



Prediksi Risiko Penyakit Parkinson Menggunakan Seleksi Fitur Algoritma Genetika dan SMOTE-XGBoost

Verina Tita Nabila^{1*}, Rewan Jayadi²

¹Fakultas Teknologi Informasi, Teknik Informatika, Universitas Catur Insan Cendekia, Cirebon, Indonesia

²Fakultas Sains dan Teknologi, Departemen Matematika, Universitas Airlangga, Surabaya, Indonesia

Email: ^{1*}verina.titanabila@cic.ac.id, ²jayadirewan6@gmail.com

(* : coressponding author)

Abstrak—Penyakit Parkinson merupakan gangguan neurodegeneratif progresif yang memengaruhi sistem saraf pusat dan ditandai oleh penurunan kontrol gerakan serta perubahan kualitas suara akibat gangguan fungsi motorik. Diagnosis penyakit Parkinson hingga saat ini bergantung pada keahlian dan pengalaman klinisi. Persebaran klinisi yang tidak merata di berbagai wilayah menjadi kendala utama dalam memberikan diagnosis dan penanganan yang tepat. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model prediksi risiko penyakit Parkinson berbasis *machine learning* dengan memanfaatkan seleksi fitur menggunakan algoritma genetika, penanganan ketidakseimbangan data melalui pendekatan SMOTE, dan prediksi yang dilakukan menggunakan metode XGBoost. Hasilnya menunjukkan bahwa metode tersebut menghasilkan performa yang sangat baik dengan nilai akurasi sebesar 95%, sensitivitas sebesar 93%, spesifisitas sebesar 100%, presisi sebesar 100%, *F1-Score* sebesar 97%, dan nilai AUC sebesar 97%. Beberapa fitur yang terpilih, seperti frekuensi dasar suara, kestabilan pitch, ukuran variasi amplitudo antar siklus suara pada rata-rata 5 periode, rasio harmonik terhadap *noise*, dan ukuran penyebaran spektral 2.

Kata Kunci: Parkinson; Algoritma Genetika; SMOTE; XGBoost

Abstract—*Parkinson's disease is a progressive neurodegenerative disorder that affects the central nervous system and is characterized by reduced motor control and changes in voice quality due to impaired motor function. To date, the diagnosis of Parkinson's disease largely depends on the expertise and clinical experience of specialists. The uneven distribution of clinicians across regions remains a major challenge in providing accurate diagnosis and appropriate treatment. Therefore, this study aims to develop a machine learning-based model for predicting the risk of Parkinson's disease by incorporating feature selection using a genetic algorithm, handling data imbalance through the SMOTE approach, and performing prediction using the XGBoost method. The results indicate that the proposed method achieves excellent performance, with an accuracy of 95%, sensitivity of 93%, specificity of 100%, precision of 100%, an F1-score of 97%, and an AUC value of 97%. Several selected features include fundamental voice frequency, pitch stability, amplitude perturbation across voice cycles averaged over five periods, harmonic-to-noise ratio, and spectral spread measure 2.*

Keywords: Parkinson; Genetic Algorithm; SMOTE; XGBoost

1. PENDAHULUAN

Penyakit Parkinson merupakan salah satu gangguan neurodegeneratif progresif yang menyerang sistem saraf pusat yang berpengaruh terhadap kemampuan motorik dan non motorik pada individu (Bloem *et al.*, 2021). Gejala utama penyakit Parkinson berupa gangguan motorik yang meliputi tremor, kekakuan otot, bradikinesia atau gerakan melambat, dan gangguan postur. Selain itu, penderita Parkinson juga mengalami gejala non-motorik seperti gangguan tidur, depresi, gangguan kognitif, dan perubahan pada suara. Gejala tersebut seringkali muncul lebih awal tetapi tidak disadari oleh pasien maupun tenaga medis karena sifatnya yang tidak spesifik (Weintraub *et al.*, 2022). Hal ini menyebabkan keterlambatan pada diagnosis dan penanganan penyakit tersebut. Salah satu gejala awal yang paling sering dijumpai pada pasien Parkinson adalah gangguan suara atau disfoni. Sebanyak 90% pasien Parkinson mengalami perubahan kualitas suara bahkan sebelum munculnya gejala motorik yang khas (Oguz & Genc, 2023).

