BINER
Jurnal Ilmiah Informatika dan Komputer
Vol. 5 | No. 2 | Juli 2026

p-ISSN : 2828-0210 | e-ISSN : 2828-0229
https://ojs.unsiq.ac.id/index.php/biner

OPTIMASI SELEKSI FITUR PADA RANDOM FOREST UNTUK PREDIKSI RISIKO DIABETES TAHAP AWAL MENGGUNAKAN SFS, SBS, DAN PSO

Mediana Nurlailli[1)], Sucipto [2)], Anita Sari Wardani [3)]
[1,2,3)] Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri
Jl. Ahmad Dahlan No.76, Mojoroto, Kec. Mojoroto, Kota Kediri, Jawa Timur 64112
Email: mediananurlailli@gmail.com [1)], sucipto@unpkediri.ac.id[2)], anita@unpkediri.ac.id [1)]

Dikirim: 10 Mei 2026; Disetujui: 4 Juli 2026; Dipublikasikan: 31 Juli 2026

ABSTRAK

Preprocessing merupakan tahapan penting dalam machine learning karena kualitas data mempengaruhi performa model klasifikasi. Salah satu teknik *preprocessing* yang sering digunakan adalah feature selection untuk memilih atribut yang relevan dan mengurangi fitur yang kurang berpengaruh terhadap proses klasifikasi. Penelitian ini bertujuan melakukan optimasi seleksi fitur menggunakan metode *Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), dan Particle Swarm Optimization (PSO*) pada algoritma *Random Forest* untuk prediksi risiko diabetes tahap awal. Dataset yang digunakan adalah *Early Stage Diabetes Risk Prediction Dataset* yang ditransformasikan ke bentuk numerik menggunakan Label *Encoding*. Hasil penelitian menunjukkan bahwa *Random Forest* tanpa seleksi fitur menghasilkan accuracy sebesar 98.27% dengan 16 fitur, metode SFS menghasilkan *accuracy* sebesar 98.65% dengan 11 fitur, sedangkan metode SBS memperoleh *accuracy* tertinggi sebesar 99.23% menggunakan 11 fitur dan PSO memperoleh *accuracy* sebesar 98.85% menggunakan 11 fitur. Selain meningkatkan *accuracy*, seleksi fitur juga mampu mengurangi jumlah atribut sehingga proses klasifikasi menjadi lebih efisien.

Kata Kunci : feature selection, Sequential Forward Selection, Sequential Backward Selection, Particle Swarm Optimization, Random Forest, diabetes.

ABSTRACT

Preprocessing is an important stage in machine learning because data quality affects the performance of classification models. One preprocessing technique commonly used is feature selection to identify relevant attributes and reduce features that have less influence on the classification process. This study aims to optimize feature selection using Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), and Particle Swarm Optimization (PSO) methods on the Random Forest algorithm for early-stage diabetes risk prediction. The dataset used in this study is the Early Stage Diabetes Risk Prediction Dataset, which was transformed into numerical form using Label Encoding. The results show that Random Forest without feature selection achieved 98.27% accuracy using 16 features, SFS achieved 98.65% accuracy using 11 features, SBS produced the highest accuracy of 99.23% using 11 features, and PSO achieved 98.85% accuracy using 11 features. In addition to improving accuracy, feature selection also reduced the number of attributes, making the classification process more efficient.

Keywords *: feature selection, Sequential Forward Selection, Sequential Backward Selection, Particle Swarm Optimization, Random Forest, diabetes.*

1. PENDAHULUAN

Perkembangan teknologi informasi menyebabkan jumlah dan ukuran data terus meningkat setiap waktu. Data yang dihasilkan akan menjadi kurang bermanfaat apabila tidak diolah dengan baik, sehingga diperlukan penerapan *machine learning* untuk mengubah data menjadi informasi dan pengetahuan yang dapat digunakan dalam mendukung pengambilan keputusan. Dalam proses *machine learning*, data yang digunakan tidak selalu berada dalam kondisi siap pakai sehingga diperlukan tahap *preprocessing* sebelum dilakukan proses klasifikasi atau prediksi. Tahap *preprocessing* memiliki peran penting karena kualitas data sangat mempengaruhi performa model *machine learning* yang dihasilkan. [1]

Diabetes merupakan salah satu penyakit kronis yang terus mengalami peningkatan setiap tahunnya dan menjadi perhatian penting dalam bidang kesehatan. Penyakit ini terjadi ketika kadar gula darah dalam tubuh meningkat akibat gangguan produksi maupun fungsi insulin sehingga dapat menyebabkan berbagai komplikasi serius apabila tidak ditangani sejak dini. Oleh karena itu, deteksi dini terhadap risiko diabetes menjadi hal yang penting agar penanganan dapat dilakukan lebih cepat dan tepat. Dengan berkembangnya teknologi informasi dan *machine learning*, proses prediksi penyakit saat ini dapat dilakukan secara lebih efektif melalui pemanfaatan data kesehatan pasien [1].

