e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 8 No. 2, Desember, 2024, **Hal. 575-584**
DOI: 10.29408/edumatic.v8i2.27788

Peningkatan Akurasi Deteksi Dini Penyakit Parkinson melalui Pendekatan Ensemble Learning dan Seleksi Fitur Optimal

Kang Andini Wulandari [1,*], Adhitya Nugraha [1], Ardytha Luthfiarta [1], Laila Rahmatin Nisa [1]

[1] Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Indonesia
[*] Correspondence: kangandiniw@gmail.com

Received: 13 Oktober 2024 | Revised: 22 Oktober 2024 | Accepted: 12 November 2024 | Published: 19 Desember 2024

Abstrak
Deteksi dini penyakit Parkinson (PD) penting untuk meningkatkan kualitas hidup pasien melalui intervensi yang lebih cepat. Penelitian ini bertujuan mengembangkan model prediksi melalui pendekatan ensemble learning dan seleksi fitur optimal. Penelitian eksperimental ini menggunakan tiga algoritma machine learning*: random forest*, *XGBoost*, dan *extra trees*, yang dioptimalkan melalui hyperparameter tuning, teknik seleksi fitur, dan *Kernel Principal Component Analysis* (KPCA) untuk reduksi dimensi. Penelitian ini menggunakan UCI *machine learning* Parkinson Dataset, yang terdiri dari 80 sampel dan 44 fitur akustik yang diekstraksi dari suara pasien saat mengucapkan vokal "/a/" selama lima detik. Hasil temuan kami menunjukkan bahwa *XGBoost* menghasilkan akurasi tertinggi sebesar 88,93% setelah *tuning* dan KPCA, diikuti oleh *extra trees* dengan akurasi 86,15% dan *random forest* sebesar 85,47%. Penerapan KPCA terbukti mampu menurunkan dimensi data tanpa mengorbankan akurasi, sehingga meningkatkan efisiensi pemodelan. Selain itu, penggunaan data suara memiliki potensi besar dalam deteksi dini PD, dan bahwa pemilihan algoritma serta teknik reduksi dimensi yang tepat sangat penting dalam mengoptimalkan model diagnostik berbasis data.

Kata kunci: *ensemble learning; filter-based; hyperparameter tuning*; klasifikasi parkinson

Abstract
Early detection of Parkinson's disease (PD) is essential to enhance patient quality of life through timely intervention. This research aims to develop a predictive model using an ensemble learning approach and optimal feature selection. This experimental study employs three machine learning algorithms: random forest, XGBoost, and extra trees, optimized through hyperparameter tuning, feature selection techniques, and Kernel Principal Component Analysis (KPCA) for dimensionality reduction. The study utilizes the UCI Machine Learning Parkinson Dataset, which consists of 80 samples and 44 acoustic features extracted from patients' voices as they sustain the vowel sound "/a/" for five seconds. Results show that XGBoost achieved the highest accuracy at 88.93% after tuning and KPCA, followed by extra trees with 86.15%, and random forest with 85.47%. The application of KPCA successfully reduced data dimensionality without sacrificing accuracy, thereby improving modeling efficiency. These findings suggest that voice data holds significant potential for early PD detection and that selecting appropriate algorithms and dimensionality reduction techniques is crucial for optimizing data-driven diagnostic models.

Keywords:*, ensemble learning; filter-based; hyperparameter tuning; parkinson classification*

PENDAHULUAN

Deteksi dini adalah proses penting dalam identifikasi penyakit pada tahap awal sebelum gejala yang lebih parah muncul. Dalam kasus penyakit Parkinson (PD), deteksi dini sangat krusial untuk memungkinkan pengobatan yang lebih cepat dan efektif serta memperlambat perkembangan penyakit. Penanganan dini dapat mengurangi keparahan gejala, memperpanjang

kualitas hidup pasien, dan menekan biaya perawatan jangka panjang (Aprilita et al., 2023). PD merupakan gangguan neurodegeneratif progresif (Desiani et al., 2023) yang ditandai oleh kematian sel-sel saraf penghasil dopamin di area substantia nigra otak, yang menyebabkan gangguan kontrol gerakan tubuh (Alalayah et al., 2023; Pramanik et al., 2023). Gejala motorik utama PD meliputi tremor, kekakuan otot, dan gangguan keseimbangan (Fahira et al., 2023), sementara gejala non-motorik seperti depresi dan gangguan tidur juga umum terjadi (Iyer et al., 2023; Malekroodi et al., 2024). Mengingat meningkatnya prevalensi global PD (Farida et al., 2023), deteksi dini sangat penting untuk memperlambat perkembangan penyakit dan meningkatkan kualitas hidup pasien.

Deteksi dini PD menghadapi tantangan sulitnya mengenali gejala awal, baik motorik maupun non-motorik, yang sering tersembunyi atau disalahartikan (Nainggolan et al., 2023). Pendekatan berbasis data mining (Ananda et al., 2024) menawarkan solusi efektif dengan algoritma machine learning (Mondol et al., 2023; Ibarra et al., 2023) seperti *XGBoost*, *random* forest, dan *extra trees* yang mampu mengklasifikasikan pasien PD menggunakan data klinis dan suara. *XGBoost* mengoptimalkan prediksi melalui *boosting* berbasis gradien, sementara *random forest* dan *extra trees* menggunakan bagging dengan acakan untuk mempercepat pelatihan dan mengurangi *overfitting*.

