Deteksi Penderita Diabetes dengan Algoritma *Random Forest* dan *Backward Elimination*

Vic Jeremy Prajogo[1], Siska Narulita[2*]

[1,2]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Nasional Karangturi, Semarang, Indonesia
Email: [2]siskanarulita84@gmail.com

ABSTRACT

Diabetes is a chronic condition for which appropriate treatment depends on precise early detection. By combining the random forest algorithm with the backward elimination methodology as one of the feature selection methods, this work seeks to optimize the diabetic detection process. The dataset, which includes 768 samples with 9 attributes such as blood pressure, body mass index, glucose levels, and other risk factors was obtained from a public database on Kaggle. Data preparation to guarantee dataset quality, pre-processing with Backward Elimination for the best feature selection, Random Forest algorithm implementation for classification, and performance evaluation with a confusion matrix comprise the four primary phases of the research technique. The results showed a significant improvement in model performance after the implementation of backward elimination, with accuracy increasing from 83.08% to 99.78% and precision from 79.37% to 99.67%, while recall remained consistent at 100%. Optimization using backward elimination proved effective in eliminating features that contribute less to prediction accuracy, resulting in a more efficient and accurate model. These findings indicate that the combination of Random Forest with Backward Elimination not only improves the accuracy of diabetes detection substantially but also has the potential to be implemented in clinical decision support systems to aid early diagnosis of diabetes.

***Keywords**: Backward Elimination; Random Forest; Diabetes Detection; Machine Learning; Feature Selection; Classification; Model Optimization*

.

ABSTRAK

Diabetes merupakan salah satu penyakit kronis yang membutuhkan deteksi dini secara akurat untuk penanganan yang tepat. Penelitian ini bertujuan untuk mengoptimalkan proses deteksi penderita diabetes menggunakan kombinasi algoritma *Random Forest* dengan teknik *Backward Elimination* sebagai salah satu metode *feature selection*. *Dataset* yang digunakan berasal dari *database* publik yang diambil dari Kaggle, terdiri dari 768 sampel dengan 9 atribut, termasuk kadar glukosa, tekanan darah, indeks massa tubuh, dan faktor risiko lainnya. Metodologi penelitian meliputi empat tahap utama, data preparation untuk memastikan kualitas *dataset*, *pre-processing* menggunakan *Backward Elimination* untuk seleksi fitur optimal, implementasi algoritma *Random Forest* untuk klasifikasi, dan evaluasi performa menggunakan *confusion matrix*. Hasil penelitian menunjukkan peningkatan signifikan dalam performa model setelah implementasi *Backward Elimination*, dengan peningkatan *accuracy* dari 83,08% menjadi 99,78%, *precision* dari 79,37% menjadi 99,67%, sementara *recall* tetap konsisten pada 100%. Optimasi menggunakan *Backward Elimination* terbukti efektif dalam mengeliminasi fitur-fitur yang kurang berkontribusi terhadap akurasi prediksi, menghasilkan model yang lebih efisien dan akurat. Temuan ini mengindikasikan bahwa kombinasi *Random Forest* dengan *Backward Elimination* tidak hanya meningkatkan akurasi deteksi penderita diabetes secara substansial, tetapi juga berpotensi untuk diimplementasikan dalam sistem pendukung keputusan klinis untuk membantu diagnosis dini diabetes.

***Kata kunci:** Backward Elimination; Random Forest;* Deteksi Diabetes*; Machine Learning;* Seleksi Fitur*;* Klasifikasi*;* Optimalisasi Model

PENDAHULUAN

Diabetes merupakan salah satu diantara penyakit kronis yang semakin meningkat nilai prevalensinya di seluruh dunia, dengan estimasi mencapai 425 juta kasus pada tahun 2020.[1] Penyakit ini dikenal dengan sebutan "*mother of diseases*" dikarenakan dapat menjadi pemicu atas berbagai komplikasi penyakit serius, seperti penyakit jantung, stroke, dan neuropati diabetik, yang dapat berujung kepada kematian jika tidak segera terdeteksi dan ditangani dengan baik.[2][3] Di Indonesia, nilai prevalensi diabetes mencapai 10,6%, yaitu sekitar 19,47 juta orang terdiagnosis diabetes.[4] Meningkatnya angka kejadian penyakit diabetes menuntut perhatian serius dalam hal deteksi dini dan pengelolaan penyakit ini.