Diagnosis penyakit Parkinson hingga saat ini masih bersifat klinis dan sangat bergantung pada pengamatan gejala oleh dokter spesialis saraf (Solana-Lavalle & Rosas-Romero, 2021). Oleh karena itu, keakuratan diagnosis sangat dipengaruhi oleh keahlian dan pengalaman klinisi. Sementara itu, prevalensi penyakit Parkinson di Indonesia terus mengalami peningkatan seiring dengan bertambahnya jumlah lansia. Menurut Perhimpunan Dokter Spesialis Neurologi Indonesia



(2024) mencatat bahwa jumlah penderita Parkinson di Indonesia diperkirakan mencapai lebih dari 100.000 orang. Namun, keterbatasan jumlah dokter spesialis saraf yang hanya sekitar 1.400 orang di seluruh Indonesia dengan persebaran yang tidak merata menjadi kendala utama dalam memberikan diagnosis dan penanganan yang tepat (Dewi *et al.*, 2023). Kondisi tersebut menyebabkan banyak penderita Parkinson terutama yang berada di wilayah terpencil dan pedesaan tidak mendapatkan diagnosis secara cepat dan tepat. Padahal, deteksi dini sangat penting untuk memperlambat perkembangan penyakit dan meningkatkan kualitas hidup pasien. Oleh karena itu, diperlukan sistem prediksi risiko penyakit Parkinson berbasis teknologi yang dapat menjangkau populasi lebih luas secara efisien.

Dalam beberapa tahun terakhir, pendekatan berbasis kecerdasan buatan, khususnya *machine learning* telah menunjukkan hasil yang signifikan dalam sistem prediksi berbasis data. Algoritma *machine learning* dapat dilatih menggunakan data rekaman suara yang telah diekstraksi menjadi fitur numerik untuk membedakan antara individu sehat dan pasien Parkinson dengan akurasi yang tinggi (Lamba *et al.*, 2022). Salah satu algoritma yang sering digunakan dalam prediksi risiko penyakit adalah *Extreme Gradient Boosting* (XGBoost). XGBoost dapat diimplementasikan pada dataset yang tidak terlalu besar dan mampu menghasilkan performa yang baik (Islam *et al.*, 2022). Namun, kinerja XGBoost sangat dipengaruhi oleh kualitas fitur yang digunakan. Oleh karena itu, diperlukan proses seleksi fitur untuk memilih subset fitur yang paling relevan dan informatif untuk meningkatkan akurasi prediksi dan mengurangi kompleksitas komputasi.

Salah satu metode seleksi fitur yang efektif untuk digunakan adalah Algoritma Genetika (AG). AG merupakan algoritma optimasi berbasis populasi yang meniru proses evolusi biologis untuk menemukan solusi optimal dalam ruang pencarian yang kompleks (Alhijawi & Awajan, 2024). Dalam konteks seleksi fitur, AG dapat mengeksplorasi berbagai kombinasi fitur untuk menemukan konfigurasi terbaik yang menghasilkan performa prediksi tertinggi.

Selain pemilihan fitur, kinerja XGBoost juga sangat dipengaruhi oleh karakteristik data yang digunakan. Salah satu permasalahan yang umum terjadi pada persoalan klasifikasi adalah distribusi kelas yang tidak seimbang. Kondisi tersebut terjadi ketika jumlah sampel pada kelas tertentu lebih sedikit dibandingkan kelas lainnya sehingga berpotensi menimbulkan bias prediksi. Oleh karena itu, pada penelitian ini, juga dilakukan penyeimbangan distribusi kelas menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE) guna meningkatkan proporsi data pada kelas minoritas sehingga model mampu menghasilkan kinerja yang lebih optimal (Sulistiyono *et al.*, 2021).

Beberapa penelitian terdahulu yang meneliti mengenai prediksi risiko penyakit Parkinson menggunakan *machine learning* telah banyak dilakukan, seperti Kurnia *et al.* (2023) memprediksi risiko penyakit Parkinson dengan pendekatan *hybrid* antara algoritma *Particle Swarm Optimization* (PSO) untuk seleksi fitur, SMOTE untuk menyeimbangkan kelas pada data, dan XGBoost untuk prediksi. Hasilnya menunjukkan bahwa metode tersebut mencapai nilai AUC sebesar 94,83%. Sementara itu, Ali *et al.* (2023) juga memprediksi risiko penyakit Parkinson menggunakan *hybrid* AG dan XGBoost. Penelitian tersebut menunjukkan bahwa *hybrid* AG dan XGBoost memperoleh presisi sebesar 94%.

