Aplikasi Prediksi Diabetes Berbasis Android Menggunakan Algoritma Random Forest dengan SMOTE dan Feature Selection

Sardiarinto [1], Vadlya Ma'arif [2*], Eko Saputro [3]

[1] *Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika Indonesia*
[2] *Teknologi Komputer, Universitas Bina Sarana Informatika Indonesia*
[3] *Sistem Informasi, Universitas Bina Sarana Informatika Indonesia*

** vadlya.vlr@bsi.ac.id*

Abstrak

Diabetes mellitus merupakan penyakit kronis dengan prevalensi yang terus meningkat secara global, sehingga deteksi dini menjadi aspek penting untuk mencegah komplikasi lebih lanjut. Perkembangan *machine learning* memberikan peluang signifikan dalam mendukung pengambilan keputusan medis melalui prediksi yang lebih akurat. Penelitian ini bertujuan mengembangkan model prediksi diabetes berbasis algoritma Random Forest yang dipadukan dengan Synthetic Minority Oversampling Technique (SMOTE) serta metode seleksi fitur. Dataset yang digunakan adalah *Pima Indians Diabetes Database* yang terdiri atas 768 sampel dengan delapan atribut prediktor. Tahapan penelitian meliputi pra-pengolahan data, penerapan SMOTE untuk mengatasi ketidakseimbangan kelas, seleksi fitur berbasis *Information Gain*, serta evaluasi model menggunakan skema *10-fold cross-validation*. Hasil eksperimen menunjukkan bahwa integrasi SMOTE dan seleksi fitur mampu meningkatkan kinerja model dibandingkan baseline, dengan akurasi sebesar 87,93%, nilai *recall* 91,3% untuk kelas positif, dan area ROC 0,949. Model terbaik kemudian diimplementasikan dalam aplikasi Android berbasis *standalone* yang memungkinkan pengguna melakukan prediksi risiko diabetes secara mandiri melalui input tujuh atribut utama, yaitu *Glucose, Pregnancies, Age, BMI, Insulin, Diabetes Pedigree Function,* dan *Blood Pressure*. Aplikasi ini menghasilkan prediksi yang cepat, mudah diakses, serta berpotensi menjadi solusi praktis untuk mendukung upaya deteksi dini diabetes.

Kata kunci: diabetes melitus, pembelajaran mesin, hutan acak, SMOTE, pemilihan fitur, aplikasi android

Abstract

Diabetes mellitus is a chronic disease with a steadily increasing prevalence globally, making early detection crucial for preventing further complications. Advances in machine learning offer significant opportunities to support medical decision-making through more accurate predictions. This study aims to develop a diabetes prediction model based on the Random Forest algorithm combined with the Synthetic Minority Oversampling Technique (SMOTE) and a feature selection method. The dataset used is the Pima Indians Diabetes Database, consisting of 768 samples with eight predictor attributes. The research steps included data preprocessing, applying SMOTE to address class imbalance, Information Gain-based feature selection, and model evaluation using

a 10-fold cross-validation scheme. Experimental results show that the integration of SMOTE and feature selection improved model performance compared to the baseline, with an accuracy of 87.93%, a recall of 91.3% for the positive class, and an ROC area of 0.949. The best model was then implemented in a standalone Android application that allows users to independently predict diabetes risk by inputting seven key attributes: Glucose, Pregnancies, Age, BMI, Insulin, Diabetes Pedigree Function, and Blood Pressure. This application produces fast, easily accessible predictions and has the potential to be a practical solution to support early diabetes detection efforts.

