

Deteksi Dini Diabetes Melitus Tipe 2 Berdasarkan Data Rekam Medis Pasien RSUD Pintu Padang Kabupaten Tapsel Menggunakan Algoritma Random Forest

Adzkia Nur Nasution^{1*}, Zulfahmi Indra²

^{1,2}Matematika dan Ilmu Pengetahuan Alam, Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia

Email: ^{1*}adzkiiaanur41@gmail.com, ²zulfahmi.indra@unimed.ac.id

(*Email Corresponding Author: adzkiiaanur41@gmail.com)

Received: 3 Juli 2026 | Revision: 10 Juli 2026 | Accepted: 13 Juli 2026

Abstrak

Diabetes melitus tipe 2 merupakan penyakit kronis yang kasusnya terus bertambah dari tahun ke tahun, termasuk di wilayah Tapanuli Selatan. Persoalan yang muncul di lapangan adalah banyaknya pasien yang baru mengetahui kondisinya setelah penyakit berkembang cukup jauh, sebagian karena sistem skrining yang ada masih bersifat konvensional dan bergantung sepenuhnya pada evaluasi klinis manual. Penelitian ini berupaya menjawab persoalan tersebut dengan membangun model deteksi dini berbasis machine learning menggunakan data rekam medis pasien RSUD Pintu Padang. Algoritma yang digunakan adalah random forest, sebuah pendekatan ensemble learning yang menggabungkan banyak pohon keputusan (decision tree) untuk menghasilkan prediksi yang lebih stabil dan akurat. Hasil evaluasi model menunjukkan kinerja yang cukup baik. Algoritma random forest menghasilkan nilai akurasi sebesar 80,82%, yang berarti model mampu mengklasifikasikan sebagian besar data secara tepat. Nilai presisi dan recall masing-masing berada di angka 75%, yang mencerminkan keseimbangan antara kemampuan model dalam mendeteksi kasus positif secara benar serta meminimalkan prediksi yang keliru. F1-score yang diperoleh juga sebesar 75%, mengonfirmasi bahwa model memiliki keseimbangan yang baik antara presisi dan recall dalam kondisi data yang tidak sepenuhnya seimbang. Penelitian ini menunjukkan bahwa pemanfaatan data rekam medis yang telah tersedia di fasilitas kesehatan daerah, jika diolah dengan pendekatan yang tepat, dapat menghasilkan alat bantu prediksi yang praktis dan bernilai klinis. Sistem yang dikembangkan diharapkan mampu mendukung pengambilan keputusan medis lebih awal serta berkontribusi pada peningkatan kualitas layanan kesehatan di tingkat daerah.

Kata Kunci: Diabetes Melitus Tipe 2, Deteksi Dini, Machine Learning, Random Forest, Rekam Medis, SMOTE, Streamlit

Abstract

Type 2 diabetes mellitus is a chronic disease whose cases continue to increase year after year, including in the South Tapanuli region. A problem that arises in the field is that many patients only discover their condition after the disease has progressed considerably, partly because the existing screening system is still conventional and relies entirely on manual clinical evaluation. This study seeks to address this issue by building a machine learning-based early detection model using patient medical records at Pintu Padang Regional Hospital. The algorithm used is random forest, an ensemble learning approach that combines multiple decision trees to produce more stable and accurate predictions. The model evaluation results showed quite good performance. The Random Forest algorithm produced an accuracy value of 80.82%, indicating the model was able to correctly classify most of the data. The precision and recall values each reached 75%, reflecting the balance between the model's ability to correctly detect positive cases and minimize false predictions. The F1-score also reached 75%, confirming that the model strikes a good balance between precision and recall in conditions with imbalanced data. This study demonstrates that utilizing medical record data readily available at regional health facilities, when processed with the right approach, can produce a practical and clinically valuable prediction tool. The developed system is expected to support earlier medical decision-making and contribute to improving the quality of healthcare services at the regional level.

Keywords: Type 2 Diabetes Mellitus, Early Detection, machine learning, random forest, Medical Records, SMOTE, Streamlit

1. PENDAHULUAN

Diabetes melitus merupakan salah satu penyakit kronis yang menjadi perhatian serius di dunia karena prevalensinya yang terus meningkat setiap tahun. Berdasarkan data *International Diabetes Federation* (IDF), Indonesia menempati peringkat ke-7 di dunia dengan jumlah penderita diabetes sekitar 10,7 juta jiwa pada tahun 2019, dan angka ini diproyeksikan meningkat menjadi sekitar 13,7 juta jiwa pada tahun 2030 [1]. Secara medis, diabetes melitus tergolong penyakit kronis yang menyerang organ pankreas, sehingga organ tersebut tidak mampu memproduksi insulin dalam

jumlah yang mencukupi kebutuhan tubuh [2]. Kondisi ini semakin diperburuk oleh faktor gaya hidup, salah satunya pola konsumsi makanan tidak sehat yang memicu peningkatan berat badan [3].

Di antara jenis-jenis diabetes yang ada, Diabetes Melitus Tipe 2 (DMT2) merupakan bentuk yang paling banyak ditemukan. Berbeda dengan Diabetes Melitus Tipe 1 yang bersifat autoimun dan umumnya terjadi pada usia muda, DMT2 ditandai oleh resistensi insulin serta disfungsi sel beta pankreas yang sangat berkaitan dengan faktor metabolik seperti obesitas dan hipertensi [4]. Penelitian terdahulu memperkirakan sekitar satu dari sebelas orang dewasa di dunia mengalami DMT2, dan sekitar 75% di antaranya tinggal di negara berkembang dengan sumber daya kesehatan yang masih terbatas [5]. Kondisi serupa juga terjadi di tingkat regional, khususnya di Provinsi Sumatera Utara, di mana faktor risiko seperti hipertensi dan obesitas berkontribusi terhadap meningkatnya kasus diabetes [6].

