Penerapan Algoritma Catboost Dalam Prediksi Dini Diabetes Melitus Berbasis Integrasi Faktor Risiko Klinis Dan Gaya Hidup

Fatia Amalia Maresti*, Resmi Maharani DwiPutri Nugraha, Nawang Eka Arsaningtyas, Kiki Mustaqim
Universitas Logistik dan Bisnis Internasional (ULBI)

Received: Accepted: Published:

*resmiputrinugraha10@gmail.com

Kata Kunci: Diabetes Melitus, CatBoost, Prediksi Dini, Gaya Hidup

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak : Diabetes Melitus (DM) merupakan ancaman kesehatan global dengan perkembangan laten yang memerlukan deteksi dini melalui integrasi parameter klinis dan gaya hidup. Penelitian ini bertujuan mengimplementasikan algoritma CatBoost (Categorical Boosting) untuk memprediksi risiko DM pada populasi usia produktif (20–35 tahun) menggunakan 12 parameter kesehatan heterogen. Menggunakan dataset sekunder hasil *purposive sampling* sebanyak 418 catatan medis bersih dengan rasio *split* 80:20, model dievaluasi secara komparatif. Hasil pengujian menunjukkan algoritma CatBoost mampu menangani data kategorikal secara langsung dan dirancang untuk meredam *overfitting*, menghasilkan akurasi global sebesar 85,71%, skor AUC-ROC sebesar 0,90, serta nilai presisi dan sensitivitas (*recall*) stabil di rentang 0,83–0,88. Analisis kepentingan fitur (*Feature Importance*) menempatkan Gula Darah Puasa (38%) dan HbA1c (31%) sebagai prediktor paling dominan. Hasil ini bersifat awal karena hanya berdasarkan satu dataset sekunder dan perlu divalidasi lebih lanjut pada data klinis nyata.

Pendahuluan

Di abad ke-21 diabetes mellitus (DM) telah berkembang menjadi salah satu masalah kesehatan yang paling signifikan dan berkembang pesat. Penyakit ini adalah gangguan metabolik kronis yang ditandai dengan hiperglikemia dan dapat menyebabkan kerusakan jangka panjang pada berbagai organ tubuh jika tidak ditangani dengan benar. Data penelitian yang dikumpulkan oleh International Diabetes Federation (IDF) menunjukkan bahwa populasi penderita diabetes di seluruh dunia diperkirakan akan meningkat drastis dari 537 juta kasus pada tahun 2021 hingga 783 juta jiwa pada tahun 2045 (estimasi IDF edisi ke-10). Dengan sekitar 19,5 juta orang yang menderita diabetes dan angka mortalitas lebih dari 236.000 kematian per tahun, Indonesia saat ini menempati posisi kelima di dunia. Lebih dari 33% dana yang dialokasikan untuk pembiayaan penyakit tidak menular dialokasikan ke sistem jaminan kesehatan nasional karena fenomena ini, yang mengakibatkan penurunan ekspektasi hidup dan produktivitas penderitanya. Estimasi terbaru IDF edisi ke-11 menunjukkan bahwa 589 juta orang dewasa berusia 20-79 tahun hidup dengan diabetes pada tahun 2024, jumlah tersebut diproyeksikan meningkat menjadi 853 juta pada tahun 2050, dan sekitar 252 juta di antaranya belum terdiagnosis[1].

Tingginya mortalitas akibat diabetes utamanya disebabkan oleh sifat penyakit yang berkembang secara laten, di mana banyak individu tidak menyadari kondisi medis mereka hingga munculnya komplikasi stadium lanjut[2]. Oleh karena itu, strategi prediksi dini menjadi instrumen krusial dalam memitigasi risiko sebelum memasuki fase klinis yang ireversibel (tidak dapat disembuhkan). Secara patofisiologi, evaluasi risiko tidak dapat hanya bersandar pada parameter klinis laboratorium saja (Fasting Blood Sugar, HbA1c), melainkan harus diintegrasikan dengan aspek antropometri fisik (BMI, Age, Gender), riwayat medis (Family History Diabetes, Gestational Diabetes), gejala awal tubuh (Excessive Thirst), serta determinan perilaku atau gaya hidup harian pasien (Diet Quality, Sleep Quality, Physical Activity).

Tantangannya, pendekatan skrining konvensional linier seringkali gagal memetakan korelasi non-linier tersembunyi antara data laboratorium yang bersifat eksak dengan data kebiasaan gaya hidup yang cenderung bersifat subjektif dan kategorikal[3]. Keterbatasan ini ditegaskan dalam sebuah studi ulasan komprehensif oleh Yousefi dkk. (2026) menunjukkan bahwa mayoritas alat skrining tradisional di fasilitas kesehatan primer masih sangat bergantung pada variabel antropometri sederhana serta memiliki sensitivitas yang terbatas jika tidak diintegrasikan dengan platform digital yang mampu mengolah pola data perilaku secara cerdas.

