Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1694~1706
e-ISSN: 2548-6861

Two-Level Ensemble with Four Meta-Features for Diabetes Classification on Clinical Tabular Data

Farid Ma'ruf[1], Ifnu Wisma Dwi Prasetya[2], Ita Aristia Sa'ida [3]
[123] Teknik Informatika, Universitas Nahdlatul Ulama Sunan Giri Bojonegoro, Indonesia
[1]ferdianfarid393@gmail.com , [2]ifnudwi867@gmail.com, [3]itaaristia@unugiri.ac.id

Article Info

Article history:

Received 2026-01-12
Revised 2026-03-10
Accepted 2026-04-10

Keyword:

Diabetes Classification,
Clinical Tabular Data,
Two-Level Ensemble,
Meta-Features,
Xgboost,
Deep Neural Network,
SHAP,
LIME,
Calibration.

ABSTRACT

Diabetes mellitus remains a major global public health challenge due to its increasing prevalence, high risk of chronic complications, and growing burden on healthcare systems. In this context, early detection supported by artificial intelligence has become increasingly important, particularly for large-scale clinical tabular data. However, no single model consistently performs best across all clinical tabular datasets, and models with strong discriminative ability do not always provide reliable probability estimates or sufficient interpretability. This study proposes a *two-level ensemble* model with four *meta-features* for diabetes classification on clinical tabular data. At the first level, XGBoost and a *baseline Deep Neural Network* (DNN) were used as heterogeneous *base learners*. Their prediction probabilities were then transformed into four *meta-features*, namely the XGBoost probability, the DNN probability, the absolute difference between the two probabilities, and their product, which were subsequently modeled using *Logistic Regression* at the second level. The proposed model was evaluated against XGBoost, *Random Forest*, and *Baseline DNN* using *Stratified 5-Fold Cross-Validation* and an independent *hold-out test*. Performance was assessed using ROC-AUC, accuracy, precision, recall, F1-score, specificity, *Brier score*, *confusion matrix*, *threshold optimization* for screening mode, isotonic probability calibration, SHAP, LIME, and DeLong statistical testing. On the *hold-out test*, the proposed Meta Level-2 LR (4 features) achieved a ROC-AUC of 0.979451, accuracy of 0.97170, F1-score of 0.806562, precision of 0.962480, specificity of 0.997486, and the best *Brier score* of 0.022476. Although XGBoost obtained the highest ROC-AUC (0.979969), the proposed model demonstrated the most balanced overall performance, particularly in terms of F1-score, precision, specificity, calibration quality, and suitability for clinical decision support. SHAP and LIME further indicated that the most influential features were clinically plausible, especially *HbA1c_level*, *blood_glucose_level*, *age*, and *BMI*. These findings indicate that the proposed *two-level ensemble* provides a strong balance among discriminative performance, probability reliability, and interpretability, and therefore has strong potential for clinical decision support in diabetes classification.

I. PENDAHULUAN

Diabetes mellitus merupakan salah satu penyakit tidak menular yang masih menjadi tantangan utama dalam kesehatan masyarakat global [1]. Peningkatan prevalensi yang konsisten menunjukkan bahwa diabetes telah berkembang menjadi masalah sistemik dengan dampak luas terhadap pelayanan kesehatan, biaya pengobatan, dan risiko komplikasi kronis. Laporan *IDF Diabetes Atlas* edisi ke-11 menegaskan bahwa jumlah penyandang diabetes di dunia terus meningkat dan diproyeksikan bertambah signifikan pada dekade mendatang [2]. Di Indonesia, prevalensi diabetes

juga diperkirakan meningkat dari 9,19% pada tahun 2020 menjadi 16,09% pada tahun 2045, disertai kenaikan angka kematian akibat diabetes [3]. Kondisi ini menegaskan pentingnya deteksi dini yang akurat, cepat, dan efisien untuk mengurangi keterlambatan diagnosis dan risiko komplikasi.

Dalam praktik klinis, diagnosis diabetes umumnya dilakukan melalui pemeriksaan glukosa darah, HbA1c, dan evaluasi faktor risiko klinis. Meskipun pendekatan tersebut telah digunakan luas, pengembangan sistem bantu prediksi tetap diperlukan untuk mendukung proses skrining yang lebih cepat, konsisten, dan skalabel. Perkembangan *Artificial Intelligence* (AI), khususnya *machine learning* dan *deep learning*, telah membuka peluang untuk membangun model klasifikasi diabetes berbasis data klinis dengan performa yang kompetitif [4]. Berbagai algoritma, seperti *Support Vector Machine* (SVM) [5], *Random Forest* (RF) [6], *Logistic Regression* (LR) [7], dan *Artificial Neural Network* (ANN) [8], telah diterapkan pada klasifikasi diabetes dengan hasil yang cukup baik [9], [10]. Namun, tantangan pada data klinis tabular masih besar, terutama terkait ketidakseimbangan kelas, heterogenitas fitur, stabilitas prediksi, dan kemampuan generalisasi model [11].

Pada data klinis tabular, performa model sangat dipengaruhi oleh karakteristik dataset. Model berbasis pohon, seperti RF dan XGBoost, umumnya kuat dalam menangani pola nonlinier dan interaksi fitur, sedangkan model *deep learning* dapat memberikan hasil baik pada kondisi tertentu tetapi cenderung lebih sensitif terhadap konfigurasi pelatihan dan distribusi data [12], [13]. Kajian mutakhir juga menunjukkan bahwa prediksi diabetes berbasis AI masih menghadapi kendala pada aspek generalisasi, stabilitas prediksi, *class imbalance*, dan validasi pada data klinis nyata [4], [12]. Oleh karena itu, evaluasi model *deep learning* pada data tabular perlu disertai pembandingan langsung dengan model tabular yang telah dikenal kuat, khususnya XGBoost dan RF.

Salah satu pendekatan untuk meningkatkan stabilitas prediksi adalah *ensemble learning*, khususnya *stacking*, yang menggabungkan keluaran beberapa *base learner* untuk menghasilkan keputusan akhir yang lebih robust. Pada prediksi diabetes, *stacked ensemble* telah dilaporkan mampu meningkatkan akurasi dibandingkan model tunggal [14]. Namun, sebagian besar penelitian sebelumnya masih menggunakan *stacking* konvensional yang hanya menggabungkan probabilitas keluaran *base learner* melalui *meta-classifier*. Pendekatan tersebut belum banyak mengeksplorasi *meta-feature* yang secara eksplisit merepresentasikan hubungan antarprobabilitas prediksi, seperti tingkat kesepakatan, selisih, dan interaksi antar-model. Padahal, informasi tersebut berpotensi memperkuat keputusan klasifikasi pada data klinis tabular.

Penelitian ini mengusulkan arsitektur *two-level ensemble* untuk klasifikasi diabetes pada data klinis tabular dengan empat *meta-features* pada level kedua. Pada level pertama, dua *base learner* dengan karakteristik berbeda, yaitu XGBoost dan *baseline* DNN, digunakan untuk menghasilkan probabilitas prediksi. Pada level kedua, dibentuk empat *meta-features*, yaitu probabilitas XGBoost (p_{XGB}), probabilitas DNN (p_{DNN}), selisih absolut ($| p_{XGB} - p_{DNN} |$), dan hasil kali ($p_{XGB} \times p_{DNN}$), yang kemudian dimodelkan menggunakan *Logistic Regression* sebagai *meta-classifier*. Kebaruan penelitian ini terletak pada perancangan *meta-level* yang ringkas dan terarah untuk menangkap komplementaritas dua model berbeda pada data tabular klinis [12], [13]. Untuk memperjelas kontribusinya, model yang diusulkan dievaluasi terhadap *baseline* DNN, XGBoost tunggal, dan RF.

Selain performa prediktif, interpretabilitas juga menjadi aspek penting dalam penerapan AI di bidang kesehatan. Model *deep learning* dan *ensemble* umumnya bersifat *black-box*, sehingga sulit dijelaskan secara langsung. Dalam konteks klinis, hal ini menjadi kendala karena keluaran prediksi harus dapat dipahami dan dipertanggungjawabkan oleh tenaga medis. Oleh sebab itu, *Explainable Artificial Intelligence* (XAI) digunakan untuk meningkatkan transparansi dan akuntabilitas model [15]. Metode SHAP dan LIME dapat menjelaskan kontribusi fitur pada tingkat global maupun lokal, sehingga membantu menjembatani kebutuhan antara akurasi dan interpretabilitas [16], [17]. Dalam prediksi diabetes, integrasi kedua metode tersebut juga dilaporkan mampu memperjelas relevansi fitur klinis utama [17].

