Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 1058~1065
e-ISSN: 2548-6861

Comparison of LightGBM and CatBoost Algorithms for Diabetes Prediction Based on Clinical Data

Muhammad Sidik Latuconsina [1]*, Majid Rahardi [2]*
* Informatika, Universitas Amikom Yogyakarta
latuconsinamsidik@students.amikom.ac.id [1], majid@amikom.ac.id [2]

Article Info

Article history:

Received 2026-01-05
Revised 2026-01-26
Accepted 2026-01-30

Keyword:

Diabetes Mellitus,
SHAP,
SMOTE,
LightGBM,
CatBoost.

ABSTRACT

Diabetes Mellitus presents a global health challenge necessitating accurate early detection to prevent fatal complications. However, clinical data often exhibit imbalanced class distributions, hindering standard prediction models from effectively detecting positive patients. This study aims to compare the performance of two modern Gradient Boosting algorithms, LightGBM and CatBoost, in predicting diabetes risk. Random Forest and Logistic Regression algorithms were included as baseline models to benchmark effectiveness. To address data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied during the training data preprocessing stage. The dataset was sourced from the Kaggle public repository (Diabetes Prediction Dataset), comprising 100,000 patient medical records with clinical attributes such as age, body mass index (BMI), and HbA1c levels. Performance evaluation utilized Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC) metrics. Experimental results demonstrated a tight competition, where LightGBM achieved the highest Accuracy of 97.16%. However, CatBoost demonstrated superior sensitivity (Recall) of 69.71% and the highest F1-Score of 80.48%. This makes CatBoost the most reliable model in minimizing False Negatives compared to LightGBM and Random Forest, whereas Logistic Regression showed the lowest performance. Furthermore, interpretability analysis using SHAP (SHapley Additive exPlanations) revealed that HbA1c and blood glucose levels were the most dominant features in detection, validating the model's alignment with clinical diagnosis. This study concludes that the CatBoost algorithm combined with SMOTE offers a more sensitive, transparent, and efficient diabetes prediction for medical screening.

I. PENDAHULUAN

Diabetes Melitus (DM) merupakan penyakit metabolik kronis yang menjadi tantangan kesehatan global utama dengan prevalensi yang terus meningkat signifikan. Berdasarkan data dari *IDF Diabetes Atlas* dan studi global oleh Ong et al. (2023) [1], beban penyakit DM diproyeksikan akan terus meningkat signifikan hingga tahun 2050, menjadikannya penyebab utama morbiditas dan mortalitas. Di Indonesia, tren peningkatan kasus DM juga sangat mengkhawatirkan, diperparah dengan risiko komplikasi fatal. Studi terbaru menunjukkan bahwa kadar gula darah yang tidak terkontrol pada DM Tipe 2 memiliki hubungan erat dengan kejadian hipertensi [2], yang meningkatkan risiko gagal ginjal terminal (*End-Stage Renal Disease*) [3]. Mengingat fatalnya risiko tersebut, pengembangan sistem deteksi dini yang cepat dan akurat sangatlah mendesak.

Dalam konteks diagnosis, teknologi *Machine Learning* (ML) telah banyak diterapkan untuk memprediksi risiko DM berdasarkan data klinis [4]. Meskipun algoritma konvensional seperti *Random Forest* (RF) [5] sering digunakan sebagai *baseline* [6], metode ini kerap menghadapi kendala skalabilitas dan efisiensi ketika berhadapan dengan data rekam medis berskala besar [7]. Selain Random Forest (RF), metode statistik klasik seperti Logistic Regression (LR) juga tetap relevan dijadikan standar dasar (*baseline*) karena

kesederhanaan dan interpretabilitasnya yang tinggi dalam analisis risiko medis [8] [9].

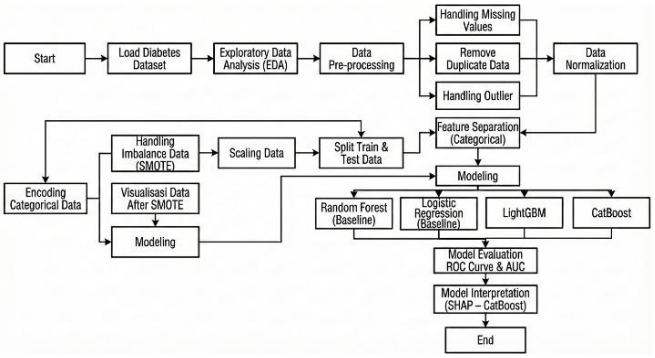
Tantangan kritis lainnya adalah masalah ketidakseimbangan kelas (*class imbalance*), di mana jumlah data pasien sehat jauh lebih banyak daripada pasien diabetes. Kondisi ini menyebabkan model klasifikasi, termasuk RF, cenderung bias ke kelas mayoritas dan memiliki sensitivitas (*Recall*) yang rendah dalam mendeteksi kasus positif yang sebenarnya [10][11].

Untuk mengatasi bias ini, teknik *resampling* seperti *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan guna menyeimbangkan distribusi data latih dan telah terbukti signifikan meningkatkan kinerja model pada metrik *Recall* dan *F1-Score* [12], [13], serta efektif meningkatkan sensitivitas deteksi pada dataset diabetes yang tidak seimbang [14], [15].

Untuk menjawab tantangan akurasi dan efisiensi pada data tidak seimbang, penelitian ini mengusulkan penggunaan algoritma *Gradient Boosting* modern, yaitu LightGBM (*Light Gradient Boosting Machine*) [16] dan CatBoost (*Categorical Boosting*) [17]. Kedua algoritma ini menawarkan efisiensi komputasi yang lebih tinggi dan performa superior dibandingkan model tradisional [18]. Penelitian ini juga menerapkan strategi validasi silang *K-Fold Cross-Validation* untuk memastikan evaluasi performa model yang lebih objektif dan *robust* terhadap variasi data [19], [20]. Dengan mengkomparasi performa LightGBM-SMOTE dan CatBoost-SMOTE terhadap Logistic Regression dan Random Forest (*baseline*) yang juga diolah dengan SMOTE, penelitian ini bertujuan mengidentifikasi model prediksi terbaik yang paling akurat dan robust, khususnya pada metrik *Recall*, untuk direkomendasikan sebagai sistem pendukung keputusan klinis.

