Journal of Informatics, Electrical and Electronics Engineering
ISSN 2807-9507 (Media Online)
Vol 5, No 4, Juni 2026
Hal 297-305
https://djournals.com/jieee

Penerapan SMOTE pada Algoritma LightGBM dan XGBoost untuk Klasifikasi Penyakit Diabetes

Regita Audyna Siregar[1,*], Solikhun[2], Timbo Faritcan P. Siallagan[2]

[1]Program Studi Sistem Informasi, Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematangsiantar, Indonesia
[2] Program Studi Teknik Informatika, Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematangsiantar, Indonesia
Email: [1]regitasiregar30@gmail.com, [2]solikhun@amiktunasbangsa.ac.id, [3]timbofaritcansiallagan@gmail.com
Email Penulis Korespondensi: regitasiregar30@gmail.com

Abstrak−Diabetes melitus merupakan penyakit metabolik kronis yang prevalensinya terus meningkat dan sering terlambat terdiagnosis akibat gejala awal yang kurang disadari. Salah satu tantangan dalam klasifikasi diabetes menggunakan *machine learning* adalah ketidakseimbangan kelas pada dataset yang dapat menyebabkan model cenderung bias terhadap kelas mayoritas sehingga kemampuan mendeteksi pasien diabetes menjadi rendah. Penelitian ini bertujuan membandingkan kinerja algoritma Light Gradient Boosting Machine (LightGBM) dan Extreme Gradient Boosting (XGBoost) yang dipadukan dengan *Synthetic Minority Oversampling Technique* (SMOTE) untuk klasifikasi penyakit diabetes. Penelitian menggunakan Pima Indians Diabetes Dataset yang terdiri atas 768 data dengan 8 atribut dan mengikuti tahapan *Machine Learning Life Cycle* (MLLC), meliputi analisis eksplorasi data, normalisasi *MinMax Scaler*, seleksi fitur menggunakan Mutual Information, penanganan ketidakseimbangan kelas dengan SMOTE, pelatihan model, serta evaluasi menggunakan *accuracy*, *precision, recall*, *F1-score*, dan *ROC-AUC*. Hasil penelitian menunjukkan bahwa penerapan SMOTE mampu meningkatkan performa kedua model secara signifikan. Model LightGBM dengan SMOTE memperoleh kinerja terbaik dengan *accuracy* 85,33%, *precision* 81,55%, *recall* 91,33%, *F1-score* 86,16%, dan *ROC-AUC* 90,49%, sedangkan XGBoost dengan SMOTE memperoleh *accuracy* 83,00%, *precision* 79,64%, *recall* 88,67%, *F1-score* 83,91%, dan *ROC-AUC* 90,04%. Hasil tersebut menunjukkan bahwa kombinasi LightGBM dan SMOTE lebih efektif dalam mendeteksi kasus diabetes dibandingkan XGBoost serta berpotensi digunakan sebagai sistem pendukung keputusan untuk skrining dini diabetes.

Kata Kunci: Diabetes Melitus; Klasifikasi; LightGBM; XGBoost; SMOTE; Mutual Information

Abstract−Diabetes mellitus is a chronic metabolic disease whose prevalence continues to increase worldwide and is often diagnosed late owing to insufficient recognition of its early symptoms. One of the major challenges in diabetes classification using machine learning is class imbalance in the dataset, which may cause the model to disproportionately favor the majority class and consequently compromise its capacity to accurately identify diabetic patients. Accordingly, this research seeks to compare the classification effectiveness of Light Gradient Boosting Machine (LightGBM) and Extreme Gradient Boosting (XGBoost) algorithms combined with the Artificial Minority Oversampling Approach (SMOTE) applied to diabetes classification task. The study utilized the Pima Indians Diabetes Dataset consisting of 768 instances and 8 attributes and followed the Machine Learning Life Cycle (MLLC), including exploratory data analysis, MinMax Scaler normalization, feature selection using Mutual Information, class imbalance handling with SMOTE, model training, and evaluation using accuracy, precision, recall, F1-score, and ROC-AUC as quantitative measures. Analysis of the results demonstrates that the application of SMOTE significantly improved the performance of both models. LightGBM with SMOTE achieved the best performance, with an accuracy of 85.33%, precision of 81.55%, recall of 91.33%, F1-score of 86.16%, and ROC-AUC of 90.49%, while XGBoost with SMOTE achieved an accuracy of 83.00%, precision of 79.64%, recall of 88.67%, F1-score of 83.91%, and ROC-AUC of 90.04%. These findings demonstrate that the combination of LightGBM and SMOTE is more effective than XGBoost in detecting diabetes cases and holds considerable promise for deployment as a clinical decision-making tool in early diabetes screening.

Keywords: Diabetes Mellitus; Classification; LightGBM; XGBoost; SMOTE; Mutual Information

1. PENDAHULUAN

Hiperglikemia yang disebabkan oleh penurunan produksi insulin, resistensi insulin, atau keduanya merupakan ciri dari khas diabetes mellitus, yaitu suatu gangguan metabolik kronis [1], [2]. Penyakit ini memberikan dampak yang serius di seluruh dunia dan kondisinya terus memburuk.. International Diabetes Federation (IDF) mencatat bahwa pada tahun 2021, sekitar 10,5% populasi dewasa usia 20–79 tahun telah terdiagnosis diabetes, dan hampir separuhnya tidak menyadari kondisi tersebut. Proyeksi IDF memperkirakan jumlah penderita diabetes akan mencapai 783 juta jiwa pada tahun 2045, sementara World Health Organization (WHO) melaporkan angka kematian akibat diabetes mencapai satu juta jiwa per tahun [3].

Deteksi dini merupakan faktor krusial dalam penatalaksanaan diabetes guna mencegah komplikasi serius seperti kerusakan organ, stroke, dan penyakit jantung [4], [5]. Namun demikian, gejala khas diabetes seperti polidipsia, polifagia, dan poliuria kerap tidak disadari atau diabaikan oleh penderita, sehingga diagnosis sering terlambat ditegakkan [6], [7]. Kondisi ini mendorong kebutuhan akan sistem deteksi dini yang akurat berbasis teknologi kecerdasan buatan.

