

Prediksi Penyakit Jantung Menggunakan *Cost-Sensitive Random Forest* dan SHAP

Heart Disease Prediction Using Cost-sensitive Random Forest and SHAP

¹Arya Chova Pratama*, ²Eko Arip Winanto, ³Beny

^{1,2,3}Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dinamika Bangsa

^{1,2,3}Jl. Raden Wijaya, Kota Jambi 36131, Indonesia

*e-mail: aryachovapratama@gmail.com

(received: 25 May 2026, revised: 26 June 2026, accepted: 15 September 2026)

Abstrak

Riset prediksi penyakit jantung berbasis *machine learning* umumnya melaporkan akurasi tinggi, tetapi mayoritas memakai *dataset* kecil yang distribusinya hampir seimbang seperti Cleveland UCI. Kondisi semacam itu jauh dari distribusi populasi nyata, yang ketidakseimbangannya berat dan konsekuensi kesalahannya tidak setara: kasus positif yang luput menghilangkan peluang intervensi dini, sementara kesalahan sebaliknya masih dapat dikoreksi pada pemeriksaan lanjutan. Studi ini menyusun kerangka evaluasi metodologis untuk klasifikasi profil risiko kardiovaskular berbasis *self-report* dengan memadukan *Cost-sensitive Random Forest* (CSRF) dan SHapley Additive exPlanations (SHAP). Data berasal dari Behavioral Risk Factor Surveillance System (BRFSS) 2022 versi no_nans rilis CDC yang dikurasi Pytlak, mencakup 246022 responden, 39 variabel prediktor, dan rasio kelas 1:17,31. Delapan konfigurasi diuji pada anggaran *tuning* yang seragam, meliputi *baseline Random Forest*, varian *Cost-sensitive*, pendekatan *resampling*, serta pembandingan *gradient boosting*, lalu dievaluasi dengan *cross-validation*, *bootstrap confidence interval*, validasi *leave-one-state-out*, optimasi *threshold* berbasis *cost matrix*, serta audit kalibrasi dan *fairness*. CSRF 1:3 menekan *false negative* dari 2117 menjadi 1360 pada *test set* 49205 responden, dengan *recall* 49,4% (95% CI: 47,5–51,2%) dan F1 0,483. Nilai F1 tersebut tertinggi pada kelompok *Random Forest* dan *resampling*, sementara CS-XGBoost dan CS-LightGBM memberi pembandingan *robustness* dengan F1 sedikit lebih tinggi. Interpretasi SHAP menempatkan HadAngina dan ChestScan sebagai prediktor dominan, dua variabel yang secara klinis sangat dekat dengan target, sehingga capaian performa lebih tepat dibaca sebagai asosiasi *cross-sectional* berbasis *self-report*, bukan prediksi prospektif klinis murni.

Kata kunci: *cost-sensitive random forest*, *decision curve analysis*, explainable AI, *false negative*, SHAP

Abstract

Machine-learning studies on heart disease prediction often report high accuracy, yet most rely on small *datasets* with near-balanced class distributions such as Cleveland UCI. Such conditions are far from real population data, where imbalance is severe and the costs of errors are asymmetric: a missed positive case removes the chance for early intervention, while a false alarm can usually be corrected during follow-up. This work develops a methodological evaluation framework for *self-reported* cardiovascular risk classification by combining *Cost-sensitive Random Forest* (CSRF) with SHapley Additive exPlanations (SHAP). The data come from the Behavioral Risk Factor Surveillance System (BRFSS) 2022 no_nans release by CDC and curated by Pytlak, comprising 246,022 respondents, 39 predictor variables, and a 1:17.31 class ratio. Eight configurations were compared under a uniform *tuning* budget, including *baseline Random Forest*, *Cost-sensitive* variants, *resampling* approaches, and gradient boosting comparators, then evaluated with *cross-validation*, *bootstrap confidence intervals*, *leave-one-state-out* validation, cost-matrix *threshold* optimisation, and calibration and fairness audits. CSRF 1:3

reduced *false negatives* from 2,117 to 1,360 on a *test set* of 49,205 respondents, with recall 49.4% (95% CI: 47.5–51.2%) and F1 0.483. That F1 score is the highest within the *Random Forest* and *resampling* group, while *CS-XGBoost* and *CS-LightGBM* obtained slightly higher F1 scores as robustness comparators. The SHAP analysis identifies HadAngina and ChestScan as dominant predictors, two variables clinically very close to the target, so the performance should be read as a *cross-sectional, self-report*-based association rather than pure prospective clinical prediction.