Salah satu proses p*reprocessing* yang sering digunakan adalah *dimensionality reduction* atau reduksi dimensi. Reduksi dimensi dilakukan untuk mengurangi jumlah fitur pada dataset sehingga model dapat bekerja lebih optimal. *Dimensionality reduction* terdiri dari *feature extraction* dan *feature selection. Feature selection* merupakan proses memilih fitur yang paling relevan dan menghilangkan fitur yang tidak relevan maupun redundan pada dataset. Salah satu metode *feature selection* yang banyak digunakan adalah *wrapper method*, yaitu metode yang melakukan evaluasi fitur berdasarkan performa model klasifikasi secara langsung menggunakan proses *training* dan *testing* secara bertahap [2].

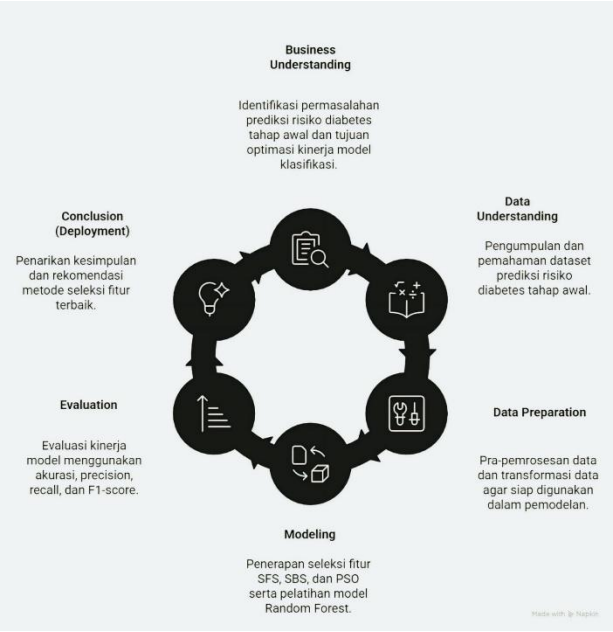
Dalam bidang kesehatan, *machine learning* banyak dimanfaatkan untuk membantu proses prediksi penyakit, salah satunya pada prediksi risiko diabetes tahap awal. Dataset diabetes dapat digunakan untuk memprediksi apakah seseorang berpotensi mengalami diabetes atau tidak berdasarkan gejala yang dimiliki. Salah satu algoritma klasifikasi yang sering digunakan dalam kasus kesehatan adalah *Random Forest* karena memiliki kemampuan yang baik dalam menghasilkan akurasi tinggi dan mampu mengurangi *overfitting*. Random Forest bekerja dengan menggabungkan beberapa *decision tree* untuk menghasilkan prediksi yang lebih stabil melalui proses *voting* dari setiap pohon keputusan [3].

Pada penelitian sebelumnya yang dilakukan oleh Nugroho, Fanani, dan Shidik, metode *Sequential Forward Selection* (SFS) dan *Sequential Backward Selection* (SBS) digunakan sebagai *wrapper feature selection* pada algoritma *Random Forest* untuk prediksi diabetes tahap awal. Penelitian tersebut menunjukkan bahwa penggunaan seleksi fitur mampu meningkatkan performa model klasifikasi dibandingkan tanpa seleksi fitur. Selain itu, metode *Particle Swarm Optimization* (PSO) juga dapat digunakan untuk mencari kombinasi fitur optimal melalui proses optimasi berbasis populasi sehingga model dapat bekerja lebih efektif dan efisien [8].

Penelitian ini diharapkan mampu menghasilkan model klasifikasi dengan performa yang lebih baik serta dapat membantu proses deteksi dini diabetes secara lebih akurat.

