JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analisis Kontribusi Fitur Bertahap dan Stabilitas Penjelasan SHAP pada Prediksi Risiko Diabetes Menggunakan XGBoost

Juni Ismail[1,*], Raja Anan Nasution[2,*], Muhammad Nasri Gea[2], Evi Handayani[2]

[1] Program Studi D3 Teknik Komputer, Politeknik Bisnis Indonesia, Simalungun, Indonesia
[2] Program Studi Manajemen Informatika, Akademi Manajemen Informatika dan Komputer ITMI, Deli Serdang, Indonesia
Email: [1,*]juniismailll@gmail.com, [2,*]rajanasutionnnn@gmail.com, [3]muhammadnasrigea@gmail.com, [4]handayanievi744@gmail.com
Email Penulis Korespondensi: juniismailll@gmail.com, rajanasutionnnn@gmail.com
Submitted **02-10-2026**; Accepted **30-10-2026**; Published **31-10-2026**

Abstrak

Model prediksi diabetes berbasis data tabular kerap dilaporkan dengan akurasi di atas 96%, padahal rendahnya prevalensi kasus membuat akurasi tersebut condong pada kelas mayoritas. Selain itu, fitur non-invasif dan penanda diagnostik umumnya digabungkan dalam satu model, sementara penjelasan SHAP jarang diuji kestabilannya. Penelitian ini bertujuan mengukur kontribusi marginal setiap kelompok fitur sekaligus menguji stabilitas penjelasan model. Sebanyak 96.128 rekam data dengan prevalensi diabetes 8,82% dimodelkan dalam empat tahap fitur, yaitu demografi dan riwayat penyakit (T0), penambahan indeks massa tubuh (T1), glukosa darah (T2), dan HbA1c (T3). XGBoost yang dioptimasi dengan Optuna dibandingkan dengan ResMLP dan FT-Transformer melalui validasi silang berulang 5×3 serta data uji terpisah, dengan AUPRC, *balanced accuracy*, dan Brier *score* sebagai metrik utama. AUPRC XGBoost mencapai 0,266; 0,325; 0,713; dan 0,889 untuk T0 hingga T3, dan setiap kenaikan antartahap signifikan (uji t terkoreksi, $p < 0,001$). Glukosa darah memberikan kenaikan AUPRC terbesar (0,389), sedangkan kontribusi usia menyusut dari 58,9% menjadi 12,7%. Peringkat fitur SHAP tetap stabil (Kendall $\tau = 0,93–1,00$; Jaccard tiga fitur teratas = 1,0) walaupun kinerja model berbeda jauh antartahap. XGBoost setara dengan model *deep learning* pada T0–T1, unggul pada T2–T3, dan lebih terkalibrasi. Audit menemukan 23,1% label tidak sesuai ambang ADA, dan seluruh kesalahan T3 berada pada zona label tidak konsisten. Fitur non-invasif berpotensi menjadi alat triase, sedangkan diagnosis membutuhkan penanda glikemik; temuan dari data publik berpola sintetis ini memerlukan validasi klinis eksternal.

Kata Kunci: Prediksi Diabetes; XGBoost; FT-Transformer; SHAP; Stabilitas Penjelasan; Data Tidak Seimbang

Abstract

Tabular diabetes prediction models are often reported with accuracies above 96%, yet the low prevalence of cases biases such accuracy toward the majority class. Moreover, non-invasive features and diagnostic markers are usually combined in a single model, and SHAP explanation stability is rarely examined. This study quantifies the marginal contribution of each feature group and assesses explanation stability. In total, 96,128 records (prevalence 8.82%) were modelled in four feature stages: demographics and medical history (T0), plus body mass index (T1), blood glucose (T2), and HbA1c (T3). Optuna-tuned XGBoost was compared with ResMLP and FT-Transformer using 5×3 repeated cross-validation and an independent hold-out set, with AUPRC, balanced accuracy, and Brier score as primary metrics. XGBoost achieved AUPRC of 0.266, 0.325, 0.713, and 0.889 for T0–T3, and every stage-wise gain was significant (corrected t-test, $p < 0.001$). Adding blood glucose produced the largest AUPRC gain (0.389), whereas the contribution of age fell from 58.9% to 12.7%. SHAP feature rankings remained stable (Kendall's $\tau = 0.93–1.00$; top-three Jaccard = 1.0) despite large performance differences across stages. XGBoost matched the deep learning models at T0–T1, was better at T2–T3, and was better calibrated. A label audit found 23.1% of labels inconsistent with ADA thresholds, with all T3 errors in the inconsistent-label zone. Non-invasive features could serve for triage, whereas diagnosis requires glycemic markers; findings from this public, synthetic-pattern dataset require external clinical validation.

Keywords: Diabetes Prediction; XGBoost; FT-Transformer; SHAP; Explanation Stability; Imbalanced Data

1. PENDAHULUAN

Diabetes melitus merupakan penyakit tidak menular yang bebannya terus meningkat secara global. *International Diabetes Federation* memperkirakan bahwa pada tahun 2024 terdapat sekitar 589 juta penduduk dewasa berusia 20–79 tahun yang hidup dengan diabetes [1]. Indonesia menempati kelompok lima negara dengan jumlah penyandang terbanyak, yakni sekitar 19,5 juta orang dewasa pada tahun 2021 [2]. Masalah yang lebih mendesak di tingkat nasional adalah besarnya kasus yang belum terdiagnosis. Survei Kesehatan Indonesia 2023 mencatat prevalensi diabetes berdasarkan pemeriksaan kadar gula darah sebesar 11,7% pada penduduk berusia 15 tahun ke atas, jauh melampaui prevalensi berdasarkan diagnosis dokter yang hanya 2,2% [3]. Kesenjangan ini memperlihatkan perlunya mekanisme skrining yang murah dan efisien, terutama pada layanan kesehatan primer yang tidak selalu memiliki akses terhadap pemeriksaan laboratorium.

Pendekatan *machine learning* menawarkan solusi melalui pemodelan risiko dari data rekam kesehatan yang telah tersedia. Salah satu data publik yang banyak digunakan memuat 100.000 rekam data dengan atribut demografi, riwayat hipertensi dan penyakit jantung, riwayat merokok, indeks massa tubuh (BMI), kadar HbA1c, serta glukosa darah [4]. Fauzi dan Budiman (2025) [5] membandingkan *Random Forest*, XGBoost, dan *Logistic Regression* pada data tersebut dengan penyeimbangan SMOTE dan memperoleh akurasi tertinggi 96,88% pada XGBoost. Khurshid dkk. (2025) [6] mengoptimasi XGBoost dengan *Bayesian optimization* pada data berukuran 100.000 partisipan dan melaporkan akurasi 97,26%. Li, Peng, dan Xiong (2024) [7] memanfaatkan algoritma genetika untuk menyetel hiperparameter XGBoost pada dataset yang memuat HbA1c dan glukosa darah. Hidayah dan Ikhsan (2026) [8] membandingkan SMOTE dan SMOTE-TomekLink pada XGBoost dan menemukan bahwa HbA1c serta glukosa darah mendominasi nilai SHAP. Iftikhar dkk.

Copyright © 2026 The **Juni Ismail**, Page 1478
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

(2026) [9] menggabungkan SMOTE dan SHAP pada dua dataset diabetes dan menempatkan glukosa, BMI, dan usia sebagai prediktor kunci. Penelitian lain mengembangkan *ensemble gradient boosting* dengan rekayasa fitur lanjutan [10], optimasi berbasis *Marine Predators Algorithm* yang dilengkapi SHAP [11], prediksi diabetes pada populasi anak [12], serta model *stacking* berbasis GA-XGBoost [13]. Pada konteks layanan primer di Indonesia, Taufiqurrahman dan Pratama (2026) [14] mengklasifikasikan risiko penyakit metabolik pada data Posbindu menggunakan model *ensemble*, *Decision Tree*, dan *Logistic Regression*.

Kajian terhadap penelitian terdahulu tersebut memperlihatkan empat celah penelitian. Celah pertama berkaitan dengan cara pelaporan kinerja. Dengan prevalensi diabetes sekitar 9%, model yang selalu menebak kelas "non-diabetes" sudah mencapai akurasi sekitar 91%, sehingga akurasi 97% hanya mencerminkan tambahan sekitar enam poin dan dapat menutupi rendahnya *recall*. Pada kondisi data tidak seimbang, kurva *precision–recall* dan luas di bawah kurva tersebut (AUPRC) lebih informatif daripada akurasi maupun ROC-AUC [15]. Teknik penyeimbangan seperti SMOTE [16] juga perlu dievaluasi dampaknya terhadap kalibrasi, karena koreksi ketidakseimbangan terbukti dapat merusak kalibrasi probabilitas model risiko klinis [17].

