan	Mean ± SD	*Range*	Keterangan
1	Pregnancies	Count	3.8 ± 3.4	0 – 17	Jumlah kehamilan
2	Glucose	mg/dL	120.9 ±32.0	0 – 199	Kadar glukosa plasma 2 jam setelah tes toleransi glukosa
3	BloodPressure	mmHg	69.1 ± 19.4	0 – 122	Tekanan darah diastolik
4	SkinThickness	mm	20.5 ± 16.0	0 – 99	Ketebalan *fold* kulit trisep
5	Insulin	µU/mL	79.8 ± 115.2	0 – 846	Kadar insulin serum 2 jam setelah tes toleransi glukosa
6	BMI	kg/m²	32.0 ± 7.9	0.0 –67.1	Indeks massa tubuh
7	DiabetesPedigreeFunction	-	0.5 ± 0.3	0.1 – 2.4	Skor risiko genetik berdasarkan riwayat keluarga
8	Age	Tahun	33.2 ± 11.8	21 – 81	Usia pasien
9	Outcome	Biner	-	-	Kelas target (0 = *Non-diabetes*, 1 = *Diabetes*)

Berdasarkan *domain knowledge* fisiologis, nilai nol pada lima fitur *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, dan *BMI* ditandai sebagai *missing values* mengikuti praktik standar pada dataset Pima Indians [20], karena nilai-nilai tersebut tidak mungkin terjadi secara biologis pada individu hidup (Tabel 2). Keputusan ini didasarkan pada konsensus literatur medis, bukan pada hasil analisis eksploratif terhadap dataset penelitian, untuk menghindari potensi kebocoran data. Proporsi *missing values* berkisar dari 0,7% pada fitur *Glucose* hingga 48,7% pada fitur *Insulin*. Seluruh fitur tetap dipertahankan sesuai dengan standar delapan fitur klinis pada dataset Pima Indians [1]-[4].

Tabel 2. Justifikasi Fisiologis dan Proporsi *Missing Values*

Variabel	Alasan Fisiologis	Nilai Minimum Fisiologis	*Missing (%)*
Glucose	Hipoglikemia berat terjadi pada <40 mg/dL; nilai 0 mg/dL mengindikasikan kematian	~40 mg/dL	5 (0.7%)
BloodPressure	Tekanan darah 0 mmHg = *cardiac arrest*, tidak kompatibel dengan kehidupan	~60/40 mmHg	35 (4.6%)
BMI	BMI = berat/(tinggi²); nilai 0 hanya mungkin jika berat=0 atau tinggi=0 (matematis *impossible*)	~10 kg/m²	11 (1.4%)
Insulin	Bahkan pada pankreatektomi total, insulin serum minimal terdeteksi pada level rendah	~2 µU/mL	374 (48.7%)
SkinThickness	Ketebalan epidermis manusia minimal 0.05 mm; nilai 0 = *measurement error*	~0.5 mm	227 (29.6%)

Imputasi *missing values* dilakukan menggunakan KNN *Imputer* (k = 5) yang diintegrasikan ke dalam *pipeline* pra-pemrosesan. Pemilihan k = 5 mempertimbangkan keseimbangan antara bias dan variansi pada data biomedis serta efisiensi komputasi [10].

Dataset dibagi menggunakan metode *stratified random split* dengan rasio 80:20 dan *random_state* = 42 menjadi data latih (n = 614) untuk proses *nested cross-validation* dan data uji (n = 154) sebagai *hold-out set* independen. Stratifikasi mempertahankan proporsi kelas yang konsisten antara data latih (34,9%) dan data uji (35,1%), sehingga keterwakilan distribusi kelas terjaga pada kedua subset.

2.2. Nested Cross Validation

Seperti yang telah diuraikan pada subbab sebelumnya, seluruh data latih (n = 614) digunakan dalam skema *nested cross-validation* untuk menghasilkan estimasi performa yang tidak bias. Metode ini diterapkan untuk mencegah bias optimistis yang muncul ketika data validasi yang sama digunakan sekaligus untuk *hyperparameter tuning* dan evaluasi performa model (*meta-overfitting*) [10]. Caranya adalah dengan memisahkan tahap pemilihan model

(*model selection*) dan tahap penilaian performa (*model assessment*) secara tegas melalui dua tingkat evaluasi yang berjenjang.

Outer loop dirancang menggunakan *stratified* 5-*fold cross-validation* pada data latih (n = 614). Pada setiap iterasi, data dibagi menjadi *outer training set* (empat *fold*) dan *outer validation set* (satu *fold*). *Outer validation set* hanya digunakan untuk evaluasi performa akhir dan sama sekali tidak terlibat dalam proses pemilihan *hyperparameter*, sehingga estimasi performa yang dihasilkan benar-benar independen dan tidak bias.

Inner loop berjalan di dalam setiap *outer training set*. Proses *hyperparameter tuning* dilakukan menggunakan *GridSearchCV* dengan 3-*fold cross-validation*, di mana *outer training set* dipecah lagi menjadi tiga bagian untuk mencari kombinasi *hyperparameter* terbaik dari ruang pencarian yang telah ditentukan (Tabel 3). Konfigurasi optimal dipilih berdasarkan nilai *mean AUC-ROC* tertinggi. Setelah *hyperparameter* terbaik ditemukan, model dilatih ulang menggunakan seluruh *outer training set*, kemudian dievaluasi pada *outer validation set* yang sebelumnya tidak tersentuh. Prosedur ini diulang untuk setiap *outer fold*, sehingga diperoleh estimasi performa akhir dalam bentuk *mean* ± SD. Pemilihan konfigurasi 5-*fold* pada *outer loop* dan 3-*fold* pada *inner loop* didasarkan pada pertimbangan keseimbangan bias-variansi dan efisiensi komputasi, mengingat ruang pencarian mencakup 167 kombinasi *hyperparameter* (Tabel 3).

Untuk mencegah kebocoran data, seluruh tahapan pra-pemrosesan diintegrasikan ke dalam *pipeline* scikit-learn [17] yang dieksekusi secara independen pada setiap *outer fold*. *Pipeline* terdiri dari tiga tahap berurutan: imputasi menggunakan *KNNImputer* (k = 5), standarisasi fitur dengan *StandardScaler*, dan klasifikasi menggunakan salah satu algoritma yang dievaluasi (*Logistic Regression*, *Random Forest*, atau *XGBoost*). *Pipeline* di-*fit* hanya pada *outer training set*, sementara *outer validation set* ditransformasikan menggunakan parameter yang sama tanpa *re-fitting*. Pendekatan ini memastikan bahwa informasi dari data validasi tidak bocor ke proses pelatihan pada tahap manapun.

