JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 1, JANUARI, 2026, Pp. 1 – 12
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Deteksi Diabetes Mellitus dengan Menggunakan Teknik Ensemble XGBoost dan LightGBM

Naufal Adhi Pratama [(1)], Danang Wahyu Utomo [(2)*]
Departemen Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia.
e-mail : 111202113590@mhs.dinus.ac.id, danang.wu@dsn.dinus.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 28 Desember 2024, direvisi 30 April 2025, diterima 15 Mei 2025, dan dipublikasikan 25 Januari 2026.

Abstract

Diabetes mellitus is a metabolic disease characterized by elevated blood sugar levels due to impaired insulin secretion, insulin action, or both. The disease has a major impact on public health and contributes to high morbidity and mortality rates in many countries. Prevention and early detection are essential to reduce the adverse effects of this disease. This study aims to analyze and apply machine learning algorithms in detecting diabetes mellitus, focusing on the use of XGBoost and LightGBM algorithms. The dataset used in this study includes various features related to diabetes risk factors, such as age, gender, body mass index (BMI), hypertension, smoking history, and HbA1c and blood glucose levels. Preprocessing was performed to clean and balance the data using the SMOTE-Tomek technique. Next, the model was built and evaluated using the K-Fold cross-validation method to measure the accuracy and stability of the model. The results showed that the XGBoost model achieved 97.31% accuracy, while the LightGBM model produced 97.26% accuracy. Combining the two models through blending techniques resulted in an accuracy of 97.51%, indicating that the combination of models can improve prediction performance. This study shows the great potential of machine learning algorithms, especially XGBoost and LightGBM, in detecting diabetes mellitus accurately and efficiently. Hopefully, the results of this study can contribute to the development of decision support systems for more effective early diagnosis of diabetes.

Keywords: Diabetes Mellitus, Machine Learning, XGBoost, LightGBM, Early Detection

Abstrak

Diabetes mellitus adalah penyakit metabolik yang ditandai dengan peningkatan kadar gula darah akibat gangguan sekresi insulin, kerja insulin, atau keduanya. Penyakit ini memiliki dampak besar terhadap kesehatan masyarakat dan berkontribusi pada tingginya angka morbiditas dan mortalitas di banyak negara. Pencegahan dan deteksi dini sangat penting untuk mengurangi dampak buruk dari penyakit ini. Penelitian ini bertujuan untuk menganalisis dan menerapkan algoritma machine learning dalam mendeteksi diabetes mellitus, dengan fokus pada penggunaan algoritma XGBoost dan LightGBM. Dataset yang digunakan dalam penelitian ini mencakup berbagai fitur terkait faktor risiko diabetes, seperti usia, jenis kelamin, indeks massa tubuh (BMI), hipertensi, riwayat merokok, serta kadar HbA1c dan glukosa darah. Proses preprocessing dilakukan untuk membersihkan dan menyeimbangkan data menggunakan teknik SMOTE-Tomek. Selanjutnya, model dibangun dan dievaluasi dengan menggunakan metode K-Fold cross-validation untuk mengukur akurasi dan kestabilan model. Hasil penelitian menunjukkan bahwa model XGBoost mencapai akurasi 97.31%, sementara model LightGBM menghasilkan akurasi 97.26%. Penggabungan kedua model melalui teknik blending menghasilkan akurasi 97.51%, yang menunjukkan bahwa kombinasi model dapat meningkatkan performa prediksi. Penelitian ini menunjukkan potensi besar dari algoritma machine learning, khususnya XGBoost dan LightGBM, dalam mendeteksi diabetes mellitus secara akurat dan efisien. Diharapkan, hasil penelitian ini dapat berkontribusi pada pengembangan sistem pendukung keputusan untuk diagnosis dini diabetes yang lebih efektif.

Kata Kunci: Diabetes Mellitus, Pembelajaran Mesin, XGBoost, LightGBM, Deteksi Dini

1. PENDAHULUAN

Diabetes mellitus, menurut definisi World Health Organization (WHO), adalah penyakit degeneratif kronis yang disebabkan oleh produksi insulin yang tidak mencukupi di pankreas atau ketidakmampuan tubuh untuk secara efektif menggunakan insulin yang diproduksi (Tanwar & Bhatia, 2024). *Hyperglycemia* (peningkatan kadar glukosa darah) menjadi indikator utama dari penyakit ini. Insulin sendiri adalah hormon yang berfungsi mengatur kadar gula darah. *Diabetes mellitus* merupakan salah satu penyakit dengan pertumbuhan tertinggi di dunia. Diperkirakan sekitar 537 juta orang dewasa di seluruh dunia, yang berusia antara 20 hingga 79 tahun, menderita diabetes, yang setara dengan 10,5% dari seluruh populasi pada rentang usia tersebut. Pada tahun 2030, jumlah penderita diabetes diperkirakan akan mencapai 643 juta orang secara global, dan meningkat menjadi 783 juta pada tahun 2045 (Mujumdar & Vaidehi, 2019). Menurut edisi ke-10 International Diabetes Federation (IDF), insiden *diabetes* di negara-negara Asia Tenggara (SEA) telah meningkat selama lebih dari 20 tahun, dan perkiraan saat ini telah melampaui seluruh proyeksi sebelumnya (Ogurtsova et al., 2017).

Menurut artikel yang diterbitkan oleh Kementerian Kesehatan Republik Indonesia, prevalensi diabetes di Indonesia mencapai 11,7% pada tahun 2023 (Rif'at et al., 2023). Ini menunjukkan peningkatan dibandingkan dengan tahun-tahun sebelumnya, di mana prevalensi diabetes tercatat sebesar 8,5% pada tahun 2018. Oleh karena itu, pemanfaatan teknologi dalam manajemen diabetes menjadi sangat penting. Dengan penerapan teknologi digital, kita dapat memprediksi dan mengidentifikasi kekurangan dalam perawatan pasien melalui penggunaan Machine Learning.

