Buletin Sistem Informasi dan Teknologi Islam
Vol 7, No 2 (2026), pp 139-149
ISSN: 2721-0901
DOI: https://dx.doi.org/10.33096/busiti.v7i2.3413

Gradient Boosting Teroptimasi untuk Klasifikasi Diabetes dengan Analisis Eksplanabilitas Menggunakan SHAP dan LIME

Optimized Gradient Boosting for Diabetes Classification with Explainability Analysis Using SHAP and LIME

Angga Kurniawan[a1,*], Nisrina Hanifa Setiono[a,2], Afin Muhammad Nurtsani[b,3], Andi Hisyam Helmi Faalih Fakhruddin[a,4], Muhammad Naufal Farabbia[a,5], Nadia Syahda Fitriani[b,6]

[a] *Sains Data, Universitas Telkom Kampus Purwokerto, Banyumas, Indonesia*
[b] *Teknik Biomedis, Universitas Telkom Kampus Purwokerto, Banyumas, Indonesia*
[1]*anggakurniawan@telkomuniversity.ac.id;* [2]*nisrinahanifas@telkomuniversity.ac.id;*
[3]*afinmuhammadn@telkomuniversity.ac.id;* [4]*andihsym@student.telkomuniversity.ac.id;*
[5]*muhammadnaufalfarabb@student.telkomuniversity.ac.id;* [6]*nanadiaa@ student.telkomuniversity.ac.id*
**corresponding author*

Informasi Artikel

Diserahkan : 8 April 2026
Diterima : 16 Mei 2026
Direvisi : 28 Mei 2026
Diterbitkan : 29 Mei 2026

Kata Kunci:
Gradient Boosting
Optuna
SHAP
LIME
Prediksi Diabetes

ABSTRAK

Penelitian ini mengusulkan kerangka klasifikasi diabetes yang mengintegrasikan tiga algoritma gradient boosting, yaitu XGBoost, LightGBM, dan CatBoost, yang dioptimasi secara otomatis menggunakan Optuna berbasis optimasi Bayesian dengan 50 percobaan dan *cross validation 5-folds*. Dataset yang digunakan adalah Pima Indians Diabetes Dataset dengan 768 sampel dan 8 fitur klinis. Preprocessing dilakukan menggunakan beberapa metode meliputi penggantian nilai nol sebagai nilai hilang, imputasi median dari data pelatihan, serta penambahan fitur indikator *Insulin_missing*. Ketidakseimbangan kelas ditangani melalui parameter *scale_pos_weight* yang dioptimasi bersama hiperparameter model dalam satu ruang pencarian. Evaluasi model menggunakan metrik AUC-ROC, akurasi, F1-score, recall, dan presisi dengan ambang batas klasifikasi 0,6. Analisis eksplanabilitas dilakukan menggunakan SHAP pada level global dan lokal serta LIME pada level instans untuk meningkatkan transparansi model. Hasil optimasi menunjukkan CatBoost mencapai AUC-ROC *cross validation* tertinggi sebesar 0.8471, diikuti LightGBM sebesar 0.8409 dan XGBoost sebesar 0.8406. Pada data uji, XGBoost mencapai AUC-ROC 0,8246, akurasi 0,753, F1-Score 0,683, dan recall kelas diabetes 0,759, sedangkan CatBoost mencapai recall tertinggi 0,888 dengan F1-Score 0,690. Analisis SHAP secara konsisten mengidentifikasi Glucose, BMI, Age, dan DiabetesPedigreeFunction sebagai empat prediktor paling berpengaruh di ketiga model, selaras dengan pengetahuan klinis mengenai faktor risiko utama diabetes tipe 2.

***Keywords*:**
Gradient Boosting
Optuna
SHAP
LIME
Diabetes Prediction

ABSTRACT

This study proposes a diabetes classification framework integrating three gradient boosting algorithms, namely XGBoost, LightGBM, and CatBoost, automatically optimized using Optuna based on Bayesian optimization with 50 trials and 5-fold stratified cross-validation. The Pima Indians Diabetes Dataset consisting of 768 samples and 8 clinical features was used. Preprocessing employed a data leakage prevention approach including zero-value replacement as missing values, median imputation from training data only, and addition of an Insulin_missing indicator feature. Class imbalance was addressed through the scale_pos_weight parameter optimized jointly with model hyperparameters within a unified search space. Model evaluation used AUC-ROC, accuracy, F1-score, recall, and precision with a classification threshold of 0.6. Explainability analysis was conducted using SHAP at global and local levels and LIME at instance level to enhance model transparency. CatBoost achieved the highest cross-validation AUC-ROC of 0.8471, followed by LightGBM at 0.8409 and XGBoost at 0.8406. On the test data, XGBoost achieved an AUC-ROC of 0.8246, an accuracy of 0.753, an F1-score of 0.683, and a diabetes class recall of 0.759, while CatBoost achieved the highest recall of 0.888 with an F1-score of 0.690. SHAP analysis consistently identified Glucose, BMI, Age, and DiabetesPedigreeFunction as the four most influential predictors in all three models, consistent with clinical knowledge regarding the primary risk factors for type 2 diabetes.

I. Pendahuluan

Diabetes melitus merupakan penyakit metabolik kronis yang ditandai dengan kadar glukosa darah yang tinggi akibat gangguan produksi atau fungsi insulin. Berdasarkan laporan *International Diabetes Federation* (IDF), jumlah penderita diabetes di seluruh dunia pada tahun 2021 mencapai 537 juta jiwa dan diproyeksikan meningkat menjadi 783 juta jiwa pada tahun 2045 [1]. Indonesia sendiri menempati posisi kelima negara dengan jumlah penderita diabetes terbanyak di dunia, menjadikan deteksi dini dan penanganan yang tepat sebagai prioritas kesehatan nasional [1]. Komplikasi yang ditimbulkan oleh diabetes, seperti penyakit kardiovaskular, gagal ginjal, neuropati, dan kebutaan, tidak hanya menurunkan kualitas hidup penderita tetapi juga memberikan beban ekonomi yang signifikan bagi sistem layanan kesehatan [2].

Pendekatan *machine learning* telah banyak diterapkan untuk membantu diagnosis dan prediksi diabetes secara otomatis berdasarkan data klinis [3]. Di antara berbagai algoritma yang tersedia, metode *ensemble* berbasis *gradient boosting* seperti XGBoost, LightGBM, dan CatBoost telah menunjukkan kinerja yang unggul pada data tabular medis karena kemampuannya menangani fitur dengan distribusi tidak normal, nilai hilang, dan interaksi antar fitur yang kompleks [4], [5]. Meskipun demikian, sebagian besar penelitian terdahulu hanya berfokus pada peningkatan akurasi prediksi tanpa mempertimbangkan aspek interpretabilitas model, yang justru sangat krusial dalam konteks klinis agar keputusan prediksi dapat dipahami dan dipercaya oleh tenaga medis [6].

