VOL. 3. NO. 01 JANUARI 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

SISTEM KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST BERDASARKAN DATA MEDIS

Hafist Rezallul Fikry[1], Rudi Kurniawan[2], Bani Nurhakim[3], Umi Hayati[4].

Program Studi Teknik Informatika[124]
Program Studi Manajemen Informatika[3]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
fikry.zalul77@gmail.com

(*) Corresponding Author : fikry.zalul77@gmail.com
Published : 30 Januari 2026

Abstract—*This study discusses the development of a classification model for early detection of diabetes mellitus using the Random Forest algorithm. Diabetes is a chronic disease with an increasing prevalence, making early detection crucial to prevent complications. The use of data mining and machine learning enables more accurate analysis of medical data to support the diagnosis process. The research approach uses a supervised learning method with a dataset containing health attributes such as glucose levels, blood pressure, BMI, age, and other relevant medical variables. The data is processed through pre-processing stages, including cleaning, transformation, normalization, and division of data into 80% training data and 20% test data. The Random Forest model was built as a classification algorithm, while its performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics. The results of the experiment showed that the model achieved an accuracy of 96.76%, with variations in precision and recall between classes indicating data imbalance and feature complexity. Although the accuracy is not yet optimal, this study confirms that Random Forest remains a potential base model in medical decision support systems. The model's performance can still be improved through parameter optimization, data balancing, and improved pre-processing quality. This approach is expected to be the first step in developing data analytics for early detection of diabetes*

Keywords : *Random Forest, Early Detection, Diabetes Mellitus, Medical Data, Classification, machine learning*

Abstrak—Penelitian ini membahas pembangunan model klasifikasi untuk deteksi dini diabetes mellitus menggunakan algoritma Random Forest. Diabetes merupakan penyakit kronis dengan prevalensi yang terus meningkat, sehingga deteksi awal menjadi krusial untuk mencegah komplikasi. Pemanfaatan data mining dan machine learning memungkinkan analisis data medis secara lebih akurat dalam mendukung proses diagnosis. Pendekatan penelitian menggunakan metode supervised learning dengan dataset berisi atribut kesehatan seperti kadar glukosa, tekanan darah, BMI, usia, dan variabel medis relevan lainnya. Data diproses melalui tahapan pre-processing, meliputi pembersihan, transformasi, normalisasi, serta pembagian data menjadi 80% data latih dan 20% data uji. Model Random Forest dibangun sebagai algoritma klasifikasi, sementara kinerjanya dievaluasi dengan metrik akurasi, presisi, recall, F1-score, dan confusion matrix. Hasil eksperimen menunjukkan bahwa model mencapai akurasi 96,76%, dengan variasi presisi dan recall antar kelas yang mengindikasikan adanya ketidakseimbangan data dan kompleksitas fitur. Meskipun akurasinya belum optimal, penelitian ini menegaskan bahwa Random Forest tetap potensial sebagai model dasar dalam sistem pendukung keputusan medis. Kinerja model masih dapat ditingkatkan melalui optimasi parameter, penyeimbangan data, serta peningkatan kualitas pre-processing. Pendekatan ini diharapkan menjadi langkah awal pengembangan analitik data untuk deteksi dini diabetes.

Kata Kunci : Random Forest, Deteksi Dini, Diabetes Mellitus, Data Medis, Klasifikasi, machine learning

INTRODUCTION

Dalam penelitian yang berfokus pada deteksi awal diabetes menggunakan algoritma *Random Forest*, sejumlah tantangan utama muncul. [1] menyoroti pentingnya metode seleksi fitur yang berfungsi untuk meningkatkan akurasi dalam pengklasifikasian data diabetes. Namun, mereka juga mencatat bahwa keterbatasan dalam data yang tersedia dapat mempengaruhi hasil analisis ini. [2] menemukan bahwa analitik *big data* dapat menawarkan akurasi yang moderat hingga tinggi dalam diagnosis diabetes, tetapi

