Volume 1 No. 2 Juni 2026
E-ISSN: 3123-9218

Jurnal Kecerdasan Buatan

KLASIFIKASI DEMENSIA ALZHEIMER DENGAN *ENSEMBLE LEAKAGE-FREE* DAN INTERPRETASI *SHAP DATASET OASIS LONGITUDINAL*

Edwin Rudini[1], Moch Anwar Abas[2]
[1, 2] STMIK Pamitran, Jl. Bharata Raya Blok k, Sukaluyu Karawang
Email : *edwin.rudini@stmikpamitran.ac.id[1] anwar.pamitran@gmail.com[2]*

INFO ARTIKEL
Sejarah Artikel:
Diterima: 05-06-2026
Diperbaiki: 15-06-2026
Disetujui: 22-06-2026
Tersedia daring: 30-06-2026
E-ISSN: 3123-9218
DOI: –

Abstrak: D*emensia Alzheimer (DA) merupakan salah satu penyebab utama demensia sehingga diperlukan metode klasifikasi yang akurat dan dapat diinterpretasikan. Penelitian ini bertujuan menganalisis pengaruh strategi imputasi missing data terhadap performa ensemble machine learning (ML) pada dataset OASIS longitudinal. Dataset terdiri atas 373 observasi dari 150 subjek. Fitur label Group terdiri atas 190 kategori Nondemented dan 183 kategori Demented setelah kategori Converted digabungkan ke dalam kelas Demented agar menjadi klasifikasi biner. Fitur CDR dihapus untuk mengurangi risiko target leakage, sedangkan nilai hilang pada fitur SES dan MMSE ditangani menggunakan imputasi Median, KNN, dan Iterative Imputation. Model algoritma terdiri atas Random Forest (RF), AdaBoost Classifier (ADC), Gradient Boosting Classfier (GBC), dan XGBoost classifier (XGBC) yang dikombinasikan menggunakan weighted soft voting dengan optimasi bobot berbasis Bayesian Optimization melalui Optuna. Evaluasi dilakukan menggunakan 10-Fold Stratified Group Cross-Validation pada tingkat subjek untuk mencegah kebocoran data. Hasil menunjukkan bahwa imputasi KNN memberikan performa rata-rata terbaik dengan ROC-AUC 85,33%, sedangkan pada agregasi prediksi model algoritma diperoleh accuracy 76,14%, precision 80,13%, recall 68,31%, dan F1-score 73,75%. Uji Wilcoxon menunjukkan tidak terdapat perbedaan performa yang signifikan antar strategi imputasi. Analisis Shapley Additive explanations(SHAP) menunjukkan bahwa fitur MMSE merupakan fitur paling dominan dalam hasil prediksi model.*

Kata kunci :
Demensia Alzheimer, Imputation Strategy, Ensemble Learning, Leakage-Free, SHAP.

1. PENDAHULUAN

Demensia Alzheimer (DA) merupakan bentuk demensia yang paling umum dan ditandai oleh penurunan fungsi kognitif yang dapat mengganggu kemampuan individu dalam menjalankan aktivitas sehari-hari. Secara global, demensia menjadi salah satu masalah kesehatan yang semakin penting. Menurut data *(World Health Organization, 2026)*, sekitar 57 juta orang hidup dengan demensia di seluruh dunia, dengan lebih dari 60% di antaranya berada di negara berpendapatan rendah dan menengah. DA merupakan penyebab sekitar 60–70% dari seluruh kasus demensia. Peningkatan jumlah penderita tersebut menjadikan deteksi dan diagnosis DA secara lebih dini sebagai salah satu fokus

penting dalam penelitian. Pendekatan berbasis *machine learning* (*ML*) telah banyak dikembangkan untuk membantu mengidentifikasi pola yang berkaitan dengan DA melalui data klinis, kognitif, genetik, maupun *neuroimaging*. Kajian sistematis menunjukkan bahwa *ML* memiliki potensi dalam memprediksi risiko dan perkembangan DA, tetapi performanya masih dipengaruhi oleh keterbatasan ukuran sampel, jumlah prediktor, validasi model, dan konsistensi pelaporan kinerja (Rowe et al., 2021). Penerapan *ML* untuk deteksi dini DA juga telah menunjukkan potensi dalam mengombinasikan berbagai sumber data dan algoritma klasifikasi. Meskipun demikian, penulis menekankan bahwa aspek generalisasi dan interpretabilitas model masih menjadi tantangan penting sebelum penerapan ML dapat digunakan secara lebih luas dalam konteks klinis (Diogo et al., 2022).

Pengembangan model *ML* untuk DA juga menghadapi kendala dalam memperoleh data medis yang memadai. Data *neuroimaging* dan data klinis umumnya tersimpan secara terdesentralisasi pada institusi kesehatan serta dibatasi oleh aspek privasi dan keamanan data, sehingga akses terhadap *dataset* berlabel dalam jumlah besar menjadi terbatas. Kondisi tersebut dapat mempersulit pengembangan model yang stabil dan memiliki kemampuan generalisasi yang baik. Selain keterbatasan data, *overfitting* merupakan permasalahan penting dalam pengembangan model prediksi berbasis *ML*. Model dapat menunjukkan performa yang tinggi pada data pelatihan tetapi mengalami penurunan kinerja ketika digunakan pada data yang belum pernah dilihat sebelumnya. Tinjauan sistematis mengenai penerapan *ML* untuk prediksi DA menunjukkan bahwa ukuran sampel yang relatif kecil, ketidakseimbangan antara jumlah prediktor dan sampel, penggunaan satu sumber data, serta validasi yang kurang memadai dapat meningkatkan risiko *overfitting* dan menghasilkan estimasi performa yang terlalu optimistis (Viswan et al., 2024). Oleh karena itu, penggunaan strategi validasi yang tepat, seperti *cross-validation* dan *nested cross-validation* menjadi penting untuk memperoleh estimasi performa model yang lebih andal.