Salah satu tantangan utama dalam pemodelan prediksi diabetes adalah ketidakseimbangan kelas (*class imbalance*), di mana jumlah sampel non-diabetes secara signifikan lebih banyak dibanding sampel diabetes [7]. Kondisi ini menyebabkan model cenderung bias terhadap kelas mayoritas dan menghasilkan nilai *recall* yang rendah untuk kelas positif diabetes — padahal dalam konteks skrining medis, kemampuan mendeteksi penderita diabetes (*recall*) jauh lebih penting daripada sekadar akurasi keseluruhan. Berbagai pendekatan telah diusulkan untuk mengatasi masalah ini, di antaranya teknik *oversampling* seperti *Synthetic Minority Over-sampling Technique* (SMOTE) [7], penggunaan parameter pembobotan kelas, serta kombinasi keduanya. Namun demikian, penentuan nilai bobot yang optimal masih sering dilakukan secara manual atau menggunakan nilai *default*, tanpa proses optimasi yang sistematis.

Optimasi hiperparameter merupakan tahapan kritis dalam pengembangan model *machine learning* yang menentukan kinerja akhir model secara signifikan [8]. Pendekatan konvensional seperti *grid search* dan *random search* memiliki keterbatasan dalam hal efisiensi komputasi karena mengevaluasi ruang hiperparameter secara seragam tanpa memanfaatkan informasi dari percobaan sebelumnya [8]. Optuna, sebagai *framework* optimasi hiperparameter berbasis *Tree-structured Parzen Estimator* (TPE), menawarkan pendekatan optimasi Bayesian yang secara adaptif memfokuskan pencarian pada wilayah ruang hiperparameter yang menjanjikan, sehingga lebih efisien dan mampu menemukan konfigurasi optimal dengan jumlah percobaan yang lebih sedikit [9].

Di sisi lain, model *gradient boosting* yang telah dioptimasi pada umumnya bersifat *black-box*, sehingga sulit untuk menjelaskan mengapa suatu prediksi dihasilkan. Hal ini menjadi hambatan adopsi model *machine learning* di dunia medis, di mana klinisi membutuhkan justifikasi yang dapat dipahami untuk setiap keputusan prediksi [6]. Teknik *Explainable Artificial Intelligence* (XAI) seperti SHAP (*SHapley Additive exPlanations*) dan LIME (*Local Interpretable Model-Agnostic Explanations*) hadir sebagai solusi untuk membuka kotak hitam model tersebut [10], [11]. SHAP memberikan penjelasan berbasis teori permainan Shapley yang konsisten secara matematis dan dapat digunakan pada level global maupun lokal, sedangkan LIME memberikan aproksimasi lokal yang intuitif per instans prediksi [10], [11].

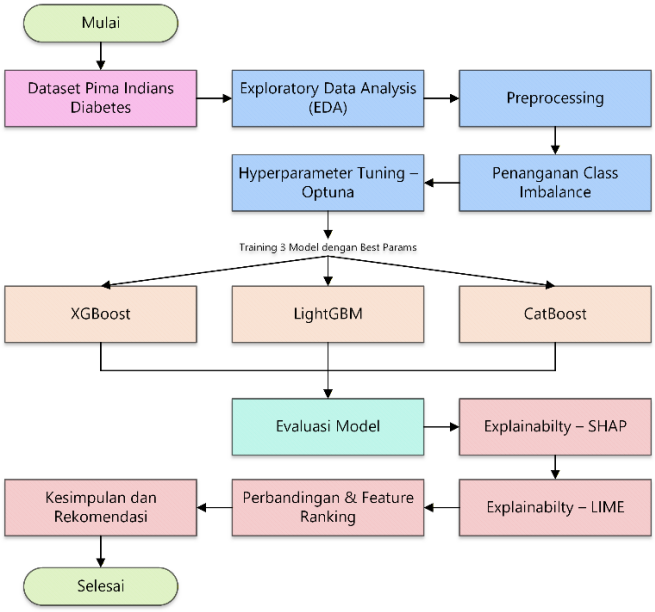
Beberapa penelitian terdahulu telah mengeksplorasi kombinasi sebagian dari elemen-elemen tersebut. Afsaneh et al. [3] melakukan tinjauan sistematis penggunaan *machine learning* untuk prediksi diabetes dan menemukan bahwa model *ensemble* secara konsisten mengungguli model tunggal. Osman et al. [12] membandingkan XGBoost dan LightGBM pada dataset diabetes dan melaporkan keunggulan LightGBM dalam hal kecepatan pelatihan, namun tidak menerapkan optimasi hiperparameter yang sistematis. Daram Sriharsha [13] menerapkan SHAP untuk menginterpretasikan model *random forest* pada prediksi diabetes, tetapi tidak mengintegrasikan analisis LIME sebagai metode komplementer. Loyola-González [14] mengkaji metode XAI pada data kesehatan secara umum, namun tidak secara khusus membahas kombinasi SHAP dan LIME pada model *gradient boosting* yang dioptimasi. Penggabungan Optuna dengan teknik eksplanabilitas ganda (SHAP dan LIME) dalam satu kerangka kerja memberikan kontribusi yang berbeda dari penelitian-penelitian sebelumnya karena tiga alasan. Pertama, Optuna memungkinkan optimasi bobot kelas *(weight_mult)* dilakukan bersama hiperparameter model dalam satu ruang pencarian, sehingga keseimbangan antara presisi dan recall ditemukan secara otomatis tanpa intervensi manual — sesuatu yang tidak dilakukan pada penelitian Osman et al.[12] maupun Daram Sriharsha[13]. Kedua, kombinasi SHAP (penjelasan global dan lokal berbasis teori permainan Shapley) dengan LIME (aproksimasi lokal per instans) menghasilkan transparansi model yang lebih komprehensif dibanding penggunaan salah satu metode secara terpisah: SHAP memberikan pandangan konsisten terhadap seluruh populasi dataset, sedangkan LIME memungkinkan verifikasi penjelasan pada level individu pasien yang sangat krusial bagi klinisi dalam pengambilan keputusan. Ketiga, sinkronisasi antara

temuan model AI dengan bukti klinis yang ada dapat divalidasi secara multidisiplin melalui perbandingan peringkat fitur SHAP dengan literatur medis yang telah mapan. Sementara itu, penelitian yang secara eksplisit mengintegrasikan optimasi hiperparameter Optuna, penanganan *class imbalance* melalui bobot yang turut dioptimasi, dan analisis eksplanabilitas ganda (SHAP + LIME) dalam satu kerangka yang kohesif masih sangat terbatas [9], [15].

