SMARTICS Journal, Vol.12 No.02, 2026. p165-175
ISSN online: 2476-9754, ISSN print: 2623-0429
DOI : https://doi.org/10.21067/smartics.v12i2.14756

Pengaruh ANOVA F-score terhadap Performa dan Interpretabilitas XGBoost untuk Prediksi Penyakit Jantung

Ubaidillah Ubaidillah[a*], Muhammad Taufik Roseno[b], Hadi Syaputra[c]
[a,b,c]Universitas Sumatera Selatan
[*]correspondence email : ubai@uss.ac.id

Abstract*— Heart disease remains a major health problem and has encouraged the use of machine learning for data-driven prediction. This study investigates the effect of ANOVA F-score feature selection on the predictive performance and interpretability of XGBoost using the UCI Heart Disease Cleveland subset with 303 observations and 13 predictor features. Two scenarios were evaluated: XGBoost using all features and XGBoost using features selected by ANOVA F-score. Model evaluation employed nested stratified cross-validation with 10 outer folds and 5 inner folds, while hyperparameter optimization was performed using Optuna with 100 trials. XGBoost using all features achieved an accuracy of 0.8346 ± 0.0684, an F1-score of 0.8149 ± 0.0759, and a ROC-AUC of 0.9020 ± 0.0447. XGBoost with ANOVA F-score achieved an accuracy of 0.8149 ± 0.0532, an F1-score of 0.7948 ± 0.0527, and a ROC-AUC of 0.9033 ± 0.0381. Feature selection reduced the transformed input space from 22 to an average of 17.8 features, equivalent to a 19.09% reduction. The Wilcoxon signed-rank test showed no statistically significant performance differences at the 0.05 significance level. SHAP analysis identified cp, ca, thal, slope, and oldpeak as the dominant features in both models. Overall, ANOVA F-score produced a more compact feature representation while preserving the main interpretability pattern, but did not significantly improve XGBoost predictive performance.

***Index Terms*—Heart disease; XGBoost; ANOVA F-score; Feature selection; SHAP.**

Abstrak*— Penyakit jantung merupakan salah satu permasalahan kesehatan utama yang mendorong pemanfaatan machine learning untuk mendukung prediksi berbasis data. Penelitian ini menganalisis pengaruh seleksi fitur ANOVA F-score terhadap performa dan interpretabilitas XGBoost menggunakan UCI Heart Disease subset Cleveland yang terdiri atas 303 observasi dan 13 fitur prediktor. Dua skenario dibandingkan, yaitu XGBoost menggunakan seluruh fitur dan XGBoost menggunakan fitur hasil seleksi ANOVA F-score. Evaluasi dilakukan menggunakan nested stratified cross-validation dengan 10 outer fold dan 5 inner fold, sedangkan optimasi hiperparameter dilakukan menggunakan Optuna sebanyak 100 trial. XGBoost dengan seluruh fitur memperoleh accuracy 0,8346 ± 0,0684, F1-score 0,8149 ± 0,0759, dan ROC-AUC 0,9020 ± 0,0447. Model dengan ANOVA F-score memperoleh accuracy 0,8149 ± 0,0532, F1-score 0,7948 ± 0,0527, dan ROC-AUC 0,9033 ± 0,0381. Seleksi fitur mengurangi ruang masukan dari 22 menjadi rata-rata 17,8 fitur atau sebesar 19,09%. Uji Wilcoxon signed-rank menunjukkan bahwa perbedaan performa kedua model tidak signifikan pada tingkat signifikansi 0,05. Analisis SHAP menunjukkan bahwa cp, ca, thal, slope, dan oldpeak tetap menjadi fitur dominan pada kedua model. Secara keseluruhan, ANOVA F-score menghasilkan representasi fitur yang lebih ringkas dengan tetap mempertahankan pola interpretasi utama, tetapi tidak memberikan peningkatan performa prediktif XGBoost yang signifikan.

***Kata Kunci*—Penyakit jantung; XGBoost; ANOVA F-score; Seleksi fitur; SHAP.**

I. PENDAHULUAN

Penyakit kardiovaskular atau *cardiovascular diseases* (CVDs), termasuk penyakit jantung, masih menjadi salah satu permasalahan kesehatan utama di dunia. Penyakit kardiovaskular menyebabkan sekitar 19,8 juta kematian pada tahun 2022 atau sekitar 32% dari seluruh kematian global [1], [2]. Tingginya angka kematian tersebut menunjukkan bahwa deteksi dini dan penilaian risiko penyakit jantung menjadi bagian penting dalam mendukung pemberian intervensi yang tepat. Dalam praktik klinis, diagnosis dilakukan dengan mempertimbangkan gejala, riwayat kesehatan, pemeriksaan fisik, dan hasil pemeriksaan penunjang. Namun, hubungan yang kompleks di antara berbagai faktor klinis dapat menyulitkan proses identifikasi pola penyakit secara cepat dan konsisten. Oleh karena itu, pendekatan berbasis data dapat digunakan untuk membantu mengidentifikasi pola yang berkaitan dengan keberadaan penyakit jantung.

Perkembangan artificial intelligence telah mendorong penerapan machine learning dalam analisis data kesehatan. Metode machine learning dapat mempelajari hubungan antara karakteristik pasien dan status penyakit berdasarkan data historis, kemudian menggunakan pola tersebut untuk menghasilkan prediksi

pada data baru. Berbagai algoritma machine learning telah diterapkan pada bidang kesehatan, salah satunya penelitian Ningrum dkk. yang menerapkan algoritma Support Vector Machine (SVM) dengan optimasi Grid Search Cross-Validation untuk klasifikasi status stunting pada balita. Penelitian tersebut menunjukkan bahwa pendekatan berbasis machine learning dapat mendukung proses klasifikasi dan pengambilan keputusan pada sektor kesehatan [3]. Selain itu, sejumlah algoritma juga telah digunakan untuk memprediksi penyakit jantung, antara lain Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, dan Extreme Gradient Boosting atau XGBoost [4], [5], [6]. Di antara metode tersebut, XGBoost memiliki kemampuan untuk memodelkan hubungan nonlinier dan interaksi antar fitur pada data tabular. XGBoost juga dilengkapi mekanisme regularisasi, penyusutan bobot, dan pengambilan sampel fitur yang dapat membantu mengendalikan kompleksitas model dan mengurangi risiko overfitting [7].