Guna menjembatani kompleksitas tersebut, pemanfaatan teknik komputasi berbasis machine learning hadir sebagai solusi mutakhir untuk menghasilkan pemodelan prediksi dengan presisi tinggi. Hal ini relevan karena data kesehatan terintegrasi umumnya bersifat heterogen, yaitu mencampurkan tipe data numerik kontinu (*Age*, *Fasting Blood Sugar*, *HbA1c*, *BMI*) dengan banyak variabel kategorikal (*Gender*, *Diet Quality*, *Sleep Quality*, *Physical Activity*, *Family History*, *Excessive Thirst*, *Gestational Diabetes*). Dalam domain *gradient boosting*, algoritma CatBoost (*Categorical Boosting*) menunjukkan efisiensi komputasi yang lebih baik dibandingkan model *ensemble* konvensional lainnya [4]. Keunggulan utama CatBoost terletak pada kemampuannya mengolah *categorical features* secara langsung lewat algoritma *Ordered Target Statistics* tanpa memerlukan pra-pemrosesan eksternal seperti *one-hot encoding* intensif yang kerap kali merusak dimensi data atau memicu terjadinya *overfitting*.

Berdasarkan landasan ilmiah tersebut, terdapat urgensi besar untuk membangun sebuah model prediksi diabetes dini yang komprehensif. Kebaruan (novelty) dari penelitian ini terletak pada cakupan variabelnya yang secara simultan menyatukan 12 parameter esensial, mulai dari biomarker laboratorium (Fasting Blood Sugar, HbA1c), kondisi fisik (BMI, *Age, Gender*), rekam medis akut (Excessive Thirst), hingga tiga pilar gaya hidup utama *(Diet, Quality, Sleep Quality, Physical Activity).*

Tinjauan literatur

Diabetes Melitus(DM) merupakan kelompok gangguan metabolik yang ditandai oleh hiperglikemia akibat gangguan sekresi insulin, kerja insulin, atau keduanya. Pada diabetes tipe 2, kondisi ini umumnya berkaitan dengan resistensi insulin dan penurunan fungsi sel beta pankreas. Deteksi dini penting karena diabetes tipe 2 dapat mengalami fase tanpa gejala yang jelas, sehingga identifikasi individu berisiko melalui penilaian faktor risiko dan pemeriksaan glikemik dapat membantu mendukung intervensi lebih awal[5].

Prediksi diabetes dapat memanfaatkan gabungan faktor demografis, antropometri, klinis, dan gaya hidup. Variabel seperti usia, jenis kelamin, indeks massa tubuh (BMI), riwayat diabetes dalam keluarga, kebiasaan merokok, aktivitas fisik, pola makan, serta indikator glikemik dapat digunakan sebagai prediktor sesuai ketersediaan data dan tujuan penelitian. Integrasi berbagai kelompok variabel memungkinkan model mempertimbangkan informasi yang lebih luas daripada model yang hanya menggunakan satu jenis indikator. Namun, kontribusi setiap variabel terhadap kinerja prediksi perlu diuji pada dataset yang digunakan[6].

Secara patofisiologis, diabetes tipe 2 melibatkan interaksi antara resistensi insulin dan disfungsi sel beta pankreas. Pada resistensi insulin, jaringan tubuh tidak memberikan respons yang memadai terhadap insulin sehingga pengaturan glukosa menjadi terganggu. Pada tahap awal, pankreas dapat meningkatkan sekresi insulin sebagai mekanisme kompensasi. Ketika kemampuan sel beta tidak lagi mencukupi kebutuhan tubuh, kadar glukosa darah dapat meningkat secara persisten. Proses ini juga dipengaruhi oleh interaksi metabolik antarorgan dan berbagai faktor biologis maupun lingkungan[7].

Pemanfaatan machine learning dalam prediksi diabetes memungkinkan pengolahan berbagai variabel secara bersamaan untuk mengidentifikasi pola yang berkaitan dengan risiko penyakit, termasuk variabel klinis, demografis, antropometri, dan gaya hidup. Namun, kemampuan model dalam menghasilkan prediksi tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh karakteristik populasi, kualitas data, pemilihan variabel, dan metode validasi[8]. Oleh karena itu, pengembangan model prediksi diabetes perlu mempertimbangkan kesesuaian algoritma dengan karakteristik data serta mengevaluasi kinerjanya secara empiris [7], [9].