Celah kedua menyangkut komposisi fitur. Sebagian besar studi yang memakai dataset tersebut [5], [6], [7], [8] memasukkan glukosa darah dan HbA1c bersama fitur lain dalam satu model, padahal kedua variabel tersebut merupakan kriteria diagnosis diabetes [18]. Model seperti itu pada dasarnya mempelajari ulang aturan diagnosis, sehingga kinerjanya tidak menggambarkan kemampuan skrining sebelum pengambilan sampel darah. Studi berbasis fitur non-invasif pada data survei nasional Amerika Serikat melaporkan batas kinerja ROC-AUC sekitar 0,82 [19], tetapi kontribusi marginal setiap kelompok fitur, mulai dari demografi hingga penanda glikemik, belum diukur secara bertahap dengan uji statistik. Celah ketiga berkaitan dengan interpretabilitas. Penjelasan SHAP [20] umumnya dilaporkan dari satu kali pelatihan tanpa menguji apakah peringkat fitur bertahan ketika model dilatih ulang. Peringkat kepentingan fitur dapat berubah akibat variasi *sampling* [21], berbagai metode XAI kerap memberikan penjelasan yang saling bertentangan untuk model yang sama [22], dan pada domain klinis lain kestabilan penjelasan dilaporkan berbeda menurut tahap penyakit [23]. Studi XGBoost-SHAP pada dataset diabetes yang sama [8], [9], [11] melaporkan peringkat fitur dari satu model akhir, sedangkan kajian kestabilan penjelasan yang ada [21], [22], [23] dilakukan pada domain lain atau pada satu konfigurasi fitur. Dengan demikian, belum diketahui apakah penjelasan SHAP tetap stabil ketika model hanya menerima fitur non-invasif bersinyal lemah dibandingkan ketika penanda glikemik tersedia. Celah keempat menyangkut pemilihan arsitektur. FT-Transformer [24] makin sering diterapkan pada data tabular, sementara studi *benchmark* menunjukkan bahwa model berbasis pohon masih unggul pada sebagian besar data tabular [25], [26]. Perbandingan keduanya pada data klinis berdimensi rendah, khususnya dari sisi kalibrasi dan konsistensi penjelasan, masih terbatas.

Berdasarkan celah tersebut, penelitian ini bertujuan menganalisis kontribusi fitur bertahap dan stabilitas penjelasan SHAP pada prediksi diabetes menggunakan XGBoost yang dibandingkan dengan dua arsitektur *deep learning* tabular, yaitu ResMLP dan FT-Transformer. Kontribusi penelitian ini mencakup empat hal. Pertama, kerangka evaluasi empat tahap (T0–T3) yang mengukur kontribusi marginal fitur non-invasif dan penanda glikemik melalui validasi silang berulang dan uji statistik yang mengoreksi ketergantungan antar-*fold*. Kedua, analisis kestabilan penjelasan menggunakan Kendall τ dan indeks Jaccard pada level antar-*run*, antar-*seed*, antarmodel, dan antarmetode XAI, yang diukur pada setiap tahap fitur sehingga kestabilan dapat dibandingkan antara kondisi sinyal lemah dan sinyal kuat. Ketiga, perbandingan *gradient boosting* dan *deep learning* tabular yang tidak hanya menilai diskriminasi, tetapi juga kalibrasi, disertai uji statistik antarmodel. Keempat, evaluasi ulang praktik pelaporan akurasi melalui replikasi studi acuan dan analisis kesalahan klasifikasi. Hasil penelitian diharapkan dapat menjadi bahan awal perancangan skrining dua tingkat, yaitu triase non-invasif di tingkat komunitas yang diikuti pemeriksaan laboratorium bagi individu berisiko tinggi, setelah divalidasi pada data klinis.

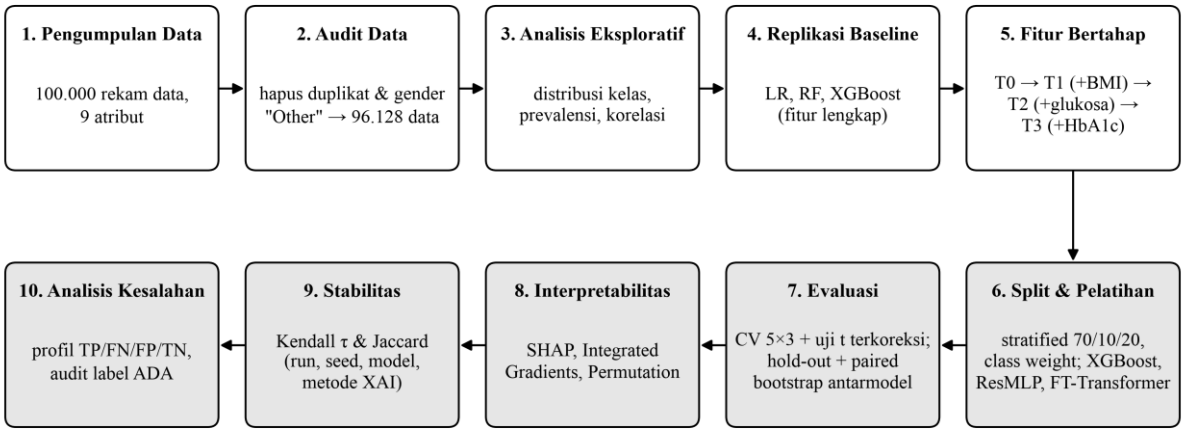
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini merupakan studi kuantitatif eksperimental dengan desain penambahan fitur bertahap (*stage-wise feature addition*). Alur penelitian ditunjukkan pada Gambar 1 dan terdiri atas sepuluh tahap yang saling berurutan, mulai dari pengumpulan data hingga analisis kesalahan klasifikasi.

Copyright © 2026 The **Juni Ismail**, Page 1479
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 1. Tahapan penelitian

Tahap pertama sampai keempat menyiapkan data dan memvalidasi *pipeline* melalui replikasi studi acuan [5]. Tahap kelima menyusun konfigurasi fitur bertahap, tahap keenam melatih tiga model, dan tahap ketujuh mengevaluasi kinerja. Tahap kedelapan dan kesembilan menganalisis interpretabilitas serta kestabilannya, sedangkan tahap kesepuluh menelusuri karakteristik kasus yang salah diklasifikasikan.

2.2 Dataset dan Pra-pemrosesan

Data yang digunakan adalah *Diabetes Prediction Dataset* [4] yang berisi 100.000 rekam data dan sembilan atribut, yaitu *gender*, *age*, *hypertension*, *heart_disease*, *smoking_history*, *bmi*, *HbA1c_level*, *blood_glucose_level*, dan label biner *diabetes*. Dataset ini dipublikasikan di Kaggle tanpa dokumentasi mengenai fasilitas asal data, periode pengumpulan, maupun persetujuan etik, sehingga representativitasnya terhadap populasi tertentu, termasuk populasi Indonesia, tidak dapat dipastikan. Audit data tidak menemukan nilai hilang, tetapi mendapati 3.854 baris duplikat identik dan 18 baris dengan kategori *gender* "Other". Baris duplikat dihapus untuk mencegah kebocoran informasi antara data latih dan data uji, sedangkan kategori "Other" dihapus karena jumlahnya terlalu kecil untuk dipelajari. Kategori "No Info" pada riwayat merokok (35.816 baris) dipertahankan sebagai kategori tersendiri. Data akhir berjumlah 96.128 baris dengan prevalensi diabetes 8,82%. Kesesuaian label terhadap ambang diagnosis ADA, yaitu HbA1c ≥ 6,5% atau glukosa plasma sewaktu ≥ 200 mg/dL [18], juga diaudit pada seluruh data bersih. Audit ini hanya memeriksa konsistensi label terhadap aturan ambang, bukan menilai benar atau salahnya diagnosis, karena diagnosis klinis memerlukan konfirmasi pemeriksaan ulang dan konteks klinis yang tidak tersedia pada dataset.

Data dibagi secara terstratifikasi menjadi data latih (±70%; 67.289 baris), data validasi (±10%) untuk *early stopping* dan pemilihan *threshold*, serta data uji (±20%; 19.226 baris). Fitur kategorikal dikodekan dengan *one-hot encoding*, sedangkan fitur kontinu distandardisasi untuk model *deep learning*. Seluruh transformasi hanya di-*fit* pada data latih. Data uji tidak digunakan dalam optimasi hiperparameter, *early stopping*, pemilihan *threshold*, maupun pelatihan model akhir. Empat konfigurasi fitur yang diuji disajikan pada Tabel 1. Tahap T0–T1 mewakili skenario skrining komunitas, sedangkan T2–T3 mewakili skenario diagnostik klinis, sehingga kinerja T3 diperlakukan sebagai batas atas (*upper bound*).

Tabel 1. Konfigurasi fitur bertahap

Tahap	Kelompok fitur	Fitur yang ditambahkan	Jumlah	Sifat
T0	Demografi + riwayat	gender, age, hypertension, heart_disease, smoking_history	5	Non-invasif
T1	T0 + antropometri	bmi	6	Non-invasif
T2	T1 + glikemik	blood_glucose_level	7	Invasif
T3	T2 + glikemik	HbA1c_level	8	Invasif (diagnostik)

2.3 Penanganan Ketidakseimbangan Kelas dan Pemodelan

Rasio kelas data sekitar 1 : 10,3. Empat strategi dibandingkan melalui studi ablasi pada T1 dan T3, yaitu tanpa penyeimbangan, pembobotan kelas (*class weight*), SMOTE [16], dan ADASYN. Teknik *oversampling* hanya diterapkan pada data latih. Ablasi menggunakan XGBoost dengan konfigurasi bawaan (300 pohon, kedalaman 5, laju pembelajaran 0,1) dan dievaluasi pada data validasi, sehingga data uji tidak terlibat dalam pemilihan strategi. Strategi terpilih diterapkan sebagai parameter *scale_pos_weight* pada XGBoost, dengan besaran bobot yang dipilih melalui optimasi hiperparameter, serta sebagai *WeightedRandomSampler* pada model *deep learning*.