Keywords: *cost-sensitive random forest, decision curve analysis, explainable AI, false negative, SHAP*

1. Pendahuluan

Penyakit kardiovaskular masih menjadi penyebab kematian terbesar di dunia. Berdasarkan Global Burden of Disease 2023, ada sekitar 19,2 juta kematian per tahun akibat penyakit ini, meningkat dari 13,1 juta pada 1990, dan iskemia jantung menjadi kontributor terbesar [1]. Sekitar delapan dari sepuluh kematian tersebut berasal dari faktor risiko yang sebenarnya dapat dimodifikasi, sehingga skrining berbasis profil risiko membuka ruang yang cukup luas untuk menekan beban penyakit [1]. Persoalan terasa lebih berat di Indonesia. Akses pemeriksaan kardiologi yang canggih masih terpusat di kota-kota besar, sedangkan layanan primer di banyak daerah harus bekerja dengan data klinis seadanya dan dituntut mengambil keputusan secara cepat [2].

Machine learning hadir sebagai alat bantu untuk mengenali pola yang sukar ditangkap aturan statistik konvensional. *Random Forest* kerap menjadi pilihan karena stabil pada data klinis berdimensi menengah [3], dan literatur belakangan ini umumnya melaporkan akurasi di atas 90% pada subset Cleveland [13]–[15]. Akan tetapi, angka tersebut menyimpan dua catatan yang perlu dicermati. Pertama, Cleveland hanya berisi sekitar 300 instance dengan distribusi kelas mendekati seimbang, sehingga tidak mewakili situasi populasi nyata. Pada survei berskala populasi, prevalensi infark miokard hanya sekitar 5%, artinya ketidakseimbangan yang harus dihadapi jauh lebih berat [2], [16]. Kedua, sebagian besar studi tersebut memperlakukan *false negative* dan *false positive* seakan-akan setara, padahal pada diagnosis kardiovaskular kedua kesalahan ini membawa konsekuensi yang sangat berbeda.

Catatan kedua menyangkut interpretabilitas. Model ensemble seperti *Random Forest* sering dilihat sebagai kotak hitam, sementara tenaga kesehatan membutuhkan justifikasi yang dapat dipahami sebelum mempercayai keluaran model. SHAP yang berakar pada teori permainan kooperatif menyediakan dekomposisi prediksi yang konsisten dan additif [9]. Namun, perpaduan SHAP dengan strategi penanganan *imbalance* jarang dikerjakan secara terstruktur dalam satu studi [17]. Literatur prediksi penyakit jantung pada umumnya juga berhenti pada SHAP global untuk merangking fitur, sedangkan pola pada kelompok kasus *false negative*, yang justru paling berguna untuk perbaikan model, jarang ditelaah secara sistematis.

Penelitian ini diarahkan ke ketiga celah tersebut. Evaluasi dijalankan pada skala populasi melalui *dataset* BRFSS 2022 yang dirilis Centers for Disease Control and Prevention dan dikurasi Pytlak di Kaggle [2], yang berisi 246022 responden dari 54 yurisdiksi dengan rasio ketidakseimbangan 1:17,31. Rasio sebesar itu jauh lebih merepresentasikan kondisi populasi sesungguhnya. Karena BRFSS adalah survei *cross-sectional* berbasis *self-report*, target HadHeartAttack adalah riwayat infark miokard yang dilaporkan responden, sehingga studi ini diposisikan sebagai klasifikasi profil risiko kardiovaskular *self-report* yang berfungsi sebagai instrumen *risk stratification* operasional. Asimetri biaya klinis dimasukkan secara eksplisit lewat *Cost-sensitive Random Forest* yang menjadikan bobot kelas sebagai parameter optimasi terhadap F-beta, sedangkan SHAP dipakai baik untuk interpretasi global maupun untuk diseksi kelompok *false negative* secara terstruktur.