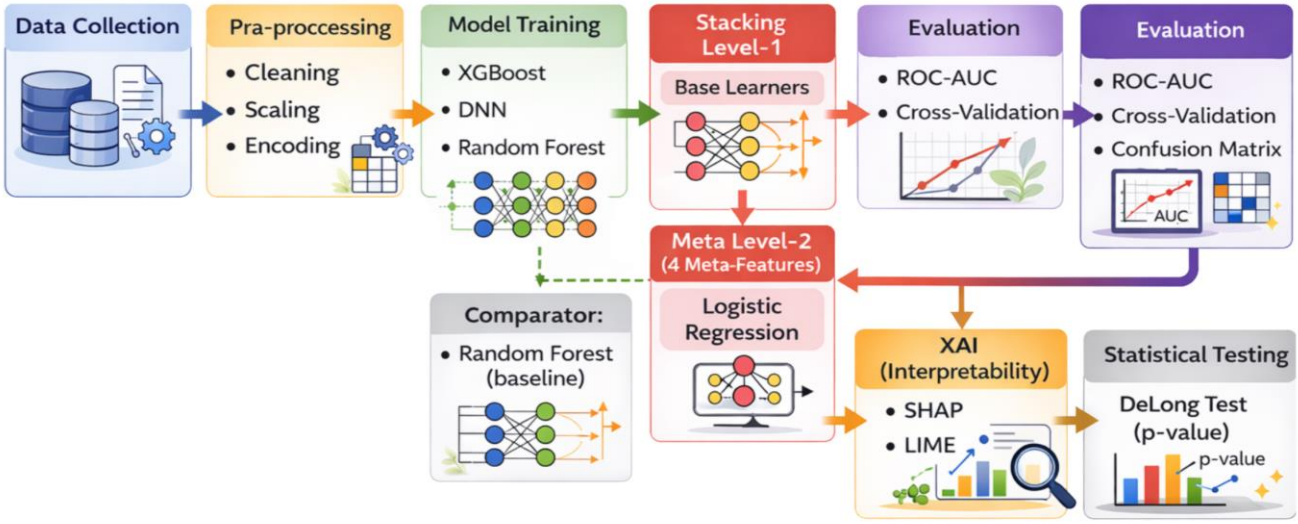
Penelitian ini berkontribusi pada tiga aspek. Pertama, mengusulkan model *two-level ensemble* dengan empat *meta-features* yang dirancang khusus untuk data klinis tabular. Kedua, menyediakan evaluasi yang lebih kuat melalui pembandingan langsung dengan XGBoost, RF, dan *baseline* DNN. Ketiga, mengintegrasikan SHAP dan LIME untuk mendukung interpretabilitas hasil prediksi. Dengan demikian, pendekatan yang diusulkan diharapkan mampu memberikan keseimbangan yang lebih baik antara performa, stabilitas, dan interpretabilitas dalam mendukung sistem pendukung keputusan medis untuk klasifikasi diabetes.

II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimen komputasional untuk membangun dan mengevaluasi model klasifikasi diabetes berbasis data klinis tabular. Fokus utama penelitian adalah pengembangan model *two-level ensemble* dengan empat *meta-features* pada level kedua, yang menggabungkan kekuatan model berbasis pohon dan jaringan saraf pada data tabular. Untuk memastikan posisi kontribusi model lebih jelas, performa model utama dibandingkan secara sistematis dengan beberapa model pembanding, yaitu XGBoost tunggal, *Random Forest* (RF), dan *baseline Deep Neural Network* (DNN).

Alur penelitian mencakup pengumpulan dataset publik, *preprocessing* data, pelatihan model pembanding, pembentukan *stacking* level-1, pembentukan meta level-2 dengan empat *meta-features*, evaluasi menggunakan *cross-validation* dan *hold-out test*, kalibrasi probabilitas, optimasi *threshold*, analisis interpretabilitas menggunakan SHAP dan LIME, serta uji statistik perbandingan ROC-AUC

menggunakan uji DeLong. Rancangan ini disusun untuk menjawab masukan reviewer terkait kejelasan sumber data, pencegahan *data leakage*, pembandingan dengan model tabular, kualitas kalibrasi probabilitas, analisis interpretabilitas global dan lokal, serta pengujian signifikansi statistik peningkatan performa [18], [19], [20].



Gambar 1 Alur penelitian

A. Sumber Data

Dataset yang digunakan adalah Diabetes Prediction Dataset yang tersedia secara publik melalui Kaggle [21]. Dataset ini terdiri atas 100.000 observasi dengan sembilan atribut, yang mencakup variabel numerik, kategorikal, biner, dan satu variabel target. Dataset dipilih karena memiliki ukuran sampel besar, representasi fitur klinis yang relevan, dan telah banyak digunakan pada studi prediksi diabetes berbasis data tabular [22], [23].

Variabel target adalah diabetes, dengan label 0 untuk non-diabetes dan 1 untuk diabetes. Fitur numerik meliputi age, bmi, HbA1c_level, dan blood_glucose_level; fitur kategorikal meliputi gender dan smoking_history; sedangkan fitur biner meliputi hypertension dan heart_disease.

TABEL I
ATRIBUT FITUR PADA DATASET DIABETES

Jenis fitur	Nama fitur
Numerik	age, bmi, HbA1c_level, blood_glucose_level
Kategorikal	gender, smoking_history
Biner	hypertension, heart_disease
Target	diabetes (0/1)

B. Pra-pemrosesan Data

1) Pembersihan data

Pembersihan data dilakukan menggunakan *listwise deletion*, yaitu penghapusan baris yang memiliki nilai hilang pada salah satu variabel. Secara formal, data bersih dinyatakan sebagai

$$D_{\text{clean}} = \{x_i \in D \mid x_i \text{ tidak mengandung nilai hilang}\} \quad (1)$$

Pendekatan ini dipilih karena sesuai untuk dataset berskala besar dengan proporsi *missing value* yang rendah, sehingga penghapusan sejumlah kecil observasi tidak mengubah representativitas data secara berarti [22], [24].

2) Pembagian data

Dataset dibagi menjadi data latih dan data uji menggunakan stratified split dengan rasio 80:20 untuk mempertahankan proporsi kelas pada kedua subset. Secara matematis, pembagian ini dituliskan sebagai

$$(D_{\text{train}}, D_{\text{test}}) = \text{Split}(D_{\text{clean}}, \text{test\_size} = 0.2) \quad (2)$$

dengan kendala

$$P(y = 1)_{\text{train}} \approx P(y = 1)_{\text{test}} \quad (3)$$

Strategi ini digunakan karena distribusi kelas pada kasus diabetes tidak seimbang, sehingga stratifikasi diperlukan agar evaluasi tidak bias terhadap kelas mayoritas [18], [25].

3) Standardisasi fitur numerik

Seluruh fitur numerik dinormalisasi menggunakan Z-score standardization yang difit pada data latih dan diterapkan ke data validasi serta data uji melalui *pipeline* yang sama. Transformasi dilakukan dengan

$$x' = \frac{x - \mu_x}{\sigma_x} \quad (4)$$

dengan $\mu_x$adalah rata-rata fitur dan $\sigma_x$adalah simpangan baku fitur pada data latih. Standardisasi diperlukan agar proses optimasi model, khususnya DNN, berlangsung lebih stabil dan tidak didominasi oleh fitur dengan skala besar [26].

4) One-hot encoding fitur kategorikal

Fitur kategorikal dikonversi menjadi representasi biner menggunakan *one-hot encoding* (OHE). Untuk setiap kategori $c_j$pada fitur kategorikal C, representasi OHE dapat ditulis sebagai

$$\mathrm{OHE}(C_i) = \begin{cases} 1, & \text{jika } C_i = c_j \\ 0, & \text{lainnya} \end{cases} \quad (5)$$

Transformasi dilakukan dengan *ColumnTransformer* agar seluruh proses *preprocessing* dipelajari hanya dari data latih dan diterapkan secara konsisten pada lipatan *cross-validation* maupun data uji akhir. Langkah ini penting untuk mencegah *data leakage* [22], [27].