Pendekatan berbasis *data mining* (Handayani et al., 2021) menawarkan solusi efektif dengan algoritma *machine learning* (Govindu., 2023) seperti *XGBoost*, *random forest*, dan *extra trees* yang mampu mengklasifikasikan pasien PD menggunakan data klinis dan suara. *XGBoost* mengoptimalkan prediksi melalui *boosting* berbasis gradien, di mana model dibangun secara bertahap, dengan setiap langkah bertujuan mengurangi kesalahan dari langkah sebelumnya. Algoritma ini memperkuat kinerjanya melalui fungsi objektif terregularisasi yang menambahkan penalti untuk kompleksitas model, membantu mencegah *overfitting* pada data kompleks. Sementara itu, *random forest* dan *extra trees* menggunakan pendekatan *bagging*, yaitu pengacakan dan penggabungan beberapa pohon keputusan independen. Teknik ini mempercepat pelatihan dan mengurangi *overfitting* dengan memanfaatkan variasi yang diperoleh dari sampel yang berbeda pada setiap pohon, sehingga menghasilkan model yang lebih stabil dan andal untuk deteksi dini PD. Ketiga algoritma ini cocok untuk data kompleks, memungkinkan deteksi dini PD yang lebih akurat. Teknik seleksi fitur dan feature scaling juga berperan dalam meningkatkan akurasi model (Deepa & Khilar, 2024; Scimeca et al.,2023).

Algoritma seperti *XGBoost, random forest,* dan *extra trees* telah terbukti efektif dalam penelitian sebelumnya. Penelitian yang dilakukan oleh Karabayir et al. (2020) melaporkan akurasi *XGBoost* sebesar 81,6% dan *random forest* mencapai 81,8%, sementara Fahim et al. (2021) mencatat *extra trees* mencapai 77,08%. Dalam penelitian ini, kami mengoptimalkan model-model tersebut menggunakan data rekaman suara, serta menerapkan teknik *feature scaling* dan pemilihan fitur untuk meningkatkan akurasi. Penelitian oleh Deepa & Khilar (2024) menunjukkan bahwa teknik ini dapat secara signifikan meningkatkan akurasi klasifikasi, dengan contoh peningkatan akurasi *random forest* dari 79,74% menjadi 83,39% dan *extra trees* dari 82,39% menjadi 85,45%. Dalam penelitian ini, *Kernel Principal Component Analysis* (KPCA) diterapkan untuk mereduksi dimensi data suara, meningkatkan akurasi model klasifikasi, dan mengurangi *overfitting*. Selain itu, KPCA sebagai pengembangan dari PCA, digunakan untuk menangani data non-linear dengan memetakan data ke ruang fitur berdimensi lebih tinggi. Teknik ini membantu mengurangi dimensi data yang kompleks, mengeliminasi *noise*, dan meningkatkan efisiensi komputasi tanpa mengorbankan informasi penting.

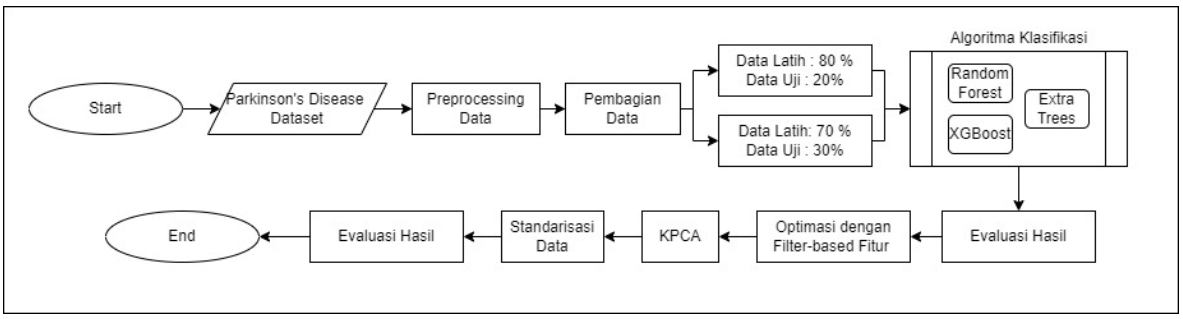
Deepa & Khilar (2024) dan Yudha & Muhammad (2023) telah menggunakan metode *ensemble* dan pemilihan fitur dalam mendeteksi dan mengklasifikasikan PD berdasarkan fitur suara. Namun, teknik pengurangan dimensi yang efektif, seperti KPCA, belum diintegrasikan secara menyeluruh untuk menjaga efisiensi model tanpa mengorbankan akurasi. Penelitian ini

mengusulkan penggunaan KPCA untuk mengurangi dimensi data sambil mempertahankan informasi penting, dengan harapan dapat mengatasi tantangan pengelolaan fitur dan menawarkan solusi yang lebih efisien dalam deteksi dini PD.

