JIIFKOM (Jurnal Ilmiah Informatika & Komputer) STTR Cepu
E-Mail: jiifkom@sttrcepu.ac.id

Pengembangan Model *Hybrid Long Short-Term Memory – Extreme Gradient Boosting* untuk Prediksi Diabetes dan Kolesterol

Naila Nadra[a], Ade Silvia Handayani[b*], Mohammad Fadhli[c]
e-mail: ade_silvia@polsri.ac.id

Politeknik Negeri Sriwijaya[abc]

Kata kunci:
LSTM, XGBoost, Diabetes, Kolesterol, SMOTE, Machine Learning

Intisari
Diabetes dan *hiperkolesterolemia* adalah dua masalah kesehatan yang dapat meningkatkan risiko berbagai penyakit kronis jika tidak diidentifikasi dan ditangani dengan segera. Pendekatan prediktif diperlukan untuk memastikan risiko penyakit dengan menggunakan data kesehatan pasien. Penelitian ini bertujuan untuk mengembangkan model prediksi diabetes dan kolesterol menggunakan metode *hybrid Long Short-Term Memory* (LSTM) dan *Extreme Gradient Boosting* (XGBoost). Dataset yang digunakan diperoleh dari platform Kaggle dan mencakup berbagai parameter kesehatan, seperti *age, gender, body mass index* (BMI), *blood pressure*, HbA1c, *blood glucose*, serta riwayat kesehatan dan gaya hidup pasien. Tahapan penelitian meliputi pengumpulan data, *preprocessing*, pembagian data pelatihan dan pengujian, penyeimbangan data menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE), ekstraksi fitur menggunakan LSTM, serta proses klasifikasi menggunakan XGBoost. Berdasarkan hasil pengujian, model prediksi diabetes tersebut memperoleh nilai accuracy sebesar 74,42%, *precision* sebesar 20,20%, *recall* sebesar 68,06%, dan F1-score sebesar 31,15%. Sementara itu, model prediksi kolesterol memperoleh nilai *accuracy* sebesar 71,73%, *precision* sebesar 45,16%, *recall* sebesar 60,73%, dan F1-*score* sebesar 51,80%. Hasil tersebut menunjukkan bahwa metode *hybrid* LSTM–XGBoost mampu mengidentifikasi pola pada data kesehatan pasien dan menghasilkan performa klasifikasi yang cukup baik dalam mendukung deteksi dini risiko diabetes dan kolesterol tinggi.

Tentang naskah:
-diterima, 16 Juni 2026
-diterbitkan, 1 Juli 2026

Abstract
Diabetes and high cholesterol are two health conditions that can increase the risk of various chronic diseases if not detected and treated early. Therefore, a prediction method is needed that can help identify disease risks based on patient health data. This study aims to develop a diabetes and cholesterol prediction model using a hybrid Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) method. The dataset used was obtained from the Kaggle platform and includes various health parameters, such as age, gender, body mass index (BMI), blood glucose levels, HbA1c levels, blood pressure, as well as the patient's medical history and lifestyle. The research stages include data collection, data pre-processing, training and testing data division, data balancing using the Synthetic Minority Over-sampling Technique (SMOTE), feature extraction using LSTM, and the classification process using XGBoost. Based on the test results, the diabetes prediction model obtained an accuracy value of 74.42%, precision of 20.20%, recall of 68.06%, and an F1-score of 31.15%. Meanwhile, the cholesterol prediction model achieved an accuracy of 71.73%, a precision of 45.16%, a recall of 60.73%, and an F1-score of 51.80%. These results demonstrate that the hybrid LSTM–XGBoost method is capable of identifying patterns in patient health data and producing quite good classification performance to support early detection of diabetes and high cholesterol risks.

Keywords:
LSTM, XGBoost, Diabetes, Cholesterol, SMOTE, Machine Learning

1. Pendahuluan

Diabetes melitus dan dislipidemia merupakan dua penyakit metabolik yang prevalensinya terus meningkat di berbagai negara dan menjadi faktor risiko utama terjadinya komplikasi kardiovaskular, gagal ginjal, stroke, serta peningkatan angka mortalitas. Kondisi tersebut mendorong perlunya deteksi dini dan pemantauan kesehatan secara berkelanjutan agar intervensi dapat dilakukan lebih cepat sebelum terjadi komplikasi yang lebih serius. Pemanfaatan

teknologi digital dalam bidang kesehatan menjadi salah satu pendekatan yang menjanjikan untuk mendukung proses prediksi penyakit secara lebih efektif dan efisien (Afsaneh et al., 2022)

Banyak penelitian telah mulai menggunakan data klinis dan riwayat kesehatan pasien untuk mendukung prediksi penyakit metabolik karena semakin banyaknya ketersediaan data kesehatan digital. Risiko diabetes dan masalah kolesterol diketahui berkorelasi dengan informasi kesehatan seperti usia, jenis kelamin, indeks massa tubuh (BMI), tekanan darah, kadar glukosa darah, kadar HbA1c, dan riwayat medis. Dengan menggunakan data ini, model prediktif yang dapat membantu mengidentifikasi risiko penyakit secara lebih cepat dan objektif dapat dikembangkan (J. J. Khanam & Foo, 2022).

