Jurnal Qua Teknika, Vol. 15 No. 2 September 2025
ISSN 2088-2424 (cetak); ISSN 2527-3892 (elektronik)
Fakultas Teknik Universitas Islam Balitar, Blitar
Https://ejournal.unisbablitar.ac.id/index.php/qua; Email: quateknika@unisbablitar.ac.id

PENERAPAN RANDOM FOREST UNTUK DETEKSI DINI PENYAKIT PARKINSON'S DENGAN DATA FREKUENSI SUARA

Mohammad Annan Makruf Mustofa[1)], Sucipto[2)], Arie Nugroho[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: makrufmustofa79@gmail.com[1)],
sucipto@unpkediri.ac.id[2)], arienugroho@unpkediri.ac.id[3)]

ABSTRACT
Parkinson's Disease is a progressive neurological disorder that affects motor functions and verbal communication of the patients. Early detection of this disease is crucial to improving patients' quality of life. This study aims to develop an early detection system for Parkinson's Disease by utilizing sound frequency as the primary feature. The algorithm employed in this research is Random Forest, with the analysis process following the CRISP-DM approach, which includes six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Based on the test results, the developed model achieved an accuracy of 94.92% on the dataset used. These findings indicate that the Random Forest algorithm can be effectively implemented as an early detection system for Parkinson's Disease using sound frequency data.This model demonstrates strong potential to be implemented as an effective, non-invasive screening tool to support early clinical diagnosis of Parkinson's Disease.

Keywords: Parkinson's Disease, Random Forest, Voice Frequency, Machine Learning, Early Detection.

ABSTRAK
Penyakit Parkinson merupakan gangguan neurologis progresif yang memengaruhi fungsi motorik dan kemampuan komunikasi verbal pada penderitanya. Deteksi dini terhadap penyakit ini sangat penting untuk meningkatkan kualitas hidup pasien. Penelitian ini bertujuan untuk mengembangkan sistem deteksi dini penyakit Parkinson dengan memanfaatkan frekuensi suara sebagai fitur utama. Algoritma yang digunakan dalam penelitian ini adalah Random Forest, dengan proses analisis yang mengikuti pendekatan CRISP-DM yang mencakup enam tahapan: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Berdasarkan hasil pengujian, model yang dikembangkan berhasil mencapai akurasi sebesar 94,92% pada dataset yang digunakan. Temuan ini menunjukkan bahwa algoritma Random Forest dapat diimplementasikan secara efektif sebagai sistem deteksi dini penyakit Parkinson berbasis data frekuensi suara. Model ini menunjukkan potensi yang kuat untuk diimplementasikan sebagai alat skrining non-invasif yang efektif dalam mendukung diagnosis awal penyakit Parkinson.

Kata Kunci: Penyakit Parkinson's, Random Forest, Frekuensi Suara, Machine Learning, Deteksi Dini.

PENDAHULUAN

Parkinson's Disease merupakan gangguan neurodegeneratif yang menyerang sistem saraf pusat [1], terutama pada neuron dopaminergik di substantia nigra, sehingga berdampak pada kontrol gerakan motorik secara bertahap. Salah satu gejala utama yang sering muncul pada pasien Parkinson adalah perubahan pada pola bicara seperti suara yang lebih rendah, tremor vokal, serta kesulitan dalam artikulasi dan intonasi [2]. Oleh karena itu, analisis data frekuensi suara menjadi salah satu metode alternatif yang dapat digunakan untuk mendeteksi adanya penyakit ini secara dini. Dengan pendekatan teknologi informasi dan *machine learning*, deteksi awal dapat dilakukan secara efisien dan akurat [3], membantu tenaga medis dalam proses diagnosis serta meningkatkan kualitas hidup pasien melalui intervensi lebih cepat.

Pendekatan berbasis data memiliki potensi besar dalam membangun sistem pendeteksi penyakit berbasis suara [4]. Data frekuensi suara bersifat numerik dan mudah diperoleh melalui rekaman audio yang kemudian diekstraksi menjadi fitur-fitur statistik yang relevan. Fitur tersebut dapat mencerminkan karakteristik vokal seseorang, baik yang sehat maupun yang mengalami gangguan neurologis seperti *Parkinson*. Dengan menggunakan model *machine learning* yang tepat, hubungan antara fitur suara dan status kesehatan individu

dapat dipelajari dan digunakan untuk melakukan prediksi secara otomatis tanpa memerlukan pemeriksaan fisik langsung. Hal ini menjadikan metode berbasis suara sebagai solusi non-invasif, murah, dan mudah diimplementasikan.

Salah satu algoritma *machine learning* yang populer dalam tugas klasifikasi adalah *Random Forest*. Algoritma ini menggunakan ensemble dari pohon keputusan untuk membuat prediksi yang stabil dan akurat [5]. Selain itu, *Random Forest* memiliki kemampuan yang baik dalam menangani *overfitting* [6], *robust* terhadap data yang tidak seimbang, dan mampu memberikan estimasi pentingnya fitur dalam proses klasifikasi. Dalam konteks deteksi penyakit *Parkinson*, *Random Forest* sangat layak digunakan karena dataset yang digunakan biasanya memiliki banyak fitur numerik yang saling berkorelasi dan membutuhkan model yang kuat namun tetap *interpretable* untuk memberikan hasil yang dapat dipercaya. Algoritma ini sangat cocok untuk dataset yang bertipe numerik. Semakin banyak fitur yang digunakan, semakin banyak variasi pohon yang ada [7].