Secara keseluruhan, penelitian ini bertujuan untuk meningkatkan akurasi deteksi dini PD dengan mengatasi tantangan dalam pengelolaan jumlah fitur yang besar, yang kerap kali menyebabkan *overfitting* dan menurunkan kemampuan generalisasi model *machine learning* pada data baru. Dengan mengelola kompleksitas data melalui teknik seleksi fitur dan pengurangan dimensi seperti KPCA, penelitian ini diharapkan dapat meningkatkan ketepatan diagnosis, mempercepat proses deteksi dini, dan meminimalkan risiko *overfitting*. Dampak dari penelitian ini tidak hanya berfokus pada peningkatan kualitas perawatan pasien PD, tetapi juga membuka peluang untuk penggunaan lebih luas teknologi berbasis data dalam diagnostik medis. Pengembangan model ini dapat menjadi landasan bagi pendekatan berbasis data yang lebih canggih dan efisien dalam meningkatkan diagnosis penyakit neurodegeneratif di masa mendatang, serta mempercepat adopsi teknologi *machine learning* dalam sistem kesehatan global.

METODE

Penelitian ini merupakan jenis penelitian eksperimental yang bertujuan untuk meningkatkan akurasi prediksi penyakit PD melalui kombinasi algoritma klasifikasi dan teknik pengurangan dimensi. Algoritma utama yang digunakan dalam penelitian ini adalah random forest, *XGBoost*, dan *extra trees*. *random forest* merupakan metode ensemble yang bekerja dengan membuat banyak pohon keputusan dan mengambil hasil rata-rata untuk meningkatkan stabilitas prediksi. *XGBoost* adalah algoritma boosting yang bekerja dengan memperbaiki kesalahan prediksi dari model sebelumnya, membuatnya sangat efisien dan mampu menangani *overfitting*. *extra trees*, serupa dengan *random forest*, menggunakan pohon keputusan tetapi dengan cara pembelahan data yang lebih acak, sehingga sering kali lebih cepat dan lebih akurat pada *dataset* besar. Setelah diterapkan pada dataset, kinerja ketiga algoritma dibandingkan untuk menentukan algoritma terbaik.



Gambar 1. Tahapan penelitian

Gambar 1 merupakan tahapan penelitian ini yang dimulai dari pengumpulan data hingga penerapan model klasifikasi dan *hyperparameter tuning*. Dataset ini diperoleh dari *University of California Irvine* (UCI) *Machine Learning Repository* (Mittal & Sharma, 2021; Nijhawan et al., 2023; Khotiah et al., 2023) dan telah disetujui oleh Komite Bioetika dari University of Extremadura (Fahim et al., 2021). Dataset terdiri dari 80 subjek berusia di atas 50 tahun, dengan 40 pasien PD (27 pria, 13 wanita) dan 40 individu sehat (22 pria, 18 wanita). Dataset ini mencakup 44 fitur akustik dari analisis suara vokal selama pengucapan suara "/a/" selama 5 detik, yang diulang tiga kali per subjek. Setelah pengumpulan, data diproses melalui pembersihan, imputasi nilai yang hilang, dan normalisasi. Algoritma *random forest*, *XGBoost*, dan *extra trees* diterapkan sebagai model dasar tanpa optimasi *hyperparameter*, dan performanya digunakan sebagai *baseline*.

Selanjutnya, tuning dilakukan menggunakan *grid search* untuk mengoptimalkan kombinasi hyperparameter, sementara seleksi fitur berbasis filter dan KPCA diterapkan untuk mengurangi dimensi data, mengurangi *noise*, dan meningkatkan efisiensi model. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* untuk menilai kinerja model secara komprehensif dalam mendeteksi penyakit PD. Tahap *hyperparameter tuning* bertujuan meningkatkan performa model dengan mencari kombinasi parameter optimal untuk memaksimalkan metrik yang telah ditetapkan, sehingga prediksi menjadi lebih akurat dan stabil dibandingkan model *baseline*. Setelah tuning, optimasi fitur dilakukan dengan metode *filter-based* untuk menyaring fitur yang paling relevan dari dataset, membantu meningkatkan efisiensi model dan mengurangi risiko *overfitting* dengan mengeliminasi fitur-fitur yang kurang signifikan terhadap prediksi. Kombinasi kedua teknik ini bertujuan mengoptimalkan kecepatan komputasi serta akurasi prediksi, menghindari overfitting, dan menjaga kemampuan generalisasi model.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Evaluasi performa model dilakukan menggunakan berbagai metrik, yaitu akurasi, precision, *recall*, dan *F1-score*, untuk memberikan gambaran yang lebih holistik. Akurasi, yang dihitung dengan Persamaan (1) mengukur seberapa baik model memprediksi kasus yang benar, di mana *True Positives (TP)* dan *True Negatives (TN)* mencerminkan prediksi yang benar, sedangkan *False Positives (FP)* dan *False Negatives (FN)* adalah prediksi yang salah. *Precision* mengukur relevansi prediksi positif, *recall* menilai kemampuan model dalam mendeteksi seluruh instance kelas positif, dan *F1-score* memberikan keseimbangan antara *precision* dan *recall*, yang sangat penting dalam konteks dataset dengan ketidakseimbangan kelas.