II. METODE

Penelitian ini dilakukan secara sistematis untuk membandingkan performa model klasifikasi dalam memprediksi risiko diabetes. Tahapan penelitian ini meliputi akuisisi data, prapemrosesan, penanganan ketidakseimbangan data, penerapan algoritma, hingga evaluasi akhir. Sistematika pelaksanaan penelitian ini dapat dilihat pada diagram alir yang disajikan dalam Gambar 1.



Gambar 1. Alur penelitian.

A. Data Acquisition

Penelitian ini memanfaatkan Diabetes Prediction Dataset yang diperoleh dari repositori publik Kaggle (tersedia di: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data). Dataset ini mencakup 100.000 rekam medis pasien yang terdiri dari delapan fitur independen (demografis dan klinis) serta satu variabel target biner. Rincian karakteristik setiap atribut beserta tipe datanya disajikan secara ringkas pada Tabel I. Berdasarkan analisis distribusi data, ditemukan tantangan berupa ketidakseimbangan kelas (class imbalance) yang ekstrem, di mana prevalensi kasus positif diabetes hanya berkisar 8,5% dari total populasi. Kondisi ini dikhawatirkan dapat memicu bias prediksi pada kelas mayoritas jika tidak ditangani dengan strategi resampling yang tepat pada tahap pra-pemrosesan [5], [6].

TABEL I
KARAKTERISTIK DATASET

Atribut	Keterangan
Sumber Data	Kaggle – Diabetes Prediction Dataset
Jumlah Data	100,000.
Variabel Independen	8
Variabel Target	Diabetes (Biner)
Distribusi Kelas	8,5% Diabetes; 91,5% Non-Diabetes
Fitur Utama	HbA1c, Kadar Glukosa Darah, BMI

B. Data Preprocessing & Experimental Setup

Tahap pra-pemrosesan diawali dengan pembersihan data (data cleaning) untuk menjamin integritas statistik input model. Berdasarkan pemeriksaan kualitas data, tidak ditemukan missing values, namun prosedur eliminasi duplikat diterapkan untuk menghapus baris identik guna mencegah redundansi informasi. Analisis pencilan (outlier) pada fitur vital seperti BMI dan Blood Glucose tetap dipertahankan karena merepresentasikan kondisi fisiologis riil pasien diabetes (kasus ekstrem).

Setelah data dipastikan bersih, fitur kategorikal (Gender, Smoking History) dikonversi menggunakan One-Hot Encoding dengan parameter drop_first=True untuk menghindari multikolinearitas.

Sebelum dilakukan penskalaan fitur, dataset terlebih dahulu dibagi menjadi dua partisi independen: 80% sebagai Data Latih (Training Set) dan 20% sebagai Data Uji (Testing Set) menggunakan teknik Stratified Sampling. Langkah ini krusial untuk memastikan rasio kelas tetap konsisten dan mencegah kebocoran data (data leakage) dari data uji ke dalam proses pelatihan.

Selanjutnya, normalisasi fitur numerik menggunakan Min-Max Scaler dilakukan. Penting untuk dicatat bahwa scaler hanya dilatih (fit) pada Data Latih, lalu parameternya digunakan untuk mentransformasi Data Uji. Persamaan normalisasi adalah:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

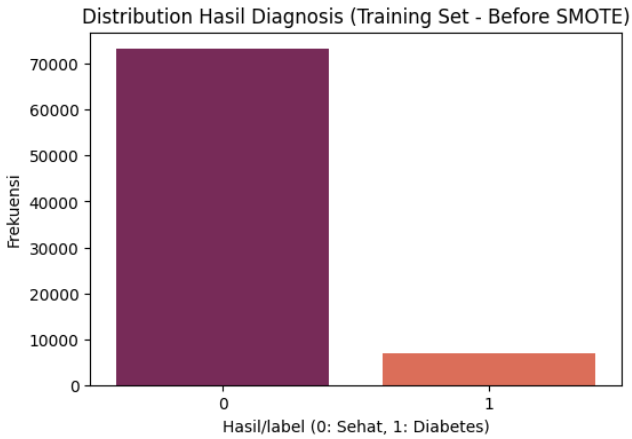
Selain pembagian data statis, penelitian ini juga menerapkan Stratified K-Fold Cross-Validation ($k = 20$) selama proses pelatihan untuk memvalidasi konsistensi performa model dan mengurangi bias variansi.

TABEL II
RINGKASAN PRA-PEMROSESAN DATA

Jenis Fitur	Metode
Numerik (Usia, BMI, HbA1c, Glukosa)	Min–Max Scaling
Kategorikal (Jenis Kelamin, Riwayat Merokok)	One-Hot Encoding
Biner (Hipertensi, Penyakit Jantung)	Tanpa Transformasi
Missing Values	Tidak Ada