Berdasarkan uraian diatas, penelitian ini bertujuan untuk memprediksi risiko penyakit Parkinson menggunakan algoritma klasifikasi XGBoost. Namun, sebelum proses tersebut, dilakukan seleksi fitur menggunakan AG dan penyeimbangan kelas pada data menggunakan SMOTE. Selain itu, data yang digunakan pada penelitian ini berupa data karakteristik suara.

2. METODE

2.1 Data dan Sumber Data

Data yang digunakan pada penelitian ini diperoleh dari *website UCI Machine Learning Repository*. Data tersebut memuat informasi mengenai karakteristik vokal dari 147 individu yang menderita Parkinson dan 48 individu yang tidak menderita Parkinson. Data tersebut memuat 23 variabel dengan 1 variabel sebagai variabel target dan 22 variabel sebagai fitur. Deskripsi variabel tersebut disajikan pada Tabel 1.

Tabel 1. Deskripsi Variabel yang Digunakan pada Penelitian

No.	Fitur	Definisi
1	D2	Dimensi fraktal sinyal – mengukur kompleksitas dinamis
2	DFA	<i>Detrended Fluctuation Analysis</i> – deteksi korelasi jangka panjang
3	HNR	<i>Harmonic-to-Noise Ratio</i> – rasio harmonik terhadap noise
4	Jitter: DDP	<i>Difference of Differences of Periods</i> - variasi frekuensi antar siklus getaran suara secara rata-rata
5	MDVP: Fo (Hz)	Frekuensi dasar suara
6	MDVP: Flo (Hz)	Frekuensi suara terendah
7	MDVP: Fhi (Hz)	Frekuensi suara tertinggi
8	MDVP: Jitter (%)	Tingkat variabilitas frekuensi (<i>pitch</i>) dalam persen
9	MDVP: Jitter (Abs)	Jitter absolut – variasi <i>pitch</i> dalam angka mutlak
10	MDVP: APQ	<i>Average Perturbation Quotient</i> - parameter akustik yang mengukur variasi amplitudo antar siklus getaran suara.
11	MDVP: PPQ	<i>Pitch Period Perturbation Quotient</i> – variasi periodik <i>pitch</i>
12	MDVP: RAP	<i>Relative Average Perturbation</i> – kestabilan <i>pitch</i>
13	MDVP: Shimmer	Variabilitas amplitudo suara
14	MDVP: Shimmer (dB)	<i>Shimmer</i> dalam skala desibel
15	NHR	<i>Noise-to-Harmonic Ratio</i> – rasio noise terhadap suara harmonik
16	PPE	<i>Pitch Period Entropy</i> – ketidakteraturan <i>pitch</i>
17	RPDE	<i>Recurrence Period Density Entropy</i> – kompleksitas sinyal suara
18	Shimmer: APQ3	<i>Amplitude Perturbation Quotient</i> - ukuran variasi amplitudo antar siklus suara (rata-rata 3 periode)
19	Shimmer: APQ5	<i>Amplitude Perturbation Quotient</i> - ukuran variasi amplitudo antar siklus suara (rata-rata 5 periode)
20	Shimmer: DDA	<i>Average Absolute Difference of Differences Between Amplitudes</i> - rata-rata variasi amplitudo antar siklus getaran suara
21	spread1	Ukuran penyebaran spektral 1
22	spread2	Ukuran penyebaran spektral 2
23	Status (Target)	Label 0: Tidak menderita Parkinson Label 1: Menderita Parkinson

2.2 Normalisasi Data

Normalisasi data dilakukan dengan menskalakan nilai atribut agar berada pada rentang tertentu (Cabello-Solorzano *et al.*, 2023). Salah satu teknik yang banyak digunakan adalah *min-max scaling*, yaitu transformasi linear yang mengubah skala data asli sehingga nilai setiap fitur dipetakan ke interval [0,1] (Henderi *et al.*, 2021). Proses tersebut dilakukan menggunakan persamaan berikut:

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

dengan

- x'_i : nilai fitur hasil transformasi ke- i ;
- x_i : nilai fitur ke- i ;
- $x_{\min}$: nilai minimum dari fitur;
- $x_{\max}$: nilai maksimum dari fitur.

2.3 Pembagian Data

Pembagian data dilakukan untuk membagi data menjadi dua sub data, yaitu data *training* dan data *testing*. Pembagian data dilakukan berdasarkan rasio pembagian data yang telah ditetapkan. Pada penelitian ini, terdapat tiga skenario pembagian data seperti yang disajikan pada Tabel 2.