Meskipun XGBoost mampu memanfaatkan banyak fitur secara langsung, penggunaan seluruh fitur belum tentu selalu menghasilkan representasi data yang paling efektif. Sebagian fitur dapat memiliki hubungan individual yang lemah dengan target atau memberikan informasi yang terbatas terhadap proses klasifikasi. Seleksi fitur digunakan untuk mengidentifikasi subset fitur yang lebih relevan sehingga ruang masukan model dapat disederhanakan. Namun, pengurangan fitur tidak selalu menghasilkan peningkatan performa karena fitur yang memiliki relevansi individual rendah masih dapat berkontribusi melalui hubungan nonlinier atau interaksi dengan fitur lainnya [8]. Oleh karena itu, pengaruh seleksi fitur terhadap model perlu dievaluasi secara empiris.

Berbagai pendekatan seleksi fitur telah digunakan dalam penelitian prediksi penyakit jantung. El-Shafiey dkk. mengembangkan GAPSO-RF dengan mengombinasikan Genetic Algorithm, Particle Swarm Optimization, dan Random Forest untuk memilih fitur optimal. Metode tersebut menghasilkan akurasi sebesar 95,6% pada dataset Cleveland dan 91,4% pada dataset Statlog [9]. Yang dan Guan menerapkan seleksi fitur berbasis information gain, penyeimbangan kelas menggunakan SMOTE-ENN, dan XGBoost untuk memprediksi kejadian kardiovaskular [10]. Penelitian lain menggunakan supervised infinite feature selection dan Improved Weighted Random Forest untuk menangani proses pemilihan fitur dan ketidakseimbangan kelas [11]. Hasil-hasil tersebut menunjukkan bahwa strategi pemilihan fitur dapat memengaruhi karakteristik dan performa model, tetapi manfaatnya bergantung pada dataset dan algoritma yang digunakan.

Pathan dkk. secara khusus menganalisis penggunaan ANOVA F-test untuk seleksi fitur pada prediksi penyakit jantung dengan membandingkan sejumlah algoritma machine learning pada dataset CVD dan Framingham [12]. Hasil penelitian tersebut menunjukkan bahwa dampak seleksi fitur terhadap XGBoost tidak konsisten. Pada dataset CVD, akurasi XGBoost meningkat dari 0,65 menjadi 0,69 setelah seleksi fitur, sedangkan pada dataset Framingham akurasinya menurun dari 0,62 menjadi 0,57. Temuan tersebut memperlihatkan bahwa seleksi fitur tidak dapat diasumsikan selalu meningkatkan performa XGBoost dan perlu dievaluasi berdasarkan karakteristik data yang digunakan.

Selain performa prediksi, interpretabilitas menjadi aspek penting dalam penggunaan machine learning pada data kesehatan. Model yang memberikan prediksi tanpa informasi mengenai fitur yang mendasari keputusan akan lebih sulit dipahami oleh pengguna. SHapley Additive exPlanations (SHAP) dapat digunakan untuk mengukur arah dan besarnya kontribusi setiap fitur terhadap keluaran model, baik secara global maupun pada prediksi individual [13]. Mushtaq et al. mengombinasikan XGBoost, Recursive Feature Elimination, SHAP, dan algoritma Peter–Clark pada UCI Heart Disease subset Cleveland. Model tersebut menghasilkan akurasi 91%, F1-score 0,90, dan ROC-AUC 0,95. Analisis SHAP menunjukkan bahwa fitur seperti tipe nyeri dada, thal, jumlah pembuluh darah utama, dan oldpeak memberikan kontribusi penting terhadap prediksi model [14].

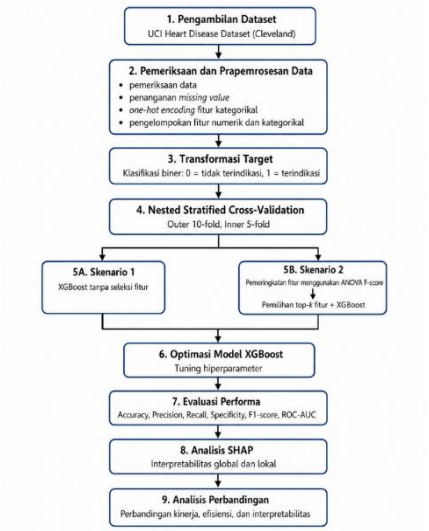
Meskipun penelitian sebelumnya telah mengkaji XGBoost, seleksi fitur, dan SHAP, analisis mengenai pengaruh ANOVA F-score terhadap performa sekaligus struktur interpretabilitas XGBoost pada UCI Heart Disease subset Cleveland masih terbatas. Penelitian yang menggunakan ANOVA F-test belum secara khusus mengevaluasi perubahan interpretabilitas XGBoost setelah fitur dikurangi, sedangkan penelitian XGBoost dan SHAP pada dataset Cleveland menggunakan strategi seleksi fitur yang berbeda seperti Recursive Feature Elimination. Dengan demikian, masih terdapat kebutuhan untuk melakukan perbandingan terkontrol antara XGBoost yang menggunakan seluruh fitur dan XGBoost yang menggunakan fitur hasil seleksi ANOVA F-score.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan menganalisis pengaruh seleksi fitur ANOVA F-score terhadap performa dan interpretabilitas XGBoost dalam prediksi penyakit jantung. Dua skenario eksperimen dibandingkan, yaitu XGBoost dengan seluruh fitur sebagai baseline dan XGBoost menggunakan fitur dengan peringkat ANOVA F-score tertinggi. Evaluasi dilakukan melalui nested stratified cross-validation dan optimasi hiperparameter menggunakan Optuna, sedangkan performa diukur menggunakan accuracy, precision, recall, specificity, F1-score, dan ROC-AUC. Uji Wilcoxon signed-rank digunakan untuk mengevaluasi signifikansi perbedaan performa, sementara SHAP digunakan untuk membandingkan pola kontribusi fitur secara global dan lokal. Melalui rancangan tersebut, penelitian ini tidak hanya mengevaluasi apakah seleksi fitur meningkatkan performa XGBoost, tetapi juga menilai sejauh mana pengurangan fitur memengaruhi kompleksitas masukan dan pola interpretasi model.