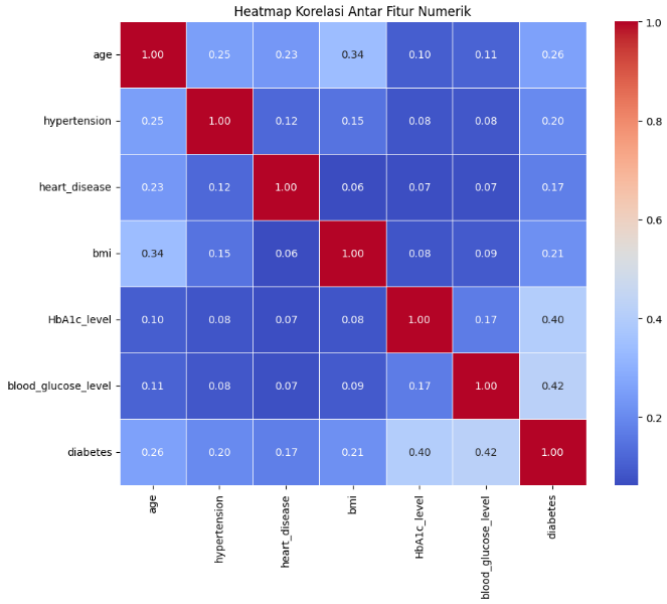
C. Exploratory Data Analysis (EDA)

Analisis Eksplorasi Data (EDA) dilakukan untuk memetakan distribusi kelas dan mengidentifikasi fitur yang paling berpengaruh terhadap diagnosis. Visualisasi pada Gambar 3 mengungkap adanya ketimpangan kelas (class imbalance) yang ekstrem, di mana pasien positif diabetes hanya mencakup 8,5% dari total populasi, sedangkan 91,5% sisanya adalah pasien non-diabetes. Disparitas rasio ini mengonfirmasi urgensi penerapan teknik resampling (SMOTE) pada data latih guna mencegah bias prediksi model terhadap kelas mayoritas [12], [13].



Gambar 2. Distribusi Kelas Target pada Dataset Diabetes

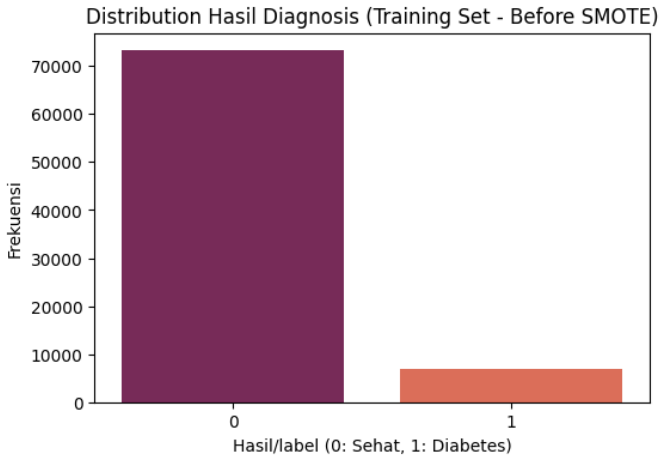
Selanjutnya, analisis hubungan antar fitur dievaluasi menggunakan *Heatmap Correlation Matrix* (Gambar 4). Hasil menunjukkan bahwa fitur *blood_glucose_level* dan *HbA1c_level* memiliki korelasi positif terkuat terhadap variabel target dengan koefisien masing-masing sebesar 0.42 dan 0.40. Nilai ini mengindikasikan bahwa indikator glukosa darah merupakan prediktor yang jauh lebih dominan dibandingkan fitur demografis (seperti *gender* atau *smoking history*). Temuan ini sejalan dengan literatur medis yang menempatkan kadar gula darah sebagai parameter diagnostik utama DM Tipe 2 [2].



Gambar 3. Matriks Korelasi Antar Variabel Klinis

D. Handling Imbalance with SMOTE

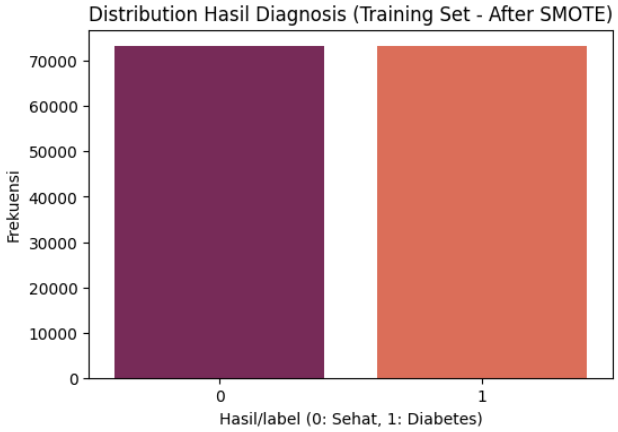
Visualisasi EDA (Gambar 3) mengungkap ketimpangan kelas ekstrem pada dataset (8,5% vs 91,5%). Guna memitigasi bias prediksi terhadap kelas mayoritas, teknik Synthetic Minority Over-sampling Technique (SMOTE) diterapkan.Sesuai kaidah metodologi yang ketat, SMOTE diterapkan secara eksklusif hanya pada Data Latih (Training Set) setelah proses pembagian data (splitting). Data Uji dibiarkan tetap pada distribusi aslinya untuk menjamin evaluasi model yang realistis. Dalam implementasinya, algoritma SMOTE dikonfigurasi menggunakan parameter $k_{neighbords} = 5$ dan $random_{state} = 42$.



Gambar 4. Distribusi Kelas Dataset Sebelum SMOTE

Secara kuantitatif, proses ini meningkatkan jumlah sampel kelas minoritas (Diabetes) pada data latih secara signifikan hingga setara dengan kelas mayoritas, yaitu mencapai 73.200 sampel per kelas. Akibatnya, total volume data latih meningkat menjadi 146.400 sampel dengan distribusi yang

seimbang sempurna (50:50). Strategi ini krusial untuk mencegah model "terbiasa" hanya memprediksi kelas mayoritas dan terbukti meningkatkan sensitivitas (Recall) dalam mendeteksi pasien positif diabetes [10], [21].