Salah satu pendekatan yang dapat digunakan adalah *Gradient Boosting Decision Tree* (GBDT), yaitu metode ansambel yang membangun sejumlah pohon keputusan secara bertahap untuk memperbaiki

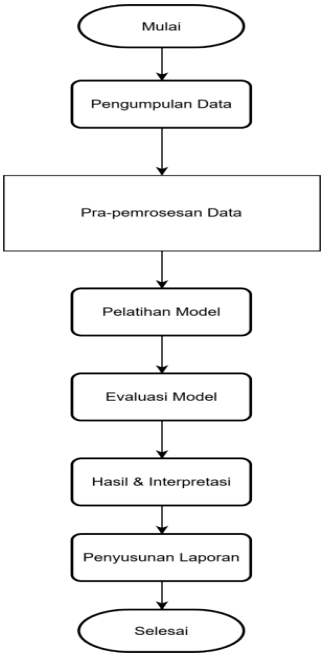
kesalahan prediksi model sebelumnya. Proses pembelajaran dilakukan dengan mengoptimalkan fungsi kerugian *(loss function)*, sehingga model dapat mempelajari hubungan antara variabel prediktor dan target secara iteratif [10]. Pendekatan ini menjadi dasar pengembangan berbagai algoritma gradient boosting, termasuk CatBoost, yang dirancang untuk menangani data tabular dengan kombinasi fitur numerik dan kategorikal.

CatBoost mengembangkan pendekatan gradient boosting melalui mekanisme *Ordered Target Statistics* untuk mengolah fitur kategorikal dan *Ordered Boosting* untuk mengurangi *prediction shift* yang berkaitan dengan kebocoran informasi target dalam proses pelatihan. Kedua mekanisme tersebut diperkenalkan oleh Prokhorenkova dkk. sebagai bagian dari upaya meningkatkan proses pembelajaran model. Dalam prediksi diabetes, CatBoost dapat diterapkan pada data yang menggabungkan variabel numerik, seperti usia, indeks massa tubuh, dan kadar glukosa, dengan variabel kategorikal, seperti jenis kelamin dan riwayat keluarga.

Model prediksi diabetes perlu dinilai menggunakan metrik yang sesuai dengan tujuan penelitian. Untuk klasifikasi, metrik yang dapat dipertimbangkan antara lain *accuracy, precision, recall* atau sensitivitas, *specificity, F1-score*, dan area *under the receiver operating characteristic curve* (ROC-AUC). Pemilihan metrik perlu mempertimbangkan konsekuensi kesalahan klasifikasi, terutama risiko individu yang sebenarnya berisiko tinggi tetapi diklasifikasikan sebagai tidak berisiko. Selain itu, pemisahan data pelatihan dan pengujian serta validasi yang tepat diperlukan untuk menilai kemampuan model menghadapi data baru. Interpretasi kontribusi variabel juga dapat membantu menjelaskan pola yang dipelajari model, tetapi hasil tersebut tidak dengan sendirinya membuktikan hubungan sebab-akibat [11].

Metode Penelitian

Metode penelitian berfungsi sebagai pedoman untuk menyusun langkah-langkah secara sistematis dan terarah dalam pengumpulan data guna mencapai tujuan penelitian. Pada penelitian ini, metode yang digunakan adalah data kuantitatif karena fokus utama adalah pada pengolahan data numerik, analisis statistik, dan penerapan algoritma *machine learning*. Adapun alur pengerjaan penelitian dapat dilihat dari struktur diagram alur penelitian.



Gambar 1. Alur Penelitian

Tahapan penelitian dirancang secara terstruktur untuk memastikan validitas dan keandalan model prediktif yang dibangun. Alur kerja penelitian dimulai dari pengumpulan data klinis, pembersihan data, pra-pemrosesan, pemodelan menggunakan algoritme *CatBoost*, hingga tahap evaluasi performa model. Kerangka kerja penelitian ini direpresentasikan melalui tahapan sistematis sebagai berikut:

1. Pengumpulan dan Validasi Data
   Penelitian ini menggunakan dataset rekam medis kesehatan yang mencakup 418 data bersih dari pasien usia produktif (20–35 tahun), yang terbebas dari nilai kosong (*missing values*) maupun duplikasi data. Dataset ini memuat 12 parameter esensial yang terdiri dari variabel laboratorium, antropometri, riwayat medis, gejala klinis, serta pilar gaya hidup.
2. Pra-Pemrosesan Data (*Data Preprocessing*)
   Sebelum dilatih ke dalam model, data melalui tahap normalisasi dan pemetaan tipe variabel. Karena algoritma *CatBoost* memiliki kemampuan bawaan (*native handling*) dalam mengolah variabel kategorikal, data dapat diproses secara optimal tanpa memerlukan transformasi *one-hot encoding* ekstensif yang berisiko memperbesar dimensi data.
3. Pembagian Dataset (*Data Splitting*)
   Dataset dibagi menjadi dua bagian utama secara proporsional, yaitu data latih (*training set*) sebesar 80% untuk proses pembelajaran pola dan kalibrasi parameter model, serta data uji (*test set*) sebesar 20% untuk menguji kemampuan generalisasi model terhadap data baru yang belum pernah dilihat sebelumnya.
4. Pemodelan Menggunakan *CatBoost*
   Algoritme *CatBoost (Categorical Boosting)* diterapkan untuk melakukan klasifikasi risiko diabetes. Arsitektur ini memanfaatkan teknik *Ordered Target Statistics* dan pohon simetris (*symmetric trees*) guna meredam potensi *overfitting* serta memaksimalkan efisiensi komputasi pada data heterogen.
5. Evaluasi Performa Model
   Tahap akhir dari metodologi ini adalah pengukuran kinerja model menggunakan metrik evaluasi global yang meliputi Akurasi, Presisi, *Recall*, nilai AUC-ROC dari kurva *Receiver Operating Characteristic*, serta analisis tingkat kepentingan fitur (*Feature Importance*) untuk mengidentifikasi variabel yang paling dominan mempengaruhi risiko diabetes.