Berdasarkan kesenjangan penelitian tersebut, penelitian ini mengusulkan kerangka yang mengintegrasikan empat kontribusi utama: (1) perbandingan tiga algoritma *gradient boosting* (XGBoost, LightGBM, CatBoost) pada dataset *Pima Indians Diabetes*; (2) optimasi hiperparameter otomatis menggunakan Optuna dengan *cross validation 5-folds*; (3) penanganan *class imbalance* melalui parameter *weight_mult* yang dioptimasi bersama hiperparameter model dalam satu ruang pencarian; dan (4) analisis eksplanabilitas dua tingkat menggunakan SHAP pada level global dan lokal serta LIME pada level instans. Penelitian ini bertujuan menghasilkan model prediksi diabetes yang tidak hanya akurat secara kuantitatif, tetapi juga dapat diinterpretasikan secara klinis untuk mendukung pengambilan keputusan medis yang transparan dan bertanggung jawab.

II. Metode

Penelitian ini merupakan penelitian eksperimen komputasional yang membandingkan kinerja tiga algoritma gradient boosting dalam klasifikasi diabetes sekaligus menerapkan teknik eksplanabilitas untuk menginterpretasikan hasil prediksi. Alur penelitian secara keseluruhan mengikuti tahapan yang ditunjukkan pada Gambar 1, meliputi pengumpulan dataset, eksplorasi data, preprocessing, penanganan class imbalance, pemodelan dengan optimasi hiperparameter Optuna, evaluasi model, dan analisis eksplanabilitas menggunakan SHapley Additive exPlanations (SHAP) dan Local Interpretable Model-Agnostic Explanations (LIME).



Gambar 1. Alur Penelitian

A. Dataset

Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang bersumber dari National Institute of Diabetes and Digestive and Kidney Diseases dan tersedia secara publik melalui repositori Kaggle. Dataset ini terdiri dari 768 sampel dengan 8 fitur klinis numerik dan 1 variabel target biner (Outcome), di mana nilai 1 menunjukkan positif diabetes dan nilai 0 menunjukkan negatif diabetes. Kedelapan fitur klinis tersebut adalah Pregnancies (jumlah kehamilan), Glucose (konsentrasi glukosa plasma 2 jam), BloodPressure (tekanan darah diastolik dalam mm Hg), SkinThickness (ketebalan lipatan kulit trisep dalam mm), Insulin (kadar insulin serum 2 jam dalam mu U/ml), BMI (indeks massa tubuh dalam kg/m²), DiabetesPedigreeFunction (fungsi silsilah diabetes), dan Age (usia dalam tahun). Statistik deskriptif dataset ditunjukkan pada Tabel 1.

Tabel 1. Statistik Deskriptif Dataset Pima Indians Diabetes

Fitur	n	Mean	Std	Min	Q1	Q3	Max
Pregnancies	768	3.86	3.37	0	1.00	6.00	17
Glucose	768	120.89	31.97	0	99.00	140.25	199
BloodPressure	768	69.11	19.36	0	62.00	80.00	122
SkinThickness	768	20.54	15.95	0	0.00	32.00	99
Insulin	768	79.80	115.24	0	0.00	127.25	846
BMI	768	31.99	7.88	0	27.30	36.60	67.1
DiabetesPedigreeFunction	768	0.47	0.33	0.33	0.24	0.63	2.42
Age	768	33.24	11.76	11.76	24.00	41.00	81

Distribusi kelas menunjukkan ketidakseimbangan dengan 500 sampel (65,10%) berkelas negatif dan 268 sampel (34,90%) berkelas positif diabetes. Rasio ketidakseimbangan ini adalah sekitar 1,87:1, yang tergolong moderat namun tetap berpotensi menyebabkan model bias terhadap kelas mayoritas dan menghasilkan recall rendah untuk kelas positif diabetes — yang justru paling kritis dalam konteks skrining klinis. Dalam penelitian ini, ketidakseimbangan kelas ditangani melalui pendekatan pembobotan kelas *(class weighting)* secara native pada tiap algoritma gradient boosting, bukan menggunakan teknik oversampling seperti SMOTE. Pendekatan ini dipilih karena pembobotan kelas tidak mengubah distribusi data asli sehingga menghindari risiko overfitting terhadap sampel sintetis yang umum menjadi kelemahan SMOTE pada dataset berukuran kecil-sedang. Lebih lanjut, nilai bobot *(weight_mult)* tidak ditetapkan secara manual melainkan dioptimasi secara otomatis oleh Optuna bersama hiperparameter model dalam satu ruang pencarian, memungkinkan model menemukan keseimbangan presisi-recall yang optimal secara data-driven.

B. Analisis Data Eksploratif

Analisis data eksploratif (Exploratory Data Analysis/EDA) dilakukan untuk memahami karakteristik dan pola distribusi data sebelum pemodelan. EDA mencakup analisis distribusi fitur menggunakan histogram dan kurva Kernel Density Estimation (KDE), deteksi outlier menggunakan metode Interquartile Range (IQR), identifikasi nilai nol sebagai missing value implisit pada fitur medis, serta analisis korelasi antar fitur menggunakan heatmap.

Hasil deteksi outlier menggunakan metode IQR dengan batas bawah Q1 − 1,5×IQR dan batas atas Q3 + 1,5×IQR ditunjukkan pada Tabel 2. Fitur BloodPressure memiliki jumlah outlier tertinggi sebanyak 45 sampel, diikuti Insulin (34 sampel) dan DiabetesPedigreeFunction (29 sampel). Outlier pada dataset ini tidak dihapus karena pada konteks medis nilai ekstrem seringkali merupakan informasi klinis yang valid dan signifikan.

Tabel 2. Statistik Deskriptif Dataset Pima Indians Diabetes

Fitur	Jumlah *Outlier*
Pregnancies	4
Glucose	5
BloodPressure	45
SkinThickness	1
Insulin	34
BMI	19
DiabetesPedigreeFunction	29
Age	9

Selain outlier, ditemukan nilai nol yang secara medis tidak mungkin pada beberapa fitur klinis. Nilai nol pada fitur Glucose, BloodPressure, SkinThickness, Insulin, dan BMI mengindikasikan data hilang (missing value implisit). Proporsi nilai nol per fitur ditunjukkan pada Tabel 3. Fitur Insulin memiliki proporsi nilai nol tertinggi yaitu 48,70%, yang berarti hampir separuh sampel tidak memiliki pengukuran insulin yang valid.