Gambar 5. Distribusi Kelas Dataset Sesudah SMOTE

E. Classification Algorithms

Setelah tahap *resampling* SMOTE, dataset dipartisi menjadi data latih (*training*) dan data uji (*testing*) dengan rasio 80:20. Penelitian ini mengimplementasikan tiga algoritma berbasis *ensemble learning* yang dipilih berdasarkan keunggulannya dalam menangani data medis terstruktur yang kompleks.

1) Rundom Forest: Random Forest adalah algoritma ensemble yang membangun sekumpulan pohon keputusan (decision trees) secara paralel menggunakan teknik Bootstrap Aggregating (Bagging). Setiap pohon dilatih pada subset data dan fitur yang dipilih secara acak guna menciptakan keragaman (diversity) model. Prediksi akhir ditentukan melalui mekanisme majority voting untuk klasifikasi. Pendekatan ini terbukti efektif dalam mereduksi varians model tunggal dan meningkatkan resistensi terhadap noise, sehingga meminimalisir risiko overfitting [4], [22].

2) Logistic Regression: Logistic Regression adalah metode statistik linear yang digunakan sebagai model *baseline* untuk klasifikasi biner[8]. Algoritma ini memodelkan probabilitas terjadinya suatu peristiwa (kelas positif diabetes) dengan memetakan *output* persamaan linear ke dalam rentang probabilitas [0, 1] menggunakan fungsi transformasi Sigmoid. Keunggulan utama metode ini terletak pada efisiensi komputasi dan interpretabilitasnya yang tinggi dalam menjelaskan hubungan antar variabel. Pendekatan ini telah terbukti efektif dalam memprediksi luaran klinis pada studi diabetes sebelumnya, [5]. Secara matematis, prediksi probabilitas diformulasikan menggunakan notasi ringkas berikut:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

Dengan:

- $P(y = 1|x)$: Probabilitas pasien terdiagnosis positif diabetes berdasarkan input data.
- β_0 : *Intercept* (konstanta bias) model.
- β_i : Koefisien regresi (bobot) untuk fitur ke-i yang menunjukkan signifikansi pengaruhnya.
- x_i : Nilai variabel prediktor (fitur) ke-i (seperti *Age*, *BMI*, *Glucose*).
- n : Jumlah total fitur input yang digunakan.

3) Light Gradient Boosting Machine (LightGBM): LightGBM adalah kerangka kerja gradient boosting yang mengoptimalkan efisiensi komputasi melalui strategi pertumbuhan pohon berbasis daun (leaf-wise growth). Berbeda dengan algoritma konvensional yang tumbuh secara level-wise, strategi leaf-wise memilih daun dengan penurunan kerugian (loss) terbesar untuk diekspansi, yang mempercepat konvergensi. Fungsi objektif LightGBM untuk meminimalkan kesalahan prediksi didefinisikan sebagai berikut:

$$Obj^t = \sum L(y_i, \widehat{\ } \{y\}_i) + \sum \Omega(f_k)$$

Dengan:

- $L(y_i, \hat{y}_i)$: Fungsi kerugian (loss function) yang mengukur selisih antara target asli dan prediksi.
- $\Omega(f_k)$: Suku regularisasi untuk mengontrol kompleksitas struktur pohon guna mencegah overfitting.

4) CatBoost: CatBoost merupakan algoritma gradient boosting mutakhir yang dirancang untuk menangani pergeseran prediksi melalui skema *Ordered Boosting*. Algoritma ini membangun pohon simetris (*oblivious trees*) yang menjamin eksekusi inferensi sangat cepat dan stabil. Secara matematis, pembaruan prediksi pada setiap iterasi untuk meminimalkan residual diformulasikan sebagai:

$$F_{t(x)} = F_{(t-1)(x)} + \alpha \cdot h_{t(x)}$$

Dengan:

- $F_t(x)$: Nilai prediksi model pada iterasi ke-t.
- $F_t{}^{-1}(x)$: Prediksi dari iterasi sebelumnya.
- α : Learning rate (laju pembelajaran).
- $h_t(x)$: Fungsi dasar (pohon baru) yang meminimalkan residual dari langkah sebelumnya.

F. Performance Evaluation

Evaluasi model dilakukan menggunakan Confusion Matrix yang memetakan prediksi ke dalam empat kuadran: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Karena dataset memiliki ketidakseimbangan kelas dan kesalahan diagnosis diabetes berisiko fatal, metrik Akurasi semata tidak cukup representatif.

Oleh karena itu, penelitian ini memprioritaskan metrik Recall (Sensitivitas) untuk meminimalkan *False Negative* (pasien positif yang salah didiagnosis sehat), serta F1-Score untuk mengukur keseimbangan harmonis antara presisi dan sensitivitas. Definisi matematis metrik evaluasi dinyatakan sebagai berikut:

1) Accuracy Mengukur rasio prediksi benar secara keseluruhan.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Precision Mengukur ketepatan prediksi positif (penting untuk menghindari diagnosis berlebih/cemas palsu).

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

3) Recall (Sensitivity) Mengukur kemampuan model mendeteksi seluruh pasien positif (metrik terpenting dalam skrining penyakit).

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

4) F1-Score Rata-rata harmonis antara Precision dan Recall, memberikan gambaran performa yang objektif pada data tidak seimbang.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Hasil perhitungan dari seluruh metrik di atas selanjutnya akan dikomparasi secara komprehensif untuk menentukan algoritma yang memiliki performa diagnostik paling optimal dalam memprediksi risiko diabetes.