Pendekatan *Machine Learning* banyak digunakan dalam studi epidemiologi klinis terkait diabetes, karena keunggulannya dalam memprediksi dan mengklasifikasikan karakteristik pasien dengan mengenali pola dalam kumpulan data (Mengcan et al., 2021). Machine learning sendiri menggunakan berbagai jenis model algoritma, yang umumnya terbagi menjadi tiga kategori, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Masing-masing kategori memiliki karakteristik dan pendekatan yang berbeda dalam proses pembelajaran. Banyaknya penelitian yang dilakukan menunjukkan tingginya minat dalam mengembangkan metode serta merancang model machine learning yang mampu memprediksi dan mengklasifikasikan karakteristik pasien *diabetes mellitus* secara akurat.

Penelitian oleh Butt et al. (2021) menunjukkan bahwa model *Multilayer Perceptron* (MLP) mampu mengungguli pengklasifikasi lainnya dengan akurasi 86,08%. Selain itu, model *Long Short-Term Memory* (LSTM) meningkatkan performa prediksi secara signifikan dengan akurasi mencapai 87,26%. Penelitian tersebut juga melakukan analisis komparatif dengan teknik lain yang sudah ada, dan menunjukkan bahwa pendekatan yang digunakan memiliki kemampuan adaptasi yang tinggi dalam berbagai aplikasi di bidang kesehatan.

Penelitian yang dilakukan oleh Chang et al. (2023) menunjukkan bahwa model klasifikasi Naive Bayes mampu memberikan hasil yang lebih baik dibandingkan dengan model Random Forest dan Decision Tree (J48) dalam hal akurasi prediksi. Pada subset data yang hanya melibatkan tiga faktor, performa model Naive Bayes tetap kompetitif dan sebanding dengan hasil yang diperoleh oleh model Random Forest pada dataset lengkap. Model Naive Bayes mencapai akurasi sebesar 79,13%, sedangkan model Random Forest memperoleh 79,57%, keduanya menjadi akurasi tertinggi yang dicapai dalam percobaan tersebut.

Penelitian oleh Saxena et al. (2022) melakukan penelitian dengan membandingkan performa *Multilayer Perceptron*, *Decision Tree*, *K-Nearest Neighbour*, dan *Random Forest*, dengan penerapan beberapa teknik pemilihan fitur untuk mendeteksi *diabetes* pada tahap awal. Proses *preprocessing* dilakukan terhadap data mentah, termasuk penghilangan *outlier* dan imputasi nilai yang hilang berdasarkan rata-rata, serta optimasi *hyperparameter*. Eksperimen dilakukan pada dataset PIMA India menggunakan Weka 3.9. Akurasi yang dicapai oleh masing-masing model adalah sebagai berikut: *Multilayer Perceptron* sebesar 77,60%, *Decision Tree* sebesar 76,07%,

K-Nearest Neighbour sebesar 78,58%, dan *Random Forest* sebesar 79,8%, yang merupakan performa terbaik di antara model yang dibandingkan.

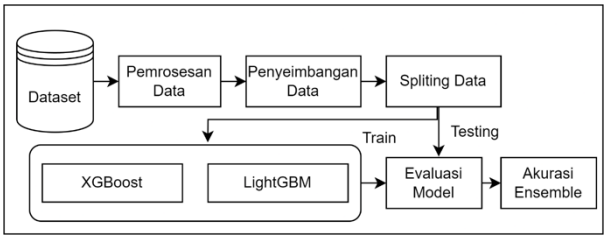
Selanjutnya penelitian oleh Lai et al. (2019) membangun model prediktif menggunakan teknik *Logistic Regression* dan *Gradient Boosting Machine* (GBM). Evaluasi performa dilakukan dengan menggunakan *receiver operating characteristic curve* (ROC). Penyesuaian ambang batas dan penerapan metode *class weighting* digunakan untuk meningkatkan sensitivitas, yaitu proporsi pasien *diabetes mellitus* yang terdeteksi secara tepat oleh model. Hasilnya, *Area Under the ROC Curve* (AUC) untuk model GBM mencapai 84,7% dengan sensitivitas 71,6%, sedangkan model *Logistic Regression* menghasilkan AUC sebesar 84,0% dengan sensitivitas 73,4%. Kedua model tersebut menunjukkan performa yang lebih baik dibandingkan dengan model *Decision Tree* dan *Random Forest*.

Berdasarkan berbagai studi tersebut, dapat disimpulkan bahwa akurasi dari model *Machine Learning* sangat bergantung pada pemilihan arsitektur algoritma yang tepat, terutama dalam mengidentifikasi kondisi medis seperti *diabetes mellitus*. Salah satu kemajuan terbesar dalam pengembangan *Machine Learning* adalah pendekatan *Ensemble Learning*. *Machine Learning* sendiri merupakan proses yang digunakan untuk menganalisis dan mengekstraksi informasi dari data dalam jumlah besar guna menemukan pengetahuan yang berguna.

Penelitian ini berfokus pada penerapan algoritma *Extreme Gradient Boosting* (XGBoost) dan *Light Gradient Boosting Machine* (LightGBM). Dataset yang digunakan mencakup berbagai fitur yang berkaitan dengan faktor risiko *diabetes*. Melalui proses *preprocessing*, termasuk penanganan data tidak seimbang menggunakan teknik *SMOTE-Tomek*, model dikembangkan dan dievaluasi dengan menggunakan metode *K-Fold Cross-Validation*. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model *Machine Learning* yang mampu memprediksi risiko *diabetes mellitus* dengan akurasi tinggi, serta mengidentifikasi faktor-faktor risiko utama yang berkontribusi terhadap kondisi tersebut melalui algoritma XGBoost dan LightGBM.

2. METODE PENELITIAN

Alur metode penelitian yang digunakan digambarkan pada Gambar 1. Penelitian dimulai dengan pengumpulan *dataset* yang berisi berbagai *fitur* terkait faktor risiko diabetes mellitus. Langkah pertama adalah *data preprocessing*, yang mencakup pembersihan data dan penanganan nilai yang hilang. Selanjutnya, dilakukan penyimbangan data menggunakan teknik SMOTE-Tomek untuk mengatasi ketidakseimbangan kelas dalam *dataset*. Setelah itu, *dataset* dibagi menjadi dua bagian: data pelatihan (*training*) dan data pengujian (*testing*). Tahap berikutnya adalah klasifikasi menggunakan dua algoritma *machine learning*, yaitu XGBoost dan LightGBM, untuk memprediksi risiko diabetes. Evaluasi kinerja model dilakukan menggunakan metrik seperti akurasi. Untuk meningkatkan performa prediksi, dilakukan *ensemble learning* dengan menggabungkan hasil dari kedua model.



