JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analisis Komparatif Prediksi Remaining Useful Life Mesin Turbofan Menggunakan XGBoost Dan Bi-LSTM Berbasis Explainable AI

Andri Laksono[1*], Majid Rahardi[2]

Fakultas Ilmu Komputer, Informatika, Universitas Amikom, Yogyakarta, Indonesia
Email: [1]andrilaksono@students.amikom.ac.id, [2]majid@amikom.ac.id,
Email Penulis Korespondensi: andrilaksono@students.amikom.ac.id
Submitted **04-06-2026**; Accepted **02-10-2026**; Published **31-10-2026**

Abstrak

Prediksi Remaining Useful Life (RUL) mesin turbofan merupakan permasalahan penting dalam pemeliharaan prediktif karena kompleksitas kondisi operasi dan pola degradasi dapat memengaruhi kinerja model secara berbeda. Selain itu, penelitian pada dataset C-MAPSS umumnya mengevaluasi subset secara terbatas dan belum banyak membandingkan model *tree-based* dan *deep learning* berbasis *attention* secara seragam serta menganalisis konsistensi interpretasi antar-metode Explainable Artificial Intelligence (XAI). Penelitian ini bertujuan membandingkan performa eXtreme Gradient Boosting (XGBoost) dan Bidirectional Long Short-Term Memory (Bi-LSTM) berbasis *attention* dalam memprediksi RUL mesin turbofan pada empat subset C-MAPSS serta menganalisis kontribusi sensor menggunakan XAI. Pendekatan meliputi pelabelan RUL *piecewise-linear* dengan ambang 125 siklus, normalisasi *condition-aware* untuk subset multi-kondisi FD002 dan FD004, serta segmentasi *sliding window* sepanjang 30 siklus untuk Bi-LSTM. Interpretasi model dilakukan menggunakan SHAP TreeExplainer untuk XGBoost dan Integrated Gradients untuk Bi-LSTM. Hasil menunjukkan bahwa Bi-LSTM mengungguli XGBoost pada tiga dari empat subset, yaitu FD001 dengan RMSE 15,35, FD002 dengan RMSE 27,92, dan FD003 dengan RMSE 17,33. Sebaliknya, XGBoost memberikan hasil lebih baik pada subset FD004 yang paling kompleks dengan RMSE 30,22 dibandingkan 33,40 pada Bi-LSTM. Analisis konsensus XAI menunjukkan kesepakatan 80% pada subset multi-kondisi dan 20% pada subset kondisi tunggal, sedangkan sensor Ps30, NRc, Nc, dan BPR teridentifikasi sebagai indikator degradasi yang konsisten. Kontribusi penelitian ini adalah menyediakan komparasi sistematis kedua paradigma model pada seluruh subset C-MAPSS serta analisis konsensus XAI untuk mengidentifikasi konsistensi kontribusi sensor terhadap prediksi RUL.

Kata kunci: Remaining Useful Life; C-MAPSS; XGBoost; Bi-LSTM; Explainable AI

Abstract

Predicting the Remaining Useful Life (RUL) of turbofan engines is an important problem in predictive maintenance because the complexity of operating conditions and degradation patterns can affect model performance differently. In addition, studies using the C-MAPSS dataset generally evaluate limited subsets and have rarely compared tree-based and attention-based deep learning models systematically while analyzing interpretation consistency across Explainable Artificial Intelligence (XAI) methods. This study aims to compare eXtreme Gradient Boosting (XGBoost) and attention-based Bidirectional Long Short-Term Memory (Bi-LSTM) for RUL prediction across four C-MAPSS subsets and to analyze sensor contributions using XAI. The approach includes piecewise-linear RUL labeling with a 125-cycle threshold, condition-aware normalization for multi-condition subsets FD002 and FD004, and 30-cycle sliding-window segmentation for Bi-LSTM. Model interpretation uses SHAP TreeExplainer for XGBoost and Integrated Gradients for Bi-LSTM. The results show that Bi-LSTM outperformed XGBoost on three of four subsets, namely FD001 with an RMSE of 15.35, FD002 with an RMSE of 27.92, and FD003 with an RMSE of 17.33. In contrast, XGBoost performed better on FD004, with an RMSE of 30.22 compared with 33.40 for Bi-LSTM. XAI consensus analysis showed 80% agreement on multi-condition subsets and 20% on single-condition subsets, while Ps30, NRc, Nc, and BPR were identified as consistent degradation indicators. The contribution of this study is a systematic comparison of both model paradigms across all C-MAPSS subsets and XAI consensus analysis to identify consistent sensor contributions to RUL prediction.

Keywords: Remaining Useful Life; C-MAPSS; XGBoost; Bi-LSTM; Explainable AI

1. PENDAHULUAN

Meskipun berbagai model machine learning dan deep learning telah diusulkan untuk prediksi RUL, permasalahan utama yang belum terpecahkan adalah minimnya evaluasi yang konsisten antar-model pada seluruh tingkat kompleksitas data operasional, serta kurangnya transparansi pengambilan keputusan model yang menjadi prasyarat kritis sebelum diadopsi dalam sistem pemeliharaan pesawat berisiko tinggi. Industri 4.0 menuntut transformasi paradigma pemeliharaan dari pendekatan reaktif dan terjadwal menuju pemeliharaan prediktif (*predictive maintenance*) yang berbasis kondisi aktual mesin. Pada sistem mesin kritis seperti mesin turbofan pesawat, kegagalan tak terduga dapat menyebabkan kerugian operasional yang besar bahkan risiko keselamatan jiwa, sehingga prediksi *Remaining Useful Life* (RUL) menjadi komponen kunci dalam kerangka *Prognostics and Health Management* (PHM) [1]. Program pemeliharaan prediktif berbasis prediksi RUL telah diterapkan pada mesin pesawat untuk meminimalkan *downtime* dan mengoptimalkan jadwal perawatan [2]. RUL didefinisikan sebagai durasi sisa pakai komponen sebelum mencapai ambang kegagalan, dan estimasi yang akurat memungkinkan perencana pemeliharaan menjadwalkan tindakan pencegahan secara optimal.

Dataset *Commercial Modular Aero-Propulsion System Simulation* (C-MAPSS) yang dirilis NASA telah menjadi *benchmark de facto* untuk evaluasi metode prediksi RUL [1]. Dataset ini terdiri dari empat subset (FD001-FD004) dengan

Copyright © 2026 The **Andri Laksono**, Page 1348
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

tingkat kompleksitas berbeda berdasarkan kombinasi kondisi operasi (1 atau 6 *operating conditions*) dan mode kerusakan (1 atau 2 *fault modes*), sebagaimana dirangkum pada Tabel 1.

Tabel 1. Karakteristik Empat Subset C-MAPSS

subset	Operating Conditions	Fault Modes	Mesin Train	Mesin test
FD001	1	1 (HPC)	100	100
FD002	6	1 (HPC)	260	259
FD003	1	2 (HPC, Fan)	100	100
FD004	6	2 (HPC, Fan)	249	248

Berbagai pendekatan telah diusulkan untuk prediksi RUL pada C-MAPSS. Pendekatan awal menggunakan *recurrent neural network* (RNN) dengan pengenalan ambang RUL *piecewise-linear* sebesar 125 siklus yang kini menjadi standar [3]. Pendekatan Dual-Channel LSTM dengan diferensiasi fitur waktu juga dikembangkan untuk meningkatkan akurasi pada subset multi-kondisi [4]. Arsitektur CAELSTM yang mengintegrasikan *convolutional autoencoder* dan *attention-based* LSTM melaporkan hasil RMSE 14,44 pada FD001 [5]. Kerangka DAB-LSTM dengan *dual attention block* diperkenalkan untuk menangkap dependensi temporal kompleks secara lebih efektif [6]. Model *lightweight* berbasis *dual attention* LSTM dengan *exponential smoothing* juga dikembangkan untuk menyeimbangkan akurasi dan efisiensi komputasi [7]. Batas SOTA kemudian didorong lebih jauh melalui kerangka *Bidirectional Mamba* yang dipadukan *causal discovery* LSTM [8]. Tren ini menunjukkan dominasi *deep learning* dalam prediksi RUL [9].