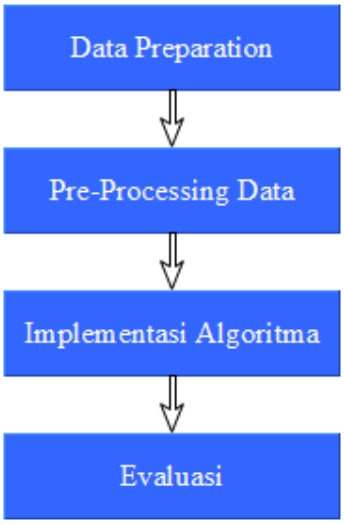
Deteksi dini terhadap penderita diabetes sangat penting sebagai upaya untuk pencegahan komplikasi penyakit yang lebih serius. Penelitian yang dilakukan menunjukkan bahwa intervensi yang tepat pada tahap awal dapat mengurangi terjadinya risiko komplikasi hingga 50%.[1] Oleh karena itu, penelitian ini berfokus pada optimasi metode *data mining Random Forest* untuk deteksi penderita diabetes. *Data mining* itu sendiri merupakan proses ekstraksi informasi atau pengetahuan dari kumpulan data yang kompleks dan besar.[5][6] Sedangkan algoritma *Random Forest* merupakan salah satu teknik dalam *data mining* yang telah terbukti efektif dalam mengklasifikasi data dengan akurasi yang tinggi. Cara kerja *Random Forest* dengan membuat banyak pohon keputusan, kemudian hasilnya digabungkan untuk meningkatkan akurasi hasil prediksi serta mengurangi adanya risiko *overfitting*.[7][8]

Beberapa penelitian terdahulu yang menggunakan algoritma *Random Forest* untuk deteksi penderita diabetes, antara lain dilakukan oleh Mawarni *et al*., (2023) yang berjudul Deteksi Dini Gejala Awal Penyakit Diabetes Menggunakan Algoritma *Random Forest*. Penelitian ini melaporkan akurasi sebesar 90,38%, dengan tingkat *precision* 100%, *recall* 84,38%, serta nilai AUC 1,00. Penelitian ini juga menggunakan *dataset* publik dari Kaggle yang terdiri dari 520 *record* dengan 17 atribut untuk klasifikasi diabetes.[9] Penelitian kedua oleh Iskandar *et al*., (2024) berjudul Klasifikasi Menggunakan Metode *Random Forest* untuk Awal Deteksi Diabetes Melitus Tipe 2. Dalam studi ini, model *Random Forest* dapat mencapai akurasi rata-rata sebesar 97% melalui proses K-*Fold Cross Validation*, dengan nilai *precision* 95%, *recall* 97%, serta F1-*Score* 96%. Penelitian ini menggunakan *dataset private* dari UPTD Puskesmas Jatiroto yang terdiri dari 1.111 data.[10] Penelitian ketiga oleh Junus *et al*., (2024) berjudul Klasifikasi Menggunakan Metode *Support Vector Machine* dan *Random Forest* untuk Deteksi Awal Risiko Diabetes Melitus. Penelitian ini menunjukkan bahwa algoritma *Random Forest* menghasilkan akurasi sebesar 98,08%, *recall* 97,87%, *precision* 98.92%, dan F1-*Score* 88,40%. Data yang digunakan berasal dari Sylhet Diabetic Hospital di Bangladesh dengan total 520 data pasien diabetes.[11] Penelitian keempat dilakukan oleh Suryanegara *et al*., (2021) berjudul Peningkatan Hasil Klasifikasi pada Algoritma *Random Forest* untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. Penelitian ini menemukan bahwa model dengan *min-max normalization* mencapai akurasi tertinggi sebesar 95,45%, diikuti oleh model dengan z-*score normalization* sebesar 95%. Penelitian ini menekankan pentingnya normalisasi data dalam meningkatkan performa algoritma *Random Forest*.[12] Berdasarkan penelitian-penelitian tersebut, dapat dilihat bahwa algoritma *Random Forest* menunjukkan performa yang sangat baik dalam mendeteksi diabetes, dengan akurasi yang bervariasi antara 90% hingga lebih dari 98%. Hal ini menunjukkan potensi besar algoritma ini untuk deteksi dini terhadap penyakit diabetes.