Tabel 3. Proporsi Nilai Nol (Missing Value Implisit) pada Fitur Medis

Fitur	Jumlah Nol	Proporsi (%)
Glucose	5	0.65
BloodPressure	35	4.56
SkinThickness	227	29.56
Insulin	374	48.70
BMI	11	1.43

Analisis rata-rata fitur berdasarkan kelas menunjukkan bahwa penderita diabetes (Outcome=1) secara konsisten memiliki nilai rata-rata lebih tinggi dibanding non-diabetes (Outcome=0), khususnya pada fitur Glucose (141,26 vs 109,98), BMI (35,14 vs 30,30), dan Age (37,07 vs 31,19), yang mengindikasikan ketiga fitur tersebut berpotensi menjadi prediktor kuat dalam klasifikasi.

C. Preprocessing Data

Tahap preprocessing dirancang mengikuti prinsip anti-kebocoran data (data leakage prevention) untuk memastikan estimasi kinerja model yang valid dan tidak bias. Seluruh operasi preprocessing diimplementasikan dalam kelas *MedicalPreprocessor* yang diterapkan secara terpisah di dalam setiap lipatan validasi silang, sehingga statistik imputasi hanya dihitung dari data pelatihan dan tidak terpapar ke data validasi maupun data uji.

Langkah-langkah preprocessing yang diterapkan adalah sebagai berikut. Pertama, data dipisah menggunakan *stratified train-test split* dengan rasio 80:20 dan random_state=42, sehingga data uji dikunci sejak awal dan sama sekali tidak terlibat dalam proses pelatihan maupun *tuning*. Kedua, ditambahkan fitur indikator biner *Insulin_missing* sebelum transformasi nilai nol dilakukan, di mana nilai 1 menandakan bahwa nilai Insulin asli adalah nol (data tidak tersedia). Langkah ini penting karena ketidaktersediaan insulin dapat memiliki makna klinis tersendiri. Ketiga, nilai nol pada fitur medis (Glucose, BloodPressure, SkinThickness,

Insulin, dan BMI) diubah menjadi NaN. Keempat, imputasi median dilakukan menggunakan nilai median yang dihitung dari data pelatihan pada setiap lipatan, kemudian diterapkan pada data validasi. Distribusi kelas setelah pemisahan data ditunjukkan pada Tabel 4. Stratified split berhasil mempertahankan proporsi kelas yang konsisten antara data pelatihan dan data uji.

Tabel 4. Distribusi Kelas pada Data Pelatihan dan Data Uji

Kelas	Jumlah Train	Jumlah Test	Proporsi Train (%)
0 (Tidak Diabetes)	400	100	65.15
1 (Diabetes)	214	54	34.85
Total	614	154	100.0

D. Penanganan Ketidakseimbangan Kelas

Ketidakseimbangan kelas ditangani menggunakan parameter pembobotan kelas yang tersedia secara native pada ketiga algoritma gradient boosting. Bobot dasar dihitung menggunakan rasio jumlah sampel negatif terhadap jumlah sampel positif dalam setiap lipatan pelatihan, sebagaimana ditunjukkan pada Persamaan (1).

$$scale_{pos_{weight}} = \frac{jumlah_{negatif}}{jumlah_{positif}} \; x \; weight_{mult} \quad (1)$$

Parameter *weight_mult* merupakan faktor pengali yang juga dioptimasi oleh Optuna dalam rentang [0,7; 2,5], sehingga model secara otomatis menemukan bobot terbaik yang menyeimbangkan antara presisi dan recall kelas positif. Untuk XGBoost dan LightGBM, bobot diterapkan melalui parameter *scale_pos_weight*, sedangkan untuk CatBoost diterapkan melalui *class_weights*=[1.0, spw]. Integrasi optimasi bobot ke dalam ruang pencarian hiperparameter menjadi salah satu kebaruan (novelty) pendekatan yang diusulkan dalam penelitian ini.

E. Model Gradient Boosting

Penelitian ini menggunakan tiga algoritma gradient boosting yang merepresentasikan varian utama dari keluarga tersebut, yaitu:

XGBoost *(Extreme Gradient Boosting)* [16] adalah algoritma boosting berbasis pohon keputusan dengan regularisasi L1 dan L2 eksplisit. XGBoost menggunakan pendekatan histogram-based *(tree_method="hist")* yang efisien secara komputasi. Hiperparameter yang dioptimasi meliputi *n_estimators* [300–2500], *max_depth* [2–10], *learning_rate* [0,001–0,3], *subsample* [0,6–1,0], *colsample_bytree* [0,6–1,0], *min_child_weight* [1–20], *gamma* [0–10], *reg_alpha* [0–10], dan *reg_lambda* [0–10].

LightGBM (*Light Gradient Boosting Machine*) [4], [5] menggunakan strategi pertumbuhan pohon berbasis *leaf-wise* dengan parameter *num_leaves* sebagai pengontrol kompleksitas, yang memungkinkan konvergensi lebih cepat dibanding pendekatan *level-wise* konvensional. Hiperparameter yang dioptimasi meliputi *n_estimators* [300–4000], *learning_rate* [0,001–0,2], *num_leaves* [8–256], *max_depth* [-1–12], *min_child_samples* [5–100], *subsample* [0,6–1,0], *colsample_bytree* [0,6–1,0], *reg_alpha* [0–10], dan *reg_lambda* [0–10].

CatBoost [9], [15] menggunakan teknik *ordered boosting* dan *oblivious trees* yang membuat model lebih *robust* terhadap *overfitting* khususnya pada dataset berukuran sedang. Hiperparameter yang dioptimasi meliputi *iterations* [400–4000], *depth* [3–10], *learning_rate* [0,001–0,2], *l2_leaf_reg* [0,001–30], *random_strength* [0–5], dan *bagging_temperature* [0–1].

F. Optimasi Hiperparameter dengan Optuna

Optuna adalah *framework* optimasi hiperparameter otomatis berbasis algoritma *Tree-structured Parzen Estimator* (TPE) yang merupakan implementasi dari optimasi Bayesian. Berbeda dengan *grid search* atau *random search* yang mengevaluasi ruang hiperparameter secara seragam, TPE secara adaptif memfokuskan pencarian pada wilayah yang diprediksi menghasilkan nilai objektif terbaik berdasarkan percobaan-percobaan sebelumnya.