Tabel 2. Banyaknya Data pada Skenario Pembagian Data

No.	Rasio pembagian data		Banyaknya data	
	Persentase data <i>training</i>	Persentase data <i>testing</i>	Banyaknya data <i>training</i>	Banyaknya data <i>testing</i>
1	90%	10%	175	20
2	80%	20%	156	39
3	70%	30%	136	59

Berdasarkan Tabel 2, terdapat tiga skenario pembagian data, yaitu pembagian data dengan 90% dari keseluruhan data sebagai data *training* dan 10% dari keseluruhan data sebagai data *testing* atau 90:10. Total data yang digunakan pada penelitian ini sebanyak 195 data sehingga dengan rasio pembagian data tersebut, 175 data digunakan sebagai data *training* dan 20 data digunakan sebagai data *testing*. Selain itu, pembagian data juga dilakukan dengan 80% dari keseluruhan data sebagai data *training* dan 20% dari keseluruhan data sebagai data *testing* atau 80:20 sehingga terdapat 156 data digunakan sebagai data *training* dan 39 data digunakan sebagai data *testing*. Skenario pembagian data berikutnya dilakukan dengan membagi data menjadi 70% dari keseluruhan data sebagai data *training* dan 30% dari keseluruhan data sebagai data *testing* atau 70:30 sehingga terdapat 136 data digunakan sebagai data *training* dan 59 data digunakan sebagai data *testing*.

2.4 Seleksi Fitur Menggunakan Algoritma Genetika

Algoritma Genetika (AG) merupakan metode optimasi berbasis populasi yang terinspirasi dari mekanisme evolusi biologis. Pendekatan ini bekerja dengan merepresentasikan solusi sebagai individu dalam suatu populasi kemudian melakukan proses seleksi dan reproduksi untuk menghasilkan solusi yang semakin optimal dari generasi ke generasi. Setiap individu biasanya direpresentasikan dalam bentuk kromosom yang berisi sekumpulan gen, dengan setiap gen merepresentasikan parameter atau keputusan tertentu dalam permasalahan yang dihadapi (Alhijawi & Awajan, 2024).

Proses pencarian solusi pada algoritma genetika dimulai dengan pembentukan populasi awal secara acak. Selanjutnya, kualitas setiap individu dievaluasi menggunakan fungsi *fitness* yang mencerminkan seberapa baik solusi tersebut dalam memenuhi tujuan optimasi. Individu dengan nilai *fitness* yang lebih tinggi memiliki peluang lebih besar untuk dipilih dalam proses seleksi. Tahap berikutnya melibatkan operator genetika, yaitu *crossover* dan *mutation*. Operator *crossover* bertujuan untuk menggabungkan informasi dari dua individu induk sehingga dapat dihasilkan keturunan baru sedangkan *mutation* digunakan untuk mempertahankan variasi individu di dalam populasi sehingga membantu algoritma menghindari terjebak terlalu cepat pada solusi lokal yang belum tentu optimal (Lamba *et al.*, 2022).

Pada penelitian ini, proses seleksi fitur dilakukan menggunakan algoritma genetika. Beberapa parameter yang digunakan pada algoritma genetika, antara lain ukuran populasi (*population size*) sebesar 30, jumlah *generations* sebanyak 20, laju *crossover* sebesar 0,8, dan laju *mutation* sebesar 0,1. Tujuan dari seleksi fitur tersebut, yaitu untuk memilih 5 fitur terbaik dari total 22 fitur yang tersedia dengan mempertimbangkan kontribusi masing-masing fitur terhadap peningkatan kinerja klasifikasi.

2.5 Penyeimbangan Kelas Menggunakan SMOTE

Pada penelitian ini, metode *Synthetic Minority Over-sampling Technique* (SMOTE) digunakan untuk menangani ketidakseimbangan kelas pada data *training*, khususnya dalam meningkatkan jumlah sampel pada kelas minoritas. Teknik ini bekerja dengan cara mensintesis data baru berdasarkan interpolasi antara sampel minoritas yang ada sehingga distribusi kelas menjadi lebih seimbang dan performa model klasifikasi dapat ditingkatkan. SMOTE dapat dilakukan menggunakan persamaan berikut (Khdair & Dasari, 2021):



$$x_{syn} = x_i + |x_{nn} - x_i|\delta \quad (2)$$

dengan

x_{syn} : data hasil replikasi

x_i : data ke- i dari kelas minoritas

x_{nn} : data dari kelas minoritas yang memiliki jarak terdekat dengan x_i

δ : parameter acak.