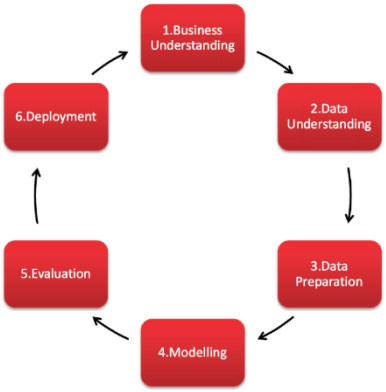
Permasalahan utama yang melatarbelakangi penelitian ini adalah belum tersedianya sistem deteksi dini yang bersifat non-invasif dan dapat diimplementasikan secara praktis pada ranah klinis berbasis data suara. Selain itu, ketidakseimbangan data pada dataset *Parkinson* seringkali menghambat performa model klasifikasi, sehingga diperlukan pendekatan algoritmik yang tidak hanya akurat tetapi juga mampu menangani kondisi tersebut secara adil. *Random Forest* menjadi alternatif yang potensial, namun studi yang menguji efektivitasnya dengan penyesuaian *class_weight='balanced'* masih terbatas. Oleh karena itu, penelitian ini dirancang untuk menjawab kebutuhan akan metode klasifikasi yang andal dan efisien dengan akurasi tinggi pada kasus deteksi dini *Parkinson* berbasis frekuensi suara.

Penelitian ini menerapkan *hyperparameter* seperti `*class_weight='balanced'*` untuk mengatasi ketidakseimbangan data, hasilnya *Random Forest* mampu mencapai akurasi sebesar 94.92%, pendekatan ini terbukti efektif dalam meningkatkan sensitivitas dan akurasi prediksi pada dataset yang tidak seimbang [8]. Sebagai perbandingan, studi oleh Elshewey et al. (2023) menerapkan model *Support Vector Machine* (*SVM*) dengan optimasi *Bayesian* (*BO-SVM*) untuk klasifikasi penyakit *Parkinson* menggunakan dataset yang sama [9]. Dalam penelitian tersebut, model *SVM* yang dioptimalkan melalui *Bayesian Optimization* berhasil mencapai akurasi sebesar 92.3%, dan menunjukkan keunggulan dibandingkan model lain seperti *Logistic Regression*, *Decision Tree*, dan *Naive Bayes* dalam kondisi pengujian serupa.

Perbandingan ini menunjukkan bahwa baik *Random Forest* maupun *BO-SVM* memiliki potensi tinggi dalam klasifikasi penyakit Parkinson. Namun, implementasi *Random Forest* dalam penelitian ini tidak hanya mampu menghasilkan akurasi yang lebih tinggi, tetapi juga lebih sederhana dari sisi tuning dan interpretasi model, menjadikannya pilihan yang sangat tepat untuk digunakan dalam sistem deteksi dini. Proses pengembangan model dilakukan dengan mengacu pada metodologi *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*), yang mencakup enam tahapan utama mulai dari pemahaman bisnis hingga implementasi sistem [10]. Dengan demikian, penelitian ini tidak hanya menghadirkan model prediktif yang unggul, tetapi juga memperkuat kontribusi pada kajian deteksi dini penyakit *Parkinson* melalui pendekatan machine learning yang adaptif dan aplikatif.

METODE PENELITIAN

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 [11]. 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 [12].



Gambar 1. Alur CRISP-DM

A. Business Understanding

Pada tahap pemahaman bisnis, beberapa aktivitas yang dilakukan meliputi penetapan tujuan bisnis, evaluasi situasi yang ada, serta penentuan tujuan data mining [13]. Dalam konteks ini, tujuan utama penelitian adalah mengembangkan model deteksi dini penyakit *Parkinson* berdasarkan karakteristik suara pasien, dengan menggunakan algoritma *Random Forest*. Fokus utama adalah menciptakan model klasifikasi yang akurat dan dapat digunakan sebagai sistem pendukung keputusan dalam diagnosis awal.

B. Data Understanding

Data Understanding merupakan fase dalam proses *data mining* yang bertujuan untuk mengumpulkan data awal, mempelajari karakteristik data, mengenali data yang akan digunakan, mengidentifikasi permasalahan terkait kualitas data, serta mendeteksi subset data yang relevan untuk membentuk hipotesis awal [14]. Dataset dalam penelitian ini berasal dari sumber terbuka (Kaggle) yang memuat 195 data rekaman suara pasien, terdiri dari 23 fitur akustik serta label target '*status*' yang menunjukkan apakah individu tersebut menderita *Parkinson* atau tidak. Proses ini melibatkan analisis distribusi data, identifikasi nilai yang hilang, *outlier*, serta pemahaman korelasi antar fitur.

C. Data Preparation

Tahap *Data Preparation* mencakup seluruh aktivitas untuk membangun kumpulan data akhir yang akan digunakan dalam pemodelan, mulai dari data mentah. Tahap ini biasanya dilakukan berulang kali dan tidak harus mengikuti urutan tertentu. Kegiatannya meliputi pemilihan tabel, record, dan atribut, serta transformasi dan pembersihan data [15]. Tahapan ini memiliki peran yang sangat menentukan terhadap keberhasilan proses klasifikasi, karena kualitas data memiliki dampak langsung terhadap kinerja model. Adapun langkah-langkah yang dilakukan pada tahap ini mencakup hal-hal berikut:

1. Penghapusan Atribut Tidak Relevan

   Pada tahap awal, dilakukan penghapusan atribut yang tidak memiliki kontribusi signifikan terhadap proses klasifikasi. Dalam hal ini, kolom name dihapus dari dataset karena hanya berisi informasi identitas pasien dan tidak memiliki pengaruh terhadap hasil prediksi.

2. Pemisahan Fitur dan Target

   Setelah atribut tidak relevan dihapus, data kemudian dipisahkan menjadi dua bagian utama, yaitu fitur (X) dan target (y). Fitur terdiri dari 22 variabel numerik terkait karakteristik suara, sedangkan target adalah label status yang menunjukkan apakah seorang individu menderita *Parkinson* (1) atau tidak (0).

