Berajah Journal
Published by: CV Lafadz Jaya | ISSN: 2797-1805 | E-ISSN: 2797-1813

Volume 6 No.4 (2026)
Page: 2194 – 2204
https://ojs.berajah.com/index.php/go/

PENGEMBANGAN SISTEM *AI HEALTH AGENT* UNTUK PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN *XGBOOST* DAN *EXPLAINABLE AI*

DEVELOPMENT OF AN AI HEALTH AGENT SYSTEM FOR PREDICTING HEART DISEASE USING XGBOOST AND EXPLAINABLE AI

I Putu Mahendra Putra[1]*, Adie Wahyudi Oktavia Gama[2]
[1,2]Teknologi Informasi, Fakultas Teknik dan Informatika, Universitas Pendidikan Nasional
***Email Correspondence**: worksheet.mahendra@gmail.com

Received: 15-07-2026 | **Revised**: 20-07-2026 | **Accepted**: 30-07-2026 | **Published**: 13-08-2026

Abstract

Opaque model decisions remain a practical barrier to the use of machine learning in heart disease screening. This study presents HeartXplain, a web-based AI Health Agent designed as a Clinical Decision Support System for early screening. The system combines XGBoost, Particle Swarm Optimization (PSO), SHAP, LIME, and interpretative narratives generated by a Large Language Model (LLM). The model was trained on 69,105 records represented by 19 clinical and engineered features. PSO optimized a recall-oriented objective, followed by decision-threshold adjustment. At a threshold of 0.4997, the model achieved 71.01% accuracy, 80.08% recall, 67.39% precision, a 73.19% F1-score, 79.29% AUC-ROC, and 78.05% PR-AUC; false negatives decreased from 1,367 to 1,360. SHAP and LIME exposed global and patient-level feature contributions, while the LLM translated those outputs into a more readable narrative. MAPIE reached 95.01% empirical coverage and 83.02% accuracy on singleton sets. Fairlearn indicated minor sex-based differences but wider gaps across age groups. HeartXplain therefore provides a transparent supporting layer for early screening, not a substitute for clinical examination, diagnosis, or professional judgment. External validation on real clinical data remains necessary.

Keywords: *Heart Disease Prediction, XGBoost, Explainable Artificial Intelligence (XAI), Particle Swarm Optimization (PSO), Clinical Decision Support System (CDSS).*

Abstrak

Keputusan model machine learning yang sulit ditelusuri masih menjadi kendala dalam skrining risiko penyakit jantung. Penelitian ini mengembangkan HeartXplain, AI Health Agent berbasis web yang dirancang sebagai Clinical Decision Support System untuk skrining awal. Sistem menggabungkan XGBoost, Particle Swarm Optimization (PSO), SHAP, LIME, dan narasi interpretatif berbasis Large Language Model (LLM). Model dilatih menggunakan 69.105 observasi yang direpresentasikan oleh 19 fitur klinis dan fitur turunan. PSO mengoptimasi fungsi objektif berorientasi recall, lalu decision threshold disesuaikan. Pada threshold 0,4997, model menghasilkan accuracy 71,01%, recall 80,08%, precision 67,39%, F1-score 73,19%, AUC-ROC 79,29%, dan PR-AUC 78,05%; jumlah false negative turun dari 1.367 menjadi 1.360. SHAP dan LIME memperlihatkan kontribusi fitur pada tingkat global dan individual, sedangkan LLM menyajikannya sebagai narasi yang lebih mudah dibaca. MAPIE mencapai empirical coverage 95,01% dan akurasi 83,02% pada singleton set. Fairlearn menunjukkan perbedaan kecil berdasarkan jenis kelamin, tetapi kesenjangan antarkelompok usia lebih besar. HeartXplain memberikan lapisan pendukung yang lebih transparan untuk skrining awal, bukan pengganti pemeriksaan, diagnosis, atau pertimbangan tenaga medis. Validasi eksternal pada data klinis nyata masih diperlukan.

Kata Kunci: Prediksi Penyakit Jantung, XGBoost, Explainable Artificial Intelligence (XAI), Particle Swarm Optimization (PSO), Clinical Decision Support System (CDSS).