Metrik AUC-ROC dipilih sebagai metrik optimisasi karena relatif tidak terpengaruh oleh ketidakseimbangan kelas [21], sehingga evaluasi performa tetap stabil meskipun distribusi kelas tidak seimbang. Selain itu, *AUC-ROC* mengevaluasi performa model pada seluruh *decision threshold*, yang memungkinkan penyesuaian ambang keputusan sesuai konteks klinis tanpa perlu melatih ulang model.

Tabel 3. *Hyperparameter Search Space* untuk *Grid Search*

Model	Hyperparameter	Nilai yang Diuji
Logistic Regression	C	{0.01, 0.1, 1, 10, 100}
	penalty	{l2}
	solver	{lbfgs}
Random Forest	n_estimators	{50, 100, 200}
	max_depth	{5, 10, 15, None}
	min_samples_split	{2, 5, 10}
	min_samples_leaf	{1, 2, 4}
XGBoost	n_estimators	{50, 100, 200}
	max_depth	{3, 5, 7}
	learning_rate	{0.01, 0.1, 0.3}
	subsample	{0.8, 1.0}

2.3. Algoritma Machine Learning

Penelitian ini membandingkan tiga algoritma dengan tingkat kompleksitas yang berbeda untuk mendapatkan gambaran performa yang komprehensif. Pertama, *Logistic Regression* sebagai model linear *baseline* yang mudah diinterpretasikan. Kedua, *Random Forest* sebagai metode *ensemble* yang tangguh terhadap *overfitting*. Ketiga, *XGBoost* sebagai metode *gradient boosting* yang lebih canggih untuk data tabular. Ketiga algoritma diimplementasikan menggunakan scikit-learn versi 1.3.2 untuk *Logistic Regression* dan *Random Forest*, serta *XGBoost* versi 2.0.1, dengan *random_state* = 42 untuk memastikan hasil eksperimen yang konsisten. Ruang pencarian *hyperparameter* untuk masing-masing algoritma disajikan pada Tabel 3.

Penelitian ini tidak menggunakan SMOTE karena tiga pertimbangan. Pertama, *imbalance ratio* sebesar 1:2 tergolong sedang dan memerlukan pertimbangan cermat terhadap *trade-off* antara performa dan validitas klinis. Kedua, sampel sintetis SMOTE dapat bersifat ambigu, tidak merepresentasikan variabilitas pasien di dunia nyata, serta berpotensi menghasilkan *feature importance* yang bertentangan dengan pengetahuan medis [22], [23]. Ketiga, untuk memfasilitasi perbandingan yang adil dengan penelitian sebelumnya yang menggunakan distribusi

kelas natural dari dataset yang sama [1]-[4]. Sebagai gantinya, model dievaluasi menggunakan metrik yang komprehensif dan sensitif terhadap ketidakseimbangan kelas, yaitu *Precision*, *Recall*, *F1-score*, dan *PR-AUC*, yang penjelasannya diuraikan lebih lanjut pada Subbab 2.4.

2.4. Metrik Evaluasi

Sesuai dengan yang disebutkan pada Subbab 2.3, model dievaluasi menggunakan beberapa metrik yang komprehensif untuk memberikan gambaran kinerja yang lebih lengkap dan tidak menyesatkan pada dataset dengan ketidakseimbangan kelas. Selain *AUC-ROC* yang digunakan sebagai metrik optimisasi pada proses *hyperparameter tuning*, penelitian ini melaporkan lima metrik tambahan.

Akurasi mengukur proporsi prediksi yang benar secara keseluruhan, namun dapat menyesatkan pada dataset dengan ketidakseimbangan kelas karena cenderung condong ke kelas mayoritas. *Precision* mengukur proporsi prediksi diabetes yang benar metrik ini penting untuk menghindari *false alarm* dan pemeriksaan diagnostik yang tidak perlu. *Recall* mengukur proporsi kasus diabetes yang berhasil terdeteksi metrik ini krusial untuk memastikan intervensi dini pada pasien berisiko. *F1-score* merupakan rata-rata harmonik dari *Precision* dan *Recall* yang menyeimbangkan kedua aspek tersebut. Terakhir, *PR-AUC* mengukur *trade-off* antara *Precision* dan *Recall* pada berbagai *decision threshold* dan lebih sensitif terhadap kelas minoritas dibandingkan *AUC-ROC* [21].

Dalam konteks skrining diabetes, *trade-off* antara *Precision* dan *Recall* perlu dipertimbangkan secara cermat berdasarkan prioritas klinis [18]. *Recall* yang tinggi penting untuk mencegah komplikasi jangka panjang melalui deteksi dini, sementara *Precision* yang rendah dapat menyebabkan *overdiagnosis* dan meningkatkan kecemasan pasien yang tidak perlu. Oleh karena itu, penelitian ini melaporkan seluruh metrik di atas untuk memfasilitasi pengambilan keputusan klinis yang lebih tepat sesuai konteks skrining.

2.5 Uji Signifikansi Statistik

Untuk menentukan apakah perbedaan performa antar model signifikan secara statistik, penelitian ini menggunakan uji statistik non-parametrik sesuai dengan praktik terbaik [10]. *Friedman test* digunakan sebagai *omnibus test* untuk membandingkan performa ketiga model berdasarkan lima skor *AUC-ROC* dari *outer folds* pada *nested cross-validation*. *Friedman test* dipilih karena ukuran sampel yang kecil (n = 5 *folds*) tidak memungkinkan validasi asumsi normalitas secara memadai, dan skor *cross-validation* memang tidak dapat diasumsikan berdistribusi normal sejak awal.

Apabila *Friedman test* menunjukkan hasil yang signifikan ($p < 0,05$), dilakukan perbandingan berpasangan menggunakan *Wilcoxon signed-rank test* dengan koreksi Bonferroni ($\alpha\_corrected = 0,0167$ untuk tiga perbandingan: LR vs RF, LR vs XGB, dan RF vs XGB). Selain signifikansi statistik, penelitian ini juga melaporkan besaran perbedaan performa (*mean difference* ± SD) untuk memberikan interpretasi yang lebih praktis. Perlu dicatat bahwa signifikansi statistik tidak selalu berarti *practical significance*. Apabila perbedaan performa antar model sangat kecil (*mean difference* < 1%), pemilihan model juga mempertimbangkan aspek interpretabilitas, efisiensi komputasi, dan kegunaan klinis [10]. Dengan demikian, model yang terpilih tetap layak untuk penerapan klinis.

2.6 Analisis Interpretabilitas SHAP

Untuk memahami kontribusi setiap fitur terhadap prediksi model, penelitian ini menggunakan SHAP (*SHapley Additive exPlanations*) [16]. Pendekatan SHAP telah banyak diterapkan dalam studi prediksi diabetes untuk memberikan penjelasan yang bermakna secara klinis [11], [13]-[15], menjadikannya pilihan yang tepat untuk tujuan validasi klinis dalam penelitian ini. Analisis SHAP dilakukan pada model terbaik yang dilatih ulang menggunakan seluruh data latih (n = 614) dengan *library* SHAP versi 0.43.0.