Seiring berkembangnya teknologi kecerdasan buatan, berbagai algoritma *machine learning* dan *deep* learning telah diterapkan untuk memprediksi risiko diabetes. Kajian sistematis menunjukkan bahwa metode berbasis *machine learning* mampu meningkatkan kemampuan identifikasi pola kompleks pada data kesehatan dibandingkan pendekatan statistik konvensional (Fregoso-Aparicio et al., 2021) Selain itu, proses prapengolahan data yang tepat serta pemanfaatan model pembelajaran mendalam juga terbukti dapat menghasilkan performa prediksi diabetes yang tinggi pada berbagai jenis dataset kesehatan (J. J. Khanam & Foo, 2022).

Salah satu arsitektur *deep learning* yang banyak digunakan pada data berurutan adalah *Long Short-Term Memory* (LSTM). *Long Short-Term Memory* (LSTM) memiliki kemampuan dalam mempelajari ketergantungan jangka panjang pada data *time series* sehingga sesuai digunakan untuk menganalisis perubahan kondisi fisiologis dari waktu ke waktu. Penelitian sebelumnya menunjukkan bahwa model *Long Short-Term Memory* (LSTM) mampu menghasilkan prediksi kadar glukosa dengan tingkat kesalahan yang rendah dan kestabilan prediksi yang baik pada data pasien diabetes (M. F. Rabby et al., 2021). Penelitian lain juga memperlihatkan bahwa LSTM memiliki performa yang kompetitif dalam tugas prediksi diabetes berbasis data kesehatan.

Di sisi lain, *Extreme Gradient Boosting* (XGBoost) merupakan algoritma ensemble berbasis decision tree yang dikenal memiliki kemampuan generalisasi yang baik, tahan terhadap overfitting, serta efektif dalam menangani hubungan nonlinier antarfitur. Pemanfaatan *Extreme Gradient Boosting* (XGBoost) pada sistem deteksi diabetes berbasis parameter fisiologis telah menunjukkan tingkat akurasi yang tinggi dengan kompleksitas komputasi yang relatif efisien (Kavitha, 2021) Selain itu, pendekatan berbasis *Extreme Gradient Boosting* (XGBoost) juga terbukti efektif dalam mengintegrasikan berbagai variabel fisiologis dan klinis untuk meningkatkan kemampuan klasifikasi risiko diabetes (Firthous & Murugeswari, 2026).

Beberapa penelitian terbaru mulai mengembangkan pendekatan *hybrid* dengan mengombinasikan kemampuan ekstraksi pola temporal dari LSTM dan kemampuan klasifikasi *Extreme Gradient Boosting* (XGBoost). Kombinasi kedua metode tersebut dilaporkan mampu meningkatkan performa prediksi risiko diabetes dibandingkan penggunaan model tunggal karena dapat memanfaatkan keunggulan masing-masing algoritma dalam proses pembelajaran data (Mufidah & Handayani, 2025)

Meskipun demikian, sebagian besar penelitian terdahulu masih berfokus pada penggunaan model tunggal atau hanya menerapkan salah satu metode dalam proses prediksi. Oleh karena itu, diperlukan penelitian lebih lanjut untuk mengevaluasi efektivitas model hybrid dalam meningkatkan performa prediksi pada dataset kesehatan dengan karakteristik yang beragam.

Berdasarkan keadaan tersebut, penelitian ini menggunakan dataset kesehatan masyarakat dari platform Kaggle untuk membuat model prediksi diabetes dan kolesterol menggunakan teknik *Hybrid Long Short-Term Memory* (LSTM)– *Extreme Gradient Boosting (*XGBoost). Sejumlah indikator kesehatan pasien, termasuk *age*, *gender*, *body mass index* (BMI), *blood pressure*, HbA1c, *blood glucose*, serta riwayat medis dan gaya hidup pasien, disertakan dalam dataset. Untuk meningkatkan proses diagnosis dini masalah diabetes dan kolesterol, diharapkan model yang diusulkan dapat mengidentifikasi pola korelasi antar faktor kesehatan dengan lebih tepat.

2. Kerangka Teori

2.1. Long Short-Term Memory (LSTM)

(Hochreiter & Schmidhuber, 1997), mengusulkan *Long Short-Term Memory* (LSTM) sebagai peningkatan dari *Recurrent Neural Networks* (RNN) untuk mengurangi masalah *vanishing* gradients dalam pembelajaran data sekuensial. LSTM memiliki kemampuan untuk mempertahankan informasi dalam jangka waktu yang lama melalui mekanisme memori internal, yang memungkinkan mereka untuk menyimpan data penting dari *input* sebelumnya sambil mengabaikan informasi yang kurang relevan. Karakteristik ini memungkinkan *Long Short-Term Memory* (LSTM) untuk unggul dalam mengidentifikasi pola dengan ketergantungan temporal.

Menurut (Goodfellow et al., 2021), *Long Short-Term Memory* (*LSTM*) menggunakan mekanisme gerbang untuk mengatur aliran informasi yang masuk, disimpan, dan dikeluarkan dari memori.

Struktur tersebut memungkinkan model untuk mempelajari hubungan yang kompleks pada data berurutan dengan lebih baik dibandingkan RNN konvensional. Oleh karena itu, *Long Short-Term Memory* (LSTM) banyak diterapkan dalam berbagai bidang, seperti pengolahan bahasa alami, prediksi deret waktu, pengenalan suara, hingga analisis data kesehatan.