Gambar 1 Metode Penelitian

2.1 Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari data medis dan demografis sekitar 100.000 pasien, mencakup status diabetes positif maupun negatif, dan tersedia secara publik di

Kaggle. Dataset ini berisi berbagai fitur klinis dan gaya hidup yang relevan untuk prediksi diabetes, seperti jenis kelamin, usia, riwayat hipertensi, penyakit jantung, kebiasaan merokok, indeks massa tubuh (BMI), kadar HbA1c, serta kadar glukosa darah. Berdasarkan studi sebelumnya oleh Alam et al. (2014), fitur-fitur ini terbukti memiliki korelasi yang signifikan terhadap risiko diabetes dan sering digunakan sebagai indikator utama dalam model prediksi medis berbasis machine learning.

Tabel 1 Struktur Dataset

	Gender	Age	Hypertension	Heart_disease	Smoking_history	Bmi	hbA1 Level	Blood_Glucose_Level	Diabetes
Count	961	961	961	961	961	961	961	961	961
Mean	0.41	41.7	0.07	0.04	2.23	27.	5.53	138	0.08
Std	0.49	22.4	0.26	0.19	1.87	6.7	1.07	40.9	0.28
Min	0.00	0.00	0.00	0.00	0.00	10	3.50	80.0	0.00
Max	2.00	80.0	1.00	1.00	5.00	95	9.00	300	1.00

Pada Tabel 1 terdapat gambaran lengkap mengenai faktor-faktor risiko yang berkaitan dengan diabetes mellitus, termasuk data medis dan demografis seperti usia, indeks massa tubuh, tekanan darah, kadar glukosa, dan riwayat keluarga. Informasi ini mencerminkan keragaman atribut yang relevan dalam konteks diagnosis dan klasifikasi diabetes, sehingga menyediakan landasan yang kuat untuk proses pelatihan model pembelajaran mesin dalam studi ini.

2.2 Pemrosesan Data

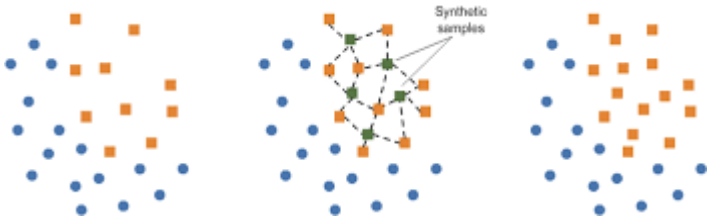
Pemrosesan data adalah langkah penting dalam mempersiapkan data sebelum digunakan untuk membangun model pembelajaran mesin (Galicia-garcia et al., 2020). Proses ini melibatkan serangkaian tahapan untuk membersihkan, mengubah, dan memformat data agar sesuai dengan kebutuhan model (Kharis & Zili, 2022). Langkah pertama adalah penanganan data yang hilang, dimana nilai yang hilang pada dataset diatasi dengan cara mengganti, menghapus, atau memperkirakan nilai yang hilang berdasarkan data yang ada. Selanjutnya, dilakukan normalisasi atau standardisasi untuk memastikan bahwa fitur-fitur numerik berada dalam skala yang sama, sehingga model tidak terpengaruh oleh perbedaan skala antar fitur. Kemudian, penanganan data tidak seimbang dilakukan, terutama jika dataset memiliki distribusi kelas yang tidak merata, menggunakan teknik seperti SMOTE atau undersampling untuk menyeimbangkan jumlah data antar kelas. Setelah itu, dilakukan encoding variabel kategorikal untuk mengubah data kategorikal menjadi format yang dapat diproses oleh model, seperti one-hot encoding atau label encoding. Langkah terakhir adalah pembagian dataset menjadi data pelatihan (training) dan data pengujian (testing) untuk memastikan bahwa model dapat dievaluasi secara efektif. Semua tahapan preprocessing ini bertujuan untuk meningkatkan kualitas data dan memaksimalkan kinerja model dalam melakukan prediksi.

2.3 Smote-Tomeks

Dalam penelitian ini, SMOTE (Synthetic Minority Over-sampling Technique) dan Tomek Links digunakan sebagai metode untuk mengatasi permasalahan ketidakseimbangan kelas dalam dataset (Fareed et al., 2022). SMOTE berfungsi dengan menambahkan sampel sintetis pada kelas minoritas untuk meningkatkan representasi data, sehingga model pembelajaran tidak bias terhadap kelas mayoritas. Sementara itu, Tomek Links digunakan untuk melakukan undersampling dengan menghapus pasangan data yang saling berdekatan namun berasal dari kelas berbeda, guna memperjelas batas keputusan antar kelas dan meningkatkan performa model secara keseluruhan (Muljono et al., 2024).

Gambar 2 menunjukkan teknik Smote-Tomeks bekerja dengan mengambil sampel acak dari kelas minoritas, mencari tetangga terdekatnya, kemudian menghasilkan data sintetis dengan mengurangkan sampel dari tetangga terdekat dan mengalikannya dengan bilangan acak. Sebaliknya, *Tomek Links* adalah strategi undersampling yang mengurangi jumlah sampel di kelas

mayoritas. Metode ini menemukan pasangan sampel dari kelas yang berbeda yang merupakan tetangga terdekat satu sama lain (Wang et al., 2019). Pasangan sampel ini, yang dikenal sebagai *Tomek Links*, kemudian dieliminasi dari kelas mayoritas.