II. METODE PENELITIAN

A. Desain Penelitian

Penelitian menggunakan desain eksperimen komparatif dengan dua skenario, yaitu XGBoost menggunakan seluruh fitur sebagai baseline dan XGBoost dengan seleksi fitur ANOVA F-score. Kedua skenario menggunakan dataset, pembagian fold, prapemrosesan, ruang pencarian hiperparameter, dan metrik evaluasi yang sama. Tahapan penelitian meliputi prapemrosesan, transformasi target, nested stratified cross-validation, seleksi fitur, optimasi XGBoost, evaluasi performa, serta analisis SHAP sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian yang Diusulkan

B. Dataset Penelitian

Penelitian ini menggunakan Heart Disease Dataset yang diperoleh dari UCI Machine Learning Repository [15]. UCI Heart Disease Dataset mencakup empat basis data yang berasal dari Cleveland, Hungary, Switzerland, dan VA Long Beach. Penelitian ini secara khusus menggunakan basis data Cleveland yang terdiri atas 303 observasi dan 13 fitur prediktor. Penelitian ini menggunakan subset Cleveland karena merupakan salah satu dataset yang paling banyak digunakan dalam penelitian klasifikasi penyakit jantung.

Dataset Cleveland terdiri atas 303 observasi pasien, 13 fitur prediktor, dan satu atribut target. Fitur prediktor mencakup karakteristik demografis, hasil pemeriksaan klinis, dan hasil pemeriksaan jantung. Deskripsi fitur ditunjukkan pada Tabel 1.

Tabel 1. Deskripsi Fitur UCI Heart Disease Subset Cleveland

No.	Fitur	Deskripsi
1	age	Usia pasien dalam tahun
2	sex	Jenis kelamin pasien
3	cp	Tipe nyeri dada
4	trestbps	Tekanan darah saat istirahat
5	chol	Kadar kolesterol serum

No.	Fitur	Deskripsi
6	fbs	Kondisi gula darah puasa
7	restecg	Hasil elektrokardiografi saat istirahat
8	thalach	Denyut jantung maksimum yang dicapai
9	exang	Angina yang dipicu oleh aktivitas fisik
10	oldpeak	Depresi segmen ST akibat aktivitas
11	slope	Kemiringan segmen ST pada aktivitas puncak
12	ca	Jumlah pembuluh darah utama yang terdeteksi
13	thal	Hasil pemeriksaan dengan kategori normal, fixed defect, dan reversible defect
14	num	Target diagnosis penyakit jantung

Atribut target num pada dataset asli memiliki nilai 0 sampai 4. Penelitian ini mengubah target tersebut menjadi klasifikasi biner. Nilai 0 ditetapkan sebagai kelas negatif atau tidak terindikasi penyakit jantung, sedangkan nilai 1 sampai 4 ditetapkan sebagai kelas positif atau terindikasi penyakit jantung. Transformasi tersebut dilakukan karena penelitian berfokus pada prediksi keberadaan penyakit jantung dan bukan pada pengelompokan tingkat keparahannya.

C. Prapemrosesan Data

Dataset Cleveland terdiri atas 303 observasi, 13 fitur prediktor, dan atribut target num. Target ditransformasikan menjadi klasifikasi biner dengan num=0 sebagai kelas negatif dan num=1–4 sebagai kelas positif. Nilai hilang ditemukan pada ca dan thal dan ditangani menggunakan imputasi modus yang dihitung hanya dari data pelatihan setiap fold.

Fitur age, trestbps, chol, thalach, dan oldpeak diperlakukan sebagai numerik kontinu, sedangkan ca sebagai numerik diskret. Fitur sex, fbs, dan exang dipertahankan sebagai biner, sedangkan cp, restecg, slope, dan thal ditransformasikan menggunakan one-hot encoding. Seluruh prapemrosesan ditempatkan di dalam pipeline untuk mencegah data leakage. Standardisasi dan SMOTE tidak digunakan karena model berbasis pohon tidak bergantung pada skala fitur dan distribusi kelas tidak menunjukkan ketidakseimbangan ekstrem.

D. Seleksi Fitur ANOVA F-score

Seleksi fitur dilakukan menggunakan ANOVA F-score melalui fungsi f_classif dan SelectKBest. Metode tersebut termasuk pendekatan *filter-based feature selection* karena menghitung hubungan setiap fitur dengan kelas target tanpa bergantung pada algoritma klasifikasi tertentu [12].

ANOVA F-score mengukur rasio variasi nilai fitur antarkelas terhadap variasi nilai fitur di dalam kelas. Fitur yang memiliki perbedaan rata-rata besar antara kelas positif dan negatif serta memiliki variasi intrakelas yang rendah akan memperoleh nilai F yang lebih tinggi. Nilai F untuk setiap fitur dihitung menggunakan Persamaan (1).

$$F_j = \frac{MSB_j}{MSW_j}$$

Dengan (F_j) merupakan ANOVA F-score fitur ke-(j), (MSB_j) merupakan variasi antarkelas, dan (MSW_j) merupakan variasi di dalam kelas. Semakin besar nilai (F_j), semakin besar kemampuan fitur dalam membedakan kelas.