tantangan signifikan masih ada, termasuk kebutuhan akan sumber data yang baik dan terstruktur serta keamanan data pasien. Selain itu, penelitian oleh Suryanegara et al. [3] menunjukkan bahwa pengolahan data yang tidak sesuai dapat mengurangi efektivitas algoritma *Random Forest*, sebab diabetes yang semakin meningkat di Indonesia memerlukan deteksi dini yang lebih akurat. Hal serupa didiskusikan oleh Hadi dan Sirodj [4], yang menekankan pentingnya penggunaan metode pengoptimalan dalam *Random Forest* untuk menanggulangi permasalahan klasifikasi. Keterbatasan pada dataset juga menjadi tantangan utama karena dapat mempengaruhi kemampuan klasifikasi dalam penelitian diabetes [5].

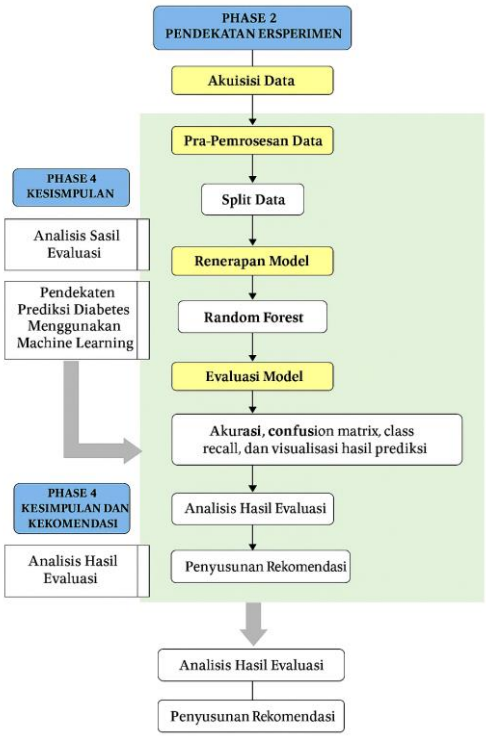
machine learning (ML) telah secara signifikan meningkatkan diagnosis medis, terutama dalam pengelolaan diabetes, di mana algoritma seperti *Random Forest* (RF) menunjukkan efektivitas tinggi dalam pemodelan prediktif. [6] Melaporkan bahwa klasifikasi RF mencapai akurasi 97,77% dalam prediksi risiko diabetes pada tahap awal, menunjukkan potensi pendekatan ML dalam meningkatkan strategi skrining dan pengelolaan dalam perawatan diabetes. Selain itu, Rohman dkk. [7] menekankan peran penyesuaian hiperparameter dalam mengoptimalkan kinerja RF, menunjukkan bahwa teknik seperti *Grid Search* dapat meningkatkan akurasi, meskipun studi mereka terutama melaporkan akurasi maksimum 75% sebelum penyesuaian. Hal ini menyarankan potensi perbaikan tetapi tidak memberikan tingkat peningkatan kinerja yang sama seperti klaim awal 98%. Bukti kolektif ini menyoroti peran integral *machine learning* dalam deteksi dini dan pengelolaan diabetes, memperkuat aplikasinya dalam pengaturan klinis.

Studi-studi terbaru telah menunjukkan potensi berbagai jenis data medis dalam memprediksi diabetes menggunakan teknik *machine learning*, khususnya melalui penggunaan algoritma *Random Forest*. Wang [8] menyoroti keefektifan model *machine learning* seperti *Random Forest* dalam memprediksi diabetes, menunjukkan kinerjanya yang unggul dibandingkan teknik tradisional di lingkungan perawatan kesehatan. Demikian pula, Li [9] membandingkan beberapa model berbasis pohon keputusan, menegaskan bahwa *Random Forest* menghasilkan hasil yang menjanjikan dalam mengidentifikasi faktor risiko diabetes dari dataset yang kompleks. Ramadhan dkk. [10] memperdalam pembahasan dengan fokus pada pentingnya prapemrosesan data dalam meningkatkan hasil pembelajaran mesin untuk deteksi diabetes, menekankan peran masukan medis yang beragam. Noviyanti dan Alamsyah [11] mengonfirmasi temuan ini dengan menyajikan data yang menyoroti kapasitas prediktif *Random Forest*, menggambarkan ketahanannya dalam deteksi dini diabetes. Selain itu, Jumanto dkk. [12] menegaskan relevansi memasukkan berbagai indikator kesehatan untuk memperkuat model prediksi diabetes, menunjukkan pendekatan komprehensif dalam memanfaatkan data medis secara efektif.