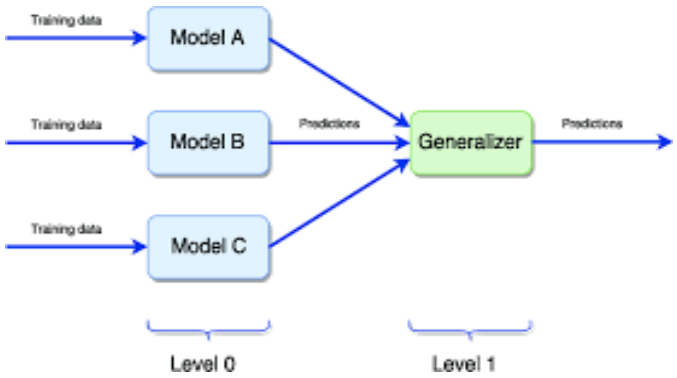
Gambar 2 Smote-Tomeks

2.4 Pembagian Data

Langkah berikutnya adalah membagi data menjadi subset pelatihan dan pengujian menggunakan metode stratifikasi label dengan rasio 80:20 setelah menyelesaikan prapemrosesan data dan penanganan ketidak seimbangan data. Mempersiapkan data untuk model pembelajaran mesin adalah langkah krusial. Train-Test Split memisahkan data yang telah diproses menjadi dua subset: data pelatihan (train) dan data pengujian (test) (Kahloot & Ekler, 2021). Dengan rasio 80:20, 80% dari data digunakan untuk melatih model, sementara 20% sisanya digunakan untuk menguji efektivitas model. Metode stratifikasi label memastikan bahwa distribusi kelas dalam dataset asli tetap terjaga di kedua subset, sehingga data pengujian secara akurat mewakili semua kelas.

2.5 Teknik Boosting

Boosting adalah model *ensemble* yang dibentuk dari beberapa model dasar yang secara berurutan model dasar tersebut dilatih dan digabungkan dalam prediksi. Algoritma boosting membangun model secara bertahap dengan mengoptimalkan suatu fungsi kerugian (Manconi et al., 2022). Pada Gambar 3 dijelaskan bahwa metode boosting menggunakan data masukan untuk melatih model boosting lemah, kesalahan pada klasifikasi, dan melatih model tersebut dengan set yang diterapkan pada klasifikasi sebelumnya peneliti lain menyatakan bahwa metode *boosting* ini berfokus pada pengurangan bias daripada variasi dengan meningkatkan boosting awal dasar yang dimiliki bias tinggi. Secara dalam analisis data prediktif dan meningkatkan kinerja model pada data kompleks atau sulit diklasifikasikan dan dapat bekerja baik dengan berbagai algoritma.

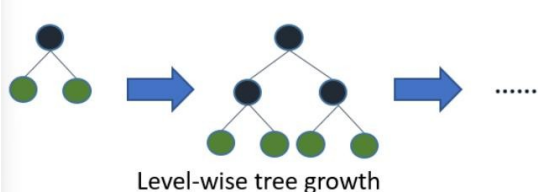


Gambar 3 Teknik Boosting

2.6 Teknik XGBoost

XGBoost merupakan implementasi dari *Gradient Boosting* untuk meningkatkan performa kinerja dan stabilitas, seperti yang ditunjukkan pada Gambar 4. Pada kasus klasifikasi dan regresi,

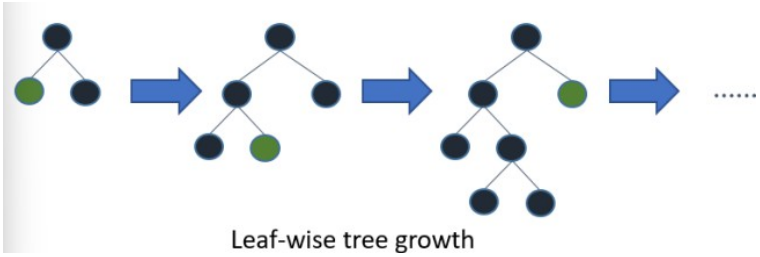
algoritma *XGBoost* tepat diusulkan karena mengacu pada pohon keputusan terbaik. Menurut Jafarzadeh et al menyatakan bahwa algoritma *XGBoost* adalah algoritma terbaik jika dibandingkan dengan algoritma ML lainnya (Sepbriant & Utomo, 2024). *XGBoost* menggunakan pohon keputusan sebagai model dasar dan menerapkan teknik boosting untuk meningkatkan kinerja model secara bertahap. Keunggulan utama *XGBoost* termasuk regularisasi untuk mencegah *overfitting*, *parallel processing* untuk mempercepat pelatihan model, dan kemampuan untuk menangani nilai yang hilang secara otomatis. Metode ini telah terbukti sangat efektif dalam berbagai kompetisi dan aplikasi machine learning, memberikan hasil prediksi yang akurat dan efisien (Azmi & Baliga, 2020).



Gambar 4 Teknik XGBoost

2.7 Teknik LightGBM

Teknik *LightGBM* dirancang untuk meningkatkan efisiensi model dan mengurangi penggunaan memori, seperti yang ditunjukkan pada Gambar 5. Dibandingkan dengan metode lainnya, *LightGBM* mengimplementasikan dua teknik inovatif, yaitu *Gradient-based One Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) (Sari et al., 2023). GOSS berfokus pada sampling data yang lebih relevan, mengurangi jumlah data yang perlu diproses, sementara EFB menggabungkan fitur eksklusif untuk mengurangi dimensi dan mempercepat proses pelatihan (Machado et al., 2019). Teknik-teknik ini memungkinkan *LightGBM* untuk bekerja lebih cepat dan lebih efisien, terutama dalam menangani dataset besar, sekaligus menjaga akurasi yang tinggi. *LightGBM* sangat cocok digunakan dalam berbagai aplikasi *machine learning*, terutama yang membutuhkan kecepatan dan skalabilitas.



Gambar 5 Teknik LightGBM

2.8 Matrix Evaluasi

Matrix Evaluasi adalah sebuah alat dalam pembelajaran mesin dan statistik yang digunakan untuk menilai kinerja algoritma klasifikasi (Zhang et al., 2021). Ini adalah matriks persegi yang sering digunakan untuk merangkum hasil dari. suatu masalah klasifikasi. Matriks kebingungan memberikan rincian terperinci tentang prediksi yang benar dan salah yang dibuat oleh model klasifikasi. Alat ini sangat berguna untuk masalah klasifikasi biner (dua kelas), namun juga dapat digunakan untuk klasifikasi multi-kelas (Thohari et al., 2024). Dalam matriks kebingungan, terdapat empat istilah yang menggambarkan hasil proses klasifikasi. Seperti yang terlihat pada Tabel II, istilah-istilah tersebut adalah True Positive (TP), False Positive (FP), True Negative (TN), dan False Negative (FN). TP dan TN menunjukkan hasil klasifikasi yang benar, sedangkan FP dan FN menunjukkan hasil klasifikasi yang salah. Rumus-rumus yang digunakan dalam metrik evaluasi ditunjukkan pada Tabel 2.