Dalam bidang kesehatan, *Long Short-Term Memory* (LSTM) telah dimanfaatkan untuk memprediksi kondisi pasien berdasarkan data fisiologis yang dikumpulkan secara berkala. Penelitian oleh (M. F. F. Rabby, 2021) menunjukkan bahwa model *Long Short-Term Memory* (LSTM) mampu menghasilkan prediksi kadar glukosa darah dengan tingkat kesalahan yang rendah karena kemampuannya dalam mempelajari perubahan pola data pasien dari waktu ke waktu. Selain itu, (J. Khanam & Foo, 2021) menyatakan bahwa model berbasis LSTM memiliki performa yang baik dalam prediksi diabetes karena mampu mengidentifikasi hubungan nonlinier dan ketergantungan temporal pada data kesehatan.

2.2. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) adalah teknik pembelajaran *ensemble* yang berasal dari metodologi *Gradient Boosting Decision Tree* (GBDT). XGBoost bekerja dengan membangun sejumlah *decision tree* secara bertahap, di mana setiap pohon baru digunakan untuk memperbaiki kesalahan prediksi dari pohon sebelumnya sehingga menghasilkan model yang lebih akurat (Sharma et al., 2023).

XGBoost memiliki beberapa keunggulan, seperti kemampuan menangani data berukuran besar, mengurangi risiko *overfitting* melalui teknik regularisasi, serta mampu memodelkan hubungan nonlinier antar variabel dengan baik (Verma et al., 2022) Oleh karena itu, algoritma ini banyak digunakan dalam berbagai tugas klasifikasi dan prediksi, termasuk pada bidang kesehatan. Pada penelitian ini, XGBoost digunakan sebagai algoritma klasifikasi setelah proses ekstraksi fitur menggunakan *Long Short-Term Memory* (LSTM). Fitur yang dihasilkan oleh LSTM selanjutnya digunakan sebagai masukan bagi *Extreme Gradient Boosting* (XGBoost) untuk melakukan prediksi diabetes dan kolesterol sehingga diharapkan dapat meningkatkan performa model secara keseluruhan.

2.3. Confusion Matrix

Confusion Matrix membandingkan prediksi model klasifikasi dengan kondisi aktual untuk mengukur kinerja model. Salah satu keluaran yang berguna dari metode ini adalah matriks yang merinci persentase prediksi akurat dan tidak akurat, yang dapat memberikan gambaran tentang kinerja keseluruhan model. Keberhasilan klasifikasi dan jenis kesalahan prediksi lainnya dapat diukur dengan menggunakan Matriks *confusion* (Sathyanarayanan & Tantri, 2024).

Struktur empat bagian ini merupakan dasar dari matriks kebingungan klasifikasi biner, yang juga mencakup *false positives* (FP), *true negatives* (FN), dan true positives (TP). Data ini digunakan untuk menghitung beberapa ukuran penilaian, seperti skor F1, *recall*, spesifisitas, *accuracy*, dan *precision*. Peneliti dapat memperoleh pemahaman yang lebih baik tentang kemampuan klasifikasi suatu model dan model yang paling tepat untuk pekerjaan tertentu dengan menggunakan ukuran-ukuran ini untuk menilai kinerja model. (Sathyanarayanan & Tantri, 2024).

2.4. Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) adalah pendekatan oversampling yang digunakan untuk memperbaiki ketidakseimbangan kelas dalam dataset klasifikasi. Ketidakseimbangan kelas terjadi ketika jumlah data dalam satu kelas jauh lebih sedikit daripada di kelas lain, sehingga model menghasilkan prediksi yang bias yang menguntungkan kelas mayoritas dan kesulitan mengidentifikasi pola di kelas minoritas (Kumar et al., 2022). SMOTE awalnya dikembangkan oleh (Chawla et al., 2002) dan terus digunakan serta disempurnakan secara luas dalam beberapa penelitian, khususnya yang berkaitan dengan data kesehatan yang ditandai dengan distribusi kelas yang tidak seimbang.

SMOTE beroperasi dengan menciptakan sampel sintetis baru dari kelas minoritas melalui metode interpolasi antara sampel minoritas dan k-tetangga terdekatnya. Strategi ini bertujuan untuk meningkatkan representasi kelas minoritas dalam distribusi data pelatihan tanpa hanya menduplikasi data yang ada (Sowjanya & Mrudula, 2023).

Dalam bidang kesehatan, ketidakseimbangan data sering ditemukan karena jumlah pasien yang terdiagnosis suatu penyakit umumnya lebih sedikit dibandingkan pasien dengan kondisi normal. Kondisi tersebut dapat menyebabkan model memiliki akurasi yang tinggi tetapi sensitivitas yang rendah dalam mendeteksi kasus positif. Oleh karena itu, penerapan teknik penyeimbangan data seperti SMOTE banyak digunakan untuk meningkatkan kemampuan model dalam mengidentifikasi kelas minoritas dan memperbaiki performa klasifikasi pada data medis (Araf et al., 2024).

Sebelum melatih model *hybrid* LSTM–XGBoost, penelitian ini menggunakan teknik SMOTE selama persiapan data untuk menyeimbangkan distribusi kelas dalam dataset diabetes dan

kolesterol. Distribusi data yang lebih merata diharapkan dapat meningkatkan kemampuan model untuk membedakan atribut dari setiap kelas, sehingga meningkatkan efektivitas prediksinya.