Penggunaan kecerdasan buatan (AI) dalam manajemen dan pencegahan diabetes telah berkembang pesat, terutama dalam pengembangan alat prediksi dan diagnosis yang lebih baik. Referensi [13] menjelaskan bagaimana AI berperan dalam manajemen pasien diabetes selama pandemi, dengan fokus pada pemantauan glukosa darah menggunakan perangkat yang terhubung Bluetooth dan analisis data secara real-time. Selain itu, referensi [14] menyoroti bagaimana AI menganalisis multi-level dari data glukosa, yang mendukung terapi yang lebih tepat dan pencegahan diabetes lebih dini. Kemajuan dalam metodologi AI juga telah berlaku untuk pengelolaan komplikasi diabetes, seperti yang dibahas oleh referensi [15], yang meninjau aplikasi AI untuk mendeteksi kaki diabetes. Perkembangan ini tidak hanya menguntungkan diagnosis dini, tetapi juga meningkatkan perawatan pasien dengan memanfaatkan *big data* untuk analisis risiko, seperti yang dijelaskan oleh referensi [16] dan [17]. Dengan semakin banyaknya penemuan di bidang ini, ada potensi besar untuk meningkatkan kualitas hidup dan outcome kesehatan bagi pasien diabetes.

MATERIALS AND METHODS

Metode yang digunakan dalam penelitian ini adalah pendekatan kuantitatif eksperimental, yang bertujuan untuk menguji hipotesis melalui analisis terkontrol dengan memanfaatkan data numerik dan pemodelan berbasis *machine learning*. Penelitian ini secara khusus mengimplementasikan algoritma *Random Forest* untuk melakukan prediksi klasifikasi pasien diabetes, memanfaatkan dataset uji yang dibaca melalui RapidMiner Studio [18]. Desain penelitian dibagi ke dalam 4 (empat) tahapan, yakni 1) Tahap Studi Pendahuluan, 2) Tahap Pencarian Literature, 3) Tahap Pemilihan Metode, 4) Tahap Kesimpulan dan Rekomendasi. Keempat tahapan tersebut dapat disajikan pada Gambar 1.



Gambar 1. Tahapan Desain Penelitian

Berdasarkan Gambar 1 diagram tersebut menggambarkan alur penelitian eksperimental dalam klasifikasi diabetes menggunakan algoritma *Random Forest*. Penelitian ini terdiri dari empat fase utama: studi pendahuluan, kajian literatur, pendekatan eksperimen, serta penarikan kesimpulan dan rekomendasi.

Pada Fase pertama, Studi Pendahuluan, Meliputi perumusan masalah, penentuan tujuan penelitian, dan penyusunan pertanyaan penelitian untuk memprediksi kategori pasien diabetes.

Fase Kedua, Kajian Literatur, Mengulas penelitian terkait penggunaan algoritma *machine learning* seperti *Random Forest* dan Decision Tree untuk prediksi diabetes, yang menjadi dasar pemilihan metode dan evaluasi model.