2. METODE

Penelitian ini menggunakan pendekatan *CRISP-DM(Cross-Industry Standard Process for Data Mining*), yang merupakan metodologi standar dalam proses penggalian data (data mining) yang sistematis dan terstruktur. *CRISP-DM* terdiri dari enam tahapan utama, yaitu: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Pendekatan ini telah terbukti efektif dalam meningkatkan akurasi prediksi dan memastikan konsistensi dalam proses pengembangan model [4]. Metodologi ini dikembangkan oleh sebuah konsorsium perusahaan di bawah inisiatif Komisi Eropa pada tahun 1996, sebagai kerangka kerja umum untuk memandu penerapan analisis data secara efektif [5].



Gambar 1. Alur CRISP-DM

2.1. Business Understanding

Pada tahap *Business Understanding* dilakukan proses identifikasi permasalahan serta penentuan tujuan penelitian. Permasalahan utama dalam penelitian ini adalah masih adanya fitur yang kurang relevan pada dataset prediksi diabetes tahap awal sehingga dapat mempengaruhi performa model klasifikasi. Penggunaan seluruh atribut pada dataset juga dapat meningkatkan kompleksitas model dan memperlambat proses komputasi. Oleh karena itu, penelitian ini bertujuan untuk melakukan optimasi seleksi fitur menggunakan metode *Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), dan Particle Swarm Optimization (PSO)* pada *algoritma Random Forest* agar diperoleh model klasifikasi dengan performa yang lebih baik dan lebih efisien. Tahap ini juga menentukan tujuan *data mining* yaitu menghasilkan model prediksi risiko diabetes tahap awal yang memiliki nilai *accuracy, precision, recall, dan F1-score* yang optimal [8].

2.2. Data Understanding

Tahap Data Understanding bertujuan untuk memahami karakteristik dataset yang digunakan dalam penelitian. Dataset yang digunakan berasal dari *Early Stage Diabetes Risk Prediction Dataset* yang diperoleh dari *UCI Machine Learning Repository*. Dataset terdiri dari 520 data dengan 17 atribut yang meliputi usia, jenis kelamin, serta berbagai gejala diabetes *seperti polyuria, polydipsia, weakness, visual blurring, alopecia,* dan atribut target berupa *class* yang menunjukkan status risiko diabetes.

Tabel 1. Deskripsi Dataset

Fitur		Type Data
Nomer	**Nama**	
0	*Age*	*Numeric*
1	*Sex*	*Nominal (Male or Female)*
2	*Polyuria*	*Nominal (yes or no)*
3	*Polydipsia*	*Nominal (yes or no)*
4	*Sudden Weight Loss*	*Nominal (yes or no)*
5	*Weakness*	*Nominal (yes or no)*
6	*Polyphagia*	*Nominal (yes or no)*
7	*Genetial Thrush*	*Nominal (yes or no)*
8	*Visual Blurring*	*Nominal (yes or no)*
9	*Itching*	*Nominal (yes or no)*
10	*Irritability*	*Nominal (yes or no)*
11	*Delayed Healing*	*Nominal (yes or no)*
12	*Partial Paresis*	*Nominal (yes or no)*
13	*Musele*	*Nominal (yes or no)*
14	*Stiffness*	*Nominal (yes or no)*
15	*Obesity*	*Nominal (yes or no)*
16	*Class*	*Nominal (positive or negative)*

Pada tahap ini dilakukan analisis struktur data, identifikasi tipe data, eksplorasi distribusi data, serta pengecekan kualitas data seperti *missing value* dan duplikasi data. Tahapan ini dilakukan untuk memastikan dataset siap digunakan pada proses pemodelan [6].

2.3. Data Preparation

Tahap Data *Preparation* mencakup seluruh proses persiapan data sebelum dilakukan pemodelan. Tahapan ini memiliki peran penting karena kualitas data sangat mempengaruhi performa model klasifikasi yang dihasilkan. Proses pertama yang dilakukan adalah transformasi data menggunakan Label *Encoding* untuk mengubah data kategorikal menjadi data numerik agar dapat diproses oleh algoritma *machine learning*. Setelah proses transformasi selesai, data dipisahkan menjadi fitur (X) dan target (y). Fitur terdiri dari seluruh atribut selain *class*, sedangkan target merupakan label klasifikasi risiko diabetes.