Permasalahan ini turut dirasakan di wilayah Tapanuli Selatan. Berdasarkan observasi lapangan di RSUD Pintu Padang Kabupaten Tapanuli Selatan, tantangan yang dihadapi bukan hanya peningkatan jumlah pasien DMT2, melainkan juga keterbatasan sistem deteksi dini yang masih bersifat konvensional dan sepenuhnya bergantung pada evaluasi klinis manual oleh tenaga medis. Akibatnya, banyak pasien baru terdiagnosis setelah penyakit memasuki tahap lanjut, yang ditandai dengan kadar glukosa darah tidak terkontrol dan meningkatnya risiko komplikasi kronis. Padahal, data rekam medis pasien sesungguhnya telah menyimpan berbagai variabel klinis penting seperti usia, jenis kelamin, berat badan, tekanan darah, dan riwayat hipertensi yang berpotensi memberikan informasi risiko secara lebih komprehensif apabila dianalisis secara terpadu [7]. Deteksi dini menjadi krusial karena upaya pencegahan penyakit terbukti jauh lebih mudah dan ekonomis dibandingkan pengobatan hiperglikemia beserta komplikasinya [8].

Seiring kemajuan teknologi, diagnosis diabetes kini dapat dilakukan dengan memanfaatkan pendekatan komputasional [9], khususnya *machine learning* yang telah memberi dampak signifikan dalam klasifikasi dan deteksi penyakit di bidang kesehatan [10]. Sebagai cabang penting dari kecerdasan buatan, *machine learning* memungkinkan sistem komputer belajar dari data historis tanpa perlu diprogram secara eksplisit untuk setiap kasus, sehingga mampu mendukung otomatisasi dengan meminimalkan keterlibatan manusia [11]. Oleh karena itu, dibutuhkan metode yang mampu melakukan prediksi sekaligus mengidentifikasi variabel-variabel paling berpengaruh terhadap risiko DMT2 secara akurat. *Random Forest* (RF) menjadi salah satu pilihan yang relevan karena kemampuannya mengklasifikasikan data meskipun terdapat amatan yang tidak lengkap (*missing value*) [12].

Random Forest merupakan pendekatan *ensemble learning* yang banyak diterapkan di bidang kesehatan berkat kemampuannya menangani data klinis yang kompleks dan multidimensi. Melalui mekanisme penggabungan sejumlah *decision tree*, algoritma ini terbukti mampu menurunkan varians, mencegah *overfitting*, serta meningkatkan akurasi diagnostik dibandingkan metode klasifikasi tunggal, sebagaimana ditunjukkan pada berbagai kasus prediksi penyakit jantung, diabetes melitus, dan kanker [13]. Studi oleh Buani (2024) yang membandingkan sembilan algoritma klasifikasi di antaranya *Logistic Regression*, *Support Vector Machine*, *Naïve Bayes*, *Decision Tree*, *K-Nearest Neighbors*, *AdaBoost*, *Gradient Boosting*, dan *XGBoost*—menemukan bahwa *Random Forest* memberikan performa terbaik dengan akurasi 98,72% dan nilai AUC 0,98 (kategori *excellent*), mengungguli *Gradient Boosting* dan *XGBoost* yang masing-masing berada di kisaran 97% [14]. Temuan serupa dilaporkan oleh Iskandar et al. (2024), yang menerapkan *Random Forest* untuk klasifikasi awal DMT2 menggunakan data sekunder dari Puskesmas Jatiroto dan memperoleh akurasi sebesar 97,48%. Kedua studi tersebut menguatkan bahwa *Random Forest* merupakan pendekatan yang efektif dan konsisten dalam permasalahan deteksi dini diabetes. [15]

Meskipun berbagai penelitian terdahulu telah menunjukkan performa *Random Forest* yang unggul, penerapannya pada data rekam medis dari fasilitas kesehatan daerah, khususnya rumah sakit di wilayah Tapanuli Selatan, masih belum banyak dikaji. Kesenjangan ini menjadi dasar bagi penelitian ini, yang bertujuan membangun model deteksi dini DMT2 berbasis *Random Forest* menggunakan data rekam medis pasien RSUD Pintu Padang Kabupaten Tapanuli Selatan, sekaligus mengembangkan sistem deteksi dini berbasis aplikasi web menggunakan Streamlit sebagai antarmuka. Melalui evaluasi kinerja model berdasarkan *confusion matrix*, akurasi, presisi, *recall*, *F1-score*, dan AUC, penelitian ini diharapkan dapat memberikan kontribusi ilmiah dalam pemanfaatan data rekam medis daerah sebagai alat bantu prediksi yang praktis, mendukung pengambilan keputusan klinis lebih awal, serta meningkatkan mutu pelayanan kesehatan di tingkat daerah.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode analisis data sekunder untuk membangun model deteksi dini Diabetes Melitus Tipe 2 menggunakan algoritma *Random Forest*. Pendekatan kuantitatif dipilih karena penelitian berfokus pada pengolahan data numerik, proses klasifikasi, serta pengukuran kinerja model secara objektif menggunakan berbagai metrik evaluasi. Seluruh proses analisis dilakukan melalui tahapan *machine learning* yang meliputi pengumpulan data, praproses data, pembangunan model, evaluasi performa, hingga implementasi model ke dalam aplikasi berbasis web.

Sumber data penelitian berasal dari data rekam medis pasien RSUD Pintu Padang Kabupaten Tapanuli Selatan yang telah terdokumentasi selama periode April 2023 hingga Februari 2026. Penggunaan data rekam medis dipilih karena memuat informasi klinis yang relevan untuk proses deteksi dini Diabetes Melitus Tipe 2, seperti usia, jenis kelamin, berat badan, tinggi badan, tekanan darah sistolik, dan tekanan darah diastolik. Seluruh data yang digunakan telah melalui proses anonimisasi sehingga identitas pasien tidak dapat dikenali, sesuai dengan prinsip kerahasiaan data kesehatan dan ketentuan perlindungan data pribadi yang berlaku.

Populasi penelitian terdiri atas 3.825 data rekam medis pasien. Dari populasi tersebut dipilih sebanyak 242 data menggunakan teknik random sampling, dengan komposisi 93 data pasien Diabetes Melitus Tipe 2 dan 149 data pasien non-diabetes. Teknik pengambilan sampel ini diterapkan agar setiap data yang memenuhi kriteria memiliki peluang yang sama untuk dipilih sebagai sampel penelitian sehingga dapat mengurangi potensi bias dalam proses pembentukan model klasifikasi.