Tiga model dibandingkan. XGBoost [27] dioptimasi terpisah untuk setiap tahap menggunakan Optuna [28] dengan *sampler* TPE sebanyak 30 *trial*. Fungsi objektifnya adalah rata-rata AUPRC pada validasi silang internal 3-*fold* atas

Copyright © 2026 The **Juni Ismail**, Page 1480
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

gabungan data latih dan validasi. Model XGBoost akhir untuk setiap tahap kemudian dilatih hanya pada data latih dengan hiperparameter terbaik tersebut. Optimasi mencakup delapan hiperparameter, yaitu *n_estimators*, *max_depth*, *learning_rate*, *subsample*, *colsample_bytree*, *min_child_weight*, *reg_lambda*, dan *scale_pos_weight*, dengan rentang *scale_pos_weight* antara 1 dan rasio kelas negatif terhadap positif (±10,3). ResMLP adalah *multilayer perceptron* dengan koneksi residual yang berperan sebagai *baseline deep learning* tabular [29], sedangkan FT-Transformer [24] mengubah setiap fitur menjadi *token embedding* yang diproses dengan mekanisme *self-attention*. ResMLP terdiri atas tiga blok residual berdimensi 256 dengan *dropout* 0,2, sedangkan FT-Transformer memakai dimensi *embedding* 64, tiga lapisan *encoder*, empat *attention head*, dan *dropout* 0,1. Keduanya dilatih dengan AdamW (laju pembelajaran 10^{-3}, *weight decay* 10^{-4}), penjadwal *OneCycle*, ukuran *batch* 1.024, maksimum 40 *epoch*, serta augmentasi berupa derau Gaussian, *feature masking*, MixUp, dan CutMix tabular. Pelatihan dihentikan lebih awal apabila AUPRC validasi tidak membaik selama enam *epoch*. Setiap model dilatih dengan tiga *seed* acak untuk mengukur variabilitas. Konfigurasi hiperparameter XGBoost terbaik untuk setiap tahap disajikan pada Tabel 2.

Tabel 2. Hiperparameter XGBoost terbaik hasil optimasi Optuna

Hiperparameter	T0	T1	T2	T3
n_estimators	362	200	395	268
max_depth	7	4	5	6
learning_rate	0,0126	0,0491	0,0236	0,0304
subsample	0,877	0,895	0,722	0,915
colsample_bytree	0,632	0,719	0,834	0,748
min_child_weight	5,41	5,88	4,10	1,76
reg_lambda	6,137	0,121	0,024	9,389
scale_pos_weight	2,63	1,43	9,00	10,30

2.4 Protokol Evaluasi

Evaluasi dilakukan pada dua lapis. Lapis pertama adalah *repeated stratified cross-validation* 5-*fold* × 3 pengulangan untuk XGBoost, sehingga setiap tahap menghasilkan 15 nilai metrik. Validasi silang ini dijalankan pada seluruh data bersih dan digunakan untuk membandingkan tahap secara relatif serta mengukur stabilitas penjelasan. Karena hiperparameter disetel pada data latih dan validasi yang turut menjadi bagian *fold*, nilai absolut lapis pertama cenderung sedikit optimistis, sedangkan estimasi kinerja yang tidak bias diperoleh dari lapis kedua. Perbedaan AUPRC antartahap berurutan diuji secara statistik dengan prosedur yang diuraikan pada subbab 2.6. Lapis kedua adalah evaluasi seluruh model pada data uji dengan dua *threshold*, yaitu 0,5 dan *threshold* teroptimasi yang memaksimalkan skor F1 pada data validasi, sejalan dengan praktik penyesuaian *threshold* pada data tidak seimbang [30]. *Threshold* ditentukan sekali pada data validasi, lalu diterapkan tanpa perubahan pada data uji. Metrik utama adalah AUPRC, *balanced accuracy*, dan Brier *score*, sedangkan akurasi, presisi, *recall*, F1, dan ROC-AUC berfungsi sebagai metrik pendukung [15]. *Balanced accuracy* dihitung dengan persamaan (1), AUPRC sebagai *average precision* dengan persamaan (2), dan Brier *score* dengan persamaan (3).

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{1}$$

$$\text{AUPRC} = \sum_{n} (R_n - R_{n-1}) P_n \tag{2}$$

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N} (\hat{p}_i - y_i)^2 \tag{3}$$

Pada persamaan tersebut, TP, TN, FP, dan FN masing-masing menyatakan *true positive*, *true negative*, *false positive*, dan *false negative*; P_n dan R_n adalah presisi dan *recall* pada *threshold* ke-n; $\hat{p}_i$ adalah probabilitas prediksi; dan y_i adalah label aktual dari N sampel.

2.5 Interpretabilitas dan Stabilitas Penjelasan

Kepentingan fitur dihitung dengan SHAP [20] untuk XGBoost, *Integrated Gradients* [31] untuk ResMLP dan FT-Transformer, serta *permutation importance* sebagai pembanding yang tidak bergantung pada model [32]. Nilai SHAP dihitung dengan algoritma TreeSHAP pada 2.000 sampel acak dari setiap *fold* uji validasi silang, *Integrated Gradients* memakai 32 langkah dengan *baseline* rata-rata data latih, dan *permutation importance* diukur sebagai penurunan AUPRC rata-rata dari lima kali pengacakan. Kontribusi relatif fitur ke-j dihitung sebagai rata-rata nilai absolut SHAP yang dinormalisasi terhadap seluruh fitur, sebagaimana persamaan (4).

$$C_j = \frac{\sum_{i=1}^{N} \left|\varphi_{ij}\right|}{\sum_{k=1}^{K}\sum_{i=1}^{N} \left|\varphi_{ik}\right|} \tag{4}$$

Copyright © 2026 The **Juni Ismail**, Page 1481
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Kestabilan peringkat fitur diukur dengan koefisien Kendall τ pada persamaan (5), dengan n_c dan n_d menyatakan jumlah pasangan fitur yang konkordan dan diskordan dari n fitur. Konsistensi tiga fitur teratas diukur dengan indeks Jaccard pada persamaan (6). Untuk XGBoost, τ dihitung pada seluruh 105 pasangan dari 15 *run* validasi silang. Karena setiap *run* muncul pada 14 pasangan, pasangan-pasangan tersebut tidak independen. Oleh karena itu, interval kepercayaan 95% diperoleh melalui *bootstrap* persentil sebanyak 2.000 kali pada level *run*: 15 *run* diambil ulang dengan pengembalian, kemudian rata-rata τ dihitung ulang hanya dari pasangan *run* yang berbeda. Kedua ukuran diterapkan pada empat level, yaitu antar-*run* validasi silang XGBoost, antar-*seed* model *deep learning*, antarmodel, dan antarmetode (SHAP terhadap *permutation importance*).

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{5}$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

2.6 Pengujian Statistik

Karena setiap pasangan *run* validasi silang berulang berbagi sebagian besar data latih, nilai metrik antar-*fold* tidak independen sehingga uji berpasangan biasa cenderung menghasilkan nilai p yang terlalu kecil. Oleh karena itu, perbedaan AUPRC antartahap berurutan diuji dengan *corrected resampled t-test* [33] menggunakan statistik pada persamaan (7).

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_{\text{uji}}}{n_{\text{latih}}}\right) s^2}} \tag{7}$$

Pada persamaan (7), $\bar{d}$ dan s^2 adalah rata-rata dan varians selisih AUPRC antara dua tahap dari k = 15 *run*, n_{uji}/n_{latih} = 1/4 merupakan rasio ukuran data uji terhadap data latih pada validasi silang 5-*fold*, dan derajat bebas uji t adalah k − 1. Uji peringkat bertanda Wilcoxon berpasangan dilaporkan sebagai uji pelengkap untuk menunjukkan konsistensi arah perbedaan. Perbedaan kinerja antarmodel diuji pada data uji dengan *paired bootstrap* terstratifikasi sebanyak 2.000 *resample*. Pada setiap *resample*, AUPRC dan Brier *score* XGBoost serta model *deep learning* (rata-rata tiga *seed*) dihitung pada sampel yang sama, lalu interval kepercayaan 95% selisihnya diambil dari persentil 2,5 dan 97,5. Nilai p dua sisi dikoreksi dengan metode Holm untuk delapan perbandingan pada setiap metrik (α = 0,05).

Seluruh eksperimen dijalankan menggunakan Python 3.13, PyTorch 2.11, XGBoost 3.4.1, scikit-learn 1.6.1, Optuna 5.0, imbalanced-learn 0.14.2, dan SHAP 0.54.1 pada GPU NVIDIA Tesla T4. Uji statistik tambahan dihitung ulang dari *checkpoint* model dengan pembagian data yang sama menggunakan CPU, dengan selisih AUPRC terhadap hasil awal ≤ 0,001, sedangkan nilai τ antar-*run* XGBoost pada Tabel 8 diperoleh dari pelatihan ulang 15 model validasi silang dengan pembagian *fold* yang sama.