Penelitian yang mengenai klasifikasi diabetes dengan menggunakan algoritma pembelajaran mesin mengalami perkembangan yang pesat.. Mulyani et al. membandingkan kinerja algoritma K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM) menggunakan teknik SMOTE pada dataset Pima Indians Diabetes. Hasil penelitian tersebut menunjukkan bahwa KNN dengan k=3 mencapai akurasi 83,33%, presisi 85,00%, recall 83,95%, dan F1 Score 84,47%, sedangkan SVM dengan kernel RBF memperoleh akurasi 81,67%, presisi 85,91%, recall 79,01%, dan F1 Score 82,32% [8]. Meskipun keduanya menunjukkan performa yang cukup baik, nilai recall yang belum optimal mengindikasikan masih terdapat sejumlah kasus diabetes yang gagal terdeteksi oleh model.

Persoalan recall yang rendah dalam klasifikasi diabetes bersifat kritis dari perspektif klinis, sebab kasus *false negative* yakni pasien diabetes yang tidak terdeteksi oleh sistem berpotensi menimbulkan konsekuensi medis yang jauh

lebih serius dibandingkan *false positive* [9], [10], [11]. Oleh karena itu, pengembangan model klasifikasi dengan nilai recall yang lebih tinggi merupakan prioritas utama dalam konteks deteksi dini berbasis data kesehatan.

Salah satu tantangan mendasar dalam klasifikasi diabetes adalah ketidakseimbangan kelas (*class imbalance*) pada dataset, ketika ukuran dari sampel kelompok mayoritas (yang tidak menderita diabetes) secara signifikan lebih besar daripada ukuran sampel kelompok minoritas (yang menderita diabetes) [12]. Pada dataset Pima Indians Diabetes, rasio ketidakseimbangan mencapai 1,87:1 antara kelas non-diabetes (500 sampel) dan kelas diabetes (268 sampel) [13]. Status kinerja dari deteksi kelas minoritas menurun mengakibatkan adanya bias model yang mengutamakan kelas mayoritas. Masalah dari hal ini telah berhasil diatasi dengan Teknik *Oversampling* Minoritas Sintetis (SMOTE), yang menghasilkan sampel sintetis untuk kelas minoritas menggunakan interpolasi k-tetangga terdekat. [14], [15].

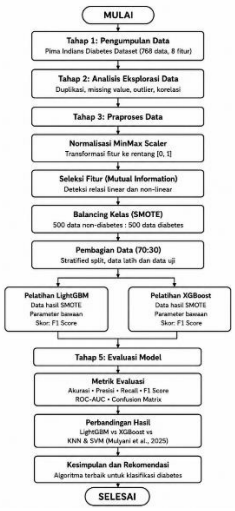
Di samping persoalan ketidakseimbangan kelas, pemilihan algoritma klasifikasi yang tepat juga berpengaruh signifikan terhadap kinerja model. Dua algoritma berbasis gradient boosting yang banyak dikaji dalam literatur machine learning adalah Extreme Gradient Boosting (XGBoost) dan Light Gradient Boosting Machine (LightGBM). Pemodelan LightGBM menawarkan efisiensi komputasi yang lebih tinggi melalui teknik pertumbuhan pohon secara bertahap per daun, XGBoost dikenal karena kemampuan regulasi yang efisien dalam mencegah overfitting. [16], [17], [18]. Perbedaan karakteristik tersebut menjadikan kedua algoritma menarik untuk dibandingkan dalam klasifikasi diabetes, khususnya pada data yang mengalami ketidakseimbangan kelas. Oleh karena itu, penelitian ini memanfaatkan kedua algoritma tersebut untuk mengevaluasi efektivitasnya setelah diterapkan teknik penyeimbangan data menggunakan SMOTE.

Meskipun berbagai penelitian telah menerapkan algoritma machine learning untuk klasifikasi diabetes, masih terdapat beberapa keterbatasan dalam penelitian sebelumnya. Sebagian penelitian masih berfokus pada algoritma konvensional seperti K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree, dan Random Forest, sedangkan penerapan algoritma gradient boosting, khususnya LightGBM dan XGBoost, pada dataset diabetes yang memiliki ketidakseimbangan kelas masih relatif terbatas. Selain itu, penelitian terdahulu cenderung menggunakan akurasi sebagai indikator utama kinerja, sehingga kemampuan model dalam mendeteksi kelas minoritas, terutama melalui recall, F1-Score, dan ROC-AUC, belum selalu dievaluasi secara komprehensif. Kondisi tersebut menunjukkan adanya research gap berupa kebutuhan untuk membandingkan kinerja LightGBM dan XGBoost pada data diabetes yang tidak seimbang dengan mempertimbangkan penanganan ketidakseimbangan kelas menggunakan SMOTE serta evaluasi menggunakan beberapa metrik kinerja secara bersamaan.

Berdasarkan research gap tersebut, penelitian ini memberikan beberapa kontribusi. Pertama, penelitian ini melakukan perbandingan langsung antara LightGBM dan XGBoost untuk klasifikasi penyakit diabetes pada dataset yang mengalami ketidakseimbangan kelas dengan menerapkan SMOTE sebagai teknik penyeimbangan data. Kedua, penelitian ini menggunakan Mutual Information untuk menganalisis tingkat keterkaitan setiap fitur terhadap variabel target sehingga dapat memberikan gambaran mengenai fitur yang paling informatif dalam klasifikasi diabetes. Ketiga, penelitian ini mengevaluasi kinerja model menggunakan accuracy, precision, recall, F1-Score, dan ROC-AUC sehingga performa model tidak hanya dinilai berdasarkan ketepatan prediksi, tetapi juga kemampuan dalam mendeteksi kasus diabetes dan membedakan kelas secara keseluruhan. Kontribusi tersebut diharapkan dapat memberikan perbandingan yang lebih komprehensif mengenai kinerja LightGBM dan XGBoost serta membantu menentukan model yang lebih sesuai untuk mendukung penelitian klasifikasi diabetes berbasis machine learning.