ANOVA F-score dihitung setelah proses imputasi dan *one-hot encoding*. Setiap kolom hasil transformasi diberi nilai F dan diurutkan dari skor tertinggi sampai terendah. Jumlah fitur terpilih dinyatakan sebagai (k). Nilai (k) tidak ditentukan secara subjektif, tetapi dipilih berdasarkan hasil validasi silang pada data pelatihan. Kandidat nilai (k) diuji mulai dari satu fitur sampai seluruh fitur hasil transformasi. Nilai yang menghasilkan ROC-AUC rata-rata tertinggi pada validasi bagian dalam dipilih sebagai jumlah fitur optimal.

Proses penghitungan ANOVA F-score dilakukan secara terpisah pada setiap *fold* pelatihan. Data pengujian tidak digunakan untuk menghitung nilai F maupun menentukan jumlah fitur. Prosedur tersebut mencegah informasi kelas dari data pengujian memengaruhi pemilihan fitur.

Kolom-kolom hasil *one-hot encoding* yang berasal dari variabel klinis yang sama tetap dipertahankan secara terpisah dalam proses seleksi. Pada tahap interpretasi SHAP, kontribusi kolom-kolom tersebut digabungkan kembali berdasarkan variabel asalnya agar hasil analisis dapat dijelaskan pada tingkat fitur klinis.

E. Pemodelan XGBoost

XGBoost digunakan sebagai algoritma klasifikasi dengan fungsi objektif binary:logistic. Probabilitas kelas positif dikonversi menjadi kelas menggunakan ambang 0,5. Optimasi hiperparameter dilakukan menggunakan Optuna pada inner cross-validation dengan rata-rata ROC-AUC sebagai fungsi objektif. Ruang pencarian hiperparameter ditunjukkan pada Tabel 2. Sebanyak 100 trial dilakukan pada setiap outer fold dengan random seed 42 untuk meningkatkan keterulangan eksperimen.

Tabel 2. Ruang Pencarian Hiperparameter Xgboost

Hiperparameter	Rentang pencarian
n_estimators	50–500
max_depth	2–8
learning_rate	0,01–0,30
min_child_weight	1–10
subsample	0,60–1,00
colsample_bytree	0,60–1,00
gamma	0–5
reg_alpha	0–5
reg_lambda	0,1–10

F. Desain Eksperimen dan Validasi

UCI Heart Disease subset Cleveland hanya memiliki 303 observasi. Penggunaan satu kali pembagian data latih dan data uji dapat menghasilkan estimasi performa yang sangat dipengaruhi oleh komposisi pembagian data. Oleh karena itu, penelitian menggunakan *nested stratified cross-validation*.

Pada validasi bagian luar digunakan *stratified 10-fold cross-validation*. Data dibagi menjadi sepuluh *fold* dengan proporsi kelas positif dan negatif yang relatif sama. Sembilan *fold* digunakan untuk pelatihan dan satu *fold* digunakan sebagai data pengujian. Proses tersebut diulang sampai seluruh *fold* pernah digunakan sebagai data pengujian.

Pada setiap data pelatihan bagian luar, digunakan *stratified 5-fold cross-validation* sebagai validasi bagian dalam. Validasi bagian dalam digunakan untuk menentukan hiperparameter terbaik pada model *baseline*. Pada model dengan seleksi fitur, validasi bagian dalam digunakan secara bersamaan untuk menentukan nilai (k) dan hiperparameter XGBoost terbaik.

Eksperimen terdiri atas dua skenario sebagaimana ditunjukkan pada Tabel 3.

Tabel 3. Skenario Eksperimen

Skenario	Seleksi fitur	Model	Keterangan
S1	Tidak digunakan	XGBoost	Menggunakan seluruh fitur hasil prapemrosesan
S2	ANOVA F-score	XGBoost	Menggunakan top-(k) fitur hasil seleksi

Data pengujian pada setiap *outer fold* tidak digunakan dalam proses imputasi, pengodean, penghitungan ANOVA F-score, penentuan nilai (k), maupun optimasi hiperparameter. Data tersebut hanya digunakan satu kali untuk mengevaluasi model akhir pada *fold* yang bersangkutan.

G. Evaluasi dan Analisis Statistik

Performa model dievaluasi menggunakan confusion matrix, accuracy, precision, recall, specificity, F1-score, dan ROC-AUC. Hasil disajikan dalam bentuk rata-rata ± standar deviasi dari sepuluh outer fold. ROC-AUC digunakan sebagai metrik utama untuk mengevaluasi kemampuan diskriminasi model, sedangkan recall diperhatikan karena berkaitan dengan jumlah false negative. Perbedaan performa kedua model diuji menggunakan Wilcoxon signed-rank test terhadap accuracy, recall, F1-score, dan ROC-AUC dari outer fold yang sama dengan tingkat signifikansi 0,05.

H. Analisis SHAP

Interpretabilitas model dianalisis menggunakan SHAP TreeExplainer pada data pengujian setiap outer fold. Analisis global menggunakan rata-rata nilai absolut SHAP dan beeswarm plot untuk mengidentifikasi fitur dominan, sedangkan analisis lokal menggunakan waterfall plot untuk menjelaskan kontribusi fitur terhadap satu prediksi individual. Untuk fitur yang mengalami one-hot encoding, kontribusi kolom yang berasal dari variabel yang sama diagregasikan kembali sehingga interpretasi dapat disajikan berdasarkan 13 fitur asli. Perbandingan kedua skenario dilakukan berdasarkan jumlah fitur, urutan kepentingan fitur, kesamaan fitur dominan, serta perubahan kontribusi SHAP pada tingkat global dan lokal.