2.5 Literature Review

Prediksi diabetes telah menjadi subjek beberapa studi yang menggunakan teknik *deep learning* dan machine *learning*. Dengan menggunakan seleksi fitur *hybrid,* penelitian (Kavitha, 2021) menunjukkan bahwa algoritma XGBoost dapat mencapai kinerja klasifikasi yang baik dalam prediksi diabetes. Pada saat yang sama, (M. F. Rabby et al., 2021) menunjukkan bahwa pendekatan berbasis *deep learning* dapat berhasil mempelajari pola temporal dalam data kesehatan dengan memprediksi kadar glukosa darah menggunakan model LSTM.

(J. J. Khanam & Foo, 2022) meneliti beberapa algoritma ML untuk prediksi diabetes dan menemukan bahwa seleksi fitur dan *preprocessing* yang akurat sangat meningkatkan kinerja model. Selain itu, XGBoost dapat secara efektif menangani korelasi nonlinier dalam data kesehatan, memiliki kemampuan generalisasi yang tinggi, dan tahan terhadap *overfitting* (Sharma et al., 2023).

Sejumlah proyek penelitian telah mulai mengerjakan metode *hybrid* yang menggabungkan kemampuan ekstraksi fitur model pembelajaran mendalam dengan kemampuan klasifikasi algoritma *ensemble*. Menurut (Mufidah & Handayani, 2025), menggabungkan LSTM dengan XGBoost dapat meningkatkan akurasi prediksi risiko diabetes jika dibandingkan dengan hanya menggunakan satu model. Namun demikian, penelitian saat ini hanya melihat prediksi diabetes dan belum mempertimbangkan untuk menerapkan metode yang sama untuk memprediksi kadar kolesterol.

Sebagian besar studi masih menggunakan satu model atau hanya memprediksi satu jenis penyakit metabolik berdasarkan penelitian sebelumnya. Belum banyak diketahui tentang penggunaan dataset kesehatan masyarakat untuk prediksi gabungan diabetes dan kolesterol menggunakan algoritma LSTM dan XGBoost. Untuk meningkatkan kemampuan prediksi model LSTM-XGBoost pada dataset diabetes dan kolesterol dengan distribusi kelas yang tidak merata, penelitian ini menyarankan pendekatan *hybrid* yang menggabungkan SMOTE.

3. Metodologi

3.1. Pengumpulan Dataset

Penelitian ini menggunakan dua dataset publik yang diperoleh dari platform Kaggle dalam format .csv, yaitu dataset prediksi diabetes dan dataset penyakit kardiovaskular. Kaggle merupakan platform yang banyak dimanfaatkan oleh praktisi dan peneliti ilmu data untuk menyediakan dataset terbuka serta mendukung pengembangan model pembelajaran mesin (Hayashi et al., 2021). Penggunaan dataset publik memungkinkan penelitian dilakukan secara transparan, mudah direplikasi, serta memudahkan perbandingan hasil dengan penelitian sebelumnya.

Dapat dilihat pada gambar 3.1, dataset diabetes terdiri dari 100.000 data dengan 9 atribut, yaitu gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, dan diabetes sebagai variabel target. Variabel diabetes memiliki dua kelas, yaitu nilai 0 yang menunjukkan tidak menderita diabetes dan nilai 1 yang menunjukkan menderita diabetes.

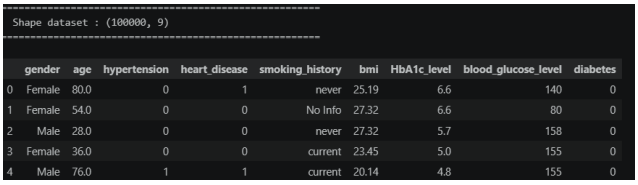
Sementara itu, gambar 3.2 menunjukkan bahwa data prediksi kolesterol menggunakan dataset penyakit kardiovaskular yang terdiri dari 70.000 data dengan 14 atribut, yaitu Unnamed: 0, id, *age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active,* dan *cardio*. Pada penelitian ini, atribut *cholesterol* digunakan sebagai variabel target untuk prediksi kolesterol, sedangkan atribut lainnya digunakan sebagai variabel prediktor. Atribut *cholesterol* terdiri atas tiga kategori, yaitu nilai 1 yang menunjukkan kadar kolesterol normal, nilai 2 yang menunjukkan kadar kolesterol di atas normal, dan nilai 3 yang menunjukkan kadar kolesterol jauh di atas normal.

Tahap pengumpulan data dilakukan dengan mengunduh kedua dataset dari Kaggle, kemudian mengidentifikasi atribut yang relevan untuk proses pemodelan. Pemilihan fitur mempertimbangkan kualitas data, konsistensi nilai, serta kontribusi masing-masing atribut terhadap kemampuan model dalam melakukan prediksi. Proses pemilihan fitur yang tepat dapat meningkatkan kinerja model dengan mempertahankan atribut yang paling informatif serta mengurangi kompleksitas data yang tidak diperlukan (Pudjihartono et al., 2022).