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Data dan Validasi Pipeline

Analisis eksploratif memperlihatkan bahwa prevalensi diabetes jauh lebih tinggi pada individu dengan hipertensi (±28% berbanding ±7%) dan penyakit jantung (±32% berbanding ±8%). Prevalensi pada laki-laki (±10%) sedikit lebih tinggi daripada perempuan (±8%). Pada riwayat merokok, prevalensi tertinggi terdapat pada kategori *former* (±17%) dan terendah pada kategori "No Info" (±4%). Rendahnya prevalensi pada kategori "No Info" diduga berkaitan dengan usia, karena kategori ini banyak memuat individu muda yang belum memiliki riwayat merokok.

Sebelum eksperimen utama, *pipeline* divalidasi dengan mereplikasi studi acuan [5] menggunakan seluruh fitur. Hasilnya disajikan pada Tabel 3. Akurasi XGBoost hasil replikasi (96,91%) hanya berselisih 0,03 poin dari studi acuan (96,88%), demikian pula *Random Forest* dan *Logistic Regression*, sehingga *pipeline* dinilai sahih.

Tabel 3. Replikasi studi acuan dengan fitur lengkap

Model	Akurasi	Bal. Acc	Presisi	Recall	ROC-AUC	AUPRC	Akurasi acuan [5]
Logistic Regression	0,885	0,880	0,427	0,874	0,960	0,812	0,890
Random Forest	0,955	0,866	0,741	0,757	0,964	0,855	0,957
XGBoost	0,969	0,854	0,918	0,713	0,976	0,877	0,969

Replikasi ini sekaligus menunjukkan kelemahan pelaporan berbasis akurasi. Pada akurasi 96,9%, *recall* XGBoost hanya 0,713, yang berarti hampir 29% kasus diabetes tidak terdeteksi. Temuan ini mendasari penggunaan AUPRC dan *balanced accuracy* sebagai metrik utama pada eksperimen berikutnya.

3.2 Pengaruh Strategi Penyeimbangan Kelas

Hasil ablasi strategi penyeimbangan pada XGBoost disajikan pada Tabel 4. Tanpa penyeimbangan, model T1 hanya mencapai *recall* 0,058 walaupun akurasinya 91,2%, sehingga praktis tidak bermanfaat untuk skrining. Pembobotan kelas menaikkan *recall* T1 menjadi 0,789 dan *recall* T3 menjadi 0,899 tanpa menurunkan AUPRC. Sebaliknya, SMOTE dan

Copyright © 2026 The **Juni Ismail**, Page 1482
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

ADASYN justru menurunkan AUPRC T1 menjadi 0,279 dan 0,258. Penurunan ini diduga terjadi karena sampel sintetis dibangkitkan di wilayah tumpang-tindih antarkelas ketika sinyal fitur non-invasif lemah.

Tabel 4. Ablasi strategi penyeimbangan kelas pada XGBoost (data validasi)

Tahap	Strategi	Akurasi	Bal. Acc	Recall	AUPRC	Brier
T1	Tanpa penyeimbangan	0,912	0,526	0,058	0,315	0,069
T1	*Class weight*	0,734	0,759	0,789	0,317	0,168
T1	SMOTE	0,845	0,686	0,494	0,279	0,107
T1	ADASYN	0,842	0,679	0,480	0,258	0,110
T3	Tanpa penyeimbangan	0,972	0,857	0,717	0,896	0,022
T3	*Class weight*	0,918	0,909	0,899	0,895	0,051
T3	SMOTE	0,971	0,861	0,726	0,895	0,023
T3	ADASYN	0,969	0,865	0,738	0,891	0,025

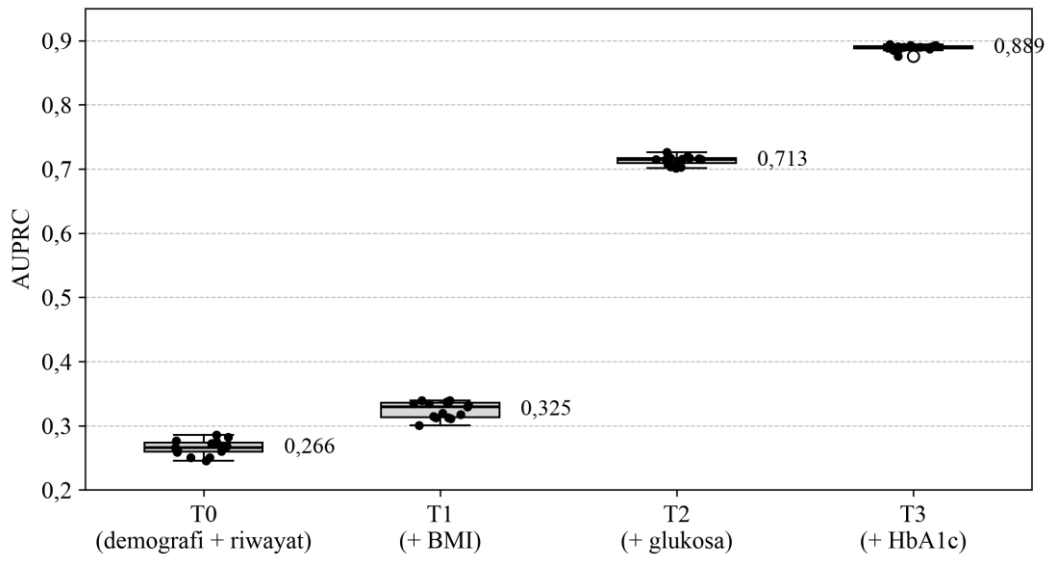
Penyeimbangan kelas pada dasarnya menggeser titik operasi model, bukan meningkatkan kemampuan diskriminasinya, karena selisih AUPRC antarstrategi tidak lebih dari 0,04. Pembobotan kelas memiliki konsekuensi berupa kenaikan Brier *score* T1 dari 0,069 menjadi 0,168, yang menandakan menurunnya kalibrasi probabilitas. Pola ini konsisten dengan temuan van den Goorbergh dkk. [17]. Pembobotan kelas dipilih karena menghasilkan *recall* tertinggi tanpa membangkitkan data sintetis dan tanpa menurunkan AUPRC. Penurunan kalibrasi yang menyertainya dapat diterima karena tujuan tahap non-invasif adalah menjaring sebanyak mungkin kasus positif, dan kalibrasi probabilitas masih dapat diperbaiki melalui kalibrasi ulang. Ketiga metrik tersebut mengukur aspek yang berbeda: AUPRC mengukur kemampuan diskriminasi yang tidak bergantung pada *threshold*, Brier *score* mengukur kesesuaian probabilitas dengan frekuensi kejadian, dan *recall* mencerminkan titik operasi yang dipilih. Oleh karena itu, strategi penyeimbangan dinilai terutama dari AUPRC, sedangkan *recall* dan kalibrasi diatur menurut tujuan penggunaan, yaitu *recall* tinggi untuk triase dan probabilitas terkalibrasi untuk estimasi risiko. Ablasi ini hanya menentukan jenis strategi, sedangkan besaran *scale_pos_weight* pada setiap tahap dipilih oleh Optuna, sehingga bobot T1 relatif kecil (1,43) dan titik operasinya diatur kembali melalui *threshold* teroptimasi.

3.3 Kinerja Prediksi Bertahap

Kinerja XGBoost pada validasi silang berulang beserta hasil uji statistik antartahap disajikan pada Tabel 5, sedangkan sebaran AUPRC pada 15 *fold* ditampilkan pada Gambar 2.

Tabel 5. Kinerja XGBoost per tahap pada validasi silang 5×3 (rata-rata ± simpangan baku)

Tahap	ROC-AUC	AUPRC	Bal. Acc	Brier	ΔAUPRC	p (uji t terkoreksi)
T0	0,805 ± 0,005	0,266 ± 0,012	0,588 ± 0,007	0,085 ± 0,001	–	–
T1	0,833 ± 0,004	0,325 ± 0,013	0,552 ± 0,003	0,070 ± 0,001	+0,059	$5{,}6\times10^{-10}$
T2	0,935 ± 0,002	0,713 ± 0,007	0,847 ± 0,004	0,098 ± 0,001	+0,389	$3{,}5\times10^{-20}$
T3	0,979 ± 0,001	0,889 ± 0,004	0,913 ± 0,003	0,059 ± 0,001	+0,176	$2{,}0\times10^{-18}$



Gambar 2. Sebaran AUPRC XGBoost per tahap pada validasi silang 5×3

Copyright © 2026 The **Juni Ismail**, Page 1483
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

AUPRC meningkat secara monoton di setiap tahap, dan tahap yang lebih lengkap unggul pada seluruh 15 *fold*. Setelah varians dikoreksi terhadap tumpang-tindih data latih, setiap kenaikan tetap signifikan ($t = 14{,}90$; 82,10; dan 61,44 dengan derajat bebas 14; $p < 10^{-9}$). Uji Wilcoxon pelengkap menghasilkan $p = 6{,}1\times10^{-5}$ untuk ketiga perbandingan, yaitu nilai terkecil yang dapat dicapai dengan 15 pasangan; karena *fold* tidak sepenuhnya independen, nilai ini dibaca sebagai bukti konsistensi arah perbedaan, bukan sebagai probabilitas yang presisi. Simpangan baku yang kecil (≤ 0,013) memperlihatkan bahwa estimasi kinerja stabil. Kontribusi setiap kelompok fitur sangat timpang. BMI hanya menambah AUPRC sebesar 0,059, sedangkan glukosa darah menambah 0,389 atau 6,6 kali lipat kontribusi BMI, dan HbA1c masih menambah 0,176. Pola ini menunjukkan bahwa fitur demografi dan antropometri hanya membawa sinyal risiko, sementara glukosa dan HbA1c membawa sinyal diagnostik.