HASIL DAN PEMBAHASAN

Hasil

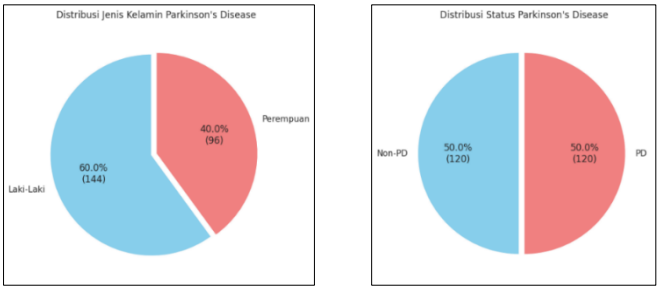
Dataset ini diperoleh dari *UCI Machine Learning Repository* dan telah disetujui oleh Komite Bioetika dari *University of Extremadura*. Dataset ini terdiri dari 80 subjek berusia di atas 50 tahun, yang terbagi menjadi dua kelompok: 40 subjek sehat (kontrol) yang terdiri dari 22 pria dan 18 wanita, serta 40 pasien PD, yang terdiri dari 27 pria dan 13 wanita. Semua pasien PD dalam dataset ini memiliki durasi penyakit lima tahun atau kurang, sesuai dengan skala *Unified Parkinson's Disease Rating Scale (UPDRS).* Dataset ini mencakup 44 fitur akustik yang dihasilkan dari analisis suara vokal berkelanjutan pada pengucapan suara "/a/" selama 5 detik, yang diulang tiga kali untuk setiap subjek. Setiap kategori fitur ini digunakan untuk menganalisis pola suara subjek guna mengidentifikasi kemungkinan adanya penyakit PD.

	ID	Recording	Status	Gender	Jitter_rel	Jitter_abs	Jitter_RAP	Jitter_PPQ	Shim_loc	Shim_dB	...	Delta3	Delta4	Delta5	Delta6	Delta7	Delta8	Delta9	Delta10	Delta11	Delta12
0	CONT-01	1	0	1	0.25546	0.000015	0.001467	0.001673	0.030256	0.26313	...	1.407701	1.417218	1.380352	1.420670	1.451240	1.440295	1.403678	1.405495	1.416705	1.354610
1	CONT-01	2	0	1	0.36964	0.000022	0.001932	0.002245	0.023146	0.20217	...	1.331232	1.227338	1.213377	1.352739	1.354242	1.365692	1.322870	1.314549	1.318999	1.323508
2	CONT-01	3	0	1	0.23514	0.000013	0.001353	0.001546	0.019338	0.16710	...	1.412304	1.324674	1.276088	1.429634	1.455996	1.368882	1.438053	1.388910	1.305469	1.305402
3	CONT-02	1	0	0	0.29320	0.000017	0.001105	0.001444	0.024716	0.20892	...	1.501200	1.534170	1.323993	1.496442	1.472926	1.643177	1.551286	1.638346	1.604008	1.621456
4	CONT-02	2	0	0	0.23075	0.000015	0.001073	0.001404	0.013119	0.11607	...	1.508468	1.334511	1.610694	1.685021	1.417614	1.574895	1.640088	1.533666	1.297536	1.382023

Gambar 2. Sampel data *parkinson's disease*

Gambar 2 menampilkan sampel data dari penelitian ini, mencakup informasi demografis subjek yang terlibat, yaitu individu sehat dan pasien PD. Dalam gambar 2, terlihat distribusi subjek berdasarkan status kesehatan, yang menunjukkan perbedaan signifikan antara jumlah individu sehat dan pasien PD, memberikan gambaran awal tentang populasi yang

diteliti. Sementara itu, Gambar 3 terbagi menjadi dua bagian: di sebelah kiri, distribusi berdasarkan jenis kelamin menunjukkan ketidakseimbangan antara jumlah pria dan wanita dalam masing-masing kelompok, yang dapat mempengaruhi hasil analisis suara; dan di sebelah kanan, distribusi status kesehatan mengilustrasikan jumlah individu sehat (0) dan terdiagnosis PD (1), yang menegaskan fokus dataset pada karakteristik suara yang berkaitan dengan diagnosis PD. Secara keseluruhan, gambar 2 dan 3 memberikan konteks penting tentang populasi yang diteliti dan distribusi subjek berdasarkan jenis kelamin dan status kesehatan, yang menjadi langkah awal dalam analisis lebih lanjut untuk mendeteksi pola suara yang berhubungan dengan penyakit PD.



Gambar 3. Hasil distribusi data berdasarkan jenis kelamin dan status

Normalisasi amplitudo dilakukan untuk mencegah variasi volume bicara mengganggu deteksi pola suara terkait Parkinson, yang dapat menurunkan akurasi prediksi akibat perbedaan tingkat suara antar pasien. Proses *data cleaning* meliputi pengecekan nilai hilang (*missing values*) dan nilai duplikat (*duplicate values*), dan analisis menunjukkan bahwa dataset tidak memiliki data yang hilang atau duplikat. Setelah memastikan data bersih, dataset dibagi menjadi dua bagian: 70% untuk data latih dan 30% untuk data uji, serta 80% untuk data latih dan 20% untuk data uji. Pembagian ini memastikan model mendapatkan variasi yang cukup dalam pelatihan dan pengujian untuk menghasilkan performa optimal.