2. METODOLOGI PENELITIAN

Penelitian ini mengacu pada kerangka Machine Learning Life Cycle (MLLC) yang mencakup lima tahap utama: pengumpulan data, analisis eksplorasi data, praproses data, pelatihan dan model, serta evaluasi dan perbandingan model. Setiap tahap dirancang secara sistematis untuk memastikan validitas dan reproduktibilitas hasil eksperimen.



Gambar 1. Alur Kerja Penelitian

Berdasarkan Gambar 1, Penelitian diawali dengan pengumpulan data menggunakan Pima Indians Diabetes Dataset yang terdiri atas 768 data dan delapan atribut. Selanjutnya dilakukan *Exploratory Data Analysis* (EDA) untuk mengidentifikasi duplikasi data, *missing value*, *outlier*, dan hubungan antarvariabel. Tahap praproses meliputi normalisasi menggunakan MinMax Scaler, seleksi fitur dengan Mutual Information, serta penanganan ketidakseimbangan kelas menggunakan SMOTE. Data kemudian dibagi menjadi data latih dan data uji dengan rasio 70:30 menggunakan *stratified split*. Model LightGBM dan XGBoost dilatih menggunakan data hasil penyeimbangan kelas, kemudian dievaluasi menggunakan *confusion matrix*, akurasi, presisi, *recall*, F1-Score, dan ROC-AUC. Hasil evaluasi selanjutnya dibandingkan dengan penelitian terdahulu untuk menentukan model terbaik dalam klasifikasi diabetes, yang menjadi dasar penyusunan kesimpulan dan rekomendasi penelitian.

2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah Pima Indians Diabetes Dataset yang bersumber dari platform Kaggle. Dataset ini terdiri atas 768 rekam data dengan delapan atribut numerik dan satu label target biner (Outcome). Delapan atribut tersebut meliputi jumlah kehamilan (Pregnancies), kadar glukosa darah (Glucose), tekanan darah diastolik (BloodPressure), ketebalan kulit (SkinThickness), kadar insulin (Insulin), indeks massa tubuh (BMI), fungsi riwayat diabetes (DiabetesPedigreeFunction), dan usia (Age). Pasien dengan diabetes memiliki label target 1, sedangkan mereka yang tidak menderita kondisi tersebut memiliki label target 0. Dengan 500 titik data pada kelas 0 (bukan diabetes) dan 268 titik data pada kelas 1 (diabetes), distribusi kelas pada kumpulan data tersebut menunjukkan ketidakseimbangan yang cukup besar, dengan rasio ketidakseimbangan sebesar 1,87:1.

2.2 Analisis Eksplorasi Data

Tahap analisis eksplorasi data (Exploratory Data Analysis/EDA) dilakukan untuk memahami karakteristik dataset sebelum praproses. Tahap ini mencakup pemeriksaan duplikasi data, deteksi nilai hilang (missing value), analisis distribusi kelas, identifikasi outlier menggunakan visualisasi boxplot, serta analisis korelasi antar atribut menggunakan Pearson Correlation [19]. Analisis korelasi dilakukan untuk mengetahui hubungan linear antar fitur terhadap variabel target. Nilai korelasi yang mendekati +1 menunjukkan hubungan positif yang kuat, mend ekati −1 menunjukkan hubungan negatif yang kuat, sedangkan nilai mendekati 0 menunjukkan tidak adanya hubungan linear antar variabel.

2.3 Praproses Data

2.3.1 Normalisasi MinMax Scaler

Normalisasi dilakukan menggunakan MinMax Scaler untuk mentransformasi nilai atribut ke rentang [0,1] sehingga perbedaan skala antar fitur tidak memengaruhi proses pelatihan model. Persamaan normalisasi ditunjukkan pada Persamaan berikut:

$$x' = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (1)$$

dimana (x') sebagai nilai hasil normalisasi, (x) nilai atribut asli, (x_{min}) nilai minimum atribut, dan (x_{max}) nilai maksimum atribut.

2.3.2 Seleksi Fitur dengan Mutual Information

Seleksi fitur dilakukan menggunakan Mutual Information (MI) karena mampu mengukur hubungan linear maupun non-linear antara fitur dan variabel target [20], [21]. Seluruh delapan fitur tetap digunakan karena LightGBM dan XGBoost memiliki mekanisme seleksi fitur internal yang mampu mengidentifikasi fitur paling informatif selama proses pelatihan.

2.3.3 Penanganan Ketidakseimbangan Kelas dengan SMOTE

Ketidakseimbangan kelas ditangani menggunakan *Synthetic Minority Oversampling Technique* (SMOTE), yaitu metode yang membangkitkan sampel sintetis pada kelas minoritas berdasarkan tetangga terdekatnya [22], [23], [24]. Persamaan SMOTE ditunjukkan pada Persamaan berikut:

$$x_{syn} = x_i + \lambda \times (x_{nn} - x_i) \quad (2)$$

dimana (x_{syn})sebagai sampel sintetis, (x_i) sampel kelas minoritas, (x_{nn}) tetangga terdekat, dan (λ) bilangan acak pada rentang [0,1]. Setelah penerapan SMOTE, setiap kelas memiliki 500 titik data, sehingga menghasilkan distribusi kelas yang kurang seimbang.

2.3.4 Pembagian Data

Dataset yang telah diproses dipisahkan dengan menggunakan pembagian bertingkat menjadi data pelatihan dan data uji dengan perbandingan 70:30 untuk mempertahankan proporsi kelas di kedua kumpulan data tersebut..

2.4 Pelatihan Model

Penelitian ini menggunakan dua algoritma berbasis gradient boosting, yaitu LightGBM dan XGBoost. Gradient Boosting adalah teknik ensembel yang secara bertahap membangun model melalui serangkaian pohon keputusan untuk

meminimalkan kesalahan prediksi [25], [26]. Dalam penelitian ini digunakan LightGBM dan XGBoost yang memiliki karakteristik berbeda. LightGBM menerapkan strategi leaf-wise growth sehingga lebih efisien dalam penggunaan memori dan waktu komputasi, sedangkan XGBoost menggunakan mekanisme regularisasi untuk mengurangi overfitting dan meningkatkan kemampuan generalisasi model [27], [28].