DOI: https://doi.org/10.47353/bj.v6i4.687

PENDAHULUAN

Penyakit kardiovaskular atau Cardiovascular Disease (CVD) tetap menjadi penyebab utama kematian di dunia. World Health Organization mencatat sekitar 19,8 juta kematian akibat CVD pada 2022, setara dengan 32% dari seluruh kematian global [1]. Besarnya beban tersebut membuat skrining awal relevan, terutama bila dapat dilakukan dengan variabel klinis rutin. Individu yang terindikasi berisiko dapat lebih cepat diarahkan ke pemeriksaan lanjutan. Namun, keluaran komputasional tetap harus dibaca sebagai informasi pendukung, bukan pengganti anamnesis, pemeriksaan fisik, pemeriksaan penunjang, atau keputusan tenaga medis.

Berbagai pendekatan machine learning telah dicoba untuk memodelkan hubungan nonlinier pada data kardiovaskular. Kerangka explainable machine learning, misalnya, digunakan untuk mendukung diagnosis dan prognosis [2], sedangkan studi berbasis beberapa dataset menggabungkan ensemble dengan penjelasan model [3]. Pendekatan transformer tabular juga dikembangkan untuk menangkap interaksi antarfaktor klinis [4]. Pada jalur lain, voting ensemble dipadukan dengan SHAP dan LIME agar alasan di balik prediksi lebih mudah ditelusuri [5], [6]. Hasil-hasil tersebut menjanjikan, tetapi angka kinerjanya tidak bisa disandingkan begitu saja karena sumber data, preprocessing, resampling, fitur, dan skema validasinya berbeda.

Masalah black box menjadi lebih sensitif ketika model ditempatkan dalam sistem pendukung keputusan klinis. Pengguna perlu mengetahui faktor yang mendorong suatu skor, bukan hanya menerima label akhir. Perspektif multidisipliner menempatkan explainability sebagai bagian dari tata kelola AI kesehatan [7]. Tinjauan sistematis juga menunjukkan bahwa penjelasan lokal dan global harus dibaca bersama konteks data serta validasi eksternal [8]; bahkan, model yang dapat dijelaskan belum tentu otomatis dapat dipercaya [9]. Karena itu, ketidakpastian prediksi dan disparitas antarkelompok demografis ikut diperiksa. LLM memang dapat membantu menerjemahkan keluaran teknis ke bahasa alami [10], tetapi pemakaiannya pada domain medis tetap memerlukan pembatasan, keterlacakan sumber, dan pengawasan manusia [11].

Atas dasar persoalan tersebut, HeartXplain dikembangkan sebagai AI Health Agent berbasis web untuk skrining awal risiko penyakit jantung. Alurnya menghubungkan pembentukan estimator XGBoost, optimasi PSO yang memprioritaskan recall, penyesuaian decision threshold, penjelasan SHAP dan LIME, serta penyajian narasi melalui LLM. MAPIE dan Fairlearn ditambahkan untuk membaca ketidakpastian dan disparitas kelompok. Evaluasi tidak berhenti pada performa klasifikasi, tetapi juga menilai apakah faktor prediksi dapat ditelusuri dan apakah keluaran model perlu diperlakukan dengan kehati-hatian. Posisi HeartXplain tetap sebagai alat bantu skrining awal.

METODE

Penelitian menggunakan dataset MyCardio yang disusun melalui harmonisasi sejumlah sumber data kardiovaskular publik, termasuk Cardiovascular Disease Dataset pada IEEE Dataport [12], serta dataset Cleveland dan Hungarian. Penyamaan atribut dilakukan sebelum pemeriksaan missing value, penghapusan duplikasi, validasi tekanan darah, transformasi, encoding, dan feature engineering. Rangkaian ini menghasilkan 69.105 observasi dengan cardio sebagai target biner. Batas JNC-7 digunakan sebagai dasar historis ketika memvalidasi tekanan darah dan membentuk kategori hipertensi [13].