G. K-Fold Cross-Validation

Penelitian ini menerapkan skema 5-Fold Cross-Validation (k = 5) untuk memitigasi risiko overfitting dan menjamin evaluasi yang lebih objektif dibandingkan metode pembagian statis (hold-out). Mekanisme ini mempartisi dataset menjadi lima bagian (folds), di mana setiap bagian digunakan secara bergantian sebagai data validasi sementara empat bagian lainnya berfungsi sebagai data latih. Pendekatan ini memastikan model mempelajari seluruh distribusi data sehingga bias varians dapat diminimalkan [23]. Performa akhir model dihitung berdasarkan rata-rata aritmatika dari setiap iterasi:

$$CV_{Score} = \frac{1}{k}\sum_{i=1}^{k} Score_i$$

Di mana $Score_i$ adalah metrik evaluasi pada iterasi ke-i. Metode ini merupakan standar validasi yang *robust* dalam studi komparasi algoritma pada data medis [24], [19].

III. Result And Discussion

Pengujian model dilakukan menggunakan dataset yang telah melalui proses penyeimbangan data (*data balancing*) menggunakan teknik SMOTE. Data dibagi dengan rasio 80:20, di mana model dilatih pada 80% data latih dan dievaluasi pada 20% data uji. Berikut adalah hasil evaluasi komparatif antara algoritma CatBoost, LightGBM, dan Random Forest.

A. Model Performance Comparison

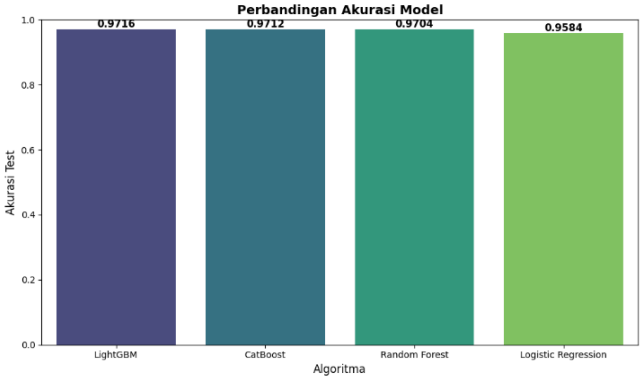
Evaluasi awal dilakukan dengan membandingkan metrik standar klasifikasi yang meliputi Akurasi, Presisi, *Recall*, dan F1-*Score*. Ringkasan performa ketiga algoritma disajikan pada Tabel II.

TABEL III
PERBANDINGAN PERFORMA ALGORITMA

Algoritma	Akurasi	Presisi	Recall	F1 Score
LightGBM	97,16	96,70	68,88	80,45
CatBoost	97,12	95,18	69,71	80,48
Random Forest	97,04	94,82	68.94%	79,84
Random Forest	95,84	84,50	62,53	71,87

Berdasarkan Tabel II, algoritma berbasis *gradient boosting* (LightGBM dan CatBoost) menunjukkan dominasi performa yang signifikan dibandingkan Random Forest dan Logistic Regression. LightGBM mencatatkan akurasi pengujian tertinggi sebesar 97,16%, disusul sangat ketat oleh CatBoost dengan 97,12%, dan Random Forest sebesar 97,04%. Sementara itu, Logistic Regression tertinggal dengan akurasi 95,84%, yang mengindikasikan keterbatasan model linear dalam menangkap pola kompleks pada data medis ini.
Selisih performa akurasi yang sangat marjinal (<0,1%) antara LightGBM dan CatBoost menunjukkan persaingan ketat. Meskipun LightGBM unggul dalam metrik akurasi global dan presisi (96,70%), analisis lebih mendalam pada metrik sensitivitas (*Recall*) dan *F1-Score* menyingkap keunggulan CatBoost.

Model CatBoost berhasil mencatatkan nilai Recall tertinggi sebesar 69,71% dan F1-Score tertinggi sebesar 80,48%, mengungguli LightGBM (*Recall* 68,88%) dan Random Forest (*Recall* 68,94%). Dalam konteks diagnosis medis, keunggulan *Recall* pada CatBoost ini sangat krusial karena merepresentasikan kemampuan model yang lebih baik dalam mendeteksi seluruh kasus positif (meminimalkan *False Negative*), meskipun harus sedikit mengorbankan presisi. Oleh karena itu, keseimbangan performa yang ditunjukkan oleh tingginya *F1-Score* menjadikan CatBoost kandidat model yang paling robust.



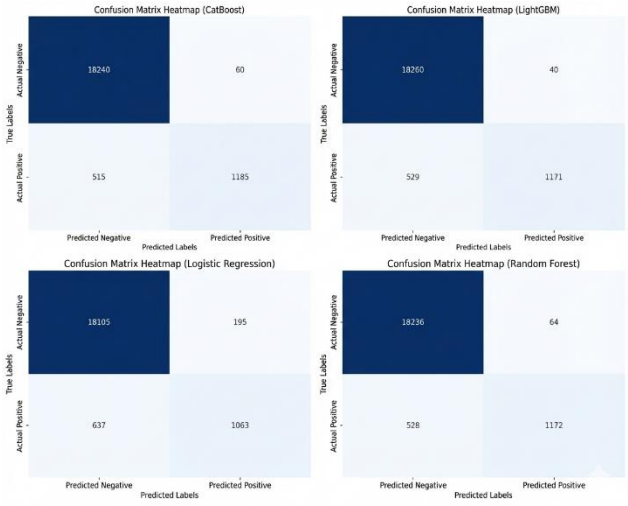
Gambar 6. Grafik Perbandingan Akurasi Antar Model

Namun, temuan menarik dan krusial terlihat pada metrik *Recall*. Meskipun LightGBM mencatatkan akurasi global tertinggi, CatBoost justru menunjukkan keunggulan dalam sensitivitas deteksi. Berdasarkan data *Confusion Matrix*,

CatBoost terbukti lebih andal dalam menangkap kasus positif (Diabetes) dibandingkan LightGBM dan Random Forest. Hal ini mengindikasikan bahwa meskipun Random Forest dan LightGBM memiliki presisi yang kompetitif, CatBoost memiliki karakteristik yang paling sensitif dalam meminimalkan *False Negative*, sebuah atribut yang sangat vital dalam diagnosis medis.