Hasil Penelitian

Penelitian menggunakan dataset Diabetes Health Dataset Analysis yang diperoleh dari Kaggle. Dataset awal terdiri atas 1.879 data, kemudian dilakukan penyaringan berdasarkan rentang usia 20–35 tahun sehingga diperoleh 418 data dengan 12 variabel prediktor dan satu variabel target, yaitu Diagnosis. Berdasarkan pemeriksaan data, tidak ditemukan nilai yang hilang maupun baris duplikat. Dataset selanjutnya dibagi menjadi data latih sebanyak 334 data (80%) dan data uji sebanyak 84 data (20%) untuk membangun dan mengevaluasi model CatBoost.

```
EVALUASI MODEL USIA PRODUKTIF 20-35 TAHUN
Jumlah sampel data testing: 84 baris
Akurasi Model: 85.71%

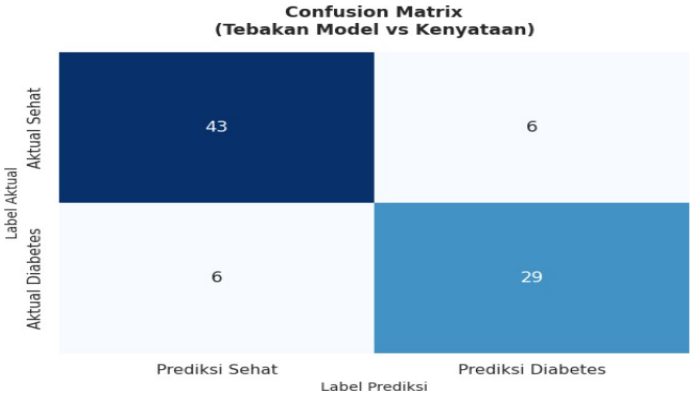
Laporan Klasifikasi Detail:
              precision    recall  f1-score   support

   Sehat (0)       0.88      0.88      0.88        49
Diabetes (1)       0.83      0.83      0.83        35

    accuracy                           0.86        84
   macro avg       0.85      0.85      0.85        84
weighted avg       0.86      0.86      0.86        84
```

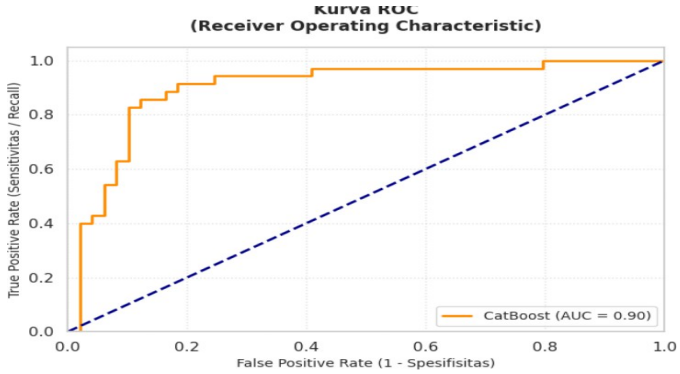
Gambar 2. Hasil Evaluasi Model CatBoost

Berdasarkan hasil pengujian pada 84 data uji, algoritma CatBoost memperoleh akurasi sebesar 85,71%. Nilai tersebut menunjukkan bahwa sebagian besar data uji berhasil diklasifikasikan sesuai dengan label aktualnya. Hasil evaluasi juga menunjukkan *nilai precision, recall,* dan *F1-score* sebesar 0,88 pada kelas non-diabetes serta 0,83 pada kelas diabetes.



