Prosiding Seminar Nasional Ilmu Teknik
Volume 2 Nomor 2 Desember 2025
e-ISSN: 3063-5713; p-ISSN: 3063-4709, Hal 559-570
DOI: https://doi.org/10.61132/prosemnasproit.v2i2.62
Tersedia: https://prosiding.aritekin.or.id/index.php/PROSEMNASPROIT/

Deteksi Risiko Diabetes Pada Wanita Hamil Menggunakan Algoritma *Random Forest*
(Studi Kasus: Pima Indian Dataset)

Yazid Ichwanuddin[1*], Maria Rosario B[2], Erissya Rasywir[3]
[1-3] Dapartemen, Fakultas Ilmu Komputer, Universitas Dinamika Bangsa, Negara
Email: ichwanyazid18@gmail.com[1], diamar_ros@yahoo.com[2]*

Alamat: Jalan Jendral Sudirman, Thehok Jambi
**Penulis Korespondensi: ichwanyazid18@gmail.com*

Abstract. *Gestational Diabetes Mellitus (GDM) is a pregnancy-related metabolic disorder that poses health risks to both mother and fetus if not detected early, requiring accurate prediction methods for early screening and clinical decision-making. This study applies the Random Forest algorithm to detect GDM risk using clinical data from the Pima Indian Dataset. Data preprocessing included handling missing values, standardization, feature engineering, and a 70:30 train–test split. Two models were developed: a baseline and an optimized model using GridSearchCV hyperparameter tuning, validated with 5-fold cross-validation. Performance was assessed using a classification report, confusion matrix, and ROC–AUC. Results show that the optimized model outperforms the baseline, achieving 88% accuracy, an AUC of 93%, and average recall of 81%–85%. Compared to previous studies, this approach demonstrates improved predictive performance. The findings indicate that combining Random Forest with comprehensive preprocessing, feature engineering, and model optimization is effective and feasible for developing a medical decision support system for early GDM risk screening.*

Keywords: *Decision Support System, Feature Engineering, Gestational Diabetes Mellitus, Random Forest, Risk Prediction*

Abstrak. *Gestational Diabetes Mellitus* (GDM) merupakan gangguan metabolik pada kehamilan yang berisiko terhadap kesehatan ibu dan janin jika tidak terdeteksi sejak dini, sehingga diperlukan metode prediksi yang akurat untuk skrining awal dan pengambilan keputusan klinis. Penelitian ini menerapkan algoritma *Random Forest* untuk mendeteksi risiko GDM menggunakan data klinis dari Pima Indian Dataset. Tahapan pra-pemrosesan data mencakup penanganan nilai hilang, standardisasi, rekayasa fitur, dan pembagian data dengan skema *train–test* 70:30. Dua model dikembangkan, yaitu model *baseline* dan model yang dioptimasi melalui *hyperparameter tuning* dengan *GridSearchCV*, kemudian divalidasi menggunakan *5-fold cross-validation* untuk memastikan stabilitas dan kemampuan generalisasi. Kinerja model dievaluasi menggunakan *classification report, confusion matrix*, dan kurva ROC dengan nilai AUC. Hasil menunjukkan model teroptimasi lebih unggul dibanding *baseline*, dengan akurasi 88%, AUC 93%, dan *recall* rata-rata 81%–85%. Pendekatan ini menunjukkan peningkatan kinerja prediksi yang signifikan dan terbukti efektif untuk pengembangan sistem pendukung keputusan medis dalam skrining awal risiko GDM.