3. Membangun Data Pelatihan dan Pengujian

   Dataset kemudian dibagi menjadi data pelatihan dan data pengujian dengan proporsi 70:30 menggunakan metode *train-test split*. Hal ini dilakukan untuk memastikan bahwa model diuji pada

data yang belum pernah dilihat sebelumnya, sehingga hasil evaluasi dapat merepresentasikan performa model secara umum.

4. Penanganan Ketidakseimbangan Kelas

   Untuk mengatasi distribusi kelas yang tidak seimbang antara penderita dan non-penderita *Parkinson*, digunakan parameter *class_weight='balance*d' dalam algoritma *Random Forest*. Penyesuaian bobot ini memungkinkan model untuk memberikan perhatian yang proporsional terhadap kelas minoritas, sehingga dapat meningkatkan sensitivitas dan akurasi keseluruhan model.

D. Modeling

Tahap *Modeling* merupakan proses pembangunan dan pelatihan model klasifikasi berdasarkan data yang telah disiapkan sebelumnya. Penelitian ini menggunakan algoritma *Random Forest* karena memiliki kinerja yang baik dalam menangani fitur berdimensi tinggi serta memberikan prediksi yang stabil dan akurat. Seperti yang dijelaskan dalam [16], *Random Forest* menunjukkan performa kuat dalam membedakan pasien *Parkinson* dan individu sehat berdasarkan fitur suara, dengan akurasi tinggi dan ketahanan terhadap *overfitting*. Model dilatih menggunakan data latih hasil *train-test split* dan dioptimalkan dengan parameter *n_estimators=100*, *random_state=42*, dan *class_weight='balanced'*. Penyesuaian parameter ini bertujuan untuk memastikan bahwa model mampu mengenali pola dari kedua kelas secara adil, terutama dalam kondisi data yang tidak seimbang. Proses pelatihan menghasilkan model yang siap untuk digunakan dalam prediksi data uji guna mengukur performanya secara objektif.

E. Evaluation

Tahap kelima adalah evaluasi performa model. Evaluasi ini dilakukan untuk menilai sejauh mana model mampu mengklasifikasikan data secara akurat, khususnya dalam konteks deteksi penyakit yang sensitif terhadap kesalahan prediksi. Dalam penelitian ini, performa model dinilai menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengukur efektivitas klasifikasi secara menyeluruh. Pendekatan evaluasi serupa juga digunakan dalam studi oleh Airlangga [17], di mana algoritma *Random Forest* dibandingkan dengan beberapa algoritma lain dalam tugas identifikasi pembicara. Hasilnya menunjukkan bahwa *Random Forest* memiliki performa yang kompetitif dengan nilai evaluasi yang konsisten tinggi pada keempat metrik tersebut. Metrik evaluasi yang digunakan memberikan gambaran lengkap mengenai ketepatan prediksi, keseimbangan performa antar kelas, dan konsistensi model dalam mengklasifikasikan data baru. Evaluasi ini juga membantu menilai kelayakan model untuk diterapkan dalam sistem deteksi dini yang akurat dan andal.

Secara struktural, *Confusion Matrix* terdiri dari empat komponen utama:

1. *True Positive* (TP) : Jumlah kasus di mana model memprediksi individu sebagai penderita *Parkinson* (kelas 1), dan hasil aktualnya memang positif (penderita *Parkinson*).
2. *True Negative* (TN) : Jumlah kasus di mana model memprediksi individu sebagai sehat (kelas 0), dan hasil aktualnya memang negatif (sehat).
3. *False Positive* (FP) : Jumlah kasus di mana model memprediksi individu sebagai penderita *Parkinson* (kelas 1), tetapi hasil aktualnya adalah sehat (kelas 0). Hal ini sering disebut sebagai *Type I error* .
4. *False Negative* (FN) : Jumlah kasus di mana model memprediksi individu sebagai sehat (kelas 0), tetapi hasil aktualnya adalah penderita *Parkinson* (kelas 1). Hal ini dikenal sebagai *Type II error* .

F. Deployment

Tahap terakhir adalah merancang implementasi model ke dalam lingkungan nyata atau sistem pendukung keputusan klinis. Meskipun pada penelitian ini tahap *deployment* masih bersifat konseptual, hasil dan model yang diperoleh dapat diintegrasikan ke dalam aplikasi berbasis web atau sistem klinik sebagai alat bantu diagnosis non-invasif. Dengan demikian, proses CRISP-DM memastikan bahwa seluruh tahapan penelitian dilakukan secara sistematis dan dapat dipertanggungjawabkan secara ilmiah.

HASIL DAN PEMBAHASAN

A. Business Understanding

Pada tahap ini, dilakukan pemahaman terhadap kebutuhan bisnis atau masalah yang ingin diselesaikan melalui penerapan teknik *data mining*. Sub-tahapan dalam CRISP-DM untuk bagian ini adalah sebagai berikut:

a. Determine Business Objectives

Tujuan utama dari penelitian ini adalah pengembangan sistem deteksi dini penyakit *Parkinson* menggunakan fitur frekuensi suara berbasis algoritma *Random Forest*. Sistem ini dirancang untuk memberikan prediksi awal terhadap risiko penyakit *Parkinson* hanya dengan menganalisis parameter vokal pasien.

b. Assess Situation

Masalah klinis yang menjadi fokus adalah ketiadaan alat pendeteksi dini yang bersifat non-invasif, mudah diakses, dan akurat. Dengan memanfaatkan *machine learning* dan data suara, solusi dapat dikembangkan tanpa memerlukan prosedur medis invasif. Dataset yang digunakan berasal dari hasil rekaman suara pasien *Parkinson* dan subjek sehat, sehingga layak untuk analisis lebih lanjut.

c. Determine Data Mining Goals

Dari perspektif *data mining*, tujuan utamanya adalah pembangunan model klasifikasi biner yang mampu membedakan antara pasien *Parkinson* (*status*=1) dan subjek sehat (*status*=0). Model ini diharapkan memiliki performa tinggi dalam hal *precision* dan *recall* agar tidak menghasilkan banyak *false negative*.