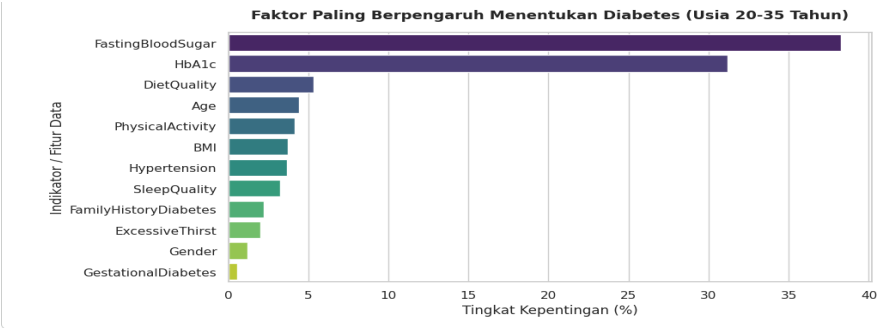
Gambar 3. *Output Confusion Matrix*

Berdasarkan *confusion matrix*, model CatBoost berhasil mengklasifikasikan 43 data non-diabetes dan 29 data diabetes dengan benar. Sementara itu, terdapat enam data non-diabetes yang diprediksi sebagai diabetes serta enam data diabetes yang diprediksi sebagai non-diabetes. Hasil ini menunjukkan bahwa model masih menghasilkan kesalahan klasifikasi pada kedua kelas.



Gambar 4. Output Kurva ROC.

Hasil evaluasi menunjukkan nilai Area Under the Receiver Operating Characteristic Curve (AUC-ROC) sebesar 0,90. Nilai ini menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan kelas diabetes dan non-diabetes berdasarkan skor prediksinya. Kurva ROC melengkapi nilai akurasi karena memperlihatkan hubungan antara true positive rate dan false positive rate pada berbagai ambang klasifikasi.



Gambar 5. *Output Feature Importance*

Hasil analisis *feature importance* menunjukkan bahwa *Fasting Blood Sugar* memiliki tingkat kepentingan tertinggi, yaitu sekitar 38%, diikuti oleh HbA1c sebesar 31%. Selanjutnya, *Diet Quality* berkontribusi sekitar 5%, sedangkan fitur lain, seperti usia, aktivitas fisik, indeks massa tubuh, dan hipertensi, memiliki tingkat kepentingan yang lebih rendah. Hasil tersebut memperlihatkan bahwa kedua indikator klinis tersebut merupakan fitur yang paling dominan dalam proses prediksi model CatBoost pada dataset

yang digunakan. Meskipun demikian, nilai feature importance menunjukkan kontribusi relatif fitur terhadap prediksi model dan tidak secara langsung membuktikan hubungan sebab-akibat dengan diabetes.

Tabel 1. Perbandingan Model Proyek Diabetes

No	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
4	CatBoost	0.8989	0.9118	0.8267	0.8671	0.9368
3	XGBoost	0.8910	0.8811	0.8400	0.8601	0.9226
2	Random Forest	0.8617	0.8603	0.7800	0.8182	0.9093
1	Decision Tree	0.8218	0.7677	0.7933	0.7803	0.8170
0	Logistic Regression	0.6223	0.5571	0.2600	0.3545	0.5880

Hasil perbandingan lima algoritma machine learning berdasarkan metrik Accuracy, Precision, Recall, F1-Score, dan ROC-AUC menunjukkan bahwa CatBoost memiliki performa terbaik. CatBoost memperoleh akurasi 89,89%, ROC-AUC 0,9368, dan presisi 91,18%. XGBoost menempati posisi kedua dengan akurasi 89,10% dan ROC-AUC 0,9226. Random Forest memperoleh akurasi 86,17% dan ROC-AUC 0,9093, sedangkan Decision Tree mencapai akurasi 82,18% dan ROC-AUC 0,8170. Logistic Regression menunjukkan performa terendah, dengan akurasi 62,23% dan recall 26%. Hasil ini menunjukkan bahwa CatBoost menjadi model dengan performa terbaik di antara algoritma yang dibandingkan untuk skrining awal diabetes pada dataset.

Pembahasan

Interpretasi performa model dalam konteks klinis. Model CatBoost memperoleh akurasi 85,71% dan AUC-ROC 0,90 pada 84 data uji. Berdasarkan *confusion matrix*, sensitivitas model terhadap kelas diabetes adalah 82,86% (29 dari 35 kasus) dan spesifisitasnya 87,76% (43 dari 49 kasus non-diabetes). Dalam konteks klinis, nilai tersebut mengindikasikan bahwa model berpotensi dimanfaatkan sebagai alat bantu skrining awal untuk memprioritaskan individu yang perlu menjalani pemeriksaan glukosa darah konfirmatif, tetapi belum memadai sebagai alat diagnostik mandiri. Pedoman ADA menegaskan bahwa diagnosis diabetes ditegakkan berdasarkan kriteria glikemik dan, pada individu tanpa gejala klasik hiperglikemia, memerlukan pengujian ulang untuk konfirmasi [5]. Selain itu, estimasi performa masih memiliki ketidakpastian yang besar akibat jumlah data uji yang kecil: interval kepercayaan Wilson 95% untuk akurasi berkisar antara 76,7% dan 91,6%, sedangkan untuk sensitivitas berkisar antara 67,3% dan 91,9%.