Tahap berikutnya adalah pembagian data menjadi data latih dan data uji menggunakan metode *train-test split* dengan proporsi 70:30. Pembagian ini bertujuan agar model dapat diuji menggunakan data yang belum pernah dipelajari sebelumnya sehingga hasil evaluasi mampu menggambarkan performa model secara umum. Selain itu, pada tahap ini juga dilakukan proses seleksi fitur menggunakan metode *Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), dan Particle Swarm Optimization (PSO)* untuk memperoleh kombinasi atribut yang paling optimal dalam proses klasifikasi [4].

2.4. Modeling

Tahap Modeling merupakan proses pembangunan dan pelatihan model klasifikasi menggunakan algoritma *Random Forest*. Algoritma ini dipilih karena memiliki kemampuan yang baik dalam menangani data klasifikasi serta mampu menghasilkan prediksi yang stabil dan akurat. *Random Forest* bekerja menggunakan kumpulan *decision tree* untuk menghasilkan prediksi akhir melalui proses voting sehingga mampu mengurangi *overfitting* pada model klasifikasi [10].

Pada penelitian ini dilakukan empat skenario pemodelan, yaitu *Random Forest* tanpa seleksi fitur, *Random Forest* dengan *Sequential Forward Selection (SFS)*, *Random Forest* dengan *Sequential Backward Selection (SBS*), dan *Random Forest* dengan *Particle Swarm Optimization (PSO)*. Metode SFS melakukan proses penambahan fitur secara bertahap berdasarkan kontribusi terbaik terhadap model, sedangkan SBS melakukan pengurangan fitur yang kurang relevan dari keseluruhan atribut yang tersedia. Sementara itu, PSO digunakan untuk mencari kombinasi fitur optimal melalui proses iterasi partikel dan evaluasi *fitness function* sehingga diperoleh subset fitur terbaik untuk proses klasifikasi [7].

2.5. Evaluation

Tahap *Evaluation* dilakukan untuk mengukur performa model klasifikasi yang telah dibangun. Evaluasi model dilakukan menggunakan beberapa metrik yaitu *accuracy, precision, recall,* dan *F1-score* agar hasil klasifikasi dapat dianalisis secara lebih menyeluruh. Selain itu, penelitian ini juga menggunakan *confusion matrix* untuk melihat jumlah prediksi benar dan salah yang dihasilkan model.

Confusion matrix terdiri dari empat komponen utama yaitu *True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).* TP menunjukkan jumlah data positif yang berhasil diprediksi dengan benar, TN menunjukkan jumlah data negatif yang diprediksi dengan benar, FP menunjukkan data negatif yang diprediksi positif, sedangkan FN menunjukkan data positif yang diprediksi negatif. Penggunaan metrik evaluasi tersebut bertujuan untuk mengetahui tingkat ketepatan model dalam melakukan prediksi risiko diabetes tahap awal [9].

2.6. Deployment

Tahap *Deployment* merupakan tahap akhir dalam metodologi *CRISP-DM* yang bertujuan untuk merancang implementasi model ke dalam sistem nyata. Pada penelitian ini tahap deployment masih bersifat konseptual, namun model yang dihasilkan berpotensi untuk diterapkan pada sistem prediksi risiko diabetes berbasis web atau aplikasi kesehatan sebagai alat bantu deteksi dini. Dengan adanya implementasi tersebut, hasil penelitian diharapkan dapat membantu proses pengambilan keputusan secara lebih cepat dan efektif dalam bidang kesehatan (Wirth & Hipp, 2000).

3. HASIL DAN PEMBAHASAN

Bagian ini membahas hasil eksperimen dan perbandingan performa model *Random Forest* tanpa seleksi fitur dan dengan seleksi fitur menggunakan metode *Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), dan Particle Swarm Optimization (PSO*). Proses penelitian dilakukan menggunakan dataset *Early Stage Diabetes Risk Prediction* dengan pendekatan *CRISP-DM* mulai dari tahap preprocessing data hingga evaluasi model klasifikasi. Pengujian dilakukan untuk mengetahui pengaruh seleksi fitur terhadap peningkatan performa model *Random Forest* dalam prediksi risiko diabetes tahap awal [8].

3.1. Data Understanding dan Transformasi Data

Tahap awal penelitian dilakukan dengan membaca dan memahami dataset yang digunakan pada proses klasifikasi. Dataset yang digunakan adalah *Early Stage Diabetes Risk*

Prediction Dataset yang berisi beberapa atribut gejala diabetes serta satu atribut target berupa *class* sebagai penentu hasil klasifikasi. Bentuk data sebelum dilakukan preprocessing ditampilkan pada Tabel 2, sedangkan hasil data setelah proses transformasi dapat dilihat pada Tabel 3.