2.5 Evaluasi Model

Kinerja model dievaluasi menggunakan confusion matrix yang terdiri atas True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Berdasarkan nilai tersebut dihitung metrik Accuracy, Precision, Recall, F1-Score, dan ROC-AUC menggunakan persamaan berikut:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{3}$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{4}$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \tag{5}$$

$$\text{F1-Score} = \frac{2\ x\ (Precision\ x\ Recall)}{(Precision+Recall)} \tag{6}$$

$$\text{AUC-ROC} = \int_0^1 TPR\ d(FPR) \tag{7}$$

Akurasi mengukur ketepatan prediksi model, precision mengukur ketepatan prediksi kelas positif, recall mengukur kemampuan model mendeteksi kasus positif, sedangkan F1-Score memberikan keseimbangan antara precision dan recall. ROC-AUC digunakan untuk mengevaluasi kemampuan diskriminasi model secara keseluruhan. Hasil evaluasi LightGBM dan XGBoost kemudian dibandingkan dengan hasil KNN dan SVM pada penelitian sebelumnya sebagai baseline perbandingan.

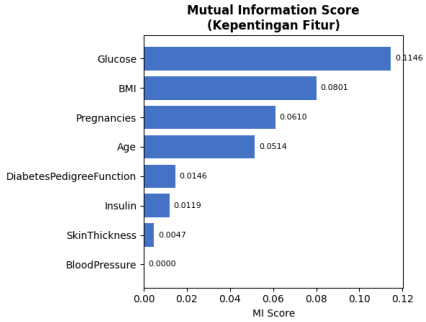
3. HASIL DAN PEMBAHASAN

3.1 Analisis Karakteristik Dataset

Pemeriksaan kualitas data menunjukkan bahwa dataset Pima Indians Diabetes tidak mengandung data duplikat maupun nilai hilang (*missing value*), sehingga dataset dinyatakan bersih dan siap digunakan tanpa memerlukan proses imputasi. Dataset terdiri atas 768 sampel dengan delapan atribut prediktor, yaitu Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, dan Age, serta satu variabel target biner (Outcome). Distribusi kelas menunjukkan ketidakseimbangan yang signifikan, yakni 500 data kelas 0 (non-diabetes) sebesar 65,1% dan 268 data kelas 1 (diabetes) sebesar 34,9%, dengan rasio ketidakseimbangan 1,87:1. Kondisi ini berpotensi menyebabkan model *machine learning* cenderung bias terhadap kelas mayoritas sehingga kemampuan deteksi kelas diabetes yang merupakan kelas minoritas dapat terganggu.

Analisis korelasi Pearson menunjukkan bahwa fitur Glucose memiliki korelasi tertinggi terhadap label Outcome sebesar 0,4666, diikuti oleh BMI (0,2927), Age (0,2384), dan Pregnancies (0,2219), sedangkan BloodPressure memiliki korelasi terendah sebesar 0,0651. Nilai korelasi positif pada keempat fitur dominan tersebut mengindikasikan bahwa individu dengan kadar glukosa lebih tinggi, indeks massa tubuh lebih besar, usia lebih tua, serta riwayat kehamilan lebih banyak cenderung memiliki risiko diabetes yang lebih tinggi. Temuan ini sejalan dengan karakteristik klinis penyakit diabetes yang ditandai oleh tingginya kadar glukosa darah sebagai penanda utama hiperglikemia kronis.

Selanjutnya, dilakukan seleksi fitur menggunakan metode *Mutual Information* untuk mengukur ketergantungan statistik antara setiap fitur dan variabel target, mencakup hubungan linier maupun non-linier. Hasil analisis menunjukkan bahwa fitur Glucose memperoleh skor *Mutual Information* tertinggi sebesar 0,1146, diikuti BMI (0,0801), Pregnancies (0,0610), dan Age (0,0514), sementara BloodPressure memperoleh skor terendah sebesar 0,0000. Distribusi skor *Mutual Information* secara lengkap disajikan pada Gambar 2.

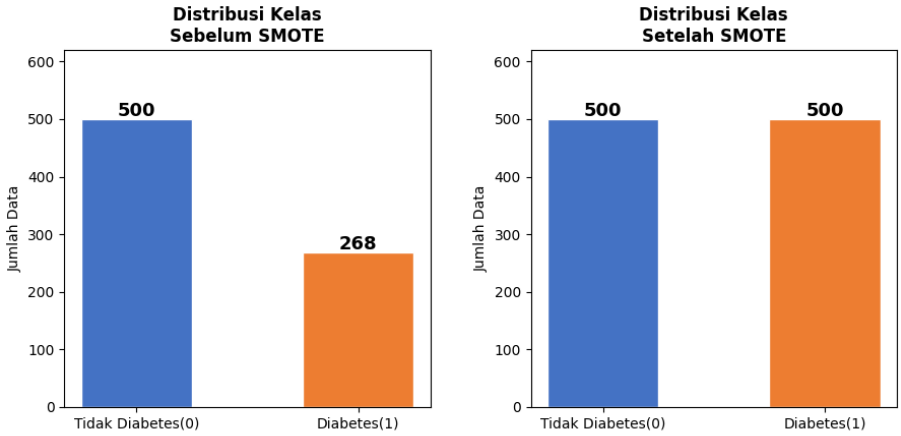


Gambar.2 Distribusi skor *mutual information*

Gambar 2 memperlihatkan bahwa Glucose memberikan kontribusi informasi terbesar terhadap proses klasifikasi diabetes, yang mengonfirmasi bahwa kadar glukosa darah merupakan faktor paling determinan dalam membedakan pasien diabetes dan non-diabetes. Pola konsistensi antara hasil korelasi Pearson dan skor *Mutual Information* memperkuat validitas kedua metode dalam mengidentifikasi fitur-fitur berpengaruh. Meskipun sejumlah fitur memiliki tingkat kontribusi yang relatif rendah, seluruh fitur dipertahankan dalam proses pemodelan karena algoritma LightGBM dan XGBoost memiliki mekanisme seleksi fitur internal yang mampu mengevaluasi kepentingan setiap atribut secara otomatis selama proses pelatihan model, sehingga fitur dengan kontribusi rendah tidak akan mendominasi keputusan klasifikasi.