2. LANDASAN TEORI

2.1 Demensia *Alzheimer* (DA)

DA merupakan penyakit *neurodegeneratif progresif* yang terutama memengaruhi memori dan fungsi kognitif. Progresivitas tersebut menjadikan deteksi dan klasifikasi kondisi secara dini penting dalam pengelolaan penyakit (Liu et al., 2020).

2.2 Machine Learning *(ML)* untuk klasifikasi

ML telah digunakan untuk klasifikasi dan prediksi DA berbasis data klinis maupun *neuroimaging*. Potensinya tetap perlu diuji dengan validasi yang ketat agar estimasi kinerja dapat digeneralisasi (Ranson et al., 2023). Penelitian ini menggunakan ML untuk klasifikasi biner yaitu *Demented* dan *Nondemented* pada *dataset OASIS longitudinal*.

2.3 *Ensemble Learning*

Ensemble learning menggabungkan beberapa *classifier* untuk memperoleh prediksi yang lebih andal. Penelitian ini menggunakan algoritma RF, ADC, GBC, dan XGBC dalam *weighted soft voting*, yaitu penggabungan probabilitas kelas dengan bobot yang dioptimalkan (Rojarath and Songpan, n.d.).

a. XGBoost Classifier (XGBC)

XGBC merupakan metode *ensemble learning* berbasis *boosting* yang membangun model secara bertahap untuk memperbaiki kesalahan prediksi model sebelumnya (Mushtaq et al., 2022). Melalui *residual learning* dan pemrosesan paralel, XGBC meningkatkan efisiensi pembelajaran dengan menggabungkan model-model secara *iteratif* untuk meminimalkan kesalahan prediksi. Formulasi matematis XGBC dijelaskan pada bagian berikut.

$$J = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

b. Random Forest (RF)

RF merupakan algoritma *supervised learning* yang populer, fleksibel, dan mudah digunakan serta mampu menghasilkan kinerja prediksi yang baik (Song et al., 2022). RF membangun sejumlah pohon keputusan dengan melakukan pemisahan pada setiap *node* untuk menentukan kelas target pada *leaf node*. Proses pembentukan model dilakukan dengan memilih pemisahan optimal berdasarkan karakteristik data pada setiap *node*. Algoritma RF dapat dirumuskan sebagai berikut:

$$Gini = N_L \sum_{k} Pk_L \ (1 - Pk_L) + N_R \sum_{k} Pk_R \ (1 - Pk_R)$$

c. AdaBoost Classifier (ADC)

ADC merupakan algoritma *supervised learning* yang dapat digunakan untuk tugas klasifikasi maupun regresi (Yongcharoenchaiyasit et al., 2023). Algoritma ini

menggabungkan beberapa *weak learners* secara bertahap menjadi model yang lebih kuat dengan memberikan perhatian lebih besar pada kesalahan prediksi dari model sebelumnya. Proses pembentukan model ADC dapat dirumuskan sebagai berikut:

$$F_T(\pi) = \sum_{t=1}^{T} \alpha t \quad f t \quad (\pi)$$

d. Gradient Boosting Classifier (GBC)

GBC merupakan metode *ensemble learning* yang menggabungkan beberapa *weak learners*, umumnya berupa *decision tree*, untuk menghasilkan model dengan kinerja prediksi yang lebih baik (Thien and Yeo, 2022). Model dibangun secara bertahap dengan setiap *learner* baru berfungsi mengurangi kesalahan prediksi pada iterasi sebelumnya. Proses *boosting* tersebut mengoptimalkan kinerja model melalui tiga komponen utama, yaitu *loss function*, *weak learner*, dan *additive model*. Formulasi matematis GBC dijelaskan sebagai berikut:

$$F(x) = \sum_{m=0}^{M} f_m(x)$$

e. *Nested Cross-Validation*

Algoritma dasar yang digunakan memiliki karakteristik komplementer. Pemodelan *XGBC* dan *GBC* membangun model secara bertahap dengan memperbaiki kesalahan prediksi sebelumnya, sedangkan RF mengagregasi sejumlah pohon keputusan melalui proses *bootstrap* dan pemilihan *subset* fitur. Model ADC meningkatkan kontribusi *learner* yang sebelumnya kurang tepat sehingga membentuk model *boosting* secara iteratif. Keempat algoritma dipilih karena mewakili pendekatan *ensemble* berbasis *bagging* dan *boosting*, lalu probabilitas prediksinya digabungkan melalui *weighted soft voting*. Digunakan untuk memperoleh evaluasi model yang lebih *reliabel* sekaligus mengurangi risiko kebocoran informasi. Pendekatan *nested cross-validation* memisahkan proses optimasi dari evaluasi model sehingga estimasi kinerja menjadi lebih *robust* (Ghiasi et al., 2026).