Tabel 2 Matrix Evaluasi

Precision	Sensitivity	F1-Score	Accuracy
$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{Precision \times Sensitivity}{Precision + Sensitivity}$	$\frac{TP+TN}{TP+TN+FP+FN}$

Akurasi digunakan untuk menghitung proporsi total prediksi yang benar terhadap keseluruhan data, sehingga memberikan gambaran umum tentang seberapa tepat model dalam mengklasifikasikan semua kelas. Sementara itu, sensitivitas atau recall digunakan untuk mengevaluasi sejauh mana model mampu mengenali kelas positif dengan benar, yang sangat penting ketika fokus utama adalah meminimalkan kesalahan dalam mendeteksi kasus positif. Di sisi lain, F1-score digunakan untuk menyeimbangkan antara presisi dan sensitivitas, dan menjadi metrik yang sangat berguna dalam situasi di mana distribusi kelas tidak seimbang, karena mempertimbangkan baik kesalahan positif maupun kesalahan negatif.

2.9 Ensemble Learning

Ensemble learning adalah pendekatan dalam pembelajaran mesin yang menggabungkan beberapa *weak learners* untuk membentuk satu model yang lebih kuat dengan kinerja prediktif yang lebih baik. Sebuah *weak learner* merupakan model yang performanya hanya sedikit lebih baik dibandingkan dengan tebakan acak (Gomes et al., 2018). Tujuan utama dalam perancangan *ensemble* adalah memastikan bahwa setiap anggota dalam *ensemble* memiliki karakteristik kesalahan klasifikasi yang berbeda. Jika anggota-anggota *ensemble* cenderung melakukan kesalahan pada *instance* yang tidak saling tumpang tindih, maka kombinasi prediksi dapat saling melengkapi dan menghasilkan kinerja yang lebih tinggi dibandingkan model individual. Dalam *ensemble learning*, istilah seperti *kombinasi* atau *voting* digunakan untuk menggambarkan mekanisme penggabungan prediksi dari berbagai anggota untuk memperoleh prediksi akhir. Arsitektur *ensemble* merujuk pada susunan *klasifikator* di dalam sistem *ensemble*, sementara metode *voting* menentukan bagaimana hasil prediksi dari setiap anggota digunakan dalam proses pengambilan keputusan akhir (Kumar et al., 2022). Meskipun peningkatan keberagaman dalam *ensemble* tidak selalu menjamin perbaikan kinerja, pendekatan ini tetap menjadi fokus utama dalam penelitian pembelajaran mesin. Berbagai taksonomi dan metode baru terus dikembangkan untuk meningkatkan akurasi dan efektivitas model, baik dalam konteks pembelajaran *batch* maupun pada aliran data (*data streams*).

3. HASIL DAN PEMBAHASAN

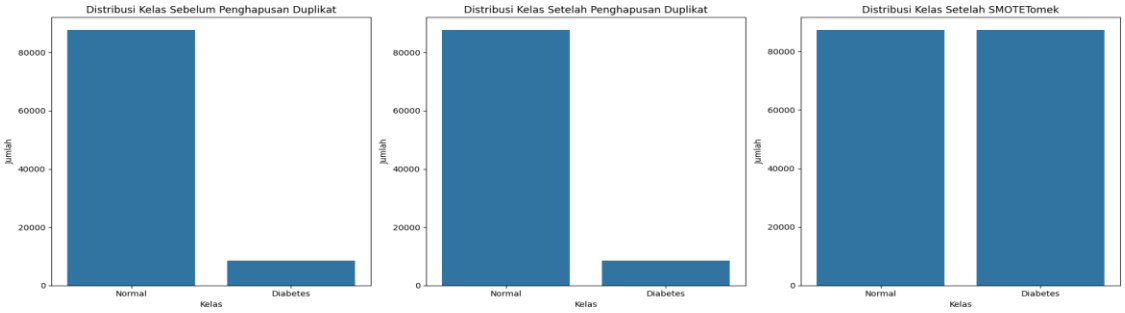
Dala penelitian ini, tahap pertama yang dilakukan adalah preprocessing dataset, yang dimulai dengan menghapus data duplikat. Tabel 3 menunjukkan bahwa pada tahap awal, dataset yang digunakan mengandung sejumlah data duplikat. Adanya duplikasi dalam dataset dapat memberikan pengaruh yang signifikan terhadap hasil analisis, karena data yang sama diulang-ulang dapat memberikan bobot yang tidak proporsional pada model, yang akhirnya menurunkan akurasi model. Oleh karena itu, langkah pertama yang diambil adalah menghapus data duplikat. Berdasarkan hasil yang didapat, dataset yang memiliki 100.000 entri awal, setelah penghapusan data duplikat, berkurang menjadi 96.146 entri. Hal ini memastikan bahwa model yang dibangun menggunakan data yang bersih dan lebih representatif. Selanjutnya dolakukan tranformasi data. Pada tahap transformating data kolom kategorikal smoking history dalam data frame diubah menjadi format numerik, sehingga data dapat digunakan dalam analisis lebih lanjut. Pada tabel menunjukkan data smoking history telah diubah menjadi format numerik.

Tabel 3 Penghapusan Data Duplikat

Jumlah Data	Data Duplikat	Setelah Penghapusan Data
100000	3854	96146

Dataset yang digunakan dalam penelitian ini awalnya berjumlah 100,000 entri, yang terdiri dari dua kelas, yaitu *Normal* dan *Diabetes*. Dari keseluruhan data, sebanyak 91,500 entri

dikategorikan sebagai *Normal* dan 8,500 entri lainnya diklasifikasikan sebagai *Diabetes*. Ketidakseimbangan kelas ini dapat menyebabkan model yang dibangun lebih cenderung untuk mengklasifikasikan data ke kelas mayoritas (*Normal*), sehingga berpotensi menurunkan performa model dalam mendeteksi kelas minoritas (*Diabetes*). Untuk mengatasi hal ini, langkah pertama yang dilakukan adalah menghapus data duplikat. Pada tahap ini, data duplikat diidentifikasi dan dihapus, menghasilkan 96,146 entri data yang unik. Penghapusan data duplikat sangat penting untuk memastikan bahwa setiap entri dalam dataset hanya mewakili informasi yang valid, yang akan mencegah distorsi dalam analisis dan model pembelajaran mesin.