Tabel 2. Dataset Sebelum Transformasi

Age	Gender	Polyura	…	Class
40	Male	No	…	Positive
58	Male	No	…	Positive
…	…	…	…	…
42	Male	No	…	negative

Berdasarkan dataset yang digunakan, terdapat 17 atribut yang terdiri dari atribut numerik dan atribut kategorikal. Atribut *age* merupakan data bertipe numerik, sedangkan atribut lainnya sebagian besar berupa data nominal dengan dua kategori nilai seperti *Yes* dan *No* maupun *Positive* dan *Negative*. Karena algoritma *machine learning* tidak dapat memproses data kategorikal secara langsung, maka diperlukan tahap transformasi data terlebih dahulu.

Tabel 3. Dataset Setelah Transformasi

Age	Gender	Polyura	…	Class
40	1	0	…	1
58	1	0	…	1
…	…	…	…	…
42	1	0	…	0

Proses transformasi dilakukan menggunakan teknik Label *Encoding* untuk mengubah data kategorikal menjadi bentuk numerik. Pada tahap ini, nilai Male dikonversi menjadi 1 dan Female menjadi 0. Selain itu, atribut class yang memiliki nilai Positive diubah menjadi 1, sedangkan Negative diubah menjadi 0. Setelah proses transformasi selesai, seluruh atribut pada dataset telah berada dalam format numerik sehingga data dapat digunakan pada tahap pemodelan menggunakan algoritma *Random Forest* serta metode seleksi fitur *Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), dan Particle Swarm Optimization (PSO*).

3.2. Split Data

Tahap selanjutnya dilakukan proses pemisahan dataset hasil transformasi menjadi data fitur dan data label. Proses ini bertujuan untuk menentukan atribut yang digunakan sebagai variabel input dan atribut yang digunakan sebagai target klasifikasi pada model *machine learning*. Pada penelitian ini, atribut fitur dimulai dari age hingga *obesity*, sedangkan atribut *class* digunakan sebagai label atau target prediksi untuk menentukan apakah data termasuk ke dalam kategori positif diabetes atau negatif diabetes.

Berdasarkan struktur dataset yang digunakan, atribut pada indeks 0 sampai 15 dijadikan sebagai fitur penelitian, sementara atribut pada indeks ke-16 digunakan sebagai label klasifikasi. Seluruh fitur tersebut berisi informasi mengenai usia, jenis kelamin, serta berbagai gejala awal diabetes yang digunakan dalam proses prediksi risiko diabetes tahap awal.

Setelah proses pemisahan fitur dan label selesai dilakukan, dataset kemudian dibagi menjadi data latih dan data uji menggunakan metode *train-test split* dengan rasio 70:30. Data latih digunakan untuk membangun dan melatih model klasifikasi, sedangkan data uji digunakan untuk mengukur performa model terhadap data yang belum pernah dipelajari sebelumnya. Pembagian dataset ini dilakukan agar proses evaluasi model dapat menggambarkan kemampuan klasifikasi secara lebih objektif dan mengurangi kemungkinan terjadinya *overfitting* pada model *Random Forest* yang digunakan dalam penelitian ini.

3.3. Feature Section

Feature selection dilakukan untuk memilih atribut yang paling relevan terhadap proses klasifikasi menggunakan metode *wrapper* dengan algoritma *Random Forest* sebagai *learning algorithm*. Pengujian dilakukan menggunakan *cross validation* untuk mengetahui performa model pada setiap kombinasi fitur yang dihasilkan. Selain pengujian menggunakan seleksi fitur, penelitian ini juga melakukan pengujian *Random Forest* tanpa seleksi fitur sebagai model awal atau *baseline* untuk membandingkan perubahan performa setelah dilakukan optimasi fitur.

Hasil pengujian menunjukkan bahwa *Random Forest* tanpa seleksi fitur menghasilkan *accuracy* sebesar 98.27% dengan menggunakan seluruh atribut pada dataset. Selanjutnya dilakukan proses seleksi fitur menggunakan *metode Sequential Forward Selection (SFS), Sequential Backward Selection (SBS),* dan *Particle Swarm Optimization (PSO).* Penggunaan metode seleksi fitur memberikan pengaruh terhadap peningkatan performa model

klasifikasi karena beberapa atribut yang kurang *relevan* dapat dikurangi sehingga model menjadi lebih optimal.