f. Evaluasi

Menurut (Rainio et al., 2024), *accuracy* merupakan persentase seluruh *instance* yang diklasifikasikan dengan benar dari keseluruhan *instance* pada tugas klasifikasi *biner*. Akurasi dihitung menggunakan jumlah prediksi *True Positive* (TP) dan *True Negative* (TN)

dibandingkan dengan seluruh observasi, yaitu TP, TN, *False Positive* (*FP*), dan *False Negative (FN)*. Rumus yang digunakan adalah:

$$Accuracy = \frac{TP + TN}{Total\ Sample}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$\mathrm{W} = \min(W^+, W^-)$$

2.4 *Leakage-Free*

Data leakage terjadi ketika informasi yang seharusnya tidak tersedia saat prediksi masuk ke proses pelatihan atau evaluasi model. Pada *dataset longitudinal*, *subject information leakage* terjadi ketika data subjek yang sama terdapat pada data pelatihan dan pengujian sehingga model berpotensi mempelajari karakteristik spesifik subjek. (Ghiasi et al., 2026) menekankan pentingnya *group-aware* data *splitting* pada klasifikasi DA berbasis OASIS *longitudinal* serta menghindari fitur *CDR* yang berpotensi menimbulkan target *leakage*. Dalam penelitian ini, pendekatan *leakage-free* diterapkan dengan memisahkan subjek secara konsisten antara data pelatihan dan pengujian serta memastikan *preprocessing* tidak menggunakan informasi dari data pengujian sehingga diperoleh estimasi kinerja model yang lebih realistis, yang dinyatakan dengan persamaan:

$$S_{train} \cap S_{test} = \emptyset$$

2.5 Explainable Artificial Intelligence dan SHAP

Ensemble learning, dapat menghasilkan prediksi yang sulit dijelaskan. *Explainable Artificial Intelligence (XAI)* digunakan untuk mengidentifikasi faktor yang memengaruhi keputusan model. Salah satu metode yang banyak digunakan adalah *Shapley Additive Explanations (SHAP)*, yang menghitung kontribusi setiap fitur terhadap prediksi

berdasarkan konsep *Shapley value*. *SHAP* dapat menjelaskan pentingnya fitur secara global maupun individual dan telah diterapkan dalam prediksi DA (Dong et al., 2022). Hal ini dapat dijabarkan menggunakan persamaan:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

2.6. Bayesian Optimization with Optuna

Bayesian Optimization memilih konfigurasi berikutnya berdasarkan hasil evaluasi sebelumnya. Optuna dengan *sampler* TPE digunakan untuk mengoptimalkan bobot model RF, ADC, GBC, dan XGBC berdasarkan *ROC-AUC* pada *inner cross-validation* (Watanabe and Hutter, 2023). Dalam penelitian ini, *recall* kelas *Demented* menjadi penting karena *false negative* dapat menunjukkan subjek yang sebenarnya *Demented* tetapi diprediksi *Nondemented*.

3. METODE PENELITIAN

3.1 Desain Penelitian

Dataset OASIS longitudinal terdiri atas 373 catatan pemeriksaan dari 150 subjek dan 15 fitur. Kategori *Converted* digabungkan ke *Demented* diperoleh distribusi 190 observasi pada kelas *Nondemented* (50,94%) dan 183 observasi pada kelas *Demented* (49,06%).

