

## Perbandingan Akurasi Algoritma Machine Learning untuk Prediksi Diabetes Berdasarkan Data Skrining Pasien Lansia di Puskesmas Sempaja

Nelson Sompaa Arifin<sup>1\*</sup>, Yohanes Andriano Teras<sup>2</sup>, Gede Enos Karli<sup>3</sup>, Norsianalara<sup>4</sup>,  
Lisnawati<sup>5</sup>, Aldi Bastiatul Fawait<sup>6</sup>

<sup>1,2,3,4,5</sup> Program Studi Ilmu Komputer, Fakultas Ilmu dan Informatika, Universitas Widya Gama Mahakam samarinda, indonesia.

\* Corresponding Author. E-mail: [nelsonsompaarifin0111@gmail.com](mailto:nelsonsompaarifin0111@gmail.com).

Receipt: Juni, 2026 .; Accepted: .....

**Abstract:** Diabetes melitus merupakan salah satu penyakit tidak menular yang memerlukan deteksi dini untuk mengurangi risiko komplikasi dan meningkatkan kualitas hidup pasien. Penelitian ini bertujuan membandingkan performa beberapa algoritma machine learning dalam memprediksi diabetes berdasarkan data skrining pasien lansia di Puskesmas Sempaja. Metode yang digunakan adalah data mining dengan tahapan Knowledge Discovery in Databases (KDD), meliputi seleksi data, prapemrosesan data, transformasi data, pemodelan, dan evaluasi. Dataset dibagi menjadi 80% data latih dan 20% data uji. Algoritma yang digunakan meliputi Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Naïve Bayes, Support Vector Machine (SVM), dan Logistic Regression. Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa Naïve Bayes, SVM, dan Logistic Regression memperoleh performa terbaik dengan nilai akurasi sebesar 88% dan F1-score sebesar 83%. Sementara itu, Decision Tree menunjukkan kemampuan yang lebih baik dalam mendeteksi kelas positif dibandingkan model lainnya. Hasil penelitian menunjukkan bahwa algoritma machine learning dapat dimanfaatkan untuk mendukung prediksi diabetes, dengan Naïve Bayes, SVM, dan Logistic Regression sebagai model yang memiliki performa terbaik pada dataset yang digunakan.

**Abstrak:** Diabetes mellitus is a non-communicable disease that requires early detection to reduce the risk of complications and improve patients' quality of life. This study aims to compare the performance of several machine learning algorithms in predicting diabetes based on health screening data of elderly patients at Puskesmas Sempaja. The research applied a data mining approach using the Knowledge Discovery in Databases (KDD) process, including data selection, preprocessing, transformation, modeling, and evaluation. The dataset was divided into 80% training data and 20% testing data. The algorithms used were Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results showed that Naïve Bayes, SVM, and Logistic Regression achieved the best performance with an accuracy of 88% and an F1-score of 83%. Meanwhile, Decision Tree demonstrated a better capability in identifying positive cases compared to the other models. The findings indicate that machine learning algorithms can be effectively utilized to support diabetes prediction, with Naïve Bayes, SVM, and Logistic Regression providing the highest overall performance on the dataset used in this study.

**Keywords** : Machine Learning; Prediksi Diabetes; Klasifikasi; Data Mining; Skrining Lansia.

### How to Cite this Article (IEEE):

S. A. Alfianto, Yudiansyah, and Aldi, "Article Title Here," *Journal of Multidisciplinary Informatics, Artificial Intelligence, and Business Innovation*., vol. 1, no. 21, p. 1-5, 2026, doi: 10.24903/minabis.v1i1.xxx