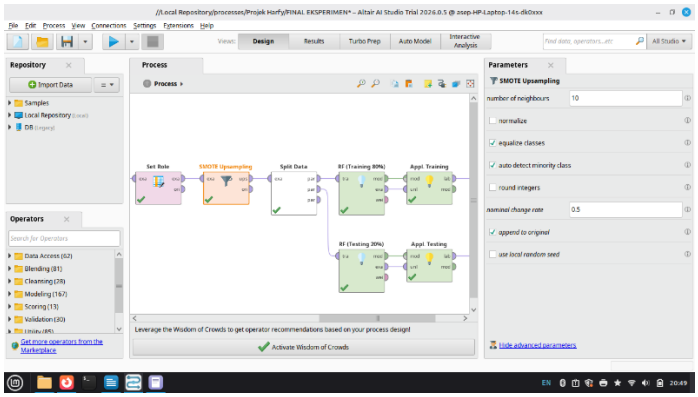
Fase ketiga Pendekatan Eksperimen, Dataset pasien diabetes diproses di RapidMiner melalui tahapan akuisisi data, pembagian data (80% training, 20% testing), pembangunan model *Random Forest*, dan evaluasi menggunakan akurasi, *confusion matrix*, serta *class recall*. Hasil eksperimen menunjukkan akurasi 96,76%.

Fase Keempat Kesimpulan dan Rekomendasi, Menganalisis hasil evaluasi, menarik kesimpulan, dan memberikan rekomendasi pengembangan model lanjutan seperti *hyperparameter tuning* atau perbandingan dengan algoritma lain.

RESULTS AND DISCUSSION

Tahap *data mining* merupakan proses utama dalam penelitian ini yang bertujuan untuk membangun model klasifikasi menggunakan algoritma *Random Forest*. Sebelum proses pemodelan dilakukan, terlebih dahulu diterapkan teknik penyeimbangan data menggunakan metode *Synthetic Minority Oversampling Technique (SMOTE)*.

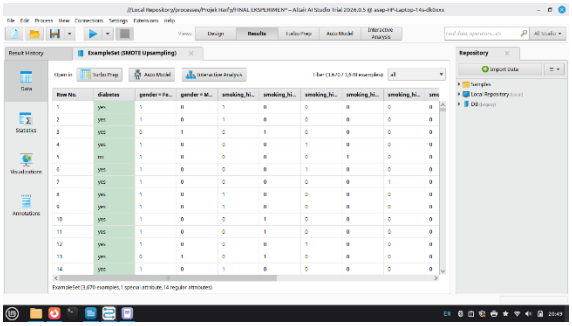
Proses ini dilakukan menggunakan operator *SMOTE Upsampling* untuk mengatasi ketidakseimbangan jumlah data antar kelas pada dataset. Parameter yang digunakan pada operator ini yaitu *number of neighbours* sebesar 10, strategi *equalize classes*, serta pengaturan *auto detect minority class* untuk mendeteksi kelas minoritas secara otomatis. Selain itu digunakan parameter *nominal change rate* sebesar 0.5 dengan opsi *append to original* sehingga data hasil sintesis akan ditambahkan ke dataset asli.



Gambar 2 Metode *SMOTE* untuk *Upsampling*

Dataset awal yang digunakan dalam penelitian ini berjumlah 1.600 data. Setelah dilakukan proses *SMOTE upsampling*, jumlah data meningkat menjadi 3.670 data. Hal ini bertujuan untuk menghasilkan distribusi kelas yang lebih seimbang sehingga model klasifikasi dapat mempelajari pola data dengan lebih baik.

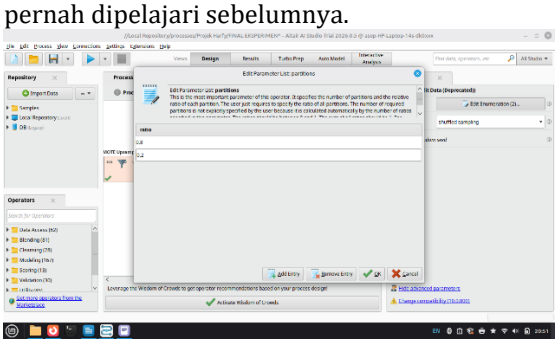
Setelah proses penyeimbangan data selesai dilakukan, dataset kemudian dibagi menjadi data latih dan data uji menggunakan operator Split Data. Proporsi pembagian data yang digunakan dalam penelitian ini adalah 80% data training dan 20% data testing, dengan metode *sampling* menggunakan *shuffled sampling* agar data diacak sebelum proses pembagian.