Gambar 2. Proses SFS

Metode *Sequential Forward Selection* (SFS) melakukan proses pemilihan fitur dengan menambahkan atribut secara bertahap dimulai dari satu fitur yang memiliki kontribusi terbaik terhadap model klasifikasi. Pada proses awal, fitur *Polyuria* dipilih sebagai atribut pertama dan menghasilkan *accuracy* sebesar 82.31%. Setelah itu dilakukan penambahan atribut lain seperti *Gender* dan *Polydipsia* sehingga nilai *accuracy* mengalami peningkatan secara bertahap pada setiap iterasi.

Berdasarkan hasil seleksi fitur yang ditunjukkan pada Gambar 3.1, performa model terus meningkat seiring bertambahnya jumlah fitur yang digunakan. *Accuracy* tertinggi diperoleh ketika model menggunakan 11 fitur dengan nilai rata-rata *cross validation* sebesar 98.65%. Setelah jumlah fitur melebihi 11 atribut, peningkatan tidak mengalami perubahan yang signifikan. Hal tersebut menunjukkan bahwa beberapa fitur tambahan memiliki pengaruh yang lebih kecil terhadap proses klasifikasi diabetes tahap awal.

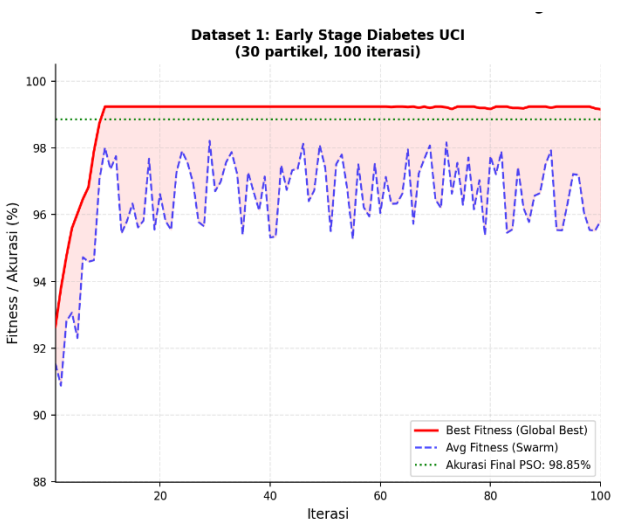
Selain meningkatkan accuracy model, metode SFS juga mampu mengurangi jumlah atribut yang digunakan dibandingkan seluruh fitur pada dataset. Pengurangan jumlah fitur tersebut membuat proses training dan testing menjadi lebih efisien karena kompleksitas model dapat diminimalkan tanpa mengurangi performa klasifikasi secara signifikan.

Gambar 3. Proses SBS

Metode *Sequential Backward Selection* (SBS) melakukan proses seleksi fitur dengan cara mengurangi atribut secara bertahap dari seluruh fitur yang tersedia pada dataset. Proses dimulai menggunakan semua atribut, kemudian fitur yang dianggap kurang relevan dihapus satu per satu sambil menghitung performa model pada setiap iterasi.

Berdasarkan hasil pengujian pada Gambar 3.2, *Random Forest* dengan seluruh fitur menghasilkan accuracy awal sebesar 98.27%. Setelah dilakukan pengurangan beberapa atribut, performa model mengalami peningkatan dibandingkan penggunaan seluruh fitur. Accuracy tertinggi diperoleh ketika model menggunakan 11 fitur dengan rata-rata *cross validation* mencapai 99.23%.

Hasil tersebut menunjukkan bahwa tidak semua atribut pada dataset memiliki hubungan yang signifikan terhadap proses klasifikasi. Penghapusan fitur yang kurang relevan mampu meningkatkan performa model sekaligus mengurangi kompleksitas komputasi. Dibandingkan metode SFS, metode SBS menghasilkan *accuracy* yang lebih stabil dan lebih tinggi sehingga metode ini dinilai lebih optimal pada penelitian prediksi risiko diabetes tahap awal menggunakan algoritma *Random Forest.*



Gambar 4. Proses PSO

Metode *Particle Swarm Optimization* (PSO) digunakan untuk melakukan optimasi seleksi fitur dengan tujuan memperoleh kombinasi atribut terbaik yang mampu menghasilkan performa klasifikasi paling optimal. PSO bekerja menggunakan mekanisme pencarian berbasis populasi melalui sejumlah partikel yang bergerak untuk mencari solusi terbaik berdasarkan nilai fitness dari *accuracy* model *Random Forest.*