B. Data Understanding

a. Collect Initial Data

Data yang digunakan adalah dataset publik yang didapat dari Kaggle dan dapat diakses melalui tautan berikut ini: *https://www.kaggle.com/datasets/jainaru/parkinson-disease-detection*. Dataset ini memiliki 196 sample data dan 24 fitur, Dapat dilihat pada tabel 1.

Tabel 1. Dataset Parkinson's

NO	Dataset Parkinson's						
	Name	**MDVP: Fo(Hz)**	**MDVP: Fhi(Hz)**	**…..**	***spread2***	***D2***	***PPE***
1	phon_R01_S01_1	119.992	157.302	…..	0.266482	2.301442	0.284654
2	phon_R01_S01_2	122.4	148.65	…..	0.33559	2.486855	0.368674
3	phon_R01_S01_3	116.682	131.111	…..	0.311173	2.342259	0.332634
4	phon_R01_S01_4	116.676	137.871	…..	0.334147	2.405554	0.368975
…	…...	…..	…..	…..	…..	…..	…..
193	phon_R01_S50_3	209.516	253.017	…..	0.129303	2.784312	0.168895
194	phon_R01_S50_4	174.688	240.005	…..	0.158453	2.679772	0.131728
195	phon_R01_S50_5	198.764	396.961	…..	0.207454	2.138608	0.123306
196	phon_R01_S50_6	214.289	260.277	…..	0.190667	2.555477	0.148569

b. Describe Data

Untuk memahami karakteristik awal dari data yang digunakan, dilakukan eksplorasi terhadap tipe data, distribusi nilai, serta nilai statistik deskriptif dari masing-masing atribut. Langkah ini penting untuk memperoleh gambaran umum mengenai pola dan kecenderungan data sebelum dilakukan pemodelan lebih lanjut.

Tabel 2. Deskripsi Data

No	Dataset Parkinson's		
	Nama Fitur	**Tipe Data**	**Keterangan**
1	Name	Kategorikal	Identitas atau nama unik tiap sampel individu.
2	MDVP:Fo(Hz)	Numerik	Frekuensi dasar suara (*pitch*); mencerminkan getaran pita suara.
3	MDVP:Fhi(Hz)	Numerik	Frekuensi tertinggi dari suara pasien.
4	MDVP:Flo(Hz)	Numerik	Frekuensi terendah dari suara pasien.
5	MDVP:Jitter(%)	Numerik	Variasi relatif dari frekuensi dasar (ketidakteraturan suara).
6	MDVP:Jitter(Abs)	Numerik	Variasi absolut dari frekuensi dasar.
7	MDVP:RAP	Numerik	Perturbasi rata-rata relatif frekuensi (*Refined Jitter measurement*).
8	MDVP:PPQ	Numerik	*Jitter* kuantisasi lima titik.
9	Jitter:DDP	Numerik	Derivatif dari RAP, mencerminkan ketidakstabilan frekuensi.
10	MDVP:Shimmer	Numerik	Variasi amplitudo suara (ketidakteraturan dalam kekuatan suara).
11	MDVP:Shimmer(dB)	Numerik	*Shimmer* dalam satuan decibel
12	Shimmer:APQ3	Numerik	*Amplitude Perturbation Quotient* dalam 3 periode.
13	Shimmer:APQ5	Numerik	*Amplitude Perturbation Quotient* dalam 5 periode.
14	MDVP:APQ	Numerik	Rata-rata variasi amplitudo dalam jangka waktu pendek.
15	Shimmer:DDA	Numerik	Derivatif dari APQ, menilai fluktuasi amplitudo
16	NHR	Numerik	*Noise-to-Harmonics Ratio*, perbandingan noise terhadap suara utama.
17	HNR	Numerik	*Harmonics-to-Noise Ratio*, kebalikan dari NHR.
18	RPDE	Numerik	*Recurrence Period Density Entropy*; mengukur kekacauan sistem suara.
19	DFA	Numerik	*Detrended Fluctuation Analysis*; mengukur kompleksitas sinyal suara.
20	Spread1	Numerik	Nilai distribusi pertama dari transformasi frekuensi suara.
21	Spread2	Numerik	Nilai distribusi kedua dari transformasi frekuensi suara.
22	D2	Numerik	Dimensi korelasi; menggambarkan kompleksitas sinyal suara.

No	Dataset Parkinson's		
	Nama Fitur	**Tipe Data**	**Keterangan**
23	PPE	Numerik	*Pitch Period Entropy*; mengukur ketidakteraturan dalam periode *pitch*.
24	Status	Kategorikal	Label target: 1 = pasien *Parkinson*, 0 = sehat.

c. Explore Data

Pada tahap ini dilakukan eksplorasi awal terhadap dataset untuk memahami karakteristik data, distribusi fitur, serta potensi hubungan antar variabel. Dataset yang digunakan berasal dari file *parkinsons data.csv*, berisi 195 sampel dan 24 kolom, termasuk label kelas bernama status. Setelah kolom *name* dihapus karena tidak relevan sebagai fitur prediktif, tersisa 23 kolom numerik yang menjadi input model machine learning.