Gambar 6 Preprocessing Dataset

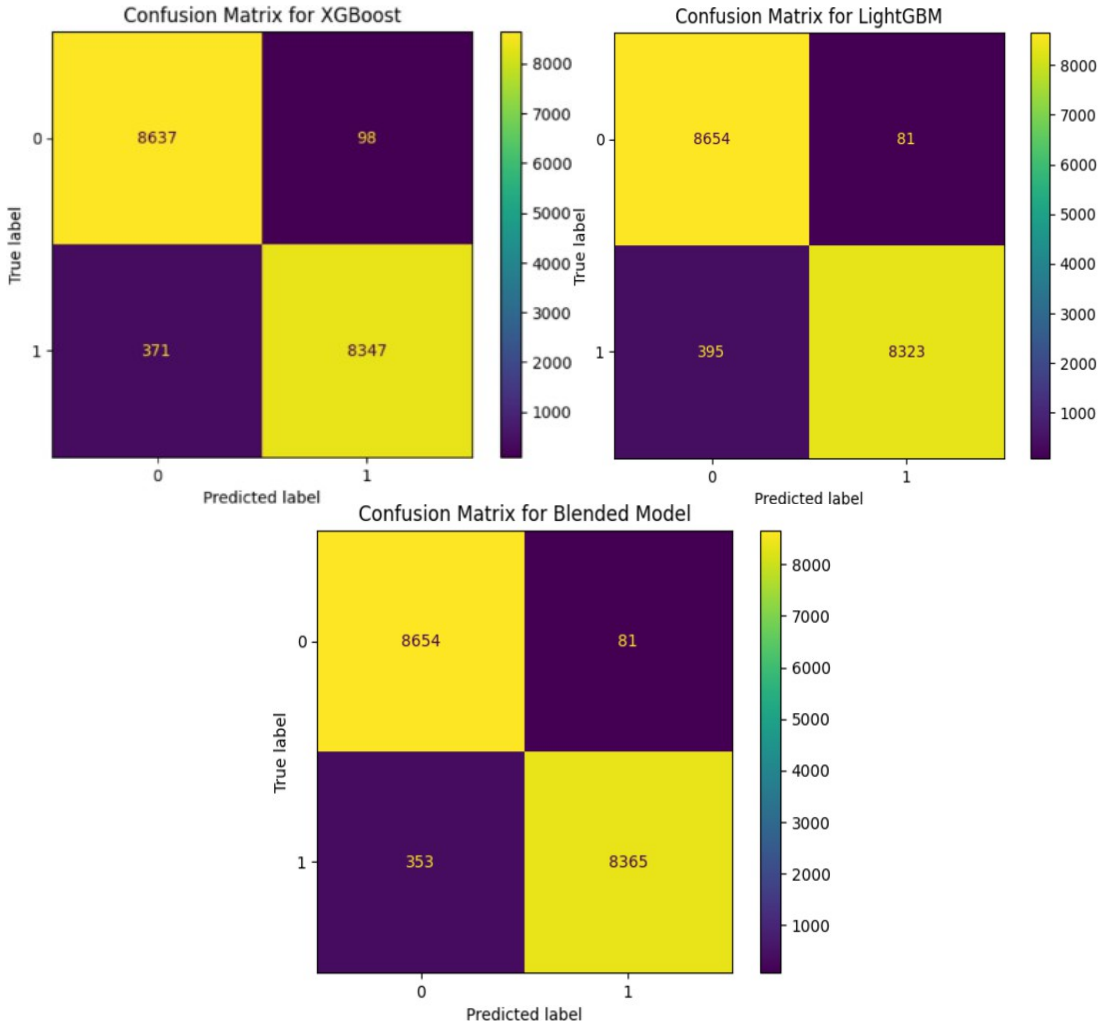
Setelah data duplikat berhasil dihapus, langkah selanjutnya adalah penyeimbangan dataset, karena terdapat ketidakseimbangan antara jumlah data untuk kelas *Normal* dan *Diabetes*. Ketidakseimbangan kelas ini dapat mempengaruhi akurasi model, karena model cenderung lebih banyak mengklasifikasikan data ke kelas yang lebih banyak jumlahnya. Untuk mengatasi masalah ini, diterapkan teknik SMOTETomek, yang menggabungkan dua pendekatan: *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Tomek links*. SMOTE menghasilkan sampel sintetis dari kelas minoritas (*Diabetes*), sementara teknik Tomek links berfungsi untuk menghapus data yang tumpang tindih atau saling mendekati antara kelas mayoritas dan kelas minoritas, yang dapat mengurangi potensi kesalahan dalam klasifikasi.

Hasil dari penerapan SMOTETomek adalah terbentuknya distribusi kelas yang lebih seimbang, dengan masing-masing kelas (*Normal* dan *Diabetes*) memiliki 87,269 entri data. Dengan distribusi data yang seimbang, model yang dibangun diharapkan dapat memprediksi kedua kelas dengan lebih akurat, serta mengurangi bias yang mungkin terjadi terhadap kelas mayoritas. Langkah-langkah ini bertujuan untuk memaksimalkan kinerja model pembelajaran mesin dalam mendeteksi diabetes secara efektif dan tepat. Setelah proses persiapan data selesai, pelatihan model dilakukan menggunakan tiga pendekatan: XGBoost, LightGBM, dan Blended Model (gabungan XGBoost dan LightGBM). Evaluasi kinerja dilakukan menggunakan metrik akurasi, *precision*, *sensitivity* (recall), dan *F1- score*.

Gambar 7 dan Tabel 4 menunjukkan hasil evaluasi dari masing-masing model. Model XGBoost menunjukkan performa yang cukup baik, dengan akurasi sebesar 97,31%, *precision* sebesar 98,84%, *sensitivity* sebesar 95,74%, dan *F1-score* sebesar 97,27%. Berdasarkan *confusion matrix*, model ini mengklasifikasikan dengan benar 8.637 data sebagai Normal dan 8.347 sebagai Diabetes, sementara terdapat 98 kesalahan klasifikasi untuk data Normal (*false positive*) dan 371 kesalahan untuk data Diabetes (*false negative*).