B. Confusion Matrix Analysis

Untuk menganalisis lebih dalam mengenai *trade-off* antara Presisi dan *Recall*, dilakukan evaluasi menggunakan *Confusion Matrix* yang memetakan distribusi kesalahan prediksi (*False Positive* dan *False Negative*). Visualisasi untuk ketiga model ditampilkan pada Gambar 6.



Gambar 7. Komparasi Confusion Matrix

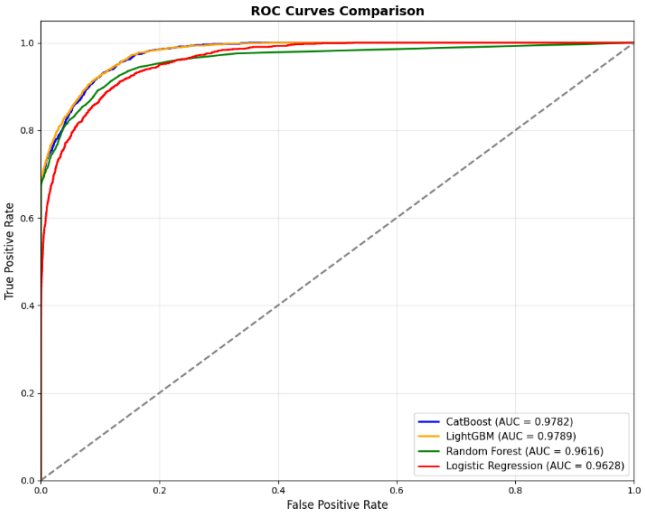
Analisis *Confusion Matrix* mengungkap temuan krusial yang menjadi dasar pemilihan model:

1) *LightGBM (Akurasi Tertinggi):* Meskipun memiliki akurasi total tertinggi, model ini mencatatkan jumlah True Positive (pasien diabetes terdeteksi benar) sebesar 1.171 orang.
2) *CatBoost (Sensitivitas Terbaik):* Meskipun akurasinya sedikit di bawah LightGBM, CatBoost berhasil mengidentifikasi 1.185 kasus positif secara tepat. Artinya, CatBoost mampu menyelamatkan 14 pasien lebih banyak dari risiko kesalahan diagnosis (False Negative) dibandingkan LightGBM.
3) *Logistic Regression & Random Forest:* Kedua model ini menghasilkan tingkat kesalahan klasifikasi yang lebih tinggi dibandingkan metode boosting, sehingga kurang ideal untuk diterapkan pada sistem diagnosis kritis.

Secara konseptual, keunggulan LightGBM dalam akurasi disebabkan oleh strategi pertumbuhan pohon leaf-wise yang agresif meminimalkan loss function. Namun, pendekatan ini terkadang kurang stabil pada data minoritas dibandingkan strategi symmetric trees yang digunakan CatBoost. Dalam konteks medis, meminimalkan False Negative (pasien sakit yang tidak terdeteksi) jauh lebih prioritas daripada mengejar akurasi rata-rata. Oleh karena itu, CatBoost ditetapkan sebagai model terbaik dalam penelitian ini karena memiliki keseimbangan optimal antara akurasi tinggi dan sensitivitas (Recall) yang superior.

C. ROC-AUC Performance Analysis

Validasi stabilitas model dilakukan dengan menganalisis kurva *Receiver Operating Characteristic* (ROC) dan nilai *Area Under Curve* (AUC). Sebagaimana terlihat pada Gambar (Kurva ROC), LightGBM dan CatBoost mencatatkan nilai AUC yang hampir identik, masing-masing sebesar 0.9789 dan 0.9782.



Gambar 8. Kurva ROC dan Nilai AUC

Nilai AUC yang mendekati angka 1.0 ini membuktikan bahwa kedua model memiliki kemampuan diskriminasi yang "sangat baik" (*excellent classification*) dalam membedakan kelas diabetes dan non-diabetes pada berbagai ambang batas (*threshold*). Meskipun LightGBM unggul tipis dalam probabilitas peringkat, stabilitas deteksi kelas positif pada CatBoost tetap menjadikannya pilihan yang lebih aman untuk implementasi klinis.

D. Discussion: LightGBM vs CatBoost

Perbedaan performa antara LightGBM dan CatBoost dapat dijelaskan melalui arsitektur dasar dan mekanisme penanganan data pada kedua algoritma. LightGBM menggunakan teknik Gradient-based One-Side Sampling (GOSS) yang memprioritaskan sampel data dengan gradien (error) besar untuk mempercepat pelatihan. Pendekatan ini memungkinkan LightGBM belajar sangat cepat dan agresif pada pola dominan, yang menjelaskan mengapa model ini mencapai skor akurasi tertinggi (97,16%) dalam eksperimen ini. Namun, keunggulan CatBoost dalam metrik sensitivitas (*Recall*) dan stabilitas model didasari oleh dua inovasi konseptual utama yang mengatasi kelemahan algoritma *boosting* tradisional:

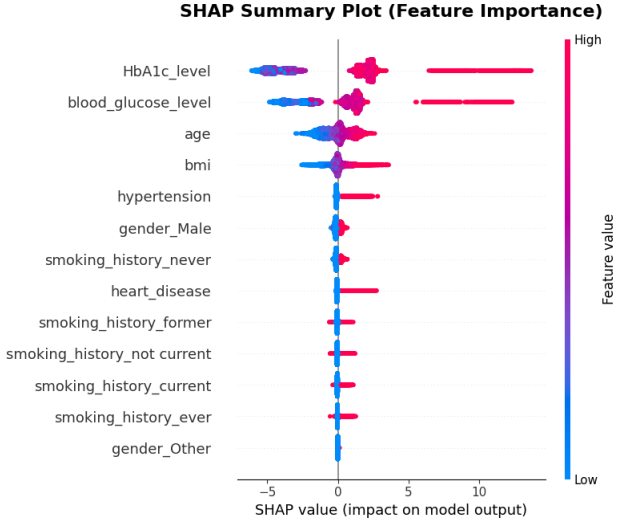
1) *Ordered Boosting & Penanganan Overfitting*: Berbeda dengan LightGBM yang rentan terhadap *target leakage* pada dataset kecil hingga menengah, CatBoost

menerapkan konsep Ordered Boosting. Mekanisme ini mengatasi masalah *prediction shift* dengan melakukan permutasi acak pada urutan data saat menghitung residu. Hal ini mencegah model "menghafal" target data latih (*overfitting*), sehingga model memiliki kemampuan generalisasi yang lebih baik saat menghadapi data uji yang belum pernah dilihat (seperti pasien baru).