### © Author(s), 2026. Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## **PENDAHULUAN**

Diabetes melitus merupakan penyakit tidak menular yang menjadi salah satu penyebab utama peningkatan beban kesehatan masyarakat di berbagai negara. Peningkatan jumlah penderita diabetes, terutama pada kelompok lansia, berkaitan erat dengan perubahan gaya hidup, pola makan yang kurang sehat, serta minimnya aktivitas fisik[1][2]. Permasalahan dalam pengendalian diabetes melitus tidak hanya terletak pada meningkatnya jumlah penderita, tetapi juga pada rendahnya tingkat deteksi dini terhadap penyakit tersebut. Banyak kasus diabetes baru teridentifikasi ketika kondisi pasien telah memasuki tahap lanjut atau mengalami komplikasi. Keterlambatan diagnosis ini berpotensi meningkatkan risiko terjadinya komplikasi kronis, seperti penyakit kardiovaskular, gagal ginjal, neuropati, dan gangguan kesehatan lainnya yang dapat memperburuk kondisi pasien serta meningkatkan beban pelayanan kesehatan[3][4]. Deteksi dini diabetes menjadi langkah strategis dalam upaya pencegahan dan pengendalian penyakit ini [5]. Namun, proses diagnosis konvensional masih bergantung pada pemeriksaan laboratorium dan sering kali dilakukan setelah gejala klinis muncul [6]. Seiring berkembangnya teknologi komputasi dan ketersediaan data kesehatan dalam jumlah besar, machine learning berkembang sebagai pendekatan yang menjanjikan untuk mendukung deteksi dini diabetes[7]. Machine learning mampu mempelajari pola kompleks dari data kesehatan dan menghasilkan model prediksi yang lebih adaptif dibandingkan metode statistik tradisional, khususnya dalam menangani hubungan nonlinier antarvariabel klinis[8]. Berbagai algoritma machine learning telah diterapkan dalam penelitian prediksi diabetes, di antaranya Decision Tree, K-Nearest Neighbor (KNN), Random Forest, Naïve Bayes, Support Vector Machine (SVM), dan Logistic Regression [6][9]. Algoritma-algoritma tersebut memiliki karakteristik yang berbeda dalam proses pembelajaran, pemilihan fitur, serta mekanisme klasifikasi data. Perbedaan karakteristik tersebut dapat memengaruhi tingkat akurasi model yang dihasilkan, terutama ketika diterapkan pada data kesehatan yang memiliki kompleksitas dan variasi atribut yang tinggi [10].

Penelitian sebelumnya menunjukkan bahwa tidak terdapat satu algoritma yang selalu menghasilkan performa terbaik pada setiap dataset diabetes. Akurasi model sangat dipengaruhi oleh kualitas data, jumlah atribut, distribusi kelas, serta teknik prapemrosesan yang digunakan [11]. Oleh karena itu, evaluasi dan perbandingan performa beberapa algoritma machine learning perlu dilakukan untuk menentukan metode yang paling sesuai dalam memprediksi diabetes berdasarkan karakteristik data yang tersedia[12].

Pada fasilitas pelayanan kesehatan tingkat pertama, seperti puskesmas, kegiatan skrining kesehatan lansia menghasilkan berbagai data yang berpotensi dimanfaatkan untuk pengembangan model prediksi diabetes. Data tersebut meliputi karakteristik demografis, kondisi kesehatan, hasil pemeriksaan fisik, serta faktor risiko yang berkaitan dengan kejadian diabetes melitus [13]. Pemanfaatan data skrining melalui pendekatan machine learning diharapkan dapat membantu tenaga kesehatan dalam mengidentifikasi individu yang berisiko mengalami diabetes secara lebih cepat dan efisien, sehingga tindakan pencegahan dan penanganan dapat dilakukan lebih dini [14].