Pada *threshold* 0,5, *balanced accuracy* dan Brier *score* tidak berubah secara monoton. *Balanced accuracy* T1 (0,552) lebih rendah daripada T0 (0,588), dan Brier *score* T2 (0,098) lebih tinggi daripada T1 (0,070). Hal ini disebabkan oleh nilai *scale_pos_weight* hasil optimasi yang berbeda antartahap, yaitu 2,63 (T0), 1,43 (T1), 9,00 (T2), dan 10,30 (T3), sehingga titik operasi dan kalibrasi model ikut bergeser. Oleh karena itu, perbandingan antartahap didasarkan pada metrik yang tidak bergantung pada *threshold*, yaitu AUPRC dan ROC-AUC.

3.4 Perbandingan Model pada Data Uji

Kinerja ketiga model pada data uji dengan *threshold* teroptimasi ditampilkan pada Tabel 6. Nilai ResMLP dan FT-Transformer merupakan rata-rata dari tiga *seed*.

Tabel 6. Kinerja model pada data uji dengan threshold teroptimasi

Tahap	Model	Akurasi	Bal. Acc	Presisi	Recall	AUPRC	Brier
T0	XGBoost	0,821	0,671	0,244	0,488	0,253	0,086
T0	ResMLP	0,830	0,667	0,252	0,469	0,250	0,201
T0	FT-Transformer	0,819	0,677	0,245	0,505	0,249	0,192
T1	XGBoost	0,848	0,702	0,297	0,524	0,312	0,071
T1	ResMLP	0,855	0,691	0,303	0,492	0,307	0,186
T1	FT-Transformer	0,853	0,694	0,301	0,502	0,311	0,176
T2	XGBoost	0,929	0,778	0,598	0,594	0,704	0,100
T2	ResMLP	0,933	0,748	0,657	0,523	0,679	0,129
T2	FT-Transformer	0,935	0,750	0,678	0,526	0,685	0,120
T3	XGBoost	0,971	0,848	0,966	0,699	0,883	0,061
T3	ResMLP	0,968	0,845	0,918	0,696	0,865	0,072
T3	FT-Transformer	0,968	0,847	0,922	0,699	0,869	0,068

XGBoost memperoleh AUPRC tertinggi di seluruh tahap, tetapi selisihnya relatif kecil, yaitu ≤ 0,005 pada T0–T1 dan 0,014–0,025 pada T2–T3. Uji *paired bootstrap* pada Tabel 7, dengan selisih dihitung sebagai XGBoost dikurangi rata-rata tiga *seed* model *deep learning*, menunjukkan bahwa selisih AUPRC pada T0–T1 tidak signifikan (p Holm ≥ 0,36; interval kepercayaan 95% mencakup nol), sehingga ketiga model dapat dianggap setara pada tahap non-invasif. Pada T2–T3, keunggulan AUPRC XGBoost signifikan (p Holm < 0,01), sedangkan Brier *score* XGBoost lebih rendah secara signifikan di seluruh tahap. Mekanisme *attention* pada FT-Transformer tidak memberikan keuntungan berarti pada data dengan lima sampai delapan fitur. Perbedaan yang lebih tegas terlihat pada kalibrasi. Brier *score* model *deep learning* pada T0–T1 (0,18–0,20) lebih dari dua kali lipat XGBoost (0,07–0,09). Penyebabnya, *WeightedRandomSampler* membuat model dilatih seolah-olah prevalensi kelas positif mendekati 50%, sehingga probabilitas keluaran bergeser ke atas. Pergeseran ini juga tampak dari *threshold* teroptimasi model *deep learning* yang tinggi (0,66–0,85) meskipun prevalensi sebenarnya hanya 8,8%.

Tabel 7. Uji *paired bootstrap* selisih kinerja XGBoost terhadap model *deep learning* pada data uji

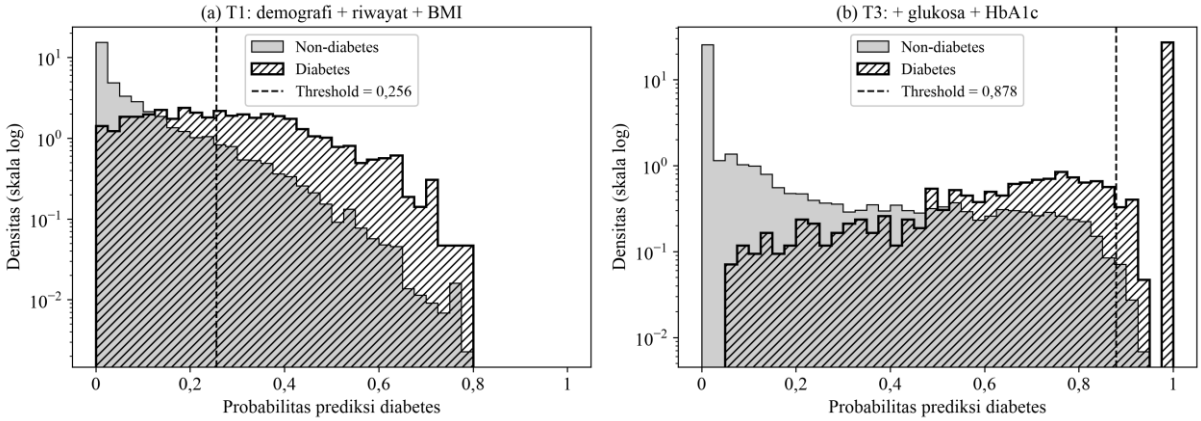
Tahap	Pembanding	ΔAUPRC (IK 95%)	p Holm	ΔBrier (IK 95%)	p Holm
T0	ResMLP	+0,003 (−0,004; 0,010)	0,73	−0,115 (−0,117; −0,113)	< 0,01
T0	FT-Transformer	+0,004 (−0,001; 0,011)	0,37	−0,106 (−0,108; −0,104)	< 0,01
T1	ResMLP	+0,005 (−0,001; 0,010)	0,36	−0,115 (−0,118; −0,113)	< 0,01
T1	FT-Transformer	+0,001 (−0,005; 0,007)	0,81	−0,105 (−0,107; −0,103)	< 0,01
T2	ResMLP	+0,025 (0,021; 0,030)	< 0,01	−0,029 (−0,030; −0,028)	< 0,01
T2	FT-Transformer	+0,019 (0,015; 0,023)	< 0,01	−0,020 (−0,021; −0,019)	< 0,01
T3	ResMLP	+0,018 (0,014; 0,022)	< 0,01	−0,011 (−0,012; −0,011)	< 0,01
T3	FT-Transformer	+0,014 (0,011; 0,017)	< 0,01	−0,007 (−0,008; −0,006)	< 0,01

Akurasi tidak informatif sebagai pembanding utama. Akurasi T0 dan T1 masing-masing 9,1 dan 6,3 poin di bawah *baseline* kelas mayoritas (91,18%), sedangkan akurasi T2 dan T3 hanya 1,7 dan 5,9 poin di atasnya. Sebaliknya, *balanced accuracy* meningkat dari 0,67 menjadi 0,85, yang mencerminkan perbaikan nyata dalam mengenali kedua kelas. Gambar

Copyright © 2026 The **Juni Ismail**, Page 1484
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

3 memperjelas perbedaan tersebut. Pada T1, distribusi probabilitas kedua kelas sangat tumpang-tindih sehingga tidak ada *threshold* yang mampu memisahkannya dengan baik. Pada T3, distribusi menjadi bimodal, dengan sebagian besar kasus non-diabetes terkumpul di sekitar nol dan sekelompok besar kasus diabetes terkumpul di sekitar satu, tetapi masih tersisa zona tumpang-tindih pada rentang 0,2–0,9.