5) Penanganan ketidakseimbangan kelas

Karena jumlah sampel non-diabetes lebih besar daripada sampel diabetes, penelitian ini menggunakan class weight balancing pada model yang mendukung pembobotan kelas. Bobot kelas dihitung menggunakan :

$$w_c = \frac{N}{K \cdot N_c} \quad (6)$$

dengan Nadalah jumlah sampel, Kadalah jumlah kelas, dan $N_c$adalah jumlah sampel pada kelas ke-c. Strategi ini digunakan untuk meningkatkan sensitivitas model terhadap kelas minoritas tanpa melakukan *oversampling* pada data mentah [28], [29].

C. Model yang Dievaluasi

Penelitian ini mengevaluasi tiga model pembanding dan satu model utama, yaitu:

1. XGBoost tunggal
2. *Random Forest*
3. *Baseline* DNN
4. Model utama: *two-level ensemble* dengan empat *meta-features* dan *Logistic Regression* sebagai *meta-classifier*

XGBoost dan RF dipilih sebagai pembanding utama karena keduanya dikenal kuat pada data tabular. *Baseline* DNN digunakan untuk menunjukkan kontribusi penambahan mekanisme ensembel terhadap model saraf Tunggal [12], [13].

D. Arsitektur Baseline DNN

Arsitektur *baseline* DNN yang digunakan pada penelitian ini adalah jaringan *feed-forward* sederhana untuk klasifikasi biner. Arsitektur akhir terdiri atas satu *input layer*, dua *hidden layer*, satu *dropout layer*, dan satu *output layer*. Aktivasi pada *hidden layer* menggunakan ReLU, sedangkan *output layer* menggunakan sigmoid. Berdasarkan implementasi eksperimen, DNN dilatih menggunakan *Adam optimizer* dengan *learning rate* 0,001, *binary cross-entropy loss*, *batch size* 32, *validation split* 0,2, dan maksimum 5 *epoch*. Pembobotan kelas diterapkan selama pelatihan untuk mengatasi *class imbalance* [29], [30].

TABEL II
ARSITEKTUR DAN HIPERPARAMETER BASELINE DNN

Komponen	Spesifikasi
Input layer	Jumlah neuron = jumlah fitur setelah *preprocessing*
Hidden layer 1	64 neuron, ReLU
Dropout	$p = 0.3$
Hidden layer 2	32 neuron, ReLU
Output layer	1 neuron, sigmoid
Optimizer	Adam
Learning rate	0,001
Loss function	*Binary cross-entropy*
Batch size	32
Epoch	5
Validation split	0,2
Class handling	*Class weight balancing*

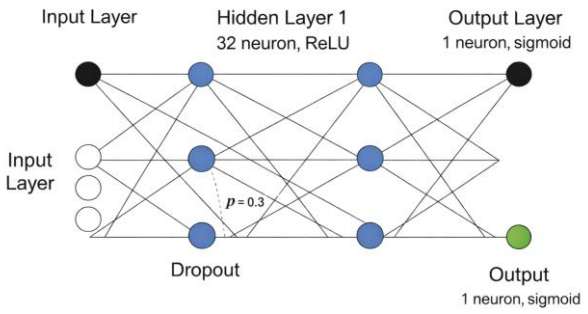
Fungsi aktivasi sigmoid pada *output layer* digunakan untuk memetakan keluaran model ke dalam rentang 0 sampai 1 sehingga dapat ditafsirkan sebagai probabilitas kelas positif. Dalam penelitian ini, probabilitas tersebut merepresentasikan kemungkinan suatu sampel termasuk ke dalam kelas diabetes. Fungsi ini sesuai untuk tugas klasifikasi biner karena menghasilkan keluaran probabilistik yang mudah diinterpretasikan.:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (7)$$

Sedangkan fungsi kerugian yang digunakan adalah *binary cross-entropy* (BCE):

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (8)$$

dengan $y_i$adalah label aktual dan $\hat{y}_i$adalah probabilitas prediksi kelas positif.



Gambar 2. Arsitektur baseline DNN

E. Pembentukan Model Utama: Two-Level Ensemble dengan Empat Meta-Features

Model utama pada penelitian ini adalah two-level ensemble dengan empat meta-features pada level kedua. Pada level pertama, dua base learner dengan karakteristik berbeda, yaitu XGBoost dan baseline DNN, digunakan untuk menghasilkan probabilitas prediksi terhadap kelas diabetes pada setiap sampel. Rancangan ini dipilih untuk menggabungkan kekuatan model berbasis pohon dan jaringan saraf pada data klinis tabular.

Untuk mencegah data leakage, pembentukan meta-feature dilakukan menggunakan out-of-fold (OOF) prediction dari data latih melalui Stratified 5-Fold Cross-Validation. Dengan demikian, probabilitas yang digunakan untuk melatih meta-classifier selalu berasal dari model yang tidak melihat sampel validasi pada lipatan tersebut. Skema ini penting untuk memastikan bahwa informasi dari data uji maupun dari lipatan validasi tidak bocor ke tahap pelatihan meta-level [19], [20].

Misalkan p_{XGB} adalah probabilitas prediksi dari XGBoost dan p_{DNN} adalah probabilitas prediksi dari DNN. Empat *meta-features* yang dibentuk pada level kedua adalah:

$$z_1 = p_{XGB} \quad (9)$$
$$z_2 = p_{DNN} \quad (10)$$
$$z_3 = | p_{XGB} - p_{DNN} | \quad (11)$$
$$z_4 = p_{XGB} \times p_{DNN} \quad (12)$$

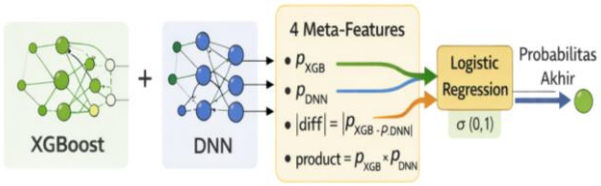
Keempat fitur tersebut kemudian disusun sebagai vektor *meta-feature*

$$\mathrm{z} = [z_1, z_2, z_3, z_4] \quad (13)$$

Selanjutnya, *meta-classifier* LR digunakan untuk menghasilkan probabilitas akhir:

$$\hat{p} = \sigma(\beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4) \quad (14)$$

Rancangan ini dipilih karena empat *meta-features* tersebut tidak hanya mewakili probabilitas dari masing-masing *base learner*, tetapi juga merepresentasikan tingkat kesepakatan dan interaksi prediksi antar-model. Dengan demikian, level kedua tidak sekadar menggabungkan dua probabilitas mentah, tetapi memanfaatkan informasi komplementer yang lebih kaya untuk meningkatkan stabilitas prediksi pada data tabular klinis [14], [31].



Gambar 3. Arsitektur two-level ensemble dengan empat meta-features pada level kedua.

F. Protokol Evaluasi dan Pencegahan Data Leakage

Protokol evaluasi dirancang dalam dua tahap. Tahap pertama adalah internal validation menggunakan Stratified 5-Fold Cross-Validation pada data latih untuk memperoleh OOF prediction, nilai rerata, dan simpangan baku performa antar-lipatan. Tahap kedua adalah hold-out test pada 20% data yang dipisahkan sejak awal dan tidak digunakan sama sekali pada proses pelatihan model maupun pembentukan *meta-feature*. Dengan desain ini, pengukuran performa akhir dilakukan pada data yang benar-benar tidak terlihat selama pelatihan [32].

Seluruh *preprocessing* termasuk standardisasi dan OHE difit hanya pada data latih di setiap lipatan, lalu diterapkan ke data validasi dan data uji menggunakan *pipeline* yang sama. Demikian pula, *meta-classifier* level kedua dilatih hanya menggunakan OOF prediction dari data latih. Pendekatan ini digunakan untuk memastikan bahwa tidak terjadi *data leakage* antara *base learner*, *meta-model*, dan data uji akhir [32], [33].

G. Metrik Evaluasi

Evaluasi model dilakukan menggunakan metrik klasifikasi biner yang umum dipakai pada studi medis, yaitu *accuracy*, *precision*, *recall* (*sensitivity*), *specificity*, *F1-score*, ROC-AUC, *confusion matrix*, dan *Brier score*. Metrik ini dipilih agar performa model dapat dinilai tidak hanya dari ketepatan klasifikasi, tetapi juga dari kemampuan diskriminatif dan kualitas probabilitas yang dihasilkan [20], [29].