Selanjutnya, validasi pendahuluan dilakukan pada kedua dataset dengan menilai kuantitas data, kompatibilitas tipe data, dan menemukan nilai yang tidak konsisten atau anomali yang dapat berdampak buruk pada proses pelatihan model. Fase ini bertujuan untuk menjamin bahwa data yang digunakan memiliki kualitas yang memadai untuk memfasilitasi pembuatan model *hybrid* LSTM–XGBoost untuk prediksi terbaik kadar diabetes dan kolesterol.

Untuk mendukung seluruh tahapan penelitian, mulai dari preprocessing data, pembagian data pelatihan dan pengujian, pelatihan model, hingga evaluasi kinerja, penelitian ini menggunakan bahasa

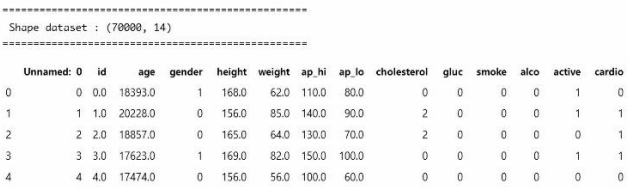
pemrograman Python versi 3.10.0 pada lingkungan pengembangan Visual Studio Code (VS Code). Proses normalisasi data, pembagian data, serta penerapan metode Synthetic Minority Over-sampling Technique (SMOTE) dilakukan menggunakan pustaka Scikit-Learn versi 1.5.1 dan Imbalanced-Learn versi 0.13.0. Model Long Short-Term Memory (LSTM) dikembangkan menggunakan Keras versi 3.9.1 dengan dukungan TensorFlow versi 2.19.0 sebagai backend, sedangkan proses klasifikasi dilakukan menggunakan pustaka XGBoost versi 3.2.0.



Shape dataset : (100000, 9)

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Gambar 3.1 Dataset Diabetes.



Shape dataset : (70000, 14)

	Unnamed: 0	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	0.0	18393.0	1	168.0	62.0	110.0	80.0	0	0	0	0	1	0
1	1	1.0	20228.0	0	156.0	85.0	140.0	90.0	2	0	0	0	1	1
2	2	2.0	18857.0	0	165.0	64.0	130.0	70.0	2	0	0	0	0	1
3	3	3.0	17623.0	1	169.0	82.0	150.0	100.0	0	0	0	0	1	1
4	4	4.0	17474.0	0	156.0	56.0	100.0	60.0	0	0	0	0	0	0

Gambar 3.2 Dataset kolesterol.

3.2. Preprocessing

Fase pembersihan data dilakukan untuk menjamin bahwa dataset yang digunakan dalam penelitian ini berkualitas tinggi dan siap untuk proses pelatihan model. Metode ini melibatkan verifikasi kuantitas data, mendeteksi data yang tidak akurat, dan memeriksa atribut fitur di dalam setiap dataset.

Dapat dilihat pada gambar 3.3 dataset diabetes awalnya terdiri dari 100.000 entri. Setelah proses pembersihan, semua data dianggap sesuai untuk digunakan, dan tidak ada data yang dibuang. Dataset tersebut terdiri dari delapan variabel prediktor: *age, gender, body mass index* (BMI), *blood pressure, HbA1c, blood glucose*, bersama dengan satu variabel target, diabetes. Temuan identifikasi mengungkapkan bahwa variabel tersebut berpotensi memainkan peran penting dalam proses prediksi diabetes.

Sementara itu, pada gambar 3.4 dataset awal yang digunakan untuk analisis kolesterol terdiri dari 70.000 entri. Setelah prosedur pembersihan data, jumlah total titik data berkurang menjadi 68.730, yang mengakibatkan penghapusan 1.270 titik data. Dataset ini terdiri dari beberapa variabel prediktor: *age, gender, height, weight*, ap_hi, ap_lo, gluk, merokok, alkohol, aktif, dan kardio, bersama dengan variabel target kolesterol_label. Temuan identifikasi fitur menunjukkan bahwa atribut usia, tinggi badan, berat badan, ap_hi, ap_lo, dan BMI memiliki signifikansi yang lebih besar daripada atribut lainnya.

Hasil pembersihan data dan identifikasi fitur menunjukkan bahwa kedua dataset memiliki kualitas data yang memadai untuk diterapkan dalam fase pemodelan. Teknik ini menawarkan ringkasan komprehensif tentang atribut yang secara signifikan memengaruhi prediksi diabetes dan kolesterol. Selanjutnya, dataset yang telah melalui proses preprocessing dibagi menjadi data pelatihan dan data pengujian. Sebelum digunakan dalam proses pelatihan model, data pelatihan terlebih dahulu dinormalisasi untuk menyamakan rentang nilai antarfitur. Mengingat distribusi kelas pada kedua dataset masih menunjukkan kondisi *class imbalance*, dilakukan penyeimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) untuk meningkatkan kemampuan model dalam mengenali kelas minoritas.

Jumlah data awal : 100,000
Jumlah setelah cleaning: 100,000
Data yang dihapus : 0

=== FITUR SETELAH CLEANING ===

Feature	Popularity
gender	Rendah
age	Tinggi
hypertension	Rendah
heart_disease	Rendah
smoking_history	Rendah
bmi	Tinggi
HbA1c_level	Tinggi
blood_glucose_level	Tinggi
diabetes	Rendah

Gambar 3.3 Hasil preprocessing dataset diabetes setelah proses cleaning data.