3.2 Pengaruh Praproses Data dan SMOTE

Tahap praproses data diawali dengan normalisasi menggunakan *MinMax Scaler* yang berhasil mentransformasi seluruh nilai atribut ke rentang [0, 1], sehingga perbedaan skala antar fitur tidak memengaruhi proses pelatihan model. Selanjutnya, penyeimbangan kelas dilakukan menggunakan metode SMOTE. Distribusi kelas berhasil diseimbangkan dengan menggunakan SMOTE dari 500:268 menjadi 500:500, sehingga total sampel meningkat dari 768 menjadi 1.000 sampel. Pembagian data dengan rasio 70:30 menggunakan *stratified split* menghasilkan 700 sampel data latih (350:350) dan 300 sampel data uji (150:150) dengan proporsi kelas yang tetap seimbang.



Gambar 3. Visualisasi distribusi kelas sebelum dan sesudah penerapan SMOTE

Gambar 3 mengilustrasikan perubahan distribusi kelas secara visual sebelum dan sesudah penerapan SMOTE. Sebelum *oversampling*, dataset menunjukkan ketidakseimbangan yang nyata antara kelas non-diabetes dan kelas diabetes. Setelah penerapan SMOTE, Ukuran sampel kelompok minoritas berhasil ditingkatkan hingga setara dengan ukuran sampel kelompok mayoritas. Diharapkan bahwa distribusi yang seimbang ini akan mengurangi bias model yang menguntungkan kelompok mayoritas dan meningkatkan sensitivitasnya dalam mengidentifikasi kasus diabetes. Berbeda dengan teknik duplikasi sederhana, SMOTE menghasilkan sampel sintetis melalui interpolasi berbasis *k-nearest neighbor* sehingga keragaman data latih tetap terjaga dan risiko *overfitting* dapat diminimalkan

3.3 Evaluasi Kinerja Model LightGBM dan XGBoost

Pengujian awal kedua algoritma dilakukan menggunakan parameter bawaan (*default*) pada data yang belum diseimbangkan. Hasil pengujian menunjukkan kinerja yang masih terbatas, terutama pada metrik *recall*, sebagaimana disajikan pada Tabel 2. Pada kondisi ini, LightGBM dan XGBoost cenderung bias terhadap kelas mayoritas (non-diabetes), sehingga sebagian besar kasus diabetes belum berhasil teridentifikasi oleh model.

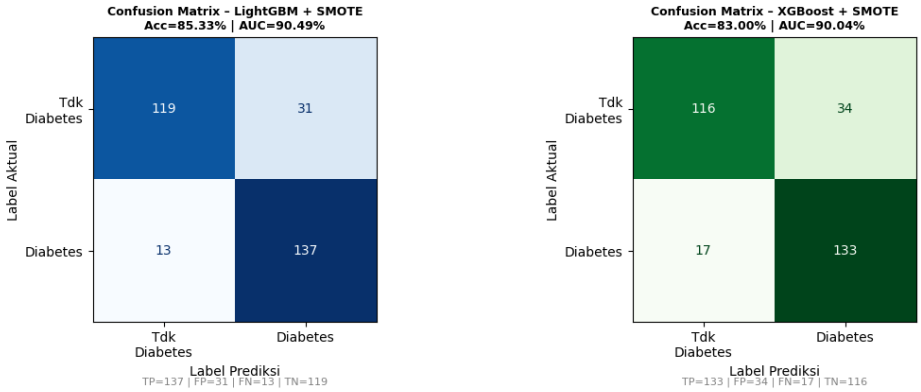
Tabel 2. Kinerja Model Baseline Sebelum dan Sesudah SMOTE

Model	Accuracy	Precision	Recall	F1-Score	AUC
LightGBM	73,59%	64,29%	55,56%	59,60%	80,27%
XGBoost	75,32%	67,65%	56,79%	61,74%	79,06%
LightGBM + SMOTE	85,33%	81,55%	91,33%	86,16%	90,49%
XGBoost + SMOTE	83,00%	79,64%	88,67%	83,91%	90,04%

Tabel 2 menunjukkan bahwa sebelum penerapan SMOTE, LightGBM hanya menghasilkan nilai *recall* sebesar 55,56% dan XGBoost sebesar 56,79%, yang berarti hampir setengah dari keseluruhan kasus diabetes gagal terdeteksi oleh model. Kondisi ini secara langsung merupakan konsekuensi dari Model ini mengoptimalkan perkiraan untuk kelas mayoritas akibat ketidakseimbangan kelas. Setelah menerapkan SMOTE, seluruh metrik evaluasi mengalami peningkatan yang signifikan pada kedua model. *Recall* LightGBM meningkat drastis menjadi 91,33% dan *recall* XGBoost menjadi 88,67%, yang menunjukkan bahwa SMOTE berhasil memberikan representasi data yang lebih seimbang selama proses pelatihan sehingga model memperoleh kemampuan yang jauh lebih baik dalam mengidentifikasi pasien diabetes sebagai

kelas minoritas. Dalam konteks klinis, peningkatan *recall* ini memiliki signifikansi yang sangat tinggi karena secara langsung mengurangi jumlah kasus *false negative*, yakni pasien diabetes yang terlewat tanpa terdeteksi oleh sistem.

Perbandingan langsung antara kedua algoritma setelah penerapan SMOTE menunjukkan bahwa LightGBM + SMOTE secara konsisten unggul pada seluruh metrik evaluasi. LightGBM + SMOTE mencapai *accuracy* 85,33%, *precision* 81,55%, *recall* 91,33%, *F1-Score* 86,16%, dan *ROC-AUC* 90,49%, sementara XGBoost + SMOTE menghasilkan *accuracy* 83,00%, *precision* 79,64%, *recall* 88,67%, *F1-Score* 83,91%, dan *ROC-AUC* 90,04%. Keunggulan LightGBM pada *recall* dan *F1-Score* mengindikasikan bahwa strategi pertumbuhan pohon berbasis *leaf-wise* yang dimiliki LightGBM mampu menangkap pola-pola kompleks pada data yang telah diseimbangkan dengan lebih efektif dibandingkan pendekatan *depth-wise* pada XGBoost. Berdasarkan keunggulan yang konsisten pada seluruh metrik, LightGBM + SMOTE ditetapkan menjadi teknik final yang dibuat dalam studi ini.