Keywords: *diabetes mellitus, machine learning, random forest, SMOTE, feature selection, android application*

1. Introduction

Diabetes mellitus merupakan salah satu masalah kesehatan global yang prevalensinya terus meningkat setiap tahun [1]. Penelitian global terbaru memperkirakan bahwa prevalensi diabetes tipe-2 terus meningkat secara signifikan, bahkan pada individu dengan BMI yang tidak tinggi, sebagaimana dilaporkan dalam studi global beban T2DM yang bukan disebabkan obesitas[2] diperkirakan bahwa prevalensi dan kematian akibat T2DM akan terus bertambah hingga tahun 2050 dalam sebuah kajian sistematik global[3]. Kondisi ini menuntut adanya strategi deteksi dini yang lebih akurat untuk membantu tenaga medis dalam pengambilan keputusan klinis [4]. Perkembangan metode *machine learning* telah membuka peluang besar dalam membangun model prediksi berbasis data kesehatan.

Berbagai penelitian sebelumnya menunjukkan bahwa penggunaan algoritma *ensemble learning* mampu meningkatkan performa prediksi dibandingkan model tunggal. Patel [5] menegaskan bahwa pendekatan *ensemble*, khususnya Random Forest, efektif dalam menganalisis data medis yang kompleks. Hal ini diperkuat oleh Brown[6] yang menunjukkan bahwa Random Forest unggul dibandingkan algoritma lain dalam membangun model prediksi klinis.

Namun, permasalahan *class imbalance* pada dataset medis sering menjadi kendala utama karena dapat mengurangi sensitivitas model terhadap kelas minoritas. García et al. [7] merekomendasikan penggunaan metode oversampling seperti Synthetic Minority Oversampling Technique (SMOTE) untuk memperbaiki distribusi kelas sehingga performa model lebih seimbang. Selain itu, kompleksitas jumlah atribut juga memengaruhi kualitas model. Zhou dan Liu[8] menjelaskan bahwa penerapan feature selection pada data kesehatan dapat mengurangi kompleksitas, mempercepat

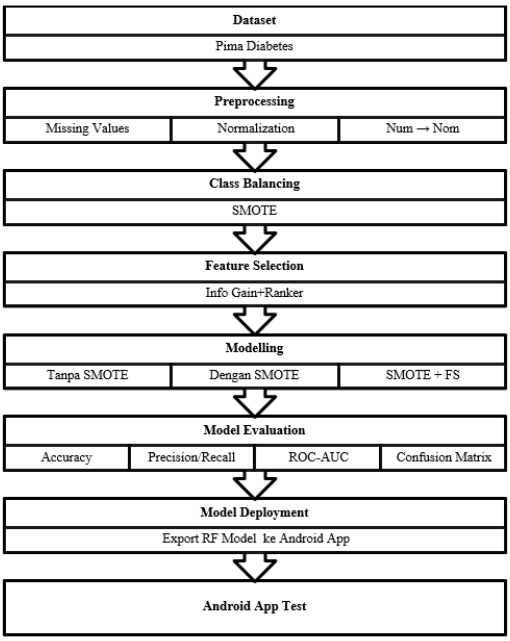
komputasi, serta meningkatkan akurasi prediksi.

Di samping pengembangan model, tren terkini juga menekankan pentingnya implementasi hasil prediksi ke dalam aplikasi mobile berbasis Android. Hendawi[9] misalnya, mengembangkan aplikasi **XAI4Diabetes** dengan fitur interpretabilitas yang membantu pengguna memahami hasil prediksi, sedangkan El-Sofany[10] berhasil mengintegrasikan model *machine learning* dengan antarmuka Android untuk mendukung prediksi diabetes secara *real-time*. Lebih jauh lagi, efektivitas aplikasi mobile berbasis Android juga dibuktikan pada bidang kesehatan lain, seperti hipertensi dan penyakit jantung. Studi terbaru menunjukkan bahwa aplikasi Android dengan fitur monitoring dan intervensi mampu meningkatkan pengelolaan tekanan darah[11]. Selain itu aplikasi mobile dapat menjadi metode tambahan efektif dalam pencegahan penyakit kardiovaskular melalui tracking faktor risiko inklusif seperti BMI, glukosa, dan gaya hidup[12].