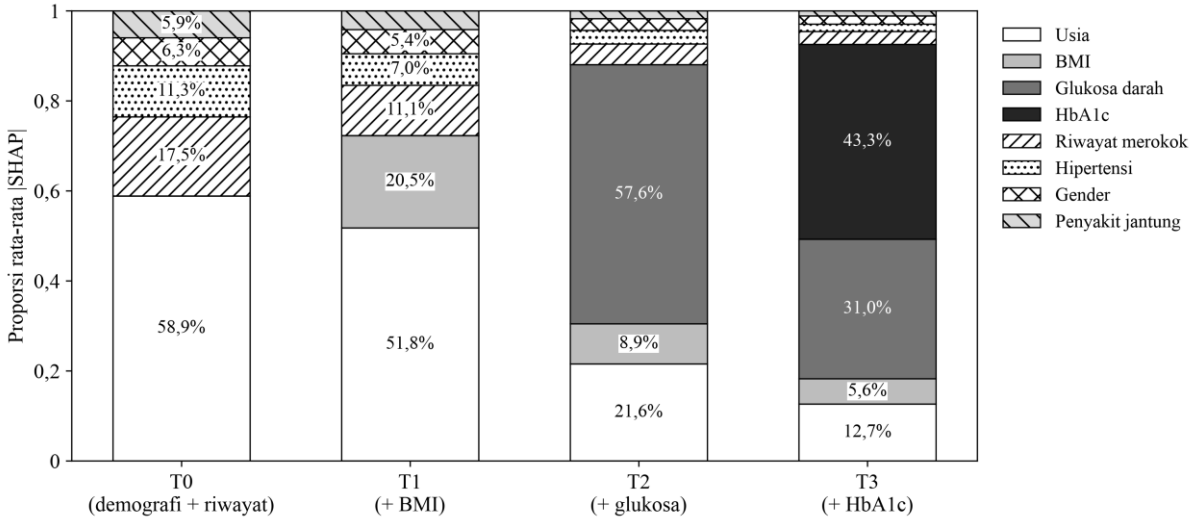


Gambar 3. Distribusi probabilitas prediksi XGBoost pada T1 dan T3

Pemilihan *threshold* juga berpengaruh besar terhadap profil kesalahan. Pada T2, XGBoost dengan *threshold* 0,5 menghasilkan *recall* 0,848 dan presisi 0,334, sedangkan *threshold* teroptimasi (0,753) membalik komposisi tersebut menjadi *recall* 0,594 dan presisi 0,598. Pada T3, *threshold* teroptimasi (0,878) menaikkan presisi dari 0,470 menjadi 0,966, tetapi menurunkan *recall* dari 0,916 menjadi 0,699. Pergeseran ini menunjukkan bahwa *threshold* yang memaksimalkan keseimbangan presisi dan *recall* belum tentu sesuai dengan kebutuhan skrining. Untuk tujuan deteksi dini, kehilangan kasus positif lebih merugikan daripada rujukan berlebih, sehingga *threshold* perlu ditentukan berdasarkan target *recall* minimum yang disepakati secara klinis, bukan semata-mata berdasarkan metrik gabungan.

3.5 Pergeseran Kontribusi Fitur

Proporsi kontribusi fitur berdasarkan persamaan (4) diperlihatkan pada Gambar 4. Usia menjadi prediktor dominan pada T0 dan T1 dengan kontribusi 58,9% dan 51,8%. Begitu glukosa darah dimasukkan, kontribusi usia turun menjadi 21,6% dan kemudian 12,7% pada T3. Glukosa darah langsung menjadi prediktor utama pada T2 (57,6%), lalu posisinya digantikan HbA1c pada T3 (43,3%), dengan kontribusi gabungan keduanya mencapai 74,3%. Kontribusi BMI menyusut dari 20,5% pada T1 menjadi 5,6% pada T3, sedangkan riwayat merokok menyusut dari 17,5% menjadi 2,8%.



Gambar 4. Pergeseran proporsi kontribusi fitur (rata-rata |SHAP|) antartahap

Penurunan kontribusi usia, BMI, dan riwayat penyakit tidak berarti fitur tersebut tidak relevan. Penurunan tersebut menunjukkan bahwa sebagian besar informasi prediktif fitur-fitur ini tumpang-tindih dengan informasi yang dibawa penanda glikemik. Karena SHAP hanya mengukur kontribusi fitur terhadap keluaran model, pola ini tidak dapat ditafsirkan sebagai hubungan sebab-akibat maupun mediasi. Secara deskriptif, pola tersebut selaras dengan pengetahuan klinis bahwa usia dan obesitas berkaitan dengan resistensi insulin yang tercermin pada kadar glukosa dan HbA1c. Fitur teratas juga konsisten lintas metode. Pada T0–T1, keempat metode (SHAP, *permutation importance*, serta *Integrated Gradients* pada ResMLP dan FT-Transformer) menempatkan usia di peringkat pertama. Pada T2 seluruh metode

Copyright © 2026 The **Juni Ismail**, Page 1485
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

menempatkan glukosa darah di peringkat pertama, dan pada T3 seluruh metode menempatkan HbA1c di peringkat pertama. Dominasi HbA1c dan glukosa ini selaras dengan temuan Hidayah dan Ikhsan [8].

3.6 Stabilitas Penjelasan

Hasil pengukuran kestabilan peringkat fitur disajikan pada Tabel 8. Kolom τ antar-*run* dihitung dari 15 model validasi silang XGBoost dengan interval kepercayaan hasil *bootstrap* pada level *run*, kolom τ antar-*seed* dari tiga model *deep learning* dengan *seed* berbeda, sedangkan kolom τ antarmodel membandingkan peringkat fitur XGBoost, ResMLP, dan FT-Transformer secara berpasangan.

Tabel 8. Stabilitas peringkat fitur berdasarkan Kendall τ

Tahap	n fitur	τ antar-run XGB (IK 95%)	τ min	Jaccard top-3	τ antar-seed DL	τ antarmodel	τ SHAP vs Perm.
T0	5	0,931 (0,886–1,000)	0,800	1,0	0,867	1,000	0,440
T1	6	1,000 (1,000–1,000)	1,000	1,0	0,911	0,867–1,000	0,591
T2	7	0,987 (0,964–1,000)	0,905	1,0	0,873	0,905–1,000	0,733
T3	8	0,970 (0,958–0,990)	0,929	1,0	0,952	0,929–1,000	0,752

Penjelasan model sangat stabil pada tiga level: antar-*run* XGBoost ($\tau \geq 0{,}93$; batas bawah IK 95% $\geq 0{,}88$), antar-*seed* model *deep learning* ($\tau \geq 0{,}87$), dan antarmodel ($\tau \geq 0{,}87$). Tiga fitur teratas identik pada seluruh *run* (Jaccard = 1,0). Sebagai gambaran, $\tau = 0{,}93$ berarti hanya sekitar 3,5% pasangan fitur yang urutannya bertukar antar-*run*, dan batas bawah IK 95% terendah (0,886 pada T0) masih setara dengan kurang dari 6% pasangan yang bertukar. Kestabilan ini tidak mengikuti tingkat kinerja. Ketika AUPRC naik dari 0,27 menjadi 0,89, nilai τ tetap berada di rentang 0,93–1,00. Dengan demikian, model berkinerja rendah tetap dapat menghasilkan penjelasan yang stabil, sehingga kestabilan penjelasan dan kemampuan prediksi merupakan dua properti yang harus dievaluasi secara terpisah.

Ketidakstabilan relatif hanya muncul pada T0 ($\tau = 0{,}931$; minimum 0,80; IK 95% 0,886–1,000). Penyebabnya adalah tiga fitur berkontribusi rendah, yaitu *gender* (0,063), *heart_disease* (0,059), dan *hypertension* (0,113), yang nilainya berdekatan sehingga urutannya mudah bertukar antar-*run*. Artinya, ketidakstabilan terpusat pada ekor peringkat, bukan pada fitur dominan. Kesepakatan SHAP dengan *permutation importance* rendah pada T0 ($\tau = 0{,}44$) dan meningkat seiring menguatnya sinyal hingga 0,75 pada T3. Pola ini merupakan wujud *disagreement problem* [22], karena *permutation importance* bernilai hampir nol untuk fitur lemah (misalnya *gender* 0,003) sehingga peringkatnya didominasi derau. Implikasinya, pada model dengan sinyal lemah, penjelasan perlu divalidasi silang dengan lebih dari satu metode XAI sebelum ditafsirkan secara klinis [21].

3.7 Analisis Kesalahan Klasifikasi

Profil rata-rata fitur pada setiap jenis prediksi XGBoost disajikan pada Tabel 9.

Tabel 9. Profil rata-rata fitur menurut jenis prediksi XGBoost pada data uji

Tahap	Jenis	n	Usia	BMI	Glukosa (mg/dL)	HbA1c (%)
T1	TP	889	66,3	35,0	–	–
T1	FN	807	54,1	29,0	–	–
T1	FP	2.108	65,9	33,5	–	–
T1	TN	15.422	36,5	26,1	–	–
T3	TP	1.185	60,8	32,6	213,8	7,32
T3	FN	511	59,8	31,1	150,0	6,13
T3	FP	42	68,2	39,9	154,1	6,18
T3	TN	17.488	40,0	26,9	132,6	5,39

Pada T1, model pada dasarnya mempelajari profil risiko. Kasus *false positive* (n = 2.108) hampir identik dengan *true positive*, yaitu berusia sekitar 66 tahun, mengalami obesitas, dan kerap disertai hipertensi serta penyakit jantung. Sebaliknya, kasus *false negative* lebih muda dan lebih kurus, bahkan beberapa kasus dengan probabilitas terendah adalah anak berusia 4–11 tahun dengan BMI normal yang kemungkinan merupakan diabetes tipe 1. Kesalahan T1 dengan demikian mencerminkan batas informasi fitur non-invasif, bukan kegagalan algoritma.