Beberapa rumus utama yang digunakan adalah sebagai berikut.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$
$$\text{Precision} = \frac{TP}{TP + FP} \quad (16)$$
$$\text{Recall} = \frac{TP}{TP + FN} \quad (17)$$
$$\text{Specificity} = \frac{TN}{TN + FP} \quad (18)$$
$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

ROC-AUC digunakan untuk menilai kemampuan model membedakan kelas positif dan negatif pada seluruh kemungkinan *threshold*. Pada penelitian ini, selain melaporkan hasil pada *threshold* 0,5, model juga dievaluasi pada skenario *screening* berbasis optimasi *threshold*.

H. Optimasi Threshold untuk Skenario Screening

Penelitian ini menambahkan evaluasi pada skenario *screening*. Nilai *threshold* tidak ditentukan langsung dari data uji, tetapi dipilih dari OOF prediction pada data latih agar tetap bebas *leakage*. *Threshold* optimum ditetapkan berdasarkan target *recall* klinis yang tinggi, sehingga model tetap sensitif dalam mendeteksi pasien yang berisiko diabetes.

Setelah *threshold* dipilih dari OOF prediction, nilai tersebut diterapkan satu kali pada *hold-out test* untuk memperoleh metrik akhir pada skenario *screening*. Dengan prosedur ini, evaluasi tetap adil dan tidak menggunakan informasi dari data uji untuk menentukan keputusan klasifikasi [33], [34].

I. Kalibrasi Probabilitas

Pada aplikasi klinis, probabilitas prediksi tidak cukup hanya diskriminatif, tetapi juga harus terkalibrasi dengan baik agar dapat ditafsirkan sebagai estimasi risiko. Oleh sebab itu, penelitian ini menggunakan Isotonic Regression untuk kalibrasi probabilitas. Sesuai prinsip anti-*leakage*, model kalibrasi difit menggunakan probabilitas OOF dari data latih, lalu diterapkan pada probabilitas data uji.

Kualitas kalibrasi dievaluasi menggunakan *Brier score* dan *calibration curve/reliability diagram*. *Brier score* dihitung dengan :

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \qquad (20)$$

dengan $p_i$adalah probabilitas prediksi dan $y_i$adalah label aktual. Nilai yang lebih kecil menunjukkan probabilitas yang lebih akurat dan lebih terkalibrasi [29].

J. Uji Statistik Perbandingan Kinerja

Untuk menilai apakah perbedaan ROC-AUC antar-model bermakna secara statistik, penelitian ini menggunakan uji DeLong pada probabilitas prediksi di *hold-out test*. Uji ini digunakan khususnya untuk membandingkan model utama dengan XGBoost tunggal dan *baseline* DNN, sehingga peningkatan performa tidak hanya dilaporkan secara deskriptif, tetapi juga diuji signifikansinya [35].

K. Interpretabilitas Model

Interpretabilitas dianalisis menggunakan dua pendekatan XAI, yaitu SHAP dan LIME. SHAP digunakan untuk menghasilkan *global feature importance ranking* dan menilai arah kontribusi fitur utama terhadap prediksi model. Dengan cara ini, Analisis ini memungkinkan identifikasi kesesuaian antara fitur-fitur dominan yang dipelajari model dan pengetahuan klinis yang telah established, terutama pada variabel HbA1c_level, blood_glucose_level, bmi, dan age. [15], [17].

Secara umum, nilai SHAP untuk fitur ke-i dinyatakan sebagai :

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)] \qquad (21)$$

Sementara itu, LIME digunakan untuk menjelaskan prediksi pada level individu, khususnya pada contoh *true positive* dan *false positive*, agar dapat dianalisis apakah alasan prediksi model masuk akal secara klinis. Secara formal, LIME dinyatakan sebagai

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (22)$$

dengan $f$adalah model kompleks, $g$adalah model interpretable lokal, $\pi_x$adalah fungsi kedekatan lokal, dan $\Omega(g)$adalah penalti kompleksitas model sederhana.

Penggunaan SHAP dan LIME secara bersamaan dimaksudkan untuk memberikan penjelasan pada dua level sekaligus: global untuk pemeringkatan fitur dan lokal untuk penjelasan keputusan pada kasus individual [15], [17].

III. HASIL DAN PEMBAHASAN

A. Hasil Evaluasi Model pada Hold-out Test

Evaluasi utama dilakukan pada data uji (hold-out test) menggunakan threshold 0,5. Sesuai fokus penelitian, model utama yang dianalisis adalah Meta Level-2 LR (4 fitur), sedangkan model pembanding meliputi XGBoost tunggal, Random Forest, Baseline DNN (MLP). Hasil pengujian menunjukkan bahwa model utama memberikan keseimbangan performa terbaik pada metrik klasifikasi utama, khususnya F1-score, precision, specificity, dan Brier score, meskipun ROC-AUC tertinggi masih dicapai oleh XGBoost tunggal dengan selisih yang sangat kecil.

TABEL IV
HASIL EVALUASI MODEL PADA DATA UJI (HOLD-OUT TEST) DENGAN THRESHOLD 0,5

Model	TEST_AUC	ACC	F1	Precision	Recall	Specificity	Brier
XGBoost single	0.979969	0.90640	0.626198	0.474002	0.922353	0.904918	0.058080
Meta Level-2 LR (4 fitur)	0.979451	0.97170	0.806562	0.962480	0.694118	0.997486	0.022476
Baseline DNN (MLP)	0.976369	0.96560	0.746313	1.000000	0.595294	1.000000	0.031521
Random Forest	0.966999	0.96935	0.793115	0.930325	0.691176	0.995191	0.025093

Berdasarkan Tabel IV, XGBoost tunggal menghasilkan AUC tertinggi dan recall tertinggi, yang menunjukkan kemampuan diskriminatif dan sensitivitas yang sangat baik pada data tabular. Namun, model ini memiliki precision yang rendah dan Brier score yang paling buruk, sehingga kurang ideal apabila probabilitas prediksi hendak digunakan sebagai estimasi risiko klinis.

Sebaliknya, Meta Level-2 LR (4 fitur) memberikan F1-score tertinggi, precision yang sangat tinggi, specificity yang sangat tinggi, serta Brier score terbaik. Hasil ini menunjukkan bahwa model utama memberikan prediksi yang lebih seimbang antara kemampuan mendeteksi kasus positif dan kemampuan menekan false positive. Dalam konteks klinis, profil seperti ini lebih sesuai untuk sistem pendukung keputusan yang memerlukan ketepatan prediksi tinggi dan probabilitas yang lebih reliabel.

Model Baseline DNN menunjukkan karakter yang sangat konservatif pada threshold 0,5. Walaupun precision dan specificity mencapai nilai sangat tinggi, recall menurun cukup nyata. Hal ini menandakan bahwa model DNN tunggal cenderung hanya memberi label positif pada kasus yang sangat yakin, sehingga lebih banyak kasus positif yang terlewat.

B. Stabilitas Performa Berdasarkan Cross-Validation

Untuk menjawab permintaan reviewer terkait stabilitas model, penelitian ini melaporkan hasil Stratified 5-Fold Cross-Validation dalam bentuk rerata ± standar deviasi berdasarkan prediksi OOF (out-of-fold). Pelaporan ini penting untuk menunjukkan bahwa performa model tidak hanya bergantung pada satu pembagian data.