Namun pendekatan *tree-based* seperti XGBoost tetap kompetitif dari sisi efisiensi komputasi dan interpretabilitas [10]. Pendekatan ensemble yang memadukan XGBoost dengan SHAP juga telah dikembangkan untuk mendukung *dynamic maintenance scheduling* [11]. Di luar konteks prediksi RUL, XGBoost telah diterapkan pada berbagai permasalahan, seperti diagnosa kanker payudara [12], prediksi diabetes [13], optimasi *hyperparameter* [14], serta prediksi kualitas udara [15]. Untuk pemodelan sekuensial, komparasi LSTM dan RNN untuk prediksi penggunaan listrik menunjukkan keunggulan LSTM dalam menangani dependensi jangka panjang [16]. Studi lain memanfaatkan LSTM untuk peramalan cuaca [17], prediksi harga saham [18], serta optimasi *neural network* untuk *time-series* [19]. Selaras dengan kemajuan akurasi, kebutuhan akan *Explainable Artificial Intelligence* (XAI) semakin meningkat. SHAP (*SHapley Additive exPlanations*) menyediakan kerangka teori-permainan untuk mengkuantifikasi kontribusi fitur [20], [21], sementara *Integrated Gradients* (IG) merupakan metode atribusi berbasis integral jalur yang memenuhi aksioma sensitivitas [22]. Pada konteks C-MAPSS, kombinasi *permutation importance* dan nilai Shapley telah diterapkan pada arsitektur LSTM-GRU untuk mengidentifikasi sensor kritis [23]. Meskipun literatur RUL pada C-MAPSS telah berkembang pesat, masih terdapat beberapa celah penelitian. Pertama, sebagian penelitian hanya mengevaluasi satu atau dua subset C-MAPSS [5], [11], sehingga perbandingan kinerja model pada tingkat kompleksitas dan kondisi operasi yang berbeda belum dapat dilakukan secara seragam. Kedua, komparasi langsung antara model tree-based gradient boosting dan deep learning berbasis attention pada seluruh subset C-MAPSS masih terbatas. Ketiga, analisis konsistensi kontribusi sensor antara metode XAI yang berbeda, khususnya SHAP dan Integrated Gradients, pada subset dengan karakteristik operasi yang berbeda masih belum banyak dieksplorasi.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan membandingkan kinerja XGBoost dan Bi-LSTM berbasis attention dalam memprediksi RUL pada keempat subset C-MAPSS serta menganalisis konsistensi kontribusi sensor menggunakan SHAP dan Integrated Gradients. Kontribusi penelitian ini meliputi komparasi sistematis kedua paradigma model pada seluruh subset C-MAPSS, penerapan *condition-aware normalization* berbasis *clustering* pada subset multi-kondisi, serta analisis konsensus XAI untuk mengidentifikasi konsistensi kontribusi sensor terhadap prediksi RUL.

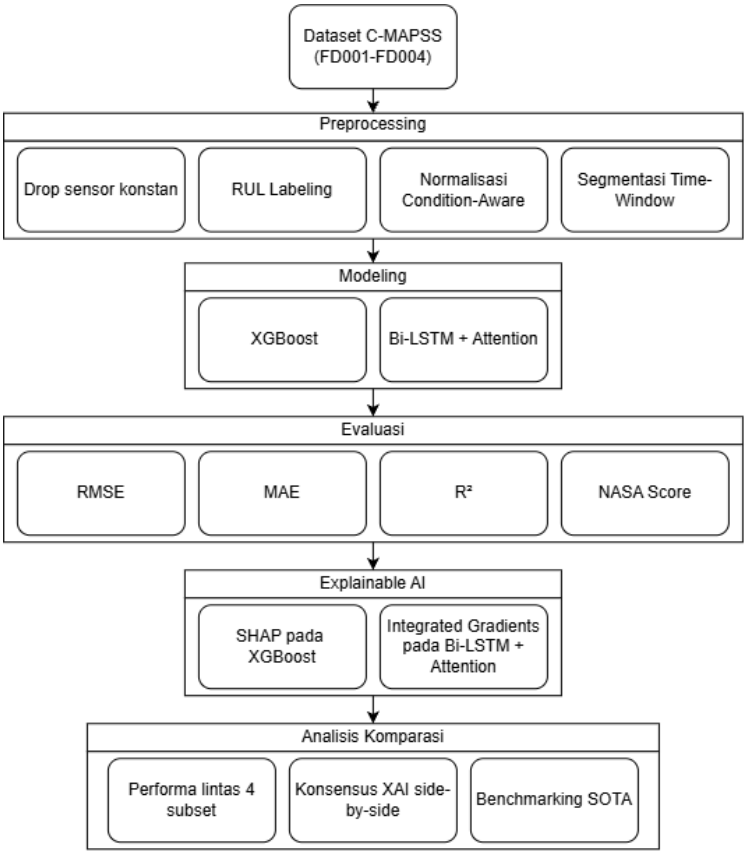
2. METODOLOGI PENELITIAN

2.1 Alur Penelitian

Penelitian ini mengikuti enam tahapan utama sebagaimana ditunjukkan pada Gambar 1: (1) pengumpulan *Dataset C-MAPSS* (FD001–FD004) dari repositori NASA *Prognostics Data Repository* [1], (2) *Preprocessing* yang terdiri atas *drop sensor konstan*, *RUL labeling*, normalisasi *condition-aware*, dan *time-window*, (3) *Modeling* menggunakan XGBoost dan Bi-LSTM, (4) *Evaluasi* dengan empat metrik (RMSE, MAE, R^2 *Score*, NASA *Score*), (5) *Explainable AI* menggunakan SHAP pada XGBoost dan *Integrated Gradients* pada Bi-LSTM, dan (6) *Analisis Komparasi* meliputi performa lintas 4 subset, konsensus XAI *side-by-side*, dan *benchmarking* SOTA. Seluruh eksperimen diimplementasikan menggunakan Python 3.10 dengan pustaka scikit-learn, XGBoost, TensorFlow, dan SHAP.

Copyright © 2025 The **Autor**, Page 1349
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 1. Alur penelitian

2.2 Preprocessing Data

Pelabelan *Remaining Useful Life* (RUL) menggunakan pendekatan *piecewise-linear* dengan batas maksimum 125 siklus [3]. Label RUL dihitung berdasarkan selisih antara siklus maksimum suatu unit mesin dan siklus pengamatan. Secara matematis, perhitungan RUL dirumuskan sebagai:

$$\mathrm{RUL}_{i,t} = \min(\mathrm{RUL}_{\max}, C_i - t) \quad (1)$$

dengan $\mathrm{RUL}_{i,t}$ merupakan nilai RUL unit mesin ke-i pada siklus ke-t, C_i merupakan siklus maksimum unit mesin ke-i, dan $\mathrm{RUL}_{\max}$ merupakan batas maksimum RUL yang ditetapkan sebesar 125 siklus. Penerapan batas maksimum tersebut bertujuan membatasi nilai RUL pada fase awal operasi ketika degradasi mesin belum menunjukkan perubahan yang signifikan.

Sensor yang memiliki nilai unik kurang dari atau sama dengan satu pada data pelatihan dihapus dari data pelatihan dan pengujian pada masing-masing subset. Berdasarkan proses tersebut, jumlah sensor yang digunakan setelah *preprocessing* adalah 18 sensor untuk FD001 dan FD003 yang memiliki satu kondisi operasi, serta 24 sensor untuk FD002 dan FD004 yang memiliki enam kondisi operasi.

Normalisasi *condition-aware* diterapkan untuk menyesuaikan perbedaan distribusi sensor akibat variasi kondisi operasi. Pada FD001 dan FD003 yang memiliki satu kondisi operasi, *MinMaxScaler* diterapkan secara global berdasarkan data pelatihan. Parameter *scaler* yang diperoleh dari data pelatihan kemudian digunakan untuk mentransformasi data pengujian.

Sementara itu, pada FD002 dan FD004 yang memiliki enam kondisi operasi, pengelompokan kondisi operasi dilakukan menggunakan *K-Means* dengan $k = 6$ berdasarkan tiga *operational settings*. Selanjutnya, normalisasi *Min-Max* diterapkan secara terpisah pada setiap kluster. Nilai hasil normalisasi dihitung menggunakan persamaan:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2)$$

Copyright © 2025 The **Autor**, Page 1350
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Dengan x' merupakan nilai sensor setelah normalisasi, x merupakan nilai sensor sebelum normalisasi, sedangkan $x_{\min}$ dan $x_{\max}$ masing-masing merupakan nilai minimum dan maksimum sensor pada kluster kondisi operasi yang bersangkutan. Parameter normalisasi ditentukan berdasarkan data pelatihan dan selanjutnya digunakan untuk mentransformasi data pengujian pada kondisi yang sesuai. Strategi ini bertujuan mengurangi pengaruh variasi pembacaan sensor yang disebabkan oleh perbedaan kondisi operasi sehingga model dapat lebih berfokus pada pola degradasi mesin Untuk Bi-LSTM, data sensor yang telah dinormalisasi kemudian disegmentasi menggunakan *sliding-window* sepanjang 30 siklus dengan *stride* 1. Pemilihan panjang 30 siklus merujuk pada literatur [3], [6] yang menunjukkan bahwa panjang tersebut memberikan keseimbangan antara konteks temporal dan jumlah sampel pelatihan. Pada fase pengujian, hanya *sequence* terakhir dari setiap unit mesin yang digunakan sebagai masukan model untuk memprediksi RUL.