Berdasarkan uraian tersebut, penelitian ini berfokus pada integrasi Random Forest dengan teknik SMOTE dan *feature selection* dalam prediksi diabetes. Selanjutnya, model terbaik diimplementasikan ke dalam aplikasi Android sehingga dapat digunakan oleh masyarakat secara langsung untuk mendukung deteksi dini.

2. Metodologi Penelitian

Keseluruhan tahapan penelitian digambarkan dalam *flowchart* yang mencakup proses mulai dari dataset, pra-pengolahan data, penerapan SMOTE, seleksi fitur, klasifikasi menggunakan Random Forest, evaluasi model, hingga implementasi model ke dalam aplikasi Android. Diagram alur ini disajikan pada Gambar 1 untuk memberikan gambaran sistematis mengenai metodologi penelitian.



Gambar 1. Flowchart Tahapan penelitian

Keseluruhan tahapan penelitian ini digambarkan dalam flowchart metodologi yang mencakup alur mulai dari dataset,

preprocessing data, penerapan SMOTE, seleksi fitur, klasifikasi dengan Random Forest, evaluasi model hingga implementasi model. Diagram alur ini membantu memberikan gambaran sistematis mengenai proses penelitian yang dilakukan.

Pemilihan Dataset

Penelitian ini menggunakan *Pima Indians Diabetes Dataset* dari UCI Machine Learning Repository yang berisi 768 sampel dengan delapan variabel prediktor dan satu variabel target. Dataset ini dipilih karena banyak digunakan dalam penelitian terdahulu sebagai standar untuk menguji model prediksi diabetes.

Pra-pengolahan Data

Tahap selanjutnya dilakukan *preprocessing data*, termasuk penanganan nilai hilang dan normalisasi agar setiap fitur berada pada skala yang sebanding. Selanjutnya, untuk meningkatkan reliabilitas hasil, validasi model dilakukan menggunakan *10-fold cross-validation* yang terbukti lebih stabil dalam mengevaluasi kinerja algoritma klasifikasi pada dataset medis [13].

Penanganan Ketidakseimbangan Kelas

Untuk mengatasi masalah ketidakseimbangan kelas, digunakan metode SMOTE yang mampu melakukan oversampling kelas minoritas dengan menciptakan data sintetis baru. Teknik ini terbukti efektif dalam meningkatkan performa model klasifikasi, khususnya pada kasus medis yang sensitif terhadap *recall* [14].

Seleksi Fitur

Selain itu, penelitian ini juga menerapkan *feature selection* menggunakan metode Information Gain dengan algoritma Ranker untuk mengidentifikasi fitur-fitur yang paling relevan. Menurut Xu dan Sun [15], seleksi fitur tidak hanya mengurangi kompleksitas model, tetapi juga dapat meningkatkan interpretabilitas dengan menunjukkan atribut mana yang paling berpengaruh terhadap hasil prediksi.

Pembangunan Model

Tahap berikutnya adalah pembangunan model prediksi menggunakan algoritma Random Forest. Algoritma ini dipilih karena bersifat *ensemble* dengan kombinasi *bagging* dan pohon keputusan yang mampu menangani data non-linear serta *robust* terhadap *overfitting*. Proses pelatihan dilakukan dengan jumlah pohon (*trees*) sebanyak 100.

Evaluasi Model

Evaluasi kinerja dilakukan menggunakan beberapa metrik, yaitu akurasi, presisi, *recall*, F-measure, dan ROC-AUC. Kombinasi metrik ini dipilih agar dapat

memberikan gambaran menyeluruh mengenai performa model, khususnya dalam mendeteksi kelas positif diabetes.

Implementasi ke Aplikasi Android

Sebagai langkah lanjutan, model terbaik di-deploy ke dalam aplikasi Android berbasis standalone. Aplikasi ini dirancang dengan antarmuka minimalis, menampilkan kolom input atribut pilihan dan menghasilkan keluaran berupa status prediksi risiko diabetes.