Kata kunci: Sistem Pendukung Keputusan; Rekayasa Fitur; Diabetes Gestasional; Random Forest; Prediksi Risiko

1. LATAR BELAKANG

Diabetes merupakan penyakit metabolik kronis yang meningkatkan risiko kesehatan ibu dan janin apabila tidak terdeteksi sejak dini (World Health Organization, 2023).Prevalensi diabetes secara global terus meningkat, dengan lebih dari 460 juta penderita pada 2021 dan diperkirakan mencapai 643 juta pada 2030 jika tidak ada intervensi preventif(International Diabetes Federation, 2024). *Gestational Diabetes Mellitus* (GDM) adalah kondisi peningkatan kadar glukosa yang pertama kali terdeteksi selama kehamilan, berbeda dengan diabetes tipe 1

Naskah Masuk: 10 Oktober 2025; Revisi: 18 Oktober 2025; Diterima: 28 November 2025; Terbit: 29 Desember 2025;

dan 2, dan berpotensi menimbulkan komplikasi obstetri serta masalah metabolik pada bayi (Mantri et al., 2024). Prevalensi GDM secara global berkisar antara 10–25%, dengan angka tertinggi terjadi di Asia Tenggara akibat obesitas dan perubahan pola makan (H et al., 2022).

Dalam memprediksi risiko GDM secara dini, *Machine Learning* terutama algoritma *Random Forest* terbukti efektif karena mampu menangkap pola *non-linier* yang sulit diidentifikasi oleh metode statistik konvensional (Zhang et al., 2022). Salah satu dataset yang relevan adalah Pima Indians, yang mencakup 768 data wanita hamil dengan delapan variabel klinis, sehingga cocok untuk studi prediksi GDM (UCI Machine learning, 2021).Algoritma *Random Forest* membangun banyak pohon keputusan, menyediakan estimasi *feature-importance,* dan meningkatkan transparansi pengaruh variabel klinis (Nassiwa & Zeng, n.d.).

Penelitian ini bertujuan mengidentifikasi risiko GDM menggunakan *Random Forest* dengan pendekatan split train/test 70 : 30 dan *Stratified K-Fold Cross Validation,* serta mengevaluasi performa model melalui ROC–AUC, *Confusion Matrix, F1-score, Accuracy, Precision,* dan *Recall.* Diharapkan hasil penelitian ini dapat menjadi dasar pengembangan *Clinical Decision Support System* (CDSS) untuk skrining dini GDM.

2. KAJIAN TEORITIS

Gestational Diabetes Mellitus

Diabetes *mellitus* merupakan penyakit metabolik kronis yang ditandai oleh hiperglikemia akibat gangguan sekresi insulin, resistensi insulin, atau kombinasi keduanya, yang berpotensi menimbulkan kerusakan organ vital seperti ginjal, jantung, mata, dan sistem saraf (World Health Organization, 2023). Secara global, diabetes menjadi salah satu penyebab utama kematian dengan jumlah penderita melebihi 460 juta orang dan diproyeksikan terus meningkat pada dekade mendatang (International Diabetes Federation, 2024). Salah satu bentuk diabetes yang spesifik pada kehamilan adalah *Gestational Diabetes Mellitus* (GDM), yaitu kondisi peningkatan kadar glukosa darah yang pertama kali terdeteksi selama masa kehamilan dan dapat meningkatkan risiko komplikasi obstetri seperti preeklamsia, kelahiran prematur, dan makrosomia (Mori & Pandey, 2022). Selain berdampak pada kehamilan, GDM juga meningkatkan risiko ibu mengalami diabetes tipe 2 di kemudian hari (Kaya et al., 2024).

Machine Learning

Machine learning (ML) merupakan cabang kecerdasan buatan yang memungkinkan sistem komputer mempelajari pola dari data historis dan melakukan prediksi atau klasifikasi tanpa pemrograman eksplisit (Alzubaidi et al., 2021). Dalam bidang kesehatan, ML banyak dimanfaatkan untuk membantu diagnosis, prediksi penyakit, serta pengembangan *Clinical Decision Support System* (CDSS) berbasis data klinis. Pendekatan ML terbukti lebih unggul dibandingkan metode statistik konvensional dalam menangkap hubungan *nonlinier* dan interaksi kompleks antarvariabel medis, termasuk dalam prediksi GDM (Wang, 2024).