Gambar 3 *Hasil Upsampling*

Setelah proses penyeimbangan data menggunakan metode *SMOTE* selesai dilakukan, langkah berikutnya adalah membagi dataset menjadi data latih dan data uji menggunakan operator *Split Data* pada RapidMiner. Proses ini bertujuan untuk memisahkan sebagian data yang digunakan untuk melatih model dengan sebagian data lainnya yang digunakan untuk menguji performa model yang telah dibangun.

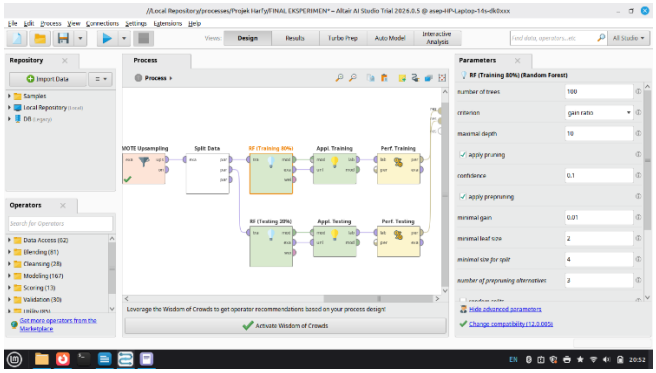
Pada penelitian ini, dataset yang telah melalui proses *SMOTE* dan berjumlah 3.670 data kemudian dibagi dengan proporsi 80% data training dan 20% data testing. *Data training* digunakan untuk membangun model klasifikasi menggunakan algoritma *Random Forest*, sedangkan data *testing* digunakan untuk mengevaluasi kemampuan model dalam melakukan prediksi terhadap data yang belum pernah dipelajari sebelumnya.



Gambar 4 *Split data*

Model klasifikasi kemudian dibangun menggunakan algoritma *Random Forest*. Parameter yang digunakan dalam pembentukan model meliputi *number of trees* sebanyak 100, *criterion* menggunakan gain ratio, serta *maximal depth* sebesar 10. Selain itu diterapkan proses *pruning* dengan nilai *confidence* sebesar 0.1 untuk mengurangi kompleksitas model. Proses *prepruning* juga diterapkan dengan parameter *minimal gain* sebesar 0.001, *minimal leaf size* sebesar 2, *minimal size of split* sebesar 4, serta *number of prepruning alternatives* sebanyak 3.

Proses pemodelan *Random Forest* pada RapidMiner dalam penelitian ini dapat dilihat pada Gambar X.



Gambar 5 *Modelling*

Berdasarkan hasil pengujian menggunakan data *training*, diperoleh nilai performa model *Random Forest* seperti yang ditunjukkan pada Tabel 1

Tabel 1 Metrik Evaluasi *Data Training*

Metric	Nilai
Accuracy	96.76%
Classification Error	3.24%
Weighted Mean Recall	96.76%
Weighted Mean Precision	96.77%

Nilai accuracy sebesar 96.76% menunjukkan bahwa model *Random Forest* mampu mengklasifikasikan *data training* dengan tingkat ketepatan yang sangat tinggi. Sementara itu nilai *classification error* sebesar 3.24% menunjukkan bahwa tingkat kesalahan model dalam melakukan klasifikasi relatif kecil.

Nilai *weighted mean recall* sebesar 96.76% menunjukkan bahwa model memiliki kemampuan yang baik dalam mendeteksi masing-masing kelas dengan mempertimbangkan distribusi jumlah data pada setiap kelas. Selain itu nilai *weighted mean precision* sebesar 96.77% menunjukkan bahwa sebagian besar prediksi yang dihasilkan oleh model merupakan prediksi yang benar.