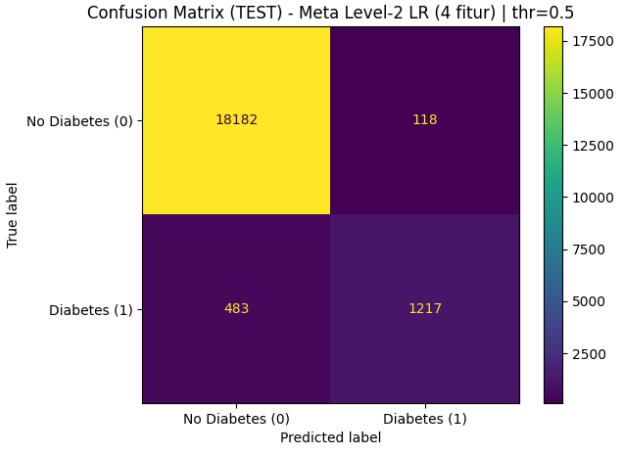
TABEL V
HASIL CROSS-VALIDATION BERUPA RERATA ± STANDAR DEVIASI BERDASARKAN PREDIKSI OOF (THRESHOLD = 0,5)

Model	AUC (mean ± SD)	ACC (mean ± SD)	F1-score (mean ± SD)	Precision (mean ± SD)	Recall (mean ± SD)	Brier (mean ± SD)
XGBoost	0,979712 ± 0,001333	0,906963 ± 0,003491	0,627286 ± 0,009324	0,475734 ± 0,010072	0,920735 ± 0,006248	0,058062 ± 0,001959
Meta Level-2 (4 fitur)	0,979427 ± 0,001380	0,970025 ± 0,001166	0,799399 ± 0,007575	0,927166 ± 0,012544	0,702647 ± 0,008198	0,023646 ± 0,000924
Random Forest	0,976234 ± 0,001270	0,915850 ± 0,002533	0,643563 ± 0,007131	0,502921 ± 0,008365	0,893529 ± 0,004269	0,055075 ± 0,001459
Base DNN	0,975429 ± 0,001651	0,965650 ± 0,000844	0,746736 ± 0,007768	1,000000 ± 0,000000	0,595882 ± 0,009928	0,031957 ± 0,000710

Tabel V menunjukkan bahwa seluruh model memiliki nilai AUC yang tinggi dengan standar deviasi yang kecil, sehingga performanya relatif stabil antar-lipatan. Model Meta Level-2 (4 fitur) menonjol karena memberikan nilai ACC, F1-score, dan Brier score terbaik. Hal ini menunjukkan bahwa model utama tidak hanya memiliki performa klasifikasi yang baik, tetapi juga kualitas probabilitas yang lebih reliabel. Dengan demikian, keunggulan model utama tidak hanya terlihat pada *hold-out test*, tetapi juga konsisten pada evaluasi *cross-validation*.

C. Analisis Confusion Matrix dan Implikasi Klinis

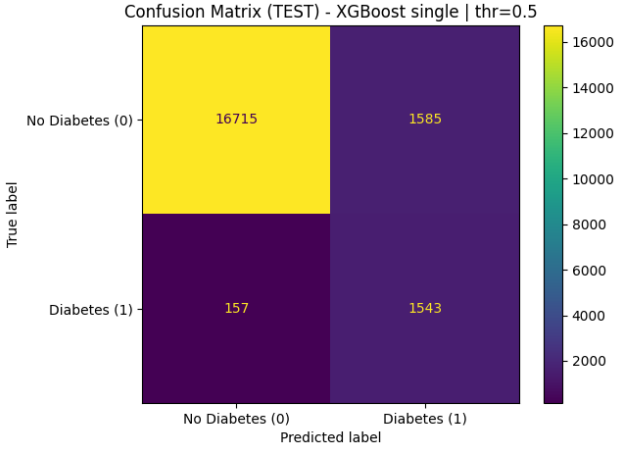
Untuk memperjelas interpretasi klinis, penelitian ini menyajikan confusion matrix dari model utama dan model pembanding pada data uji. Visualisasi ini penting karena metrik agregat seperti AUC atau F1-score belum sepenuhnya menunjukkan distribusi kesalahan klasifikasi. Melalui confusion matrix, pola kesalahan prediksi setiap model dapat diamati secara lebih rinci. Hal ini membantu menilai keseimbangan antara sensitivitas dan ketepatan klasifikasi.



Gambar 4. Confusion matrix model Meta Level-2 LR (4 fitur)

Pada model utama, distribusi prediksi menunjukkan bahwa jumlah false positive dapat ditekan secara sangat baik, sementara jumlah true positive tetap berada pada tingkat yang memadai. Karakteristik ini selaras dengan tingginya precision dan specificity yang diperoleh model.

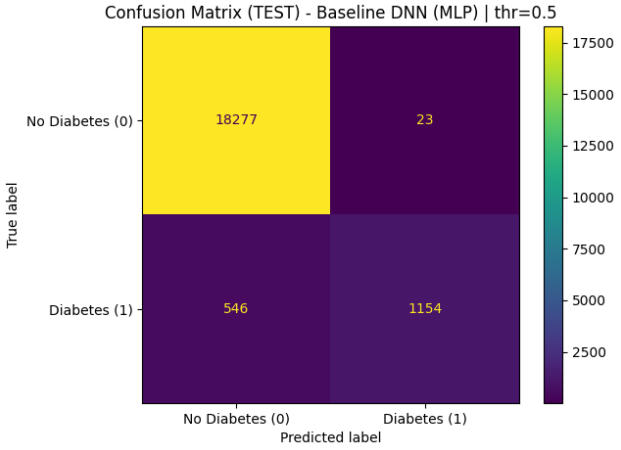
Model XGBoost tunggal menghasilkan recall yang sangat tinggi, tetapi jumlah false positive juga meningkat. Hal ini menjelaskan mengapa precision model tersebut lebih rendah dibanding model utama.



Gambar 5. Confusion matrix model XGBoost

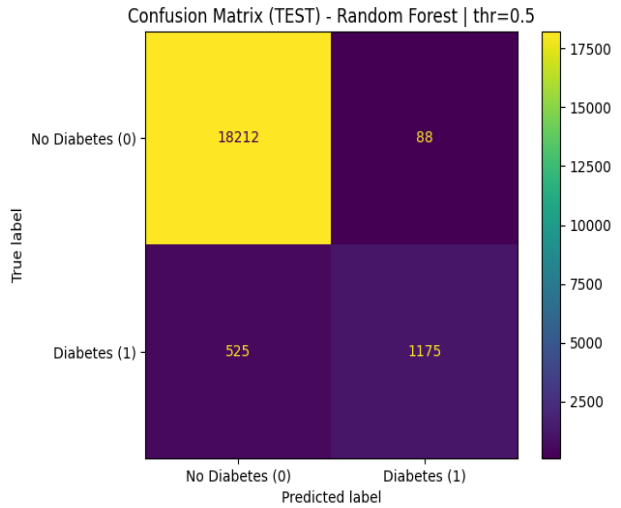
Dari perspektif klinis, pola ini dapat diterima untuk tahap skrining awal, tetapi kurang ideal untuk tahap keputusan lanjut karena berpotensi meningkatkan pemeriksaan lanjutan yang tidak perlu.

Model Baseline DNN menunjukkan pola yang berlawanan. Model ini sangat selektif dalam memprediksi kelas positif, sehingga false positive sangat rendah, tetapi jumlah false negative meningkat. Temuan ini menjelaskan rendahnya recall model DNN tunggal pada data uji.



Gambar 6. Confusion matrix model Baseline DNN (MLP)

Untuk Random Forest, diperoleh TN = 18.212, FP = 88, FN = 525, dan TP = 1.175. Hasil ini menunjukkan bahwa Random Forest cukup baik dalam menekan false positive, tetapi masih sedikit di bawah model utama dari sisi F1-score dan kualitas probabilitas. Secara keseluruhan, analisis confusion matrix memperlihatkan adanya trade-off klinis yang jelas. XGBoost lebih unggul untuk skenario yang memprioritaskan sensitivitas tinggi, sedangkan Meta Level-2 LR (4 fitur) lebih unggul ketika tujuan utamanya adalah menghasilkan keputusan yang lebih presisi dan lebih reliabel. Oleh karena itu, model Meta Level-2 layak dijadikan metode utama pada penelitian ini.



Gambar 7. Confusion matrix model Random Forest

D. Optimasi Threshold untuk Mode Skrining

Untuk menjawab masukan reviewer terkait threshold optimization, penelitian ini menambahkan evaluasi pada mode skrining, yaitu skenario ketika model ditujukan untuk menangkap sebanyak mungkin kasus positif. Pada skenario ini, threshold ditentukan dari prediksi OOF agar tetap bebas data leakage, kemudian diuji satu kali pada data uji.