2) *Symmetric Trees & Fitur Kategorikal:* CatBoost menggunakan struktur pohon simetris (*Oblivious Trees*), di mana pemisahan (*split*) yang sama diterapkan di seluruh level pohon. Struktur ini memberikan regularisasi yang lebih kuat dan mengurangi varians model, menjadikannya lebih stabil dibandingkan pohon asimetris LightGBM (*leaf-wise growth*). Selain itu, kemampuan CatBoost dalam menangani fitur kategorikal secara native (tanpa perlu *One-Hot Encoding* yang masif) meminimalkan hilangnya informasi penting antar variabel.

E. Model Interpretation with SHAP

Pendekatan Explainable AI (XAI) menggunakan kerangka kerja SHAP (SHapley Additive exPlanations) diterapkan pada model terpilih, CatBoost, untuk mengatasi sifat Black Box dan membedah mekanisme pengambilan keputusan klinis. Berdasarkan SHAP Summary Plot pada Gambar 8, fitur HbA1c_level dan blood_glucose_level teridentifikasi sebagai determinan paling dominan dengan tingkat kepentingan global tertinggi dibandingkan fitur lainnya.



Gambar 9. Interpretasi Fitur Menggunakan SHAP pada CatBoost

Analisis mendalam memperlihatkan polarisasi kontribusi yang tegas pada kedua fitur vital tersebut terhadap prediksi kelas positif (Diabetes) dan negatif (Non-Diabetes):

1) *Kontribusi terhadap Kelas Positif (Risiko Diabetes):* Visualisasi menunjukkan bahwa titik-titik berwarna merah (merepresentasikan nilai fitur tinggi) pada HbA1c_level dan blood_glucose_level terkonsentrasi kuat di sisi positif (kanan sumbu x). Hal ini mengonfirmasi bahwa peningkatan kadar gula darah dan HbA1c secara signifikan mendorong *log-odds* prediksi ke arah diagnosis diabetes. Ekor distribusi merah yang memanjang jauh ke kanan pada fitur HbA1c bahkan menegaskan bahwa nilai ekstrem pada parameter ini menjadi indikator prediktif yang nyaris mutlak bagi model dalam mendeteksi kasus positif.
2) *Kontribusi terhadap Kelas Negatif (Sehat):* Sebaliknya, nilai fitur yang rendah (ditandai dengan titik biru) berkumpul secara padat di sisi negatif (kiri). Fenomena ini mengindikasikan bahwa kadar gula darah yang berada dalam rentang normal berperan aktif dalam menurunkan skor risiko, sehingga mengarahkan prediksi model menuju kelas sehat (Non-Diabetes).

Selain indikator klinis utama, fitur demografis seperti Usia (*Age*) dan BMI juga menunjukkan tren linear yang konsisten: semakin tinggi nilainya (gradasi warna merah), semakin besar kontribusi positifnya terhadap risiko diabetes. Konsistensi logika ini memvalidasi model secara klinis, membuktikan bahwa algoritma CatBoost telah berhasil mempelajari patofisiologi diabetes yang valid—seperti pengaruh hiperglikemia kronis, penuaan, dan obesitas—sehingga aman dan layak diandalkan sebagai instrumen pendukung keputusan medis.

F. Implikasi Klinis dan Penerapan Prakti

Analisis *feature importance* melalui SHAP menyoroti dominasi HbA1c_level dan blood_glucose_level sebagai prediktor utama, yang menunjukkan koherensi kuat dengan pedoman klinis endokrinologi sekaligus memvalidasi kemampuan model dalam menangkap patofisiologi diabetes secara akurat. Dengan karakteristik sensitivitas (*Recall*) yang unggul dalam meminimalkan *False Negative*, model CatBoost ini memiliki implikasi praktis yang signifikan sebagai instrumen skrining awal (*early screening tool*) yang efisien. Penerapan algoritma ini memungkinkan fasilitas kesehatan melakukan stratifikasi risiko secara otomatis pada populasi besar, di mana pasien yang teridentifikasi berisiko tinggi dapat diprioritaskan untuk pemeriksaan laboratorium lanjutan. Pendekatan ini tidak hanya meningkatkan akurasi deteksi dini, tetapi juga menawarkan efisiensi alokasi sumber daya medis dengan memfokuskan intervensi klinis pada individu yang paling membutuhkan, didukung oleh transparansi keputusan yang disediakan oleh visualisasi SHAP untuk meningkatkan kepercayaan tenaga medis.

IV. KESIMPULAN

Penelitian ini telah berhasil mengevaluasi efektivitas empat algoritma *ensemble learning*— LightGBM, CatBoost, Random Forest, dan Logistic Regression—dalam memprediksi risiko diabetes menggunakan dataset yang diseimbangkan dengan teknik SMOTE. Berdasarkan hasil pengujian komprehensif, terdapat persaingan ketat antara dua algoritma berbasis *gradient boosting*. LightGBM mencatatkan Akurasi tertinggi sebesar 97,16%, sedikit mengungguli CatBoost yang meraih akurasi 97,12%.