Sebelas fitur primer digunakan sebagai masukan: age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, dan active. Dari fitur tersebut dibentuk delapan fitur turunan, yaitu Body Mass Index (BMI), Mean Arterial Pressure (MAP), Pulse Pressure, Hypertension Stage, Age-BMI

DOI: https://doi.org/10.47353/bj.v6i4.687

Interaction, Blood Pressure Ratio, Cholesterol-Glucose Interaction, dan BMI Category. Dengan demikian, setiap observasi direpresentasikan oleh 19 fitur. Stratified split digunakan agar proporsi kelas tetap terjaga pada data latih dan data uji. Matriks fitur dan label inilah yang kemudian masuk ke tahap pembentukan model.

Model klasifikasi dibangun dengan Extreme Gradient Boosting (XGBoost), yaitu ensemble pohon yang ditambahkan secara bertahap dan dikendalikan melalui regularisasi [14]. Hyperparameter dicari menggunakan Particle Swarm Optimization (PSO). Dalam PSO, setiap kandidat bergerak dengan mempertimbangkan posisi terbaiknya sendiri dan posisi terbaik populasi [15], [16]. Fungsi objektif menggabungkan AUC dengan penalti ketika recall belum mencapai sasaran eksperimen. Dengan rancangan ini, pencarian tidak hanya mengejar kemampuan diskriminasi umum, tetapi juga menjaga sensitivitas terhadap kelas positif. Pembaruan kecepatan dan posisi partikel dirumuskan pada Persamaan (1) dan Persamaan (2).

Persamaan (1)

$$v_i^{t+1} = w \cdot v_i^t + c_1 \cdot r_1 \cdot (pbest_i - x_i^t) + c_2 \cdot r_2 \cdot (gbest - x_i^t)$$

Persamaan (2)

$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

Pada Persamaan (1), w menyatakan inertia weight; c1 dan c2 adalah koefisien percepatan; r1 dan r2 merupakan bilangan acak pada rentang 0-1; pbest menunjukkan posisi terbaik tiap partikel; sedangkan gbest menunjukkan posisi terbaik populasi. Persamaan (2) memperbarui posisi berdasarkan posisi sebelumnya dan kecepatan baru. Konfigurasi terbaik dari PSO dipakai untuk melatih model final. Sesudahnya, decision threshold digeser dari nilai default 0,5000 untuk melihat kemungkinan peningkatan recall. Evaluasi mencakup accuracy, recall, precision, F1-score, specificity, AUC-ROC, PR-AUC, Brier Score, dan jumlah false negative.

Interpretabilitas dibaca melalui dua sudut. LIME membangun model pendekatan di sekitar satu observasi, sehingga faktor yang menaikkan atau menurunkan prediksi dapat dilihat secara lokal [17]. SHAP menghitung kontribusi aditif tiap fitur dan dapat diringkas dari tingkat individual ke tingkat global [18]. Untuk model pohon seperti XGBoost, pendekatan khusus tree model menjaga konsistensi interpretasi [19]. HeartXplain menampilkan ringkasan global, penjelasan individual, dan Clinical Insights. LLM hanya merangkai narasi dari probabilitas serta keluaran XAI yang telah tersedia; modul ini tidak membuat diagnosis atau rekomendasi terapi sendiri.

Ketidakpastian model diperiksa dengan MAPIE, pustaka conformal prediction yang membentuk prediction set pada target coverage tertentu [20], [21]. Singleton set memuat satu kelas, ambiguous set memuat {0,1}, dan empty set tidak memuat kelas. Interpretasinya perlu mempertimbangkan keseimbangan antara coverage dan ukuran set [22]. Audit fairness dilakukan melalui demographic parity difference dan equalized odds difference pada jenis kelamin serta kelompok usia [23]. Kedua metrik dipakai untuk menandai disparitas, bukan untuk menyimpulkan diskriminasi kausal. Analisis bias pada prediksi CVD tetap memerlukan pembacaan demografis dan validasi lintas populasi [24].