Table 3.1 Distribusi final fitur Group

Group	Jumlah	Persentase
Nondemented	190	50.94 %
Demented	183	49.06 %

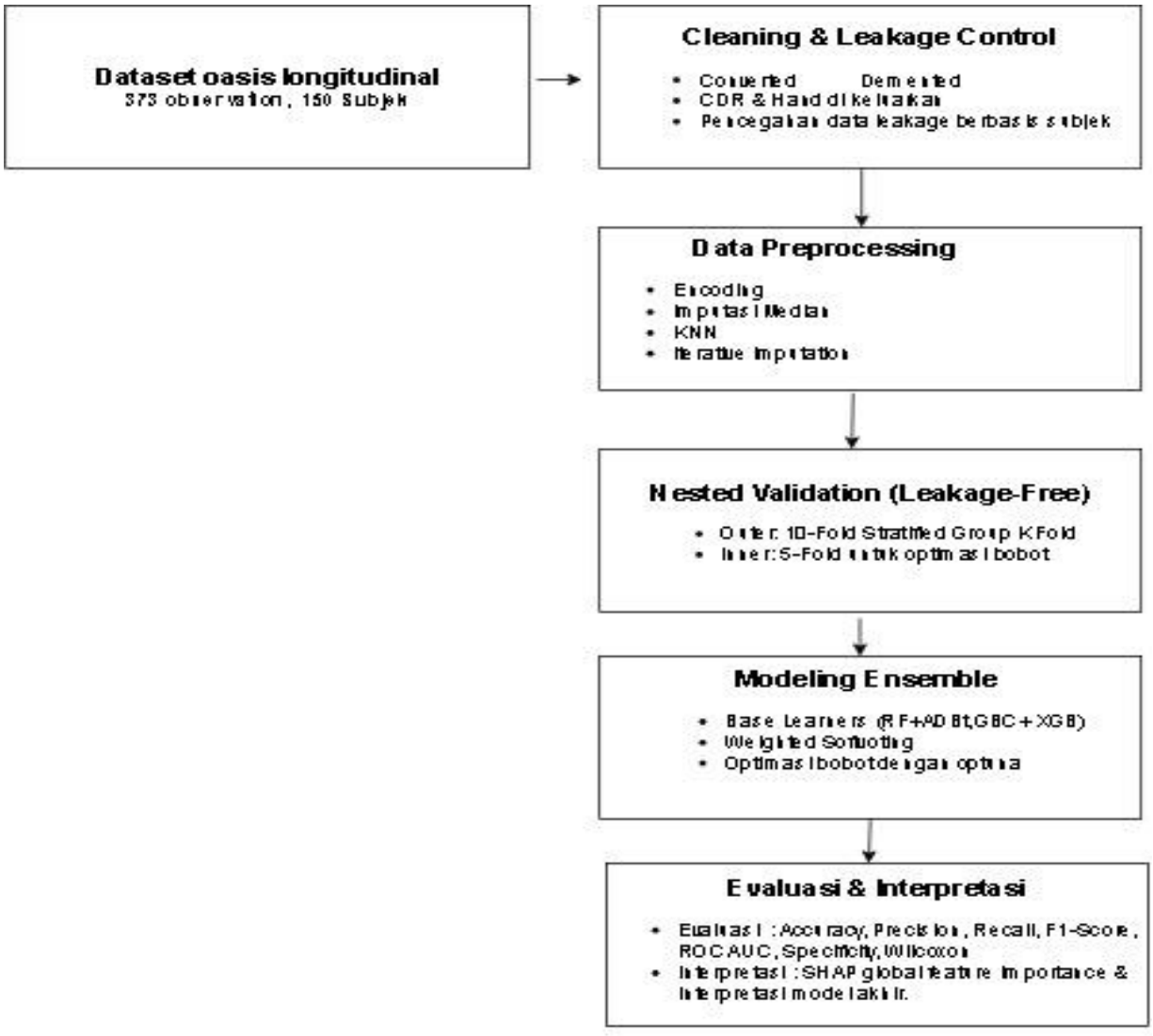
Pada fitur *Group* menunjukkan 190 observasi pada kelas *Nondemented* dan 183 observasi pada kelas *Demented*. Dari distribusi kelas tersebut menunjukkan bahwa *dataset oasis lonitudinal* memiliki proporsi kelas yang relatif seimbang. *Missing value* ditemukan pada fitur SES sebanyak 19 nilai (5,09%) dan pada fitur MMSE sebanyak 2 nilai (0,54%), sehingga terdapat 21 nilai *missing* secara keseluruhan (5,63% dari 373 observasi). Baris tidak dihapus agar jumlah observasi dan struktur *longitudinal* tetap terjaga.

Tabel 3.2 *Missing value* pada *dataset oasis longitudinal*

Fitur	Missing Value	Persentase
SES	19	5,09%
MMSE	2	0,54%

Total	21	5,63%

Fitur CDR dihapus untuk mengurangi risiko target *leakage*, sedangkan fitur *Hand* dihapus karena seluruh subjek dominan *right handed*. Group dikodekan 0 sama dengan *Nondemented* dan 1 sama dengan *Demented*. Imputasi dilakukan hanya pada data *training* setiap *fold*. Evaluasi menggunakan *nested cross-validation* berbasis subjek. *Outer Stratified Group KFold* 10-*fold* dan *inner* Group KFold 5-*fold*. Model *RF*, *ADC*, *GBC*, dan *XGBC* digabungkan melalui *weighted soft voting* dengan bobot dioptimalkan melalui optuna TPE. Metrik evaluasi meliputi *accuracy*, *precision*, *recall*, *F1-score*, *ROC-AUC*, dan specificity. Perbedaan imputasi diuji dengan *Wilcoxon* berdasarkan *ROC-AUC outer folds*. *SHAP* digunakan pada tahap akhir untuk menginterpretasikan kontribusi fitur terhadap prediksi model.



Gambar 3.1 Alur penelitian

Tabel 3.3 menyajikan informasi fitur *dataset oasis longitudinal* yang digunakan dalam penelitian. Fitur mencakup karakteristik demografis, hasil pemeriksaan kognitif, serta karakteristik struktural otak dari pemeriksaan MRI. *Subject ID* dan *MRI ID* digunakan sebagai *identifier* dan tidak menjadi prediktor, sedangkan fitur *CDR* dikeluarkan untuk mengurangi risiko target *leakage*. Fitur yang digunakan sebagai prediktor meliputi *M/F, Age, EDUC, SES, MMSE, eTIV, nWBV, ASF, Visit, dan MR Delay*. Dibawah ini penjelasan singkat mengenai setiap fitur *dataset oasis longitudinal*

Table 3.3 *Oasis longitudinal feature dataset*

No	Features	Description
1	*M/F*	Gender (Male, Female)
2	*Age*	Subject's age at the time of examination
3	*EDUC*	Years of education (Education Level) subject
4	*SES*	Socioeconomic status
5	*MMSE*	Mini mental State examination (mental state examination value)
6	*eTIV*	Estimated Total Intracranial Volume
7	*nWBV*	Normalized Whole Brain Volume
8	*ASF*	Atlas Scaling Factor
9	*Visit*	Number of subject examinations
10	*MR Delay*	The value of the subject's delay since the last examination

Empat model dasar menggunakan konfigurasi yang sama pada seluruh eksperimen untuk menjaga keterbandingan, kemudian digabungkan dalam *weighted soft voting*.