Model LightGBM menunjukkan hasil yang sedikit berbeda, dengan akurasi sebesar 97,26%, *precision* sebesar 99,08%, *sensitivity* sebesar 95,46%, dan *F1-score* sebesar 97,23%. Model ini menghasilkan 81 *false positive* dan 395 *false negative*, serta mengklasifikasikan 8.654 data sebagai Normal dan 8.323 sebagai Diabetes. Meskipun tingkat akurasi sedikit lebih rendah dibandingkan XGBoost, nilai *precision*-nya lebih tinggi, menunjukkan kemampuan yang baik dalam menghindari kesalahan klasifikasi untuk kelas minoritas.

Model terbaik diperoleh dari pendekatan *ensemble* melalui Blended Model yang menggabungkan XGBoost dan LightGBM. Model ini mencatat akurasi tertinggi sebesar 97,51%, dengan *precision* sebesar 98,95%, *sensitivity* sebesar 96,00%, dan *F1-score* sebesar 97,46%. Berdasarkan hasil *confusion matrix*, model ini berhasil mengklasifikasikan 8.654 data sebagai Normal dan 8.365 sebagai Diabetes, dengan hanya 81 *false positive* dan 353 *false negative*.



Gambar 7 Hasil Evaluasi

Tabel 4 Evaluated Models

Model	Class	Precision	Recall	F1-Score	Accuracy
XGBoost	Predicted	0.9884	0.9574	0.9727	0.9731
LightGBM	Predicted	0.9908	0.9546	0.9723	0.9726
Ensemble	Predicted	0.9895	0.9600	0.9746	0.9751

4. KESIMPULAN

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model machine learning dalam mendeteksi diabetes mellitus menggunakan algoritma XGBoost dan LightGBM, dua model pembelajaran mesin yang sering digunakan untuk klasifikasi data medis. Diabetes mellitus adalah penyakit metabolik yang berdampak besar terhadap kesehatan masyarakat, dan deteksi dini sangat penting untuk mengurangi dampak buruknya. Melalui penggunaan dataset yang mencakup berbagai fitur terkait faktor risiko diabetes, penelitian ini menguji efektivitas algoritma

XGBoost dan LightGBM dalam mendeteksi dan mengklasifikasikan data diabetes dan normal. Hasil eksperimen menunjukkan bahwa kedua model ini memiliki kinerja yang sangat baik dalam hal akurasi prediksi. Model XGBoost mencatatkan akurasi sebesar 97,31%, sedangkan LightGBM menghasilkan akurasi 97,26%.

Penerapan teknik SMOTE-Tomek untuk menangani ketidakseimbangan data menjadi salah satu faktor penting dalam meningkatkan kinerja model. Dalam dataset asli, terdapat ketidakseimbangan antara jumlah data pasien yang terdiagnosis diabetes dan yang tidak, yang berpotensi menyebabkan model kurang sensitif dalam mendeteksi kelas minoritas (diabetes). Dengan menerapkan SMOTE-Tomek, data yang tidak seimbang dapat diubah menjadi lebih seimbang, sehingga model dapat belajar lebih baik dalam membedakan antara kedua kelas. Proses preprocessing ini terbukti meningkatkan kemampuan model dalam mengenali pola-pola penting yang membedakan data diabetes dan normal, serta menghasilkan data pelatihan yang lebih representatif.

Selanjutnya, penggabungan kedua model ini melalui teknik ensemble atau blending menghasilkan peningkatan kinerja yang signifikan, dengan akurasi mencapai 97,51%. Hasil ini menunjukkan bahwa kombinasi dari kedua algoritma dapat meningkatkan kemampuan prediksi model dibandingkan jika hanya menggunakan satu model tunggal. Teknik ensemble memanfaatkan kelebihan masing-masing model, dengan mengurangi kelemahan yang ada, sehingga menghasilkan keputusan yang lebih baik. Penilaian model dilakukan menggunakan metode K-Fold cross-validation, yang memberikan hasil yang lebih stabil dan menghindari bias evaluasi yang bisa timbul jika hanya menggunakan satu data split. Dengan pendekatan ini, penelitian ini berhasil menghasilkan model yang dapat memberikan prediksi diabetes dengan tingkat keakuratan yang sangat tinggi.

Dari hasil penelitian ini, dapat disimpulkan bahwa penerapan algoritma XGBoost dan LightGBM, yang dipadukan dengan teknik SMOTE-Tomek dan ensemble, memberikan hasil yang sangat memuaskan dalam mendeteksi diabetes mellitus. Kombinasi model ini terbukti tidak hanya mampu meningkatkan akurasi, tetapi juga mengurangi kesalahan klasifikasi, seperti False Positives dan False Negatives. Oleh karena itu, model ini memiliki potensi besar untuk diimplementasikan dalam sistem pendukung keputusan medis yang dapat membantu dalam diagnosis dini diabetes mellitus secara lebih akurat dan efisien. Selain itu, penelitian ini juga memberikan wawasan baru mengenai pentingnya teknik preprocessing dan ensemble dalam meningkatkan kinerja model pembelajaran mesin, khususnya dalam konteks data medis yang sering kali tidak seimbang. Sebagai arah pengembangan penelitian selanjutnya, disarankan untuk mengeksplorasi pendekatan optimasi parameter (hyperparameter tuning) secara lebih mendalam serta penerapan metode feature selection berbasis algoritma evolusioner guna meningkatkan efisiensi dan kinerja model. Selain itu, penggunaan algoritma deep learning seperti deep neural networks atau transformer-based models dapat dipertimbangkan untuk mengakomodasi kompleksitas hubungan antar fitur yang tidak linier. Di samping itu, penelitian lanjutan sebaiknya tidak hanya terbatas pada tugas klasifikasi statis, tetapi juga mencakup prediksi longitudinal terkait progresi diabetes dan efektivitas intervensi medis. Pendekatan explainable AI (XAI) juga direkomendasikan agar hasil model dapat diterjemahkan secara interpretatif, sehingga mendukung transparansi dan kepercayaan dalam implementasi klinis, untuk lebih meningkatkan performa dalam deteksi dan prediksi diabetes mellitus.