III. HASIL DAN PEMBAHASAN

A. Karakteristik Dataset dan Hasil Prapemrosesan

Setelah transformasi target, sebanyak 164 observasi (54,1%) berada pada kelas negatif dan 139 observasi (45,9%) pada kelas positif. Nilai hilang ditemukan pada ca dan thal dan ditangani melalui imputasi modus di dalam pipeline. Proses one-hot encoding pada fitur kategorikal mengubah 13 fitur asli menjadi 22 kolom masukan. Seluruh parameter prapemrosesan hanya dipelajari dari data pelatihan pada setiap fold untuk mencegah data leakage.

B. Hasil Seleksi Fitur ANOVA F-score

Jumlah fitur optimal berbeda pada setiap outer fold, dengan nilai k antara 13–21, rata-rata 17,8, dan median 18 dari 22 kolom hasil transformasi. Dengan demikian, ANOVA F-score menghasilkan rata-rata reduksi ruang masukan sebesar 19,09%.

Pada tingkat fitur asli, age, ca, cp, exang, thalach, oldpeak, sex, slope, dan thal terpilih pada seluruh outer fold. Fitur trestbps, restecg, dan chol masing-masing terpilih pada 90%, 80%, dan 50% fold, sedangkan fbs tidak pernah terpilih. Variasi tersebut menunjukkan bahwa subset fitur optimal dipengaruhi komposisi data pelatihan dan bahwa relevansi univariat suatu fitur tidak selalu mencerminkan manfaatnya dalam interaksi yang dipelajari XGBoost.

C. Perbandingan Performa Model

Evaluasi dilakukan menggunakan nested stratified cross-validation dengan 10 outer fold dan 5 inner fold. Pada setiap outer fold, optimasi hiperparameter dilakukan menggunakan Optuna dengan 100 trial. ROC-AUC rata-rata pada validasi bagian dalam digunakan sebagai fungsi objektif. Prosedur yang sama diterapkan pada kedua skenario sehingga perbedaan performa terutama merepresentasikan pengaruh penggunaan seleksi fitur ANOVA F-score. Hasil evaluasi ditunjukkan pada Tabel 4.

Tabel 4. Perbandingan Performa Xgboost Dengan Dan Tanpa Anova F-Score

Metrik	XGBoost Seluruh Fitur	XGBoost + ANOVA F-score	Selisih
Accuracy	0,8346 ± 0,0684	0,8149 ± 0,0532	−0,0197
Precision	0,8547 ± 0,1156	0,8289 ± 0,1104	−0,0258
Recall	0,7912 ± 0,0982	0,7764 ± 0,0809	−0,0148
Specificity	0,8721 ± 0,1128	0,8474 ± 0,1178	−0,0246
F1-score	0,8149 ± 0,0759	0,7948 ± 0,0527	−0,0201
ROC-AUC	0,9020 ± 0,0447	0,9033 ± 0,0381	+0,0013

Model XGBoost yang menggunakan seluruh fitur menghasilkan nilai yang lebih tinggi pada accuracy, precision, recall, specificity, dan F1-score. Accuracy berkurang sebesar 0,0197 setelah ANOVA F-score diterapkan, sedangkan precision, recall, specificity, dan F1-score masing-masing mengalami penurunan sebesar 0,0258; 0,0148; 0,0246; dan 0,0201.

Pola berbeda terlihat pada ROC-AUC. Model dengan ANOVA F-score menghasilkan ROC-AUC sebesar 0,9033 ± 0,0381, sedikit lebih tinggi dibandingkan model seluruh fitur sebesar 0,9020 ± 0,0447. Namun, selisih sebesar 0,0013 sangat kecil sehingga peningkatan tersebut belum dapat dianggap sebagai keunggulan performa yang berarti tanpa memperhatikan hasil pengujian statistik.

Hasil ini menunjukkan bahwa pengurangan fitur tidak secara otomatis meningkatkan kinerja XGBoost. Meskipun ANOVA menghilangkan sejumlah fitur dengan relevansi individual rendah, XGBoost dapat

memperoleh manfaat dari kombinasi dan interaksi fitur yang tidak sepenuhnya ditangkap melalui analisis univariat. Oleh karena itu, seleksi fitur perlu dievaluasi secara empiris dan tidak dapat diasumsikan selalu menghasilkan peningkatan performa.

D. Hasil Uji Statistik

Uji Wilcoxon signed-rank menunjukkan bahwa perbedaan accuracy (p=0,0625), recall (p=0,5313), F1-score (p=0,1563), dan ROC-AUC (p=0,6719) tidak signifikan pada α=0,05. Dengan demikian, belum terdapat bukti statistik yang cukup bahwa seleksi fitur ANOVA F-score menghasilkan perubahan performa XGBoost yang signifikan.

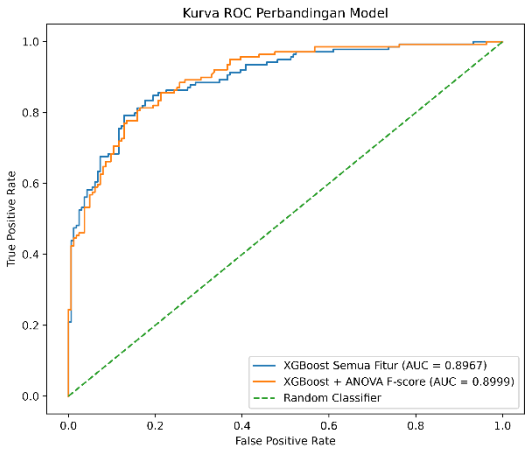
E. Analisis Confusion Matrix dan Kurva ROC

Selain menggunakan rata-rata performa outer fold, evaluasi dilakukan terhadap prediksi out-of-fold seluruh 303 observasi. Setiap observasi diprediksi ketika berada pada data pengujian sehingga hasil yang digunakan tidak berasal dari prediksi terhadap data yang digunakan untuk melatih model.