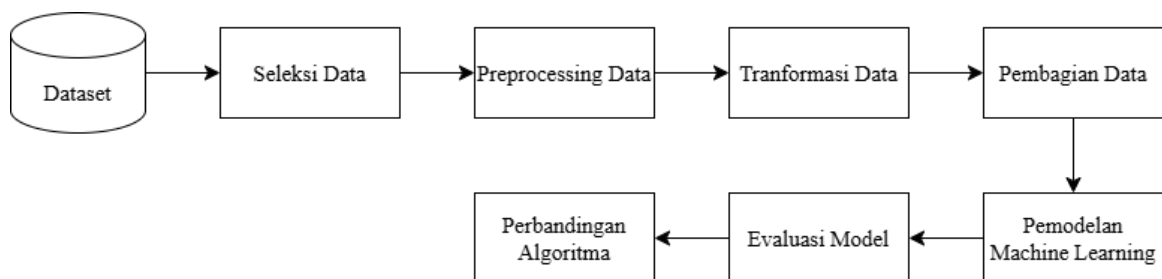
Puskesmas Sempaja merupakan salah satu fasilitas pelayanan kesehatan yang secara rutin melaksanakan kegiatan skrining kesehatan bagi pasien lansia. Data hasil skrining yang tersedia memiliki potensi untuk dimanfaatkan sebagai sumber data dalam pembangunan model prediksi diabetes berbasis machine learning. Namun, hingga saat ini belum terdapat penelitian yang membandingkan performa beberapa algoritma machine learning dalam memprediksi diabetes berdasarkan data skrining pasien lansia di Puskesmas Sempaja.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk membandingkan akurasi algoritma Decision Tree, K-Nearest Neighbor (KNN), Random Forest, Naïve Bayes, Support Vector Machine (SVM), dan Logistic Regression dalam prediksi diabetes berdasarkan data skrining pasien lansia di Puskesmas Sempaja. Evaluasi dilakukan menggunakan metrik

akurasi, precision, recall, dan F1-score untuk menentukan algoritma yang memiliki performa terbaik. Hasil penelitian diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan kesehatan serta menjadi referensi bagi penerapan machine learning dalam deteksi dini diabetes pada tingkat pelayanan kesehatan primer.

## METODE

Penelitian ini menggunakan pendekatan data mining untuk membangun dan membandingkan model prediksi diabetes berdasarkan data skrining pasien lansia di Puskesmas Sempaja. Tahapan penelitian mengacu pada proses Knowledge Discovery in Databases (KDD) yang terdiri atas seleksi data, prapemrosesan data, transformasi data, proses data mining, serta evaluasi hasil [15][12]. Pendekatan data mining banyak digunakan dalam bidang kesehatan karena mampu menemukan pola tersembunyi dari data medis dan menghasilkan model prediksi yang efektif untuk mendukung pengambilan keputusan klinis [16][17]. Seluruh proses implementasi direpresentasikan dalam Gambar 1.



Gambar 1. Metode

Tahap pertama adalah seleksi data (data selection), yaitu proses pengumpulan dan pemilihan data skrining pasien lansia yang memiliki keterkaitan dengan risiko diabetes [18]. Selanjutnya dilakukan tahap prapemrosesan data (data preprocessing) untuk menangani data yang hilang, data duplikat, serta data yang tidak konsisten sehingga kualitas data yang digunakan dalam pemodelan menjadi lebih baik [19].

Setelah proses pembersihan data selesai, dilakukan transformasi data (data transformation) melalui normalisasi dan pengkodean atribut kategorikal agar data dapat diproses oleh algoritma klasifikasi [20]. Dataset kemudian dibagi menjadi data latih dan data uji dengan tujuan mengevaluasi kemampuan model dalam melakukan prediksi terhadap data baru [21].

Pada tahap data mining, penelitian ini menerapkan enam algoritma klasifikasi yaitu Decision Tree, K-Nearest Neighbor (KNN), Random Forest, Naïve Bayes, Support Vector Machine (SVM), dan Logistic Regression. Algoritma tersebut dipilih karena telah banyak digunakan dalam penelitian prediksi diabetes dan menunjukkan performa yang baik pada berbagai jenis dataset kesehatan [15][22].

Evaluasi model dilakukan menggunakan confusion matrix dengan metrik accuracy, precision, recall, dan F1-score. Metrik tersebut digunakan untuk membandingkan performa masing-masing algoritma dan menentukan model terbaik dalam memprediksi diabetes berdasarkan data skrining pasien lansia di Puskesmas Sempaja [23].