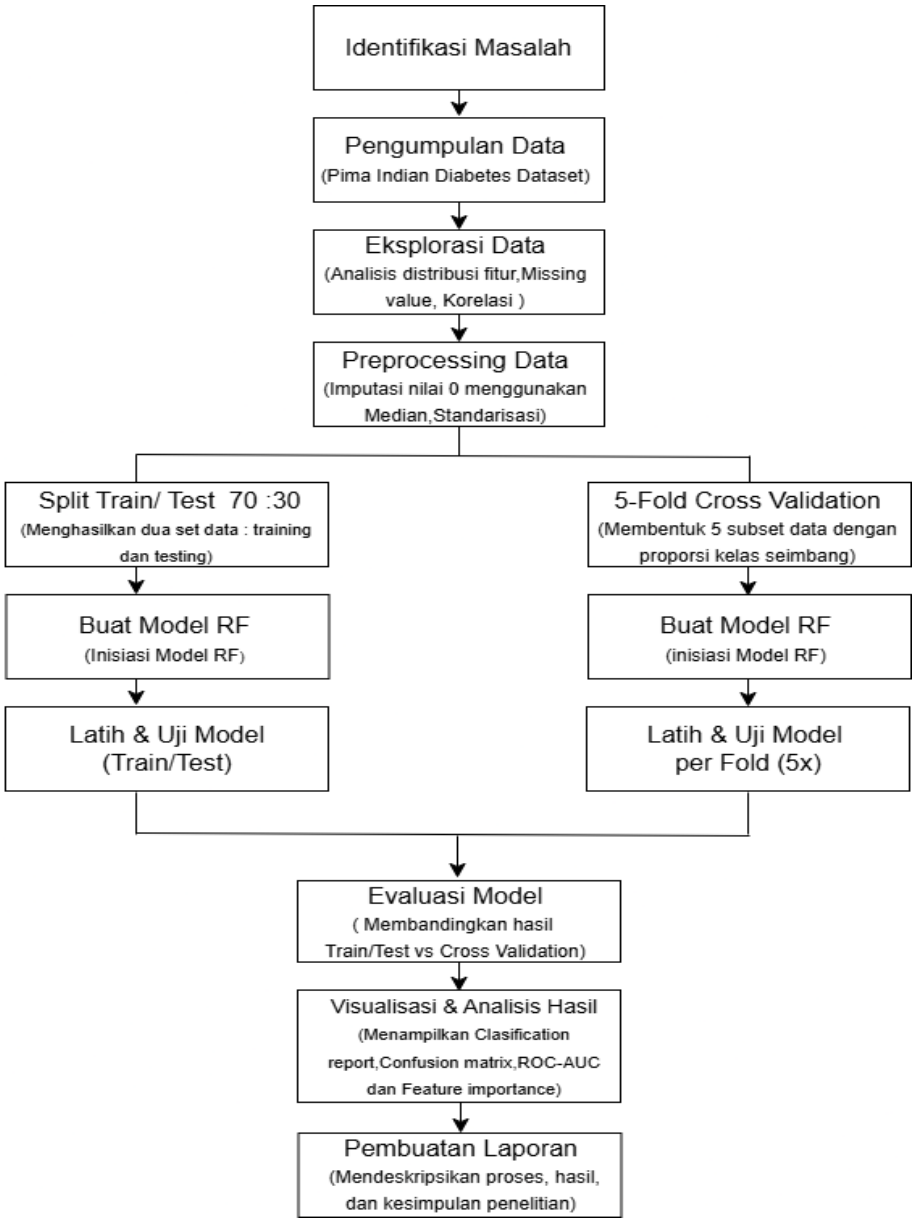
Random Forest

Random Forest merupakan algoritma *ensemble learning* berbasis *decision tree* yang membangun sejumlah pohon keputusan secara acak dan menggabungkan hasilnya melalui mekanisme *voting* untuk menghasilkan klasifikasi akhir yang lebih stabil. Pendekatan *ensemble* ini membantu mengurangi variansi model tunggal, meningkatkan kemampuan generalisasi, sekaligus tahan terhadap *overfitting* dalam kasus data kompleks dan berdimensi tinggi. Algoritma ini sering dipilih dalam domain medis karena performanya yang konsisten dalam memprediksi kondisi klinis serta kemampuannya menilai kontribusi variabel melalui *feature importance*, sehingga mendukung interpretasi faktor klinis yang berpengaruh dalam prediksi penyakit seperti diabetes dan *gestational diabetes mellitus* (GDM) (Xu, 2024).

***Preprocessing Data* Dan Evaluasi Model**

Kualitas dan kesiapan data sangat menentukan kinerja model ML. *Preprocessing data* bertujuan untuk menyiapkan dataset agar bebas dari nilai hilang, *outlier*, duplikasi, serta memastikan skala antar variabel konsisten. Tahapan ini meliputi penanganan *missing value*, standarisasi atau normalisasi, serta *feature engineering* untuk membentuk variabel prediktif baru yang lebih informatif (Reddy & Kumar, 2023). Evaluasi model klasifikasi dilakukan menggunakan metrik yang komprehensif, antara lain *accuracy, precision, recall, F1-score, confusion matrix,* dan ROC-AUC (Pham et al., 2023). *Cross-validation*, termasuk *StratifiedKFold*, digunakan untuk memperoleh estimasi performa yang stabil dengan mempertahankan distribusi kelas pada setiap *fold*, khususnya pada dataset yang tidak seimbang (Joseph & Vakayil, 2022).

3. METODE PENELITIAN



Gambar 1. Alur Penelitian

Diagram alur pada Gambar 1. menunjukkan tahapan penelitian yang dilakukan secara sistematis dalam membangun model klasifikasi risiko diabetes menggunakan algoritma *Random Forest.* Alur penelitian dimulai dari tahap identifikasi masalah hingga pembuatan laporan hasil penelitian.