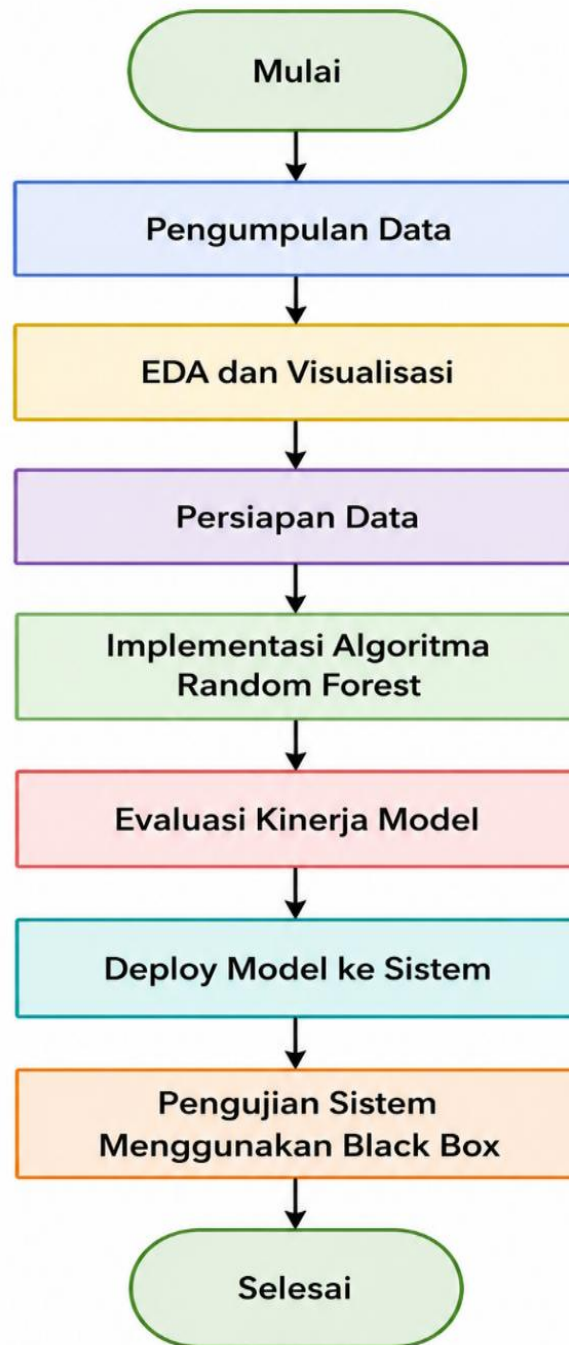
Variabel penelitian terdiri atas variabel independen (X) dan variabel dependen (Y). Variabel independen meliputi usia, jenis kelamin, tinggi badan, berat badan, indeks massa tubuh (Body Mass Index/BMI), tekanan darah sistolik, dan tekanan darah diastolik. Sementara itu, variabel dependen berupa status Diabetes Melitus Tipe 2 yang diklasifikasikan ke dalam dua kelas, yaitu pasien diabetes (1) dan pasien non-diabetes (0). Seluruh variabel tersebut dipilih berdasarkan pertimbangan klinis serta hasil penelitian terdahulu yang menunjukkan keterkaitannya terhadap risiko terjadinya Diabetes Melitus Tipe 2. Rincian variabel penelitian disajikan pada Tabel 1..

Tabel 1. Variabel Penelitian

Variabel Penelitian	Keterangan	Kategori
Kelas Pasien (y)	Label sebagai indikator status pasien	$y = \begin{cases} 1 & \text{diabetes} \\ 0 & \text{non diabetes} \end{cases}$
Karakteristik Pasien (x)	Berbagai fitur seperti usia, jenis kelamin, tinggi badan, berat badan, sistolik, diastolik	x = fitur numerik dan kategorikal

2.1 Tahapan Penelitian

Alur kerja penelitian ini mengadaptasi tahapan umum proyek *machine learning*, yakni: pengumpulan data, eksplorasi sekaligus pembersihan data (*Exploratory Data Analysis*), praproses data mencakup transformasi, pemilihan variabel, penanganan outlier, pembagian data latih-uji, dan penyeimbangan kelas lewat *Synthetic Minority Over-sampling Technique* (SMOTE) pembentukan model klasifikasi dengan algoritma *Random Forest*, evaluasi performa memakai *confusion matrix*, akurasi, presisi, *recall*, F1-score, dan *Area Under Curve* (AUC), lalu diakhiri dengan penerapan model ke sistem deteksi dini berbasis web. Rangkaian tahapan tersebut divisualisasikan pada Gambar 1.



Gambar 1. Alur tahapan penelitian

Pengambilan data rekam medis dilakukan melalui observasi dokumen, diawali dengan pengurusan izin resmi ke pihak rumah sakit. Identitas pasien disamarkan lewat proses anonimisasi mengikuti Peraturan Menteri Kesehatan Nomor 269/MENKES/PER/III/2008 tentang Rekam Medis serta Undang-Undang Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi [16]. Masuk ke tahap praproses, nilai kosong (*missing value*) pada variabel tinggi badan, berat badan, sistolik, dan diastolik diisi memakai nilai median, sebab median dinilai lebih tahan terhadap pengaruh nilai ekstrem dibanding rata-rata. Deteksi pencilan (*outlier*) memanfaatkan metode *Interquartile Range* (IQR), lalu ditangani lewat teknik *winsorizing* nilai yang berada di luar rentang wajar dipangkas ke batas atas atau batas bawah, bukan dihapus, sehingga informasi dalam dataset tetap utuh. Model dievaluasi menggunakan confusion matrix dan beberapa metrik evaluasi seperti: Accuracy, Precision, Recall, F1-score, ROC-AUC.

Dataset selanjutnya disekat menjadi data latih dan data uji dengan proporsi 70:30 melalui fungsi *train_test_split* dari *scikit-learn*, menghasilkan 169 data latih dan 73 data uji. Parameter *stratify=y* diikutsertakan agar sebaran kelas pada kedua subset tetap proporsional, sementara *random_state=42* dipakai supaya hasil pembagian dapat direplikasi. Karena rasio kelas pada data latih masih timpang dengan *Imbalance Ratio* sebesar 1,60 (149 berbanding 93). SMOTE diterapkan

husus pada data latih guna mencegah kebocoran informasi ke data uji (*data leakage*). Model klasifikasi dibangun dengan algoritma *Random Forest* memakai konfigurasi $n_estimators = 300$, $max_depth = 15$, $class_weight = balanced$, dan $random_state = 42$. Secara prinsip, *Random Forest* menyusun sejumlah T pohon keputusan dari subset data hasil *bootstrap sampling* D_t , di mana tiap simpul (*node*) hanya mempertimbangkan sebagian fitur yang dipilih secara acak bukan seluruh fitur untuk menekan korelasi antar pohon. Pemilihan pemisahan terbaik pada tiap simpul didasarkan pada minimisasi indeks Gini:

$$Gini(S) = 1 - \sum_{k=1}^K p_k^2 \quad (1)$$

dengan p_k adalah proporsi data kelas ke- k pada subset S . Setelah keseluruhan $T = 300$ pohon terbentuk, keputusan akhir $\hat{y}$ diperoleh lewat mekanisme *majority voting*

$$\hat{y} = model\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (2)$$

Mekanisme *ensemble* semacam ini membuat *Random Forest* mampu menekan varians yang lazim muncul pada pohon keputusan tunggal, sehingga prediksinya cenderung lebih stabil dan akurat.