Tabel 1. *Hyperparameter tuning* yang digunakan

Model	**Data Latih : Data Uji**	***Best Hyperparameters***						
		n_estimators	***bootstrap***	***criterion***	***max_depth***	***max_features***	***min_samples_leaf***	***min_samples_split***
Random Forest	70 : 30	50	*True*	entropy	3	sqrt	1	10
	80 : 20	100	*True*	gini	7	Log2	2	5
XGBoost	70 : 30	100	*True*	gini	7	auto	1	2
	80 : 20	50	*True*	gini	3	auto	1	2
Extra Trees	70 : 30	50	*True*	entropy	7	sqrt	2	10
	80 : 20	50	*True*	gini	7	Log2	2	2

Tabel 1 menunjukkan hasil *hyperparameter tuning* untuk tiga algoritma pembelajaran mesin: *random forest*, *XGBoost*, dan *extra trees*, dengan dua skenario pembagian data (70%:30% dan 80%:20%). Untuk pemilihan *hyperparameter*, digunakan metode *grid search* karena memberikan pencarian yang sistematis dan menyeluruh di seluruh ruang *hyperparameter*, sehingga memungkinkan identifikasi kombinasi terbaik. Pada Algoritma *random forest*, kombinasi hyperparameter terbaik berbeda pada tiap skenario, dengan *n_estimators* sebanyak 50 dan 100, serta perbedaan dalam penggunaan kriteria *entropy* atau *gini*, kedalaman maksimum pohon, dan metode pemilihan fitur (*sqrt* atau *log2*). Algoritma *XGBoost* menunjukkan hasil yang optimal dengan *n_estimators* 100 pada skenario pertama dan

50 pada skenario kedua, sementara *extra trees* mempertahankan *n_estimators* 50 di kedua skenario dengan variasi kecil dalam kriteria dan kedalaman pohon. Proses *tuning* ini menunjukkan bahwa kombinasi *hyperparameter* optimal bervariasi tergantung pada algoritma dan proporsi data latih/uji. *Tuning* yang baik memastikan model menghindari *overfitting* atau *underfitting*, sehingga kinerja model dapat ditingkatkan dalam hal metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score*.

Tabel 2. Perbandingan algoritma

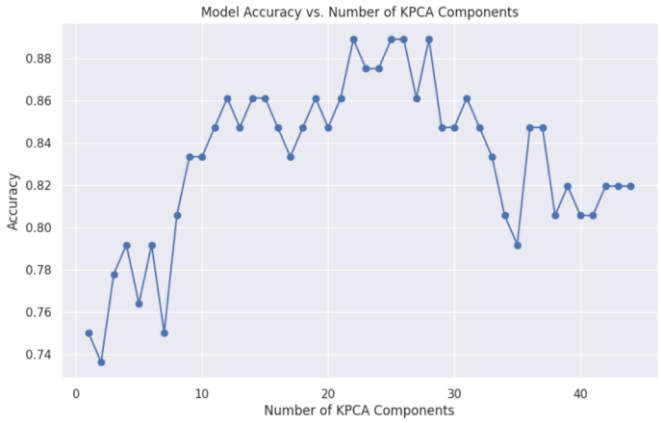
Algoritma	**Metode**	**Data Latih : Data Uji**	**Akurasi**	**Presisi**	***Recall***	***F1-Score***	***Cross Validation***
Random Forest	Normal	70 : 30	0.8055	0.8062	0.8055	0.8052	0.8085
		80 : 20	0.7916	0.7937	0.7916	0.7913	0.7533
	Hyperparameter	70 : 30	0.8055	0.8055	0.8055	0.8055	0.7942
		80 : 20	0.7916	0.7937	0.7916	0.7913	0.7733
XGBoost	Normal	70 : 30	0.8472	0.8476	0.8472	0.8472	0.7933
		80 : 20	0.7708	0.7831	0.7708	0.7683	0.7688
	Hyperparameter	70 : 30	0.8611	0.8611	0.8611	0.8611	0.7933
		80 : 20	0.8125	0.8174	0.8125	0.8117	0.7688
Extra Trees	Normal	70 : 30	0.8472	0.8476	0.8472	0.8472	0.8342
		80 : 20	0.8125	0.8130	0.8125	0.8124	0.7888
	Hyperparameter	70 : 30	0.8194	0.8195	0.8194	0.8193	0.7933
		80 : 20	0.7708	0.7713	0.7708	0.7933	0.7933

Tabel 2 menampilkan perbandingan kinerja tiga algoritma pembelajaran mesin: *random forest*, *XGBoost*, dan *extra trees*, baik sebelum maupun sesudah dilakukan *hyperparameter tuning*. Setiap algoritma diuji dengan dua proporsi data latih dan data uji yang berbeda (70%:30% dan 80%:20%) dan dievaluasi berdasarkan metrik akurasi, presisi, recall, *F1-score*, serta *cross validation*. Hasil *tuning hyperparameter* menunjukkan bahwa *XGBoost* mengalami peningkatan kinerja yang signifikan, khususnya pada pembagian data 70%:30%, dengan peningkatan akurasi dari 0.8472 menjadi 0.8611. Pencapaian ini menegaskan kemampuan *XGBoost* dalam menangani dataset besar dan kompleks, yang sangat penting dalam konteks diagnostik medis. Akurasi yang tinggi pada model ini memungkinkan lebih banyak prediksi yang benar, berpotensi meningkatkan keandalan diagnosis dan pengobatan pasien.