DOI: https://doi.org/10.47353/bj.v6i4.687

HASIL DAN PEMBAHASAN

Hasil Preprocessing dan Feature Engineering

Cleaning, validasi, encoding, dan feature engineering menyisakan 69.105 baris dengan 19 fitur. Fitur turunan merangkum hubungan yang tidak tertangkap oleh satu variabel saja. BMI, misalnya, menggabungkan tinggi dan berat; MAP, Pulse Pressure, Blood Pressure Ratio, dan Hypertension Stage meringkas karakteristik tekanan darah. Age-BMI dan Cholesterol-Glucose Interaction menangkap interaksi antarfaktor. Bentuk ini tetap menggunakan variabel rutin yang dapat dimasukkan pengguna, sekaligus memberi representasi tambahan bagi model.

Fitur turunan tersebut tidak otomatis memiliki makna kausal atau validitas klinis. Karena MyCardio dibentuk dari harmonisasi sumber publik, perbedaan definisi variabel dan karakteristik populasi masih mungkin terbawa ke model. Hasil yang dilaporkan karena itu terbatas pada dataset penelitian. Setelah batas ini ditegaskan, pembahasan dilanjutkan pada cara estimator XGBoost dibentuk dan dikembangkan menjadi model final HeartXplain.

Proses Pembentukan Model XGBoost

Setelah preprocessing, 69.105 observasi disusun menjadi matriks dengan 19 fitur dan satu label biner, cardio. Stratified split menjaga perbandingan kelas saat data dipisahkan. Data latih dipakai untuk membentuk estimator XGBoost, sementara data uji menjadi dasar evaluasi akhir. Keluaran estimator berupa probabilitas kelas positif; label keputusan baru ditetapkan setelah probabilitas tersebut dibandingkan dengan decision threshold. Probabilitas yang sama kemudian diteruskan ke modul interpretasi, uncertainty, fairness, dan penyusunan narasi.

Pengembangan Model HeartXplain

Model dikembangkan dari konfigurasi Optuna sebagai baseline, kemudian ruang hyperparameter XGBoost ditelusuri dengan PSO. Kandidat dinilai menggunakan fungsi objektif AUC yang diberi penalti apabila recall belum mencapai sasaran. Kandidat terbaik menjadi model final; decision threshold selanjutnya disesuaikan tanpa mengubah probabilitas yang dihasilkan estimator. Alur ini mengikuti mekanisme dasar XGBoost dan PSO [14]-[16]. Model final tidak berdiri sendiri: probabilitasnya menjadi masukan bagi evaluasi klasifikasi, SHAP, LIME, MAPIE, Fairlearn, dan LLM. Dengan pembagian tersebut, optimasi model, kebijakan keputusan, interpretasi, serta penyajian narasi tetap dapat dibedakan dengan jelas.

Hasil Optimasi XGBoost Menggunakan PSO

Tabel 1. Konfigurasi Hyperparameter Baseline dan PSO

Hyperparameter	Baseline (Optuna)	PSO (Final)
n_estimators	270	1000
max_depth	4	7
learning_rate	0,005313	0,007342
subsample	0,6547	1,0000
colsample_bytree	0,6103	0,5585

DOI: https://doi.org/10.47353/bj.v6i4.687

Hyperparameter	Baseline (Optuna)	PSO (Final)
min_child_weight	6	7
gamma	0,9671	0,9678
reg_alpha (L1)	5,4710	9,9674
reg_lambda (L2)	$3{,}22\times10^{-7}$	9,9674
scale_pos_weight	1,4987	1,5000
eval_metric	logloss	logloss

Sumber: Diolah oleh peneliti

Konfigurasi Optuna berfungsi sebagai baseline, sedangkan hasil PSO digunakan pada model utama. PSO memilih 1.000 pohon, max_depth 7, learning_rate 0,007342, subsample 1,0000, dan colsample_bytree 0,5585. Nilai regularisasi L1 dan L2 sama-sama 9,9674, dengan scale_pos_weight 1,5000. Baseline, sebagai pembanding, menggunakan 270 pohon dan max_depth 4. Nilai-nilai pada Tabel 1 merupakan konfigurasi XGBoost yang ditemukan dalam ruang pencarian penelitian ini [14].