Jumlah data awal : 70,000
Jumlah setelah cleaning: 68,730
Data yang dihapus : 1,270

=== FITUR SETELAH CLEANING ===

Feature	Popularity
age	Tinggi
gender	Rendah
height	Tinggi
weight	Tinggi
ap_hi	Tinggi
ap_lo	Tinggi
cholesterol	Rendah
gluc	Rendah
smoke	Rendah
alco	Rendah
active	Rendah
cardio	Rendah
bmi	Tinggi
cholesterol_label	Rendah

Gambar 3.4 Hasil preprocessing dataset Kolesterol setelah proses cleaning data.

3.3. Diabetes Dataset Splitting and Balancing

Setelah melalui tahap preprocessing, dataset diabetes dibagi menjadi data pelatihan dan data pengujian menggunakan rasio 80:20. Sebanyak 80% data digunakan untuk proses pelatihan

model, sedangkan 20% sisanya digunakan untuk pengujian. Untuk memastikan proporsi setiap kelas pada data pelatihan dan data pengujian tetap konsisten dengan distribusi kelas pada dataset asli, digunakan teknik stratified sampling melalui parameter stratify=y.

Setelah pembagian data dilakukan, fitur pada data pelatihan menjalani proses normalisasi untuk menyamakan rentang nilai antar atribut. Selanjutnya dilakukan penyeimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) karena distribusi kelas pada data pelatihan masih menunjukkan kondisi *class imbalance*. Ketidakseimbangan kelas dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas sehingga menurunkan kemampuan model dalam mengenali kelas minoritas (Kumar et al., 2022).

SMOTE hanya digunakan dalam penelitian ini dengan parameter k-neighbors = 5 pada data pelatihan yang dinormalisasi (X_train_scaled). Untuk mencapai distribusi kelas yang lebih seimbang, strategi ini menciptakan sampel sintetis tambahan di kelas minoritas berdasarkan lima tetangga terdekat (k-nearest neighbors) dari setiap sampel (Sowjanya & Mrudula, 2023). Untuk mencegah kebocoran data yang dapat memengaruhi temuan evaluasi model, SMOTE hanya diterapkan pada data pelatihan.

Hasil dari proses ini menghasilkan data pelatihan baru, yaitu X_train_smote dan y_train_smote, yang memiliki distribusi kelas lebih seimbang dibandingkan data awal. Dataset hasil penyeimbangan tersebut kemudian digunakan dalam proses pelatihan model *hybrid* LSTM–XGBoost, sedangkan data pengujian tetap menggunakan data asli tanpa proses *oversampling* untuk memastikan evaluasi model dilakukan pada kondisi data yang sebenarnya.

```
X_train_raw, X_test_raw, y_train, y_test = train_test_split(
    X, y, test_size=0.20, random_state=SEED, stratify=y
)
```

Gambar 3.5 Pembagian dataset menggunakan metode train-test split.

```
print('Distribusi label SEBELUM SMOTE (train):')
print(f'  Tidak Diabetes (0): {(y_train == 0).sum():,}')
print(f'  Diabetes       (1): {(y_train == 1).sum():,}')
print(f'  Rasio          : {(y_train==1).sum()/(y_train==0).sum()*100:.1f}% positif')

smote = SMOTE(random_state=SEED, k_neighbors=5)
X_train_smote, y_train_smote = smote.fit_resample(X_train_scaled, y_train)

print()
print('Distribusi label SETELAH SMOTE (train):')
print(f'  Tidak Diabetes (0): {(y_train_smote == 0).sum():,}')
print(f'  Diabetes       (1): {(y_train_smote == 1).sum():,}')
```

Gambar 3.6 Penyeimbangan data menggunakan metode SMOTE.

3.4 Cholesterol Dataset Splitting and Balancing

Untuk menjaga konsistensi dengan distribusi kelas dataset asli, parameter stratify=y menerapkan strategi pengambilan sampel bertingkat (*stratified sampling*) baik pada dataset pelatihan maupun pengujian. Pada penelitian ini digunakan rasio pembagian sebesar 80:20, di mana 80% data digunakan untuk pelatihan model dan 20% data digunakan untuk pengujian. Selain itu, diterapkan teknik stratified sampling melalui parameter stratify=y untuk memastikan proporsi setiap kelas pada data pelatihan dan data pengujian tetap sama dengan distribusi kelas pada dataset asli.

Setelah pembagian data dilakukan, distribusi kelas pada data pelatihan menunjukkan ketidakseimbangan antara kelas normal (0) dan kolesterol tinggi (1). Kondisi ini berpotensi menyebabkan model lebih cenderung mempelajari pola dari kelas mayoritas sehingga kemampuan dalam mengenali kelas minoritas menjadi kurang optimal (Kumar et al., 2022). Oleh karena itu, dilakukan proses penyeimbangan data menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE).

SMOTE diterapkan hanya pada data pelatihan yang telah melalui tahap normalisasi untuk menghindari terjadinya data *leakage* selama proses pengembangan model. Pada penelitian ini digunakan parameter k-neighbors = 5, sehingga pembentukan data sintetis dilakukan berdasarkan lima tetangga terdekat dari setiap sampel kelas minoritas (Sowjanya & Mrudula, 2023).