## HASIL DAN PEMBAHASAN

Penelitian ini dilakukan untuk menganalisis performa beberapa algoritma machine learning dalam mengklasifikasikan data diagnosis pasien. Pemodelan dilakukan menggunakan bahasa pemrograman Python. Dataset dibagi menjadi data pelatihan dan data pengujian, dengan 80%

data digunakan untuk proses pelatihan model dan 20% data digunakan untuk pengujian. Enam algoritma machine learning yang digunakan dalam penelitian ini yaitu Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, Support Vector Machine (SVM), dan Logistic Regression.

Evaluasi performa model dilakukan menggunakan empat parameter, yaitu Accuracy, Precision, Recall, dan F1-Score. Accuracy digunakan untuk mengukur tingkat ketepatan prediksi model secara keseluruhan. Precision menunjukkan kemampuan model dalam memberikan prediksi positif yang benar, sedangkan Recall menunjukkan kemampuan model dalam mendeteksi seluruh data positif yang sebenarnya. F1-Score merupakan rata-rata harmonik dari Precision dan Recall yang digunakan untuk mengukur keseimbangan performa model. Deskripsi data yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

**Tabel 1.** Deskripsi Data

| No | Atribut                 | Tipe data | Missing Value |
|----|-------------------------|-----------|---------------|
| 1  | Tinggi badan (TB)       | Int       | 0             |
| 2  | Berat badan (BB)        | Int       | 0             |
| 3  | Indeks Masa tubuh (IMT) | Int       | 0             |
| 4  | Tekanan darah (TD)      | Int       | 0             |
| 5  | Lingkar perut (LP)      | Int       | 0             |
| 6  | Laboratorium (LAB)      | Int       | 0             |
| 7  | Usia                    | Int       | 0             |
| 8  | Rujukan                 | Int       | 0             |
| 9  | Diagnosa                | int       | 0             |

Berdasarkan Tabel 1, seluruh atribut pada dataset memiliki tipe data numerik sehingga dapat digunakan secara langsung dalam proses pemodelan machine learning. Variabel TB , BB , IMT, TD, LP, LAB, USIA, dan RUJUKAN digunakan sebagai variabel independen (fitur), sedangkan DIAGNOSA digunakan sebagai variabel dependen atau target klasifikasi.

Tabel 2 menampilkan hasil perbandingan performa model klasifikasi yang diperoleh setelah dilakukan pengujian terhadap data diagnosis pasien.

**Tabel 2.** Perbandingan Model Klasifikasi

| Model                  | Akurasi | Precision | Recall | F1-Score |
|------------------------|---------|-----------|--------|----------|
| Decision Tree          | 79%     | 80%       | 79%    | 79%      |
| K-Nearest Neighbors    | 87%     | 78%       | 87%    | 82%      |
| Random Forest          | 85%     | 79%       | 85%    | 81%      |
| Naïve Bayes            | 88%     | 82%       | 88%    | 83%      |
| Support Vector Machine | 88%     | 78%       | 88%    | 83%      |
| Logistic Regression    | 88%     | 78%       | 88%    | 83%      |

Penelitian ini memberikan gambaran mengenai kemampuan berbagai algoritma machine learning dalam melakukan klasifikasi diagnosis pasien. Berdasarkan hasil pengujian, model Naive Bayes, Support Vector Machine (SVM), dan Logistic Regression menunjukkan performa terbaik dengan nilai akurasi sebesar 88%. Ketiga model tersebut juga memperoleh nilai F1-Score

sebesar 83%, yang menunjukkan bahwa model mampu memberikan hasil klasifikasi yang baik pada keseluruhan data pengujian.

Model K-Nearest Neighbors (KNN) menunjukkan performa yang cukup baik dengan nilai akurasi sebesar 87% dan F1-Score sebesar 82%. Nilai akurasi yang tinggi menunjukkan bahwa model mampu melakukan klasifikasi dengan baik pada sebagian besar data. Namun, berdasarkan hasil evaluasi per kelas, model KNN masih mengalami kesulitan dalam mengenali kelas positif karena nilai recall pada kelas positif sangat rendah. Kondisi ini menunjukkan bahwa model lebih banyak mengklasifikasikan data ke dalam kelas negatif.