2.3 Model XGBoost

XGBoost [10] membangun ansambel pohon Keputusan secara iteratif, di mana setiap pohon baru dilatih untuk memperbaiki residu prediksi pohon sebelumnya. Mekanisme regularisasi (*gamma, alpha, lambda*) dan teknik *column subsampling* membuat model ini tahan terhadap *overfitting*. Sementara optimasi *parallel tree construction* membuat waktu pelatihannya jauh lebih cepat dibanding *deep learning*. *Hyperparameter* yang digunakan diadaptasi dari literatur [11], [23] dan disesuaikan kompleksitas masing-masing subset, sebagaimana dirangkum pada Tabel 2.

Tabel 2. Konfigurasi Hyperparameter XGBoost per Kategori Subset

parameter	FD001/FD003	FD002/FD004
n_estimators	500	800
learning_rate	0,05	0,05
max_depth	6	8
min_child_weight	3	3
subsample	0,8	0,8
colsample_bytree	0,8	0,8
gamma	0,1	0,1
reg_alpha	0,05	0,05
reg_lambda	1,0	1,0

Konfigurasi yang lebih dalam (max_depth=8, n_estimators=800) untuk FD002 dan FD004 dimotivasi oleh variabilitas data yang lebih tinggi pada subset multi-kondisi, sehingga model membutuhkan kapasitas representasi yang lebih besar.

2.4 Model Bi-LSTM dengan Self-Attention

Arsitektur Bi-LSTM [9] yang digunakan terdiri atas dua lapisan *Bidirectional* LSTM yang diikuti oleh lapisan *Self-Attention* dan dua lapisan *Dense*, sebagaimana dirinci pada Tabel 3.

Tabel 3. Arsitektur Bi-LSTM + Attention

Lapisan	Output Shape	Keterangan
Input	(30, n_fitur)	-
Bi-LSTM 1	(30, 128)	*return_sequences=True*
Dropout 1	(30, 128)	*rate=0,2*
Bi-LSTM 2	(30, 64)	*return_sequences=True*
Dropout 2	(30, 64)	*rate=0,2*
Self-Attention	(64,)	*Bahdanau-style*
Dense 1	(32,)	*Activation=ReLU*
Dropout 3	(32,)	*Rate=0,2*
Output	(1,)	*Activation=Linear*

Lapisan *Additive Self-Attention* mengimplementasikan mekanisme *Bahdanau-style* yang menghitung bobot relevansi setiap *timestep* terhadap prediksi RUL. Mekanisme ini memungkinkan model secara otomatis memprioritaskan siklus-siklus yang mengandung sinyal degradasi kuat dan mengabaikan siklus yang kurang informatif. Model dilatih dengan *optimizer* Adam (*learning rate 1e-3*), *callback EarlyStopping* (*patience*=10, *restore_best_weights=True*), *ReduceLROnPlateau* (*factor*=0,5, *patience*=5, *min_lr*=1e-5), selama maksimum 80 *epoch* dengan *batch size* 256 dan *validation split* 10%

2.5 Metrik Evaluasi

Empat metrik yang digunakan untuk mengevaluasi performa prediksi RUL, di mana n adalah jumlah sampel, y_i adalah nilai RUL aktual, $\hat{y}_i$ adalah nilai RUL prediksi, dan $\bar{y}$ adalah rata-rata dari nilai RUL aktual: Root Mean Square Error (RMSE): Mengukur rata-rata kuadrat kesalahan prediksi, sensitif terhadap pencilan (*outlier*).

Copyright © 2025 The **Autor**, Page 1351
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

Mean Absolute Error (MAE): Mengukur rata-rata kesalahan absolut, lebih tangguh (*robust*) terhadap pencilan.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (4)$$

Koefisien Determinasi (R^2): Mengukur proporsi variansi target yang dapat dijelaskan oleh model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

NASA Asymmetric Scoring Function (S): Metrik standar C-MAPSS [1] yang memberikan penalti asimetris berdasarkan arah kesalahan prediksi ($d_i = \hat{y}_i - y_i$). *Over-estimation* ($d_i \geq 0$) mendapat penalti lebih berat karena lebih berbahaya dalam konteks operasional.

$$S = \sum_{i=1}^{n} s_i \quad (6)$$

$$s_i = \begin{cases} e^{-\frac{d_i}{13}} - 1, & d_i < 0 \\ e^{\frac{d_i}{10}} - 1, & d_i \geq 0 \end{cases}$$

Koefisien penyebut lebih kecil pada kasus *over-estimation* (10 berbanding 13) menyebabkan penalty tumbuh lebih cepat secara eksponensial.

2.6 Explainable AI

SHAP TreeExplainer [20]. Digunakan untuk menginterpretasikan XGBoost. SHAP menghitung kontribusi marginal rata-rata setiap fitur terhadap prediksi berdasarkan teori permainan koalisi (*Shapley values*) [21]. Implementasi *TreeExplainer* memungkinkan perhitungan nilai SHAP secara eksak dalam kompleksitas waktu polinomial. Tiga jenis visualisasi dihasilkan per subset: (1) *feature importance* berbasis *gain* sebagai pembanding, (2) *SHAP summary plot* yang menampilkan distribusi kontribusi fitur secara global, dan (3) *SHAP waterfall plot* yang menjelaskan prediksi individual pada unit kritis.

Integrated Gradients (IG) [22]. digunakan untuk menginterpretasikan model Bi-LSTM. IG menghitung integral jalur dari gradien *output* model terhadap *input* sepanjang interpolasi linier antara *baseline* (vektor nol) dan *input* asli. Metode ini memenuhi dua aksioma penting: sensitivitas (fitur yang memengaruhi prediksi mendapat atribusi non-nol) dan invariansi implementasi. Perhitungan dilakukan dengan *Riemann sum approximation* menggunakan 50 langkah interpolasi. Dua jenis visualisasi dihasilkan per subset: (1) IG global yang merata-ratakan absolut nilai IG atas seluruh sampel pengujian, dan (2) IG lokal yang menampilkan atribusi untuk unit kritis individual.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Komparasi Kinerja Prediktif

Evaluasi kinerja prediksi *Remaining Useful Life* (RUL) dilakukan menggunakan XGBoost dan Bi-LSTM dengan *self-attention* pada keempat subset C-MAPSS. Evaluasi menggunakan empat metrik, yaitu *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), Koefisien Determinasi (R^2), dan *NASA Scoring Function*. Untuk memberikan transparansi penuh terhadap proses pengolahan dan evaluasi data, sub-bab ini diawali dengan contoh perhitungan manual pada tahapan pelabelan *piecewise-linear*, normalisasi *condition-aware*, hingga perhitungan metrik evaluasi.

3.1.1 Contoh Perhitungan Pelabelan RUL

Pelabelan RUL dilakukan menggunakan pendekatan *piecewise-linear* dengan batas maksimum 125 siklus [3]. Sebagai contoh penerapan pada data aktual subset FD001, unit mesin nomor 1 memiliki total siklus operasi sebanyak 192 siklus sebelum mengalami kegagalan. Pada observasi siklus ke-20, nilai RUL dihitung sebagai:

$$\text{RUL} = 192 - 20 = 172$$

Karena nilai 172 melebihi batas maksimum, maka dilakukan pembatasan (*clipping*):

$$\text{RUL} = \min(172,125) = 125$$

Sebaliknya, pada observasi siklus ke-142, nilai RUL dihitung sebagai:

Copyright © 2025 The **Autor**, Page 1352
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

$$\text{RUL} = 192 - 142 = 50$$

Karena nilai 50 masih berada di bawah batas 125 siklus, maka label RUL yang digunakan adalah 50 siklus tanpa pembatasan. Dengan demikian, penerapan *piecewise-linear* menghasilkan label RUL maksimum 125 siklus pada fase awal operasi mesin di mana degradasi belum signifikan dan menurun secara linear mendekati akhir masa operasi.