3. Hasil dan Pembahasan

Penelitian ini menggunakan Pima Indians Diabetes Database yang tersedia publik melalui Kaggle, terdiri atas 768 sampel dengan delapan atribut prediktor (*Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age*) dan satu atribut target (*Outcome*). Terdapat 268 instance kelas positif (diabetes) dan 500 instance kelas negatif, menunjukkan ketidakseimbangan yang signifikan.

Sebelum eksperimen dilakukan, data diproses melalui *preprocessing* meliputi penanganan nilai hilang, normalisasi, dan konversi tipe data target ke nominal ("*Yes"/"No*"). Hal ini penting agar algoritma klasifikasi Random Forest di WEKA, termasuk *SMOTE* dan seleksi fitur, dapat bekerja optimal.

Setelah preprocessing selesai, dilakukan tiga skenario eksperimen, yaitu:

1. Model baseline menggunakan *Random Forest* tanpa SMOTE dan tanpa seleksi fitur.
2. Model dengan balancing, yaitu Random Forest setelah dilakukan oversampling menggunakan SMOTE.
3. Model integrasi balancing dengan seleksi 6 fitur dan 7 fitur, yaitu Random Forest dengan SMOTE ditambah feature selection berbasis *Information Gain*.

Hasil dari ketiga skenario tersebut dibandingkan secara komprehensif menggunakan metrik evaluasi akurasi, *precision, recall, ROC Area, dan Confusion Matrix*. Hasil pengolahan data dari tiga skenario menggunakan aplikasi WEKA dapat dilihat pada tabel 1 dan tabel 2.

Tabel 1. Perbandingan *Confusion Matrix*

Skenario	TP (Yes)	FN (Yes)	FP (No)	TN (No)	Total
Tanpa SMOTE & Tanpa Seleksi Fitur	159	109	82	418	768
Dengan SMOTE	451	85	127	373	1036
Dengan SMOTE + Feature Selection (6 fitur)	391	145	102	898	1536
Dengan SMOTE + Feature Selection (7 fitur)	979	93	157	843	2072

Pada skenario pertama (tanpa SMOTE & seleksi fitur), model cukup baik mengenali kelas *no* tetapi lemah pada kelas *yes*, ditunjukkan oleh 109 kasus positif yang salah diprediksi sebagai negatif. Setelah penerapan SMOTE (skenario kedua), sensitivitas terhadap kelas *yes* meningkat

dengan TP lebih tinggi (451), meskipun FP juga bertambah.

Dengan SMOTE + seleksi 6 fitur utama (skenario ketiga), jumlah FN memang lebih besar, tetapi FP menurun sehingga presisi meningkat. Terakhir pada SMOTE ditambah seleksi 7 fitur (skenario keempat), model mencapai kinerja terbaik dengan TP tertinggi (979) serta keseimbangan baik antara recall dan precision.

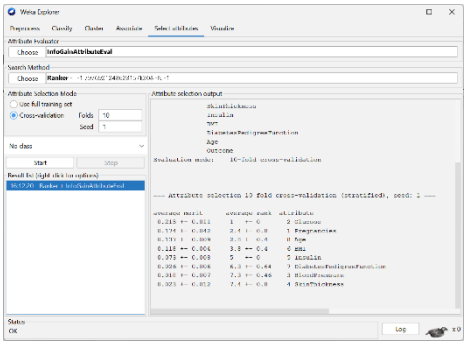
Tabel 2. Perbandingan Hasil Pengujian

Kondisi Eksperimen	Akurasi (%)	Recall (Diabetes)	Recall (Non-Diabetes)	Precision (Diabetes)	AUC
Tanpa SMOTE & tanpa seleksi fitur	75,13	59,3%	83,6%	66,0%	0,806
Dengan SMOTE	79,53	84,1%	74,6%	78,0%	0,876
SMOTE + Seleksi Fitur (6 fitur)	83,91	72,9%	89,8%	79,3%	0,915
SMOTE + Seleksi Fitur (7 fitur)	87,93	91,3%	84,3%	86,2%	0,949