Perbedaan konfigurasi pada Tabel 1 tidak berarti PSO selalu lebih unggul daripada metode optimasi lain. PSO memperbarui kandidat berdasarkan solusi terbaik individual dan global [15], dan pola pencarian serupa telah digunakan untuk pemilihan hyperparameter [16]. Pada eksperimen ini, PSO menemukan kombinasi yang sesuai dengan fungsi objektif yang ditetapkan. Kemampuannya untuk digeneralisasi masih harus diuji pada data dan ruang pencarian yang berbeda.

Evaluasi Performa Model dan Decision Threshold

Tabel 2. Performa Model PSO pada Dua Decision Threshold

Metrik	Threshold 0,5000	Threshold 0,4997	Selisih
Accuracy	0,7106	0,7101	-0,0005
Recall	0,7998	0,8008	+0,0010
Precision	0,6747	0,6739	-0,0008
F1-score	0,7320	0,7319	-0,0001
Specificity	0,6234	0,6214	-0,0020
AUC-ROC	0,7929	0,7929	0,0000
PR-AUC	0,7805	0,7805	0,0000
Brier Score	0,1947	0,1947	0,0000
False negative	1.367	1.360	-7

Sumber: Diolah oleh peneliti

Ketika threshold digeser dari 0,5000 menjadi 0,4997, recall naik dari 79,98% menjadi 80,08%. Jumlah false negative berkurang tujuh kasus, dari 1.367 menjadi 1.360. Perubahan ini disertai penurunan kecil pada

DOI: https://doi.org/10.47353/bj.v6i4.687

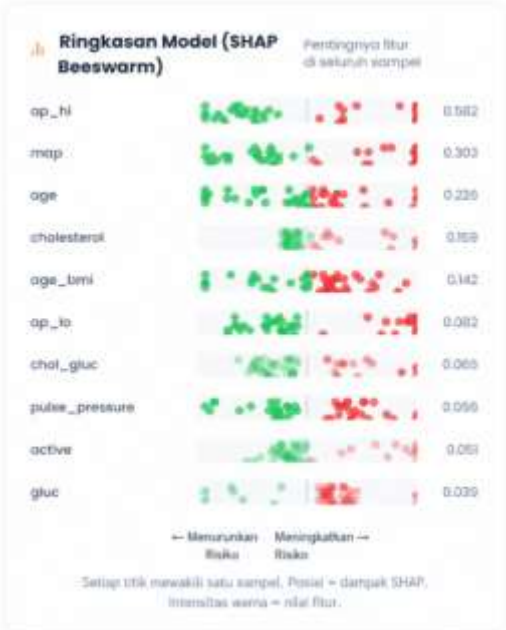
accuracy (0,05 poin persentase), precision (0,08 poin), F1-score (0,01 poin), dan specificity (0,20 poin). AUC-ROC, PR-AUC, serta Brier Score tetap karena pergeseran threshold mengubah label keputusan, bukan probabilitas model.

Recall diprioritaskan karena false negative pada konteks skrining dapat membuat individu berisiko tidak segera diarahkan ke pemeriksaan lanjutan. Konsekuensinya, threshold yang lebih sensitif dapat menambah false positive dan beban tindak lanjut. Kenaikan recall di sini hanya 0,10 poin persentase. Oleh sebab itu, threshold 0,4997 belum dapat dianggap optimal secara klinis sebelum diuji melalui validasi prospektif dan analisis konsekuensi penggunaannya.

AUC-ROC sebesar 79,29% menggambarkan kemampuan diskriminasi model, sedangkan PR-AUC 78,05% merangkum hubungan precision dan recall. Nilai tersebut tidak dibandingkan langsung dengan hasil kerangka diagnosis kardiovaskular [2], transformer tabular [4], maupun ensemble explainable AI [6]. Ketiga studi menggunakan dataset dan protokol evaluasi yang berbeda. Perbandingan kinerja yang adil memerlukan data uji dan preprocessing yang sama.