3.1.2 Contoh Perhitungan Normalisasi Condition-Aware

Pada subset multi-kondisi (FD002 dan FD004), normalisasi dilakukan secara terpisah per kluster kondisi operasi untuk menghilangkan sinyal semu. Sebagai contoh penerapan pada data aktual, setelah proses *clustering* K-Means ($k = 6$) pada subset FD002, sensor Ps30 pada kluster 0 memiliki nilai minimum 327,00 dan nilai maksimum 336,00. Untuk suatu observasi dengan nilai Ps30 sebesar 330,00, normalisasi *MinMaxScaler* dihitung sebagai:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Sehingga :

$$x' = \frac{330{,}00 - 327{,}00}{336{,}00 - 327{,}00} = \frac{3{,}00}{9{,}00} = 0{,}333$$

Dengan demikian, nilai Ps30 sebesar 330,00 ditransformasikan menjadi 0,333 sebelum digunakan sebagai masukan model. Proses ini diterapkan pada seluruh sensor dan seluruh kluster secara independen, sehingga perbedaan pembacaan sensor antar-kondisi operasi tidak memengaruhi proses pelatihan model.

3.1.3 Contoh Perhitungan Metrik Evaluasi

Untuk mengilustrasikan penerapan formula evaluasi yang telah dijabarakan pada Bab 2.5, berikut disajikan contoh perhitungan pada data sampel dari subset FD001. Tabel 4 menampilkan tiga unit mesin pengujian dengan nilai RUL aktual (y) dan prediksi (ŷ) dari masing-masing model.

Tabel 4. Contoh Data Sample Prediksi RUL pada Subset FD001

unit	RUL Aktual (y)	Prediksi XGBoost (ŷ)	Prediksi Bi-LSTM(ŷ)
33	7	6,40	10,50
55	30	28,15	27,80
72	65	60,20	62,10

Berdasarkan data pada Tabel 4, perhitungan RMSE untuk model XGBoost adalah sebagai berikut:

$$RMSE_{XGBoost} = \sqrt{\frac{1}{3}((7 - 6{,}40)^2 + (30 - 28{,}15)^2 + (65 - 60{,}20)^2)} = \sqrt{\frac{26{,}8225}{3}} = \sqrt{8{,}94} = 2{,}99$$

Sedangankan perhitungan RMSE untuk Bi-LSTM :

$$\text{RMSE}_{\text{Bi-LSTM}} = \sqrt{\frac{1}{3}((7 - 10{,}50)^2 + (30 - 27{,}80)^2 + (65 - 62{,}10)^2)} = \sqrt{\frac{25{,}50}{3}} = \sqrt{8{,}50} = 2{,}92$$

Pada contoh tiga unit ini, Bi-LSTM memperoleh RMSE yang sedikit lebih rendah (2,92) dibandingkan XGBoost (2,99). Selanjutnya, untuk menunjukkan karakteristik asimetris dari *NASA Scoring Function*, perhitungan dilakukan pada unit 33 yang memiliki RUL aktual paling rendah ($y = 7$). Pada model XGBoost dengan prediksi $\hat{y} = 6{,}40$, diperoleh $d = \hat{y} - y = 6{,}40 - 7 = -0{,}60$. Karena $d < 0$ (kasus *under-estimation*), maka skor dihitung sebagai:

$$s = \exp\left(-\frac{-0{,}60}{13}\right) - 1 = \exp(0{,}046) - 1 = 0{,}047$$

Sebaliknya, pada model Bi-LSTM dengan prediksi $\hat{y} = 10{,}50$ untuk unit yang sama, diperoleh $d = 10{,}50 - 7 = 3{,}50$. Karena $d \geq 0$ (kasus *over-estimation*), maka skor dihitung sebagai:

$$s = \exp\left(\frac{3{,}50}{10}\right) - 1 = \exp(0{,}35) - 1 = 0{,}419$$

Perbandingan kedua skor ini (0,047 untuk XGBoost berbanding 0,419 untuk Bi-LSTM) menunjukkan bahwa meskipun selisih absolut prediksi Bi-LSTM terhadap nilai aktual (3,50 siklus) tidak jauh berbeda dari XGBoost (0,60 siklus), penalti yang dihasilkan oleh NASA Scoring Function pada kasus over-estimation hampir 9 kali lebih besar. Fenomena inilah

Copyright © 2025 The **Autor**, Page 1353
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

yang menyebabkan NASA Score Bi-LSTM pada subset FD004 membengkak secara signifikan, karena model tersebut cenderung melakukan over-estimation pada banyak unit mesin secara bersamaan.

3.1.4 Hasil Pengujian pada Empat Subset C-MAPSS

Hasil evaluasi lengkap pada seluruh unit pengujian di keempat subset dirangkum pada Tabel 5.

Tabel 5. Hasil Evaluasi Komparatif XGBoost dan Bi-LSTM Lintas Empat Subset C-MAPSS

subset	Model	RMSE	MAE	R^2	NASA Score
FD001	XGBoost	18,20	13,34	0,808	847,25
	Bi-LSTM	15,35	11,61	0,864	392,02
FD002	XGBoost	29,72	20,73	0,695	14.226,59
	Bi-LSTM	27,92	18,91	0,730	9.075,76
FD003	XGBoost	21,75	16,25	0,724	2.236,63
	Bi-LSTM	17,33	12,95	0,825	709,82
FD004	XGBoost	30,22	22,73	0,693	9.423,85
	Bi-LSTM	33,40	22,28	0,625	178.903,47

Berdasarkan Tabel 5, Bi-LSTM dengan *self-attention* menunjukkan kinerja yang superior pada tiga dari empat subset C-MAPSS. Pada FD001 yang memiliki tingkat kompleksitas terendah, Bi-LSTM menghasilkan RMSE 15.35, mengungguli XGBoost yang mencatat angka 18.20. Tren keunggulan ini berlanjut pada FD002 dan FD003. Di mana Bi-LSTM secara konsisten memperoleh nilai MAE dan RMSE yang lebih rendah. Pada FD002 yang memiliki enam kondisi operasi, Bi-LSTM mampu menekan RMSE hingga 27.92, sementara pada FD003 yang memiliki dua mode kegagalan, model ini mencapai performa optimal dengan R^2 sebesar 0.825. keunggulan konsisten pada tiga subset pertama ini mengindikasikan efektivitas mekanisme *Attention* dalam membantu Bi-LSTM fokus pada sinyal sensor yang paling relevan terhadap degradasi seiring waktu.

Namun, anomali signifikan terdeteksi ketika model diuji pada sub-dataset FD004 yang merepresentasikan skenario paling kompleks, performa Bi-LSTM menurun drastis dengan RMSE melonjak menjadi 33.40, lebih buruk dibandingkan XGBoost yang mencatat angka 30.22. kesenjangan performa pada kondisi ekstrem ini menunjukkan adanya keterbatasan pada stabilitas model sekuensial ketika dihadapkan pada varians data yang sangat tinggi.

Kelemahan arsitektur berbasis *deep learning* pada kondisi FD004 semakin terlihat melalui metrik NASA Score untuk Bi-LSTM membengkak hingga mencapai 178.903, berbanding terbalik dengan XGBoost yang tetap stabil di angka 9.423. Secara matematis, metrik ini memberikan penalti yang lebih berat pada kasus *over-estimation* (prediksi RUL lebih besar dari nilai aktual) dibandingkan *under-estimation* [1]. Fenomena membengkaknya skor ini mengindikasikan bahwa model Bi-LSTM terlalu sering memprediksi sisa umur mesin yang lebih panjang dari kenyataan aslinya. Dalam konteks operasional penerbangan, perilaku *over-estimation* sangat berbahaya karena dapat memicu penundaan jadwal pemeliharaan yang berujung pada risiko kegagalan mesin secara mendadak saat beroperasi.

Dugaan penyebab utama perilaku *over-estimation* ini adalah interaksi non-linear yang sangat kompleks antara enam kondisi operasi dan dua mode kegagalan pada FD004. *Multi-condition* menyebabkan distribusi nilai sensor pada satu unit menjadi sangat fluktuatif, sementara *multi-fault* mengakibatkan pola degradasi di akhir masa pakai mesin menjadi tidak konsisten antar-unit. Bi-LSTM, meskipun dirancang untuk menangkap dependensi temporal, diduga mengalami kesulitan dalam membedakan fluktuasi akibat perubahan mode operasi dengan tren degradasi yang sebenarnya pada subset FD004. Di sisi lain XGBoost sebagai model tabular tidak terpengaruh oleh urutan temporal data, sehingga pada kasus spesifik ini justru menunjukkan ketahanan yang lebih baik terhadap variabilitas kondisi operasi. Temuan ini mengindikasikan bahwa keunggulan *deep learning* dalam pemodelan sekuensial tidak selalu berlaku pada skenario dengan kombinasi *multi-condition* dan *multi-fault* yang ekstrem.