Global feature importance dihitung dengan merata-ratakan nilai SHAP absolut untuk setiap fitur, kemudian divisualisasikan menggunakan *summary plot* untuk menunjukkan tingkat kepentingan fitur, arah pengaruh (positif atau negatif), serta magnitudo kontribusinya. *Local explanations* dianalisis menggunakan *waterfall plots* untuk menjelaskan kontribusi masing-masing fitur pada sampel individual, sehingga memberikan gambaran yang lebih granular tentang bagaimana model mengambil keputusan untuk setiap pasien.

Peringkat fitur berdasarkan SHAP kemudian dibandingkan dengan faktor risiko dan kriteria diagnostik diabetes yang sudah diakui dalam pedoman klinis WHO [18] dan ADA [24], seperti kadar glukosa tinggi sebagai kriteria diagnostik utama, obesitas (BMI ≥ 25–30 kg/m²), riwayat keluarga, usia lanjut, dan riwayat diabetes gestasional. Kesesuaian antara peringkat SHAP dan pedoman klinis ini menjadi indikator bahwa model tidak sekadar menangkap korelasi statistik, melainkan mempelajari pola yang secara medis dapat dipertanggungjawabkan sebuah syarat penting sebelum model dapat diadopsi dalam praktik klinis.

3. Hasil dan Pembahasan

3.1. Performa Model

Tabel 4 menyajikan hasil evaluasi ketiga model melalui skema *nested cross-validation* (5-*fold outer*, 3-*fold inner*) pada data latih (n = 614). Hasil ini merupakan estimasi performa yang konservatif karena setiap model dievaluasi pada data yang benar-benar tidak terlibat dalam proses *hyperparameter tuning*.

Tabel 4. Performa Model pada *Nested Cross-Validation*

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC-ROC*	*PR-AUC*
Logistic Regression	0.775 ± 0.014	0.748 ± 0.068	0.552 ±0.056	0.630 ± 0.025	0.723 ± 0.016	0.566 ± 0.023
Random Forest	0.762 ± 0.030	0.691 ± 0.062	0.589 ± 0.040	0.634 ± 0.035	0.722 ± 0.027	0.550 ± 0.038
XGBoost	0.765 ± 0.033	0.703 ± 0.076	0.580 ± 0.037	0.634 ± 0.043	0.722 ± 0.031	0.555 ± 0.049

Ketiga model menunjukkan performa yang sebanding, dengan nilai *mean AUC-ROC* yang sangat berdekatan di kisaran 0,722–0,723. *Logistic Regression* mencapai akurasi tertinggi (77,5% ± 1,4%) dan *Precision* tertinggi (74,8% ± 6,8%), dengan standar deviasi terendah untuk akurasi (±0,014) dan *AUC-ROC* (±0,016) dibanding dua model lainnya. Stabilitas ini menunjukkan bahwa *Logistic Regression* menghasilkan estimasi yang paling konsisten di berbagai *fold*, menjadikannya kandidat yang dapat diandalkan untuk penerapan klinis.

Di sisi lain, *Recall* pada *Logistic Regression* (55,2% ± 5,6%) tercatat lebih rendah dibandingkan *Random Forest* (58,9% ± 4,0%) dan *XGBoost* (58,0% ± 3,7%). Ini mencerminkan *trade-off* yang inheren antara *Precision* dan *Recall Logistic Regression* lebih konservatif dalam memberikan prediksi positif, sehingga lebih sedikit kasus diabetes yang terlewat namun *Recall*-nya lebih rendah. Dalam konteks skrining klinis, *trade-off* ini perlu dipertimbangkan sesuai prioritas apakah meminimalkan *false negative* (deteksi dini) atau meminimalkan *false positive* (menghindari *overdiagnosis*) yang lebih diprioritaskan. Dari sisi *PR-AUC* yang lebih sensitif terhadap kelas minoritas, *Logistic Regression* juga mencatatkan nilai tertinggi (0,566 ± 0,023), diikuti *XGBoost* (0,555 ± 0,049) dan *Random Forest* (0,550 ± 0,038). Meskipun selisihnya kecil, pola ini konsisten dengan dominasi *Logistic Regression* pada metrik-metrik lainnya dan memperkuat argumen bahwa model ini paling stabil di antara ketiga algoritma yang diuji.

3.2. Hasil Uji Signifikansi Statistik

Tabel 5 menyajikan hasil pengujian statistik untuk menentukan apakah perbedaan performa antar model signifikan secara statistik.

Tabel 5. Hasil Uji Signifikansi Statistik Antar Model

Pengujian	Statistik	P-value	Signifikan
Friedman	X^2 = 0,4000	0,8187	Tidak
Wilcoxon: LR vs RF	W = 7,0000	1,0000	Tidak
Wilcoxon: LR vs XGBoost	W = 6,0000	0,8125	Tidak
Wilcoxon: RF vs XGBoost	W = 7,0000	1,0000	Tidak

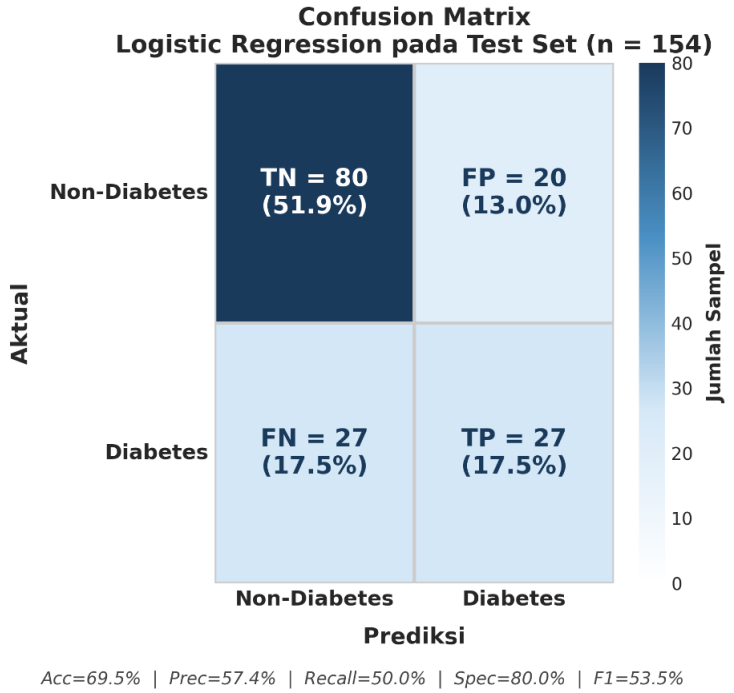
Friedman test menunjukkan tidak ada perbedaan yang signifikan secara statistik di antara ketiga model (X^2 = 0,4; p = 0,819). Untuk kelengkapan pelaporan, uji *Wilcoxon signed-rank* tetap dijalankan pada setiap pasangan model dengan koreksi Bonferroni (α_corrected = 0,0167). Hasilnya konsisten dengan temuan Friedman tidak ada pasangan model yang menunjukkan perbedaan signifikan: LR vs RF (W = 7,0; p = 1,000), LR vs XGBoost (W = 6,0; p = 0,813), dan RF vs XGBoost (W = 7,0; p = 1,000).