Pada kondisi pertama, model Random Forest diuji menggunakan dataset asli yang berjumlah 768 instance dengan delapan atribut prediktor. Hasil evaluasi menunjukkan bahwa model berhasil mengklasifikasikan data dengan akurasi sebesar 75,13%, dengan Recall untuk kelas positif (diabetes) hanya mencapai 59,3%, sementara recall untuk kelas negatif (non-diabetes) lebih tinggi, yaitu 83,6%. Hal ini menunjukkan adanya kecenderungan model untuk lebih baik mengenali pasien non-diabetes dibandingkan pasien diabetes. Penyebabnya adalah distribusi data yang tidak seimbang, di mana jumlah pasien non-diabetes lebih dominan.

Pada kondisi kedua, teknik SMOTE digunakan untuk menyeimbangkan distribusi kelas dengan melakukan oversampling pada kelas minoritas. Hasil evaluasi menunjukkan adanya peningkatan kinerja model. Akurasi model naik menjadi 79,53%, menandakan peningkatan konsistensi prediksi antara kelas positif dan negatif. Recall untuk kelas diabetes meningkat signifikan menjadi 84,1%, yang berarti model lebih mampu mengenali pasien dengan diabetes. Namun, recall untuk kelas non-diabetes sedikit menurun menjadi 74,6%, meskipun presisi tetap terjaga. Temuan ini mengonfirmasi bahwa SMOTE efektif dalam mengatasi ketidakseimbangan kelas dan meningkatkan deteksi kasus diabetes.

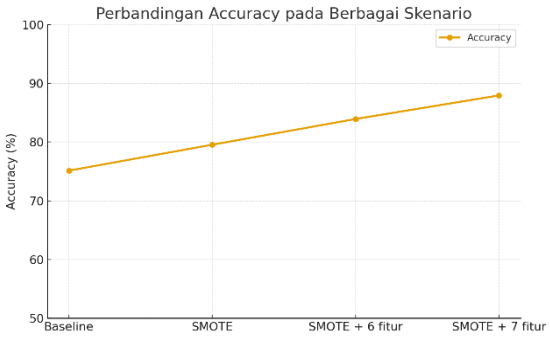
Kondisi ketiga mengombinasikan SMOTE dengan seleksi fitur menggunakan *InfoGainAttributeEval* dan metode *Ranker*. Atribut dengan kontribusi tertinggi mulai dari *Glucose*, *Pregnancies*, *Age*, *BMI*, *Insulin*, Diabetes Pedigree Function dan Blood Pressure dipilih untuk membangun model. Hasil seleksi fitur dapat di lihat pada gambar 2.



Gambar 2. Seleksi Feature dengan metode *InfoGainAttributeEval*

Eksperimen dilakukan dengan enam hingga tujuh atribut terbaik, dan hasilnya menunjukkan adanya peningkatan yang lebih signifikan dibanding dua kondisi sebelumnya. Pada saat menggunakan enam fitur terbaik model mencapai akurasi 83,91%. Recall untuk kelas positif 72,9%, sementara recall untuk kelas negatif mencapai 89,8%. Nilai AUC sebesar 0,915. Sedangkan saat menggunakan tujuh fitur terbaik model mencapai akurasi 87,93%. Recall untuk kelas positif meningkat hingga 91,3%, sementara recall untuk kelas negatif mencapai 84,3%. Nilai AUC sebesar 0,949 menunjukkan kemampuan diskriminatif model yang sangat baik.