Identifikasi Masalah

Tahap awal penelitian adalah identifikasi masalah, yang bertujuan untuk menentukan fokus penelitian dalam memprediksi risiko diabetes berdasarkan data klinis. Pada tahap ini dirumuskan permasalahan, tujuan penelitian, serta pendekatan pemodelan yang digunakan..

Pengumpulan Data

Tahap berikutnya adalah pengumpulan data, di mana data yang digunakan merupakan *Pima Indians Diabetes Dataset*. Dataset ini berisi data klinis wanita yang pernah mengalami kehamilan, dengan delapan atribut prediktor dan satu atribut target *(Outcome)* yang merepresentasikan status diabetes.

Eksplorasi Data

Setelah data dikumpulkan, dilakukan eksplorasi data untuk memahami karakteristik dataset. Tahap ini meliputi analisis distribusi fitur, identifikasi nilai tidak valid (*missing value* terselubung), serta analisis korelasi antar variabel untuk memperoleh gambaran awal hubungan data terhadap variabel target.

Pra-Pemrosesan Data

Tahap selanjutnya adalah pra-pemrosesan data, yang bertujuan meningkatkan kualitas data sebelum pemodelan. Nilai nol pada atribut klinis yang secara medis tidak valid diperlakukan sebagai *missing value* dan ditangani menggunakan teknik imputasi median. Selanjutnya dilakukan standarisasi fitur untuk menyeragamkan skala data.

Pengujian Model

Data yang telah diproses kemudian digunakan pada dua skema evaluasi, yaitu pembagian data *train-test* dengan rasio 70:30 dan *Stratified 5-Fold Cross Validation*. Pada masing-masing skema, dilakukan pembuatan model *Random Forest*, diikuti dengan proses pelatihan dan pengujian model baik pada data uji maupun pada setiap *fold*

Evaluasi Model

Tahap evaluasi model dilakukan dengan membandingkan hasil pengujian dari kedua skema evaluasi tersebut. Penilaian performa model dilakukan menggunakan metrik
Akurasi, presisi, *recall, F1-score,* serta ROC-AUC.