Interpretasi Global dan Lokal Menggunakan SHAP dan LIME

Gambar 1 merangkum kontribusi fitur di seluruh sampel melalui SHAP Beeswarm. Pembacaannya mengikuti additive feature attribution dan agregasi penjelasan pada model pohon [18], [19]. Tekanan darah sistolik (ap_hi) memiliki kontribusi agregat terbesar, disusul MAP, age, cholesterol, dan age_bmi. Sebaran titik juga menunjukkan bahwa satu fitur dapat mendorong skor ke arah berbeda ketika berinteraksi dengan karakteristik individu lain. Dominasi variabel rutin tersebut sejalan dengan variabel yang digunakan pada pengembangan model risiko kardiovaskular modern [25], tetapi keselarasan ini bukan bukti hubungan kausal maupun validasi diagnosis.



Gambar 1. Interpretasi global menggunakan SHAP Beeswarm

Ringkasan global belum menjawab mengapa satu observasi memperoleh skor tertentu. Pada contoh di Gambar 2, SHAP Waterfall menghasilkan skor akhir 34,74%. Smoke, chol_gluc, bp_ratio, dan pulse_pressure mendorong skor naik; alco, age_bmi, MAP, dan BMI mendorongnya turun. Kontribusi lokal

DOI: https://doi.org/10.47353/bj.v6i4.687

tersebut membentuk jalur dari nilai dasar menuju skor akhir [18], [19]. Arah kontribusi tidak boleh dibaca sebagai jaminan bahwa mengubah satu fitur akan langsung mengubah risiko secara kausal.

Gambar 2. Interpretasi lokal menggunakan SHAP Waterfall

Sebagai pembanding lokal, LIME membangun pendekatan di sekitar satu observasi [17]. Pada contoh yang dianalisis, kolesterol, tekanan darah sistolik, MAP, glukosa, usia, kekakuan arteri, dan BMI dikelompokkan sebagai faktor yang menaikkan atau menurunkan prediksi. SHAP mempertahankan kontribusi aditif yang dapat diagregasi dari lokal ke global; LIME menyajikan ringkasan yang lebih langsung dalam kelompok risk factors dan protective factors. Keduanya saling melengkapi, tetapi penjelasan lokal tetap melekat pada observasi yang sedang dibaca.

Integrasi XAI dan LLM pada HeartXplain

HeartXplain meneruskan probabilitas model dan faktor utama dari XAI ke modul Clinical Insights. Gambar 3 memperlihatkan contoh prediksi risiko tinggi sebesar 78,42%, disertai area yang perlu diperhatikan dan narasi AI Health Insights. Penggunaan model bahasa pada konteks medis perlu menjaga konteks serta pengawasan [10], [11]. Karena itu, LLM tidak menghitung ulang risiko. Tugasnya terbatas pada merangkai informasi yang sebelumnya telah dihasilkan oleh model dan XAI.

DOI: https://doi.org/10.47353/bj.v6i4.687

Gambar 3. Integrasi prediksi, faktor risiko, dan AI Health Insights

Pemisahan peran menjaga asal setiap keluaran tetap jelas: probabilitas berasal dari XGBoost, kontribusi fitur berasal dari SHAP atau LIME, sedangkan LLM menangani penyajian bahasa. Narasi sebaiknya selalu dibaca bersama skor, faktor kontribusi, dan peringatan penggunaan. Integrasi ini dapat membuat keluaran teknis lebih mudah dipahami, tetapi kegunaan narasinya tetap perlu diuji bersama pengguna pada konteks layanan yang nyata.

Evaluasi Pendukung Menggunakan MAPIE dan Fairlearn

Tabel 3. Hasil Evaluasi Ketidakpastian Menggunakan MAPIE

Metrik	Nilai
Confidence Level	95%
Empirical Coverage	95,01%
Average Prediction Set Size	1,71
Singleton Prediction Set	29,40%
Ambiguous Prediction Set	70,60%
Empty Prediction Set	0,00%
Singleton Prediction Accuracy	83,02%

Sumber: Diolah oleh peneliti

Empirical coverage mencapai 95,01%, sangat dekat dengan confidence level 95%. Namun, average prediction set size sebesar 1,71 memperlihatkan bahwa set prediksi masih lebar. Hanya 29,40% prediksi berupa singleton; 70,60% lainnya berupa ambiguous set {0,1}, dan tidak ditemukan empty set. Akurasi pada subset singleton mencapai 83,02%, lebih tinggi daripada accuracy keseluruhan 71,01%. Pola ini sesuai dengan prinsip conformal prediction: coverage perlu dibaca bersama ukuran prediction set [20]-[22].