Hasil proses SMOTE menghasilkan data pelatihan baru yang memiliki distribusi kelas lebih seimbang, yaitu X_train_smote sebagai fitur pelatihan dan y_train_smote sebagai label pelatihan. Dataset hasil penyeimbangan tersebut selanjutnya digunakan dalam proses pelatihan model *hybrid* LSTM–XGBoost, sedangkan data pengujian tetap menggunakan data asli tanpa proses *oversampling*. Dengan distribusi kelas yang lebih seimbang, model diharapkan mampu meningkatkan kemampuan prediksi terhadap kondisi kolesterol secara lebih akurat.

```
X_train_raw, X_test_raw, y_train, y_test = train_test_split(
    X, y, test_size=0.20, random_state=SEED, stratify=y
)
```

Gambar 3.7 Pembagian dataset menggunakan metode train-test split.

```
print('Distribusi label SEBELUM SMOTE (train):')
print(f'  Normal (0)           : {(y_train == 0).sum():,}')
print(f'  Kolesterol Tinggi (1): {(y_train == 1).sum():,}')

smote = SMOTE(random_state=SEED, k_neighbors=5)
X_train_smote, y_train_smote = smote.fit_resample(X_train_scaled, y_train)

print()
print('Distribusi label SETELAH SMOTE (train):')
print(f'  Normal (0)           : {(y_train_smote == 0).sum():,}')
print(f'  Kolesterol Tinggi (1): {(y_train_smote == 1).sum():,}')
```

Gambar 3.8 Pembagian dataset menggunakan metode train-test split.

4. Hasil dan Pembahasan

4.1 Hasil Klasifikasi Popularitas Menu

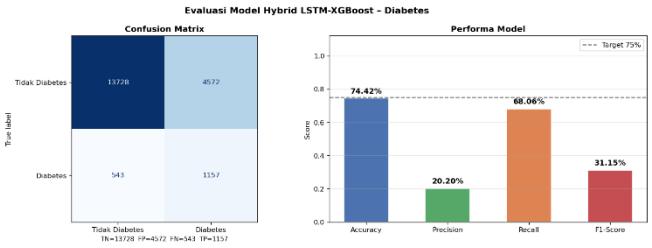
Berdasarkan data kesehatan pasien, penelitian ini memprediksi diabetes dan kolesterol menggunakan teknik *Hybrid Long Short-Term*

Memory (LSTM)–XGBoost. Preprocessing data, pemisahan data *training* dan *testing*, penyeimbangan data menggunakan *Synthetic Minority Oversampling Technique* (SMOTE), ekstraksi fitur menggunakan LSTM, dan klasifikasi menggunakan algoritma XGBoost merupakan langkah-langkah dalam proses penelitian. Matriks kebingungan, accuracy, *precision*, *recall*, dan F1-*score* digunakan untuk mengevaluasi model.

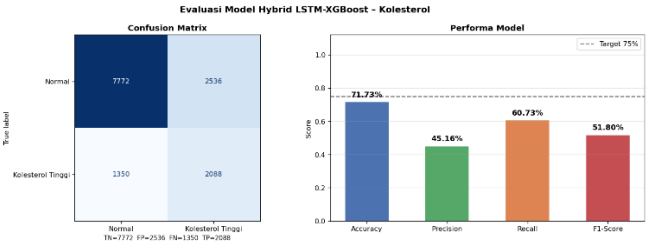
Hasil evaluasi model prediksi diabetes ditunjukkan pada Gambar 4.1. Secara keseluruhan, sebagian besar pola terkait diabetes dalam data kesehatan pasien dikenali oleh model. Model menunjukkan kemampuan yang baik untuk membedakan antara kelompok dengan dan tanpa diabetes berdasarkan matriks kebingungan. Selain itu, nilai *recall*, yang lebih besar daripada metrik lainnya, menunjukkan bahwa model tersebut dapat digunakan sebagai alat skrining awal karena memiliki kemampuan yang baik untuk mengidentifikasi orang yang mungkin menderita diabetes.

Seperti ditunjukkan pada Gambar 4.2, hasil evaluasi model prediksi kolesterol disajikan menggunakan confusion matrix dan metrik kinerja. Berdasarkan fitur kesehatan pada dataset, model menunjukkan kemampuan yang baik untuk membedakan antara orang dengan kadar kolesterol normal dan tinggi. Kemampuan klasifikasi model yang relatif konsisten di kedua kelas ditunjukkan oleh nilai *precision* dan skor F1 yang lebih seimbang.

Perbandingan hasil dari kedua dataset menunjukkan bahwa model prediksi kolesterol memberikan kinerja klasifikasi yang lebih seimbang, sementara model prediksi diabetes lebih baik dalam mengidentifikasi kasus positif. Secara keseluruhan, pendekatan *Hybrid* LSTM–XGBoost memberikan prediksi yang baik untuk membantu identifikasi dini risiko diabetes dan kolesterol tinggi dengan memanfaatkan kemampuan klasifikasi XGBoost dan kemampuan LSTM untuk mengungkap pola dari data kesehatan.



Gambar 4.1 Hasil evaluasi model hybrid LSTM–XGBoost pada prediksi diabetes.



Gambar 4.2 Hasil evaluasi model hybrid LSTM–XGBoost pada prediksi kolesterol.