Setiap model dioptimasi dengan 50 percobaan (*n_trials=50*) menggunakan fungsi yang memaksimalkan AUC-ROC rata-rata dari validasi silang stratifikasi 5-lipat (*5-fold Stratified Cross-Validation*). Desain validasi silang di dalam fungsi Optuna memastikan bahwa setiap percobaan hiperparameter dievaluasi secara robust dan tidak *overfit* terhadap satu partisi data tertentu. Alur optimasi dapat dinyatakan pada Persamaan (2).

$$\theta^* = \arg\max_{\theta} \frac{1}{k} \sum_{k=1}^{K} AUC - ROC \; (f_\theta, D_{val}^{(k)}) \quad (2)$$

di mana θ adalah set hiperparameter, $k = 5$ adalah jumlah lipatan, f_θ adalah model dengan hiperparameter θ, dan $D_{val}^{(k)}$ adalah data validasi pada lipatan ke- k.

G. Evaluasi Model

Evaluasi kinerja model dilakukan pada data uji yang telah dikunci sejak awal menggunakan ambang batas klasifikasi (*threshold*) sebesar 0,6. Penggunaan *threshold* 0,6 (lebih tinggi dari nilai *default* 0,5) bertujuan untuk meningkatkan presisi prediksi positif diabetes, mengingat konsekuensi klinis dari *false positive* dan *false negative* yang perlu dipertimbangkan. Metrik evaluasi yang digunakan meliputi: AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*), Akurasi, *F1-Score*, *Recall* (Sensitivitas), dan Presisi, sebagaimana didefinisikan pada Persamaan (3) hingga Persamaan (6).

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \tag{5}$$

$$AUC - ROC = \int_0^1 TPRd(FPR) \tag{6}$$

H. Analisis Eksplanabilitas

Analisis eksplanabilitas dilakukan menggunakan dua metode untuk memberikan interpretasi pada dua level yang berbeda. SHAP (*SHapley Additive exPlanations*) diimplementasikan menggunakan TreeExplainer dengan *model_output="probability"* sehingga nilai SHAP berada dalam skala probabilitas yang lebih mudah diinterpretasikan secara klinis. SHAP digunakan pada dua level: (1) interpretasi global menggunakan *summary plot* untuk menunjukkan kontribusi rata-rata setiap fitur terhadap prediksi model secara keseluruhan, dan (2) interpretasi lokal menggunakan *waterfall plot* untuk menganalisis kasus spesifik *True Positive* (TP) dan *False Negative* (FN). Untuk perbandingan lintas model, dihitung nilai mean |SHAP| per fitur menggunakan Persamaan (7).

$$\overline{|\phi_j|} = \frac{1}{n}\sum_{i=1}^{n}|\phi_j^{(i)}| \tag{7}$$

di mana $\phi_j^{(i)}$ adalah nilai SHAP fitur ke- j pada sampel ke- i, dan n adalah jumlah sampel uji.

LIME (*Local Interpretable Model-Agnostic Explanations*) [7] diimplementasikan menggunakan *LimeTabularExplainer* dengan mode klasifikasi dan fitur kontinu yang didiskretisasi *(discretize_continuous=True)*. LIME menghasilkan penjelasan lokal per instans dengan menampilkan 8 fitur paling berpengaruh beserta arah dan besaran kontribusinya terhadap prediksi model pada sampel tertentu.

III. Hasil dan Pembahasan

A. Hasil Optimasi Hiperparameter

Proses optimasi Optuna berhasil menemukan kombinasi hiperparameter terbaik untuk ketiga model setelah 50 percobaan masing-masing. Nilai AUC-ROC terbaik yang diperoleh selama proses validasi silang menunjukkan bahwa ketiga model mencapai kinerja yang kompetitif. Tabel 5 merangkum hiperparameter terpilih beserta nilai AUC-ROC validasi silang terbaik yang dicapai.

Tabel 5. Hiperparameter Terbaik Hasil Optimasi Optuna

Hiperparameter	XGBoost	LightGBM	CatBoost
n_estimators	1308	352	971
learning_rate	0.0017	0.0226	0.0013
max_depth	7	3	3
weight_mult	1.3468	1.9183	2.3383
AUC-ROC CV	0.8405	0.8409	0.8471

B. Kinerja Model pada Data Uji

Setelah proses *tuning*, model dilatih ulang menggunakan seluruh data pelatihan (*full training*) dengan hiperparameter terbaik, kemudian dievaluasi pada data uji. Hasil evaluasi lengkap menggunakan *threshold* = 0,6 ditunjukkan pada Tabel 6.

Tabel 6. Perbandingan Kinerja Model pada Data Uji (*Threshold* = 0,6)

Metrik	XGBoost	LightGBM	CatBoost
AUC-ROC	0.824	0.824	0.819
Akurasi	0.753	0.740	0.720
F1-Score	0.683	0.687	0.690
Recall	0.759	0.814	0.888
Presisi	0.621	0.594	0.564

Berdasarkan Tabel 6, ketiga algoritma menunjukkan kemampuan diskriminatif yang relatif sebanding dengan nilai AUC-ROC yang hampir identik, yaitu 0,824 untuk XGBoost dan LightGBM serta 0,819 untuk

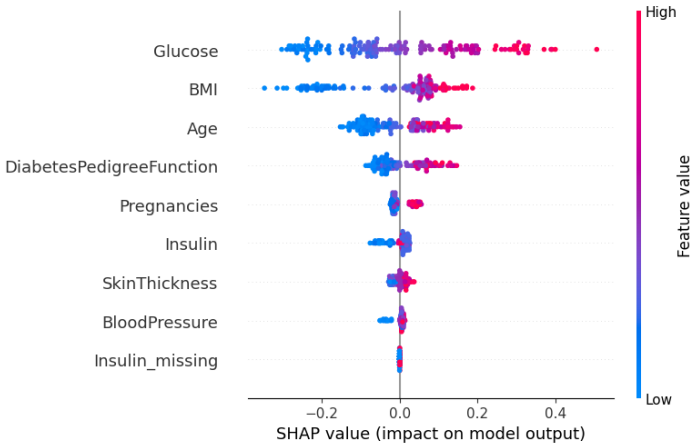
CatBoost. Akurasi tertinggi dicapai oleh XGBoost (0,753), diikuti LightGBM (0,740) dan CatBoost (0,720). Sementara itu, F1-Score yang merupakan harmonisasi presisi dan recall menunjukkan nilai yang berdekatan, dengan CatBoost sedikit unggul (0,690) dibanding LightGBM (0,687) dan XGBoost (0,683). Secara umum, ketiga model memiliki kapasitas prediktif yang setara dari perspektif klasifikasi biner.