Faktor prediktor dominan. Dominasi *Fasting Blood Sugar* (38%) dan HbA1c (31%) sejalan dengan kenyataan bahwa kedua parameter tersebut merupakan kriteria diagnostik baku diabetes [5]. Apabila label *Diagnosis* pada dataset ditetapkan berdasarkan kedua parameter tersebut, model berpotensi mempelajari kembali aturan diagnostik sehingga kontribusi variabel gaya hidup, seperti *Diet Quality* yang hanya sekitar 5%, tampak kecil. Sebagai pembanding, [12] yang hanya menggunakan variabel demografis, pemeriksaan, dan gaya hidup tanpa parameter laboratorium glukosa melaporkan bahwa asupan energi, karbohidrat, dan lemak menjadi kontributor terbesar terhadap prediksi diabetes. Hal ini menunjukkan bahwa peran variabel gaya hidup perlu dikaji lebih lanjut, misalnya melalui eksperimen tanpa fitur laboratorium dan dokumentasi yang jelas mengenai penetapan label pada dataset.

Perbandingan dengan penelitian terdahulu. Sabili dkk. [2] melaporkan akurasi CatBoost sebesar 98,63% untuk klasifikasi diabetes berdasarkan faktor risiko setelah penanganan data pencilan dan pembersihan data, yang lebih tinggi daripada hasil penelitian ini. Sumathi dkk. [4] membandingkan CatBoost dengan sejumlah *classifier* lain dan model *ensemble* serta melaporkan keunggulan CatBoost yang mampu memangkas waktu pemprosesan hingga 4,27 detik dengan mempertahankan performa akurasi dan diskriminasi AUC yang optimal. Perbedaan nilai tersebut tidak dapat dibandingkan secara langsung karena dataset, jumlah sampel, rentang usia, variabel, dan skema pembagian data berbeda. Qin dkk. [12] melaporkan akurasi 82,1% dan AUC 0,83 untuk CatBoost pada 17.833 responden NHANES dengan variabel gaya hidup, sedangkan Riveros Perez dan Avella-Molano [13] pada 29.509 orang dewasa menemukan bahwa XGBoost memberikan AUC tertinggi (0,8168) dan sensitivitas sebagian besar model terbatas. Temuan-temuan ini menunjukkan bahwa keunggulan CatBoost bergantung pada karakteristik data, sehingga hasil penelitian ini hanya berlaku pada dataset yang digunakan.

Implikasi klinis false negative. Enam kasus *false negative* (17,1% dari 35 pasien diabetes) berarti sekitar satu dari enam pasien diabetes pada data uji tidak terdeteksi. Pada skrining kesehatan masyarakat, kesalahan ini lebih berisiko daripada *false positive* karena individu yang tidak terdeteksi dapat merasa aman dan terlambat memperoleh intervensi, padahal diabetes berkembang secara laten [2]. Sebaliknya, enam kasus *false positive* umumnya hanya berdampak pada kebutuhan pemeriksaan konfirmatif tambahan. Oleh karena itu, penerapan model di fasilitas kesehatan perlu mempertimbangkan penurunan ambang klasifikasi (*threshold*) di bawah 0,5 untuk meningkatkan sensitivitas, dengan konsekuensi bertambahnya *false positive*, serta menetapkan pemeriksaan konfirmatif bagi seluruh hasil positif.

Penelitian memiliki beberapa keterbatasan. Pertama, dataset hanya terdiri atas 418 data hasil penyaringan usia 20–35 tahun dari 1.879 data awal, bersumber dari data sekunder Kaggle. Kedua, evaluasi hanya menggunakan satu kali pembagian data 80:20 dengan 84 data uji tanpa *cross-validation*, sehingga estimasi kinerja belum stabil. Ketiga, validasi hanya dilakukan pada satu dataset tanpa validasi eksternal pada data dari fasilitas kesehatan primer di Indonesia. Keempat, dominasi parameter laboratorium membatasi kesimpulan mengenai peran faktor gaya hidup. Ravi dan Olickal [6] menyimpulkan bahwa bukti mengenai model prediksi diabetes tipe 2 berbasis *machine learning* di layanan primer masih terbatas, dan Ganie [14] menekankan perlunya validasi prospektif sebelum penerapan klinis. Pelaporan penelitian lanjutan sebaiknya mengacu pada panduan TRIPOD+AI [11]. Dengan demikian, hasil penelitian ini bersifat awal dan belum dapat digeneralisasi ke populasi Indonesia.

Kesimpulan

Penelitian telah mencapai tujuan yang ditetapkan, yaitu mengimplementasikan algoritma CatBoost (*Categorical Boosting*) dalam pengembangan model prediksi dini Diabetes Mellitus berbasis integrasi faktor klinis dan gaya hidup. Berdasarkan hasil pengolahan data, pemodelan, serta evaluasi yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut.