Besaran perbedaan performa (*mean difference* ± SD) antar model juga sangat kecil: LR vs RF sebesar 0,0012 ± 0,0309, LR vs XGBoost sebesar 0,0010 ± 0,0397, dan RF vs XGBoost sebesar -0,0002 ± 0,0207. Perbedaan yang berada jauh di bawah ambang 1% ini mengonfirmasi bahwa ketiga model secara praktis setara dalam hal performa pada dataset ini. Berdasarkan kesetaraan performa yang terbukti secara statistik, pemilihan model akhir didasarkan pada prinsip parsimoni dan pertimbangan klinis. *Logistic Regression* dipilih sebagai model terbaik karena tiga alasan: (1) stabilitas tertinggi dengan standar deviasi *AUC-ROC* terendah (±0,016) dibanding *Random Forest* (±0,027) dan *XGBoost* (±0,031); (2) *Precision* tertinggi (74,8% ± 6,8%) yang penting untuk meminimalkan *false alarm* dalam konteks klinis; dan (3) kompleksitas model yang paling rendah sehingga hasilnya paling mudah diinterpretasikan dan dijelaskan kepada klinisi.

Submitted : 15-03-2026 | Reviewed : 25-03-2026 | Accepted : 31-03-2026

3.3. Evaluasi pada Data Uji Independen

Evaluasi pada *test set* independen (n = 154) menghasilkan *accuracy* 69,5%, *precision* 57,4%, *recall* 50,0%, *specificity* 80,0%, *F1-score* 53,5%, *AUC-ROC* 81,4%, dan *PR-AUC* 65,9%. Hasil ini mencerminkan performa model pada data yang benar-benar tidak terlibat dalam proses pelatihan maupun pemilihan *hyperparameter*.



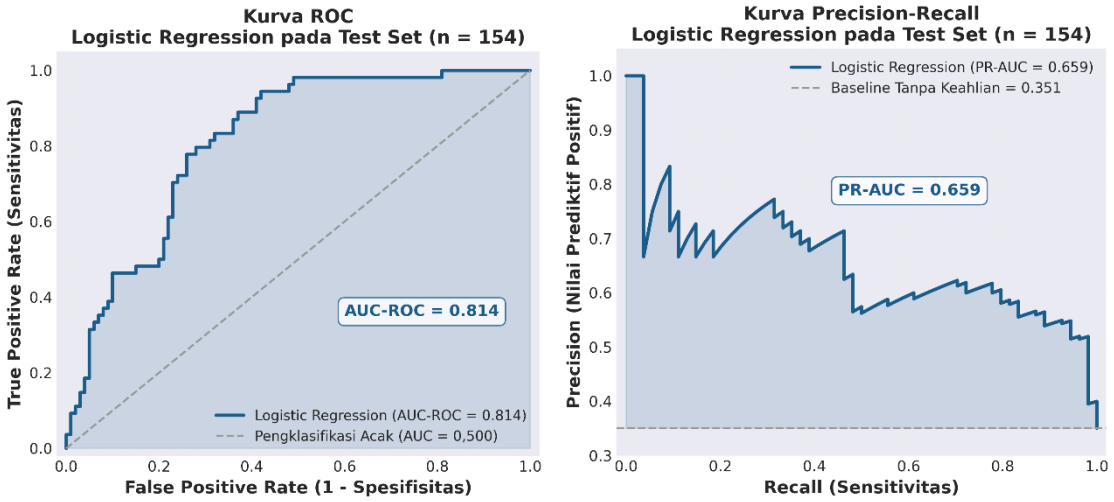
Gambar 2. *Confusion Matrix* Logistic Regression pada Test Set Independen (n = 154).

Gambar 2 menunjukkan rincian prediksi model pada level individual. Dari 100 pasien non-diabetes, model berhasil mengklasifikasikan 80 dengan benar (TN = 80, spesifisitas 80,0%), sementara 20 pasien non-diabetes diprediksi salah sebagai diabetes (FP = 20). Dari 54 pasien diabetes, model berhasil mendeteksi 27 kasus (TP = 27, *recall* 50,0%), sedangkan 27 kasus diabetes tidak terdeteksi (FN = 27). Nilai FN yang cukup tinggi ini mencerminkan *trade-off* yang terjadi pada threshold default 0,5 model cenderung lebih konservatif dalam memberikan prediksi positif, sehingga *precision* terjaga (57,4%) namun *recall* lebih rendah. Dalam konteks skrining klinis, threshold dapat diturunkan untuk meningkatkan *recall* demi deteksi dini, dengan konsekuensi *precision* yang lebih rendah.

Untuk menentukan threshold yang optimal secara klinis, dilakukan analisis pada beberapa kandidat threshold. Pada threshold default 0,5, model mencapai *precision* 57,4% dan *recall* 50,0%. Dengan menurunkan threshold ke 0,30, *recall* meningkat signifikan menjadi 83,3% dengan *precision* 58,4% lebih sesuai untuk konteks skrining populasi di mana deteksi dini diprioritaskan. Sebaliknya, threshold 0,60 menghasilkan *precision* lebih tinggi (71,4%) namun *recall* menurun ke 46,3%, lebih cocok untuk konteks konfirmasi diagnostik. Pemilihan threshold akhir sebaiknya dilakukan bersama klinisi berdasarkan prioritas antara false negative dan false positive dalam setting klinis yang spesifik. Hal ini sejalan dengan temuan penelitian terkini yang menunjukkan bahwa trade-off antara sensitivitas dan spesifisitas dalam prediksi diabetes perlu disesuaikan dengan prioritas aplikasi klinis yang spesifik [25]. perbandingan performa model pada berbagai threshold disajikan pada Tabel 6.

Tabel 6. Performa Model pada Berbagai *Threshold*

Threshold	*Precision*	*Recall*	*F1-Score*
0,30	58,4%	83,3%	68,7%
0,40	59,7%	63,0%	61,3%
0,50 (default)	57,4%	50,0%	53,5%
0,60	71,4%	46,3%	56,2%



Gambar 3. Kurva ROC dan *Precision-Recall* Logistic Regression pada Test Set Independen (n = 154). (A) Kurva ROC menunjukkan AUC-ROC = 0,814. (B) Kurva *Precision-Recall* menunjukkan PR-AUC = 0,659 dengan *baseline* tanpa keahlian = 0,351.