Perbedaan utama terlihat pada trade-off antara recall dan presisi. CatBoost mencapai recall tertinggi (0,888) namun presisi terendah (0,564), mengindikasikan bahwa model ini sangat sensitif dalam mengidentifikasi penderita diabetes (meminimalkan false negative) dengan konsekuensi menghasilkan lebih banyak false positive. Sebaliknya, XGBoost memiliki presisi tertinggi (0,621) dan recall terendah (0,759), mencerminkan pendekatan yang lebih konservatif sehingga mengurangi false positive tetapi berisiko melewatkan lebih banyak kasus positif. LightGBM berada di posisi tengah dengan recall 0,814 dan presisi 0,594. Dalam konteks skrining diabetes di mana false negative memiliki konsekuensi klinis yang lebih serius, CatBoost lebih diutamakan untuk memaksimalkan deteksi, sementara XGBoost lebih cocok jika prioritas diberikan pada pengurangan false positive.

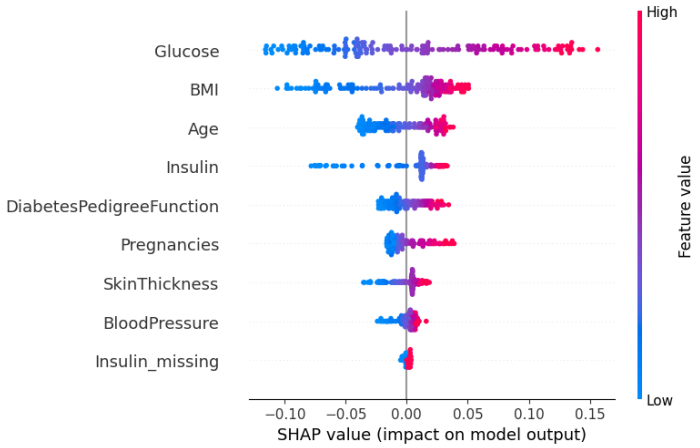
C. Analisis SHAP Global

Analisis SHAP global dilakukan menggunakan *summary plot* yang menampilkan distribusi nilai SHAP seluruh sampel uji untuk setiap fitur. Gambar 2, Gambar 3, dan Gambar 4 menunjukkan *summary plot* SHAP untuk XGBoost, LightGBM, dan CatBoost secara berurutan.

Gambar 2. SHAP Summary Plot XGBoost



Gambar 3. SHAP Summary Plot LightGBM



Gambar 4. SHAP Summary Plot CatBoost

Untuk membandingkan kepentingan fitur secara kuantitatif lintas model, dihitung nilai mean |SHAP| per fitur dan dirangkum pada Tabel 7.

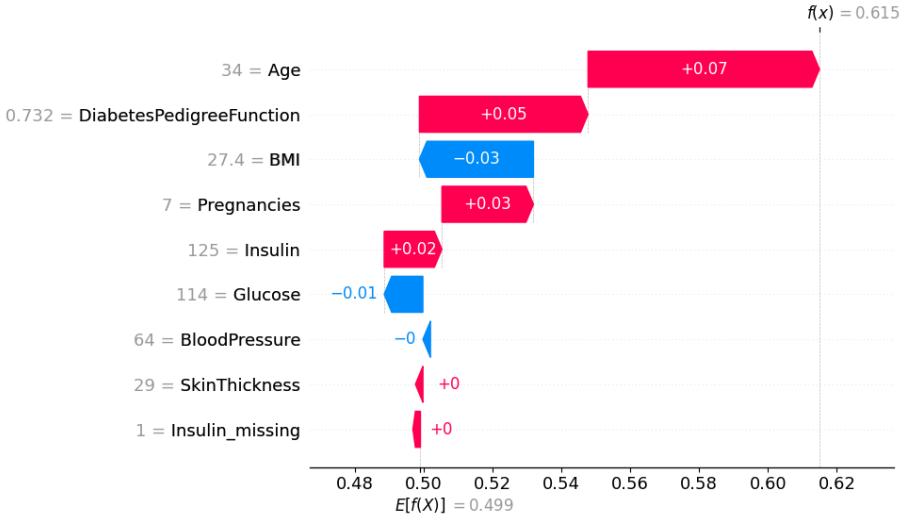
Tabel 7. Perbandingan Mean |SHAP| Lintas Model

Fitur	XGBoost	LightGBM	CatBoost	Rata-rata
Glucose	0,1509	0,1480	0,1384	0,1458
BMI	0,1131	0,0914	0,0824	0,0956
Age	0,0643	0,0700	0,0464	0,0602
DiabetesPedigreeFunction	0,0382	0,0427	0,0446	0,0418
Insulin	0,0316	0,0231	0,0300	0,0282
Pregnancies	0,0173	0,0209	0,0311	0,0231
SkinThickness	0.0087	0.0138	0.0173	0.0132
BloodPressure	0,0088	0.0069	0.0129	0.0095
Insulin_missing	0,0000	0.0000	0.0125	0.0041

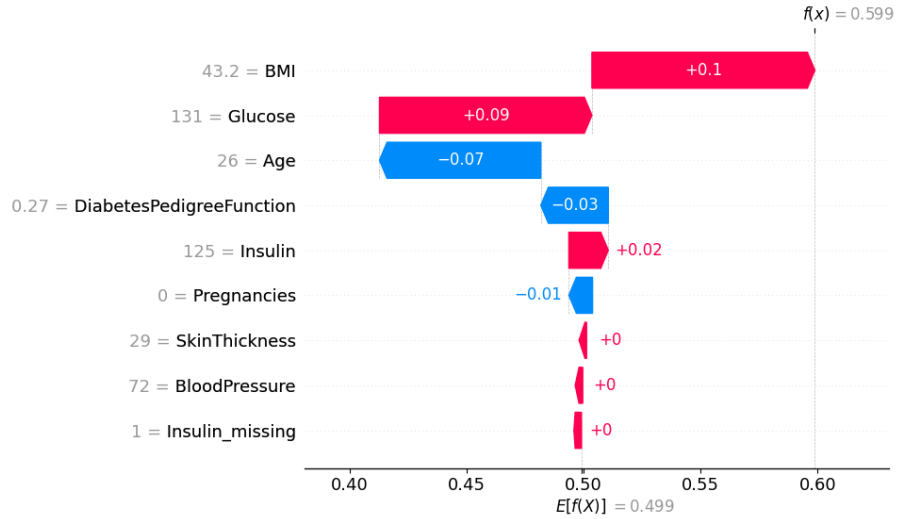
Hasil pada Tabel 7 menunjukkan bahwa *Glucose* secara konsisten menjadi fitur paling berpengaruh pada ketiga model dengan nilai mean |SHAP| rata-rata lintas model sebesar 0,1458. Hal ini selaras dengan pengetahuan klinis bahwa kadar glukosa plasma merupakan indikator utama dalam diagnosis diabetes. *BMI* berada di posisi kedua dengan nilai rata-rata 0,0956, diikuti *Age* (0,0602) dan *DiabetesPedigreeFunction* (0,0418). Konsistensi peringkat fitur di antara ketiga model mengindikasikan bahwa pola yang dipelajari bersifat stabil dan bukan artefak dari algoritma tertentu.