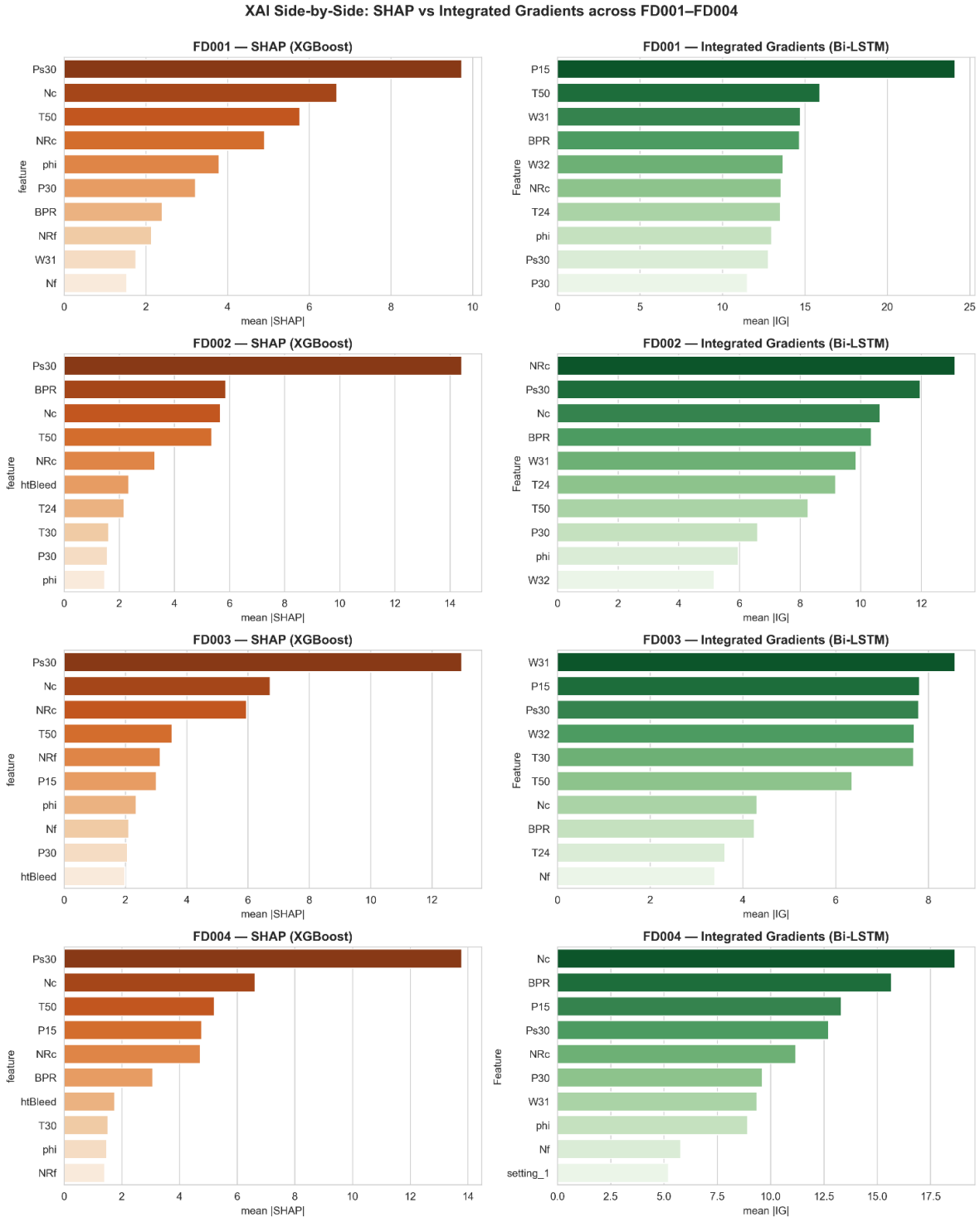
3.2 Interpretasi Model Berbasis Explainable AI

Selain evaluasi performa prediktif, penelitian ini menerapkan dua teknik XAI yang sesuai dengan karakteristik masing-masing model: SHAP TreeExplainer untuk XGBoost dan Integrated Gradients untuk Bi-LSTM. Analisis XAI dilakukan pada tiga level, yaitu global (ranking fitur lintas seluruh test set), local (interpretasi prediksi kritis), dan consensus (konvergensi fitur antar model dan subset).

Gambar 2 menampilkan komparasi side-by-side antara SHAP summary plot (XGBoost) dan Integrated Gradients global (Bi-LSTM) untuk keempat subset.

Copyright © 2025 The **Autor**, Page 1354
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 2. Komparasi side-by-side hasil SHAP (XGBoost) dan Integrated Gradients (Bi-LSTM) untuk empat subset C-MAPSS.

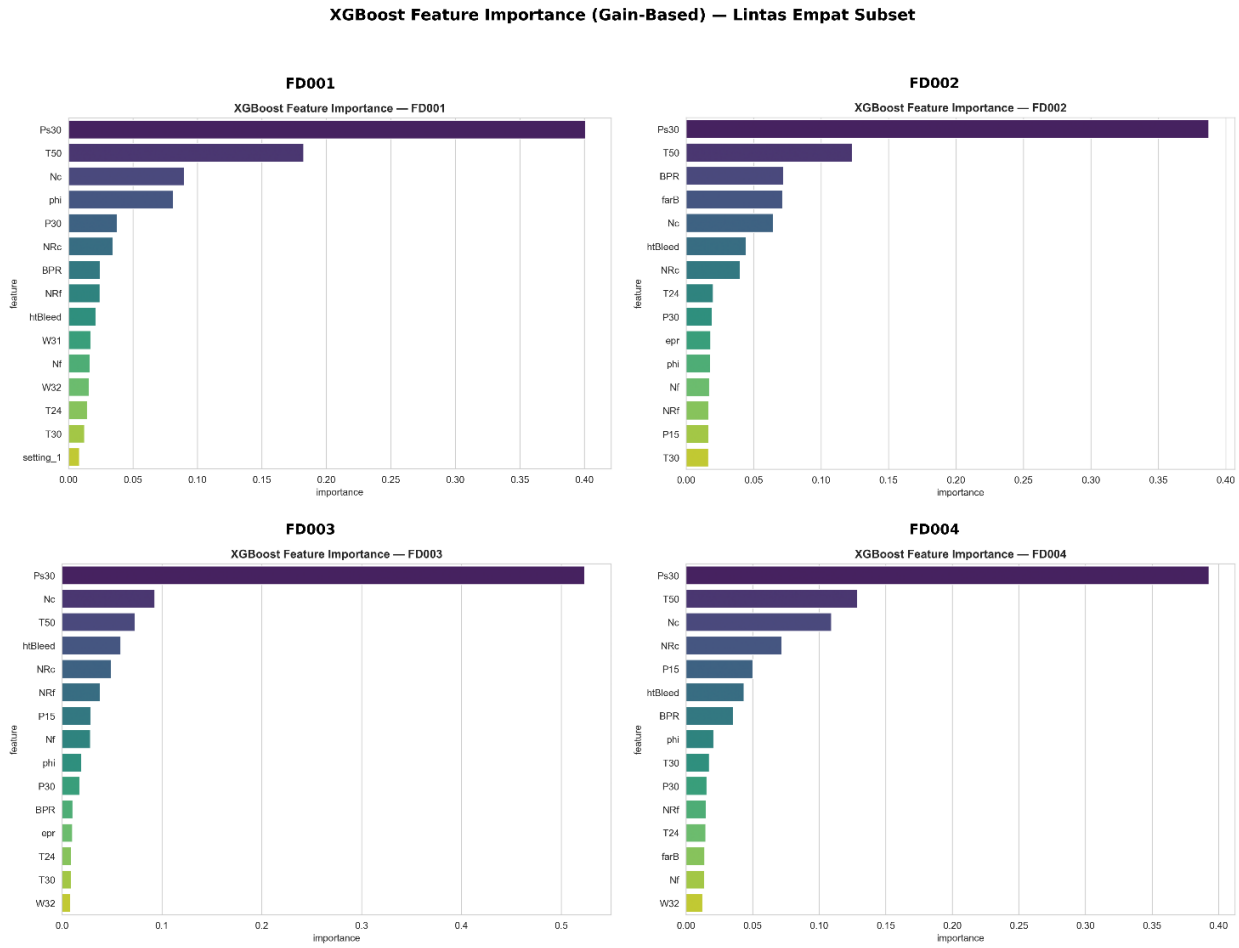
Pada sisi XGBoost, SHAP summary plot menunjukkan bahwa sensor Ps30 (tekanan statis outlet HPC) secara konsisten menempati peringkat pertama di seluruh empat subset dengan sebaran SHAP value terluas, diikuti oleh Nc (kecepatan rotasi inti) dan T50 (temperatur exhaust). Intensitas warna pada visualisasi menunjukkan Tingkat kontribusi fitur terhadap model, di mana warna oranye yang lebih gelap merepresentasikan tingkat kepentingan (*importance*) yang lebih tinggi, sedangkan warna yang lebih terang menunjukkan kontribusi yang lebih rendah. Dominasi Ps30 mengindikasikan bahwa tekanan outlet HPC memiliki pengaruh signifikan terhadap prediksi *Remaining Useful Life* (RUL), selaras dengan pemahaman termodinamika bahwa penurunan tekanan pada outlet HPC merupakan indikator kuat terjadinya degradasi kompresor.