Tabel 3. Eksplorasi Data

NO	Dataset Parkinson's				
	Nama Fitur	**Mean**	**Standar Deviasi**	**Minimum**	**Maximum**
1	MDVP:Fo(Hz)	154.228641	41.390065	88.333000	260.105000
2	MDVP:Fhi(Hz)	197.104918	91.491548	102.145000	592.030000
3	MDVP:Flo(Hz)	116.324631	43.521413	65.476000	239.170000
4	MDVP:Jitter(%)	0.006220	0.004848	0.001680	0.033160
5	MDVP:Jitter (Abs)	0.000044	0.000035	0.000007	0.000260
6	MDVP:RAP	0.003306	0.002968	0.000680	0.021440
7	MDVP:PPQ	0.003446	0.002759	0.000920	0.019580
8	Jitter:DDP	0.009920	0.008903	0.002040	0.064330
9	MDVP:Shimmer	0.029709	0.018857	0.009540	0.119080
10	MDVP:Shimmer(dB)	0.282251	0.194877	0.085000	1.302000
11	Shimmer:APQ3	0.015664	0.010153	0.004550	0.056470
12	Shimmer:APQ5	0.017878	0.012024	0.005700	0.079400
13	MDVP:APQ	0.024081	0.016947	0.007190	0.137780
14	Shimmer:DDA	0.046993	0.030459	0.013640	0.169420
15	NHR	0.024847	0.040418	0.000650	0.314820
16	HNR	21.885974	4.425764	8.441000	33.047000
17	RPDE	0.498536	0.103942	0.256570	0.685151
18	DFA	0.718099	0.055336	0.574282	0.825288
19	Spread1	-5.684397	1.090208	-7.964984	-2.434031
20	Spread2	0.226510	0.083406	0.006274	0.450493
21	D2	2.381826	0.382799	1.423287	3.671155

NO	Dataset Parkinson's				
	Nama Fitur	**Mean**	**Standar Deviasi**	**Minimum**	**Maximum**
22	PPE	0.206552	0.090119	0.044539	0.527367
23	Status	0.753846	0.431878	0.000000	1.000000

Eksplorasi data menunjukkan adanya variasi yang signifikan pada nilai tiap fitur, seperti *MDVP:Fhi(Hz)* yang memiliki rentang lebar hingga 592 Hz, serta *Jitter(%)* dan Shimmer yang mencerminkan ketidakteraturan suara. Variasi ini mengindikasikan potensi kuat dari fitur-fitur tersebut dalam membedakan suara pasien *Parkinson* dan individu sehat, sehingga relevan untuk proses klasifikasi selanjutnya.

C. Data Preparation

a. Penghapusan Fitur Tidak Relevan

Pada tahap awal, dilakukan penghapusan atribut yang tidak memiliki kontribusi signifikan terhadap proses klasifikasi. Dalam hal ini, kolom name dihapus dari dataset karena hanya berisi informasi identitas pasien dan tidak memiliki pengaruh terhadap hasil prediksi.

Tabel 4. Perbandingan Fitur Setelah Dihapus

No	Dataset Parkinson's	
	Sebelum	**Sesudah**
1	Name	MDVP:Fo(Hz)
2	MDVP:Fo(Hz)	MDVP:Fhi(Hz)
3	MDVP:Fhi(Hz)	MDVP:Flo(Hz)
4	MDVP:Flo(Hz)	MDVP:Jitter(%)
5	MDVP:Jitter(%)	MDVP:Jitter (Abs)
6	MDVP:Jitter(Abs)	MDVP:RAP
7	MDVP:RAP	MDVP:PPQ
8	MDVP:PPQ	Jitter:DDP
9	Jitter:DDP	MDVP:Shimmer
10	MDVP:Shimmer	MDVP:Shimmer(dB)
11	MDVP:Shimmer(dB)	Shimmer:APQ3
12	Shimmer:APQ3	Shimmer:APQ5
13	Shimmer:APQ5	MDVP:APQ
14	MDVP:APQ	Shimmer:DDA
15	Shimmer:DDA	NHR
16	NHR	HNR
17	HNR	RPDE
18	RPDE	DFA

No	Dataset Parkinson's	
	Sebelum	**Sesudah**
19	DFA	Spread1
20	Spread1	Spread2
21	Spread2	D2
22	D2	PPE
23	PPE	Status
24	Status	

b. Pemisahan Fitur dan Target

Dataset kemudian dipisahkan menjadi dua bagian utama, yaitu fitur (*independen*) dan target (*dependen*). Fitur terdiri dari sejumlah variabel numerik yang merepresentasikan ciri akustik suara pasien, sedangkan target adalah kolom status yang menunjukkan kondisi pasien (1 untuk *Parkinson*, 0 untuk sehat). Pemisahan ini diperlukan agar algoritma klasifikasi dapat mempelajari hubungan antara fitur dan label keluaran.

c. Membangun Data Pelatihan dan Pengujian

Setelah pemisahan fitur dan target dilakukan, dataset dibagi menjadi data pelatihan (*training set*) dan data pengujian (*testing set*). Pada penelitian ini, digunakan proporsi 70% data latih dan 30% data uji. Data pelatihan digunakan untuk melatih model klasifikasi, sedangkan data pengujian digunakan untuk mengevaluasi performa model terhadap data yang belum pernah dilihat sebelumnya. Pembagian dilakukan secara acak dengan parameter *random_state* untuk menjamin hasil yang konsisten.