Tabel 3.4 Konfigurasi Model *Base Learner* dalam penelitian

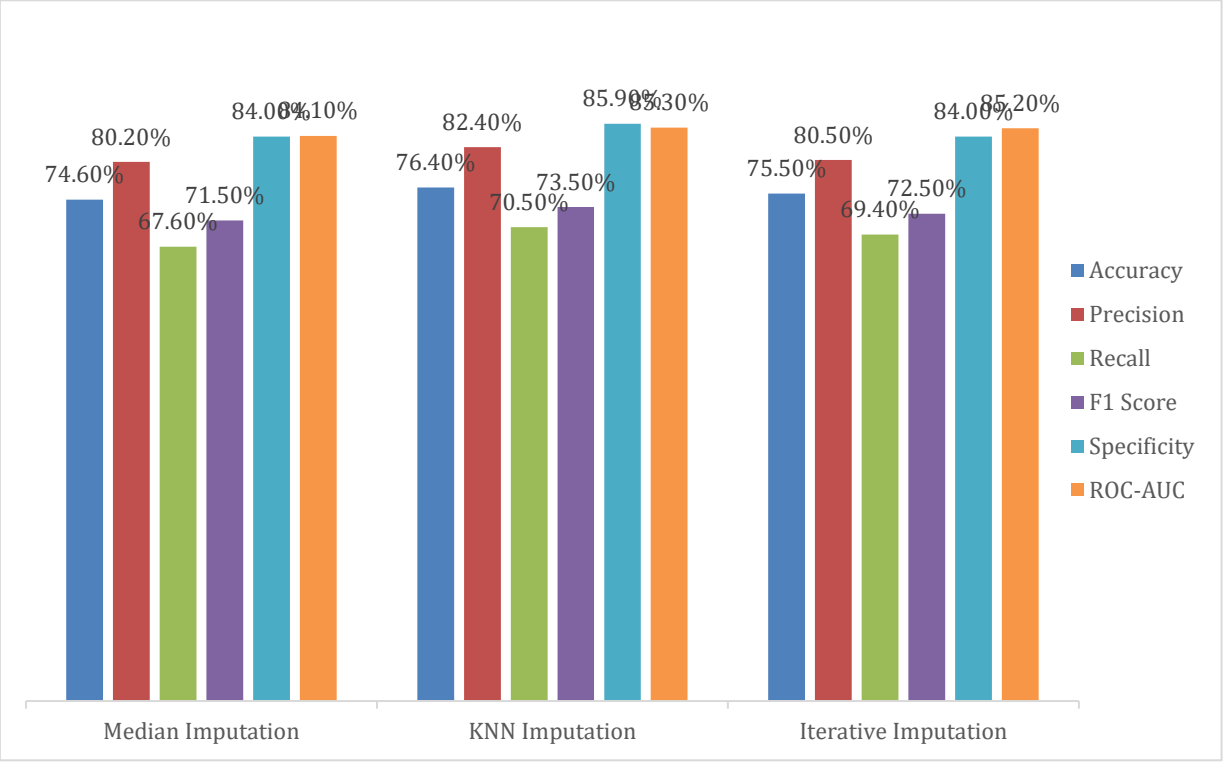
Model Algoritma	Konfigurasi utama
Random Forest (RF)	300 trees; max_features=sqrt
AdaBoost Classifier (ADC)	200 estimators; learning rate=0,05
Gradient Boosting Classifier (GBC)	200 estimators; learning rate=0,05; depth=3
XGBoost Classifier (XGBC)	300 estimators; learning rate=0,05; depth=3; subsample=0,8; colsample=0,8

4. HASIL DAN PEMBAHASAN

4.1 Hasil Perbandingan Strategi Imputasi

KNN *Imputation* memberikan kinerja deskriptif tertinggi pada seluruh metrik evaluasi, dengan *accuracy* 76,40%, *precision* 82,40%, *recall* 70,50%, *F1-score* 73,50%, *specificity*

85,90%, dan *ROC-AUC* 85,30%. Median menghasilkan kinerja yang lebih rendah pada seluruh metrik, sedangkan *iterative* menunjukkan hasil yang mendekati KNN, khususnya pada ROC-AUC (85,20% dibandingkan 85,30%). Dengan demikian, KNN merupakan strategi dengan performa deskriptif terbaik, meskipun selisih kinerjanya terhadap *iterative* relatif kecil.



Gambar 4.1 Perbandingan performa klasifikasi berdasarkan strategi imputasi

4.2 Evaluasi Kinerja pada Outer Cross-Validation

Pada *outer cross-validation*, KNN memperoleh *accuracy* 0,7638±0,0776; *precision* 0,8240±0,1799; recall 0,7045±0,1370; F1 0,7352±0,0936; ROC-AUC 0,8533±0,0833; specificity 0,8586±0,1432. *Iterative* dan Median masing-masing memperoleh *ROC-AUC* 0,8516±0,0787 dan 0,8405±0,0880. Imputasi KNN tertinggi pada seluruh metrik, tetapi variasi antar-fold tetap cukup besar.