Penelitian berhasil mengintegrasikan 12 parameter kesehatan yang terdiri atas faktor demografis, klinis, riwayat kesehatan, dan gaya hidup ke dalam dataset yang siap digunakan untuk pemodelan. Dataset akhir yang bersumber dari data sekunder Kaggle terdiri atas 418 catatan dengan rentang usia 20–35 tahun. Hasil eksplorasi dan preprocessing data menunjukkan bahwa dataset memiliki kualitas yang baik, tanpa *missing values* maupun data duplikat. Pembagian dataset dilakukan menggunakan rasio 80% data latih sebanyak 334 data dan 20% data uji sebanyak 84 data. Dengan demikian, dataset yang digunakan telah memenuhi kebutuhan pengolahan untuk membangun model klasifikasi CatBoost.

Hasil pengujian menunjukkan bahwa algoritma CatBoost mampu memberikan performa klasifikasi yang baik dalam membedakan individu dengan kategori sehat dan diabetes. Model memperoleh nilai akurasi sebesar 85,71% dengan nilai presisi, *recall*, dan *F1-score* yang berada pada rentang 0,83 hingga 0,88. Pada kelas sehat (0), ketiga metrik tersebut mencapai nilai 0,88, sedangkan pada kelas diabetes (1) mencapai nilai 0,83. Selain itu, nilai AUC-ROC sebesar 0,90 mengindikasikan bahwa model memiliki kemampuan diskriminasi yang baik dalam membedakan kedua kelas berdasarkan dataset yang digunakan. Berdasarkan hasil *Confusion Matrix* terhadap 84 data uji, model berhasil mengklasifikasikan 43 data *True Negative* dan 29 data *True Positive*, dengan masing-masing 6 data *False Positive* dan *False Negative*. Hasil tersebut mengindikasikan potensi CatBoost sebagai model pendukung dalam proses deteksi dini Diabetes Mellitus berdasarkan dataset yang digunakan. Namun, enam kasus false negative (17,1% dari 35 pasien diabetes), jumlah data uji yang kecil, penggunaan satu dataset sekunder, serta tidak adanya validasi silang maupun validasi eksternal menunjukkan bahwa model perlu divalidasi lebih lanjut pada data klinis nyata sebelum diterapkan di fasilitas kesehatan.

Berdasarkan analisis *Feature Importance*, diketahui bahwa parameter klinis memiliki kontribusi dominan terhadap hasil prediksi model. Gula darah puasa (*Fasting Blood Sugar*) menjadi variabel dengan tingkat kepentingan tertinggi sebesar 38%, diikuti oleh HbA1c sebesar 31%. Sementara itu, variabel lainnya, seperti BMI, usia, jenis kelamin, riwayat keluarga diabetes, diabetes gestasional, hipertensi, gejala rasa haus berlebihan (*Excessive Thirst*), kualitas tidur, kualitas diet, dan aktivitas fisik, turut memberikan kontribusi terhadap proses klasifikasi. Temuan ini mengindikasikan bahwa integrasi parameter klinis dan gaya hidup dapat memberikan gambaran yang lebih menyeluruh dalam mengidentifikasi faktor-faktor yang berkaitan

dengan risiko Diabetes Mellitus pada kelompok usia produktif. Namun, dominasi dua parameter klinis menunjukkan bahwa kontribusi variabel gaya hidup pada dataset ini relatif kecil.

Untuk penelitian selanjutnya, pengembangan model CatBoost perlu diarahkan pada pengujian menggunakan data klinis nyata yang diperoleh dari fasilitas pelayanan kesehatan primer di Indonesia, seperti Puskesmas dan klinik, guna mengetahui kemampuan generalisasi model terhadap kondisi populasi yang lebih beragam. Penelitian berikutnya juga perlu menerapkan stratified k-fold cross-validation dan validasi eksternal agar estimasi kinerja model lebih stabil. Penelitian berikutnya juga disarankan untuk memperluas cakupan kelompok usia, sehingga model tidak hanya terbatas pada rentang usia 20–35 tahun, tetapi dapat diuji pada kelompok usia lainnya. Selain itu, pendekatan *federated learning* dapat dipertimbangkan sebagai metode pengembangan model lintas fasilitas kesehatan dengan tetap memperhatikan kerahasiaan dan keamanan data medis yang bersifat sensitif [15], [16]. Pengembangan selanjutnya juga perlu mencakup optimalisasi nilai *threshold* atau ambang batas klasifikasi berdasarkan karakteristik populasi dan kebutuhan klinis di Indonesia, terutama untuk meningkatkan sensitivitas dalam mengidentifikasi individu yang berpotensi mengalami Diabetes Mellitus. Di samping itu, penambahan variabel gaya hidup yang lebih spesifik, seperti kebiasaan merokok, tingkat stres, konsumsi air putih, serta riwayat genetik yang lebih mendalam, dan peningkatan jumlah serta keragaman dataset diharapkan dapat meningkatkan kemampuan generalisasi dan keandalan model. Dengan pengembangan tersebut, model klasifikasi CatBoost diharapkan dapat memberikan kontribusi lebih lanjut dalam mendukung kegiatan skrining awal Diabetes Mellitus di Indonesia.