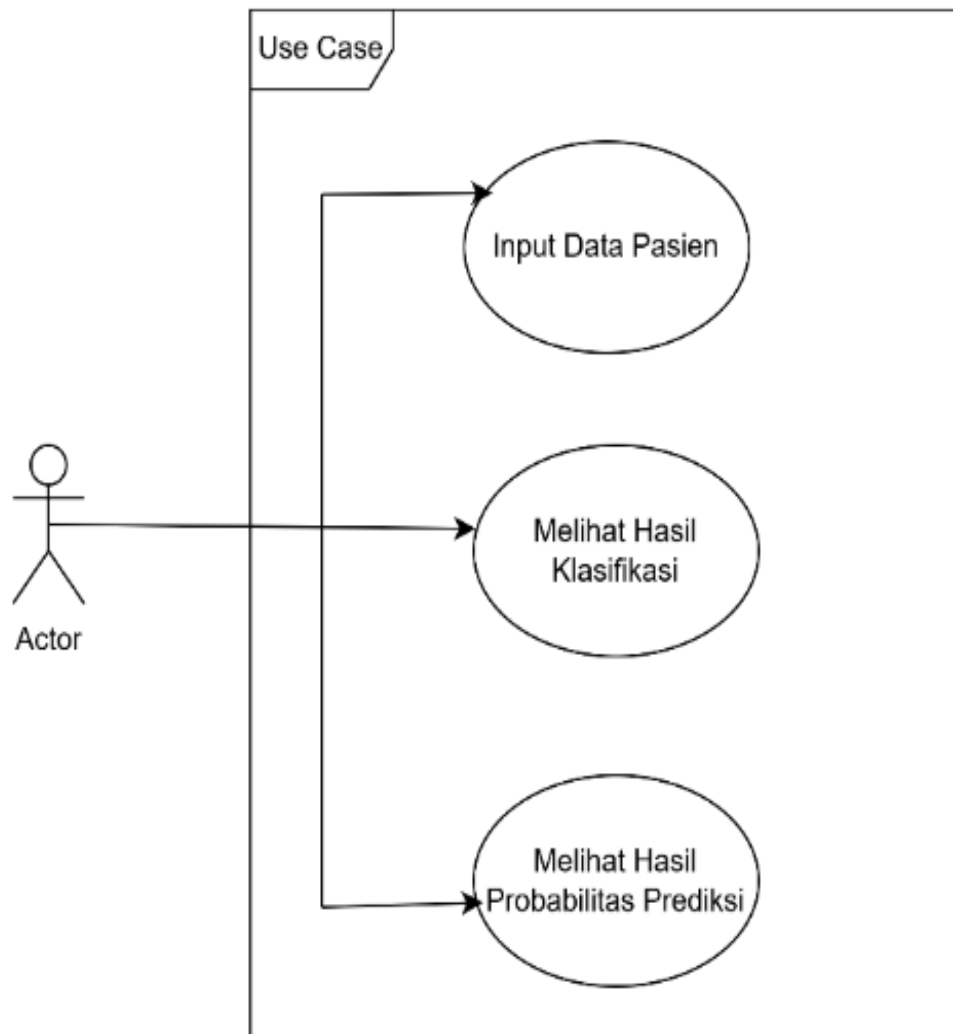
2.2 Perancangan Sistem

Perancangan sistem dilakukan setelah model klasifikasi *Random Forest* selesai dibangun dan memperoleh performa yang memenuhi kebutuhan penelitian. Tahap ini bertujuan mengimplementasikan model machine learning ke dalam sebuah aplikasi berbasis web sehingga proses deteksi dini Diabetes Melitus Tipe 2 dapat dilakukan secara lebih mudah, cepat, dan interaktif. Sistem dirancang agar dapat digunakan oleh tenaga kesehatan maupun masyarakat umum sebagai alat bantu skrining awal, tanpa menggantikan proses diagnosis yang dilakukan oleh tenaga medis.

Aplikasi dikembangkan menggunakan framework **Streamlit** karena mampu menyediakan antarmuka (user interface) yang sederhana, ringan, serta mudah diintegrasikan dengan model machine learning berbasis Python. Pengguna cukup memasukkan data pasien berupa usia, jenis kelamin, berat badan, tinggi badan, tekanan darah sistolik, dan tekanan darah diastolik. Selanjutnya sistem akan melakukan proses perhitungan Indeks Massa Tubuh (Body Mass Index/BMI) secara otomatis, melakukan praproses data sesuai dengan tahapan pelatihan model, kemudian mengirimkan data tersebut ke model *Random Forest* untuk menghasilkan prediksi tingkat risiko Diabetes Melitus Tipe 2.

Hasil prediksi yang diberikan sistem tidak hanya berupa kategori **Diabetes Melitus Tipe 2** atau **Non-Diabetes**, tetapi juga menampilkan nilai probabilitas prediksi untuk setiap kelas. Informasi probabilitas tersebut bertujuan memberikan gambaran mengenai tingkat keyakinan model terhadap hasil klasifikasi yang dihasilkan. Dengan demikian, pengguna memperoleh informasi yang lebih komprehensif sehingga hasil prediksi dapat dijadikan sebagai pendukung pengambilan keputusan awal sebelum dilakukan pemeriksaan medis lebih lanjut.