Pada prediksi *out-of-fold*, model seluruh fitur menghasilkan TN=143, FP=21, FN=29, dan TP=110, sedangkan model ANOVA menghasilkan TN=139, FP=25, FN=31, dan TP=108. XGBoost dengan seluruh fitur menghasilkan 143 true negative dan 110 true positive. Sementara itu, terdapat 21 false positive dan 29 false negative. Model dengan ANOVA F-score menghasilkan 139 true negative, 108 true positive, 25 false positive, dan 31 false negative.

Secara numerik, model seluruh fitur menghasilkan jumlah klasifikasi benar yang sedikit lebih tinggi. Model tersebut juga menghasilkan false negative lebih rendah, yaitu 29 dibandingkan 31 pada model ANOVA. Temuan ini konsisten dengan nilai recall yang lebih tinggi pada model seluruh fitur, yaitu 0,7912 dibandingkan 0,7764. Dalam konteks klasifikasi kesehatan, false negative menjadi perhatian karena menunjukkan observasi positif yang tidak berhasil dikenali oleh model.

Evaluasi berdasarkan kurva ROC terhadap seluruh prediksi out-of-fold menghasilkan AUC sebesar 0,8967 pada XGBoost seluruh fitur dan 0,8999 pada XGBoost dengan ANOVA F-score. Kedua kurva terlihat relatif berdekatan sebagaimana ditunjukkan pada Gambar 2.



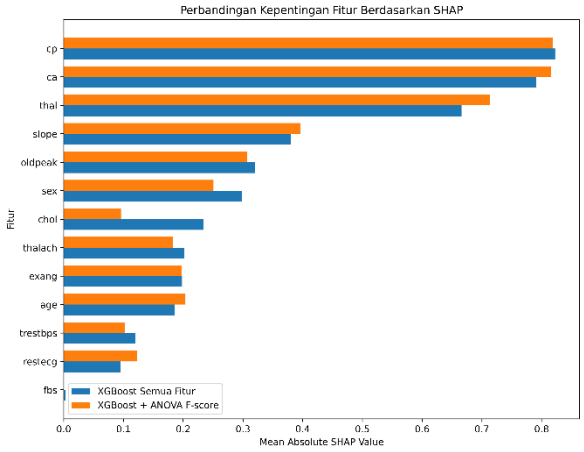
Gambar 2. Kurva ROC perbandingan XGBoost dengan dan tanpa seleksi fitur ANOVA F-score

Nilai AUC out-of-fold tersebut sedikit berbeda dari rata-rata ROC-AUC pada outer fold karena keduanya diperoleh melalui prosedur perhitungan yang berbeda. ROC-AUC pada Tabel 4 merupakan rata-rata AUC dari sepuluh outer fold, sedangkan ROC pada Gambar 2 dihitung satu kali setelah seluruh probabilitas out-of-fold digabungkan. Meskipun demikian, kedua pendekatan memberikan pola yang sama, yaitu kemampuan diskriminasi kedua model sangat berdekatan dan model ANOVA hanya menunjukkan peningkatan AUC yang sangat kecil.

F. Analisis Interpretabilitas Global Menggunakan SHAP

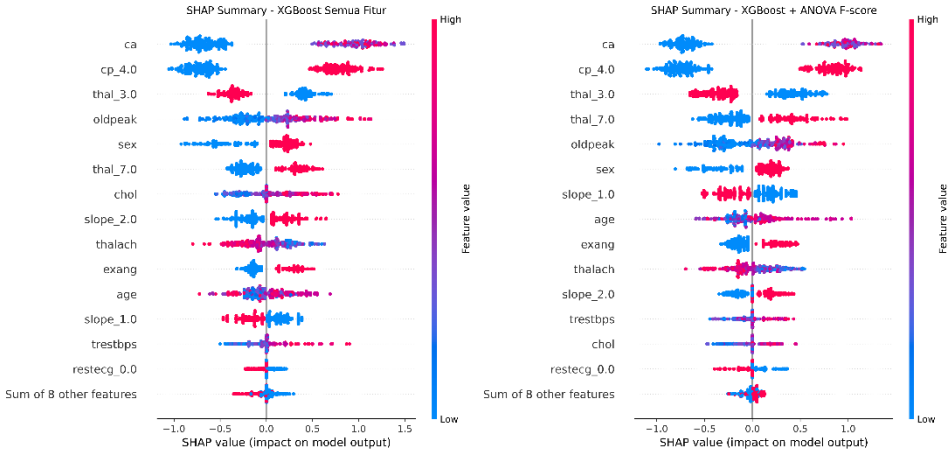
Lima fitur dengan kontribusi tertinggi pada kedua model adalah cp, ca, thal, slope, dan oldpeak. Pada model seluruh fitur, cp memiliki mean |SHAP| tertinggi sebesar 0,8224 dan tetap dominan setelah ANOVA

sebesar 0,8190. Kontribusi ca meningkat dari 0,7904 menjadi 0,8161 dan thal dari 0,6659 menjadi 0,7140. Sebaliknya, chol turun dari 0,2341 menjadi 0,0959, sedangkan fbs berubah dari 0,0026 menjadi 0, konsisten dengan hasil seleksi ANOVA.



Gambar 3. Perbandingan kepentingan fitur berdasarkan mean absolute SHAP

Visualisasi SHAP beeswarm memberikan informasi tambahan mengenai arah kontribusi setiap fitur.