D. Analisis SHAP Lokal

Analisis SHAP lokal dilakukan pada algoritma XGBoost menggunakan *waterfall plot* untuk dua jenis kasus prediksi: *True Positive* (TP) dan *False Negative* (FN). Analisis ini memberikan wawasan tentang mengapa model membuat keputusan tertentu pada level individu pasien. Pada kasus True Positive (Pasien ke-3, XGBoost, XGBoost berhasil mengidentifikasi pasien ke-3 sebagai penderita diabetes. *Waterfall plot* menunjukkan fitur-fitur yang mendorong prediksi ke arah positif, dengan *Glucose* dan *BMI* sebagai kontributor utama yang meningkatkan probabilitas prediksi di atas nilai ekspektasi dasar (*base value*) model. Sedangkan Kasus False Negative (Pasien ke-14, XGBoost, XGBoost gagal mendeteksi pasien ke-14 yang sebenarnya adalah penderita diabetes. Analisis *waterfall plot* mengungkapkan bahwa kombinasi nilai *DiabetesPedigreeFunction* yang rendah, usia muda (24–29 tahun), dan nilai *BloodPressure* rendah secara kolektif menurunkan probabilitas prediksi di bawah *threshold* 0,6, meskipun nilai *BMI* dan *Insulin* pasien tersebut berada di rentang yang mengindikasikan risiko diabetes. Gambar 5 dan Gambar 6 menunjukkan *waterfall plot* SHAP *True Positive* dan *False Negative* XGBoost secara berurutan.



Gambar 5. Waterfall Plot SHAP — *True Positive* XGBoost



Gambar 6. Waterfall Plot SHAP — *False Negative* XGBoost

E. Analisis LIME

Analisis LIME dilakukan pada sampel ke-14 (instans *False Negative* pada XGBoost) untuk ketiga model, dengan menampilkan 8 fitur paling berpengaruh. Hasil LIME untuk ketiga model ditunjukkan pada Tabel 8.

Tabel 8. Hasil Analisis LIME pada Sampel ke-14 (Instans *False Negative*)

Fitur/Kondisi	XGBoost	LightGBM	CatBoost
32.40 < BMI ≤ 36.50	+0.0802	+0.0762	+0.0697
117.00 < Glucose ≤ 140.00	+0.0469	+0.0305	+0.0529
120.00 < Insulin ≤ 125.00	+0.0370	+0.0495	+0.0375
Insulin missing (0–1)	-	-	+0.0206
DiabetesPedigreeFunction ≤ 0.24	-0.0532	-0.0608	-0.0866
24.00 < Age ≤ 29.00	-0.0601	-0.0658	-0.0368
BloodPressure ≤ 64.00	-0.0339	-0.0188	-0.0450
1.00 < Pregnancies ≤ 3.00	-0.0207	-0.0200	-0.0362
25.00 < SkinThickness ≤ 29.00	-0.0178	-0.0172	-

Tanda positif (+) menunjukkan fitur yang mendorong prediksi ke arah kelas diabetes, sedangkan tanda negatif (−) menunjukkan fitur yang mendorong ke arah non-diabetes. Hasil analisis LIME secara konsisten menunjukkan bahwa pada sampel ke-14, BMI dalam rentang 32.40–36.50 menjadi faktor pendorong terkuat ke arah diabetes di ketiga model. Sebaliknya, nilai *DiabetesPedigreeFunction* yang rendah (≤ 0,24) dan usia muda (24–29 tahun) menjadi faktor penghambat utama yang menyebabkan ketiga model gagal mengklasifikasikan sampel ini sebagai positif diabetes.

Menariknya, pada model CatBoost, fitur *Insulin_missing* muncul sebagai salah satu faktor pendorong positif (+0,0206), menunjukkan bahwa ketidaktersediaan data insulin itu sendiri merupakan sinyal yang informatif untuk prediksi diabetes — sebuah temuan yang mendukung keputusan desain untuk menambahkan fitur indikator *missing* dalam tahap *preprocessing*.

F. Peringkat Fitur Lintas Model

Untuk mendapatkan gambaran komprehensif tentang kepentingan fitur yang tidak bergantung pada satu model tertentu, dilakukan agregasi peringkat fitur berdasarkan nilai mean |SHAP| dari ketiga model. Peringkat rata-rata (*mean rank*) dihitung dengan merata-ratakan peringkat masing-masing fitur di tiap model. Hasil peringkat akhir ditunjukkan pada Tabel 9.

Tabel 9. Peringkat Fitur Berdasarkan Mean |SHAP| Lintas Model

Peringkat	Fitur	XGBoost	LightGBM	CatBoost	Rata-rata	Mean Rank
1	Glucose	0,1509	0,1480	0,1384	0,1458	1.0000
2	BMI	0,1131	0,0914	0,0824	0,0956	2.0000
3	Age	0,0643	0,0700	0,0464	0,0602	3.0000
4	DiabetesPedigreeFunction	0,0382	0,0427	0,0446	0,0418	4.0000
5	Insulin	0,0316	0,0231	0,0300	0,0282	5.3333
6	Pregnancies	0,0173	0,0209	0,0311	0,0231	5.6666
7	SkinThickness	0.0087	0.0138	0.0173	0.0132	7.3333
8	BloodPressure	0,0088	0.0069	0.0129	0.0095	7.6666
9	Insulin_missing	0,0000	0.0000	0.0125	0.0041	9.0000

Secara klinis, dominasi Glucose sebagai prediktor teratas sangat koheren dengan kriteria diagnostik diabetes dari *American Diabetes Association (ADA)*, yang menggunakan kadar glukosa plasma puasa ≥126 mg/dL atau kadar glukosa 2 jam ≥200 mg/dL sebagai ambang batas diagnosis. BMI pada posisi kedua konsisten dengan bukti epidemiologis yang menunjukkan bahwa obesitas (BMI ≥30 kg/m²) meningkatkan risiko diabetes tipe 2 hingga beberapa kali lipat melalui mekanisme resistensi insulin. Usia (Age) di posisi ketiga mencerminkan penurunan sekresi insulin dan peningkatan resistensi insulin yang terkait proses penuaan, sementara DiabetesPedigreeFunction mengkuantifikasi kontribusi faktor herediter yang telah lama diakui dalam literatur genetika diabetes. Perlu dicatat bahwa Insulin, meskipun secara fisiologis merupakan biomarker langsung diabetes, hanya menempati posisi kelima. Hal ini kemungkinan disebabkan oleh tingginya proporsi missing value (48,70%) pada fitur ini yang membatasi kontribusi informatifnya — sebuah limitasi yang penting untuk diakui dalam interpretasi klinis hasil model.