5. Simpulan

Studi ini menyimpulkan bahwa metode *hybrid* LSTM–XGBoost mampu digunakan untuk melakukan prediksi diabetes dan kolesterol berdasarkan data kesehatan pasien. Pendekatan yang digunakan memanfaatkan kemampuan *Long Short-Term Memory* (LSTM) dalam mengekstraksi fitur dari data masukan serta kemampuan XGBoost dalam melakukan proses klasifikasi. Tahapan penelitian meliputi pengumpulan dataset, *preprocessing* data, pembagian dan penyeimbangan data menggunakan SMOTE, ekstraksi fitur menggunakan LSTM, serta klasifikasi menggunakan algoritma XGBoost.

Berdasarkan hasil pengujian, menunjukkan bahwa model prediksi diabetes memperoleh nilai *accuracy* sebesar 74,42%, *precision* sebesar 20,20%, *recall* sebesar 68,06%, dan F1-*score* sebesar 31,15%. Sementara itu, model prediksi kolesterol memperoleh nilai *accuracy* sebesar 71,73%, *precision* sebesar 45,16%, *recall* sebesar 60,73%, dan F1-*score* sebesar 51,80%. Hasil tersebut menunjukkan bahwa model mampu membedakan antara pasien yang berisiko dan tidak berisiko pada kedua kategori penyakit dengan tingkat akurasi yang cukup baik.

Berdasarkan hasil evaluasi, model prediksi diabetes menghasilkan nilai akurasi yang lebih tinggi dibandingkan model prediksi kolesterol. Namun, model prediksi kolesterol menunjukkan keseimbangan performa yang lebih baik berdasarkan nilai precision dan F1-*score*. Secara keseluruhan, penerapan metode *hybrid* LSTM–XGBoost mampu memberikan performa klasifikasi yang cukup baik dalam mendukung identifikasi dini risiko diabetes dan kolesterol tinggi berdasarkan karakteristik kesehatan pasien

Penelitian selanjutnya dapat dikembangkan dengan menambahkan jumlah dan variasi data, mengintegrasikan parameter kesehatan lainnya, melakukan optimasi hiperparameter pada model LSTM maupun XGBoost, serta membandingkan performa metode *hybrid* yang diusulkan dengan algoritma pembelajaran mesin lainnya untuk memperoleh hasil prediksi yang lebih optimal.

Daftar Pustaka

Artikel jurnal:

Afsaneh, E., Sharifdini, A., Ghazzaghi, H., & Zarei Ghobadi, M. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetology & Metabolic Syndrome*, *14*, 196. https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-022-00969-9

Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: A review. *Artificial Intelligence Review*, *57*, 80. https://doi.org/10.1007/s10462-023-10652-8

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Firthous, J. J., & Murugeswari, G. (2026). Diabetes prediction using PPG signal and clinical data with XGBoost classification. *The Bioscan*, *21*(2), 764–779. https://thebioscan.com/index.php/pub/article/view/5765

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*, *13*, 148. https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9

Goodfellow, I., Bengio, Y., & Courville, A. (2021). *Deep learning*. MIT Press.

Hayashi, T., Shimizu, T., & Fukami, Y. (2021). Collaborative problem solving on a data platform kaggle. *ArXiv Preprint*. https://arxiv.org/abs/2108.08157

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kavitha, V. (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. *Computers in Biology and Medicine*, *136*, 104664. https://www.sciencedirect.com/science/article/pii/S0010482521004583

Khanam, J., & Foo, S. (2021). A Comparison of Machine Learning Algorithms for Diabetes Prediction. *ICT Express*, *7*(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004

Khanam, J. J., & Foo, S. Y. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, *220*, 106773. https://www.sciencedirect.com/science/article/pii/S0169260722001596

Kumar, V., Lalotra, G. S., & Sasikala, P. (2022). Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. *Healthcare*, *10*(7), 1293. https://doi.org/10.3390/healthcare10071293

Mufidah, S. I., & Handayani, I. (2025). Deteksi risiko diabetes berdasarkan faktor kesehatan menggunakan long short-term memory (LSTM) dan XGBoost. *INTECOMS: Journal of Information Technology and Computer Science*. https://journal.ipm2kpe.or.id/index.php/INTECOM/article/view/17291

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, *2*, 927312. https://doi.org/10.3389/fbinf.2022.927312

Rabby, M. F. F. (2021). *Blood glucose prediction using LSTM-based deep learning model*.

Rabby, M. F., Tu, Y., & Hossen, M. I. (2021). Stacked LSTM based deep recurrent neural network with Kalman smoothing for blood glucose prediction. *BMC Medical Informatics and Decision Making*, *21*, 101. https://doi.org/10.1186/s12911-021-01462-5

Sathyanarayanan, S., & Tantri, B. R. (2024). *Confusion matrix-based performance evaluation metrics*. *27*(4).

Sharma, A., Kumar, R., & Singh, P. (2023). XGBoost-based machine learning models for healthcare prediction: A review. *Healthcare Analytics*, *4*, 100215.

Sowjanya, A. M., & Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Applied Nanoscience*, *13*(3), 1829–1840. https://doi.org/10.1007/s13204-021-02063-4

Verma, A., Gupta, D., & Kaur, M. (2022). Performance analysis of XGBoost algorithm for classification and prediction tasks. *Expert Systems with Applications*, *198*, 116805.

Buku:

Goodfellow, I., Bengio, Y., & Courville, A. (2021). *Deep learning*. MIT Press.