Copyright © 2025 The **Autor**, Page 1355
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Pada sisi Bi-LSTM, pola Integrated Gradients global menunjukkan preferensi fitur yang lebih dinamis. Pada FD001, sensor P15 (tekanan booster outlet), W31, dan W32 (sensor aliran) mendominasi, sedangkan Ps30 hanya menempati peringkat ke-10. Namun pada subset multi-kondisi (FD002 dan FD004), Integrated Gradients bergeser signifikan: NRc, Ps30, dan NC naik ke peringkat atas, sementara sensor aliran (W31, W32) turun. Pergeseran ini menunjukkan bahwa Bi-LSTM lebih adaptif dalam memilih fitur tergantung kompleksitas data, namun adaptivitas ini juga berarti preferensi fiturnya kurang stabil dibanding XGBoost.

Temuan menarik lainnya muncul dari perbandingan antara *gain-based feature importance* dan SHAP pada XGBoost. *Gain-based feature* menunjukkan Ps30 mendominasi dengan proporsi 40-52% dari total gain di semua subset (Gambar 3), sementara SHAP mendistribusikan kontribusi lebih merata. Perbedaan ini mengonfirmasi bahwa *gain-based importance* cenderung *overestimate* fitur yang sering dijadikan split point, sedangkan SHAP memberikan gambaran yang lebih akurat tentang kontribusi aktual fitur terhadap prediksi.

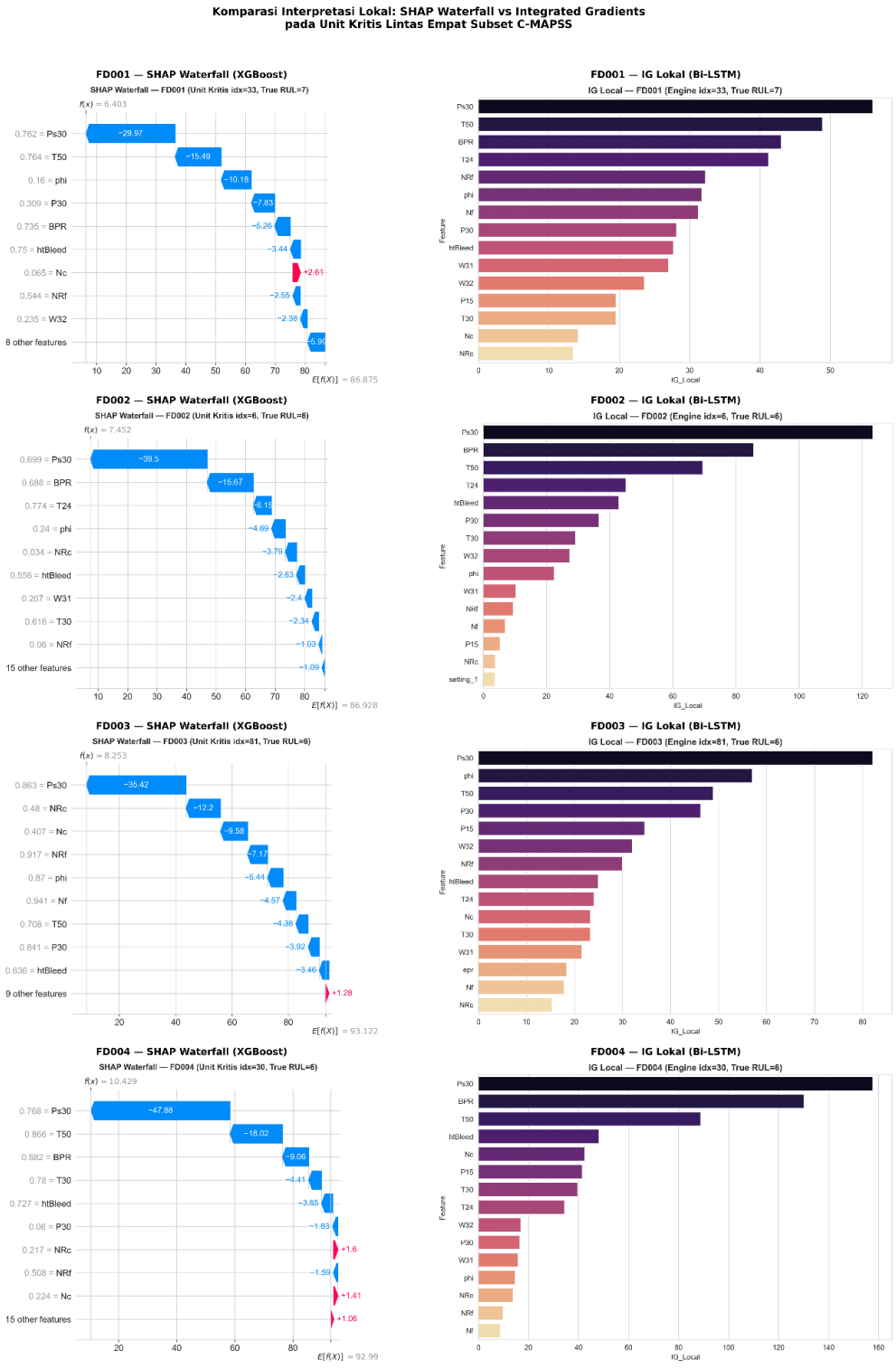


Gambar 3. Feature importance gain-based XGBoost untuk keempat subset C-MAPSS

Untuk mengkaji apakah konsensus XAI juga berlaku pada level prediksi individual, analisis lokal dilakukan pada unit kritis (unit dengan True RUL terkecil) di setiap subset. Gambar 4 menampilkan komparasi *side-by-side* antara SHAP waterfall (XGBoost, kolom kiri) dan IG lokal (Bi-LSTM, kolom kanan) untuk keempat subset, di mana setiap baris merepresentasikan unit kritis yang sama pada subset bersangkutan.

Copyright © 2025 The **Autor**, Page 1356
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom



Gambar 4. Komparasi interpretasi lokal pada unit kritis: SHAP waterfall (kiri) dan Integrated Gradients (kanan) untuk keempat subset C-MAPSS.

Gambar 4 mengungkap temuan yang sangat menarik: meskipun secara global (analisis di atas) kedua model memiliki preferensi fitur yang berbeda terutama pada FD001 dan FD003 pada level unit kritis yang mendekati kegagalan, kedua model justru menunjukkan kesepakatan yang tinggi di seluruh subset. Pada FD001, SHAP waterfall menunjukkan Ps30 (-29,97), T50 (-15,49), dan phi (-10,18) sebagai tiga kontributor utama, sedangkan IG lokal menunjukkan Ps30, T50, dan BPR. Pola serupa teramati pada FD002 (Ps30 dan BPR konsisten di kedua metode), FD003 (Ps30 konsisten), dan FD004 (Ps30, T50, dan BPR ketiganya konsisten). Secara keseluruhan, Ps30 menempati peringkat pertama di kedua metode XAI pada seluruh empat subset untuk unit yang mendekati kegagalan. Tabel 6 merangkum konsensus lokal ini secara lengkap.