Hasil *confusion matrix* dari pengujian data training dapat dilihat pada Tabel 2

Tabel 2 *Confussion Matrix Data Training*

Actual / Prediction	Yes	No
Yes	1411	54
No	41	1430

Berdasarkan tabel tersebut dapat diketahui bahwa model berhasil memprediksi 1411 data diabetes (yes) dengan benar dan 1430 data non-diabetes (no) dengan benar. Namun masih terdapat beberapa kesalahan prediksi yaitu 54 data diabetes yang diprediksi sebagai non-diabetes serta 41 data non-diabetes yang diprediksi sebagai diabetes.

Selain menggunakan data *training*, evaluasi model juga dilakukan pada data *testing* untuk mengetahui kemampuan model dalam melakukan prediksi terhadap data baru. Hasil evaluasi model

menggunakan data testing dapat dilihat pada Tabel 3

Tabel 3 Metrik Evaluasi *Data Testing*

Metric	Nilai
Accuracy	98.09%
Classification Error	1.91%
Weighted Mean Recall	98.10%
Weighted Mean Precision	98.11%

Berdasarkan hasil evaluasi tersebut, diperoleh nilai accuracy sebesar 98.09%, yang menunjukkan bahwa model *Random Forest* mampu melakukan klasifikasi dengan tingkat ketepatan yang sangat tinggi pada data yang belum pernah dipelajari sebelumnya.

Nilai *classification error* sebesar 1.91% menunjukkan bahwa tingkat kesalahan model dalam melakukan prediksi relatif sangat kecil. Selain itu nilai *weighted mean recall* sebesar 98.10% dan *weighted mean precision* sebesar 98.11% menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mendeteksi serta memprediksi kelas secara akurat.

Hasil *confusion matrix* dari pengujian data testing dapat dilihat pada Tabel 4.6.

Tabel 4 *Confussion Matrix Data Testing*

Actual / Prediction	Yes	No
Yes	359	11
No	3	361

Berdasarkan tabel tersebut dapat diketahui bahwa model berhasil memprediksi 359 data diabetes (yes) dengan benar serta 361 data non-diabetes (no) dengan benar. Sementara itu terdapat 11 data diabetes yang diprediksi sebagai non-diabetes serta 3 data non-diabetes yang diprediksi sebagai diabetes.

CONCLUSION

Penelitian ini bertujuan untuk membangun sistem klasifikasi penyakit diabetes menggunakan algoritma Random Forest berdasarkan data medis pasien. Dataset yang digunakan terdiri dari beberapa atribut penting, yaitu age, blood glucose, HbA1c level, BMI, heart disease, hypertension, smoking history, dan gender. Proses pengolahan data dilakukan melalui beberapa tahapan, meliputi data preprocessing, data transformation, data mining, dan data evaluation. Untuk mengatasi ketidakseimbangan data, diterapkan metode SMOTE sehingga jumlah data meningkat dari 1600 menjadi 3670 data dan distribusi kelas menjadi lebih seimbang.

Hasil evaluasi menunjukkan bahwa model Random Forest memiliki performa yang sangat baik dalam mendeteksi penyakit diabetes. Pada data training diperoleh nilai accuracy sebesar 96,76% dengan classification error sebesar 3,24%, sedangkan pada data testing diperoleh accuracy sebesar 98,09% dengan classification error sebesar 1,91%. Nilai precision dan recall yang tinggi serta hasil confusion matrix menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar, baik pada kelas diabetes maupun non-diabetes. Hal ini menunjukkan bahwa algoritma Random Forest berpotensi digunakan sebagai sistem pendukung dalam deteksi dini penyakit diabetes.

REFERENCE

[1] S. Raghavendra and S. K. J, "Performance Evaluation of Random Forest With Feature Selection Methods in Prediction of Diabetes," *International Journal of Electrical and Computer Engineering (Ijece)*, vol. 10, no. 1, p. 353, 2020, doi: 10.11591/ijece.v10i1.pp353-359.

[2] I. J. B. do Nascimento *et al.*, "Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future Studies (Preprint)," 2021, doi: 10.2196/preprints.27275.