Kurva ROC (Gambar 3A) menunjukkan bahwa model memiliki kemampuan diskriminasi yang baik dengan AUC-ROC = 0,814, jauh di atas pengklasifikasi acak (AUC = 0,500). Kurva *Precision-Recall* (Gambar 3B) menunjukkan PR-AUC = 0,659, hampir dua kali lipat di atas *baseline* tanpa keahlian (0,351), yang mengonfirmasi bahwa model memiliki kemampuan prediksi yang bermakna meskipun distribusi kelas tidak seimbang. Untuk mencapai *recall* di atas 70%, *precision* turun hingga sekitar 55%, sehingga pemilihan *threshold* perlu disesuaikan dengan prioritas klinis yang berlaku.

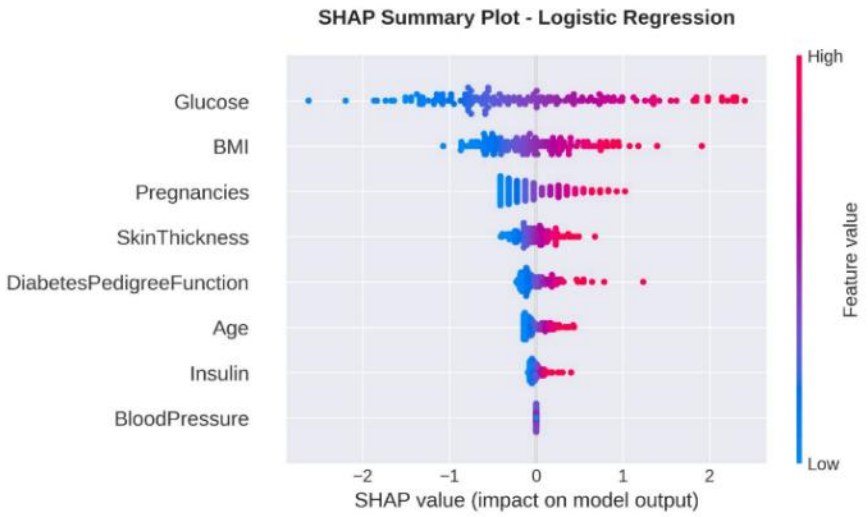
Perbedaan nilai AUC-ROC antara *nested cross-validation* (72,3% ± 1,6%) dan *test set* (81,4%) dapat dijelaskan oleh tiga faktor metodologis. Pertama, *nested cross-validation* menghasilkan estimasi yang konservatif karena pada setiap *outer fold*, model hanya dilatih menggunakan subset data latih (sekitar 491 sampel), sementara evaluasi pada *test set* menggunakan model yang dilatih pada seluruh data latih (614 sampel) sehingga model memiliki informasi yang lebih lengkap. Kedua, ukuran *test set* yang relatif kecil (n = 154) dengan proporsi kelas diabetes 34,4% dapat menyebabkan variabilitas *sampling* dan fluktuasi performa. Ketiga, *outer folds* pada *nested cross-validation* menggunakan *stratified splitting* yang berbeda pada setiap iterasi, sehingga nilai performa yang dilaporkan mewakili rata-rata tingkat kesulitan dari berbagai subset data, bukan hasil dari satu *test set* yang tetap. Dalam konteks ini, estimasi dari *nested cross-validation* (72,3%) berfungsi sebagai acuan konservatif yang lebih aman untuk perencanaan penerapan klinis, sementara AUC-ROC pada *test set* (81,4%) berada dalam rentang atas yang wajar dan konsisten dengan tujuan utama skema validasi yang diterapkan.

Untuk mengkuantifikasi ketidakpastian estimasi, dihitung 95% *confidence interval* AUC-ROC pada *test set* menggunakan metode *bootstrap* (n = 1.000), menghasilkan 95% CI = [0,737; 0,876]. Batas bawah CI (73,7%) yang masih kompatibel dengan estimasi *nested cross-validation* (72,3%) mengonfirmasi bahwa kesenjangan tersebut disebabkan oleh variabilitas *sampling*, bukan indikasi *overfitting*. Untuk penelitian lanjutan, penerapan *repeated nested cross-validation* (misalnya 5 repetisi) dapat dipertimbangkan guna lebih lanjut mengurangi variabilitas estimasi performa.

3.4 Interpretabilitas dan Validasi Klinis

Analisis SHAP pada *Logistic Regression* mengidentifikasi *Glucose* sebagai prediktor dominan dengan *mean* |SHAP| sebesar 0,864, diikuti oleh BMI (0,415), *Pregnancies* (0,289), *SkinThickness* (0,153), dan *DiabetesPedigreeFunction* (0,150). Fitur *Age*, *Insulin*, dan *BloodPressure* menunjukkan kontribusi yang relatif kecil dengan *mean* |SHAP| < 0,12. *Summary plot* (Gambar 4) menampilkan distribusi nilai SHAP untuk setiap fitur, di mana warna menunjukkan nilai fitur (merah = tinggi, biru = rendah), sedangkan posisi horizontal menggambarkan dampaknya terhadap prediksi model. *Glucose* terlihat jelas mendominasi prediksi model, ditunjukkan oleh rentang nilai SHAP yang paling lebar dibandingkan fitur lainnya.

Submitted : 15-03-2026 | Reviewed : 25-03-2026 | Accepted : 31-03-2026



Gambar 4. *Global Feature Importance* berdasarkan SHAP *Values*: *Summary Plot* untuk *Logistic Regression*