REFERENCES

[1] I. Genitsaridi *et al.*, "11th Edition Of The IDF Diabetes Atlas: Global, Regional, And National Diabetes Prevalence Estimates For 2024 And Projections For 2050," *Lancet Diabetes Endocrinol.*, vol. 14, no. 2, pp. 149–156, 2026, [Online]. Available: 10.1016/S2213-8587(25)00299-2

[2] N. L. Sabili, F. R. Umbara, and M. Melina, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Categorical Boosting Dengan Faktor Risiko Diabetes," *Jati (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 11391–11398, 2024, [Online]. Available: 10.36040/jati.v8i6.11447

[3] 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, p. 13101, 2023, [Online]. Available: 10.1038/s41598-023-40170-0

[4] M. Sumathi, S. Sahana, S. S. R. Rajeswari, V. Kruthi, and S. P. Raja, "Enhancing Diabetes Prediction With Minimal Processing Time Using Catboost: A Comparative Study," *J. Eval. Clin. Pract.*, vol. 31, no. 7, p. e70284, 2025, [Online]. Available: 10.1111/jep.70284

[5] American Diabetes Association Professional Practice Commitee, "Diagnosis And Classification Of Diabetes: Standards Of Care In Diabetes-2024," *Diabetes Care*, vol. 47, no. Suppl 1, p. S20, 2023, [Online]. Available: 10.2337/dc26-S002

[6] R. Ravi and J. J. Olickal, "Machine Learning-Based Risk Prediction Models For Type 2 Diabetes In Primary Care: A Scoping Review," *Front. Public Heal.*, vol. 14, p. 1946070, 2026, [Online]. Available: 10.3389/fpubh.2026.1946070

[7] J. M. Sanches, L. N. Zhao, A. Salehi, C. B. Wollheim, and P. Kaldis, "Pathophysiology Of Type 2 Diabetes And The Impact Of Altered Metabolic Interorgan Crosstalk," *FEBS J.*, vol. 290, no. 3, pp. 620–648, 2023, [Online]. Available: 10.1111/febs.16306

[8] M. Li, A. Tan, L. Song, X. Qiu, L. Zhou, and Z. Wen, "Performance Of Diabetes Risk Prediction Models: A Systematic Review And Meta-Analysis," *Endocr. Connect.*, vol. 14, no. 11, 2025, [Online]. Available: 10.1530/EC-25-0353

[9] K. De Silva, W. K. Lee, A. Forbes, R. T. Demmer, C. Barton, and J. Enticott, "Use And Performance Of Machine Learning Models For Type 2 Diabetes Prediction In Community Settings: A Systematic Review And Meta-Analysis," *Int. J. Med. Inform.*, vol. 143, p. 104268, 2020, [Online]. Available: 10.1016/j.ijmedinf.2020.104268

[10] 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.*, vol. 15, no. 1, p. 73, 2023, [Online]. Available: 10.1186/s13321-023-00743-7

[11] G. S. Collins *et al.*, "TRIPOD+ AI Statement: Updated Guidance For Reporting Clinical Prediction Models That Use Regression Or Machine Learning Methods," *BMJ*, vol. 385, 2024, [Online]. Available: 10.1136/bmj-2023-078378

[12] Y. Qin *et al.*, "Machine Learning Models For Data-Driven Prediction Of Diabetes By Lifestyle Type," *Int. J. Environ. Res. Public Health*, vol. 19, no. 22, p. 15027, 2022, [Online]. Available: 10.3390/ijerph192215027

[13] E. Riveros Perez and B. Avella-Molano, "Learning From The Machine: Is Diabetes In Adults Predicted By Lifestyle Variables? A Retrospective Predictive Modelling Study Of NHANES 2007–2018," *BMJ Open*, vol. 15, no. 3, p. e096595, 2025, [Online]. Available: 10.1136/bmjopen-2024-096595

[14] S. M. Ganie, "Explainable Ensemble Boosting Framework for Chronic Disease Risk Prediction Using Lifestyle and Clinical Data: A Multi-Dataset Study," *Life*, 2026, [Online]. Available: 10.3390/life16101734

[15] M. R. Hasan and J. Li, "Privacy-Preserving Collaborative Diabetes Prediction In Heterogeneous Health Care Systems: Algorithm Development And Validation Of A Secure Federated Ensemble Framework," *JMIR Diabetes*, vol. 11, p. e79166, 2026, [Online]. Available: 10.2196/79166

[16] N. Aasoum, I. Jellouli, and S. Amjad, "Enhanced Diabetes Detection via a Privacy-Preserving Federated Learning Framework," *Acta Inform. Pragensia*, vol. 15, no. 2, pp. 327–344, 2026, [Online]. Available: 10.18267/j.aip.304