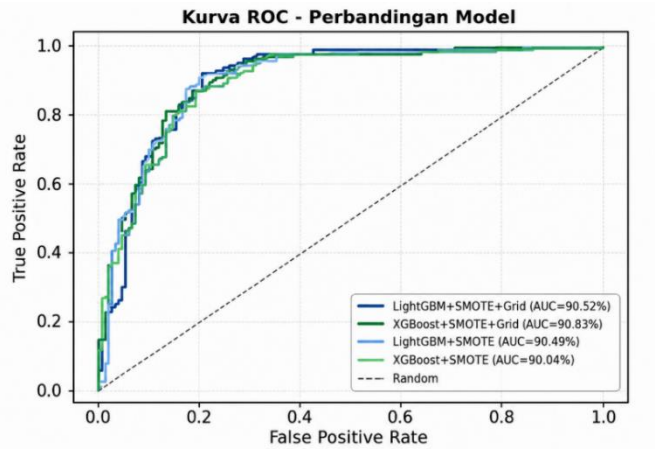


Gambar 4. Confusion matrix model LightGBM + SMOTE dan XGBoost + SMOTE

Gambar 4 menyajikan *confusion matrix* kedua model setelah penerapan SMOTE. Model LightGBM menghasilkan 119 TN, 137 TP, 31 FP, dan 13 FN, sedangkan XGBoost menghasilkan 116 TN, 133 TP, 34 FP, dan 17 FN. Nilai FN yang lebih rendah pada LightGBM (13 dibandingkan 17 pada XGBoost) menegaskan keunggulan LightGBM dalam meminimalkan kasus diabetes yang tidak terdeteksi, yang merupakan aspek paling kritis dalam sistem deteksi dini penyakit diabetes dari perspektif klinis.

3.4 Analisis ROC Curve dan Validasi Model

Studi ini juga mengevaluasi kemampuan model dalam melakukan diskriminasi dengan menggunakan metrik akurasi, presisi, recall, dan F1-score. menggunakan *Receiver Operating Characteristic* (ROC) Curve dan *Area Under the Curve* (AUC). ROC Curve menggambarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai *threshold* klasifikasi, sehingga memberikan gambaran yang lebih komprehensif tentang performa model dibandingkan evaluasi pada satu titik *threshold* tunggal. Semakin dekat kurva ROC ke sudut kiri atas dan semakin besar nilai AUC, maka semakin baik kemampuan model dalam membedakan kelas diabetes dan non-diabetes.



Gambar 5. ROC Curve model LightGBM dan XGBoost

Gambar 5 menunjukkan bahwa seluruh kurva ROC berada di atas garis diagonal yang merepresentasikan klasifikasi acak (*random classifier*), yang mengonfirmasi bahwa kedua model memiliki kemampuan diskriminasi yang baik. Model LightGBM dengan SMOTE menghasilkan nilai AUC tertinggi sebesar 90,49%, diikuti XGBoost dengan SMOTE sebesar 90,04%. Sebagai perbandingan, LightGBM tanpa SMOTE hanya memperoleh AUC sebesar 80,27% dan

XGBoost tanpa SMOTE sebesar 79,06%. Peningkatan nilai AUC yang signifikan setelah penerapan SMOTE menunjukkan bahwa penyeimbangan distribusi kelas secara substantif meningkatkan keandalan model untuk mengidentifikasi perbedaan pasien diabetes dengan kelas non-diabetes pada tiap *threshold* klasifikasi. Nilai AUC di atas 90% pada kedua model setelah SMOTE mengindikasikan tingkat kemampuan diskriminasi yang sangat baik (*excellent discrimination*) menurut kriteria umum dalam evaluasi model diagnostik.

Untuk mengevaluasi konsistensi dan kemampuan generalisasi model, digunakan skema validasi 10-*Fold Cross Validation* sebagai metode pengujian. Hasil yang didapatkan bahwa LightGBM memperoleh CV *Accuracy* sebesar 80,40% ± 3,50% dan CV *F1-Score* sebesar 81,11% ± 3,16%, sedangkan XGBoost memperoleh CV *Accuracy* sebesar 81,30% ± 3,72% dan CV *F1-Score* sebesar 82,03% ± 3,57%. Nilai deviasi standar yang relatif rendah pada kedua model mengindikasikan performa yang stabil dan konsisten pada berbagai pembagian data selama proses validasi. Hal ini menunjukkan bahwa kedua model tidak mengalami *overfitting* yang signifikan dan memiliki kemampuan generalisasi yang memadai terhadap data yang belum pernah dilihat sebelumnya. Meskipun hasil 10-*Fold Cross Validation* menunjukkan XGBoost memperoleh nilai CV yang sedikit lebih tinggi, perbedaan tersebut tidak substansial dan LightGBM tetap unggul pada evaluasi data uji independen yang lebih mencerminkan performa model pada kondisi nyata.

3.5 Perbandingan dengan Penelitian Sebelumnya dan Implikasi Penelitian

Guna menempatkan kontribusi penelitian ini dalam konteks literatur yang lebih luas, performa model yang diusulkan dibandingkan dengan hasil penelitian Mulyani et al. yang menggunakan algoritma KNN dan SVM dengan teknik SMOTE pada dataset yang sama. Perbandingan komprehensif disajikan pada Tabel 4.