TABEL VI
HASIL EVALUASI MODE SKRINING PADA MODEL META LEVEL-2 (4 FITUR), THRESHOLD DIPILIH DARI OOF

Model	*Threshold* OOF	TEST AUC	TEST Recall	TEST Precision	TEST F1-score	TEST Specificity
Meta Level-2 (4 fitur)	0,05	0,979940	0,944706	0,434877	0,595587	0,885956

Tabel VI menunjukkan bahwa penggunaan *threshold* 0,05 meningkatkan *recall* pada data uji menjadi 0,944706, sehingga model mampu mendeteksi hampir seluruh kasus positif. Akan tetapi, peningkatan sensitivitas tersebut disertai penurunan *precision* dan *specificity*, yang menunjukkan bahwa model menjadi lebih longgar dalam memberikan label positif. Hasil ini menegaskan bahwa model utama memiliki fleksibilitas penggunaan sesuai konteks klinis. Pada tahap skrining awal, *threshold* yang lebih rendah lebih sesuai untuk meminimalkan *false negative*, sedangkan pada tahap konfirmasi atau triase lanjutan, *threshold* 0,5 lebih tepat karena menghasilkan prediksi yang lebih selektif..

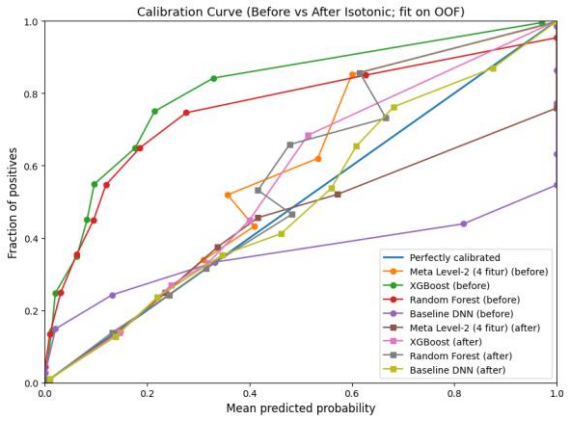
E. Kalibrasi Probabilitas

Selain diskriminasi klasifikasi, penelitian ini juga mengevaluasi kualitas probabilitas prediksi. Kalibrasi dilakukan menggunakan Isotonic Regression. Hasilnya dinilai menggunakan Brier score dan divisualisasikan dengan calibration curve.

TABEL VII
PERBANDINGAN BRIER SCORE SEBELUM DAN SESUDAH KALIBRASI

Model	Brier sebelum	Brier sesudah	Delta (sebelum-sesudah)
Meta Level-2 (4 fitur)	0.022530	0.022445	0.000085
Baseline DNN	0.031777	0.024479	0.007298
Random Forest	0.056353	0.023675	0.032677
XGBoost	0.058080	0.022379	0.035702

Berdasarkan Tabel VII, model Meta Level-2 sejak awal telah menunjukkan kualitas probabilitas yang baik, sehingga penurunan Brier score setelah kalibrasi relatif kecil. Sebaliknya, XGBoost dan Random Forest mengalami penurunan Brier score yang lebih besar setelah proses kalibrasi. Temuan ini menunjukkan bahwa model berbasis pohon memiliki kemampuan diskriminatif yang tinggi, tetapi kualitas probabilitas awalnya masih berada di bawah model utama. Dengan demikian, kalibrasi memberikan dampak yang lebih nyata pada model-model tersebut dibandingkan pada Meta Level-2. Hasil ini memperkuat bahwa model utama tidak hanya unggul dari sisi klasifikasi, tetapi juga lebih baik dalam menghasilkan estimasi probabilitas yang konsisten.



Gambar 8.Calibration curve sebelum dan sesudah kalibrasi isotonic.

Dari sudut pandang klinis, hasil ini sangat penting. Model dengan *AUC* tinggi belum tentu menghasilkan probabilitas yang akurat sebagai estimasi risiko. Karena itu, keunggulan *Meta Level-2 pada Brier score* memperkuat argumen bahwa model ini lebih sesuai digunakan sebagai alat bantu estimasi risiko, bukan hanya sebagai pengklasifikasi biner.

F. Uji Signifikansi Statistik Menggunakan DeLong

Untuk memastikan bahwa peningkatan performa tidak hanya bersifat deskriptif, penelitian ini melakukan uji DeLong pada ROC-AUC.

Hasil pada Tabel VIII menunjukkan bahwa:

- XGBoost memiliki AUC yang secara statistik lebih tinggi daripada Baseline DNN.
- Meta Level-2 memiliki AUC yang secara statistik lebih tinggi daripada Baseline DNN.
- XGBoost juga memiliki AUC yang sedikit lebih tinggi daripada Meta Level-2, dan perbedaan tersebut signifikan secara statistik.

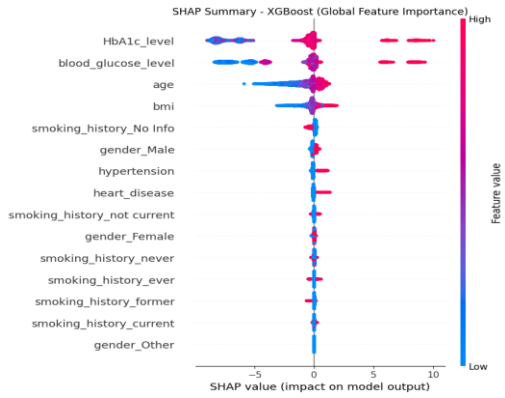
Temuan ini perlu dilaporkan secara proporsional. Oleh karena itu, penelitian ini tidak mengklaim bahwa model utama memiliki AUC tertinggi mutlak. Klaim yang lebih tepat adalah bahwa Meta Level-2 LR (4 fitur) memberikan trade-off terbaik antara AUC yang sangat tinggi, F1-score tertinggi, precision tinggi, specificity tinggi, dan Brier score terbaik.

TABEL VIII
HASIL UJI DELONG UNTUK PERBANDINGAN ROC-AUC PADA DATA UJI

Model A	Model B	AUC_A	AUC_B	Delta_A_minus_B	p-value
XGBoost	Base DNN	0.979969	0.976369	0.003600	< 0.001
Base DNN	Meta Level-2 (4 fitur)	0.976369	0.979451	-0.003082	< 0.001
XGBoost	Meta Level-2 (4 fitur)	0.979969	0.979451	0.000518	0.016

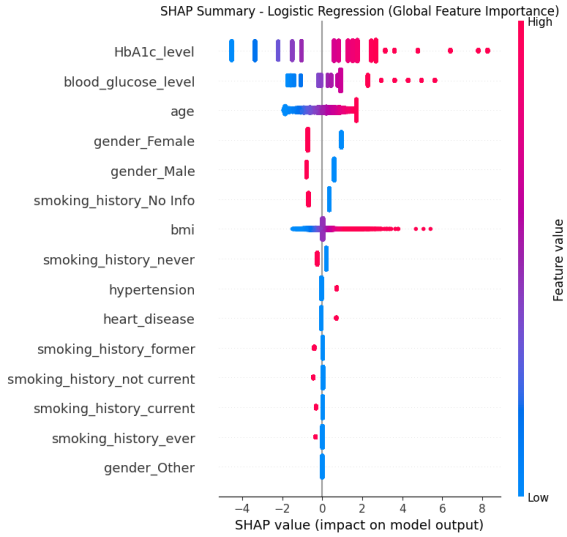
G. Analisis Interpretabilitas Global Menggunakan SHAP

Dalam mengidentifikasi feature importance ranking secara global, penelitian ini menggunakan SHAP karena metode ini dapat memberikan estimasi kontribusi setiap fitur terhadap keluaran model secara komprehensif. Visualisasi SHAP menunjukkan bahwa fitur-fitur yang paling berpengaruh terhadap prediksi model konsisten dengan pengetahuan klinis, terutama HbA1c_level, blood_glucose_level, age, dan bmi. Pada kedua model, HbA1c_level dan blood_glucose_level muncul secara konsisten sebagai fitur teratas. Temuan ini selaras dengan literatur medis karena kedua variabel tersebut merupakan indikator biokimia utama dalam evaluasi diabetes.



Gambar 9. SHAP summary plot model XGBoost.

Fitur age dan bmi juga memberikan kontribusi besar, yang sesuai dengan bukti epidemiologis bahwa peningkatan usia dan obesitas berkaitan erat dengan peningkatan risiko diabetes.



Gambar 10. SHAP summary plot model Meta Level-2 LR (4 fitur).

Arah pengaruh fitur pada plot SHAP juga konsisten secara klinis. Nilai HbA1c dan glukosa darah yang tinggi cenderung mendorong prediksi ke arah kelas diabetes. Dengan demikian, SHAP tidak hanya menunjukkan fitur terpenting secara global, tetapi juga memperkuat validitas klinis model karena pola yang dipelajari sesuai dengan mekanisme penyakit yang telah diketahui.