Visualisasi Dan Analisis Hasil

Selanjutnya dilakukan visualisasi dan analisis hasil, yang mencakup penyajian *classification report, confusion matrix*, kurva ROC-AUC, serta *feature importance* untuk mendukung interpretasi performa dan perilaku model.

Pembuatan Laporan

Tahap akhir penelitian adalah pembuatan laporan, yang menyajikan keseluruhan proses, hasil evaluasi, serta kesimpulan penelitian secara sistematis.

4. HASIL DAN PEMBAHASAN

Proses pengumpulan data

Penelitian ini menggunakan *Pima Indians Diabetes Dataset* dari repositori publik yang terdiri atas 768 *instances*, dengan delapan atribut prediktor dan satu target *(Outcome)* untuk analisis risiko diabetes. Deskripsi akan disajikan pada Tabel 1.

Tabel 1. Deskripsi Dataset

No	Nama Kolom	Deskripsi Fitur	Type Data
1	*Pregnancies*	Jumlah Kehamilan Pasien	Numerik
2	*Glucose*	Kadar Glukosa Plasma	Numerik
3	*Bloodpressure*	Tekanan Darah Diastolik	Numerik
4	*Skinthickness*	Ketebalan Lipatan Kulit Trisep	Numerik
5	*Insulin*	Kadar Insulin	Numerik
6	*BMI*	Indeks Massa Tubuh	Numerik
7	*Diabetespedigreefunction*	Risiko Diabetes Riwayat Keluarga	Numerik
8	*Age*	Usia Pasien	Numerik
9	*Outcome*	Label 1 = Diabetes,0 = Tidak Diabetes	Kategorikal

(Sumber: Data Olahan Peneliti,2025)

Hasil *Feature Engineering*

Bagian ini menyajikan *hasil feature engineering* berupa penambahan fitur turunan dari variabel asli melalui interaksi, rasio, dan transformasi *non-linear*. Fitur-fitur yang dihasilkan ditampilkan pada Tabel 2.

Tabel 2. Hasil *Feature Engineering*

Fitur Baru
Age_BMI
*Glucose_BMI_*
Preg_Age
Insulin_Glucose
Age2

(Sumber: Data Olahan Peneliti,2025)

Preprocessing Data

Pada tahap ini, dataset dibagi menjadi data pelatihan dan data pengujian menggunakan rasio 70:30 dengan teknik stratified split untuk menjaga keseimbangan kelas. Selanjutnya, fitur distandarisasi menggunakan *StandardScaler* guna memastikan konsistensi skala antar variabel.

Tabel 3. Hasil Pembagian data Latih dan Uji

Data Latih	Data Uji
537	231

(Sumber: Data Olahan Peneliti, 2025)

Pengujian *Baseline* Model

Pada tahap ini dilakukan pengujian baseline untuk mengevaluasi performa awal model *Random Forest* menggunakan data hasil *preprocessing*. Evaluasi dilakukan pada data pelatihan dan data pengujian menggunakan metrik klasifikasi.

Hasil Pengujian Pada Data Train

Pengujian pada data train dilakukan untuk mengetahui kemampuan awal model *Random Forest* dalam mempelajari pola data sebelum dilakukan proses optimasi. Evaluasi performa model didasarkan pada beberapa metrik, yaitu akurasi, *precision, recall,* dan *f1-score*, yang dihitung berdasarkan hasil prediksi terhadap data pelatihan.