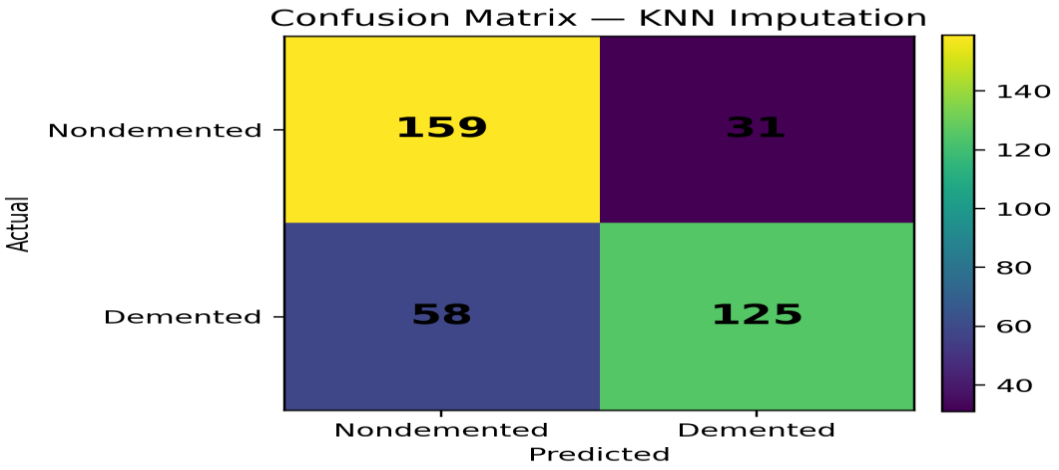
Tabel 4.1 Kinerja rata-rata ± simpangan baku pada 10-fold outer cross-validation

Imputasi	Accuracy	Precision	Recall	F1	ROC-AUC	Specificity
Median	0,7464 ± 0,0869	0,8021 ± 0,1708	0,6759 ± 0,1210	0,7151 ± 0,0982	0,8405 ± 0,0880	0,8403 ± 0,1333
KNN	0,7638 ± 0,0776	0,8240 ± 0,1799	0,7045 ± 0,1370	0,7352 ± 0,0936	0,8533 ± 0,0833	0,8586 ± 0,1432

Iterative	0,7547 ± 0,0917		0,8046 ± 0,1645		0,6944 ± 0,1498		0,7245 ± 0,1054		0,8516 ± 0,0787		0,8403 ± 0,1333	

4.4 Confusion Matrix dan Kinerja Klasifikasi

Klasifikasi prediksi *outer cross-validation* menghasilkan TN=159, FP=31, FN=58, dan TP=125, dengan *accuracy* 76,14%, *precision* 80,13%, *recall* 68,31%, dan *F1-score* 73,75%. *Recall Nondemented* 83,68% lebih tinggi daripada *Demented* 68,31%, sehingga *false negative* pada kelas *Demented* masih menjadi perhatian.



Gambar 4.2 Confusion matrix model dengan imputasi KNN pada seluruh prediksi *outer cross-validation*

4.3 Uji Statistik Perbandingan Strategi Imputasi

Metode pengujian *Wilcoxon* pada *ROC-AUC outer folds* menghasilkan p=0,16406 (Median–KNN), p=0,42578 (Median–*Iterative*), dan p=0,94531 (KNN–Iterative). Seluruh p>0,05, sehingga tidak ada perbedaan signifikan. KNN dipilih karena mean ROC-AUC tertinggi, bukan karena keunggulan statistik.

Tabel 4.2 Hasil Wilcoxon signed-rank test terhadap ROC-AUC per fold

Perbandingan	Statistik Wilcoxon	p-value	Keputusan
Median vs KNN	10,0	0,16406	Tidak signifikan
Median vs Iterative	15,0	0,42578	Tidak signifikan
KNN vs Iterative	17,0	0,94531	Tidak signifikan

.

4.4 Bobot *Weighted Soft Voting*

Optuna menghasilkan bobot model interpretasi akhir: *RF* 0,0855 (8,55%), *ADC* 0,8298 (82,98%), *GBC* 0,0025 (0,25%), dan *XGBC* 0,0822 (8,22%). Model ADC menjadi kontributor probabilistik terbesar dalam *ensemble* akhir. Bobot tersebut hanya merupakan konfigurasi model interpretasi akhir, bukan ukuran kepentingan model yang stabil atau universal. Stabilitas bobot sebaiknya diuji pada seluruh outer folds.

Tabel 4.3 Bobot Model pada Weighted Soft Voting Ensemble

Model	**Bobot**
Random Forest (RF)	0,0855 (8,55%)
AdaBoost Classifier (ADC)	0,8298 (82,98%)
Gradient Boosting Classifier (GBC)	0,0025 (0,25%)
XGBoost Classifier (XGBC)	0,0822 (8,22%)