[3] G. A. B. Suryanegara, A. Adiwijaya, and M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi Pada Algoritma Random Forest Untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[4] D. A. Hadi and D. A. N. Sirodj, "Metode Random Forest Untuk Klasifikasi Penyakit Diabetes," *Bandung Conference Series Statistics*, vol. 3, no. 2, pp. 428–435, 2023, doi: 10.29313/bcss.v3i2.8354.

[5] S. Alkhalefah, I. Al-Turaiki, and N. Altwaijry, "Advancing Diabetic Foot Ulcer Care: AI and Generative AI Approaches for Classification, Prediction, Segmentation, and Detection," *Healthcare*, vol. 13, no. 6, p. 648, 2025, doi: 10.3390/healthcare13060648.

[6] P. V. S. Kumar and N. S. Kumar, "Analysis and Comparison for Prediction of Diabetic Among Pregnant Women Using Innovative Support Vector Machine Algorithm Over Random Forest Algorithm With Improved Accuracy," *Cm*, no. 25, pp. 956–962, 2023, doi: 10.18137/cardiometry.2022.25.956962.

[7] F. N. Rohman, F. Farikhin, and B. Surarso, "Hyperparameter Tuning of Random Forest Algorithm for Diabetes Classification," *International Journal of Current Science Research and Review*, vol. 08, no. 01, 2025, doi: 10.47191/ijcsrr/v8-i1-31.

[8] M. Wang, H. Cao, Z. Ai, and Q. Zhang, "Fault diagnosis of ship ballast water system based on support vector machine optimized by improved sparrow search algorithm," *IEEE Access*, vol. 12, pp. 17045–17057, 2024, doi: 10.1109/access.2024.3351171.

[9] L. Li, "Comparative Research on Diabetes Influencing Factors Based on Random Forest and Decision Tree Models," *Highlights in Science Engineering and Technology*, vol. 72, pp. 231–242, 2023, doi: 10.54097/7m4x7j04.

[10] N. G. Ramadhan, A. Adiwijaya, W. Maharani, and A. A. Gozali, "Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review," *Ieee Access*, vol. 12, pp. 80698–80730, 2024, doi: 10.1109/access.2024.3406748.

[11] C. N. Noviyanti and A. Alamsyah, "Early Detection of Diabetes Using Random Forest Algorithm," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, 2024, doi: 10.52465/joiser.v2i1.245.

[12] J. Jumanto, M. A. Muslim, Y. Dasril, and T. Mustaqim, "Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest," *J. Inf. Syst. Explor. Res.*, vol. 1, no. 1, pp. 49–70, 2022, doi: 10.52465/joiser.v1i1.104.

[13] L. P. Sari, N. Darmawulan, F. Agustina, and N. Nursiswati, "Application of Artificial Intelligence for Managing Diabetes Mellitus Patients During the Covid-19 Pandemic," *Journal of Nursing Care*, vol. 5, no. 2, 2023, doi: 10.24198/jnc.v5i2.36971.

[14] A. Sadagopan, "Artificial Intelligence Driving Diabetes Care," *Journal for International Medical Graduates*, vol. 2, no. 1, 2023, doi: 10.56570/jimgs.v2i1.92.

[15] G. Chemello, B. Salvatori, M. Morettini, and A. Tura, "Artificial Intelligence Methodologies Applied to Technologies for Screening, Diagnosis and Care of the Diabetic Foot: A Narrative Review," *Biosensors*, vol. 12, no. 11, p. 985, 2022, doi: 10.3390/bios12110985.

[16] Z. Guan *et al.*, "Artificial Intelligence in Diabetes Management: Advancements, Opportunities, and Challenges," *Cell Reports Medicine*, vol. 4, no. 10, p. 101213, 2023, doi: 10.1016/j.xcrm.2023.101213.

[17] S. Ziajor *et al.*, "The Use of Artificial Intelligence in the Diagnosis and Detection of Complications of Diabetes," *Journal of Education Health and Sport*, vol. 65, pp. 11–27, 2024, doi: 10.12775/jehs.2024.65.001.

[18] K. Thanigainathan, "USING ENSEMBLE CLUSTERING TO IDENTIFY PHENOTYPES," no. March, 2022.