Tabel 4. *Classification Report Train*

Class	*Precision*	*Recall*	*F1-Score*	*Support*
0	0.96	0.97	0.97	350
1	0.95	0.92	0.93	187
Accurasy		0.96		537
Macro Avg	0.95	0.95	0.85	537
Weighted Avg	0.96	0.96	0.96	537

(Sumber: Data Olahan Peneliti, 2025)

Hasil Pengujian Pada Data Test

Pengujian pada data test dilakukan untuk mengevaluasi kemampuan generalisasi model *Random Forest* terhadap data yang belum pernah dilihat sebelumnya. Evaluasi performa model didasarkan pada beberapa metrik, yaitu akurasi, *precision*, dan *recall*, guna menilai efektivitas model dalam melakukan klasifikasi pada data uji.

Tabel 5. *Classification Report Test*

Class	*Precision*	*Recall*	*F1-Score*	*Support*
0	0.88	0.92	0.90	150
1	0.84	0.78	0.81	81
Accurasy		0.87		231
Macro Avg	0.86	0.85	0.85	231
Weighted Avg	0.87	0.87	0.87	231

(Sumber : Data Olahan Peneliti, 2025)

Confusion matrix

Untuk menilai kemampuan model *Random Forest* dalam mengklasifikasikan pasien dengan dan tanpa diabetes secara lebih detail, digunakan *confusion matrix.* Matriks ini menunjukkan jumlah prediksi benar dan salah pada setiap kelas untuk data latih dan data uji, sehingga memudahkan analisis performa model secara spesifik.

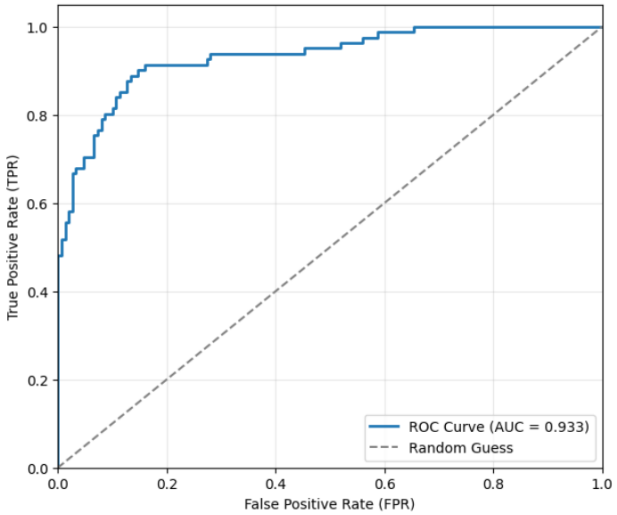
Tabel 6. *Confusion Matrix Train* Dan *Test*

Dataset	***Actual 0 Pred* 0**	***Actual 0 Pred* 1**	***Actual 1 Pred* 0**	***Actual 1 Pred* 1**
Train	341	9	15	172
Test	138	12	18	63

(Sumber : Data Olahan Peneliti, 2025)

ROC Curve

ROC Curve disajikan setelah *confusion matrix* untuk mengevaluasi kemampuan model dalam membedakan kelas diabetes dan tidak diabetes, yang diukur melalui nilai *Area Under Curve* (AUC).



Gambar 2. ROC Curve

Kurva ROC menunjukkan performa model yang sangat baik dengan nilai AUC sebesar 0,933, menandakan kemampuan diskriminasi kelas yang tinggi. Hal ini menunjukkan bahwa model memiliki probabilitas prediksi yang akurat dalam membedakan pasien diabetes dan tidak diabetes pada berbagai ambang keputusan. Dengan demikian, model *Random Forest* dapat dikatakan andal untuk digunakan sebagai alat bantu klasifikasi risiko diabetes. Namun, hasil ini masih bergantung pada karakteristik dataset yang digunakan sehingga perlu pengujian lebih lanjut pada data yang lebih beragam.

Pengujian Model Menggunakan *GridSearchCV*

Pengujian model dilakukan menggunakan *Grid Search Cross-Validation* untuk memperoleh kombinasi *hyperparameter* terbaik pada algoritma *Random Forest* dengan pendekatan *k-fold cross-validation* terstratifikasi.