2.6 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) merupakan salah satu algoritma *ensemble learning* berbasis teknik *gradient boosting* yang menggabungkan beberapa model pohon keputusan secara bertahap untuk meningkatkan kemampuan prediksi (Dachi & Sitompul, 2023). Pada pendekatan ini, model dibangun secara sekuensial dengan setiap pohon baru bertugas memperbaiki kesalahan prediksi dari model sebelumnya melalui proses optimasi fungsi objektif. Selain itu, XGBoost menerapkan strategi regularisasi untuk mengendalikan kompleksitas model sehingga mampu mengurangi risiko *overfitting* dan meningkatkan kemampuan generalisasi. Keunggulan utama XGBoost terletak pada fleksibilitasnya dalam menangani data numerik maupun kategorik yang telah ditransformasikan dan kemampuannya menghasilkan performa yang stabil pada berbagai permasalahan klasifikasi serta regresi (Islam *et al.*, 2022).

2.7 Evaluasi Model

Evaluasi model dilakukan menggunakan *confusion matrix* yang menyajikan ringkasan prediksi model terhadap kondisi aktual dalam bentuk tabel kontingensi dua dimensi. Dari komponen *confusion matrix*, evaluasi model dapat dilakukan berdasarkan akurasi, sensitivitas, spesifisitas, presisi, *F1-Score*, dan nilai *Area Under the Receiver Operating Characteristic* (ROC) Curve (AUC) (Nahm, 2022).

3. ANALISA DAN PEMBAHASAN

3.1 Analisis Hasil Seleksi Fitur Menggunakan Algoritma Genetika

Pada penelitian ini, representasi solusi dalam algoritma genetika menggunakan kromosom berupa string bit biner sepanjang 22 gen, dengan angka 1 menunjukkan bahwa suatu fitur dipilih sedangkan angka 0 menunjukkan bahwa fitur tersebut tidak dipilih. Selain itu, pada proses seleksi fitur menggunakan algoritma genetika, teknik *crossover* yang digunakan adalah *one point crossover* sedangkan teknik mutasi yang digunakan adalah *resiprocal mutation exchange*.

Setelah melalui serangkaian proses seleksi fitur menggunakan algoritma genetika yang mencakup proses seleksi, *crossover* dan mutasi, setiap individu dievaluasi berdasarkan nilai akurasi klasifikasi menggunakan algoritma XGBoost. Individu dengan nilai akurasi tertinggi merupakan individu yang terpilih sebagai solusi optimal dan fitur-fitur pada individu tersebut digunakan untuk mengekstraksi subset data. Subset fitur tersebut digunakan dalam proses klasifikasi utama yang menggunakan metode XGBoost.

3.2 Analisis Hasil Penyeimbangan Kelas Menggunakan SMOTE

Jumlah data hasil sintesis SMOTE untuk masing-masing skenario pembagian data disajikan pada Tabel 3.

Tabel 3. Banyaknya Data Hasil Sintesis SMOTE

Skenario pembagian data	Banyaknya data sebelum SMOTE		Banyaknya data setelah SMOTE	
90:10	Label 0	= 43 data	Label 0	= 132 data
	Label 1	= 132 data	Label 1	= 132 data
	Total data	= 175 data	Total data	= 264 data
	Label 0	= 38 data	Label 0	= 118 data
80:20	Label 1	= 118 data	Label 1	= 118 data
	Total data	= 156 data	Total data	= 236 data

Skenario pembagian data	Banyaknya data sebelum SMOTE	Banyaknya data setelah SMOTE
70:30	Label 0 = 33 data	Label 0 = 103 data
	Label 1 = 103 data	Label 1 = 103 data
	Total data = 136 data	Total data = 206 data

Berdasarkan Tabel 3, pada skenario pembagian data 90% data *training* dan 10% data *testing*, sebelum dilakukan proses penyeimbangan kelas menggunakan SMOTE banyaknya data pada kelas 0 sebanyak 43 data dan banyaknya data pada kelas 1 sebanyak 132 data. Hal ini menunjukkan bahwa terdapat ketidakseimbangan kelas dengan banyaknya data pada kelas 0 yang lebih sedikit dibandingkan dengan data pada kelas 1. Setelah dilakukan penyeimbangan kelas menggunakan SMOTE, banyaknya data pada kelas 0 menjadi sebanyak 132 data dan banyaknya data pada kelas 1 juga sebanyak 132 data sehingga banyaknya data pada kedua kelas label tersebut sudah seimbang. Proses yang sama juga dilakukan untuk skenario pembagian data yang lainnya.