Tabel 4. Perbandingan Performa Model dengan Penelitian Sebelumnya

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
KNN k=3	83,33%	85,00%	83,95%	84,47%	–
SVM RBF	81,67%	85,91%	79,01%	82,32%	–
LightGBM + SMOTE	85,33%	81,55%	91,33%	86,16%	90,49%
XGBoost + SMOTE	83,00%	79,64%	88,67%	83,91%	90,04%

Berdasarkan Tabel 4, model LightGBM + SMOTE yang diusulkan dalam penelitian ini menunjukkan performa yang lebih unggul dibandingkan metode KNN dan SVM dari penelitian Mulyani et al. pada hampir seluruh metrik kritis. *Accuracy* LightGBM + SMOTE sebesar 85,33% melampaui KNN (83,33%) dan SVM (81,67%). Peningkatan yang paling signifikan terlihat pada metrik *recall*: LightGBM + SMOTE mencapai 91,33%, meningkat 7,38 poin persentase dibandingkan KNN (83,95%) dan 12,32 poin dibandingkan SVM (79,01%). Peningkatan *F1-Score* dari 84,47% (KNN) menjadi 86,16% pada LightGBM + SMOTE juga menunjukkan keseimbangan yang lebih baik antara *precision* dan *recall*. XGBoost + SMOTE pun menunjukkan hasil yang kompetitif dengan *recall* 88,67% dan *F1-Score* 83,91% yang masing-masing melampaui SVM pada kedua metrik tersebut. Perlu dicatat bahwa nilai *precision* KNN (85,00%) dan SVM (85,91%) lebih tinggi dibandingkan kedua model yang diusulkan; hal ini merupakan konsekuensi yang dapat diterima mengingat peningkatan *recall* yang substansial jauh lebih bernilai dalam konteks deteksi dini diabetes dari sisi klinis.

Keunggulan performa yang diperoleh dapat dikaitkan dengan dua faktor utama. Pertama, algoritma *gradient boosting* seperti LightGBM dan XGBoost pada dasarnya memiliki kapasitas representasi yang lebih tinggi dibandingkan KNN dan SVM, karena mampu menangkap interaksi non-linier antar fitur secara hierarkis melalui ensemble pohon keputusan. Kedua, mekanisme *leaf-wise growth* pada LightGBM memungkinkan pembentukan pohon yang lebih dalam dan terspesialisasi pada pola-pola data yang kompleks, sehingga menghasilkan *recall* yang lebih tinggi pada kelas diabetes setelah distribusi kelas diseimbangkan oleh SMOTE.

Dari sudut pandang implikasi praktis, hasil penelitian ini memiliki relevansi yang tinggi sebagai fondasi pembangunan sistem bantu keputusan klinis (*clinical decision support system*) guna keperluan skrining diabetes sejak dini. Nilai *recall* sebesar 91,33% yang dicapai LightGBM + SMOTE berarti bahwa model hanya melewatkan kurang dari 9% dari seluruh kasus diabetes pada data uji, sebuah tingkat sensitivitas yang sangat memadai untuk aplikasi skrining populasi. Nilai *ROC-AUC* di atas 90% pada kedua model juga memberikan fleksibilitas dalam penyesuaian *threshold* klasifikasi sesuai konteks penggunaan, misalnya menurunkan *threshold* untuk memaksimalkan sensitivitas pada program skrining massal, atau meningkatkan *threshold* untuk mempertahankan *precision* pada konteks yang memerlukan konfirmasi diagnosis yang lebih ketat. Dengan demikian, kombinasi LightGBM dan SMOTE berpotensi diimplementasikan sebagai komponen *machine learning* pada sistem deteksi dini diabetes berbasis data klinis, yang dapat membantu tenaga medis dalam memprioritaskan pasien yang memerlukan pemeriksaan lanjutan secara lebih efisien dan akurat.

4. KESIMPULAN

Upaya pengembangan dan pengujian model klasifikasi diabetes dalam penelitian ini dilakukan melalui algoritma LightGBM dan XGBoost yang dipadukan dengan Synthetic Minority Oversampling Technique (SMOTE) untuk mengatasi ketidakseimbangan kelas pada Pima Indians Diabetes Dataset, dengan Mutual Information sebagai metode seleksi fitur dan ROC-AUC sebagai metrik evaluasi tambahan. Hasil penelitian menunjukkan bahwa ketidakseimbangan kelas menjadi salah satu faktor penyebab rendahnya recall kelas positif pada kondisi data asli, sedangkan penerapan

SMOTE mampu meningkatkan kinerja kedua model. LightGBM + SMOTE menunjukkan kinerja terbaik dengan accuracy sebesar 85,33%, precision 81,55%, recall 91,33%, F1-Score 86,16%, dan ROC-AUC 90,49%, sedangkan XGBoost + SMOTE memperoleh accuracy 83,00%, precision 79,64%, recall 88,67%, F1-Score 83,91%, dan ROC-AUC 90,04%. Hasil tersebut menunjukkan bahwa LightGBM + SMOTE lebih efektif dalam mendeteksi kasus diabetes dibandingkan XGBoost + SMOTE, terutama berdasarkan kemampuan recall yang lebih tinggi. Dibandingkan penelitian Mulyani et al. (2025) yang menggunakan KNN dan SVM, LightGBM + SMOTE menghasilkan peningkatan recall sebesar 7,38 poin persentase dibandingkan KNN terbaik (91,33% vs. 83,95%) serta peningkatan F1-Score. Dengan demikian, kombinasi LightGBM dan SMOTE berpotensi digunakan sebagai pendekatan machine learning untuk mendukung skrining awal diabetes, meskipun penerapannya pada lingkungan klinis tetap memerlukan validasi lebih lanjut menggunakan data eksternal dan data pasien yang lebih beragam.