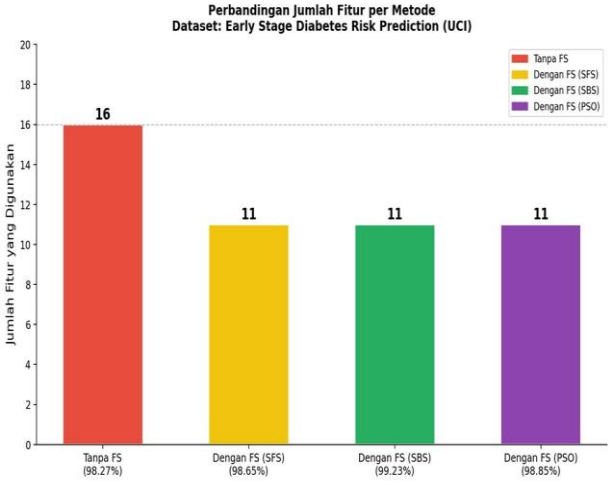
Berdasarkan hasil pengujian yang ditunjukkan pada Gambar 4, proses optimasi dilakukan selama 30 iterasi. Pada iterasi awal diperoleh *accuracy* yang terus meningkat seiring proses pencarian solusi optimal oleh setiap partikel.

Accuracy tertinggi diperoleh pada akhir proses iterasi sebesar 98.85% menggunakan 11 fitur terpilih.

Hasil optimasi menunjukkan bahwa metode PSO mampu meningkatkan performa model sekaligus mengurangi jumlah fitur yang digunakan dibandingkan penggunaan seluruh atribut dataset. Selain itu, perubahan *accuracy* pada setiap iterasi memperlihatkan bahwa proses optimasi berjalan cukup stabil hingga mencapai solusi terbaik. Dengan demikian, metode PSO dapat digunakan sebagai pendekatan optimasi seleksi fitur yang efektif untuk meningkatkan performa *Random Forest* pada prediksi risiko diabetes tahap awal.

3.4. Perbandingan Seleksi Fitur dan Akurasi Model

Tahap ini membahas hasil perbandingan performa model *Random Forest* tanpa seleksi fitur dan dengan seleksi fitur menggunakan metode *Sequential Forward Selection (SFS), Sequential Backward Selection (SBS),* dan *Particle Swarm Optimization (PSO).* Perbandingan dilakukan untuk mengetahui pengaruh seleksi fitur terhadap peningkatan *accuracy* model pada prediksi risiko diabetes tahap awal. Hasil perbandingan *accuracy* model ditunjukkan pada Gambar 5.



Gambar 5. Hasil Perbandingan

Berdasarkan hasil pengujian, Random Forest tanpa seleksi fitur menghasilkan *accuracy* sebesar 98.27% dengan menggunakan seluruh atribut pada dataset sebanyak 16 fitur. Setelah dilakukan proses seleksi fitur menggunakan metode SFS, *accuracy* model meningkat menjadi 98.65% dengan jumlah fitur yang digunakan sebanyak 11 atribut. Hasil tersebut menunjukkan bahwa proses seleksi fitur mampu meningkatkan performa model sekaligus mengurangi jumlah atribut yang digunakan.

Berbeda dengan metode SFS, penggunaan metode *Sequential Backward Selection* (SBS) menghasilkan *accuracy* tertinggi dibandingkan seluruh metode yang diujikan. Metode SBS memperoleh *accuracy* sebesar 99.23% dengan jumlah fitur sebanyak 11 atribut. Hasil tersebut menunjukkan bahwa proses pengurangan fitur secara bertahap mampu menghilangkan atribut yang kurang relevan sehingga model dapat bekerja lebih optimal.

Sementara itu, *Particle Swarm Optimization* (PSO) menghasilkan *accuracy*

sebesar 98.85% dengan jumlah fitur sebanyak 11 atribut. Ketiga metode seleksi fitur berhasil mengurangi jumlah atribut dari 16 menjadi 11 fitur sekaligus meningkatkan atau mempertahankan performa model *Random Forest.*

Secara keseluruhan, hasil perbandingan menunjukkan bahwa metode seleksi fitur memberikan pengaruh terhadap performa model klasifikasi. Metode SBS menghasilkan *accuracy* tertinggi sebesar 99.23%, diikuti PSO sebesar 98.85% dan SFS sebesar 98.65%. Ketiga metode seleksi fitur berhasil mengurangi jumlah atribut dari 16 menjadi 11 fitur, menunjukkan bahwa seleksi fitur efektif dalam meningkatkan efisiensi model tanpa menurunkan performa klasifikasi.