Copyright © 2025 The **Autor**, Page 1357
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

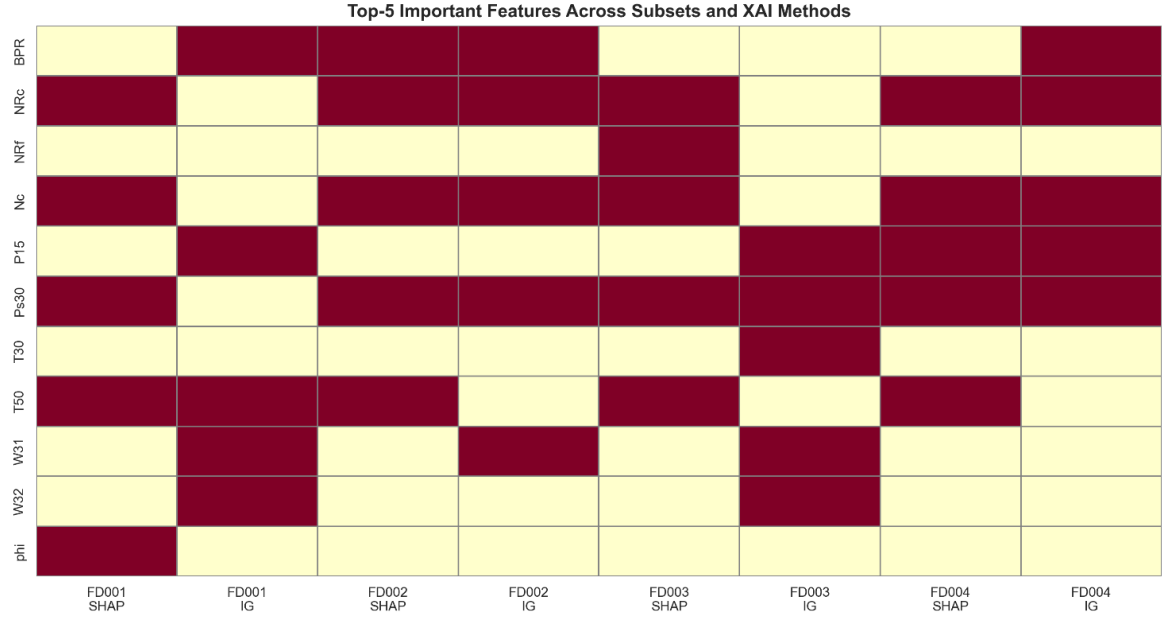
Tabel 6. Komparasi Top-3 Sensor pada Analisis Lokal (Unit Kritis) SHAP vs IG

Subset	Unit	True RUL	SHAP Waterfall Top-3	IG Lokal Top-3	Sensor Konsensu
FD001	idx=33	7	Ps30, T50, phi	Ps30, T50, BPR	Ps30, T50
FD002	idx=6	6	Ps30, BPR, T24	Ps30, BPR, T50	Ps30, BPR
FD003	idx=81	6	Ps30, NRc, Nc	Ps30, phi, T50	Ps30
FD004	idx=30	6	Ps30, T50, BPR	Ps30, BPR, T50	Ps30, T50, BPR

Pada level konsensus global, Tabel 7 merangkum kesepakatan top-5 fitur antara SHAP dan IG, sedangkan Gambar 5 memvisualisasikan kehadiran setiap fitur di top-5 kedua metode lintas empat subset.

Tabel 7. Konsensus Top-5 Fitur Global antara SHAP dan IG Lintas Empat Subset.

Subset	SHAP Top-5	IG Top-5	Fitur Konsensus	Kesepakatan
FD001	NRc, Nc, Ps30, T50, phi	BPR, P15, T50, W31, W32	T50	20%
FD002	BPR, NRc, Nc, Ps30, T50	BPR, NRc, Nc, Ps30, W31	BPR, NRc, Nc, Ps30	80%
FD003	NRc, NRf, Nc, Ps30, T50	P15, Ps30, T30, W31, W32	Ps30	20%
FD004	NRc, Nc, P15, Ps30, T50	BPR, NRc, Nc, P15, Ps30	NRc, Nc, P15, Ps30	80%



Gambar 5. Heatmap kehadiran fitur di top-5 SHAP dan IG lintas empat subset. Warna gelap menandakan fitur berada di top-5.

Pola kontra-intuitif teramati: kesepakatan 80% terjadi pada subset multi-kondisi (FD002, FD004), namun hanya 20% pada subset kondisi tunggal (FD001, FD003). Pada subset kondisi tunggal, *signal-to-noise* ratio relatif tinggi sehingga ruang fitur yang valid sebagai indikator degradasi cukup luas akibatnya XGBoost dan Bi-LSTM dapat memilih "jalur berbeda" untuk membuat prediksi. XGBoost cenderung mengandalkan sensor termodinamika inti (Nc, NRc, Ps30, T50), sementara Bi-LSTM memanfaatkan sensor aliran (W31, W32) dan tekanan (P15) yang menyimpan informasi temporal. Sebaliknya pada subset multi-kondisi, variabilitas kondisi operasi membuat sebagian besar sensor terkontaminasi noise, sehingga hanya sedikit sensor yang benar-benar reliable. Konvergensi ini memaksa kedua model untuk mengandalkan kelompok sensor inti yang sama: Ps30, NRc, Nc, dan BPR.

Dari Gambar 5, dapat diidentifikasi tiga kategori sensor berdasarkan konsistensi kemunculannya: (1) sensor universal yang muncul di top-5 kedua metode di hampir semua subset: Ps30 (8/8 slot) dan Nc (6/8 slot); (2) sensor model-specific XGBoost: T50 muncul konsisten di top-5 SHAP (4/4 subset) namun hanya 1/4 di IG; dan (3) sensor model-specific Bi-LSTM: W31, W32, dan P15 muncul konsisten di IG pada subset kondisi tunggal namun hilang pada multi-kondisi.

Temuan ini memiliki implikasi penting bagi praktik PHM industri: untuk lingkungan operasional dengan kondisi heterogen, sensor Ps30 (tekanan statis outlet HPC), NRc (kecepatan rotasi inti terkoreksi), dan Nc (kecepatan rotasi inti) merupakan indikator degradasi yang paling reliable terlepas dari arsitektur model yang digunakan.

Copyright © 2025 The **Autor**, Page 1358
This Journal is licensed under a Creative Commons Attribution 4.0 International License

JURIKOM (Jurnal Riset Komputer), Vol. 13 No. 5, Oktober 2026
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v15i5.9868
Hal 1348-1361
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

3.3 Pembahasan

Berdasarkan hasil eksperimen yang telah diuraikan, terdapat tiga temuan utama yang layak dibahas secara mendalam dan dibandingkan dengan penelitian sejenis.

Temuan pertama berkaitan dengan keunggulan Bi-LSTM pada tiga subset dan anomali pada FD004. Hasil penelitian ini menunjukkan bahwa arsitektur deep learning berbasis attention tidak selalu superior terhadap model tree-based pada seluruh tingkat kompleksitas data. Temuan serupa dilaporkan dalam [5] yang mencatat bahwa arsitektur CAELSTM juga mengalami penurunan performa pada subset multi-kondisi dibanding subset kondisi tunggal. Tabel 8 menyandingkan hasil penelitian ini dengan beberapa metode SOTA yang relevan.

Tabel 8. Komparasi RMSE Penelitian Ini dengan Metode SOTA pada Empat Subset C-MAPSS

Metode	FD001	FD002	FD003	FD004
LSTM (Heimes, 2008) [3]	13,52	29,67	17,62	31,84
DCNN (Li, 2018)	12,61	22,36	12,64	23,31
CNN-LSTM-Att (Wang, 2024)	15,97	14,45	13,91	16,64
CAELSTM (Elsherif, 2025) [5]	14,44	-	13,40	-
Ensemble+SHAP (Ozcan, 2025) [11]	6,62	-	9,71	-
Bi-cLSTM (2025)	-	13,96	-	14,25
Penelitian ini XGBoost	18,20	29,72	21,75	30,22
Penelitian ini Bi-LSTM+Attn	15,35	27,92	17,33	33,40

Berdasarkan Tabel 8, kinerja penelitian ini kompetitif namun belum mengungguli SOTA terbaru. Pada FD001, RMSE Bi-LSTM (15,35) sebanding dengan Bi-LSTM+Attention yang dilaporkan sebesar 14,50 dalam [5], tetapi masih di atas CAELSTM (14,44). Pada FD003, Bi-LSTM (17,33) berada pada level yang sama dengan LSTM dasar yang melaporkan 17,62 [3]. Perlu dicatat bahwa mayoritas penelitian SOTA hanya melaporkan hasil pada FD001 dan FD003 [5], [11], sehingga komparasi pada FD002 dan FD004 masih terbatas dalam literatur.

Temuan kedua berkaitan dengan pola konsensus XAI yang kontra-intuitif. Kesepakatan 80% pada subset multi-kondisi versus 20% pada subset kondisi tunggal menunjukkan bahwa variabilitas kondisi operasi justru mempersempit ruang fitur yang reliable, sehingga memaksa kedua model untuk konvergen pada kelompok sensor inti yang sama. Temuan ini melengkapi hasil penelitian [23]. yang menunjukkan pentingnya identifikasi sensor kritis pada dataset CMAPSS berskala besar, namun dengan pendekatan konsensus lintas dua metode XAI yang berbeda, suatu analisis yang belum dilaporkan dalam literatur sebelumnya.