Pada T3, jumlah *false positive* turun drastis dari 2.108 menjadi 42, dan kesalahan yang tersisa terpusat pada rentang prediabetes. Kasus *false negative* memiliki rata-rata HbA1c 6,13% dan glukosa 150 mg/dL, jauh di bawah kasus *true positive* (7,32% dan 214 mg/dL). Penelusuran kasus dengan kesalahan terbesar menunjukkan bahwa sebagian label tidak sesuai dengan kriteria diagnosis, sehingga kesesuaian label diaudit pada seluruh 96.128 data bersih. Hasilnya, 1.795 dari 8.482 kasus berlabel diabetes (21,2%) tidak memenuhi kriteria HbA1c ≥ 6,5% maupun glukosa ≥ 200 mg/dL [18], dan seluruhnya berada pada rentang prediabetes. Sebaliknya, 20.392 dari 87.646 kasus berlabel non-diabetes (23,3%) justru memenuhi kriteria tersebut, terutama karena memiliki HbA1c 6,5–6,6% (14.921 kasus) atau glukosa tepat 200 mg/dL (6.642 kasus). Secara keseluruhan, 23,1% label tidak sesuai dengan aturan ambang diagnosis ADA.

Copyright © 2026 The **Juni Ismail**, Page 1486
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.10234
Hal 1478-1489
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Audit juga mengungkap bahwa HbA1c dan glukosa darah masing-masing hanya memiliki 18 nilai unik, dan label mengikuti pola berbasis zona yang dirangkum pada Tabel 10.

Tabel 10. Sebaran label menurut zona nilai glikemik pada data bersih

Zona glikemik	Jumlah data	Proporsi	Diabetes	Non-diabetes	Prevalensi
HbA1c ≥ 6,8% atau glukosa ≥ 220 mg/dL	5.671	5,9%	5.671	0	100%
Zona antara	79.295	82,5%	2.811	76.484	3,5%
HbA1c < 5,7% dan glukosa < 126 mg/dL	11.162	11,6%	0	11.162	0%

Seluruh data pada zona atas berlabel diabetes dan seluruh data pada zona bawah berlabel non-diabetes. Pada zona antara, prevalensi diabetes relatif datar, yaitu 4,7–5,7% untuk setiap nilai HbA1c antara 5,7% dan 6,6% serta 4,5–5,4% untuk setiap nilai glukosa antara 126 dan 200 mg/dL, sehingga label di zona ini praktis tidak bergantung pada kadar glikemik. Pola tersebut konsisten dengan data yang dibangkitkan secara sintetis dan menjelaskan tiga temuan sebelumnya. Pertama, puncak probabilitas di sekitar satu pada Gambar 3 berasal dari zona atas yang mencakup 66,9% kasus diabetes. Kedua, kasus *false negative* T3 terpusat pada rentang prediabetes karena 33,1% kasus diabetes berada di zona antara dan tidak dapat dibedakan melalui nilai glikemik. Ketiga, AUPRC T3 tertahan pada sekitar 0,88 karena sisa sinyal di zona antara hanya berasal dari fitur non-glikemik.

Dampak ketidaksesuaian label terhadap kinerja terlihat ketika data uji dipecah menurut zona yang sama. Pada T3, seluruh 1.153 kasus diabetes di zona atas terdeteksi (*recall* 1,000), sedangkan seluruh 511 *false negative* dan 42 *false positive* berada di zona antara. Di dalam zona antara, yang memuat 15.803 data uji dengan prevalensi 3,4%, AUPRC T3 hanya 0,259 dan *recall* 0,059. Dengan demikian, kinerja tinggi T3 terutama berasal dari zona yang labelnya sepenuhnya ditentukan oleh nilai glikemik, sedangkan di zona antara model hampir tidak mampu membedakan kedua kelas. Model T0–T1 tidak memakai fitur glikemik sehingga tidak mempelajari aturan zona tersebut secara langsung, tetapi tetap dipengaruhi oleh derau label yang sama. Selain itu, ambang ADA ditetapkan untuk pemeriksaan klinis yang dikonfirmasi ulang, sedangkan dataset hanya memuat satu nilai per individu tanpa konteks klinis. Ketidaksesuaian terhadap aturan ambang tersebut menandakan inkonsistensi data, bukan kesalahan diagnosis oleh tenaga kesehatan, sehingga nilai kinerja absolut pada dataset ini tidak dapat dialihkan langsung ke populasi klinis.

3.8 Pembahasan

Hasil penelitian mengindikasikan potensi skema skrining dua tingkat, meskipun potensi ini belum diuji secara klinis. Model T0–T1 belum layak dipakai sebagai alat diagnosis karena AUPRC-nya tidak lebih dari 0,33. Rendahnya nilai ini wajar, karena dengan lima atau enam fitur yang sebagian besar biner atau kategorikal, banyak individu memiliki kombinasi fitur yang identik tetapi berstatus diabetes berbeda. Meskipun demikian, model tersebut berpotensi dimanfaatkan sebagai triase non-invasif berbiaya rendah yang merujuk individu berisiko tinggi ke pemeriksaan glukosa atau HbA1c. Nilai ROC-AUC T1 sebesar 0,833 sejalan dengan batas kinerja sekitar 0,82 yang dilaporkan pada model non-invasif berbasis data survei nasional [19], sehingga temuan ini searah dengan studi pada populasi lain. Apabila diterapkan untuk triase, *threshold* T0–T1 sebaiknya diatur untuk menghasilkan *recall* tinggi, misalnya sekitar 0,79 sebagaimana dicapai oleh pembobotan kelas pada Tabel 4, dengan konsekuensi presisi yang rendah.

Kinerja tinggi pada T3 perlu ditafsirkan dengan hati-hati. Karena glukosa dan HbA1c merupakan kriteria diagnosis, model T3 sebagian besar mempelajari ulang aturan diagnosis. Akurasi 96–97% yang dilaporkan pada dataset ini [5], [6] juga hanya sekitar enam poin di atas *baseline* kelas mayoritas. Karena itu, AUPRC dan *balanced accuracy* lebih tepat digunakan untuk data dengan prevalensi di bawah 10% [15]. Dari sisi arsitektur, FT-Transformer dan ResMLP tidak melampaui XGBoost, yaitu setara pada T0–T1 dan lebih rendah secara signifikan pada T2–T3 (Tabel 7), membutuhkan komputasi GPU dengan beberapa *seed*, dan menghasilkan probabilitas yang kurang terkalibrasi. Temuan ini konsisten dengan studi *benchmark* [25], [26] dan memperkuat pilihan XGBoost untuk data klinis berdimensi rendah, seperti yang juga dipakai pada studi prediksi lain berbasis SHAP [30].

Dibandingkan penelitian terdahulu, kebaruan penelitian ini terletak pada pemisahan sinyal risiko dan sinyal diagnostik serta pengujian kestabilan penjelasan. Studi yang menggabungkan seluruh fitur [5], [6], [7] melaporkan kinerja yang setara dengan T3 pada penelitian ini, tetapi tidak dapat menunjukkan berapa besar kinerja yang tersisa apabila penanda glikemik belum tersedia. Studi yang memanfaatkan SHAP [8], [9], [11] menemukan prediktor dominan yang serupa, tetapi melaporkan penjelasan dari satu model tanpa ukuran kestabilan. Kajian kestabilan XAI sebelumnya umumnya membandingkan metode penjelasan pada satu konfigurasi fitur [21], [22] atau membandingkan tahap penyakit [23], sedangkan penelitian ini mengukur kestabilan pada empat konfigurasi fitur dengan kekuatan sinyal berbeda sehingga kestabilan penjelasan dapat dipisahkan dari kinerja prediksi. Hasil Tabel 8 melengkapi temuan tersebut dengan menunjukkan bahwa peringkat fitur teratas dapat dipercaya, sedangkan urutan fitur berkontribusi rendah perlu ditafsirkan secara hati-hati. Temuan ini berbeda dengan studi pada penyakit Alzheimer yang mendapati kestabilan penjelasan lebih rendah pada tahap awal penyakit [23]. Pada data diabetes ini, kestabilan peringkat tetap tinggi meskipun sinyal pada tahap non-invasif lemah, sehingga penjelasan model T0–T1 berpotensi dimanfaatkan sebagai bahan edukasi faktor risiko setelah divalidasi pada data klinis.