Selanjutnya, model Random Forest memperoleh akurasi sebesar 85% dengan nilai F1-Score sebesar 81%. Model ini mampu mengenali kelas negatif dengan sangat baik, namun kemampuan dalam mendeteksi kelas positif masih terbatas. Meskipun demikian, Random Forest tetap menunjukkan performa yang cukup stabil dalam proses klasifikasi.

Di sisi lain, model Decision Tree memperoleh akurasi sebesar 79% dengan nilai F1-Score sebesar 79%. Meskipun memiliki akurasi yang lebih rendah dibandingkan model lainnya, Decision Tree menunjukkan kemampuan yang relatif lebih baik dalam mendeteksi kelas positif. Hal ini terlihat dari nilai recall pada kelas positif yang lebih tinggi dibandingkan beberapa model lainnya. Dalam konteks klasifikasi diagnosis penyakit, kemampuan mendeteksi kasus positif menjadi aspek yang penting karena berkaitan dengan proses identifikasi pasien yang membutuhkan penanganan lebih lanjut.

Secara keseluruhan, hasil penelitian menunjukkan bahwa Naive Bayes, SVM, dan Logistic Regression merupakan model dengan performa terbaik berdasarkan nilai akurasi dan F1-Score. Namun, pemilihan model terbaik tidak hanya mempertimbangkan nilai akurasi, melainkan juga kemampuan model dalam mengenali kelas positif yang ditunjukkan oleh nilai recall. Oleh karena itu, evaluasi model perlu dilakukan secara menyeluruh agar diperoleh model yang sesuai dengan tujuan klasifikasi dan karakteristik data yang digunakan.

## **SIMPULAN**

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma machine learning mampu digunakan untuk mengklasifikasikan diagnosis pasien berdasarkan atribut yang digunakan dalam penelitian. Tujuan penelitian untuk membandingkan performa beberapa algoritma klasifikasi telah tercapai, di mana Naive Bayes, Support Vector Machine (SVM), dan Logistic Regression menunjukkan performa terbaik dengan nilai akurasi sebesar 88%. Temuan penelitian juga menunjukkan bahwa meskipun memiliki akurasi yang lebih rendah, Decision Tree memiliki kemampuan yang lebih baik dalam mendeteksi kelas positif dibandingkan model lainnya. Oleh karena itu, pemilihan model klasifikasi tidak hanya didasarkan pada nilai akurasi, tetapi juga perlu mempertimbangkan kemampuan model dalam mengenali kelas positif. Untuk penelitian selanjutnya, disarankan menggunakan teknik penyeimbangan data, seperti SMOTE atau oversampling, guna meningkatkan kemampuan model dalam mendeteksi kelas positif dan menghasilkan performa klasifikasi yang lebih optimal.

## **UCAPAN TERIMA KASIH**

Penulis mengucapkan terima kasih kepada Puskesmas Sempaja yang telah memberikan izin dan dukungan dalam pelaksanaan penelitian ini. Ucapan terima kasih juga disampaikan kepada dosen pembimbing, Universitas Widya Gama Mahakam Samarinda, serta semua pihak yang telah membantu dalam proses pengumpulan data, pelaksanaan penelitian, hingga penyusunan artikel ini. Semoga hasil penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam penerapan machine learning untuk prediksi diabetes.