DOI: https://doi.org/10.47353/bj.v6i4.687

Singleton menghasilkan keluaran yang lebih tegas, sedangkan ambiguous set menunjukkan bahwa kedua kelas masih masuk akal di sekitar decision boundary.

Tabel 4. Hasil Evaluasi Fairness Menggunakan Fairlearn

Atribut Sensitif	Demographic Parity Difference	Equalized Odds Difference
Jenis Kelamin	0,0097 (0,97%)	0,0193 (1,93%)
Kelompok Usia	0,3714 (37,14%)	0,3702 (37,02%)

Sumber: Diolah oleh peneliti

Pada atribut jenis kelamin, demographic parity difference sebesar 0,0097 dan equalized odds difference 0,0193 menunjukkan selisih yang relatif kecil. Polanya berbeda pada kelompok usia, dengan nilai masing-masing 0,3714 dan 0,3702. Metrik Fairlearn berfungsi sebagai penanda awal dalam audit kelompok [23], bukan bukti diskriminasi kausal. Kajian bias pada model kardiovaskular juga menekankan perlunya memeriksa prevalensi, representasi, dan karakteristik fitur pada setiap kelompok [24]. Karena itu, kesenjangan usia pada hasil ini perlu ditelusuri dengan data eksternal yang lebih beragam.

Jika dibaca sebagai satu rangkaian, hasil penelitian bergerak dari preprocessing dan pembentukan XGBoost menuju optimasi PSO, penyesuaian threshold, interpretasi, estimasi ketidakpastian, dan audit fairness. Model lebih sensitif terhadap kelas positif dan faktor prediksinya dapat ditelusuri, sementara MAPIE serta Fairlearn menambahkan konteks yang tidak terlihat dari metrik klasifikasi saja. Di sisi lain, dataset hasil harmonisasi, belum adanya validasi pada rekam medis nyata, tingginya ambiguous set, disparitas usia, serta narasi LLM yang belum diuji secara faktual membatasi generalisasi. HeartXplain pada tahap ini masih merupakan purwarupa alat bantu skrining awal.

KESIMPULAN

HeartXplain dibangun dari estimator XGBoost yang dioptimasi dengan PSO berorientasi recall, lalu dilengkapi penyesuaian decision threshold, SHAP, LIME, LLM, MAPIE, dan Fairlearn. Pada threshold 0,4997, model memperoleh accuracy 71,01%, recall 80,08%, precision 67,39%, F1-score 73,19%, AUC-ROC 79,29%, dan PR-AUC 78,05%. False negative berkurang dari 1.367 menjadi 1.360. SHAP dan LIME memperlihatkan faktor yang memengaruhi prediksi, sementara LLM menyusunnya menjadi narasi. MAPIE mencapai coverage 95,01%; Fairlearn menemukan selisih kecil berdasarkan jenis kelamin, tetapi selisih yang lebih besar pada kelompok usia.

Temuan tersebut belum menempatkan HeartXplain sebagai alat diagnosis. Sistem tetap memerlukan pemeriksaan dan keputusan tenaga medis. Generalisasi dibatasi oleh dataset publik terharmonisasi, proporsi ambiguous prediction set yang tinggi, disparitas usia, dan belum tersedianya validasi klinis nyata. Penelitian lanjutan perlu menguji model secara eksternal dan prospektif pada data rekam medis, memeriksa stabilitas threshold, memperdalam audit fairness serta uncertainty, dan menilai akurasi faktual maupun kegunaan narasi LLM bersama tenaga kesehatan.