Interaksi antara pengguna dengan sistem dimodelkan menggunakan **Unified Modeling Language (UML)** dalam bentuk **Use Case Diagram** sebagaimana ditunjukkan pada Gambar 2. Diagram tersebut menggambarkan aktivitas utama yang dapat dilakukan pengguna, mulai dari memasukkan data pasien, menjalankan proses prediksi, hingga melihat hasil klasifikasi yang dihasilkan sistem. Setelah implementasi selesai, sistem diuji menggunakan metode **Black Box Testing** untuk memastikan seluruh fungsi berjalan sesuai kebutuhan, serta **User Acceptance Testing (UAT)** menggunakan kuesioner skala Likert yang melibatkan 10 responden guna mengetahui tingkat penerimaan dan kemudahan penggunaan sistem yang telah dikembangkan.



Gambar 2. Use Case Diagram

Kelayakan sistem diuji lewat dua metode: *black box testing* untuk memastikan tiap fungsi berjalan sesuai kebutuhan pengguna, dan *User Acceptance Testing* (UAT) memakai kuesioner skala Likert yang melibatkan 10 responden guna mengukur tingkat penerimaan terhadap sistem yang dibangun.

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Data

Total 3.825 data rekam medis terkumpul sepanjang April 2023–Februari 2026, dengan rincian per tahun pada Tabel 2. Dari jumlah tersebut, dipilih 242 data sebagai sampel penelitian, terbagi atas 93 data (38,4%) berlabel DM Tipe 2 dan 149 data (61,6%) non-diabetes (Tabel 3). Walau kelas non-diabetes lebih mendominasi, porsi kelas diabetes yang hampir menyentuh 40% tetap tergolong besar, sehingga penanganan ketidakseimbangan kelas menjadi keharusan pada tahap praproses

Tabel 2. Populasi

Tahun	Jumlah Pasien
2023	1172
2024	1626
2025	906
2026	121

Total	3825
--------------	-------------

Berikut merupakan tabel populasi data rekam medis sepanjang april 2023-Februari 2026

Tabel 3. Sampel

Kelas	Jumlah
Diabetes Melitus Tipe 2	93
Non Diabetes	149
Total	242

Eksplorasi data awal (EDA) menemukan sebaran jenis kelamin yang relatif berimbang pasien perempuan berjumlah sekitar 130 data, sedikit lebih banyak dari pasien laki-laki yang berjumlah sekitar 110 data. Rentang usia pasien terbentang dari 18 sampai 85 tahun, namun mayoritas data terkonsentrasi pada kelompok usia 50–70 tahun, mengindikasikan dominasi pasien paruh baya hingga lanjut usia. Menariknya, ketika usia dipetakan berdasarkan status diabetes, terlihat pola yang cukup jelas: median usia kelompok non-diabetes berkisar 40–45 tahun, sedangkan median usia kelompok diabetes naik ke kisaran 55–60 tahun. Pola ini mengisyaratkan bahwa makin lanjut usia seseorang, makin besar pula kecenderungan risiko DM Tipe 2 yang dihadapinya. Sebaliknya, nilai BMI pada kedua kelompok tidak menunjukkan perbedaan mencolok, sejalan dengan hasil korelasi yang hanya mencatat hubungan positif lemah antara BMI dan status diabetes ($r \approx 0,14$). Dari sisi korelasi antar variabel, berat badan dan BMI saling berkaitan cukup erat ($r \approx 0,77$), begitu pula tekanan sistolik dengan diastolik ($r \approx 0,55$). Terhadap variabel target diabetes sendiri, usia mencatat korelasi 0,38 dan sistolik 0,46 — keduanya lebih kuat dibanding BMI yang hanya 0,14. Temuan ini mengonfirmasi bahwa usia dan tekanan darah sistolik punya keterkaitan yang lebih relevan terhadap status diabetes pada dataset yang digunakan, dibanding indikator BMI.

3.2 Praproses Data

Pemeriksaan nilai kosong lewat perintah `data.isnull().sum()` mengungkap bahwa kolom tinggi badan dan berat badan masing-masing kehilangan 20 data, kolom BMI kosong pada 236 baris (karena memang belum dihitung), sementara sistolik dan diastolik masing-masing kehilangan 16 dan 17 data (Tabel 4). Seluruh nilai kosong pada variabel numerik itu diisi memakai median, sedangkan kolom BMI dihitung ulang setelah berat dan tinggi badan lengkap.

Tabel 4. Missing Value

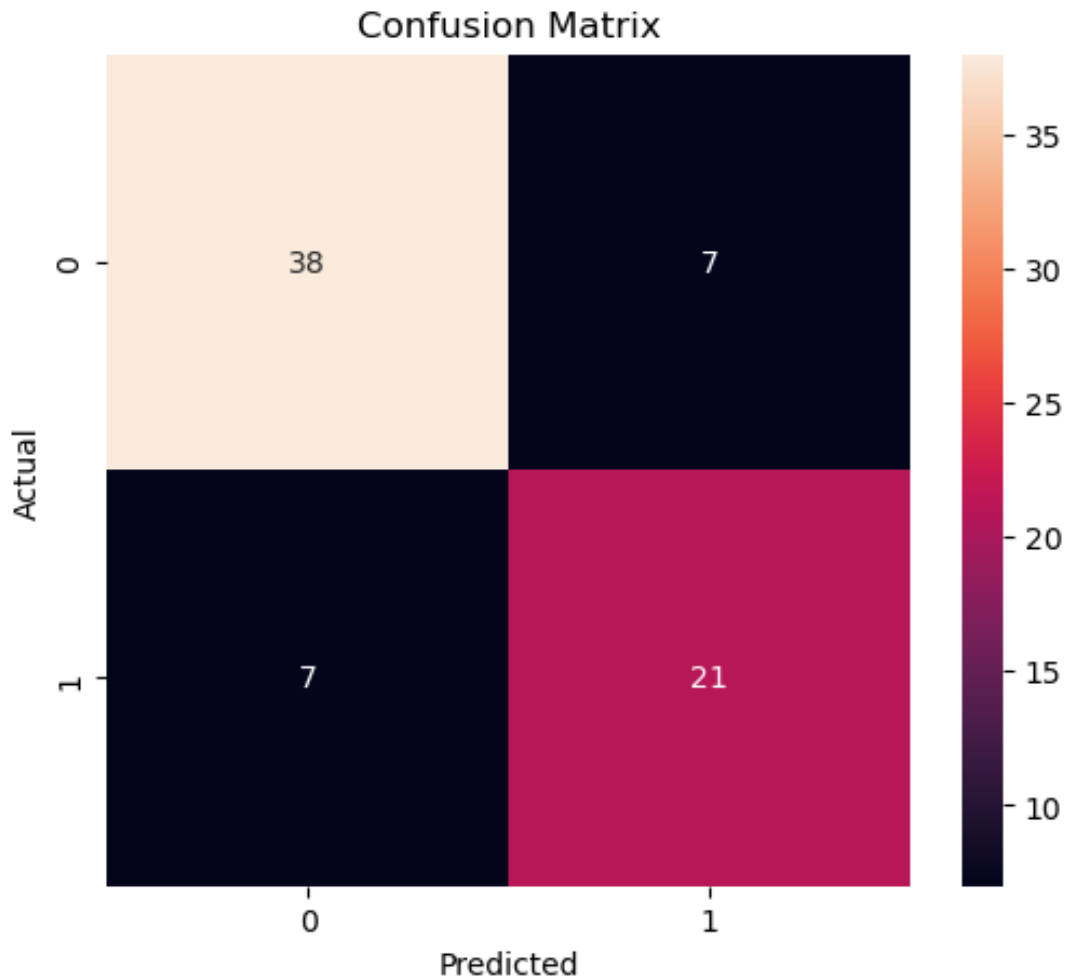
Kolom	Jumlah Missing Value
Usia	0
Jenis Kelamin	0
Tinggi Badan	20
Berat Badan	20
BMI	236
Sistolik	16
Diastolik	17
Diabetes	0