REFERENCES

[1] H. Guo, H. Wu, and Z. Li, "The Pathogenesis of Diabetes," *Int. J. Mol. Sci.*, vol. 24, no. 8, pp. 0–23, 2023, doi: 10.3390/ijms24086978.
[2] J. Harreiter and M. Roden, "Diabetes mellitus: definition, classification, diagnosis, screening and prevention (Update 2023)," *Wien. Klin. Wochenschr.*, vol. 135, pp. 7–17, 2023, doi: 10.1007/s00508-022-02122-y.
[3] International Diabetes Federation, "IDF Diabetes Atlas," Brussels, 2025. [Online]. Available: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/
[4] H. Guan *et al.*, "Advances in secondary prevention mechanisms of macrovascular complications in type 2 diabetes mellitus patients: a comprehensive review," *Eur. J. Med. Res.*, vol. 29, no. 1, pp. 1–34, 2024, doi: 10.1186/s40001-024-01739-1.
[5] X. An *et al.*, "Early effective intervention can significantly reduce all-cause mortality in prediabetic patients: a systematic review and meta-analysis based on high-quality clinical studies," *Front. Endocrinol. (Lausanne).*, vol. 15, no. March, 2024, doi: 10.3389/fendo.2024.1294819.
[6] K. Sękowski, J. Grudziąż-Sękowska, J. Pinkas, and M. Jankowski, "Public knowledge and awareness of diabetes mellitus, its risk factors, complications, and prevention methods among adults in Poland—A 2022 nationwide cross-sectional survey," *Frontiers in Public Health*, vol. 10, 2022, Art. no. 1029358, doi: 10.3389/fpubh.2022.1029358
[7] K. Dulyapach, P. Ngamchaliew, P. Vichitkunakorn, P. Sornsenee, and K. Choomalee, "Prevalence and Associated Factors of Delayed Diagnosis of Type 2 Diabetes Mellitus in a Tertiary Hospital: A Retrospective Cohort Study," *Int. J. Public Health*, vol. 67, no. November, pp. 1–9, 2022, doi: 10.3389/ijph.2022.1605039.
[8] A. Mulyani, S. Khoerunisa, and D. Kurniadi, "Comparison of KNN and SVM Algorithms Performance Using SMOTE to Classify Diabetes," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 14, no. 1, pp. 25–34, 2025, doi: 10.22146/jnteti.v14i1.15198.
[9] W. H. G. Cheng *et al.*, "Non-Laboratory-Based Risk Prediction Tools for Undiagnosed Pre-Diabetes: A Systematic Review," *Diagnostics*, vol. 13, no. 7, 2023, doi: 10.3390/diagnostics13071294.
[10] F. Mohsen, H. R. H. Al-Absi, N. A. Yousri, N. El Hajj, and Z. Shah, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," *npj Digit. Med.*, vol. 6, no. 1, pp. 1–15, 2023, doi: 10.1038/s41746-023-00933-5.
[11] S. G. Choi *et al.*, "Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-023-40170-0.
[12] Z. Rozikin and J. J. Hidayat, "Perbandingan Metode Oversampling SMOTE dan ADASYN pada Klasifikasi Diabetes Menggunakan Algoritma CatBoost," *J. Manaj. Inform. Teknol.*, vol. 6, no. 1, pp. 151–164, 2026, doi: 10.51903/mifortekh.v6i1.1157.
[13] I. Abousaber, H. F. Abdallah, and H. El-Ghaish, "Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets," *Front. Artif. Intell.*, vol. 7, 2024, doi: 10.3389/frai.2024.1499530.
[14] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: a Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *Int. J. Informatics Vis.*, vol. 8, no. 3, pp. 1310–1318, 2024, doi: 10.62527/joiv.8.3.2283.
[15] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: https://doi.org/10.1007/s10994-022-06296-4.
[16] Y. Xie and S. P. Tummala, "Machine Learning for Sensor Analytics: A Comprehensive Review and Benchmark of Boosting Algorithms in Healthcare, Environmental, and Energy Applications," *Sensors*, vol. 25, no. 23, 2025, doi: 10.3390/s25237294.
[17] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Front. Genet.*, vol. 14, no. October, pp. 1–15, 2023, doi: 10.3389/fgene.2023.1252159.
[18] S. Masood, M. A. Al-Bashrawi, M. A. Khan, and Y. K. Dwivedi, "From data to diagnosis: a systematic review on AI-driven approaches to diabetes prediction," *Artif. Intell. Rev.*, vol. 59, no. 3, 2026, doi: 10.1007/s10462-025-11485-3.
[19] H. M. Jumasa, "Model Predictive Analytics Terhadap Pasien Diabetes Menggunakan Exploratory Data Analysis dan Algoritma Random Forest," *INTEK: J. Inform. dan Teknol. Inform.*, vol. 6, no. 2, pp. 44–50, Nov. 2023, doi: 10.37729/intek.v6i2.3867.
[20] K. Li and N. Fard, "A Novel Nonparametric Feature Selection Approach Based on Mutual Information Transfer Network," *Entropy*, vol. 24, no. 9, 2022, doi: 10.3390/e24091255.
[21] A. K. Tiwari, R. Saini, A. Nath, P. Singh, and M. A. Shah, "Hybrid similarity relation based mutual information for feature selection in intuitionistic fuzzy rough framework and its applications," *Sci. Rep.*, pp. 1–21, 2024, doi: 10.1038/s41598-024-55902-z.
[22] Y. Chen, J. Zou, L. Liu, and C. Hu, "Improved Oversampling Algorithm for Imbalanced Data Based on K-Nearest Neighbor and Interpolation Process Optimization," *Symmetry*, vol. 16, no. 3, Art. no. 273, 2024, doi: 10.3390/sym16030273..
[23] N. Altwaijry, "Probability-Based Synthetic Minority Oversampling Technique," *IEEE Access*, vol. 11, pp. 28831–28839, 2023, doi: 10.1109/ACCESS.2023.3260723.
[24] SAS Institute Inc., *SAS Data Surveyor for User ' s Guide*. 2026.

[25] N. Gunasekara, B. Pfahringer, H. Gomes, and A. Bifet, "Gradient boosted trees for evolving data streams," *Mach. Learn.*, vol. 113, no. 5, pp. 3325–3352, 2024, doi: 10.1007/s10994-024-06517-y.
[26] D. Boldini, F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber, "Practical guidelines for the use of gradient boosting for molecular property prediction," *J. Cheminform.*, pp. 1–13, 2023, doi: 10.1186/s13321-023-00743-7.
[27] F. E. P. Nadya, M. F. I. Ferdiansyah, V. R. S. Nastiti, and C. S. K. Aditya, "Implementation of Feature Selection Strategies to Enhance Classification Using XGBoost and Decision Tree," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 187–194, 2024, doi: 10.15294/sji.v11i1.48145.
[28] P. B. Khokhar, C. Gravino, and F. Palomba, "Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review," *Artif. Intell. Med.*, vol. 164, no. September 2024, p. 103132, 2025, doi: 10.1016/j.artmed.2025.103132.