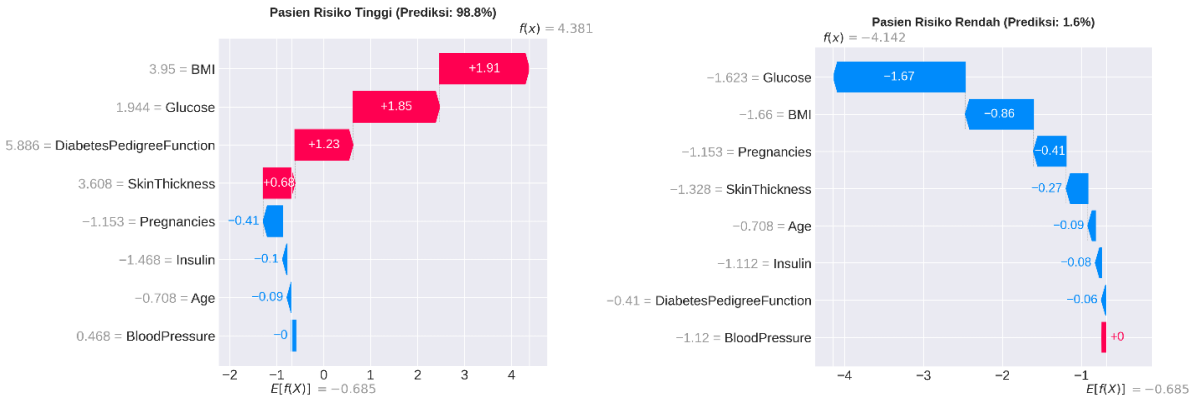
Untuk menilai apakah model mempelajari pola yang bermakna secara medis atau hanya artefak spesifik dataset, peringkat *feature importance* SHAP dibandingkan dengan faktor risiko diabetes dan kriteria diagnostik yang telah diakui dalam pedoman American Diabetes Association (ADA) [24] dan World Health Organization (WHO) [18]. *Glucose* muncul sebagai prediktor utama, mencerminkan perannya sebagai kriteria diagnostik untuk diabetes. ADA menetapkan *fasting plasma glucose* (FPG) ≥126 mg/dL atau *2-hour plasma glucose* ≥200 mg/dL selama *oral glucose tolerance test* (OGTT) sebagai ambang diagnosis diabetes [24, p. S21]. Dataset *Pima Indians* menggunakan pengukuran glukosa plasma 2 jam pasca-OGTT, yang secara langsung merepresentasikan standar diagnostik tersebut. Oleh karena itu, prioritas model terhadap *Glucose* selaras dengan perannya sebagai *gold standard* biomarker diagnostik dalam praktik klinis.

BMI menempati peringkat kedua, konsisten dengan pedoman skrining ADA yang mengidentifikasi kelebihan berat badan dan obesitas sebagai kriteria risiko utama tes skrining perlu dipertimbangkan pada orang dewasa dengan BMI ≥25 kg/m² atau ≥23 kg/m² untuk individu Asia-Amerika [24, p. S27]. WHO juga menekankan bahwa sebagian besar penderita diabetes tipe 2 memiliki kelebihan berat badan atau obesitas yang menyebabkan atau memperburuk resistensi insulin [18, p. 15]. Identifikasi BMI sebagai prediktor terpenting kedua mencerminkan pemahaman klinis yang sudah mapan mengenai obesitas sebagai faktor risiko utama yang dapat dimodifikasi pada diabetes tipe 2.

DiabetesPedigreeFunction, yang mewakili riwayat keluarga dengan diabetes, berada pada peringkat ketiga sejalan dengan kriteria ADA yang mencantumkan kerabat tingkat pertama dengan diabetes sebagai salah satu kriteria utama skrining [24, p. S27]. WHO juga mengakui adanya kecenderungan familial yang kuat, kemungkinan bersifat genetik atau epigenetik, pada diabetes tipe 2 [18, p. 15]. Studi klinis menunjukkan bahwa individu dengan kerabat tingkat pertama penderita diabetes memiliki risiko 2–6 kali lipat lebih tinggi, sehingga mendukung identifikasi fitur ini sebagai prediktor yang signifikan. Fitur tambahan *Age*, *Pregnancies* (riwayat diabetes gestasional), dan *BloodPressure* berada pada peringkat yang lebih rendah, namun tetap diakui dalam pedoman ADA dan WHO sebagai faktor risiko yang relevan.

Secara keseluruhan, tiga prediktor teratas menunjukkan kesesuaian yang kuat dengan kriteria klinis. Kesesuaian ini menjadi indikator bahwa model tidak sekadar menangkap korelasi statistik, melainkan mempelajari pola yang secara medis dapat dipertanggungjawabkan sebuah syarat penting sebelum sistem berbasis AI dapat diadopsi secara tepercaya dalam praktik klinis [26].

Local explanations melalui *waterfall plots* (Gambar 5) memberikan gambaran kontribusi fitur pada level pasien individual dengan karakteristik yang kontras.



Gambar 5. SHAP *Waterfall Plots* untuk Kontribusi Fitur pada Prediksi Individual. (A) Kasus *High-Risk* (prediksi: 98,8%). (B) Kasus *Low-Risk* (prediksi: 1,6%).

Gambar 5A menunjukkan kasus berisiko tinggi dengan probabilitas prediksi 98,8% (f(x) = 4,381). Pada pasien ini, BMI tinggi (3,95), *Glucose* tinggi (1,944), dan *DiabetesPedigreeFunction* tinggi (5,886) menghasilkan kontribusi SHAP positif yang kuat: BMI (+1,91), *Glucose* (+1,85), dan *DiabetesPedigreeFunction* (+1,23), mendorong probabilitas prediksi jauh di atas *baseline* E[f(x)] = −0,685. Perlu dicatat bahwa pada level global *Glucose* adalah prediktor dominan, namun pada kasus individual ini BMI memberikan kontribusi sedikit lebih besar perbedaan ini wajar karena *global feature importance* mencerminkan rata-rata di seluruh dataset, sementara *local explanation* dipengaruhi oleh nilai spesifik fitur pasien tersebut.

Sebaliknya, Gambar 5B menunjukkan kasus berisiko rendah dengan probabilitas prediksi 1,6% (f(x) = −4,142). Pasien ini memiliki *Glucose* rendah (−1,623) dan BMI dalam rentang normal (−1,66), menghasilkan kontribusi SHAP negatif yang kuat: *Glucose* (−1,67) dan BMI (−0,86), menarik probabilitas prediksi jauh di bawah *baseline*. Kedua contoh ini menunjukkan konsistensi perilaku model dengan pemahaman klinis faktor risiko yang diketahui mendorong prediksi ke arah positif pada kasus berisiko tinggi, dan tidak adanya faktor risiko tersebut menghasilkan prediksi rendah pada kasus berisiko rendah.

3.5 Implikasi dan Keterbatasan

Penelitian ini memberikan kontribusi metodologis yang dapat dibedakan dari studi-studi sebelumnya pada tiga aspek utama. Pertama, penerapan *nested cross-validation* dengan *pipeline* preprocessing terintegrasi secara eksplisit mencegah *data leakage* dan bias optimistis, sehingga estimasi performa yang dihasilkan lebih konservatif namun dapat dipercaya. Kedua, uji signifikansi statistik menggunakan Friedman (p = 0,819) dan *post-hoc* Wilcoxon membuktikan bahwa perbedaan performa antar model tidak signifikan secara statistik, sehingga pemilihan *Logistic Regression* didasarkan pada prinsip parsimoni dan interpretabilitas yang relevan untuk konteks medis bukan sekadar peringkat performa. Ketiga, validasi *feature importance* SHAP terhadap pedoman klinis ADA [24] dan WHO [18] meningkatkan tingkat kepercayaan terhadap model serta memfasilitasi potensi adopsinya dalam praktik klinis.