Perbandingan antara SHAP (level global) dan LIME (level lokal) juga mengungkap konsistensi sekaligus perbedaan penting antara kedua metode. Secara global, SHAP menunjukkan bahwa Glucose dan BMI adalah prediktor dominan yang berlaku untuk populasi secara keseluruhan. Namun pada level instans — khususnya pada sampel ke-14 yang merupakan kasus False Negative — analisis LIME mengungkap bahwa DiabetesPedigreeFunction yang rendah dan usia muda (24–29 tahun) justru menjadi faktor penghambat yang lebih dominan, meskipun nilai Glucose pasien tersebut berada dalam rentang moderat (117–140 mg/dL). Ketidakselarasan ini bukan merupakan kontradiksi, melainkan mencerminkan sifat nonlinier model gradient boosting: pengaruh suatu fitur bergantung pada kombinasi nilai fitur lainnya pada individu tertentu. Temuan ini memiliki implikasi klinis penting — profil risiko individual selalu harus dipertimbangkan secara holistik, dan analisis LIME per instans memungkinkan pendekatan tersebut secara praktis.

IV. Kesimpulan dan saran

Penelitian ini berhasil mengembangkan dan membandingkan tiga model klasifikasi diabetes berbasis *gradient boosting* (XGBoost, LightGBM, dan CatBoost) yang dioptimasi secara otomatis menggunakan Optuna dengan pendekatan optimasi Bayesian. Integrasi optimasi parameter pembobotan kelas *(weight_mult)* ke dalam ruang pencarian hiperparameter terbukti efektif dalam menangani ketidakseimbangan kelas tanpa mengubah distribusi data asli. XGBoost mencapai AUC-ROC sebesar 0.824 dengan *recall* kelas diabetes sebesar 0.759 pada *threshold* 0,6, menunjukkan kemampuan deteksi diabetes yang baik. Analisis eksplanabilitas menggunakan SHAP dan LIME secara konsisten mengidentifikasi *Glucose*, *BMI*, *Age*, dan *DiabetesPedigreeFunction* sebagai empat prediktor paling berpengaruh di ketiga model, yang selaras dengan pengetahuan klinis yang ada. Temuan menarik dari analisis LIME menunjukkan bahwa fitur *Insulin_missing* memberikan kontribusi positif pada prediksi CatBoost, mengkonfirmasi bahwa ketidaktersediaan data insulin itu sendiri merupakan sinyal klinis yang informatif. Penelitian selanjutnya disarankan untuk mengeksplorasi pendekatan *ensemble* dari ketiga model melalui teknik *stacking* atau *blending* untuk lebih meningkatkan kinerja prediksi. Selain itu, pengujian pada dataset diabetes lain yang lebih beragam secara demografis perlu dilakukan untuk memvalidasi generalisasi model. Penerapan teknik penanganan *imbalance* berbasis *sampling* seperti *SMOTE* sebagai perbandingan terhadap pendekatan pembobotan juga menjadi arah penelitian yang menarik untuk dieksplorasi. Lebih lanjut, validasi model pada dataset klinis yang lebih heterogen secara demografis misalnya data dari populasi Asia Tenggara, Afrika, atau Timur Tengah yang memiliki profil risiko diabetes berbeda dari populasi Pima Indians sangat diperlukan sebelum model dapat dipertimbangkan untuk implementasi klinis yang lebih luas. Dari perspektif translasi ke praktik medis, pengembangan antarmuka

berbasis web atau aplikasi mobile yang menyajikan hasil eksplanabilitas SHAP dan LIME dalam format yang dapat dipahami oleh tenaga medis non-teknis seperti visualisasi risiko individual dan penjelasan dalam bahasa klinis merupakan langkah kritis berikutnya untuk menjembatani kesenjangan antara kemampuan teknis model machine learning dan kebutuhan praktis di lingkungan layanan kesehatan.

Daftar Pustaka

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021. [Online]. Available: https://www.diabetesatlas.org

[2] N. H. Cho *et al.*, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, Apr. 2018, doi: 10.1016/j.diabres.2018.02.023.

[3] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," *Diabetol. Metab. Syndr.*, vol. 14, no. 1, Dec. 2022, doi: 10.1186/s13098-022-00969-9.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[5] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Mar. 27, 2026. [Online]. Available: https://github.com/Microsoft/LightGBM.

[6] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, p. e1312, Jul. 2019, doi: 10.1002/widm.1312.

[7] 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, Jun. 2002, doi: 10.1613/jair.953.

[8] BergstraJames and BengioYoshua, "Random search for hyper-parameter optimization," *The Journal of Machine Learning Research*, Feb. 2012, doi: 10.5555/2188385.2188395.

[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, Jul. 2019, Accessed: Mar. 27, 2026. [Online]. Available: http://arxiv.org/abs/1907.10902

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Neural Information Processing Systems*, 2017.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pp. 97–101, Feb. 2016, doi: 10.18653/v1/n16-3020.

[12] A. H. Osman and H. M. Aljahdali, "Diabetes Disease Diagnosis Method based on Feature Extraction using K-SVM," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 1, Jan. 2017, doi: 10.14569/ijacsa.2017.080130.

[13] S. Daram, "Explainable AI in Healthcare: Enhancing Trust, Transparency, and Ethical Compliance in Medical AI Systems," *International Journal of AI, BigData, Computational and Management Studies*, vol. 6, no. 2, pp. 11–20, Apr. 2025, doi: 10.63282/3050-9416.ijaibdcms-v6i2p102.

[14] O. Loyola-Gonzalez, "Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view," 2019, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2019.2949286.

[15] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, Accessed: Mar. 27, 2026. [Online]. Available: https://github.com/catboost/catboost

[16] A. KURNIAWAN, M. KUDIN, and A. S. AT-TAQWA, "Evaluation of Machine Learning and Deep Learning Algorithms with Feature Scaling and K-Fold Cross-Validation for Diabetes Classification," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 11, no. 1, pp. 494–512, Jan. 2026, doi: 10.36341/rabit.v11i1.7005.