Tabel 5. Spilt Dataset

Split Dataset			
Kelas	**Jumlah Data**	**Training (70%)**	**Testing (30%)**
Status = 1 (Parkinson)	147	103	44
Status = 0 (Sehat)	48	33	15
Total	195	136	59

D. Modeling

a. Inisiasi dan Pelatihan Model

Model *Random Forest* diinisialisasi dengan parameter *n_estimators=100*, *random_state=42*, serta *class_weight='balanced'* untuk menangani ketidakseimbangan kelas dalam data. Parameter ini memungkinkan model untuk memberi bobot lebih pada kelas minoritas agar proses pembelajaran tidak bias. Setelah inisialisasi, model dilatih menggunakan data latih (*training set*) yang telah dipersiapkan sebelumnya. Proses pelatihan bertujuan untuk membentuk pola hubungan antara fitur akustik dan status kesehatan pasien.

b. Prediksi Data Uji

Model yang telah dilatih kemudian digunakan untuk melakukan prediksi terhadap data uji (*testing set*). Proses ini bertujuan untuk mengevaluasi kemampuan model dalam mengenali pola pada data baru yang belum pernah dilihat sebelumnya. Output dari proses ini berupa prediksi status kesehatan pasien (*Parkinson* atau sehat) berdasarkan nilai fitur yang tersedia.

c. Output Model

Hasil keluaran model dievaluasi menggunakan beberapa metrik evaluasi, seperti *confusion matrix*, *classification report* (*precision, recall, f1-score*), serta *accuracy score*. Berdasarkan hasil yang diperoleh, model menunjukkan performa klasifikasi yang baik dalam membedakan pasien yang mengidap *Parkinson* dan yang tidak.

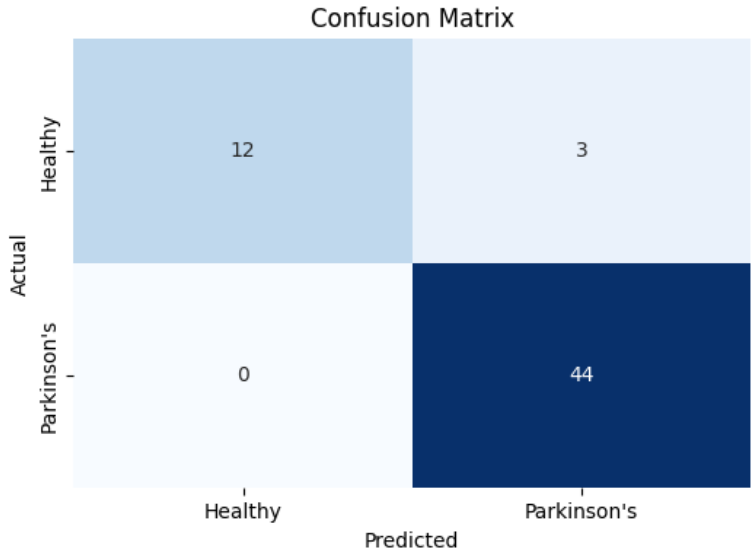
E. Evaluation

Pada tahap evaluasi dalam penelitian ini, *Confusion Matrix* digunakan sebagai salah satu metrik utama untuk mengevaluasi kinerja model klasifikasi biner yang dikembangkan berdasarkan algoritma *Random Forest* . *Confusion Matrix* memberikan representasi matriks yang menjelaskan jumlah prediksi benar (*true positive* dan *true negative*) serta prediksi salah (*false positive* dan *false negative*) dari model terhadap dataset pengujian. Dengan kata lain, matriks ini memungkinkan analisis lebih mendalam mengenai kemampuan model dalam membedakan antara dua kelas target, yaitu pasien dengan status sehat (kelas 0) dan penderita penyakit *Parkinson* (kelas 1).

Berdasarkan hasil evaluasi yang telah dilakukan, *Confusion Matrix* yang dihasilkan oleh model *Random Forest* pada data uji menunjukkan konfigurasi sebagai berikut:

a. Confusion Matrix

Confusion matrix digunakan untuk menunjukkan jumlah prediksi benar dan salah yang dilakukan oleh model. Berdasarkan hasil evaluasi, model menghasilkan matriks kebingungan sebagai berikut:



Gambar 2. *Confusion Matrix*

Interpretasi Komponen Matriks:

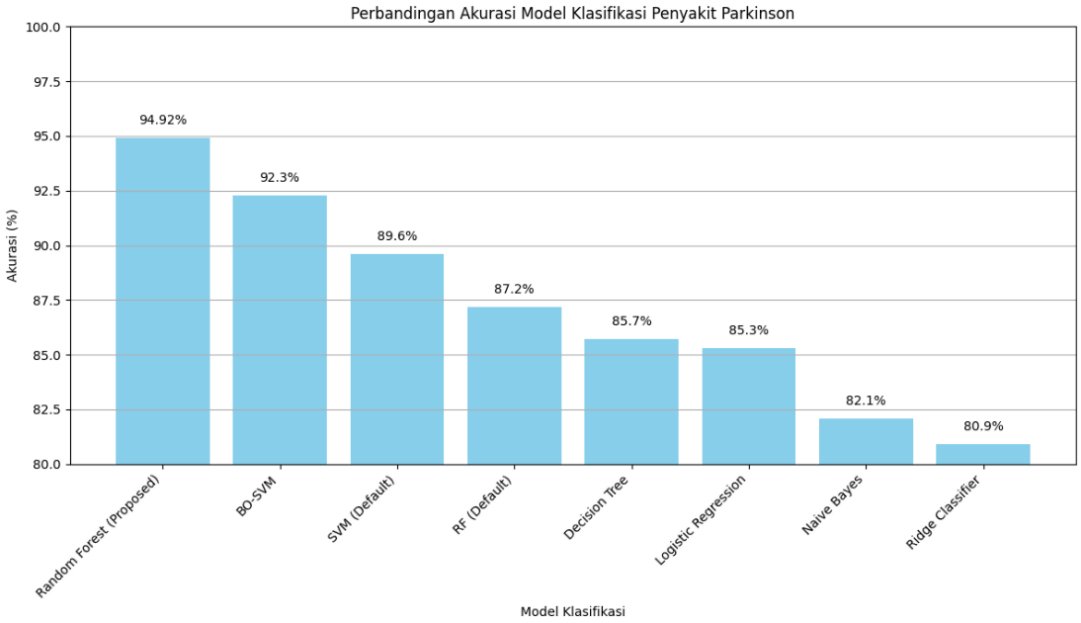
1. *True Negatives* (TN) = 12 : Terdapat 12 individu sehat yang diprediksi dengan benar oleh model sebagai sehat.
2. *False Positives* (FP) = 3 : Terdapat 3 individu sehat yang secara salah diprediksi oleh model sebagai penderita Parkinson.
3. *False Negatives* (FN) = 0 : Tidak ada individu penderita Parkinson yang secara salah diprediksi sebagai sehat oleh model.
4. *True Positives* (TP) = 44 : Semua individu penderita Parkinson berhasil diprediksi dengan benar oleh model.