Meskipun demikian, CatBoost ditetapkan sebagai model terbaik dalam penelitian ini karena keunggulannya pada metrik sensitivitas (*Recall*) yang mencapai 69,71% dan F1-Score sebesar 80,48%. Berbeda dengan temuan sebelumnya di mana Random Forest mendominasi sensitivitas, hasil eksperimen ini menunjukkan bahwa CatBoost lebih andal dalam meminimalkan *False Negative*, sebuah atribut krusial dalam diagnosis medis untuk memastikan pasien positif tidak terlewatkan. Sementara itu, Logistic Regression menunjukkan performa terendah, menegaskan perlunya model non-linear untuk data medis yang kompleks.

Penerapan kerangka kerja Explainable AI (XAI) melalui SHAP berhasil mengungkap transparansi model "Black Box", dengan mengidentifikasi Level HbA1c dan Level Glukosa Darah sebagai fitur paling dominan. Temuan ini memvalidasi kesesuaian logika model dengan standar medis klinis. Secara keseluruhan, integrasi CatBoost dengan teknik SMOTE dan interpretasi SHAP direkomendasikan sebagai solusi sistem pendukung keputusan klinis yang akurat, sensitif, dan dapat dijelaskan (explainable) untuk skrining awal diabetes.

DAFTAR PUSTAKA

[1] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 183, Jan. 2022, doi: 10.1016/j.diabres.2021.109119.

[2] E. Er Unja, B. Trihandini, and P. Sarjana Keperawatan dan Ners, "Journal of Nursing Invention Hubungan Kadar Gula Darah Dengan Hipertensi Pada Pasien Diabetes Melitus Tipe 2 Di Wilayah Kerja Puskesmas Teluk Tiram Kota Banjarmasin Tahun 2024", doi: 10.33859/jni.

[3] S. Rumondang, B. P. Sedli, and O. R. H. Umboh, "Pengaruh Inflamasi Mikro terhadap Penyakit Ginjal pada Pasien Diabetes Melitus Tipe-2 Microinflamation Influence on Renal Diseases in Type 2 Diabetes Mellitus," *Medical Scope Journal*, vol. 4, no. 1, pp. 40–47, 2022, doi: 10.35790/msj.v4.i1.44682.

[4] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[5] M. Fadli Kurniawan and D. Ayu Megawaty, "Comparison of Logistic Regression, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) Algorithms in Diabetes Prediction," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[6] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 706–716. doi: 10.1016/j.procs.2020.03.336.

[7] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. Part C Emerg. Technol.*, vol. 58, pp. 308–324, Sep. 2015, doi: 10.1016/j.trc.2015.02.019.

[8] R. Kaur, R. Kumar, S. Kaur, G. Singh, A. Kaur, and S. Singh, "Machine Learning for Diabetes Prediction: Performance Analysis Using Logistic Regression, Naïve Bayes, and Decision Tree Models," *Healthcraft Frontiers*, vol. 02, no. 04, pp. 169–187, Dec. 2024, doi: 10.56578/hf020401.

[9] Y. Zhang, "A Comparative Study of Logistic Regression and Machine Learning Models for Diabetes Prediction Using the BRFSS Dataset," *Applied and Computational Engineering*, vol. 196, no. 1, pp. 177–185, Oct. 2025, doi: 10.54254/2755-2721/2025.ld28522.

[10] P. Branco, I. Torgo, R. P. Ribeiro, and L. Torgo, "A Survey of Predictive Modeling on Imbalanced Do-mains," 2016.

[11] T. F. Sukamto, C. L. Prameswary, D. Royadi, and D. Sofia, "Diabetes Disease Prediction on Unbalanced Data Using SMOTE-Tomek Links and Random Forest Algorithm," *G-Tech: Jurnal Teknologi Terapan*, vol. 9, no. 3, pp. 1194–1203, Jul. 2025, doi: 10.70609/g-tech.v9i3.7164.

[12] K. Ujaran, K. Ridwan, E. Heni Hermaliani, M. Ernawati, and C. Author, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada," 2024. [Online]. Available: http://jurnal.bsi.ac.id/index.php/co-science

[13] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, Jul. 2022, doi: 10.30812/matrik.v21i3.1726.

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

[15] A. Aich, M. M. Murshed, S. Hewage, and A. Mayeaux, "CopulaSMOTE: A Copula-Based Oversampling Approach for Imbalanced Classification in Diabetes Prediction," Sep. 2025, [Online]. Available: http://arxiv.org/abs/2506.17326

[16] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." [Online]. Available: https://github.com/Microsoft/LightGBM.

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1706.09516

[18] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[19] A. Zarghani, "Comparative Analysis of LSTM Neural Networks and Traditional Machine Learning Models for Predicting Diabetes Patient Readmission."

[20] Y. Zhang, H. Zhang, D. Wang, N. Li, H. Lv, and G. Zhang, "Development of a 5-Year Risk Prediction Model for Transition From Prediabetes to Diabetes Using Machine Learning: Retrospective Cohort Study," *J. Med. Internet Res.*, vol. 27, no. 1, 2025, doi: 10.2196/73190.

[21] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.

[22] M. Rahman, D. Islam, R. J. Mukti, and I. Saha, "A deep learning approach based on convolutional LSTM for detecting diabetes," *Comput. Biol. Chem.*, vol. 88, Oct. 2020, doi: 10.1016/j.compbiolchem.2020.107329.

[23] Z. Rafie, M. S. Talab, B. E. Z. Koor, A. Garavand, C. Salehnasab, and M. Ghaderzadeh, "Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes," *BMC Public Health*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12889-025-24953-w.

[24] M. Hasan and F. Yasmin, "Predicting Diabetes Using Machine Learning: A Comparative Study of Classifiers."