4. PENUTUP

4.1. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, penerapan seleksi fitur memberikan pengaruh terhadap performa algoritma *Random Forest* dalam prediksi risiko diabetes tahap awal. *Random Forest* tanpa seleksi fitur menghasilkan *accuracy* sebesar 98.27% dengan menggunakan seluruh atribut pada dataset sebanyak 16 fitur. Metode *Sequential Forward Selection* (SFS) menghasilkan *accuracy* sebesar 98.65% dengan 11 fitur terpilih, *sedangkan Sequential Backward Selection* (SBS) dan *Particle Swarm Optimization* (SBS) memperoleh accuracy tertinggi sebesar 99.23% dengan 11 fitur, sedangkan PSO menghasilkan *accuracy* sebesar 98.85% dengan 11 fitur. Metode SBS dinilai paling optimal karena menghasilkan accuracy tertinggi dengan jumlah fitur yang efisien. Hasil penelitian menunjukkan bahwa seleksi fitur mampu membantu model dalam memilih atribut yang paling relevan sehingga performa klasifikasi menjadi lebih optimal.

4.2. Saran

Penelitian selanjutnya diharapkan dapat melakukan perbandingan metode wrapper dengan metode seleksi fitur lainnya seperti filter *method* maupun *embedded method* untuk mengetahui performa terbaik pada proses klasifikasi diabetes. Selain itu, penelitian berikutnya juga dapat menggunakan lebih dari satu dataset agar hasil pengujian model menjadi lebih bervariasi dan memiliki tingkat generalisasi yang lebih baik.

Pengembangan penelitian juga dapat dilakukan dengan menambahkan algoritma klasifikasi lain seperti *Support Vector Machine (SVM), K-Nearest Neighbor (KNN*), maupun *XGBoost* sebagai pembanding performa *Random Forest*. Selain itu, model hasil penelitian ini berpotensi untuk dikembangkan menjadi sistem prediksi berbasis web atau aplikasi kesehatan sehingga dapat membantu proses deteksi dini risiko diabetes secara lebih praktis dan efisien.

5. DAFTAR PUSTAKA

[1] Mustofa, M. A. M., Sucipto, dan Nugroho, A. 2025. “Penerapan Random Forest untuk Deteksi Dini Penyakit Parkinson’s dengan Data Frekuensi Suara.” *Jurnal Qua Teknika* 15(2): 25–37.

[2] Sucipto, Prasetya, D. D., dan Widiyaningtyas, T. 2024. “Educational Data Mining: Multiple Choice Question Classification in Vocational School.” *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer* 23(2): 379–388

[3] Triwardani, A. A., Daniati, E., dan Wardani, A. S. 2025. “Studi Komparatif Algoritma Machine Learning dalam Klasifikasi dan Prediksi Kanker Paru-Paru.” *JMSI* 6(2): 214–222.

[4] Guyon, I., dan A. Elisseeff. 2003. “An Introduction to Variable and Feature Selection.” *Journal of Machine Learning Research* 3: 1157–1182.

[5] Han, J., M. Kamber, dan J. Pei. 2012. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham: Morgan Kaufmann.

[6] Islam, M. M. F., R. Ferdousi, S. Rahman, dan H. Y. Bushra. 2020. “Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques.” Dalam *Computer Vision and Machine Intelligence in Medical Image Analysis*, 113–125. Singapore: Springer

[7] Kennedy, J., dan R. Eberhart. 1995. “Particle Swarm Optimization.” Dalam *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1942–1948. IEEE.

[8] Nugroho, A., A. Z. Fanani, dan G. F. Shidik. 2021. “Wrapper Feature Selection for

Random Forest Classifier Using Backward Elimination and Forward Selection for Diabetes Prediction." Dalam *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 178–181. IEEE.

[9] Sokolova, M., dan G. Lapalme. 2009. "A Systematic Analysis of Performance Measures for Classification Tasks." *Information Processing and Management* 45(4): 427–437.

[10] Breiman, L. 2001. "Random Forests." *Machine Learning* 45(1): 5–32.