Kontribusi penelitian dapat diuraikan menjadi tiga bagian yang saling berkait. Pertama, pemetaan empiris cara intervensi *Cost-sensitive* dan *resampling* bekerja pada data populasi nyata, dilengkapi analisis mekanisme di balik peningkatan *recall* yang teramati. Kedua, analisis forensik SHAP terhadap 1360 kasus *false negative* yang memperlihatkan profil pasien sistematis di balik kegagalan prediksi,

sebuah lapisan interpretabilitas yang melengkapi pembacaan SHAP global. Ketiga, pembacaan kritis atas struktur kontribusi fitur klinis pada SHAP global, yang menempatkan capaian performa pada konteks *self-report cross-sectional* yang lebih sesuai. Kerangka evaluasi yang menampung ketiga kontribusi tersebut memuat *bootstrap confidence interval*, *paired t-test* dengan koreksi *multiple comparison*, *leave-one-state-out* lintas 53 yurisdiksi, *threshold optimization* berbasis *cost matrix*, *decision curve analysis*, kalibrasi *isotonic*, dan audit *subgroup fairness* dalam satu studi terkendali. Rangkaian semacam ini relatif jarang dijumpai secara utuh pada literatur prediksi penyakit jantung berbahasa Indonesia.

2. Tinjauan Literatur

Random Forest dirumuskan Breiman sebagai gabungan *bootstrap aggregation* dan *random subspace selection*. Kombinasi tersebut yang menjelaskan kestabilan *Random Forest* pada beragam *dataset*, sekaligus daya tahannya terhadap *overfitting* dibandingkan pohon tunggal [3]. Reputasi pada prediksi penyakit jantung diperkuat sejumlah studi terbaru. Chandrasekhar dan Peddakrishna menggabungkan *Random Forest* dalam skema *soft voting* bersama lima algoritma lain pada subset Cleveland dan mencatat akurasi 93,44% [13]. Asif dkk. memperluas eksperimen ke gabungan tiga *dataset* Kaggle melalui Extra Trees Classifier dan mencapai akurasi 98,15% [14]. Naik dkk. mengoptimasi *hyperparameter Random Forest* dan *XGBoost* lewat Social Group Optimization dan menunjukkan peningkatan stabilitas lintas fold [15]. Sadr dkk. memadukan *machine learning* dan deep learning pada tiga *dataset* CVD dan melaporkan akurasi 94,1% pada pendekatan hybrid mereka [16]. Pola yang terlihat dari studi-studi tersebut serupa, yaitu ensemble berbasis pohon tetap menjadi tulang punggung, sementara variasi metode datang dari sisi optimasi *hyperparameter* atau skema stacking. Penelaahan lebih cermat memperlihatkan tiga keterbatasan yang saling berkait di kelompok studi tersebut. Mayoritas eksperimen masih bertumpu pada subset Cleveland UCI berukuran ratusan instance dengan distribusi kelas yang nyaris seimbang. Kondisi tersebut tidak menggambarkan keadaan populasi yang prevalensi infark miokardnya hanya sekitar lima persen, sehingga ketidakseimbangan sesungguhnya jauh lebih berat [2], [18]. Klaim akurasi sangat tinggi seperti 98,15% pada Asif dkk. [14] juga jarang disertai analisis ketidakpastian melalui *confidence interval*. Perbedaan kecil antarmodel acap kali diklaim sebagai keunggulan tanpa pengujian statistik formal. Hampir semua studi melaporkan *accuracy*, *precision*, *recall*, dan F1 secara setara tanpa membedakan implikasi *false negative* dan *false positive*, padahal pada diagnosis kardiovaskular kedua kesalahan ini menanggung konsekuensi yang sangat berbeda. Ogunpola dkk. memang menyinggung pentingnya menangani *imbalance* pada prediksi kardiovaskular, tetapi pembahasannya berhenti pada *SMOTE* tanpa pembandingan terkendali [18].