Penelitian ini memiliki beberapa keterbatasan. Pertama, dataset bersifat publik dan berpola sintetis, tanpa dokumentasi asal-usul data dan dengan 23,1% label yang tidak sesuai dengan aturan ambang ADA, sehingga validasi

Copyright © 2026 The **Juni Ismail**, Page 1487
This Journal is licensed under a Creative Commons Attribution 4.0 International License

eksternal pada data klinis Indonesia, misalnya data Posbindu [14], masih diperlukan. Kedua, kategori "No Info" pada riwayat merokok mencakup 35,8% data dan berkorelasi dengan usia muda, sehingga berpotensi menjadi *proxy* usia. Ketiga, tipe diabetes tidak dibedakan. Keempat, evaluasi data uji XGBoost dilakukan dengan satu *seed*, sedangkan variabilitasnya diwakili oleh validasi silang berulang. Kelima, validasi silang berulang dijalankan pada seluruh data bersih dengan hiperparameter yang disetel pada data latih dan validasi, sehingga nilai absolutnya sekitar 0,01 lebih tinggi daripada AUPRC data uji; ketergantungan antar-*fold* telah dikoreksi pada uji statistik, tetapi estimasi kinerja absolut tetap didasarkan pada data uji. Keenam, fitur non-invasif lain seperti lingkar pinggang, riwayat keluarga, dan aktivitas fisik belum tersedia. Ketujuh, model belum diuji dalam alur layanan klinis, sehingga manfaatnya sebagai alat triase masih bersifat potensial.

4. KESIMPULAN

Penelitian ini menganalisis kontribusi fitur bertahap dan stabilitas penjelasan SHAP pada prediksi diabetes menggunakan 96.128 rekam data publik, empat konfigurasi fitur, dan tiga model. Kontribusi setiap kelompok fitur terbukti sangat timpang: AUPRC XGBoost meningkat dari 0,266 pada T0 menjadi 0,325, 0,713, dan 0,889 pada T1 hingga T3, dengan kenaikan yang konsisten pada seluruh *fold* dan tetap signifikan setelah koreksi ketergantungan antar-*fold*. Glukosa darah memberikan tambahan terbesar (0,389), sedangkan kontribusi usia menyusut dari 58,9% menjadi 12,7% setelah penanda glikemik dimasukkan. Peringkat fitur SHAP sangat stabil antar-*run* ($\tau = 0{,}93$–$1{,}00$), antar-*seed*, dan antarmodel ($\tau \geq 0{,}87$) dengan tiga fitur teratas yang identik, dan kestabilan ini tidak mengikuti tingkat kinerja model. XGBoost setara dengan ResMLP dan FT-Transformer dalam AUPRC pada tahap non-invasif, unggul pada tahap glikemik, dan lebih terkalibrasi di seluruh tahap. Fitur demografi dan antropometri dengan demikian membawa sinyal risiko yang berpotensi dimanfaatkan untuk triase, bukan untuk diagnosis. Namun, seluruh temuan diperoleh dari dataset publik berpola sintetis yang 23,1% labelnya tidak sesuai dengan ambang ADA dan belum divalidasi pada data pasien nyata, sehingga penerapan untuk skrining memerlukan validasi eksternal pada data klinis Indonesia, penambahan fitur non-invasif seperti lingkar pinggang dan riwayat keluarga, kalibrasi ulang model, serta pengujian kestabilan penjelasan di bawah gangguan data.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada Politeknik Bisnis Indonesia dan Akademi Manajemen Informatika dan Komputer ITMI atas dukungan yang diberikan sehingga penelitian ini dapat terlaksana, serta kepada penyedia *Diabetes Prediction Dataset* yang telah membagikan data secara terbuka untuk kepentingan penelitian.

REFERENCES

[1] International Diabetes Federation, “IDF Diabetes Atlas, 11th Edition,” International Diabetes Federation, Brussels, 2025. [Online]. Available: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

[2] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, and B. B. Duncan, “IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045,” *Diabetes Research and Clinical Practice*, vol. 183, p. 109119, 2022, doi: 10.1016/j.diabres.2021.109119.

[3] Badan Kebijakan Pembangunan Kesehatan, “Survei Kesehatan Indonesia (SKI) 2023 dalam Angka,” Kementerian Kesehatan Republik Indonesia, Jakarta, 2023. [Online]. Available: https://www.badankebijakan.kemkes.go.id/hasil-ski-2023/

[4] M. Mustafa, “Diabetes Prediction Dataset,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[5] N. A. N. A. Fauzi and F. Budiman, “Analisis Perbandingan Metode Random Forest, XGBoost, dan Logistic Regression Untuk Klasifikasi Deteksi Dini Penyakit Diabetes,” *JURIKOM (Jurnal Riset Komputer)*, vol. 12, no. 6, pp. 995–1005, 2025, doi: 10.30865/jurikom.v12i6.9392.

[6] 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, p. e0310218, 2025, doi: 10.1371/journal.pone.0310218.

[7] W. Li, Y. Peng, and S. Xiong, “Diabetes Risk Prediction Model Based on Improved XGBoost with Genetic Algorithm,” in *Proceedings of the 2024 4th International Conference on Computational Modeling, Simulation and Data Analysis*, 2024, pp. 390–396. doi: 10.1145/3727993.3728059.

[8] A. B. Hidayah and A. N. Ikhsan, “Analisis Performa SMOTE dan SMOTE-TomekLink untuk Klasifikasi Diabetes: XGBoost-SHAP,” *Journal of Software Engineering and Information System (SEIS)*, vol. 6, no. 2, pp. 84–94, 2026, doi: 10.37859/seis.v6i2.11357.

[9] W. Iftikhar, M. Yaseen, G. Rahman, M. A. Nauman, and U. F. Khattak, “Improving Diabetes Prediction Accuracy and Interpretability with SMOTE and SHAP,” *Engineering, Technology & Applied Science Research*, vol. 16, no. 2, pp. 34276–34282, 2026, doi: 10.48084/etasr.16247.

[10] D. T. Ramadhan and F. Agustina, “Prediksi Diabetes Mellitus dengan Ensemble Gradient Boosting dan Advanced Feature Engineering,” *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 2, pp. 1222–1233, 2025, doi: 10.47065/bits.v7i2.8011.

Copyright © 2026 The **Juni Ismail**, Page 1488
This Journal is licensed under a Creative Commons Attribution 4.0 International License

[11] A. Nasrin, M. B. Asif, R. Naidu, A. H. Asif, M. S. Chohan, and G. A. Khan, "An Optimized and Explainable Machine Learning Framework for Diabetes Prediction Using Marine Predators Algorithm and SHAP," *Journal of Personalized Medicine*, vol. 16, no. 8, p. 423, 2026, doi: 10.3390/jpm16080423.

[12] A. E.-S. El-Bashbishy and H. M. El-Bakry, "Pediatric diabetes prediction using machine learning," *Scientific Reports*, vol. 16, no. 1, p. 1979, 2026, doi: 10.1038/s41598-025-24964-y.

[13] W. Li, Y. Peng, and K. Peng, "Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm," *PLoS One*, vol. 19, no. 9, p. e0311222, 2024, doi: 10.1371/journal.pone.0311222.

[14] M. A. Taufiqurrahman and C. D. S. Pratama, "Klasifikasi Risiko Penyakit Metabolik dengan model Ensemble, Decision Tree Dan Logistic Regression Pada Posbindu Universitas Gadjah Mada," *JURIKOM (Jurnal Riset Komputer)*, vol. 13, no. 4, pp. 1230–1245, 2026, doi: 10.30865/jurikom.v13i4.9958.

[15] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[17] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression," *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, 2022, doi: 10.1093/jamia/ocac093.

[18] American Diabetes Association Professional Practice Committee, "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025," *Diabetes Care*, vol. 48, no. Supplement_1, pp. S27–S49, 2025, doi: 10.2337/dc25-S002.

[19] B. Sun, C. Zhang, K. Ji, and X. Lu, "Beyond the Blood Draw: Explainable Machine Learning for Non-Invasive Dysglycemia Risk Screening," arXiv preprint arXiv:2606.16056, 2026, doi: 10.48550/arXiv.2606.16056.

[20] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017, pp. 4765–4774.

[21] J. Goldwasser and G. Hooker, "Statistical Significance of Feature Importance Rankings," arXiv preprint arXiv:2401.15800, 2024, doi: 10.48550/arXiv.2401.15800.

[22] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, and H. Lakkaraju, "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective," *Transactions on Machine Learning Research*, 2024, doi: 10.48550/arXiv.2202.01602.

[23] P. Guillén and E. Frias-Martinez, "Enhancing SHAP Explainability for Diagnostic and Prognostic ML Models in Alzheimer's Disease," *Computers, Materials & Continua*, vol. 87, no. 2, p. 95, 2026, doi: 10.32604/cmc.2026.076400.

[24] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting Deep Learning Models for Tabular Data," in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021, pp. 18932–18943.

[25] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Datasets and Benchmarks Track*, 2022, pp. 507–520.

[26] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022, doi: 10.1016/j.inffus.2021.11.011.

[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[29] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed.* Sebastopol, CA: O'Reilly Media, 2022.

[30] H. Gultom, A. Fauzi, and R. Ester, "Optimasi Hyperparameter XGBoost Berbasis Bayesian Optimization dan SHAP untuk Prediksi Dropout Mahasiswa Perguruan Tinggi," *JURIKOM (Jurnal Riset Komputer)*, vol. 13, no. 4, pp. 1189–1200, 2026, doi: 10.30865/jurikom.v13i4.10055.

[31] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[32] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 3rd ed.* Leanpub, 2025. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[33] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003, doi: 10.1023/A:1024068626366.

This Journal is licensed under a Creative Commons Attribution 4.0 International License