Sebagai bagian dari penelitian ini, teknik *feature selection Backward Elimination* digunakan untuk memilih fitur yang paling relevan dari sekumpulan data yang besar. Teknik ini dapat membantu mengurangi kompleksitas model dengan cara menghilangkan variabel yang tidak signifikan, sehingga dapat meningkatkan interpretabilitas dan akurasi model secara keseluruhan. Penelitian sebelumnya menunjukkan bahwa penggunaan teknik *feature selection* dapat meningkatkan performa algoritma dalam bidang Kesehatan.[3][8]

METODE

Penelitian ini dilakukan dalam beberapa tahap yaitu, *data preparation*, *pre-processing data*, implementasi algoritma, dan evaluasi. Gambar 1 berikut merupakan alur dari metode penelitian ini:



Gambar 1. Metode Penelitian

1. *Data Preparation*

Sumber data penelitian ini berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases* yang diperoleh dari *platform* Kaggle. Tujuan dari pengumpulan data ini adalah untuk dilakukan prediksi secara diagnostik apakah pasien terkena penyakit diabetes atau tidak, berdasarkan pengukuran diagnostik tertentu yang diikutsertakan dalam kumpulan data ini. Secara khusus, semua pasien dalam dataset di sini adalah perempuan berusia minimal 21 tahun yang berasal dari suku Indian Pima.

2. *Pre-Processing Data*

Pre-processing data pada penelitian ini dilakukan dengan melakukan pembagian *data training* dan *data testing* dengan persentase rasio 60% banding 40%, serta menerapkan teknik *feature selection* untuk mengoptimasi kinerja algoritma *machine learning Random Forest*. Metode *feature selection* yang digunakan adalah *Backward Elimination*.

3. Implementasi Algoritma

Setelah dilakukan *pre-processing data*, *dataset* yang telah siap tersebut diimplementasikan pada algoritma *Random Forest* untuk melakukan prediksi penderita diabetes melitus. Adapun langkah-langkah perhitungan pada algoritma *Random Forest* adalah:

a. Mencari nilai *impurity* menggunakan persamaan (1) berikut:

$$Gini(A) = 1 - \sum_{i=1}^{m} P_i^2 \quad \text{…….....…} (1)$$

Dimana:

P_i : Nilai peluang dari sebuah nilai *tuple* A pada suatu kelas

m : Jumlah label kelas

b. Menentukan nilai indeks gini menggunakan persamaan (2) berikut:

$$Gini_B(A) = \frac{|A_1|}{|A|} Gini(A_1) + \frac{|A_2|}{A} Gini(A_2) \quad (2)$$

c. Menghitung *impurity reduction* melalui persamaan (3) berikut ini:

$$\Delta Gini(B) = Gini(A) - Gini_B(A) \quad \text{…} (3)$$

4. Evaluasi

Evaluasi model pada penelitian ini menggunakan *multiple* metrik performa (*confusion matrix*), termasuk di dalamnya akurasi (*accuracy*), presisi (*precision*), dan *recall*. Nilai akurasi dapat dihitung menggunakan persamaan (4) di bawah ini:

$$\frac{(TP + TN)}{(TP + FP + FN + TN)} \quad \text{……..……...…} (4)$$

Persamaan (5) digunakan untuk menghitung tingkat presisi model:

$$\frac{(TP)}{(TP + FP)} \quad \text{……...…..…..……} (5)$$

Sedangkan untuk menghitung *recall* dapat digunakan persamaan (6) berikut ini:

$$\frac{(TP)}{(TP + FN)} \quad \text{……...…...…...……} (6)$$

Dimana:

TP : *True Positive*
TN : *True Negative*
FP : *False Positive*
FN : *False Negatif*

HASIL

1. *Data Preparation*

Proses pengumpulan data mendapatkan *dataset* diabetes dari *repository public* Kaggle, di mana dalam *dataset* diabetes tersebut terdiri atas sembilan atribut, antara lain *pregnancies*, *glucose*, *blood pressure*, *skin thickness*, insulin, BMI,

diabetes pedigree function, *age*, dan *outcome* dengan jumlah 768 *record data*, tidak ada *missing value* dalam *dataset* ini, dan untuk *class* (label) ditentukan atribut *outcome*.

2. *Pre-Processing Data*

Pada tahap *pre-processing data*, dilakukan splitting data dengan persentase rasio sebesar 60% banding 40% untuk *data training* dan *data testing*, sehingga diperoleh jumlah *record data* pada *data training* sebesar 461, sedangkan pada *data testing* sebesar 307 *record data*. Pada proses *feature selection* untuk mengoptimalkan algoritma *Random Forest* menggunakan teknik *Backward Elimination*. Hasil dari tahap ini, terjadi reduksi jumlah atribut yang semula berjumlah sembilan menjadi empat atribut, yaitu *outcome*, *glucose*, *skin thickness*, dan *age*.