Araf dkk. memetakan 173 publikasi rentang 2010–2022 untuk meninjau penanganan *imbalance* pada data medis. Kesimpulan utamanya, *Cost-sensitive learning* lebih unggul daripada *resampling* pada *dataset* berukuran kecil sampai menengah karena tidak menambahkan instance baru yang berpotensi mengubah distribusi alami data [5]. Fondasi teoritis pendekatan ini diletakkan oleh Elkan yang merumuskan optimal *Cost-sensitive decision* sebagai modifikasi *threshold* berdasarkan rasio biaya kesalahan [21], sementara Fernández dkk. menyusun tinjauan komprehensif lintas dekade tentang strategi pembelajaran untuk data tidak seimbang [22]. Zhu dkk. menerapkan kerangka tersebut secara spesifik pada *Random Forest* lewat adaptive class weights yang dihitung berdasarkan prior distribusi kelas, dan melaporkan kenaikan *recall* pada kelas minoritas tanpa menurunkan *precision* secara dramatis [8]. Di domain penyakit jantung, pendekatan ini jarang diadopsi. Sebagian besar peneliti tetap memilih *SMOTE* [4], [7] karena ketersediaan pustaka imbalanced-learn yang siap pakai, walaupun pada data klinis besar dengan *severe imbalance*, *SMOTE* kerap menaikkan variansi lintas fold tanpa perbaikan *recall* yang konsisten dan berisiko menimbulkan *overfitting* di region sintesis [22].

Tren menuju explainable AI memunculkan beragam metode interpretasi, dan SHAP menonjol berkat tiga sifat aksiomatiknya: local accuracy, missingness, dan consistency [6], [9]. Penerapannya pada domain medis berkembang cepat. Chen dkk. memakai SHAP untuk menjelaskan model prediksi luaran fungsional pasien stroke iskemik akut dan menemukan skor NIHSS serta usia sebagai prediktor

dominan yang sejalan dengan pemahaman klinis [17]. Chulde-Fernández dkk. menunjukkan *Random Forest* yang dilengkapi SHAP menghasilkan interpretasi yang sesuai dengan pemahaman fisiologis gagal jantung [19]. Tinjauan menyeluruh oleh Noor dkk. mencatat bahwa SHAP kini menjadi metode interpretasi yang paling banyak dipakai pada penelitian AI bidang kesehatan, namun integrasinya dengan strategi *Cost-sensitive* masih jarang dilakukan dalam satu rangkaian eksperimen terkendali [11]. Pada lingkup nasional, Selayanti dkk. menggunakan *Random Forest* pada *dataset* Heart Failure tanpa menyentuh persoalan *imbalance* maupun *explainability* secara eksplisit [20], dan kondisi ini menggambarkan kesenjangan adopsi metodologi mutakhir pada publikasi domestik.