## **DAFTAR PUSTAKA**

- [1] A. Astofa, P. Rosyani, and S. Apandi, "Evaluasi Komparatif Algoritma Machine Learning untuk Prediksi Dini Diabetes," vol. 6, no. 1, pp. 558–565, 2025, doi: 10.47065/bulletincsr.v6i1.859.
- [2] R. Amin and E. Tampubolon, "Penerapan Algoritma K Nearest Neighbor untuk Prediksi Akurasi Penyakit Diabetes," *Informatics Digit. Expert*, vol. 6, no. 1, pp. 87–91, 2024.
- [3] R. Ajeng, M. Farhan, and A. Cahyani, "Prediksi Terkena Diabetes menggunakan Metode K- Nearest Neighbor ( KNN ) pada Dataset UCI Machine Learning Diabetes Abstract : Abstrak :," vol. 3, no. 2, pp. 15–19, 2023, doi: 10.35472/indojam.v3i2.1577.
- [4] J. Wei *et al.*, "Metadata information and fundus image fusion neural network for hyperuricemia classification in diabetes," vol. 256, no. May, pp. 1–12, 2024, doi: 10.1016/j.cmpb.2024.108382.
- [5] M. O. Olusanya, R. E. Ogunsakin, M. Ghai, and M. A. Adeleke, "Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus : A Systematic Survey and Meta-Analysis Approach," no. MI, 2022.
- [6] L. F. Aparicio, J. Noguez, L. Montesinos, and J. A. G. García, "Machine learning and deep learning predictive models for type 2 diabetes : a systematic review," *Diabetol. Metab. Syndr.*, 2021, doi: 10.1186/s13098-021-00767-9.
- [7] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction , diagnosis , and management of diabetes : a comprehensive review," *Diabetol. Metab. Syndr.*, 2022, doi: 10.1186/s13098-022-00969-9.
- [8] A. Lim *et al.*, "Machine Learning Approaches for Type 2 Diabetes Prediction and Care Management".
- [9] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," 2021.
- [10] X. Wang *et al.*, "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," *BMC Med. Inform. Decis. Mak.*, pp. 1–14, 2021, doi: 10.1186/s12911-021-01471-4.
- [11] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI," no. September 2022, pp. 1–10, 2023, doi: 10.1049/htl2.12039.
- [12] S. G. Choi *et al.*, "Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods," *Sci. Rep.*, pp. 1–11, 2023, doi: 10.1038/s41598-023-40170-0.
- [13] M. F. Kurniawan and D. A. Megawaty, "Comparison of Logistic Regression , Random Forest , Support Vector Machine ( SVM ) and K-Nearest Neighbor ( KNN ) Algorithms in Diabetes Prediction," vol. 9, no. 5, pp. 2154–2162, 2025.
- [14] K. Huda and M. Ula, "PENERAPAN NAIVE BAYES, REGRESI LOGISTIK, RANDOM FOREST, SVM, DAN KNN UNTUK PREDIKSI DIABETES," pp. 1–10.
- [15] G. Dharmarathne, T. N. Jayasinghe, and M. Bogahawaththa, "Healthcare Analytics A novel machine learning approach for diagnosing diabetes with a self-explainable interface," *Healthc. Anal.*, vol. 5, no. January, p. 100301, 2024, doi: 10.1016/j.health.2024.100301.
- [16] O. T. Kee, H. Harun, N. Mustafa, N. Azian, A. Murad, and S. F. Chin, "Cardiovascular complications in a diabetes prediction model using machine learning : a systematic review," *Cardiovasc. Diabetol.*, pp. 1–10, 2023, doi: 10.1186/s12933-023-01741-7.
- [17] J. Hao, S. Luo, and L. Pan, "Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes," *Sci. Rep.*, pp. 1–12, 2022, doi: 10.1038/s41598-022-14143-8.
- [18] P. Sampath, G. Elangovan, K. Ravichandran, V. Shanmuganathan, S. Pasupathi, and T. Chakrabarti, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over- sampling technique," pp. 1–15, 2024.
- [19] M. T. Moghaddam, Y. Jahani, Z. Arefzadeh, A. Dehghan, and M. Khaleghi, "Predicting

diabetes in adults : identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm,” 2024.

- [20] M. Phongying and S. Hiriote, “Diabetes Classification Using Machine Learning Techniques,” 2023.
- [21] B. Feng, W. Saaveethya, S. King, H. Lim, and F. H. Juwono, “Diabetes detection based on machine learning and deep learning approaches,” pp. 24153–24185, 2024, doi: 10.1007/s11042-023-16407-5.
- [22] P. Bakhsh, C. Gravino, and F. Palomba, “Advances in Artificial Intelligence for Diabetes Prediction : Insights from a Systematic Literature Review,” 2024.
- [23] M. Hasan and F. Yasmin, “Predicting Diabetes Using Machine Learning : A Comparative Study of Classifiers,” 2025.