4.5 Interpretasi Global dengan SHAP

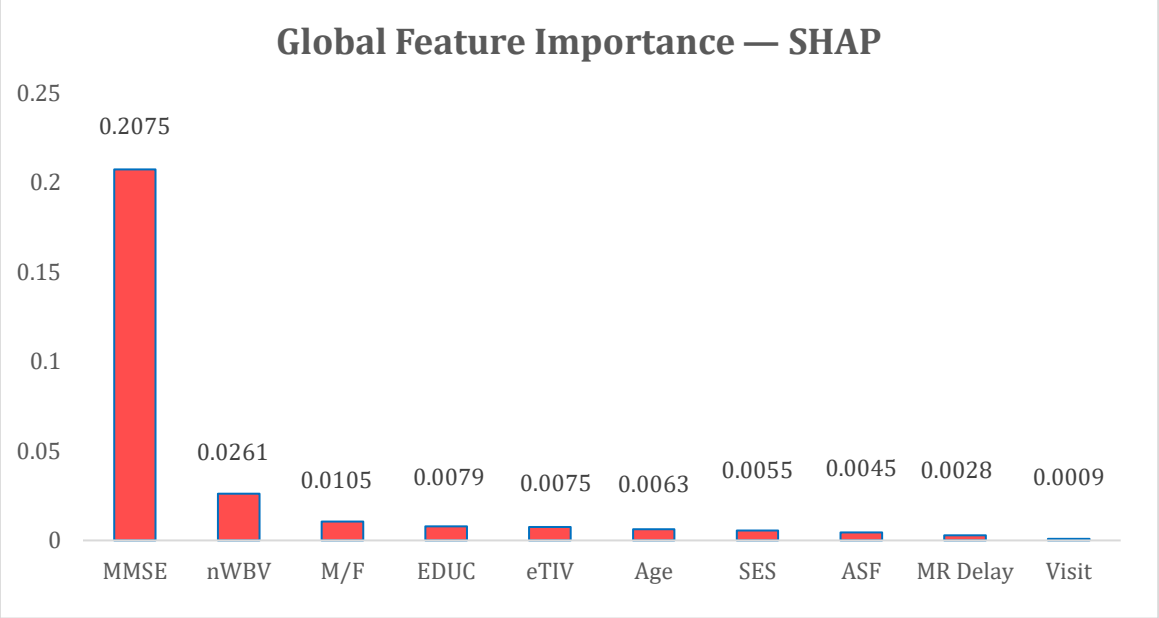
Berdasarkan tabel 4.4 dapat disimpulkan bahwa pengaruh fitur terhadap prediksi model sangat tidak merata. Nilai yang ditampilkan merupakan mean absolute SHAP value, sehingga semakin besar nilainya, semakin besar kontribusi fitur terhadap prediksi model secara global.

Tabel 4.4 Fitur paling berpengaruh terhadap prediksi model

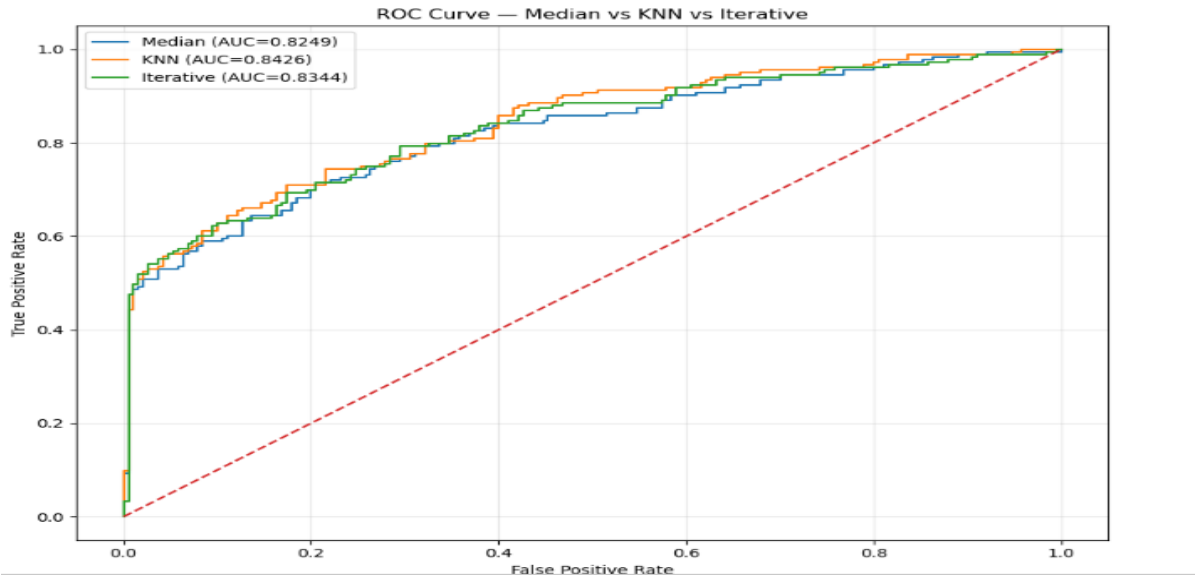
Feature	**Mean_Absolute_SHAP**	**Contribution_%**
MMSE	0.2075	74.2370
nWBV	0.0261	9.3255
M/F	0.0105	3.7610
EDUC	0.0079	2.8373
eTIV	0.0075	2.6976
Age	0.0063	2.2575
SES	0.0055	1.9737
ASF	0.0045	1.5924

Feature	Mean_Absolute_SHAP	Contribution_%
MR Delay	0.0028	0.9917
Visit	0.0009	0.3263

Berdasarkan analisis *SHAP*, fitur *MMSE* merupakan fitur yang paling dominan dalam prediksi model dengan mean absolute SHAP value sebesar 0,2075. Nilai tersebut merepresentasikan sekitar 74,24% dari total magnitudo SHAP fitur yang dianalisis, menunjukkan bahwa MMSE memberikan kontribusi paling besar terhadap keputusan model dibandingkan fitur lainnya. Fitur dengan kontribusi berikutnya adalah nWBV dengan nilai SHAP sebesar 0,0261 atau sekitar 9,34%. Sementara itu, fitur M/F, EDUC, eTIV, Age, SES, ASF, MR Delay, dan Visit memberikan kontribusi yang relatif lebih kecil.