DOI: https://doi.org/10.47353/bj.v6i4.687

DAFTAR PUSTAKA

[1] "Cardiovascular diseases (CVDs)." Accessed: Aug. 05, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] Md. E. A. Sourov, Md. S. Hossen, P. Shaha, Md. M. Siddique, Y. Sutradhar, and M. S. Iqbal, "Explainable Machine Learning Framework for Cardiovascular Disease Diagnosis and Prognosis," in *2026 IEEE International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Kelaniya, Gampaha, Sri Lanka: IEEE, Mar. 2026, pp. 1–6. doi: 10.1109/SCSE70081.2026.11499913.
[3] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Sci. Rep.*, vol. 15, no. 1, p. 13912, Apr. 2025, doi: 10.1038/s41598-025-97547-6.
[4] F. Mahmud *et al.*, "HybridTabNet-QC: A Transformer-Based Clinical Feature Fusion Framework for Heart Disease Risk Prediction," *IEEE Open J. Comput. Soc.*, vol. 7, pp. 1–13, 2026, doi: 10.1109/OJCS.2025.3637308.
[5] N. G. Rezk, S. Alshathri, A. Sayed, E. El-Din Hemdan, and H. El-Behery, "XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach," *Bioengineering*, vol. 11, no. 10, p. 1016, Oct. 2024, doi: 10.3390/bioengineering11101016.
[6] A. Adekoya, F. Saeed, W. Ghaban, and S. N. Qasem, "Ensemble learning approach with explainable AI for improved heart disease prediction," *Front. Pharmacol.*, vol. 16, p. 1654681, Dec. 2025, doi: 10.3389/fphar.2025.1654681.
[7] the Precise4Q consortium, J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 310, Dec. 2020, doi: 10.1186/s12911-020-01332-6.
[8] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, p. 107161, Nov. 2022, doi: 10.1016/j.cmpb.2022.107161.
[9] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Inform.*, vol. 113, p. 103655, Jan. 2021, doi: 10.1016/j.jbi.2020.103655.
[10] M. Moor *et al.*, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023, doi: 10.1038/s41586-023-05881-4.
[11] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.
[12] R. K. Halder, "Cardio Data." IEEE DataPort, Nov. 10, 2020. doi: 10.21227/7QM5-DZ13.
[13] A. V. Chobanian *et al.*, "Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure," *Hypertension*, vol. 42, no. 6, pp. 1206–1252, Dec. 2003, doi: 10.1161/01.HYP.0000107251.49515.c2.
[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[15] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia: IEEE, 1995, pp. 1942–1948. doi: 10.1109/ICNN.1995.488968.
[16] P. R. Lorenzo, J. Nalepa, M. Kawulok, L. S. Ramos, and J. R. Pastor, "Particle swarm optimization for hyper-parameter selection in deep neural networks," in *Proceedings of the Genetic and Evolutionary Computation Conference*, Berlin Germany: ACM, Jul. 2017, pp. 481–488. doi: 10.1145/3071178.3071208.
[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

DOI: https://doi.org/10.47353/bj.v6i4.687

Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[18] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017, *arXiv*. doi: 10.48550/ARXIV.1705.07874.

[19] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[20] V. Taquet, V. Blot, T. Morzadec, L. Lacombe, and N. Brunel, "MAPIE: an open-source library for distribution-free uncertainty quantification," 2022, *arXiv*. doi: 10.48550/ARXIV.2207.12274.

[21] A. N. Angelopoulos and S. Bates, "Conformal Prediction: A Gentle Introduction," *Found. Trends® Mach. Learn.*, vol. 16, no. 4, pp. 494–591, Mar. 2023, doi: 10.1561/2200000101.

[22] A. Fisch, T. Schuster, T. Jaakkola, and R. Barzilay, "Conformal Prediction Sets with Limited False Positives," 2022, *arXiv*. doi: 10.48550/ARXIV.2202.07650.

[23] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and Improving Fairness of AI Systems," 2023, *arXiv*. doi: 10.48550/ARXIV.2303.16626.

[24] F. Li, P. Wu, H. H. Ong, J. F. Peterson, W.-Q. Wei, and J. Zhao, "Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction," *J. Biomed. Inform.*, vol. 138, p. 104294, Feb. 2023, doi: 10.1016/j.jbi.2023.104294.

[25] S. S. Khan *et al.*, "Development and Validation of the American Heart Association's PREVENT Equations," *Circulation*, vol. 149, no. 6, pp. 430–449, Feb. 2024, doi: 10.1161/CIRCULATIONAHA.123.067626.