Hasil *Confusion Matrix* menunjukkan bahwa model memiliki sensitivitas tinggi terhadap kelas positif (penderita *Parkinson*), dengan nilai *recall* atau *true positive rate* sebesar 100%. Hal ini menegaskan bahwa model tidak melakukan kesalahan jenis kedua (*Type II error*), sehingga sangat efektif dalam mengidentifikasi semua kasus nyata penderita penyakit *Parkinson*.

Sebaliknya, meskipun jumlah *False Positive* relatif kecil (3 dari 15 total kasus sehat), hal ini tetap menjadi area perhatian karena dapat menyebabkan diagnosis keliru pada individu sehat. Dalam konteks medis, kesalahan semacam ini dapat menimbulkan kecemasan berlebihan dan kebutuhan akan tes lanjutan yang tidak diperlukan.

b. Accuracy Score

Hasil akurasi ini menunjukkan bahwa model memiliki rasio kesalahan klasifikasi yang sangat rendah. Dari total 59 data uji, hanya terdapat 3 kasus salah klasifikasi (*False Positive*), sedangkan 56 sisanya berhasil diklasifikasikan secara akurat. Model menghasilkan tingkat akurasi sebesar 94.92%, menunjukkan bahwa model dapat mengklasifikasikan data dengan tingkat kesalahan yang sangat rendah. Tingkat akurasi ini termasuk tinggi dalam konteks data medis, yang umumnya sangat menuntut ketelitian.



Gambar 3. Hasil Perbandingan Akurasi

Jika dibandingkan dengan penelitian terdahulu, misalnya studi oleh Elshewey et al. (2023) yang menerapkan algoritma *Support Vector Machine* dengan optimasi *Bayesian* (BO-SVM), model tersebut menghasilkan akurasi sebesar 92,3%. Meskipun nilai ini cukup tinggi, model *Random Forest* dalam penelitian ini berhasil melampaui performa akurasi BO-SVM sebesar 2,62%. Keunggulan ini diperoleh tanpa proses tuning parameter yang kompleks seperti pada optimasi *Bayesian*, menjadikan *Random Forest* sebagai pilihan algoritma yang tidak hanya lebih sederhana dalam implementasi, tetapi juga lebih efektif dalam menghasilkan prediksi yang akurat.

Selain itu, hasil akurasi ini juga menunjukkan peningkatan dibandingkan metode klasifikasi tradisional lainnya seperti *Decision Tree* dan *Naive Bayes*, yang pada studi sebelumnya sering kali menghasilkan akurasi di bawah 90% pada dataset yang sama. Hal ini menunjukkan bahwa penggunaan *Random Forest* dengan penyesuaian bobot kelas (*class_weight='balanced'*) mampu mengatasi ketidakseimbangan data dan memberikan kontribusi signifikan terhadap peningkatan performa model.

c. Macro dan Weighted Average

Nilai rata-rata makro (*macro avg*) dan tertimbang (*weighted avg*) untuk *precision*, *recall*, dan *f1-score* juga berada di atas 0.90. Hal ini memperkuat kesimpulan bahwa model bekerja dengan baik secara keseluruhan, baik dalam menangani ketidakseimbangan kelas maupun dalam memberikan hasil prediksi yang stabil.

F. Deployment

Model yang dikembangkan memiliki prospek untuk diimplementasikan sebagai bagian dari sistem deteksi dini penyakit *Parkinson* berbasis analisis suara. Melalui integrasi antara perangkat perekam suara dan algoritma klasifikasi, sistem ini mampu memberikan hasil skrining awal secara cepat kepada individu, bahkan sebelum dilakukan pemeriksaan langsung oleh tenaga medis. Lebih lanjut, dengan penambahan data baru dan pelatihan ulang model secara berkala, sistem ini dapat diadaptasi untuk mencakup populasi yang lebih luas dan heterogen. Struktur pengembangan yang modular juga mendukung pembaruan model secara dinamis, sehingga kinerja dan akurasinya dapat terus ditingkatkan seiring waktu dan tetap relevan terhadap perkembangan data terbaru.

SIMPULAN

Penelitian ini menunjukkan bahwa penerapan algoritma *Random Forest* dalam mendeteksi dini penyakit *Parkinson* berdasarkan data frekuensi suara mampu menghasilkan performa klasifikasi yang sangat baik. Model yang dikembangkan berhasil mencapai tingkat akurasi sebesar 94,92%, dengan nilai *precision, recall*, dan *F1-score* yang tinggi dan seimbang. Evaluasi melalui *Confusion Matrix* menunjukkan bahwa model tidak menghasilkan kesalahan klasifikasi tipe II (*False Negative*), sehingga sangat ideal untuk diterapkan dalam konteks medis yang menuntut sensitivitas tinggi.