3. Implementasi Algoritma

Pada penelitian ini, algoritma *Random Forest* diimplementasikan sebelum dan setelah *dataset* melalui *pre-processing data* (penggunaan teknik *Backward Elimination*). Tabel 1 berikut ini menunjukkan hasil *accuracy* algoritma *Random Forest* sebelum penggunaan teknik *Backward Elimination.*

Tabel 1. *Confusion Matrix Accuracy* Sebelum Penambahan *Backward Elimination*

	True 1	True 0	Class Precision
Pred. 1	83	0	100%
Pred. 0	78	300	79,37%
Class Recall	51,55%	100%	
Accuracy: 83,08%			

Tabel 2 berikut ini menunjukkan hasil *precision* algoritma *Random Forest* sebelum penggunaan teknik *Backward Elimination*.

Tabel 2. *Confusion Matrix Precision* Sebelum Penambahan *Backward Elimination*

	True 1	True 0	Class Precision
Pred. 1	83	0	100,00%
Pred. 0	78	300	79,37%
Class Recall	51,55%	100,00%	
Precision: 79,37%			

Tabel 3 berikut ini menunjukkan hasil *recall* algoritma *Random Forest* sebelum penggunaan teknik *Backward Elimination.*

Tabel 3. *Confusion Matrix Recall* Sebelum Penambahan *Backward Elimination*

	True 1	True 0	Class Precision
Pred. 1	83	0	100%
Pred. 0	78	300	79,37%
Class Recall	51,55%	100%	
Recall: 100%			

4. Evaluasi

Pada tahap sebelumnya, yaitu implementasi algoritma *Random Forest*, sudah dilakukan penerapan algoritma pada proses prediksi *dataset* sebelum ada penambahan *feature selection Backward Elimination*. Pada tahap ini dilakukan proses evaluasi terhadap implementasi algoritma *Random Forest* yang sudah ditambahkan teknik *feature selection Backward Elimination*. Tabel 4 menunjukkan hasil *accuracy* algoritma *Random Forest* setelah penggunaan teknik *Backward Elimination*.

Tabel 4. *Confusion Matrix Accuracy* Sesudah Penambahan *Backward Elimination*

	True 1	True 0	Class Precision
Pred. 1	160	0	100%
Pred. 0	1	300	99,67%
Class Recall	99,38%	100%	
Accuracy: 99,78%			

Tabel 5 berikut ini menunjukkan hasil *precision* algoritma *Random Forest* sesudah penggunaan teknik *Backward Elimination*.

Tabel 5. *Confusion Matrix Precision* Sesudah Penambahan *Backward Elimination*

	True 1	True 0	Class Precision
Pred. 1	160	0	100,00%
Pred. 0	1	300	99,67%
Class Recall	99,38%	100,00%	
Precision: 99,67%			

Tabel 6 berikut ini menunjukkan hasil *recall* algoritma *Random Forest* setelah penggunaan teknik *Backward Elimination*.

Tabel 6. *Confusion Matrix Recall* Setelah Penambahan *Backward Elimination*

	True 1	True 0	Class Precision
Pred. 1	160	0	100,00%
Pred. 0	1	300	99,67%
Class Recall	99,38%	100,00%	
Recall: 100,00%			

PEMBAHASAN

Setelah dilakukan percobaan pada dataset diabetes dengan menggunakan algoritma *Random Forest* baik sebelum maupun sesudah penggunaan teknik *feature selection Backward Elimination* pada tahap *pre-processing data*, ditemukan bahwa terdapat perbedaan akurasi yang dihasilkan antara proses yang menggunakan teknik *feature selection Backward Elimination* dan tanpa penggunaan *Backward Elimination*. Tabel 7 berikut ini memperlihatkan hasil perbandingan implementasi algoritma *Random Forest* dengan dan tanpa teknik *feature selection Backward Elimination*.