Variabel kategorikal jenis kelamin diubah ke bentuk numerik lewat proses *encoding* "L" menjadi 1 dan "P" menjadi 0. Sementara itu, deteksi *outlier* memakai metode IQR menunjukkan variabel diastolik paling banyak mengandung nilai ekstrem, yakni 82 data, disusul sistolik 15 data, BMI 7 data, berat badan 6 data, dan tinggi badan 1 data; variabel usia dan jenis kelamin bersih dari *outlier*. Nilai-nilai ekstrem tersebut ditangani lewat *winsorizing*, yakni memangkas nilai di luar

batas wajar ke titik batas terdekat tanpa membuang datanya. Tahap berikutnya, dataset dipecah menjadi 169 data latih dan 73 data uji dengan rasio 70:30, memakai parameter *stratify=y* agar proporsi kelas tetap terjaga di kedua subset, dan *random_state=42* agar hasil pembagian bisa diulang secara konsisten. Karena rasio kelas pada data latih masih timpang (*Imbalance Ratio* = 1,60), diterapkan SMOTE hanya pada data latih untuk menghindari kebocoran data ke set pengujian.

3.3 Evaluasi Kinerja Model

Kinerja model diukur lewat *confusion matrix* atas 73 data uji, menghasilkan *True Negative* (TN) = 38, *False Positive* (FP) = 7, *False Negative* (FN) = 7, dan *True Positive* (TP) = 21 (Gambar 3).



Gambar 3. Confusion Matrix

Berdasarkan nilai tersebut, metrik evaluasi dihitung sebagai berikut:

$$Akurasi = \frac{21 + 38}{21 + 38 + 7 + 7} = \frac{59}{73} \approx 0.8082 = 80,82\%$$

$$Presisi = \frac{21}{21 + 7} = \frac{21}{28} \approx 0.75 = 75\%$$

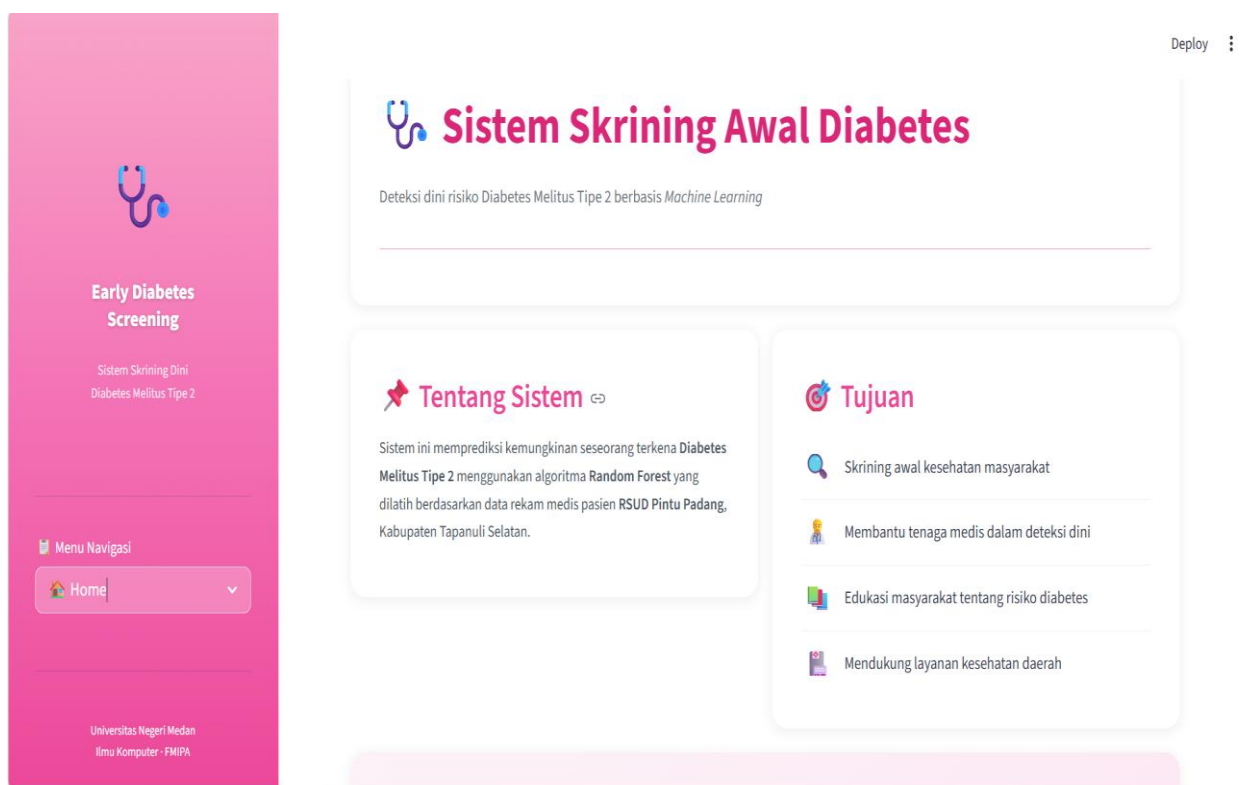
$$Recall = \frac{21}{21 + 7} = \frac{21}{28} \approx 0.75 = 75\%$$

$$F1 = 2 \times \frac{0.75 \times 0.75}{0.75 + 0.75} = 2 \times \frac{0.562}{1.5} \approx 0.75 = 75\%$$