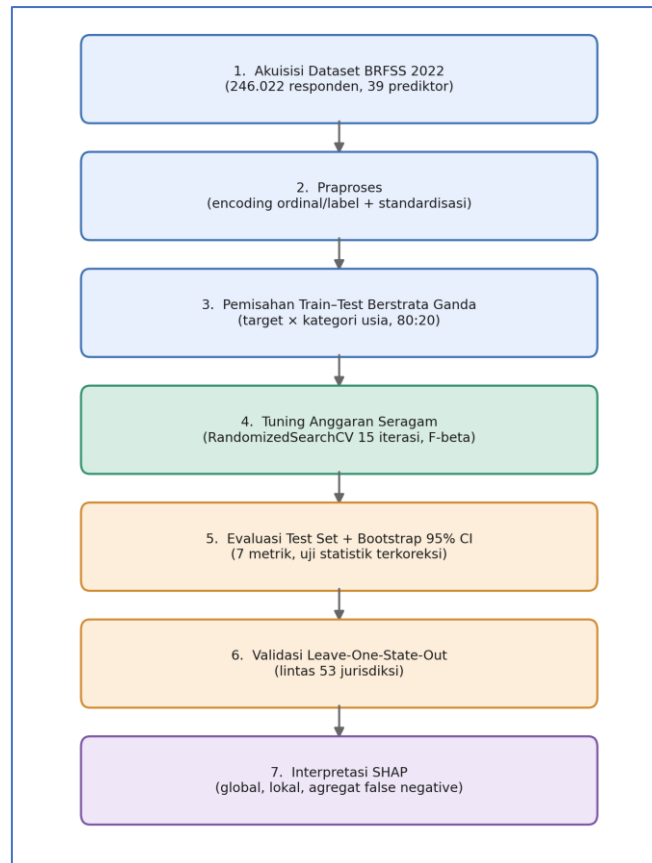
Selain pilihan algoritma, terdapat persoalan metodologis yang lebih halus pada cara studi-studi tersebut mengevaluasi modelnya. Perbandingan jarang dilakukan dengan anggaran *tuning* yang setara, karena model utama sering disesuaikan secara khusus sementara *baseline* berjalan dengan parameter *default*, sehingga keunggulan model utama berpotensi membawa bias eksperimental. Sebagian besar laporan hanya menyajikan satu pemisahan train-test atau k-fold tunggal tanpa *bootstrap confidence interval*, sehingga kekuatan klaim sulit diverifikasi. Pengujian statistik berpasangan dengan koreksi *multiple comparison* hampir absen dari literatur prediksi penyakit jantung, padahal pada perbandingan lebih dari dua model, nilai p mentah mudah keliru. Validasi geografis silang lewat *leave-one-state-out* juga jarang dilakukan, sehingga klaim transferabilitas model sulit dipertanggungjawabkan [12]. Pada lapisan interpretabilitas, mayoritas studi yang mengadopsi SHAP berhenti pada level global dengan *ranking* fitur melalui *beeswarm plot* atau bar chart *mean absolute SHAP value*, sedangkan analisis subgrup maupun audit kasus salah klasifikasi jarang dilanjutkan.

Berdasarkan sintesis tersebut, penelitian ini diarahkan ke ketiga celah yang telah diuraikan. Pengujian dijalankan dengan delapan konfigurasi pada anggaran *tuning* yang seragam, dilengkapi *bootstrap 95% confidence interval* [10], analisis power dan Cohen's d [12], koreksi Holm dan Benjamini-Hochberg, validasi *leave-one-state-out* lintas 53 yurisdiksi, *threshold optimization* berbasis *cost matrix* klinis, *decision curve analysis* [23], audit *subgroup fairness*, serta SHAP agregat pada kelompok *false negative*. Tujuh lapis evaluasi yang dipadukan dengan diseksi forensik kelompok *false negative* inilah yang membedakan posisi metodologis penelitian ini relatif terhadap literatur prediksi penyakit jantung sebelumnya.

3. Metode Penelitian

Alur penelitian disusun menjadi tujuh tahap berurutan, yaitu akuisisi *dataset* BRFSS 2022, pra-proses (*encoding* dan *scaling*), pemisahan train-test berstrata ganda, *tuning* dengan anggaran setara antarkonfigurasi, evaluasi pada *test set* dengan *bootstrap confidence interval*, validasi *leave-one-state-out* lintas yurisdiksi, serta interpretasi SHAP global dan agregat *false negative*. Skema umumnya disajikan pada Gambar 1. Implementasi keseluruhan dijalankan pada Python 3.12 dengan pustaka *scikit-learn*, *imbalanced-learn* [4], *XGBoost* [24], *LightGBM*, dan SHAP. *Random seed* ditetapkan 42 pada seluruh komponen yang menerima parameter *random_state*.

Seluruh eksperimen dijalankan pada perangkat laptop dengan prosesor AMD Ryzen 5 3500U with Radeon Vega Mobile Gfx 2,10 GHz, RAM 8,00 GB dengan 5,95 GB usable, GPU AMD Radeon Vega 8 Graphics 2 GB, sistem operasi Windows 11 Home Single Language 64-bit, dan penyimpanan 477 GB. Eksperimen tidak menggunakan GPU secara khusus karena model utama berbasis pohon dijalankan pada CPU. Waktu komputasi total untuk proses *tuning*, evaluasi *test set*, *bootstrap confidence interval*, validasi *leave-one-state-out*, dan interpretasi SHAP berkisar 6–8 jam. Informasi ini disertakan untuk memperjelas kelayakan replikasi eksperimen pada perangkat komputasi standar.