Tabel 7. Parameter Terbaik Hasil *GridSearchCV*

Komponen	Hasil
Best Cv Accurasy	0.8902
Best Max_Depth	*None*
Best_Min_Samples_Leaf	2
Best_Min_Samples_Split	10
Best N_Estimators	150
Total Kombinasi Dicoba	108
Total Proses Pelatihan (Fits)	540

(Sumber: Data Olahan Peneliti, 2025)

A. Pengujian Model Menggunakan *5 Fold Cross Validation*

Pengujian model dilakukan menggunakan *5-fold cross-validation* terstratifikasi dengan menggunakan *hyperparameter* terbaik hasil *GridSearchCV* untuk memperoleh estimasi performa model yang lebih stabil dan objektif.

Tabel 8. Hasil Uji Model Menggunakan *5-Fold Cross-Validation*

Metrik Evaluasi	Nilai Rata-rata
Accurasy	0.88
AUC	0.93
Precision	0.83
Recall	0.81
F1-Score	0.82

(Sumber: Data Olahan Peneliti. 2025)

Feature Importance

Feature importance disajikan untuk menunjukkan kontribusi relatif setiap fitur dalam proses pengambilan keputusan model *Random Forest*. Nilai *importance* yang lebih tinggi mengindikasikan bahwa fitur tersebut memiliki pengaruh yang lebih besar terhadap hasil klasifikasi risiko diabetes.

Tabel 9. Feature Importance

Feature	Importance
Insulin	0.329302
Glucose_BMI	0.113291
Glucose	0.107461
SkinThickness	0.100392
Insulin_Glucose	0.098641
Age_BMI	0.064166
Age2	0.034264
BMI	0.033389
Preg_Age	0.030958
DiabetesPedigreeFunction	0.026778
Age	0.026400
BloodPressure	0.017991
Pregnancies	0.016966

(Sumber: Data Olahan Peneliti, 2025)

B. Perbandingan Kinerja Model *Baseline* Dan Model *5-Fold Cross Validation*

perbandingan ini menyajikan perbedaan kinerja antara model *baseline* dan model *Random Forest* dengan *5-fold cross-validation* berdasarkan metrik evaluasi utama untuk menilai peningkatan performa dan stabilitas model.

Tabel 10. Perbandingan Kinerja Model

Metode Evaluasi	*Accurasy*	*Precision*	*Recall*	*F1-Score*	AUC
***Baseline* (70:30)**	0.87	0.84	0.78	0.81	0.93
5-Fold CV + GridSearch	0.88	0.83	0.81	0.82	0.93

(Sumber: Data Olahan Peneliti, 2025)

Berdasarkan Tabel 10, model *Random Forest* yang dievaluasi menggunakan pendekatan *5-fold cross-validation* dengan optimasi *GridSearch* menunjukkan peningkatan kinerja dibandingkan model *baseline*. Meskipun nilai AUC pada kedua pendekatan sama, yaitu sebesar 0,93, model dengan validasi silang memiliki akurasi, *recall*, dan *F1-score* yang lebih tinggi. Peningkatan nilai *recall* pada kelas diabetes menunjukkan bahwa pendekatan *5-fold cross-validation* mampu meningkatkan kemampuan model dalam mendeteksi kasus diabetes secara lebih konsisten dan stabil.