Sebagai langkah lebih lanjut, *XGBoost* dioptimasi menggunakan KPCA dengan *kernel* non-linear untuk menangani kompleksitas data, serta dilakukan standarisasi pada fitur sebelum diterapkan pada model. Standarisasi dan reduksi dimensi dengan KPCA diharapkan dapat meningkatkan kinerja model dengan mengeliminasi *noise* pada fitur yang tidak relevan, sehingga menghasilkan klasifikasi yang lebih akurat dan efisien. Melalui pendekatan ini, kami berharap dapat memperoleh wawasan yang lebih mendalam tentang karakteristik suara yang terkait dengan penyakit Parkinson serta meningkatkan kemampuan klasifikasi dalam konteks kesehatan. Pemilihan fitur menggunakan KPCA merupakan prosedur statistik yang sangat efektif untuk ekstraksi fitur dan pengurangan dimensionalitas, terutama dalam menangani data non-linear. Berbeda dengan PCA konvensional, KPCA memanfaatkan fungsi *kernel* untuk mengubah data asli ke dalam ruang fitur berdimensi lebih tinggi sebelum menerapkan transformasi ortogonal.

Gambar 4 menampilkan hubungan antara jumlah komponen KPCA dengan akurasi model yang dihasilkan. Berdasarkan grafik tersebut, terlihat bahwa performa model meningkat secara signifikan seiring dengan bertambahnya jumlah komponen KPCA hingga mencapai sekitar 26 komponen, di mana akurasi model berada pada titik optimalnya. Setelah titik ini,

penambahan komponen tidak lagi memberikan peningkatan yang signifikan terhadap akurasi, yang menunjukkan bahwa 26 komponen utama merupakan jumlah optimal dalam menjaga keseimbangan antara pengurangan dimensi dan retensi informasi penting. Dengan hanya menggunakan 26 komponen, kompleksitas dataset berhasil disederhanakan tanpa mengorbankan akurasi prediksi, sehingga mempercepat proses komputasi dan mengurangi risiko overfitting. Gambar 4 mengilustrasikan bahwa KPCA tidak hanya berfungsi untuk mereduksi dimensi, tetapi juga memainkan peran kunci dalam meningkatkan kinerja keseluruhan model, terutama ketika diterapkan bersama proses standarisasi data, yang memastikan setiap fitur memiliki skala yang seimbang.



Gambar 4. *Model accuracy vs. number of kpca components*

Tabel 3. Metriks evaluasi sebelum dan sesudah dilakukan kpca

Classifiers	***XGBoost***	***XGBoost + Hyperparameter***	***XGBoost + KPCA***
Akurasi	0.8472	0.8611	0.8893
Presisi	0.8476	0.8611	0.8934
Recall	0.8472	0.8611	0.8889
F1-Score	0.8472	0.8611	0.8884
Cross Validation	0.7933	0.7933	0.7914

Tabel 3 membandingkan metrik evaluasi model *XGBoost* sebelum dan sesudah dilakukan *hyperparameter tuning* serta setelah penerapan KPCA. Hasilnya menunjukkan peningkatan yang signifikan setelah KPCA diterapkan. Akurasi awal model yang sebesar 0.8472 meningkat menjadi 0.8611 setelah *hyperparameter tuning*, dan lebih lanjut meningkat menjadi 0.8893 setelah KPCA diterapkan. Selain akurasi, metrik lainnya seperti presisi, *recall*, dan *F1-Score* juga mengalami peningkatan setelah KPCA, dengan masing-masing mencapai nilai tertinggi 0.8934, 0.8889, dan 0.8884. Walaupun *cross-validation* sedikit menurun dari 0.7933 menjadi 0.7914 setelah penerapan KPCA, penurunan ini tidak mengurangi efek positif KPCA dalam meningkatkan performa keseluruhan model. Selanjutnya, setelah dilakukannya klasifikasi diagnosis, penggunaan data longitudinal untuk memprediksi progresi penyakit Parkinson dapat menjadi arah penelitian selanjutnya. Pendekatan ini dapat memberikan wawasan yang lebih dalam mengenai perkembangan penyakit dan membantu dalam membuat keputusan klinis yang lebih tepat.

Pembahasan

Penelitian ini mengevaluasi kinerja model *machine learning* untuk deteksi penyakit Parkinson menggunakan dataset dari *UCI Machine Learning Repository*, yang terdiri dari 80 subjek (40 pasien Parkinson dan 40 individu sehat). Dalam penelitian ini, kinerja model

machine learning untuk deteksi penyakit Parkinson menunjukkan variasi antara algoritma *random forest, XGBoo*st, dan *extra trees*. *XGBoost* mencapai akurasi tertinggi sebesar 0.8893 setelah *tuning hyperparameter*, sementara *random forest* dan *extra trees* memiliki akurasi lebih rendah, dengan *random forest* hanya mencapai 0.8055 pada pengujian normal. Hal ini disebabkan oleh karakteristik algoritma, di mana *XGBoost*, sebagai algoritma *boosting*, lebih baik dalam menangkap pola kompleks, sedangkan *random forest* dan *extra trees* mengandalkan beberapa pohon keputusan yang kurang efektif dalam mengatasi interaksi antar fitur. Selain itu, pengaturan *hyperparameter* yang mungkin tidak optimal juga dapat memengaruhi hasil, mencerminkan bahwa tuning tidak selalu meningkatkan performa. Secara keseluruhan, meskipun *random forest* dan *extra trees* menunjukkan hasil yang layak, mereka kurang optimal dibandingkan *XGBoost* dalam mendeteksi pola spesifik yang berkaitan dengan penyakit Parkinson.