Gambar 1 Alur penelitian tujuh tahap evaluasi *cost-sensitive random forest*

3.1 Dataset dan Praproses

Dataset utama yang dipakai adalah BRFSS 2022 versi heart_2022_no_nans yang dirilis Pytlak di Kaggle, diturunkan dari Behavioral Risk Factor Surveillance System 2022 milik CDC [2]. Versi no_nans dipilih karena semua baris bermasalah telah dieliminasi sejak hulu, sehingga interpretasi SHAP terbebas dari artefak imputasi. Data mencakup 246022 responden dari 54 jurisdiksi negara bagian Amerika Serikat dengan 39 variabel prediktor dan satu target biner HadHeartAttack yang menandai riwayat infark miokard. Distribusi kelasnya sangat tidak seimbang: 5,46% kasus positif dan 94,54% kasus negatif, atau rasio 1:17,31. Kondisi ini lebih representatif terhadap populasi nyata dibandingkan Cleveland UCI yang distribusinya mendekati 1:1. Ringkasan karakteristik *dataset* disajikan pada Tabel 1.

Tabel 1 Karakteristik *dataset* BRFSS 2022

Kategori	Nilai
Total responden	246022
Jumlah jurisdiksi	54
Variabel prediktor	39
Kasus positif	13435 (5,46%)
Kasus negatif	232587 (94,54%)
Rasio ketidakseimbangan	1 : 17,31
Proporsi perempuan	51,95%
Rerata BMI	28,67
Periode pengumpulan	Januari – Desember 2022
Lisensi data	CC0 1.0 Universal

Praproses sengaja dirancang ringkas supaya interpretabilitas model tetap mudah dilacak. Variabel target HadHeartAttack dibinarisasi menjadi {Yes $\rightarrow$ 1, No $\rightarrow$ 0}. Variabel State disimpan terpisah untuk keperluan *leave-one-state-out* (LOSO) dan tidak diikuti sebagai prediktor, sehingga model tidak menghafal jurisdiksi pelatihan. Sebanyak 32 variabel kategorikal dikodekan ordinal mengikuti urutan klinis yang logis untuk variabel yang memang ordinal, misalnya GeneralHealth dengan urutan Poor < Fair < Good < Very Good < Excellent, sedangkan variabel nominal tanpa urutan natural seperti Sex, RaceEthnicityCategory, dan SmokerStatus dikodekan dengan *label-encoding* berurutan. Pilihan ini mengikuti praktik umum pada literatur *Random Forest* dengan *dataset* BRFSS [2], [18]. Enam variabel numerik (PhysicalHealthDays, MentalHealthDays, SleepHours, HeightInMeters, WeightInKilograms, BMI) di-*standardize* dengan StandardScaler agar *pipeline* tetap konsisten lintas algoritma.

3.2 Pemisahan Data

Pemisahan train-test dijalankan secara stratified 80:20 berstrata ganda, yaitu kombinasi target dan kategori usia, supaya distribusi kelas dan distribusi usia pada *training set* dan *test set* tetap proporsional. Strategi seperti ini penting karena AgeCategory adalah salah satu prediktor terkuat sebagaimana ditunjukkan interpretasi SHAP, sehingga pergeseran distribusi usia antara train dan test akan mengganggu validitas evaluasi. Hasilnya, *training set* berisi 196817 responden dan *test set* 49205 responden. *Tuning hyperparameter* hanya dijalankan pada *training set* lewat stratified 5-fold *cross-validation*, sedangkan *test set* baru disentuh setelah *tuning* rampung.