Tabel 7. Perbandingan Model *Random Forest* Menggunakan *Feature Selection Backward Elimination* dan Tanpa *Feature Selection*

Parameter	RF	RF + BE
Accuracy	83,08%	99,78%
Precision	79,37%	99,67%
Recall	100,00%	100,00%

SIMPULAN

Penelitian ini telah berhasil menunjukkan bahwa penerapan teknik feature selection Backward Elimination pada implementasi algoritma Random Forest secara signifikan dapt meningkatkan kinerja algoritma Random Forest pada dataset diabetes. Hasil evaluasi menunjukkan peningkatan accuracy dari 83,08% menjadi 99,78%, serta peningkatan precision dari 79,37% menjadi 99,67%, sementara *recall* tetap pada 100% untuk kelas positif. Peningkatan ini menegaskan bahwa pentingnya pemilihan atribut dalam pengembangan model *machine learning*, yang pada prakteknya tidak hanya mengurangi kompleksitas tetapi juga dapat meningkatkan keandalan model klasifikasi. Berdasarkan hasil penelitian yang dilakukan, teknik *feature selection Backward Elimination* terbukti efektif dalam meningkatkan performa algoritma *Random Forest*, sehingga menjadikannya sebagai alat yang lebih handal dalam proses analisis data.

Untuk penelitian selanjutnya, dapat dilakukan penggunaan *feature selection* lainnya, seperti *Forward Selection*, *Recursive Feature Elimination*, Regresi Lasso, Regresi Ridge, dan lain sebagainya. Atau juga dapat dilakukan dengan penggunaan algoritma prediksi lain, misalnya Naïve Bayes, k-NN, C4.5, k-*Means*, Apriori, dan lain sebagainya untuk meningkatkan performansi pada deteksi penyakit diabetes.

DAFTAR PUSTAKA

1. Prasetyo SY, Santy, Yunanda R. Diabetes Risk Prediction Exploration: Uncovering Patterns and Enhancing Predictive Accuracy through Ensemble Learning. In: 2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA). Balikpapan: Global Technopreneur Campus; 2024. p. 213–8.
2. Nugraha MD, Ramdhani YN, Utami M. Hubungan Dukungan Keluarga dengan Tingkat Distres pada Lansia Penderita Diabetes Melitus Tipe 2 di Wilayah Kerja Puskesmas Kuningan Tahun 2023. Journal of Nursing Practice and Education. 2023;4(1):177–84.
3. Sami A, Naseer A, Hussain MZ, Hasan MZ, Mustafa M, Khalid A, et al. Enhancing Diabetes Detection: A Weighted Averaging Approach for Combined Model Accuracy. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). Pune, India: IEEE; 2024. p. 1–5.
4. Rastipiati, Nugraha MD, Purnama R. Pengaruh Terapi Air Putih Hangat dan Air Putih Biasa terhadap Penurunan Kadar Gula Darah Sewaktu (GDS) pada Lansia Diabetes Melitus di Desa Luragung Landeuh Kecamatan Luragung Kabupaten Kuningan Tahun 2023. National Nursing Conference. 2023;1(2):85–102.
5. Rahayu PW, Sudipa IGI, Suryani, Surachman A, Ridwan A, Darmawiguna IGM, et al. Buku Ajar Data Mining. Bandung: PT Sonpedia Publishing Indonesia; 2024.
6. Narulita S, Adi PN. Feature Selection Information Gain pada Klasifikasi Pasien Penyakit Jantung (Heart Disease). JURMIK: Jurnal Rekam Medis dan Manajemen

Informasi Kesehatan. 2024;4(1):13–9.
7. Sari L, Romadloni A, Listyaningrum R. Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. Infotekmesin. 2023;14(1):155–62.
8. Wan X, Liu Y, Yang L, Zeng C, Hao D. Sleep Apnea Detection Method Based on Improved Random Forest. International Journal of Advanced Computer Science and Applications(IJACSA). 2023;14(11):594–600.
9. Mawarni AC, Rusdah, Hin LL, Anubhakti D. Deteksi Dini Gejala Awal Penyakit Diabetes Menggunakan Algoritma Random Forest. Idealis: Indonesia Journal Information System. 2023;6(2):165–71.
10. Iskandar RFN, Gutama DH, Wijaya DP, Danianti D. Klasifikasi Menggunakan Metode Random Forest untuk Awal Deteksi Diabetes Melitus Tipe 2. JUTIN: Jurnal Teknik Industri Terintegrasi. 2024;7(3):1620–6.
11. Junus CZV, Tarno, Kartikasari P. Klasifikasi Menggunakan Metode Support Vector Machine dan Random Forest untuk Deteksi Awal Risiko Diabetes Melitus. Jurnal Gaussian Universitas Diponegoro Semarang. 2022;11(3):386–96.
12. Suryanegara GAB, Adiwijaya K, Purbolaksono MD. Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. Resti: Rekayasa Sistem dan Teknologi Informasi. 2021;5(1):114–22.