3.3 Klasifikasi Penyakit Parkinson Menggunakan *Extreme Gradient Boosting*

Dalam penelitian ini, prediksi risiko penyakit Parkinson menggunakan algoritma XGBoost dilakukan pada setiap skenario pembagian data sehingga nantinya terdapat beberapa hasil prediksi yang akan dibandingkan untuk menentukan skenario pembagian data yang terbaik. Selain itu, beberapa *hyperparameter* yang digunakan pada XGBoost antara lain *number of estimators* sebesar 100, *maximum depth* sebesar 6, dan *learning rate* sebesar 0,3.

3.4 Evaluasi Model

Proses evaluasi dilakukan terhadap setiap skenario pembagian data *training* dan data *testing*, seperti 90:10, 80:20, dan 70:30, untuk memastikan konsistensi performa model terhadap variasi jumlah data *training*. Pada penelitian ini, proses *testing* pada setiap skenario dilakukan sebanyak 10 kali percobaan. Pada setiap percobaan, model dievaluasi menggunakan enam metrik utama, yaitu akurasi, sensitivitas, spesifisitas, presisi, *F1-score*, dan nilai AUC. Hasil evaluasi metode *hybrid* algoritma genetika dan SMOTE-XGBoost pada skenario pembagian data 90% data *training* dan 10% data *testing* disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi Metode pada Skenario Pembagian Data 90:10

Percobaan ke-	Akurasi (%)	Sensitivitas (%)	Spesifisitas (%)	Presisi (%)	<i>F1-Score</i> (%)	Nilai AUC (%)
1	75	67	100	100	80	83
2	85	80	100	100	89	90
3	90	87	100	100	93	93
4	95	93	100	100	97	97
5	85	80	100	100	89	90
6	80	80	80	92	86	80
7	85	80	100	100	89	90
8	80	73	100	100	85	87
9	90	87	100	100	93	93
10	85	80	100	100	89	90

Berdasarkan Tabel 4, percobaan keempat menghasilkan performa terbaik, dengan nilai akurasi mencapai 95%, sensitivitas sebesar 93%, spesifisitas sebesar 100%, presisi sebesar 100%, *F1-score* sebesar 97%, dan nilai AUC sebesar 97%. Oleh karena itu, hasil dari percobaan keempat dipilih sebagai representasi terbaik dari prediksi risiko penyakit Parkinson berbasis algoritma genetika dan SMOTE-XGBoost pada skenario pembagian data 90% data *training* dan 10% data *testing*. Beberapa fitur yang terpilih pada percobaan keempat adalah MDVP: Fo (Hz), MDVP: RAP, Shimmer: APQ5, HNR, dan spread2. Sementara itu, hasil evaluasi metode *hybrid* algoritma genetika dan SMOTE-XGBoost pada skenario pembagian data 80% data *training* dan 20% data *testing* disajikan pada Tabel 5.

Tabel 5. Hasil Evaluasi Metode pada Skenario Pembagian Data 80:20

Percobaan ke-	Akurasi (%)	Sensitivitas (%)	Spesifisitas (%)	Presisi (%)	F1-Score (%)	Nilai AUC (%)
1	84,62	82,76	90	96	89	86
2	94,87	96,55	90	96,55	96,55	93,28
3	92,31	93,10	90	96,43	94,74	91,55
4	82,05	79,31	90	95,83	86,79	84,66
5	76,92	82,76	60	85,71	84,21	71,38
6	84,62	86,21	80	92,59	89,29	83,10
7	87,18	86,21	90	96,15	90,91	88,10
8	87,18	89,66	80	92,86	91,23	84,83
9	92,31	93,10	90	96,43	94,74	91,55
10	94,87	96,55	90	96,55	96,55	93,28

Berdasarkan Tabel 5, percobaan kedua menghasilkan performa terbaik dengan nilai akurasi mencapai 94,87%, sensitivitas sebesar 96,55%, spesifisitas sebesar 90%, presisi sebesar 96,55%, *F1-score* sebesar 96,55%, dan nilai AUC sebesar 93,28%. Fitur yang terpilih pada percobaan kedua adalah Shimmer: APQ3, MDVP: APQ, HNR, spread1, dan D2. Dengan demikian, hasil dari percobaan kedua dipilih sebagai representasi terbaik dari prediksi risiko penyakit Parkinson berbasis algoritma genetika dan SMOTE-XGBoost pada skenario pembagian data 80% data *training* dan 20% data *testing*. Sementara itu, hasil evaluasi metode *hybrid* algoritma genetika dan SMOTE-XGBoost pada skenario pembagian data 70% data *training* dan 30% data *testing* disajikan pada Tabel 6.