Keunggulan model ini semakin diperkuat dengan hasil perbandingan terhadap penelitian terdahulu. Studi oleh Elshewey et al. (2023) yang menggunakan model SVM dengan optimasi *Bayesian* hanya mencapai akurasi sebesar 92,3%, sedangkan model pada penelitian ini berhasil melampaui capaian tersebut dengan pendekatan yang lebih sederhana dan efisien. Penggunaan parameter *class_weight='balanced'* juga terbukti efektif dalam mengatasi ketidakseimbangan kelas, sehingga model dapat mengenali pola dari kedua kelas secara adil.

Dengan struktur sistem yang modular dan performa yang tinggi, model *Random Forest* ini memiliki potensi besar untuk diimplementasikan sebagai sistem deteksi dini berbasis suara. Integrasi dengan perangkat perekam suara dan pembaruan model secara berkala memungkinkan sistem tetap adaptif terhadap perubahan populasi data, menjadikannya alat skrining non-invasif yang efisien, cepat, dan dapat diandalkan, khususnya di wilayah dengan akses terbatas terhadap fasilitas kesehatan spesialis.

REFERENSI

[1] A. Iskandar, "Sistem Pakar Dalam Mendiagnosa Penyakit Parkinson Menerapkan Metode Dempster-Shafer," *Journal of Information System Research (JOSH)*, vol. 4, no. 3, pp. 847–854, Apr. 2023, doi: 10.47065/josh.v4i3.3320.

[2] S. Alia *et al.*, "Penyakit Parkinson: Tinjauan Tentang Salah Satu Penyakit Neurodegeneratif yang Paling Umum," *Jurnal Aksona*, vol. 1, no. 2, 2021.

[3] L. Judijanto, A. Amin, and L. Nurhakim, "Implementasi Teknologi Artificial Intelligence dan Machine Learning dalam Praktik Akuntansi dan Audit: Sebuah Revolusi atau Evolusi," *COSMOS: Jurnal Ilmu Pendidikan, Ekonomi dan Teknologi*, vol. 1, no. 6, pp. 3046–4846, 2024.

[4] K. A. Wulandari, A. Nugraha, A. Luthfiarta, and L. R. Nisa, "Peningkatan Akurasi Deteksi Dini Penyakit Parkinson melalui Pendekatan Ensemble Learning dan Seleksi Fitur Optimal," *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 2, pp. 575–584, Dec. 2024, doi: 10.29408/edumatic.v8i2.27788.

[5] Y. Sulistyo Nugroho and dan Nova Emiliyawati, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," *Jurnal Teknik Elektro*, 2017, [Online]. Available: http://archive.ics.uci.edu/ml/

[6] H. Tantyoko, D. Kartika Sari, and A. R. Wijaya, "PREDIKSI POTENSIAL GEMPA BUMI INDONESIA MENGGUNAKAN METODE RANDOM FOREST DAN FEATURE SELECTION," 2023. [Online]. Available:

http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/indexHenriTantyoko|http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index|

[7] A. Nugroho, A. Z. Fanani, and G. F. Shidik, "Evaluation of Feature Selection Using Wrapper For Numeric Dataset With Random Forest Algorithm," *International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2021, doi: 10.1109/iSemantic52711.2021.9573249.

[8] H. Akbar and W. K. Sanjaya, "Kajian Performa Metode Class Weight Random Forest pada Klasifikasi Imbalance Data Kelas Curah Hujan," *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, vol. 3, no. 1, Dec. 2023, doi: 10.20885/snati.v3i1.30.

[9] A. M. Elshewey, M. Y. Shams, N. El-Rashidy, A. M. Elhady, S. M. Shohieb, and Z. Tarek, "Bayesian Optimization with Support Vector Machine Model for Parkinson Disease Classification," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042085.

[10] D. Astuti, A. Rahmat Iskandar, and A. Febrianti, "Penentuan Strategi Promosi Usaha Mikro Kecil Dan Menengah (UMKM) Menggunakan Metode CRISP-DM dengan Algoritma K-Means Clustering," *Journal of Informatics, Information System, Software Engineering and Applications*, vol. 1, no. 2, pp. 60–072, 2019, doi: 10.20895/INISTA.V1I2.

[11] S. Sucipto, D. Dwi Prasetya, and T. Widiyaningtyas, "Educational Data Mining: Multiple Choice Question Classification in Vocational School," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 379–388, Mar. 2024, doi: 10.30812/matrik.v23i2.3499.

[12] N. Cholifah Sastya and D. I. Nugraha, "Penerapan Metode CRISP-DM dalam Menganalisis Data untuk Menentukan Customer Behavior di MeatSolution," *JURNAL PENDIDIKAN DAN APLIKASI INDUSTRI*, 2023, [Online]. Available: http://ejournal.unis.ac.id/index.php/UNISTEK

[13] A. P. Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," 2015.

[14] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[15] J. Brzozowska, J. Pizoń, G. Baytikenova, A. Gola, A. Zakimova, and K. Piotrowska, "DATA ENGINEERING IN CRISP-DM PROCESS PRODUCTION DATA – CASE STUDY," *Applied Computer Science*, vol. 19, no. 3, pp. 83–95, 2023, doi: 10.35784/acs-2023-26.

[16] Y. Sun, S. J. Chun, and Y. Lee, "Learned Semantic Index Structure Using Knowledge Graph Embedding and Density-Based Spatial Clustering Techniques," *Applied Sciences (Switzerland)*, vol. 12, no. 13, Jul. 2022, doi: 10.3390/app12136713.

[17] G. Airlangga, "Analysis of Machine Learning Classifiers for Speaker Identification: A Study on SVM, Random Forest, KNN, and Decision Tree," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 6, no. 1, pp. 430–438, Jan. 2024, doi: 10.47709/cnahpc.v6i1.3487.