Dibandingkan dengan penelitian terkini yang juga menerapkan SHAP untuk prediksi diabetes, terdapat beberapa perbedaan metodologis yang perlu dicatat. Netayawijit et al. [14] telah menerapkan *preprocessing* dalam setiap *fold* untuk mencegah *data leakage*, namun menggunakan *simple cross-validation* tanpa *nested loop* untuk pemilihan *hyperparameter*, sehingga estimasi performa berpotensi optimistis akibat tidak adanya pemisahan antara proses seleksi model dan evaluasi performa. Hasan et al. [11] menggunakan *simple cross-validation* tanpa penjelasan eksplisit mengenai pencegahan *data leakage* pada tahap *hyperparameter tuning*. Selain itu, tidak ada satupun dari kedua studi tersebut yang menerapkan uji signifikansi statistik antar model sehingga klaim keunggulan suatu algoritma hanya didasarkan pada selisih angka metrik tanpa konfirmasi statistik. Penelitian ini juga membedakan diri melalui validasi *feature importance* SHAP secara sistematis terhadap pedoman klinis ADA [24] dan WHO [18], sedangkan studi-studi tersebut umumnya hanya menyentuh interpretabilitas klinis secara parsial.

Perbandingan dengan beberapa studi yang menggunakan dataset yang sama disajikan pada Tabel 7.

Tabel 7. Perbandingan Performa dan Metodologi

Penelitian Validasi	Metode Validasi	Akurasi (%)	*AUC-ROC* (%)	Rigor Metodologis
Erlin et al. [1]	10 - Fold CV	77 – 82	-	Simple CV, potensi *data leakage*
Safitri et al. [3]	10 - fold CV	75.0	82.0	Simple CV, potensi *data leakage*

Pramudyantoro et al.[2]	5 - fold CV	90.6	93.3	Simple CV, potensi *data leakage*
Setiawan & Suhirman [4]	*Train-test split* 80:20	76,62	-	*Single split* tanpa CV, tanpa SHAP formal, tanpa uji statistik
Penelitian ini	Nested 5×3 CV	77.5 ± 1.4 (CV) 69.5 (test)	72.3 ± 1.6 (CV) 81.4 (test)	Nested CV + *pipeline* anti-leakage + SHAP + validasi klinis

Tabel 7 menunjukkan bahwa beberapa studi sebelumnya melaporkan akurasi yang lebih tinggi, khususnya Pramudyantoro et al. [2] dengan akurasi 90,6%. Namun perlu dicatat bahwa nilai tersebut diperoleh melalui *simple cross-validation* yang berpotensi menghasilkan estimasi optimistis akibat kebocoran informasi pada tahap seleksi model. Dalam konteks aplikasi klinis, estimasi performa yang lebih rendah namun diperoleh melalui metodologi yang ketat lebih dapat diandalkan dibandingkan performa tinggi yang berpotensi bias. Setiawan & Suhirman [4] bahkan menggunakan *single train-test split* tanpa *cross-validation* sama sekali, yang menghasilkan estimasi dengan *variance* tinggi dan tidak mencerminkan kemampuan generalisasi model secara andal.

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, penggunaan satu dataset yaitu *Pima Indians Diabetes Dataset* yang hanya mencakup perempuan keturunan Indian Pima membatasi generalisasi hasil ke populasi dengan karakteristik demografis dan genetik yang berbeda. Kedua, proporsi *missing values* yang tinggi pada variabel *Insulin* (48,7%) berpotensi mengurangi kontribusi fitur tersebut dalam model prediksi meskipun telah ditangani melalui imputasi KNN dengan k = 5 dalam *pipeline*. Ketiga, dataset hanya mencakup delapan fitur klinis standar dan tidak menyertakan biomarker tambahan yang lazim digunakan dalam praktik klinis, seperti HbA1c dan profil lipid. Keempat, beberapa faktor risiko yang tercantum dalam pedoman ADA [24] termasuk tingkat aktivitas fisik, riwayat penyakit kardiovaskular, dan *polycystic ovary syndrome* tidak tersedia dalam dataset, sehingga cakupan validasi klinis model menjadi terbatas, Kelima, terdapat keterbatasan pada *external validity* model yang perlu mendapat perhatian. Meskipun model menunjukkan performa yang baik pada data uji internal (AUC-ROC 81,4%, 95% CI [0,737; 0,876]), validitas eksternal terhadap populasi di luar kelompok Pima Indian belum dapat dipastikan. Risiko *domain shift* menjadi nyata ketika model diterapkan pada populasi dengan distribusi demografis, etnis, atau klinis yang berbeda misalnya perbedaan rata-rata BMI, kadar glukosa, atau pola faktor risiko antar etnis dapat menyebabkan penurunan performa model secara signifikan. Penelitian lanjutan disarankan untuk melakukan validasi eksternal pada dataset multi-populasi serta mengeksplorasi teknik *domain adaptation* untuk meningkatkan generalisasi model.

Penelitian selanjutnya disarankan untuk menerapkan *multi-site datasets* guna meningkatkan generalisasi model ke populasi yang lebih beragam, mengeksplorasi metode imputasi yang lebih canggih seperti *Multiple Imputation by Chained Equations* (MICE) untuk menangani *missing values* secara lebih optimal, serta mengintegrasikan biomarker tambahan seperti HbA1c dan profil lipid agar relevansi klinis model dapat ditingkatkan. Eksplorasi metode XAI lain seperti LIME dan *Partial Dependence Plots* (PDP) juga dapat memperkaya interpretasi model dari perspektif yang berbeda.

4. Kesimpulan

Penelitian ini berhasil mengevaluasi dan membandingkan kinerja tiga algoritma *machine learning Logistic Regression*, *Random Forest*, dan *XGBoost* dalam memprediksi diabetes menggunakan kerangka validasi yang lebih ketat berbasis *nested cross-validation* (5-*fold outer*, 3-*fold inner*) dengan *pipeline* preprocessing terintegrasi. Ketiga model menunjukkan performa yang sebanding dengan nilai *mean AUC-ROC* di kisaran 0,722–0,723, dan uji Friedman (p = 0,819) mengonfirmasi bahwa perbedaan antar model tidak signifikan secara statistik. Berdasarkan prinsip parsimoni dan interpretabilitas, *Logistic Regression* dipilih sebagai model terbaik karena memiliki stabilitas tertinggi (standar deviasi *AUC-ROC* ±0,016), *precision* tertinggi (74,8% ± 6,8%), dan kompleksitas terendah. Pada data uji independen (n = 154), model mencapai *AUC-ROC* 81,4% dan *PR-AUC* 65,9%, yang mengonfirmasi kemampuan generalisasi model pada data yang tidak terlibat dalam proses pelatihan maupun pemilihan *hyperparameter*.