Tabel 6. Hasil Evaluasi Metode pada Skenario Pembagian Data 70:30

Percobaan ke-	Akurasi (%)	Sensitivitas (%)	Spesifisitas (%)	Presisi (%)	F1-Score (%)	Nilai AUC (%)
1	76,27	77,27	73,33	89,47	82,93	75,30
2	93,22	90,91	100	100	95,24	95,45
3	88,14	88,64	86,67	95,12	91,76	87,65
4	88,14	86,36	93,33	97,44	91,57	89,85
5	89,83	86,36	100	100	92,68	93,18
6	84,75	81,82	93,33	97,30	88,89	87,58
7	88,14	84,09	100	100	91,36	92,05
8	93,22	93,18	93,33	97,62	95,35	93,26
9	79,66	79,55	80	92,11	85,37	79,77
10	83,05	79,55	93,33	97,22	87,50	86,44

Berdasarkan Tabel 6, percobaan kedua menghasilkan performa terbaik dengan nilai akurasi mencapai 93,22 %, sensitivitas sebesar 90,91 %, spesifisitas sebesar 100 %, presisi sebesar 100 %, *F1-score* sebesar 95,24 %, dan nilai AUC sebesar 95,45 %. Beberapa fitur yang terpilih pada percobaan kedua antara lain MDVP: Fhi (Hz), MDVP: Shimmer, MDVP: APQ, RPDE, dan spread1. Dengan demikian, hasil dari percobaan kedua dipilih sebagai representasi terbaik dari prediksi risiko penyakit Parkinson berbasis *hybrid* algoritma genetika dan SMOTE-XGBoost pada skenario pembagian data 70% data *training* dan 30% data *testing*. Selanjutnya, hasil terbaik dari setiap skenario pembagian data dibandingkan untuk menentukan skenario pembagian data yang paling efektif. Perbandingan hasil metrik evaluasi metode *hybrid* algoritma genetika dan SMOTE-XGBoost pada masing-masing skenario pembagian data disajikan pada Tabel 7.

Tabel 7. Perbandingan Hasil Evaluasi Metode pada Setiap Skenario Pembagian Data

Skenario Pembagian Data	Akurasi (%)	Sensitivitas (%)	Spesifisitas (%)	Presisi (%)	F1- Score (%)	Nilai AUC (%)
90:10	95	93	100	100	97	97



80:20	94,87	96,55	90	96,55	96,55	93,28
70:30	93,22	90,91	100	100	95,24	95,45

Berdasarkan Tabel 7, diperoleh informasi bahwa skenario pembagian data dengan 90% data *training* dan 10% data *testing* menghasilkan performa terbaik dibandingkan kedua skenario pembagian data lainnya. Dengan demikian, mekanisme terbaik untuk prediksi risiko penyakit Parkinson berbasis *hybrid* algoritma genetika dan SMOTE-XGBoost adalah menggunakan skenario pembagian data 90 % data *training* dan 10 % data *testing* dengan fitur terpilih yaitu MDVP: Fo (Hz), MDVP: RAP, Shimmer: APQ5, HNR, dan spread2.

4. KESIMPULAN

Prediksi risiko penyakit Parkinson dikembangkan melalui tiga tahapan utama, yaitu seleksi fitur menggunakan algoritma genetika, penyeimbangan kelas menggunakan metode SMOTE, dan prediksi menggunakan metode XGBoost. Hasil evaluasi metode menunjukkan bahwa pendekatan *hybrid* algoritma genetika dan SMOTE-XGBoost menghasilkan performa yang sangat baik dengan nilai akurasi sebesar 95%, sensitivitas sebesar 93%, spesifisitas sebesar 100%, presisi sebesar 100%, *F1-Score* sebesar 97%, dan nilai AUC sebesar 97%. Rasio pembagian data yang digunakan, yaitu 90% data *training* dan 10% data *testing*. Selain itu, beberapa fitur yang terpilih dari hasil seleksi fitur, antara lain frekuensi dasar suara, kestabilan *pitch*, ukuran variasi amplitudo antar siklus suara pada rata-rata 5 periode, rasio harmonik terhadap *noise*, dan ukuran penyebaran spektral 2.