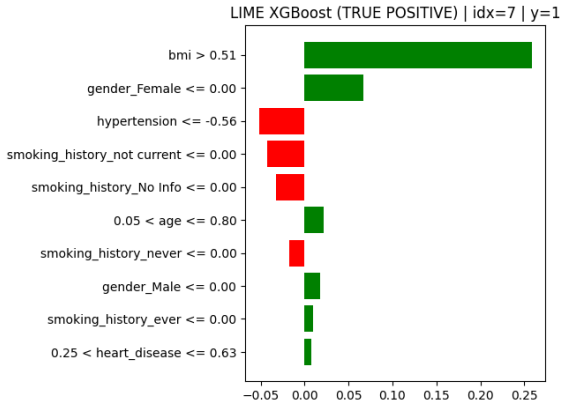
H. Analisis Interpretabilitas Lokal Menggunakan LIME

Selain analisis global, penelitian ini menggunakan LIME untuk menjelaskan prediksi pada level individu. Analisis

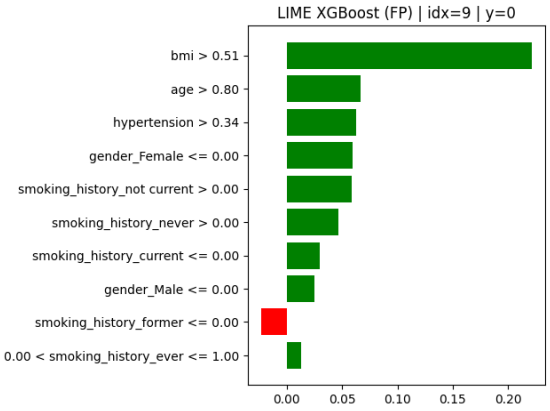
dilakukan pada contoh true positive dan false positive agar dapat dilihat apakah alasan prediksi model masih masuk akal secara klinis.

1) LIME pada model XGBoost

Pada kasus true positive, fitur dominan yang mendorong prediksi positif pada model XGBoost adalah bmi tinggi, usia, dan beberapa indikator komorbiditas. Pada kasus false positive, model tetap dipengaruhi oleh bmi tinggi, usia lanjut, dan hipertensi, meskipun label aktual adalah non-diabetes.



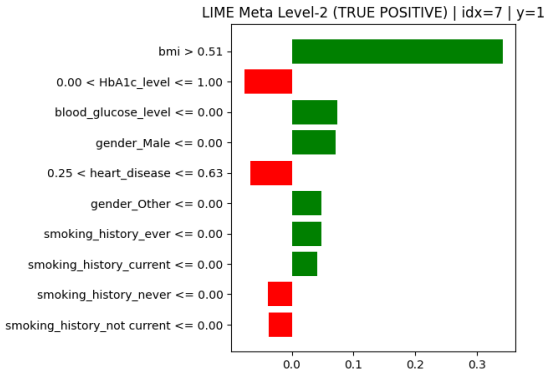
Gambar 11. Visualisasi LIME model XGBoost pada contoh true positive.



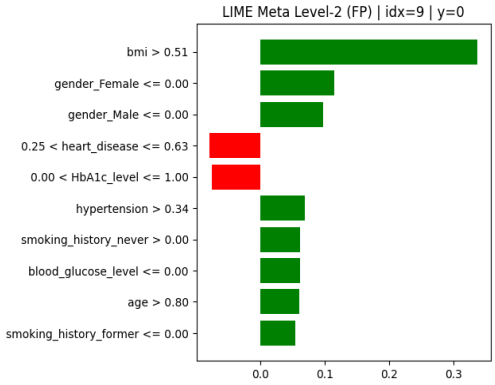
Gambar 12. Visualisasi LIME model XGBoost pada contoh false positive.

2) LIME pada model Meta Level-2

Pada model Meta Level-2, penjelasan LIME pada kasus true positive menunjukkan bahwa bmi, blood_glucose_level, dan indikator metabolik lain menjadi pendorong utama prediksi positif. Pada kasus false positive, faktor seperti bmi tinggi, gender, heart_disease, HbA1c, hypertension, blood_glucose_level, dan age tetap muncul sebagai faktor dominan.



Gambar 13. Visualisasi LIME model Meta Level-2 LR (4 fitur) pada contoh true positive



Gambar 14. Visualisasi LIME model Meta Level-2 LR (4 fitur) pada contoh false positive.

Secara umum, hasil LIME tidak menunjukkan konflik substantif dengan SHAP. Keduanya sama-sama menempatkan bmi, HbA1c_level, blood_glucose_level, dan age sebagai faktor penting. Perbedaannya hanya terletak pada level analisis. SHAP memberikan pemeringkatan global, sedangkan LIME menjelaskan alasan lokal pada kasus individual. Dengan demikian, integrasi SHAP dan LIME pada penelitian ini bersifat saling melengkapi.

I. Pembahasan

Hasil penelitian ini menunjukkan bahwa peralihan dari stacking konvensional menuju arsitektur two-level ensemble dengan empat meta-features memberikan kontribusi yang lebih jelas dan terukur. Kebaruan model tidak terletak pada penggunaan stacking semata, melainkan pada rancangan level kedua yang memanfaatkan dua probabilitas dasar serta dua fitur turunan, yaitu selisih absolut dan hasil kali probabilitas, untuk menangkap komplementaritas antara XGBoost dan DNN.

Hasil komparasi menunjukkan bahwa model berbasis pohon, khususnya XGBoost dan Random Forest, tetap memiliki kinerja yang sangat kuat pada data tabular. XGBoost bahkan menghasilkan nilai AUC tertinggi pada pengujian.

Namun, model utama Meta Level-2 LR (4 fitur) memberikan F1-score tertinggi, precision tinggi, specificity sangat tinggi, serta Brier score terbaik, sehingga menunjukkan keseimbangan performa yang lebih baik secara keseluruhan. Temuan ini menegaskan bahwa keunggulan pendekatan yang diusulkan tidak terletak pada dominasi mutlak terhadap seluruh model pembanding, tetapi pada kemampuannya mengombinasikan kekuatan model tabular dan DNN untuk menghasilkan prediksi yang lebih seimbang dan lebih relevan bagi kebutuhan klinis.

Dari perspektif klinis, model yang diusulkan memiliki dua diturunkan agar recall meningkat dan kasus positif tidak banyak terlewat. Sebaliknya, pada tahap konfirmasi atau triase lanjutan, threshold 0,5 lebih sesuai karena menghasilkan prediksi yang lebih selektif dengan jumlah false positive yang lebih rendah. Fleksibilitas ini menunjukkan bahwa model memiliki potensi untuk diintegrasikan ke dalam alur kerja klinis sesuai dengan tujuan penggunaannya.

Dari sisi kompleksitas, model utama memang lebih kompleks dibandingkan model tunggal seperti Baseline DNN, XGBoost, atau Random Forest karena melibatkan dua tingkat pemodelan. Namun, peningkatan kompleksitas tersebut diikuti oleh perbaikan performa yang nyata, terutama pada F1-score, precision, dan kualitas kalibrasi probabilitas. Selain itu, level kedua pada arsitektur yang diusulkan hanya memanfaatkan empat meta-features, sehingga tambahan kompleksitas yang diperkenalkan tetap relatif terbatas dan masih dapat dipertanggungjawabkan dari sisi implementasi.