Gambar 4.3 Kontribusi global fitur berdasarkan mean absolute SHAP



Gambar 4.4. Perbandingan Kurva ROC pada Imputasi Median, KNN, dan Iterative Imputation

Gambar 4.4 menunjukkan perbandingan kurva ROC pada strategi Median, KNN, dan Iterative Imputation. KNN menghasilkan ROC-AUC tertinggi sebesar 0,8426, diikuti Iterative Imputation 0,8344 dan Median 0,8249. Seluruh nilai AUC di atas 0,80 menunjukkan kemampuan diskriminasi yang baik. Meskipun KNN memberikan performa terbaik, selisih terhadap Iterative dan Median relatif kecil, sehingga perbedaan kinerjanya tidak terlalu signifikan. Hasil ini menunjukkan bahwa strategi imputasi dapat memengaruhi kemampuan diskriminasi model, dengan KNN lebih unggul karena mempertimbangkan kemiripan karakteristik antarobservasi dalam mengestimasi nilai yang hilang.

4.7 Pembahasan dan Keterbatasan

Hasil menunjukkan variasi performa antarimputasi, tetapi perbedaannya kecil dan tidak signifikan. KNN memiliki mean ROC-AUC tertinggi, namun hanya unggul 0,0017 dari Iterative. Penghapusan CDR, pemisahan berbasis subjek, dan nested cross-validation menghasilkan estimasi yang lebih konservatif. Kontribusi penelitian diposisikan sebagai evaluasi komparatif imputasi dalam pipeline ensemble yang leakage-free.

Secara metodologis, seluruh strategi diuji dalam pipeline identik: pembagian berbasis subjek, imputasi hanya pada training fold, optimasi ensemble pada inner fold, dan evaluasi pada outer fold. Dengan demikian, perbedaan lebih dapat dikaitkan dengan strategi imputasi dan bukan kebocoran atau pembagian data. Keterbatasan meliputi jumlah subjek yang kecil (150), penggabungan Converted ke Demented, SHAP yang ditujukan untuk interpretasi

model akhir, bobot ensemble yang belum diringkas dari seluruh outer folds, serta belum adanya validasi eksternal. Penelitian lanjutan perlu menguji kohort independen dan mempertahankan aspek temporal secara eksplisit.

5. KESIMPULAN DAN SARAN

Pipeline leakage-free berhasil dibangun melalui penghapusan CDR, pemisahan berbasis subjek, dan nested cross-validation. KNN memberikan kinerja rata-rata tertinggi (ROC-AUC 0,8533; accuracy 0,7638; F1 0,7352), tetapi tidak berbeda signifikan dari strategi lain. Bobot terbesar pada model akhir diberikan kepada AdaBoost (82,98%), sedangkan SHAP menunjukkan dominasi MMSE (74,24%) dan nWBV (9,33%). Hasil menegaskan bahwa imputasi harus diperlakukan sebagai bagian dari desain eksperimen. Penelitian berikutnya perlu menggunakan validasi eksternal, menguji stabilitas bobot, dan mempertahankan informasi temporal.

DAFTAR PUSTAKA

Diogo, V.S., Ferreira, H.A., Prata, D., 2022. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimer's Research & Therapy* 14.

Dong, S., Khattak, A., Ullah, I., Zhou, J., Hussain, A., 2022. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley additive exPlanations. *International Journal of Environmental Research and Public Health* 19.

Ghiasi, M.M., Falck, R.S., Liu-Ambrose, T., Tam, R.C., 2026. Explainable machine learning for predicting longitudinal dementia status: Establishing a leakage-free benchmark. *PLOS Digital Health* 5.

Liu, L., Zhao, S., Chen, H., Wang, A., 2020. A new machine learning method for identifying Alzheimer's disease. *Simulation Modelling Practice and Theory* 99.

Mushtaq, Z., Ramzan, M.F., Ali, S., Baseer, S., Samad, A., Husnain, M., 2022. Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. *Mobile Information Systems* 2022.

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports* 14.

Ranson, J.M., Bucholc, M., Lyall, D., Newby, D., Winchester, L., Oxtoby, N.P., Veldsman, M., Rittman, T., Marzi, S., Skene, N., Al Khleifat, A., Foote, I.F., Orgeta,
V., Kormilitzin, A., Lourida, I., Llewellyn, D.J., 2023. Harnessing the potential of machine learning and artificial intelligence for dementia research. *Brain Informatics*.

Rojarath, A., Songpan, W., n.d. Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. *Applied Intelligence*.

Rowe, T.W., Katzourou, I.K., Stevenson-Hoare, J.O., Bracher-Smith, M.R., Ivanov, D.K., Escott-Price, V., 2021. Machine learning for the life-time risk prediction of Alzheimer's disease: a systematic review. *Brain Communications*.

Song, Y., Li, H., Xu, P., Liu, D., Cheng, S., 2022. A method of intrusion detection based on WOA-XGBoost algorithm. *Discrete Dynamics in Nature and Society* 2022.

Thien, T.F., Yeo, W.S., 2022. A comparative study between PCR, PLSR, and LW-PLS on the predictive performance at different data splitting ratios. *Chemical Engineering Communications* 209, 1439–1456.

Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., Hajamohideen, F., 2024. Explainable artificial intelligence in Alzheimer's disease classification: a systematic review. *Cognitive Computation*.

Watanabe, S., Hutter, F., 2023. c-TPE: Tree-Structured Parzen Estimator with Inequality Constraints for Expensive Hyperparameter Optimization. World Health Organization, 2026. Dementia.

Yongcharoenchaiyasit, K., Arwatchananukul, S., Temdee, P., Prasad, R., 2023
Gradient boosting based model for elderly heart failure, aortic stenosis, and dementia classification. *IEEE Access* 11, 48677–48696.