KPCA bekerja dengan mengubah data asli ke dalam ruang fitur berdimensi lebih tinggi menggunakan fungsi kernel, seperti *gaussian* atau *polynomial*, untuk menangkap pola non-linear. Proses ini memungkinkan transformasi data ke dalam ruang fitur di mana hubungan kompleks dapat diungkap. Dalam penelitian ini, hingga 26 komponen KPCA ditemukan optimal karena mampu mempertahankan informasi penting tanpa *overfitting*. Setelah transformasi, PCA diterapkan untuk menemukan komponen utama yang lebih representatif terhadap variasi data. KPCA berhasil mengeliminasi *noise* dan fitur yang tidak relevan, meningkatkan informasi yang digunakan oleh algoritma pembelajaran mesin. Pentingnya standarisasi fitur sebelum penerapan KPCA juga sangat krusial, karena memastikan semua fitur berada pada skala yang seimbang, mengurangi bias dari perbedaan skala antar fitur.

Model *XGBoost* mencapai akurasi tertinggi sebesar 0.8893, melampaui hasil penelitian sebelumnya (Karabayir et al., 2020; Deepa & Khilar, 2024). Peningkatan akurasi ini disebabkan oleh pemilihan fitur yang relevan, teknik pengolahan data yang tepat, dan *hyperparameter tuning* yang efektif. Dengan variabel yang mencakup data akustik dan fitur klinis, *XGBoost* dapat menangkap pola kompleks dalam dataset. Peningkatan akurasi *random forest* dan *extra trees* juga mencerminkan kemajuan berkat penerapan teknik standarisasi sebelum KPCA. Penelitian ini menunjukkan potensi algoritma dalam mendeteksi Parkinson secara akurat, menandakan kemajuan teknologi *data mining* di bidang medis. Untuk mencegah perkembangan penyakit ini, penerapan skrining dini dengan model prediktif berbasis data sangat penting, memungkinkan identifikasi pasien berisiko tinggi sebelum gejala klinis muncul. Kombinasi deteksi dini, intervensi tepat waktu, dan program edukasi dapat menjadi strategi efektif dalam mencegah perkembangan penyakit Parkinson.

SIMPULAN

Penelitian ini menyoroti pentingnya penerapan standarisasi dan KPCA dalam meningkatkan performa model XGBoost untuk mendeteksi penyakit Parkinson berdasarkan karakteristik suara. Penerapan KPCA secara efektif mengurangi dimensi data tanpa kehilangan informasi penting, meningkatkan akurasi model dari 0.8472 menjadi 0.8893 dengan penggunaan 26 komponen utama yang optimal. Selain itu, penelitian ini juga menunjukkan bahwa algoritma *random forest* dan *extra trees* mengalami peningkatan performa signifikan setelah penerapan hyperparameter tuning dan KPCA, yang mencerminkan efektivitas metode ini dalam meningkatkan keandalan model. Dampak positif dari hyperparameter tuning dan standarisasi membantu mencegah *overfitting*, memastikan semua fitur berkontribusi proporsional dalam pembelajaran model, serta mempercepat konvergensi dan meningkatkan akurasi prediksi. Implikasi dari temuan ini adalah bahwa pendekatan terintegrasi, yang menggabungkan teknik-teknik tersebut, sangat penting dalam pengembangan model pembelajaran mesin yang efisien dan efektif untuk deteksi dini penyakit Parkinson, yang pada gilirannya dapat berkontribusi pada pengobatan dan manajemen pasien secara lebih efektif.

REFERENSI

Alalayah, K. M., Senan, E. M., Atlam, H. F., Ahmed, I. A., & Shatnawi, H. S. A. (2023). Automatic and early detection of Parkinson's disease by analyzing acoustic signals using classification algorithms based on recursive feature elimination method. *Diagnostics, 13(11)*, 1-24. https://doi.org/10.3390/diagnostics13111924

Ananda, I. K., Fanani, A. Z., Setiawan, D., & Wicaksono, D. F. (2024). Penerapan random oversampling dan algoritma boosting untuk memprediksi kualitas buah jeruk. *Edumatic: Jurnal Pendidikan Informatika, 8*(1), 282–289. https://doi.org/10.29408/edumatic.v8i1.25836

Aprilitaz, W., Akbar, R.., Cahya Prayogi, R., & Rahmaddeni. (2023). Komparasi Algoritma K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penyakit Parkinson. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, *1*(1), 188-193.

Deepa, P., & Khilar, R. (2024). Parkinson's disease detection and classification: Leveraging voice features and ensemble methods with feature selection and ERT classifier. *Procedia Computer Science, 235*, 1695–1706. https://doi.org/10.1016/j.procs.2024.04.160

Desiani, A., Narti, N., Ramayanti, I., Arhami, M., & Irmeilyana, I. (2023). Diagnosa penyakit Parkinson dengan algoritma K-Nearest Neighbor dan Decision Tree C4.5. *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, *12*(1), 47–58.