5. KESIMPULAN DAN SARAN

Hasil penelitian menunjukkan bahwa algoritma *Random Forest* efektif dalam memprediksi risiko *Gestational Diabetes Mellitus* (GDM) pada wanita hamil. Model *baseline* dengan pembagian data 70:30 memberikan performa awal yang baik, namun model yang dioptimasi menggunakan GridSearchCV dan dievaluasi dengan *5-fold cross-validation* menunjukkan kinerja yang lebih unggul dan stabil. Peningkatan nilai *recall* pada kelas diabetes mengindikasikan bahwa pendekatan validasi silang mampu meningkatkan kemampuan model dalam mendeteksi kasus GDM secara lebih optimal, sehingga mengurangi risiko kasus positif yang tidak teridentifikasi. Selain itu, model teroptimasi menghasilkan akurasi rata-rata sebesar 88% dengan nilai AUC sebesar 0,93, yang mencerminkan kemampuan diskriminasi kelas yang tinggi. Keterbatasan penelitian ini terletak pada ukuran dan karakteristik dataset yang masih terbatas, sehingga hasil penelitian perlu diinterpretasikan secara hati-hati. Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan beragam, menambahkan variabel klinis yang relevan, serta mengkaji implementasi model dalam sistem pendukung keputusan klinis untuk skrining dini GDM.

DAFTAR REFERENSI

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00444-8

H, W., N, L., T, C., M, W., H, S., L, Y., & X, Y. (2022). IDF diabetes atlas: Estimation of global and regional gestational diabetes mellitus prevalence for 2021 by International Association of Diabetes in Pregnancy Study Group's criteria. *Diabetes Research and Clinical Practice*, *183*.

International Diabetes Federation. (2024). *IDF Diabetes Atlas*. IDF.

Joseph, V. R., & Vakayil, A. (2022). SPlit: An Optimal Method for Data Splitting. *Technometrics*, *64*(2), 166–176. https://doi.org/10.1080/00401706.2021.1921037

Kaya, Y., Bütün, Z., Çelik, Ö., Salik, E. A., Tahta, T., & Yavuz, A. A. (2024). The early prediction of gestational diabetes mellitus by machine learning models. *BMC Pregnancy and Childbirth*, *24*(1). https://doi.org/10.1186/s12884-024-06783-7

Mantri, N., Goel, A. D., Patel, M., Baskaran, P., Dutta, G., Gupta, M. K., Yadav, V., Mittal, M., Shekhar, S., & Bhardwaj, P. (2024). National and regional prevalence of gestational diabetes mellitus in India: a systematic review and Meta-analysis. *BMC Public Health*, *24*(1). https://doi.org/10.1186/s12889-024-18024-9

Mori, R., & Pandey, A. (2022). Global burden of early pregnancy gestational diabetes mellitus (eGDM): prevalence, risk factors and outcomes. *Acta Diabetologica*, *59*(4), 453–462. https://pubmed.ncbi.nlm.nih.gov/34743219/

Nassiwa, F., & Zeng, J. (n.d.). *Evaluating Traditional Machine Learning Models for Predicting Diabetes Onset Using the Pima Indians Dataset*. https://ssrn.com/abstract=4878052

Pham, H. H., Nguyen, H. Q., Nguyen, H. T., Le, L. T., & Lam, K. (2023). *Evaluating the impact of an explainable machine learning system on the interobserver agreement in chest radiograph interpretation*. http://arxiv.org/abs/2304.01220

Reddy, A. A., & Kumar, P. (2023). Feature selection and feature engineering strategies for diabetes prediction. *Journal of Biomedical Informatics*.

UCI Machine learning. (2021). *Pima Indians Diabetes Database*. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?utm_source=chatgpt.com

Wang, W. (2024). *Principles of Machine Learning: The Three Perspectives* (Springer Nature).

World Health Organization. (2023). *Diabetes fact sheet*. WHO. https://www.who.int/news-room/fact-sheets/detail/diabetes

Xu, Y. (2024). Random Forest-based clinical decision support for gestational diabetes prediction and feature interpretation. *IEEE Access*, *12*.

Zhang, Z., Yang, L., Han, W., Wu, Y., Zhang, L., Gao, C., Jiang, K., Liu, Y., & Wu, H. (2022). Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis. *Journal of Medical Internet Research*, *24*(3). https://doi.org/10.2196/26634