Model juga diuji lewat kurva *Receiver Operating Characteristic* (ROC), yang memetakan hubungan antara *True Positive Rate* dan *False Positive Rate* pada berbagai ambang klasifikasi. Kurva yang dihasilkan berada konsisten di atas garis diagonal model acak (AUC = 0,5), dengan nilai *Area Under Curve* mencapai 0,85 menandakan model cukup andal membedakan pasien diabetes dari non-diabetes, dengan ketepatan klasifikasi sekitar 85%. Secara umum, kombinasi akurasi 80,82%, presisi dan *recall* yang sama-sama berada di 75%, *F1-score* 75%, dan AUC 0,85 memperlihatkan performa model yang berimbang tidak condong ke satu sisi, baik dalam mendeteksi kasus positif maupun menahan kesalahan klasifikasi. Meski begitu, masih terbuka ruang perbaikan lewat penyetelan *hyperparameter* atau penambahan variabel klinis pada pengembangan berikutnya.

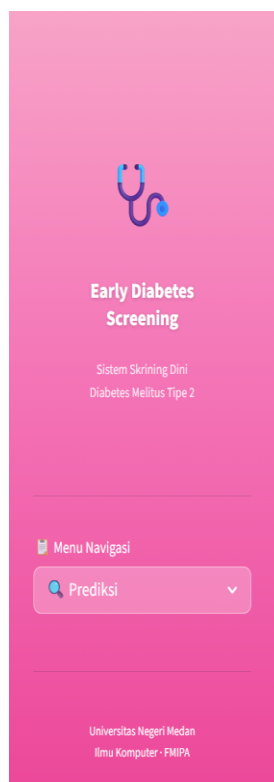
3.4 Implementasi Sistem

Model yang telah dilatih dipasang ke aplikasi web berbasis Streamlit dengan tiga menu utama: *Home* yang menyajikan gambaran umum sistem, *Prediksi* untuk memasukkan data pasien (usia, jenis kelamin, berat badan, tinggi badan, sistolik, diastolik), dan *Hasil Prediksi* yang menampilkan status klasifikasi beserta grafik probabilitas tiap kelas. Nilai BMI dihitung otomatis oleh sistem begitu berat dan tinggi badan diisi, sehingga input pengguna lebih ringkas.



Gambar 4. Menu Utama

Gambar diatas menampilkan halaman beranda (homepage) aplikasi web berbasis Streamlit untuk sistem skrining awal Diabetes Melitus Tipe 2. Pada halaman ini terdapat informasi singkat mengenai sistem, tujuan pengembangan sistem, serta menu navigasi yang memudahkan pengguna mengakses fitur utama, seperti menu beranda dan menu skrining. Halaman beranda dirancang sebagai tampilan awal untuk memberikan gambaran mengenai fungsi sistem sebelum pengguna melakukan proses deteksi dini diabetes.



Early Diabetes Screening

Isi seluruh data di bawah ini untuk mendapatkan hasil prediksi

Data Diri

Usia (tahun)

Jenis Kelamin

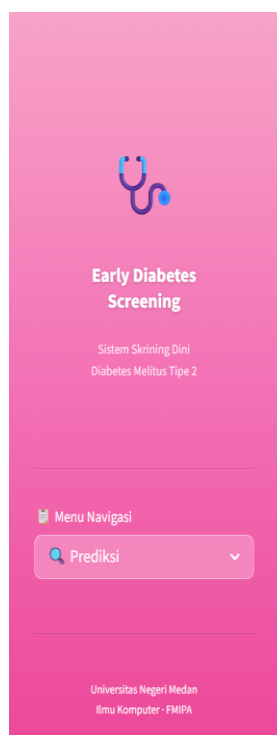
Berat Badan (kg)

Tinggi Badan (cm)

Data Tekanan Darah

Gambar 5. Menu Prediksi

Gambar 5 menampilkan halaman prediksi pada aplikasi web berbasis Streamlit. Pada halaman ini, pengguna dapat memasukkan data pasien yang meliputi usia, jenis kelamin, berat badan, tinggi badan, serta tekanan darah sistolik dan diastolik. Data yang telah diinput kemudian diproses oleh model Random Forest untuk menghasilkan prediksi apakah pasien berisiko mengalami Diabetes Melitus Tipe 2 atau tidak. Halaman ini merupakan fitur utama sistem yang digunakan dalam proses deteksi dini diabetes.



Ringkasan Data

Usia	BMI	Sistolik/Diastolik
68 tahun	21.0	170/80

Data Diri

Usia (tahun)

Jenis Kelamin

Berat Badan (kg)

Tinggi Badan (cm)

Data Tekanan Darah

Sistolik (mmHg)

Hasil Prediksi



Terindikasi Diabetes Melitus Tipe 2

Probabilitas diabetes: 82.3%

Diagram Probabilitas

Probabilitas Prediksi

100

Gambar 6. Hasil Prediksi

Gambar 6 memperlihatkan keluaran (output) sistem setelah proses prediksi selesai dilakukan. Sistem menampilkan ringkasan data pasien, hasil klasifikasi berupa status terindikasi atau tidak terindikasi Diabetes Melitus Tipe 2, serta visualisasi probabilitas prediksi sebagai pendukung dalam memahami hasil klasifikasi yang dihasilkan oleh model Random Forest..