Berdasarkan ketiga skenario tersebut, terlihat adanya peningkatan kinerja model seiring dengan penerapan SMOTE dan seleksi fitur. Tanpa balancing data, model cenderung bias terhadap kelas mayoritas sehingga banyak kasus diabetes yang tidak terdeteksi. Penerapan SMOTE berhasil mengurangi bias tersebut, meskipun terjadi sedikit penurunan kinerja pada kelas mayoritas. Penambahan seleksi fitur semakin meningkatkan performa karena hanya atribut yang relevan yang digunakan dalam pelatihan model, sehingga kompleksitas berkurang dan kemampuan generalisasi meningkat.



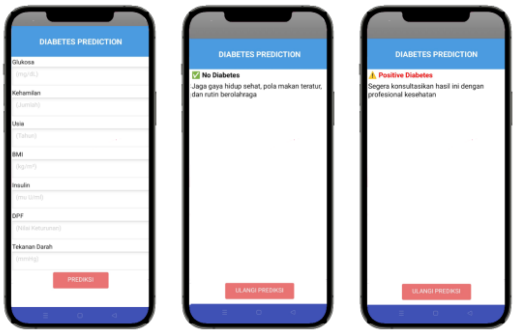
Gambar 3. Perbandingan Akurasi

Berdasarkan Gambar 2, secara keseluruhan kombinasi Random Forest dengan SMOTE dan Feature Selection terbukti memberikan hasil terbaik dengan peningkatan akurasi hampir 13 poin persentase dibandingkan kondisi awal. Hal ini menunjukkan bahwa pendekatan tersebut efektif dalam menangani masalah prediksi diabetes, khususnya pada dataset yang mengalami ketidakseimbangan kelas dan redundansi atribut.

Setelah proses pelatihan dan evaluasi model prediksi menggunakan Random Forest dengan kombinasi SMOTE dan feature

selection, tahap selanjutnya adalah model deployment. Model yang telah terlatih diekspor dalam format yang kompatibel sehingga dapat diintegrasikan ke dalam aplikasi Android.

Aplikasi kemudian dibangun dengan antarmuka sederhana yang memungkinkan pengguna memasukkan tujuh atribut prediktor, yaitu *Glucose, Pregnancies, Age, BMI, Insulin, Diabetes Pedigree Function,* dan *Blood Pressure*. Data input ini diproses langsung oleh model yang tertanam di dalam aplikasi, dan hasil prediksi ditampilkan dalam bentuk output yang mudah dipahami, seperti "Positive Diabetes" atau "No Diabetes".



Gambar 3. Pengujian Aplikasi

Tahap terakhir adalah pengujian aplikasi Android (*app test*). Pengujian dilakukan menggunakan emulator Android untuk memastikan fungsi model berjalan dengan baik. Hasil pengujian menunjukkan bahwa aplikasi mampu memberikan prediksi sesuai dengan performa model yang diuji pada tahap sebelumnya. Dengan demikian, aplikasi ini dapat dijadikan sebagai sistem pendukung keputusan berbasis mobile untuk deteksi dini diabetes.

4. Kesimpulan

Penelitian ini berhasil membangun model prediksi diabetes menggunakan algoritma Random Forest dengan kombinasi SMOTE dan seleksi fitur. Hasil eksperimen menunjukkan bahwa penerapan SMOTE mampu meningkatkan sensitivitas model terhadap kelas minoritas, sementara integrasi seleksi fitur lebih lanjut meningkatkan akurasi keseluruhan, presisi, dan recall. Model terbaik kemudian diimplementasikan ke dalam aplikasi Android berbasis standalone, yang memungkinkan pengguna melakukan prediksi risiko diabetes secara mandiri melalui input tujuh atribut utama, yaitu *Glucose, Pregnancies, Age, BMI, Insulin, Diabetes Pedigree Function,* dan *Blood Pressure*. Aplikasi ini terbukti mampu memberikan hasil prediksi secara cepat dan mudah digunakan, sehingga berpotensi menjadi alat bantu deteksi dini diabetes yang praktis bagi masyarakat maupun tenaga medis.