Temuan ketiga berkaitan dengan implikasi praktis bagi industri. Identifikasi Ps30, NRc, Nc, dan BPR sebagai sensor degradasi yang konsisten lintas model dan lintas metode XAI memberikan landasan empirik bagi perancang sistem condition monitoring. Untuk lingkungan operasional dengan kondisi heterogen, sensor-sensor tersebut merupakan indikator degradasi yang paling konsisten dalam eksperimen ini, terlepas dari arsitektur model yang digunakan. Kontribusi utama penelitian ini tidak terletak pada akurasi absolut, melainkan pada kerangka komparatif yang reproducible dan interpretable. Untuk aplikasi industri yang menuntut transparansi dan auditabilitas, kombinasi performa kompetitif dengan validasi silang antar metode XAI memberikan nilai praktis yang lebih tinggi dibandingkan optimasi RMSE absolut tanpa interpretabilitas.

4. KESIMPULAN

Penelitian ini telah melakukan komparasi sistematis antara XGBoost dan Bi-LSTM berbasis *self-attention* untuk prediksi RUL mesin turbofan pada keempat subset C-MAPSS, dilengkapi analisis XAI menggunakan SHAP dan Integrated Gradients, dengan tiga temuan utama. Pertama, Bi-LSTM dengan *self-attention* secara konsisten mengungguli XGBoost pada tiga dari empat subset dengan RMSE 15,35 pada FD001, 27,92 pada FD002, dan 17,33 pada FD003, namun pada subset terkompleks FD004, XGBoost justru memberikan hasil yang lebih baik (RMSE 30,22 berbanding 33,40) dengan NASA Score yang jauh lebih stabil (9.423 berbanding 178.903), yang mengindikasikan bahwa dominasi deep learning tidak berlaku universal pada seluruh tingkat kompleksitas data. Kedua, analisis konsensus XAI side-by-side mengungkap pola kontra-intuitif, yaitu kesepakatan top-5 fitur antara SHAP dan Integrated Gradients mencapai 80% pada subset multi-kondisi (FD002, FD004) namun hanya 20% pada subset kondisi tunggal (FD001, FD003), sementara analisis lokal pada unit kritis menunjukkan bahwa saat mesin mendekati kegagalan, kedua model konvergen pada sensor yang sama, terutama Ps30 yang menempati peringkat pertama di kedua metode XAI pada seluruh subset. Ketiga, sensor Ps30 (tekanan statis outlet HPC), NRc (kecepatan rotasi inti terkoreksi), Nc (kecepatan rotasi inti), dan BPR (bypass ratio) teridentifikasi

Copyright © 2025 The **Autor**, Page 1359
This Journal is licensed under a Creative Commons Attribution 4.0 International License

sebagai indikator degradasi paling konsisten, memberikan landasan empirik bagi perancang sistem condition monitoring. Penelitian ini memiliki beberapa keterbatasan, yaitu hyperparameter yang ditentukan berdasarkan literatur dan penyesuaian manual (bukan optimasi sistematis), arsitektur Bi-LSTM yang digunakan relatif standar, dan strategi normalisasi KMeans yang menggunakan jumlah kluster heuristik, sehingga untuk penelitian lanjutan disarankan penerapan arsitektur Transformer-based untuk mengatasi kelemahan Bi-LSTM pada FD004, penerapan Bayesian Optimization untuk hyperparameter, penambahan metode XAI ketiga seperti LIME untuk memperkuat validasi silang, serta penerapan kerangka ini pada dataset industri nyata di luar simulasi C-MAPSS.

REFERENCES

[1] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *Proc. IEEE Int. Conf. Prognostics and Health Management (PHM)*, 2008, pp. 1–9. doi: 10.1109/PHM.2008.4711414.

[2] F. Xue, G. Jin, L. Tan, C. Zhang, and Y. Yu, "Predictive maintenance programs for aircraft engines based on remaining useful life prediction," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-19957-w.

[3] F. O. Heimes, "Recurrent neural networks for remaining useful life estimation," in *2008 International Conference on Prognostics and Health Management*, IEEE, Oct. 2008, pp. 1–6. doi: 10.1109/PHM.2008.4711422.

[4] C. Peng, J. Wu, Q. Wang, W. Gui, and Z. Tang, "Remaining Useful Life Prediction Using Dual-Channel LSTM with Time Feature and Its Difference," *Entropy*, vol. 24, no. 12, Dec. 2022, doi: 10.3390/e24121818.

[5] S. M. Elsherif, B. Hafiz, M. A. Makhlouf, and O. Farouk, "A deep learning-based prognostic approach for predicting turbofan engine degradation and remaining useful life," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-09155-z.

[6] O. A. Ghoneim, A. Sleem, I. Gad, and T. Allam, "Remaining useful life prediction for turbofan engines using an attention-based data-driven deep-learning approach," *PeerJ Comput. Sci.*, vol. 11, p. e3438, Dec. 2025, doi: 10.7717/peerj-cs.3438.

[7] J. Shi, J. Zhong, Y. Zhang, B. Xiao, L. Xiao, and Y. Zheng, "A dual attention LSTM lightweight model based on exponential smoothing for remaining useful life prediction," *Reliab. Eng. Syst. Saf.*, vol. 243, p. 109821, Mar. 2024, doi: 10.1016/j.ress.2023.109821.

[8] M. Li, L. Zhu, M. Luo, and T. Ke, "Remaining Useful Life Prediction of Airplane Engine Based on Bidirectional Mamba and Causal Discovery," *Sensors*, vol. 25, no. 11, Jun. 2025, doi: 10.3390/s25113429.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[11] H. Özcan, "Interpretable ensemble remaining useful life prediction enables dynamic maintenance scheduling for aircraft engines," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-23473-2.

[12] M. Ravly Andryan *et al.*, "Komparasi Kinerja Algoritma XGBoost dan Algoritma Support Vector Machine (SVM) untuk Diagnosa Penyakit Kanker Payudara," *Jurnal Informatika dan Komputer*, vol. 6, no. 1, pp. 1–5, 2022.

[13] Muhammad Surono, Muhammad Fadli, Dian Sri Purwanti, and Erliyan Redy Susanto, "Hybrid XGBoost-SVM Model untuk Sistem Pendukung Keputusan dalam Prediksi Penyakit Diabetes," *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 4, no. 3, pp. 443–454, Jun. 2025, doi: 10.55123/insologi.v4i3.5410.

[14] G. Abdurrahman, H. Oktavianto, and M. Sintawati, "Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada Klasifikasi Penyakit Diabetes," *INFORMAL: Informatics Journal*, vol. 7, no. 3, p. 193, Dec. 2022, doi: 10.19184/isj.v7i3.35441.

[15] A. A. Nababan, M. Jannah, M. Aulina, and D. Andrian, "Prediksi Kualitas Udara Menggunakan XGBoost dengan Synthetic Minority Oversampling Technique (SMOTE) Berdasarkan Indeks Standar Pencemaran Udara (ISPU)," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 7, no. 1, 2023.

[16] N. Selle, N. Yudistira, and C. Dewi, "Perbandingan Prediksi Penggunaan Listrik dengan Menggunakan Metode Long Short Term Memory (LSTM) dan Recurrent Neural Network (RNN)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 1, pp. 155–162, 2022, doi: 10.25126/jtiik.202295585.

[17] H. T. A. Simanjuntak, A. Lumbanraja, G. Samosir, and Regita, "Prediksi Single-Step dan Multi-Step Data Cuaca Menggunakan Model Long Short-Term Memory dan Sarima," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 2, pp. 399–410, Apr. 2025, doi: 10.25126/jtiik.2025129444.

[18] Haeruddin, Rinaldo, and Gautama, "Pengembangan Sistem Prediksi Harga Saham Berbasis Web Menggunakan Model LSTM dan CNN–LSTM," 2025, doi: 10.35889/jutisi.v15i1.3518.

[19] N. P. N. P. Dewi, N. K. Kertiasih, and N. L. D. Sintiari, "Modifikasi Fruit Fly Optimiziation Algorithm untuk Optimasi General Regression Neural Network pada Kasus Prediksi Time-Series," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 192–204, Dec. 2022, doi: 10.23887/janapati.v11i3.54521.

[20] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[21] Molnar and Christoph, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Munich: Christoph Molnar, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[22] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1703.01365

Copyright © 2025 The **Autor**, Page 1360
This Journal is licensed under a Creative Commons Attribution 4.0 International License

[23] A. Junaid, A. Iqbal, A. Khan, G. Husnain, A.-R. Ahmad, and M. Al-Naeem, "Engine Failure Prediction on Large-Scale CMAPSS Data Using Hybrid Feature Selection and Imbalance-Aware Learning," *Computers, Materials & Continua*, vol. 0, no. 0, pp. 1–10, 2025, doi: 10.32604/cmc.2025.073189.

Copyright © 2025 The **Autor**, Page 1361
This Journal is licensed under a Creative Commons Attribution 4.0 International License