REFERENCES

- Alhijawi, B., & Awajan, A. (2024). Genetic algorithms: theory, genetic operators, solutions, and applications. *Evolutionary Intelligence*, 17, 1245–1256. <https://doi.org/https://doi.org/10.1007/s12065-023-00822-6>
- Ali, A. M., Salim, F., & Saeed, F. (2023). Parkinson's Disease Detection Using Filter Feature Selection and a Genetic Algorithm with Ensemble Learning. *Diagnostics*, 13(17). <https://doi.org/https://doi.org/10.3390/diagnostics13172816>
- Bloem, B. R., Okun, M. S., & Klein, C. (2021). Parkinson's disease. *The Lancet*, 397(10291), 2284–2303. [https://doi.org/10.1016/S0140-6736\(21\)00218-X](https://doi.org/10.1016/S0140-6736(21)00218-X)
- Cabello-Solorzano, K., De, I. O., Araujo, Peña, M., Correia, L., & Tallón-Ballesteros, A. J. (2023). The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis. *18th International Conference on Soft Computing Models in Industrial and Environmental Applications*, 344–353. https://doi.org/https://doi.org/10.1007/978-3-031-42536-3_33
- Dachi, J. M. A. S., & Sitompul, P. (2023). Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit. *Jurnal Riset Rumpun Matematika Dan Ilmu Pengetahuan Alam*, 2(2), 87–103. <https://doi.org/10.55606/jurrimipa.v2i2.1470>
- Dewi, R. N. V. R., Oktamianti, P., & Muliawati, D. (2023). Gambaran Kebijakan Pemenuhan Kebutuhan Tenaga Dokter Spesialis Di Indonesia. *Jurnal Cahaya Mandalika*, 3(2), 551–562. <https://doi.org/https://doi.org/10.36312/jcm.v3i2.1661>
- Henderi, Wahyuningsih, T., & Rahwanto, E. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *International Journal of Informatics and Information Systems*, 4(1), 13–20. <https://doi.org/https://doi.org/10.47738/ijiis.v4i1.73>
- Islam, S. M. S., Talukder, A., Awal, M. A., Siddiqui, M. M. U., Ahamad, M. M., Ahammed, B., Rawal, L. B., Alizadehsani, R., Abawajy, J., Laranjo, L., Chow, C. K., & Maddison, R. (2022). Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. *Frontiers in Cardiovascular Medicine*, 9(839379). <https://doi.org/10.3389/fcvm.2022.839379>
- Khdair, H., & Dasari, N. M. (2021). Exploring machine learning techniques for coronary heart disease prediction. *International Journal of Advanced Computer Science and Applications*, 12(5), 28–36. <https://doi.org/10.14569/IJACSA.2021.0120505>
- Kurnia, D., Mazdadi, M. I., Kartini, D., Nugroho, R. A., Abadi, F., Mangkurat, U. L., & Korespondensi, P. (2023). Seleksi Fitur Dengan Particle Swarm Optimization Pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, 10(5), 1083–1094. <https://doi.org/10.25126/jtiik.2023107252>



JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 3, No. 11 April Tahun 2026
ISSN 3025-0919 (media online)
Hal 3007-3015

- Lamba, R., Gulati, T., Fahad, H., & Anurag, A. (2022). A hybrid system for Parkinson's disease diagnosis using machine learning techniques. *International Journal of Speech Technology*, 25(3), 583–593. <https://doi.org/10.1007/s10772-021-09837-9>
- Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, 75(1), 25–36. <https://doi.org/10.4097/kja.21209>
- Oguz, M. S. A. O., & Genc, G. (2023). Hypokinetic Dysarthria in Parkinson's Disease: A Narrative Review. *Medical Bulletin of Sisli Etfal Hospital*, 57(2). <https://doi.org/10.14744/SEMB.2023.29560>
- Perhimpunan Dokter Spesialis Neurologi Indonesia. (2024). *Panduan Tata Laksana Penyakit Parkinson Indonesia*. UI Publishing.
- Solana-Lavalle, G., & Rosas-Romero, R. (2021). Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, 198(105793). <https://doi.org/https://doi.org/10.1016/j.cmpb.2020.105793>
- Sulistiyono, M., Pristyanto, Y., Adi, S., & Gumelar, G. (2021). Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi. *SISTEMASI: Jurnal Sistem Informasi*, 10(2), 445. <https://doi.org/10.32520/stmsi.v10i2.1303>
- Weintraub, D., Aarsland, D., Chaudhuri, K. R., Dobkin, R. D., Leentjens, A. F., & Rodriguez-Violante, M. (2022). The neuropsychiatry of Parkinson's disease: advances and challenges. *The Lancet Neurology*, 21(1), 89–102. [https://doi.org/10.1016/S1474-4422\(21\)00330-6](https://doi.org/10.1016/S1474-4422(21)00330-6)