Fahim, M. I., Islam, S., Noor, S. T., Hossain, M. J., & Setu, M. S. (2021). Machine learning model to analyze telemonitoring dysphemia factors of Parkinson's disease. *International Journal of Advanced Computer Science and Applications, 12*(8), 786–795. https://doi.org/10.14569/IJACSA.2021.0120890

Fahira, N. R., Lawi, A., & Aqsha, M. (2023). Early detection model of Parkinson's disease using random forest method on voice frequency data. *Journal of Natural Sciences and Mathematics Research, 9(1), 29-37.* https://doi.org/10.21580/jnsmr.2023.9.1.13148

Farida, Y., Ulinnuha, N., Sari, S. K., & Desinaini, L. N. (2023). Comparing support vector machine and Naïve Bayes methods with a selection of fast correlation based filter features in detecting Parkinson's disease. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, 14(2), 80-90.* https://doi.org/10.24843/LKJITI.2023.v14.i02.p02

Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using machine learning. *Procedia Computer Science, 218*, 249–261. https://doi.org/10.1016/j.procs.2023.01.007

Handayani, P. K. (2021). Penerapan algoritma support vector machine (SVM) untuk analisis pola klasifikasi pada Parkinson's dataset. *Indonesian Journal of Technology, Informatics and Science (IJTIS), 3*(1), 31–35. https://doi.org/10.24176/ijtis.v3i1.7530

Ibarra, E. J., Arias-Londoño, J. D., Zañartu, M., & Godino-Llorente, J. I. (2023). Towards a corpus (and language)-independent screening of Parkinson's disease from voice and speech through domain adaptation. *Bioengineering, 10*(11), 1-19. https://doi.org/10.3390/bioengineering10111316

Iyer, A., Kemp, A., Rahmatallah, Y., Pillai, L., Glover, A., Prior, F., Larson-Prior, L., & Virmani, T. (2023). A machine learning method to process voice samples for identification of Parkinson's disease. *Scientific Reports, 13*(1), 1-9. https://doi.org/10.1038/s41598-023-47568-w

Karabayir, I., Goldman, S. M., Pappu, S., & Akbilgic, O. (2020). Gradient boosting for Parkinson's disease diagnosis from voice recordings. *BMC Medical Informatics and Decision Making, 20*(1), 1-7. https://doi.org/10.1186/s12911-020-01250-7

Khotiah, T., Abdillah, D. F., K, I. B., Arianto, F., & Rohman, A. (2023). Comparison of machine learning techniques in the classification of Parkinson's disease sufferers.

Journal of Computer Networks, Architecture and High Performance Computing, 5(1), 129–137. https://doi.org/10.47709/cnahpc.v5i1.2035

Malekroodi, H. S., Madusanka, N., Lee, B. Il, & Yi, M. (2024). Leveraging deep learning for fine-grained categorization of Parkinson's disease progression levels through analysis of vocal acoustic patterns. *Bioengineering, 11*(3), 1-23. https://doi.org/10.3390/bioengineering11030295

Mittal, V., & Sharma, R. K. (2021). Machine learning approach for classification of Parkinson disease using acoustic features. *Journal of Reliable Intelligent Environments, 7*(3), 233–239. https://doi.org/10.1007/s40860-021-00141-6

Mondol, S. I. M. M. R., Kim, R., & Lee, S. (2023). Hybrid machine learning framework for multistage Parkinson's disease classification using acoustic features of sustained Korean vowels. *Bioengineering, 10*(8), 1-15. https://doi.org/10.3390/bioengineering10080984

Nainggolan, K. R., Purnamasari, F., & Pulungan, A. F. (2023). Prediksi penyakit Parkinson melalui dataset rekam suara dengan menggunakan algoritma deep neural network. *Jurnal Minfo Polgan, 12*(1), 401-409.

Nijhawan, R., Kumar, M., Arya, S., Mendirtta, N., Kumar, S., Towfek, S. K., Khafaga, D. S., Alkahtani, H. K., & Abdelhamid, A. A. (2023). A novel artificial-intelligence-based approach for classification of Parkinson's disease using complex and large vocal features. *Biomimetics, 8*(4), 1-19. https://doi.org/10.3390/biomimetics8040351

Pramanik, M., Pradhan, R., Nandy, P., Bhoi, A. K., & Barsocchi, P. (2023). The ForEx++ based decision tree ensemble approach for robust detection of Parkinson's disease. *Journal of Ambient Intelligence and Humanized Computing, 14*(9), 11429–11453. https://doi.org/10.1007/s12652-022-03719-x

Scimeca, S., Amato, F., Olmo, G., Asci, F., Suppa, A., Costantini, G., & Saggio, G. (2023). Robust and language-independent acoustic features in Parkinson's disease. *Frontiers in Aging Neuroscience, 15*, 1-19. https://doi.org/10.3389/fneur.2023.1198058

Yudha, E. P., & Muhammad, N. F. (2023). Sistem otomatis untuk deteksi penyakit Parkinson menggunakan fuzzy K-NN. Jurnal Teknik Komputer, 9(2), 176–184. https://doi.org/10.31294/jtk.v9i2.15933