Penelitian ini masih memiliki beberapa keterbatasan. Evaluasi dilakukan pada satu dataset publik sehingga validasi eksternal pada dataset independen tetap diperlukan untuk menilai generalisasi model secara lebih luas. Selain itu, analisis LIME masih didasarkan pada kasus-kasus terpilih sehingga belum sepenuhnya merepresentasikan seluruh variasi pola pasien. Walaupun demikian, rangkaian hasil yang diperoleh menunjukkan bahwa model utama telah memenuhi beberapa kriteria penting untuk sistem pendukung keputusan klinis, yaitu performa prediktif yang tinggi, probabilitas yang reliabel, dan interpretabilitas yang memadai.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa model *two-level ensemble* dengan empat *meta-features* efektif untuk klasifikasi diabetes pada data klinis tabular karena mampu memberikan keseimbangan yang baik antara kemampuan diskriminatif, ketepatan klasifikasi, reliabilitas probabilitas, dan interpretabilitas. Dibandingkan dengan *Baseline DNN*, XGBoost, dan *Random Forest*, model Meta Level-2 LR (4 fitur) memberikan performa paling seimbang, terutama pada F1-score, *precision*, *specificity*, dan *Brier score*, meskipun ROC-AUC tertinggi tetap dicapai oleh XGBoost. Hasil *cross-validation*, optimasi *threshold*, kalibrasi isotonic, serta analisis SHAP dan LIME semakin memperkuat bahwa model yang diusulkan memiliki kinerja yang stabil, probabilitas prediksi yang baik, dan interpretasi yang konsisten dengan pengetahuan klinis. Dengan demikian, model ini berpotensi dikembangkan sebagai dasar sistem pendukung keputusan klinis untuk klasifikasi diabetes, dengan kebutuhan validasi eksternal pada penelitian selanjutnya.

DAFTAR PUSTAKA

[1] F. Najafi *et al.*, “The incidence of diabetes mellitus and its determining factors in a Kurdish population : insights from a cohort study in western Iran,” *Sci. Rep.*, pp. 1–11, 2024, doi: 10.1038/s41598-024-66795-3.

[2] *International Diabetes Federation, IDF Diabetes Atlas, 11th ed. Brussels, Belgium: International Diabetes Federation, 2025. Available: IDF Diabetes Atlas official website.* 2025.

[3] M. Wahidin *et al.*, “Projection of diabetes morbidity and mortality till 2045 in Indonesia based on risk factors and NCD prevention and control programs,” *Sci. Rep.*, pp. 1–17, 2024, doi: 10.1038/s41598-024-54563-2.

[4] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, “Machine learning and deep learning predictive models for type 2 diabetes: a systematic review,” *Diabetol. Metab. Syndr.*, vol. 13, no. 1, 2021, doi: 10.1186/s13098-021-00767-9.

[5] G. R. D. E. A. Lgorithm, “DIABETES PREDICTION USING SUPPORT VECTOR,” vol. 8, no. 1, pp. 44–52, 2024.

[6] C. N. Noviyanti, “Journal of Information System Early Detection of Diabetes Using Random Forest Algorithm,” vol. 2, no. 1, pp. 41–48, 2024.

[7] R. Hidayat, D. Mahdiana, and A. Fergina, “Comparative Analysis of Logistic Regression , SVM , Xgboost , and Random Forest Algorithms for Diabetes Classification,” vol. 7, no. 1, pp. 281–291, 2024, doi: 10.32493/jtsi.v7i1.38258.

[8] I. N. Mahmood and H. S. Abdullah, “Analyzing the behavior of different classification algorithms in diabetes prediction,” vol. 13, no. 1, pp. 201–206, 2024, doi: 10.11591/ijai.v13.i1.pp201-206.

[9] F. Refindha, A. Harianto, Z. Alawi, and I. Aristia, “PENGARUH KOMPOSISI SPLIT DATA PADA AKURASI KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN,” vol. 8, no. 1, pp. 36–44, 2025.

[10] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,” *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, 2022, doi: 10.1016/j.cmpb.2022.106773.

[11] B. P. Pamungkas, M. J. Vikri, and I. Aristia, “Application of SMOTE-ENN Method in Data Balancing for Classification of Diabetes Health Indicators with C4 . 5 Algorithm,” vol. 14, pp. 183–188, 2025.

[12] F. Mohsen, H. R. H. Al-absi, and N. El Hajj, “OPEN A scoping review of arti fi cial intelligence-based methods for diabetes risk prediction,” pp. 1–15, doi: 10.1038/s41746-023-00933-5.

[13] O. B. Ayoade and S. Shahrestani, “Machine Learning and Deep Learning Approaches for Predicting Diabetes Progression : A Comparative Analysis,” 2025.

[14] R. S. Chhillar, “Optimized stacking ensemble for early-stage diabetes mellitus prediction,” vol. 13, no. 6, pp. 7048–7055, 2023, doi: 10.11591/ijece.v13i6.pp7048-7055.

[15] R. Alkhanbouli, H. Matar Abdulla Almadhaani, F. Alhosani, and M. C. E. Simsekler, “The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions,” *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, 2025, doi: 10.1186/s12911-025-02944-6.

[16] Z. Ganji, F. Nikparast, N. Shoeibi, A. Shoeibi, and H. Zare, “Decoding Parkinson’s diagnosis: An OCT-based explainable AI with SHAP/LIME transparency from the Persian Cohort Study,” *Photodiagnosis Photodyn. Ther.*, vol. 54, no. May, 2025, doi: 10.1016/j.pdpdt.2025.104668.

[17] S. Ahmed, M. S. Kaiser, M. Shahadat Hossain, and K. Andersson, “A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions,” *IEEE Access*, vol.

13, no. July 2024, pp. 37370–37388, 2025, doi: 10.1109/ACCESS.2024.3422319.

[18] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, no. December 2024, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[19] M. Bhandarkar, V. S. Bendre, Y. V. Bellary, and A. K. Bhole, "Ensemble stacking classifier model for prediction of diabetes," vol. 13, no. 3, pp. 499–508, 2024, doi: 10.11591/ijict.v13i3.pp499-508.

[20] S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Heliyon Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, no. 2, p. e24536, 2024, doi: 10.1016/j.heliyon.2024.e24536.

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

[22] X. Li, "Design of a Multi-Model Diabetes Risk Prediction System for Clinical Application," *Appl. Comput. Eng.*, vol. 155, no. 1, pp. 125–136, 2025, doi: 10.54254/2755-2721/2025.gl23415.

[23] H. Sadr *et al.*, "Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches," *Eur. J. Med. Res.*, vol. 30, no. 1, 2025, doi: 10.1186/s40001-025-02680-7.

[24] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured datasets : a systematic review," *BMC Med. Res. Methodol.*, 2024, doi: 10.1186/s12874-024-02310-6.

[25] B. Toleva, I. Atanasov, and I. Ivanov, "An Effective Methodology for Diabetes Prediction in the Case of Class Imbalance," pp. 1–17, 2025.

[26] A. El, S. El, and H. M. El Bakry, "Pediatric diabetes prediction using deep learning," *Sci. Rep.*, no. 0123456789, pp. 1–20, 2024, doi: 10.1038/s41598-024-51438-4.

[27] R. K. Bhujade and S. Asthana, "An Extensive review of ReLu and Sigmoid Function in Multiple Hidden Layer Back Propagation Neural Network Model," *Int. J. Appl. Eng. Technol.*, vol. 5, no. 2, pp. 67–70, 2023.

[28] B. BAKIRARAR and A. H. ELHAN, "Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research," *Turkiye Klin. J. Biostat.*, vol. 15, no. 1, pp. 19–29, 2023, doi: 10.5336/biostatic.2022-93961.

[29] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, *Handling imbalanced medical datasets: review of a decade of research*, vol. 57, no. 10. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10884-2.

[30] H. Setiawan, A. Firnanda, and U. Khair, "Enhancing the Accuracy of Diabetes Prediction Using Feedforward Neural Networks : Strategies for Improved Recall and Generalization," vol. 4, no. 1, pp. 201–207, 2024.

[31] M. Saleh, A. L. Reshan, S. Amin, and M. A. L. I. Zeb, "An Innovative Ensemble Deep Learning Clinical Decision Support System for Diabetes Prediction," *IEEE Access*, vol. 12, no. May, pp. 106193–106210, 2024, doi: 10.1109/ACCESS.2024.3436641.

[32] J. J. Eertink, M. W. Heymans, G. J. C. Zwezerijnen, J. M. Zijlstra, H. C. W. De Vet, and R. Boellaard, "External validation : a simulation study to compare cross - validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients," pp. 4–11, 2022, doi: 10.1186/s13550-022-00931-w.

[33] A. Apicella, F. Isgrò, and R. Prevete, "machine learning and transfer learning," pp. 1–58, 2025.

[34] L. Sasse, J. Dukart, S. B. Eickhoff, M. Götz, S. Hamdan, and V. Komeyer, "Overview of leakage scenarios in supervised machine learning," 2025.

[35] S. Liu, Q. Tian, Y. Liu, and P. Li, "Joint Statistical Inference for the Area under the ROC Curve and Youden Index under a Density Ratio Model," pp. 1–21, 2024.