3.5 Pengujian Model pada Data Baru

Untuk mengecek kemampuan generalisasi, model diuji lagi dengan 10 data responden masyarakat umum di luar dataset pelatihan/pengujian 5 orang berstatus diabetes terdiagnosis dokter, dan 5 orang tanpa riwayat diabetes. Sistem berhasil menebak benar 4 dari 5 kasus diabetes, dan 4 dari 5 kasus non-diabetes, sehingga akurasi pada pengujian independen ini mencapai 80%. Hasil ini menguatkan bahwa model cukup mampu menggeneralisasi pola dari data yang sama sekali belum pernah "dilihat" sebelumnya

4. KESIMPULAN

Riset ini berhasil menghasilkan model klasifikasi untuk mendeteksi dini Diabetes Melitus Tipe 2 memakai algoritma *Random Forest*, dibangun dari data rekam medis pasien RSUD Puntu Padang, Kabupaten Tapanuli Selatan, dengan variabel usia, jenis kelamin, berat badan, tinggi badan, BMI, serta tekanan darah sistolik dan diastolik. Proses penelitian menempuh serangkaian tahapan mulai dari pengumpulan data, praproses (penanganan *missing value* dan *outlier*), penyeimbangan kelas lewat SMOTE pada data latih, pembangunan model, evaluasi performa, hingga penerapan model ke sistem berbasis web memakai Streamlit. Hasil pengujian memperlihatkan performa yang cukup memuaskan: akurasi 80,82%, presisi dan *recall* pada kelas diabetes masing-masing 75%, F1-score 75%, serta AUC 0,85 nilai yang menandakan kemampuan model membedakan pasien diabetes dan non-diabetes cukup baik, jauh melampaui klasifikasi acak. Pengujian tambahan pada 10 responden masyarakat umum turut mengonfirmasi kemampuan generalisasi model dengan akurasi 80%. Temuan ini menegaskan bahwa data rekam medis dasar yang sudah tersedia di fasilitas kesehatan daerah bila diolah dengan pendekatan *machine learning* yang tepat dapat diubah menjadi alat bantu prediksi yang praktis dan bernilai klinis, sehingga berpotensi mendorong pengambilan keputusan medis yang lebih dini serta turut mengangkat mutu layanan kesehatan di tingkat daerah. Untuk pengembangan lanjutan, disarankan memperluas volume dan variasi data, menambah variabel lain seperti riwayat keluarga atau hasil laboratorium, serta melakukan penyetelan *hyperparameter* secara lebih mendalam guna mendongkrak akurasi dan daya generalisasi model.

REFERENCES

- [1] L. Kresnowati, S. Suhartono, Z. Shaluhiah, and B. Widjanarko, "Drop-out to follow up screening of diabetes mellitus in Indonesia from national health insurance data 2022-2024," *E3S Web of Conferences*, vol. 605, 2025, doi: 10.1051/e3sconf/202560502005.
- [2] Y. P. Utami, A. Triayudi, and E. T. Esthi Handayani, "Sistem Pakar Deteksi Penyakit Diabetes Mellitus (DM) menggunakan Metode Forward chaining dan Certainty factor Berbasis Android," *Jurnal JTik (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 5, no. 1, pp. 49–55, 2021, doi: 10.35870/jtik.v5i1.200.
- [3] A. Fauzi, "Aplikasi Sistem Pakar Dengan Metode Naive Bayes untuk Mendeteksi Penyakit Diabetes Expert System Application Using the Naive Bayes Method to Detect Diabetes," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 15, no. April, pp. 17–31, 2025.
- [4] B. M. Shields *et al.*, "Can clinical features be used to differentiate type 1 from type 2 diabetes? A systematic review of the literature," *BMJ Open*, vol. 5, no. 11, 2015, doi: 10.1136/bmjopen-2015-009088.
- [5] K. R. Widiyari, I. M. K. Wijaya, and P. A. Suputra, "Diabetes Melitus Tipe 2: Faktor Risiko, Diagnosis, Dan Tatalaksana," *Ganesha Medicine*, vol. 1, no. 2, pp. 114–120, 2021.
- [6] R. Riviani, E. Mutiara, H. Santoso, and A. Masyitah, "Analisis Spasial Kasus Diabetes Melitus dan Faktor Risiko di Provinsi Sumatera Utara Tahun 2023," *Jurnal Pengabdian Masyarakat dan Riset Pendidikan*, vol. 4, no. 1, pp. 972–979, 2025, doi: 10.31004/jerkin.v4i1.1683.
- [7] Muhriati and L. Lazuardi, "Meningkatkan Mutu Informasi Kesehatan Melalui Evaluasi Kualitas Data Rekam Medis Elektronik," *Journal of Information Systems for Public Health*, vol. 10, no. 02, pp. 51–63, 2025.

- [8] Erlin, Y. Nora Marlim, Junadhi, L. Suryati, and N. Agustina, "Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, no. 2, pp. 88–96, 2022, doi: 10.22146/jnteti.v11i2.3586.
- [9] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5," *Jurnal Krisnadana*, vol. 3, no. 1, pp. 25–33, 2023, doi: 10.58982/krisnadana.v3i1.470.
- [10] B. F. Sitanggang and P. Sitompul, "Deteksi Awal Kelangsungan Hidup Pasien Gagal Jantung Menggunakan Machine Learning Metode Random Forest," *INNOVATIVE: Journal Of Social Science Research*, vol. 4, no. 2, pp. 3347–3357, 2024.
- [11] W. Aprilia, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *Jurnal Sistem Informasi*, vol. 10, no. 1, p. 163, 2021, doi: 10.32520/stmsi.v10i1.1129.
- [12] F. Anisha, D. Vionanda, N. Amalita, and Zilrahmi, "Application of Random Forest for The Classification Diabetes Mellitus Disease in RSUP Dr. M. Jamil Padang," *UNP Journal of Statistics and Data Science*, vol. 1, no. 2, pp. 45–52, 2023, doi: 10.24036/ujsds/vol1-iss2/30.
- [13] L. A. Celi, M. S. Majumder, P. Ordonez, J. S. Osorio, K. E. Paik, and M. Somai, *Leveraging Data Science for Global Surgery*, vol. Part F2696. 2020. doi: 10.1007/978-3-030-83864-5_5.
- [14] D. C. P. Buani, "Deteksi Dini Penyakit Diabetes dengan Menggunakan Algoritma Random Forest," *EVOLUSI: Jurnal Sains dan Manajemen*, vol. 12, no. 1, pp. 1–8, 2024, doi: 10.31294/evolusi.v12i1.21005.
- [15] R. F. N. Iskandar, D. H. Gutama, D. P. Wijaya, and D. Danianti, "Klasifikasi Menggunakan Metode Random Forest untuk Awal Deteksi Diabetes Melitus Tipe 2," *Jurnal Teknik Industri Terintegrasi*, vol. 7, no. 3, pp. 1620–1626, 2024, doi: 10.31004/jutin.v7i3.26916.
- [16] A. Purnomo, M. Mashuri, and Humiati, "Akibat Hukum Bagi Rumah Sakit Terhadap Penyebarluasan Data Rekam Medis Pasien," *Jurnal Ners*, vol. 9, pp. 4865–4874, 2025.