5. References

[1] H. Aulia, A. Wibowo, and S. Sutrisno, “Integration of Random Forest, ADASYN, and SHAP for Diabetes Prediction and

Interpretation," *Scientific Journal of Informatics*, vol. 12, no. 2, pp. 211–222, Jun. 2025, doi: 10.15294/sji.v12i2.24314.

[2] J. Wu *et al.*, "Global burden of type 2 diabetes attributable to non-high body mass index from 1990 to 2019," *BMC Public Health*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12889-023-15585-z.

[3] H. Muhammad, U. Abid, M. Naveed, B. Afzal, and M. Azeem, "Global Prevalence and Mortality of Type-2 Diabetes", doi: 10.31703/gdddr.2024(IX-I).

[4] A. Shafqat *et al.*, "Diabetes Prediction Using Deep Learning: A Comprehensive Approach Utilizing Feature Selection and Deep Neural Networks", doi: 10.56979/801/2024.

[5] A. Wantoro, A. Fitria Yuliana, D. Yana, A. Andini, I. Awaliyani, and W. Caesarendra, "Optimizing Type 2 Diabetes Classification with Feature Selection and Class Balancing in Machine Learning," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 4, pp. 2723–3863, 2025, doi: 10.52436/1.jutif.2025.6.4.5166.

[6] A. A. Aouragh, M. Bahaj, and F. Toufik, "Diabetes Prediction: Optimization of Machine Learning through Feature Selection and Dimensionality Reduction," *International journal of online and biomedical engineering*, vol. 20, no. 8, pp. 100–114, May 2024, doi: 10.3991/ijoe.v20i08.47765.

[7] I. J. Kakoly, M. R. Hoque, and N. Hasan, "Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique," *Sustainability (Switzerland)*, vol. 15, no. 6, Mar. 2023, doi: 10.3390/su15064930.

[8] E. I. Abd El-Latif and I. A. Moneim, "Exploring Feature Selection and Machine Learning Algorithms for Predicting Diabetes Disease," *International Journal of Intelligent Systems and Applications*, vol. 16, no. 1, pp. 1–10, Feb. 2024, doi: 10.5815/ijisa.2024.01.01.

[9] R. Hendawi, J. Li, and S. Roy, "A Mobile App That Addresses Interpretability Challenges in Machine Learning–Based Diabetes Predictions: Survey-Based User Study," *JMIR Form Res*, vol. 7, no. 1, 2023, doi: 10.2196/50328.

[10] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, "A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App," *International Journal of Intelligent Systems*, vol. 2024, 2024, doi: 10.1155/2024/6688934.

[11] S. Martini *et al.*, "OPEN ACCESS EDITED BY Design and development of a smartphone app for hypertension management: An intervention mapping approach."

[12] A. Battistoni, G. Tocci, G. Gallo, G. Solfanelli, and M. Volpe, "A Mobile App-based Approach in Cardiovascular Disease Prevention:

A Prospective Randomized Study," *High Blood Pressure and Cardiovascular Prevention*, vol. 31, no. 1, pp. 93–96, Jan. 2024, doi: 10.1007/s40292-024-00625-5.

[13] J. Yang *et al.*, "Construction of a 3-year risk prediction model for developing diabetes in patients with pre-diabetes," *Front Endocrinol (Lausanne)*, vol. 15, 2024, doi: 10.3389/fendo.2024.1410502.

[14] Allani Udaya, "Interactive Diabetes Risk Prediction Using Explainable Machine Learning: A Dash-Based Approach with SHAP, LIME, and Comorbidity Insights," May 2025, doi: https://doi.org/10.48550/arXiv.2505.05683.

[15] S. Liu, "Diabetes Prediction by KNN, SVM, Random Forest and XGBoost," 2023.