DAFTAR PUSTAKA

Alam, U., Asghar, O., Azmi, S., & Malik, R. A. (2014). General Aspects of Diabetes Mellitus. In *Handbook of Clinical Neurology* (Vol. 4, pp. 211–222). https://doi.org/10.1016/B978-0-444-53480-4.00015-1

Azmi, S. S., & Baliga, S. (2020). An Overview of Boosting Decision Tree Algorithms Utilizing AdaBoost and XGBoost Boosting Strategies. *International Research Journal of Engineering and Technology*, *7*(5), 6867–6870. https://www.irjet.net/archives/V7/i5/IRJET-V7I51293.pdf

Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *Journal of Healthcare Engineering*, *2021*, 1–17. https://doi.org/10.1155/2021/9930985

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms. *Neural Computing and Applications*, *35*(22), 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

Fareed, M. M. S., Zikria, S., Ahmed, G., Mui-Zzud-Din, Mahmood, S., Aslam, M., Jillani, S. F., Moustafa, A., & Asad, M. (2022). ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans. *IEEE Access*, *10*, 96930–96951. https://doi.org/10.1109/ACCESS.2022.3204395

Galicia-garcia, U., Benito-vicente, A., Jebari, S., & Larrea-sebal, A. (2020). Costus Ignus: Insulin Plant and It's Preparations as Remedial Approach for Diabetes Mellitus. *International Journal of Molecular Sciences*, 1–34. https://doi.org/10.13040/IJPSR.0975-8232.13(4).1551-58

Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. (2018). A Survey on Ensemble Learning for Data Stream Classification. *ACM Computing Surveys*, *50*(2), 1–36. https://doi.org/10.1145/3054925

Kahloot, K. M., & Ekler, P. (2021). Algorithmic Splitting: A Method for Dataset Preparation. *IEEE Access*, *9*, 125229–125237. https://doi.org/10.1109/ACCESS.2021.3110745

Kharis, S. A. A., & Zili, A. H. A. (2022). Learning Analytics dan Educational Data Mining pada Data Pendidikan. *Jurnal Riset Pembelajaran Matematika Sekolah*, *6*(1), 12–20. https://doi.org/10.21009/jrpms.061.02

Kumar, M., Singhal, S., Shekhar, S., Sharma, B., & Srivastava, G. (2022). Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning. *Sustainability*, *14*(21), Article ID: 13998. https://doi.org/10.3390/su142113998

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive Models for Diabetes Mellitus Using Machine Learning Techniques. *BMC Endocrine Disorders*, *19*(1), Article ID: 101. https://doi.org/10.1186/s12902-019-0436-6

Machado, M. R., Karray, S., & de Sousa, I. T. (2019). LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry. *2019 14th International Conference on Computer Science & Education (ICCSE)*, 1111–1116. https://doi.org/10.1109/ICCSE.2019.8845529

Manconi, A., Armano, G., Gnocchi, M., & Milanesi, L. (2022). A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19. *Applied Sciences*, *12*(15), Article ID: 7554. https://doi.org/10.3390/app12157554

Mengcan, M., Xiaofang, C., & Yongfang, X. (2021). Constrained Voting Extreme Learning Machine and Its Application. *Journal of Systems Engineering and Electronics*, *32*(1), 209–219. https://doi.org/10.23919/JSEE.2021.000018

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction Using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047

Muljono, Wulandari, S. A., Azies, H. Al, Naufal, M., Prasetyanto, W. A., & Zahra, F. A. (2024). Breaking Boundaries in Diagnosis: Non-Invasive Anemia Detection Empowered by AI. *IEEE Access*, *12*(2023), 9292–9307. https://doi.org/10.1109/ACCESS.2024.3353788

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. E., & Makaroff, L. E. (2017). IDF Diabetes Atlas: Global Estimates for the Prevalence of Diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, *128*, 40–50. https://doi.org/10.1016/j.diabres.2017.03.024

Rif'at, I. D., Hasneli N, Y., & Indriati, G. (2023). Gambaran Komplikasi Diabetes Melitus pada Penderita Diabetes Melitus. *Jurnal Keperawatan Profesional*, *11*(1), 52–69. https://doi.org/10.33650/jkp.v11i1.5540

Sari, L., Romadloni, A., Lityaningrum, R., & Hastuti, H. D. (2023). Implementation of LightGBM and Random Forest in Potential Customer Classification. *TIERS Information Technology Journal*, *4*(1), 43–55. https://doi.org/10.38043/tiers.v4i1.4355

Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods.

Computational Intelligence and Neuroscience, *2022*(2), 1–11. https://doi.org/10.1155/2022/3820360

Sepbriant, G. D., & Utomo, D. W. (2024). Ensemble Learning pada Kategorisasi Produk E-Commerce Menggunakan Teknik Boosting. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(2), 123–133. https://doi.org/10.14421/jiska.2024.9.2.123-133

Tanwar, A., & Bhatia, P. K. (2024). A Review on Diabetes Prediction Using Machine Learning Techniques. In *Lecture Notes in Electrical Engineering* (Vol. 1185, Number 09, pp. 513–524). https://doi.org/10.1007/978-981-97-1682-1_41

Thohari, A. N. A., Karima, A., Santoso, K., & Rahmawati, R. (2024). Crack Detection in Building Through Deep Learning Feature Extraction and Machine Learning Approch. *Journal of Applied Informatics and Computing*, *8*(1), 1–6. https://doi.org/10.30871/jaic.v8i1.7431

Wang, Z., Wu, C., Zheng, K., Niu, X., & Wang, X. (2019). SMOTETomek-Based Resampling for Personality Recognition. *IEEE Access*, *7*, 129678–129689. https://doi.org/10.1109/ACCESS.2019.2940061

Zhang, H., Liu, C., Zhang, Z., Xing, Y., Liu, X., Dong, R., He, Y., Xia, L., & Liu, F. (2021). Recurrence Plot-Based Approach for Cardiac Arrhythmia Classification Using Inception-ResNet-v2. *Frontiers in Physiology*, *12*, 1–13. https://doi.org/10.3389/fphys.2021.648950