Analisis threshold menunjukkan bahwa penurunan threshold ke 0,30 meningkatkan *recall* menjadi 83,3% lebih sesuai untuk konteks skrining populasi sementara 95% *confidence interval* AUC-ROC = [0,737; 0,876] mengonfirmasi bahwa kesenjangan antara estimasi *nested cross-validation* (72,3%) dan *test set* (81,4%) disebabkan oleh variabilitas *sampling*, bukan indikasi *overfitting*.

Analisis SHAP mengidentifikasi *Glucose* sebagai prediktor dominan (*mean* |SHAP| = 0,864), diikuti BMI (0,415) dan *DiabetesPedigreeFunction* (0,150). Ketiga prediktor teratas menunjukkan kesesuaian yang kuat dengan kriteria klinis dalam pedoman ADA dan WHO *Glucose* sebagai *gold standard* diagnostik, BMI sebagai indikator obesitas dan resistensi insulin, serta *DiabetesPedigreeFunction* sebagai representasi kecenderungan familial yang bersifat genetik atau epigenetik. Kesesuaian ini menjadi bukti bahwa model tidak sekadar menangkap korelasi

statistik, melainkan mempelajari pola yang secara medis dapat dipertanggungjawabkan sebuah syarat penting sebelum sistem berbasis AI diadopsi dalam praktik klinis.

Penelitian ini menunjukkan bahwa estimasi performa yang konservatif namun dapat dipercaya lebih bernilai untuk penerapan klinis dibandingkan performa tinggi yang diperoleh melalui metodologi yang berpotensi bias. Untuk penelitian selanjutnya, disarankan penggunaan *multi-site datasets* guna meningkatkan generalisasi, eksplorasi metode imputasi yang lebih canggih seperti MICE untuk menangani *missing values*, serta integrasi biomarker tambahan seperti HbA1c dan profil lipid untuk meningkatkan relevansi klinis model.

Daftar Rujukan

[1] Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, and Nova Agustina, "Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 88–96, 2022, doi: 10.22146/jnteti.v11i2.3586.

[2] A. Pramudyantoro, E. Utami, and D. Ariatmanto, "Penggabungan K-Nearest Neighbors Dan Lightgbm Untuk Prediksi Diabetes Pada Dataset Pima Indians: Menggunakan Pendekatan Exploratory Data Analysis," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 3, pp. 1133–1144, 2024, doi: 10.29100/jipi.v9i3.4966.

[3] E. Safitri, D. Rofianto, N. Purwati, H. Kurniawan, and S. Karnila, "Prediksi Penyakit Diabetes Melitus Menggunakan Algoritma Machine Learning," *J. Sist. dan Teknol. Inf.*, vol. 12, no. 4, pp. 760–766, 2024, doi: 10.26418/justin.v12i4.84620.

[4] A. Setiawan and Suhirman, "Pengembangan Sistem Prediksi Risiko Diabetes Menggunakan Algoritma Support Vector Machine (SVM)," *J. Pustaka AI*, vol. 5, no. 3, pp. 562–572, 2025, doi: https://doi.org/10.55382/jurnalpustakaai.v5i3.1437.

[5] S. Arti and E. Suherlan, "Evaluasi Kinerja Machine Learning dalam Memprediksi Kemampuan Adaptasi Mahasiswa pada Lingkungan Pembelajaran Daring," *J. Pustaka AI*, vol. 5, no. 1, pp. 50–57, 2025, doi: https://doi.org/10.55382/jurnalpustakaai.v5i1.901.

[6] International Diabetes Federation, "IDF Diabetes Atlas 11th Edition," Brussels, Belgium, 2025.

[7] World Health Organization, "Diabetes (Fact Sheet)," Geneva, 2024.

[8] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16157–16173, 2023, doi: 10.1007/s00521-022-07049-z.

[9] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[10] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 2, pp. 1–43, 2023, doi: 10.1002/widm.1484.

[11] R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, "Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights," *Inf.*, vol. 16, no. 1, 2025, doi: 10.3390/info16010007.

[12] R. Hendawi, J. Li, and S. Roy, "A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study," *JMIR Form. Res.*, vol. 7, no. 1, pp. 1–18, 2023, doi: 10.2196/50328.

[13] M. M. Islam, H. R. Rifat, M. S. Bin Shahid, A. Akhter, M. A. Uddin, and K. M. M. Uddin, "Explainable Machine Learning for Efficient Diabetes Prediction Using Hyperparameter Tuning, SHAP Analysis, Partial Dependency, and LIME," *Eng. Reports*, vol. 7, no. 1, 2025, doi: 10.1002/eng2.13080.

[14] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support," *Healthc.*, vol. 13, no. 20, pp. 1–26, 2025, doi: 10.3390/healthcare13202588.

[15] M. Kutlu, T. B. Donmez, and C. Freeman, "Machine Learning Interpretability in Diabetes Risk Assessment: A SHAP Analysis," *Comput. Electron. Med.*, vol. 1, no. 1, pp. 34–44, 2024, doi: 10.69882/adba.cem.2024075.

[16] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[17] F. Pedregosa *et al.*, "Scikit-Learn Classifier Tuning from Complex Training Sets," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[18] World Health Organization, *Classification of diabetes mellitus*. Geneva, 2019. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/325158/9789241515702-eng.pdf

[19] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," *Proc. Annu. Symp. Comput. Appl. Med. Care*, pp. 261–265, 1988, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/

[20] A. Altamimi *et al.*, "An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques," *BMC Med. Res. Methodol.*, vol. 24, no. 1, 2024, doi: 10.1186/s12874-024-02324-0.

[21] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100994, 2024, doi: 10.1016/j.patter.2024.100994.

[22] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," *BioData Min.*, vol. 16, no. 1, pp. 1–14, 2023, doi: 10.1186/s13040-023-00330-4.

[23] S. Gholampour, "Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 827–841, 2024, doi: 10.3390/make6020039.

[24] American Diabetes Association Professional Practice Committee, "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2024," *Diabetes Care*, vol. 47, no. Supplement_1, pp. S20–S42, 2024, doi: 10.2337/dc25-S002.

[25] M. R. Khurshid, S. Manzoor, T. Sadiq, L. Hussain, M. S. Khan, and A. K. Dutta, "Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction," *PLoS One*, vol. 20, no. 1 January, pp. 1–29, 2025, doi: 10.1371/journal.pone.0310218.

[26] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Inf. Fusion*, vol. 99, no. April, p. 101805, 2023, doi: 10.1016/j.inffus.2023.101805.

Submitted : 15-03-2026 | Reviewed : 25-03-2026 | Accepted : 31-03-2026