Gambar 4. SHAP summary plot: (a) XGBoost seluruh fitur dan (b) XGBoost + ANOVA F-score

Pada kedua model, kategori cp=4 cenderung menghasilkan nilai SHAP positif sehingga mendorong keluaran model menuju kelas terindikasi. Kategori thal=3 cenderung menghasilkan kontribusi negatif, sedangkan thal=7 lebih banyak menghasilkan kontribusi positif. Nilai oldpeak yang lebih tinggi juga cenderung memberikan kontribusi positif terhadap prediksi kelas terindikasi. Sebaliknya, nilai thalach yang lebih tinggi lebih banyak ditemukan pada sisi SHAP negatif. Fitur exang=1 juga cenderung memberikan kontribusi ke arah kelas positif.

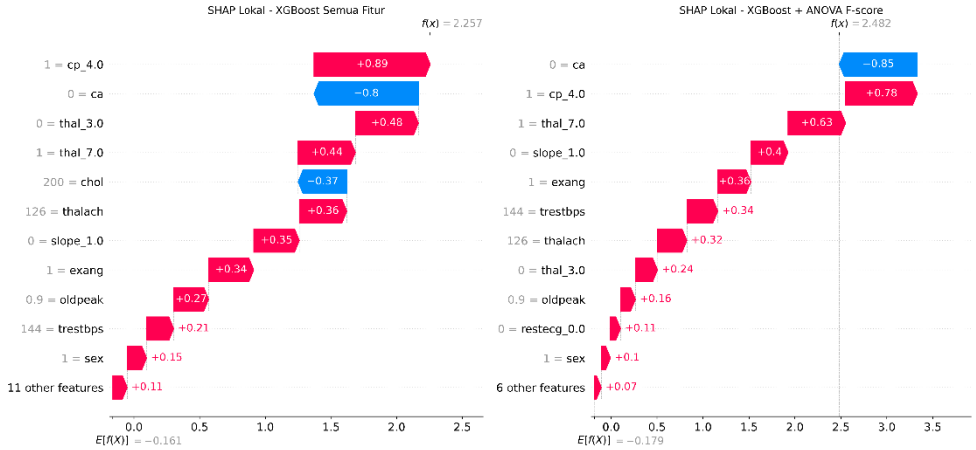
Pola tersebut merupakan hubungan yang dipelajari model dari dataset dan tidak menunjukkan hubungan sebab-akibat secara klinis. SHAP digunakan untuk menjelaskan perilaku model, bukan untuk menyatakan bahwa suatu fitur secara langsung menyebabkan penyakit. Hasil SHAP global juga memiliki kesesuaian dengan penelitian Mushtaq et al. [13], yang mengidentifikasi fitur seperti tipe nyeri dada, thal, jumlah pembuluh darah utama, dan oldpeak sebagai fitur penting pada model XGBoost berbasis UCI Cleveland. Kesamaan tersebut menunjukkan bahwa beberapa fitur dominan relatif konsisten meskipun strategi seleksi fitur dan rancangan eksperimennya berbeda.

G. Analisis Interpretabilitas Lokal Menggunakan SHAP

Selain analisis global, SHAP digunakan untuk menjelaskan satu prediksi individual. Observasi dipilih dari kelompok aktual positif yang berhasil diprediksi positif oleh kedua model. Untuk menghindari

pemilihan kasus ekstrem, observasi yang memiliki rata-rata probabilitas prediksi kedua model paling dekat dengan median kelompok true positive digunakan sebagai contoh representatif.

Observasi yang terpilih berada pada indeks 207 dan memiliki karakteristik utama berupa usia 50 tahun, sex=1, cp=4, trestbps=144, chol=200, fbs=0, restecg=2, thalach=126, exang=1, oldpeak=0,9, slope=2, ca=0, dan thal=7. Hasil SHAP lokal pada XGBoost seluruh fitur ditunjukkan pada Gambar 5 (a).



Gambar 5. SHAP lokal pada observasi representatif: (a) XGBoost seluruh fitur dan (b) XGBoost + ANOVA F-score

Pada model seluruh fitur, thal memberikan kontribusi positif terbesar sebesar 0,9161, diikuti cp sebesar 0,8853. Fitur slope, thalach, exang, oldpeak, trestbps, restecg, dan sex juga mendorong prediksi menuju kelas positif. Sebaliknya, ca=0 memberikan kontribusi negatif terbesar sebesar −0,7951, diikuti chol sebesar −0,3709 dan age sebesar −0,0746. Hasil model dengan ANOVA F-score ditunjukkan pada Gambar 5 (b).

Pola utama pada model ANOVA relatif serupa. thal dan cp tetap menjadi pendorong positif utama dengan nilai SHAP masing-masing sebesar 0,8670 dan 0,7810. Fitur slope, exang, trestbps, thalach, oldpeak, restecg, dan sex juga memberikan kontribusi positif. Sementara itu, ca=0 menghasilkan kontribusi negatif sebesar −0,8514.

Perubahan penting terlihat pada chol. Pada model seluruh fitur, chol=200 memberikan kontribusi sebesar −0,3709, sedangkan pada model ANOVA kontribusinya menjadi nol karena fitur tersebut tidak dipertahankan pada outer fold yang digunakan untuk menjelaskan observasi tersebut. Kondisi ini memperlihatkan secara langsung bagaimana seleksi fitur dapat mengubah struktur penjelasan lokal, meskipun prediksi akhir kedua model tetap berada pada kelas yang sama.

Output mentah model pada visualisasi waterfall meningkat dari (f(x)=2,257) pada XGBoost seluruh fitur menjadi (f(x)=2,482) pada model ANOVA. Nilai tersebut merupakan keluaran dalam skala raw margin atau log-odds, bukan probabilitas secara langsung. Perbandingan waterfall menunjukkan bahwa penghilangan sejumlah fitur menyebabkan kontribusi fitur yang dipertahankan didistribusikan kembali, tetapi struktur utama keputusan model tetap didominasi oleh fitur yang relatif sama.

H. Pembahasan Komparatif

Secara keseluruhan, ANOVA F-score mengurangi ruang masukan sebesar 19,09%, tetapi tidak meningkatkan performa XGBoost secara konsisten. Model seluruh fitur menghasilkan accuracy, precision, recall, specificity, dan F1-score yang lebih tinggi, sedangkan model ANOVA hanya meningkatkan ROC-AUC sebesar 0,0013. Seluruh perbedaan yang diuji menggunakan tidak signifikan pada α=0,05.

Temuan tersebut sejalan dengan Pathan et al. [11], yang menunjukkan bahwa pengaruh ANOVA F-test terhadap XGBoost berbeda antar-dataset. Salah satu penjelasannya adalah ANOVA mengevaluasi relevansi fitur secara individual, sedangkan XGBoost dapat memanfaatkan hubungan nonlinier dan interaksi antarfitur. Dengan demikian, fitur dengan relevansi univariat rendah masih dapat memberikan informasi tambahan ketika digunakan bersama fitur lainnya.

Analisis SHAP menunjukkan bahwa cp, ca, thal, slope, dan oldpeak tetap menjadi fitur dominan setelah seleksi. Temuan ini juga konsisten dengan Mushtaq et al. [13], yang mengidentifikasi beberapa fitur serupa sebagai fitur penting pada dataset Cleveland. Dengan demikian, ANOVA F-score pada penelitian ini lebih bermanfaat dalam menyederhanakan ruang fitur daripada meningkatkan performa prediktif.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa XGBoost dengan seluruh fitur menghasilkan accuracy 0,8346 ± 0,0684, F1-score 0,8149 ± 0,0759, dan ROC-AUC 0,9020 ± 0,0447, sedangkan XGBoost dengan ANOVA F-score memperoleh accuracy 0,8149 ± 0,0532, F1-score 0,7948 ± 0,0527, dan ROC-AUC 0,9033 ± 0,0381. ANOVA F-score mengurangi ruang masukan sebesar 19,09%, tetapi uji Wilcoxon menunjukkan bahwa perbedaan performa kedua model tidak signifikan pada α=0,05.

Analisis SHAP menunjukkan bahwa cp, ca, thal, slope, dan oldpeak tetap menjadi fitur dominan pada kedua model. Dengan demikian, ANOVA F-score lebih memberikan manfaat berupa representasi fitur yang lebih ringkas daripada peningkatan performa prediktif. Penelitian selanjutnya dapat mengevaluasi metode seleksi fitur lain dan dataset yang lebih besar untuk menguji konsistensi temuan.

V. DAFTAR PUSTAKA

[1] "Cardiovascular diseases (CVDs)." Diakses: 11 Agustus 2026. [Daring]. Tersedia pada: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] B. A. Stark *dkk.*, "Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990-2023," *JACC*, vol. 86, no. 22, hlm. 2167–2243, Des 2025, doi: 10.1016/j.jacc.2025.08.015.

[3] K. Budholiya, S. K. Shrivastava, dan V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, hlm. 4514–4523, Jul 2022, doi: 10.1016/j.jksuci.2020.10.013.

[4] P. Srinivas dan R. Katarya, "hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," *Biomed. Signal Process. Control*, vol. 73, Mar 2022, doi: 10.1016/j.bspc.2021.103456.

[5] A. S. Sunge, Amali, Z. Y. Ahmad Turmudi, D. K. Pramudito, A. Badruzzaman, dan Purwanto, "The model interpretability on SHAP and comparison classification selection feature for heart disease prediction," dalam *Procedia Computer Science*, Elsevier B.V., 2024, hlm. 210–219. doi: 10.1016/j.procs.2024.10.245.

[6] T. Chen dan C. Guestrin, "XGBoost," dalam *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Agu 2016, hlm. 785–794. doi: 10.1145/2939672.2939785.

[7] S. Ningrum, M. A. Barata, dan N. Mahmudah, "Klasifikasi Status Stunting Pada Balita di Kecamatan Singgahan dengan Algoritma," *SMARTICS Journal*, vol. 11, no. 1, hlm. 28–33, Apr 2025, doi: 10.21067/smartics.v11i1.11862.

[8] Guyon Isabelle dan Elisseeff André, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, Mar 2003, doi: 10.5555/944919.944968.

[9] M. G. El-Shafiey, A. Hagag, E. S. A. El-Dahshan, dan M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," *Multimed. Tools Appl.*, vol. 81, no. 13, hlm. 18155–18179, Mei 2022, doi: 10.1007/s11042-022-12425-x.

[10] J. Yang dan J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information*, vol. 13, no. 10, hlm. 475, Okt 2022, doi: 10.3390/info13100475.

[11] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, dan H. M. Gheni, "Improving the Heart Disease Detection and Patients' Survival Using Supervised Infinite Feature Selection and Improved Weighted Random Forest," *IEEE Access*, vol. 10, hlm. 67363–67372, 2022, doi: 10.1109/ACCESS.2022.3185129.

[12] M. S. Pathan, A. Nag, M. M. Pathan, dan S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, hlm. 100060, Nov 2022, doi: 10.1016/J.HEALTH.2022.100060.

[13] S. M. Lundberg dan S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, hlm. 4766–4775, Mei 2017, doi: https://doi.org/10.48550/arXiv.1705.07874.

[14] M. Mushtaq, Z. Mahmood, A. A. Butt, M. Y. Shabir, dan M. N. U. Khan, "Causal and Explainable Machine Learning Framework for Heart Disease Prediction using XGBoost and SHAP," *Journal of Computing and Biomedical Informatics*, vol. 10, no. 1, 2025, doi: 10.56979/1001/2025/1102.
[15] "Heart Disease - UCI Machine Learning Repository." Diakses: 11 Agustus 2026. [Daring]. Tersedia pada: https://archive.ics.uci.edu/dataset/45/heart+disease