3.3 Konfigurasi Model dan Mekanisme Cost-sensitive

Delapan konfigurasi diuji dalam tiga kelompok. Kelompok utama terdiri atas *Baseline* RF (titik referensi), CSRF balanced (class_weight="balanced"), CSRF 1:3 (class_weight={0:1, 1:3}), SMOTE+RF, ADASYN+RF, dan RUS+RF. Kelompok pembandingan yang dipakai untuk menguji konsistensi mekanisme adalah CS-XGBoost 1:3 dan CS-LightGBM 1:3, keduanya berbasis *gradient boosting* dengan parameter scale_pos_weight setara bobot 1:3. Setiap konfigurasi memperoleh anggaran *tuning* yang setara melalui RandomizedSearchCV dengan 15 iterasi pada ruang parameter yang proporsional, dan F-beta score ($\beta = 2$) digunakan sebagai *scoring function* yang memberi bobot lebih ke *recall* sesuai tujuan klinis menekan *false negative*. Ruang pencarian *Random Forest* mencakup n_estimators $\in \{100, 200, 300\}$, max_depth $\in \{\text{None}, 8, 12, 16\}$, min_samples_leaf $\in \{1, 5, 10, 20\}$, dan max_features $\in \{\text{"sqrt"}, \text{"log2"}, 0.5\}$. XGBoost dan LightGBM di-tune pada n_estimators $\in \{200, 300\}$, max_depth $\in \{3, 4, 5, 7\}$, dan learning_rate $\in \{0.05, 0.1\}$. Pemilihan rasio 1:3 untuk konfigurasi *Cost-sensitive* utama divalidasi lewat analisis sensitivitas yang dilaporkan pada Subbab 4.3 (Tabel 6), yang menempatkan rasio 1:3 sebagai titik dengan F1 tertinggi di antara opsi 1:3, 1:5, dan 1:10.

Mekanisme *Cost-sensitive* pada pohon keputusan bekerja melalui pembobotan sampel saat menghitung *impurity* setiap simpul. Jika sampel kelas k pada simpul t diberi bobot w_k , proporsi berbobot kelas k pada simpul itu menjadi $p_k^w = (\sum w_i \cdot 1[y_i = k]) / (\sum w_i)$, sehingga *weighted Gini impurity* didefinisikan sebagaimana pada Persamaan (1).

$$I_w(t) = 1 - \sum_k (p_k^w)^2 \quad (1)$$

Bobot yang lebih tinggi pada kelas minoritas membuat *impurity reduction* lebih mudah dicapai melalui *split* yang memurnikan kelas tersebut. Konsekuensinya, pohon-pohon hasil pelatihan menjadi lebih sensitif terhadap pasien sakit tanpa memerlukan data sintesis tambahan. Mekanisme ini berbeda secara fundamental dari SMOTE atau ADASYN yang bekerja di level data dengan mensintesis sampel baru, dan dari RUS yang bekerja melalui pengurangan sampel mayoritas.

3.4 Skema Evaluasi Berlapis

Evaluasi dirancang berlapis agar setiap klaim dapat ditelusuri pada bukti yang spesifik. Lapisan pertama menghitung tujuh metrik (accuracy, precision, recall, specificity, F1, ROC-AUC, PR-AUC) pada *test set* independen, dengan recall sebagai metrik utama untuk konteks klinis dan PR-AUC sebagai

pendamping yang lebih informatif pada data tidak seimbang daripada ROC-AUC [10]. Pada lapisan berikutnya, *bootstrap 95% confidence interval* (1000 *resampling*) dihitung untuk seluruh metrik. Stratified 5-fold *cross-validation* pada *training set* kemudian dijalankan untuk menguji perbedaan F1 antarmodel melalui *paired t-test* dan *Wilcoxon signed-rank test*, dilengkapi Cohen's d, achieved power, dan required sample size untuk power 80% mengikuti rekomendasi Dietterich [12]. *Multiple comparison* dikoreksi dengan metode Bonferroni, Holm, dan Benjamini-Hochberg. Setelah itu, *leave-one-state-out validation* dijalankan dengan melatih ulang model pada 52 yurisdiksi dan mengujinya pada satu yurisdiksi yang ditahan, bergantian untuk seluruh yurisdiksi yang $n\text{-nya} \geq 1000$ (53 yurisdiksi), untuk menilai generalisasi lintas distribusi geografis.

Lapisan berikutnya adalah *threshold optimization* berbasis *cost matrix* klinis (CFN = 5, CFP = 1) yang meminimalkan *expected cost*, sebagai translasi langsung dari asimetri konsekuensi klinis ke level keputusan operasional. Rasio CFN:CFP = 5:1 digunakan sebagai skenario heuristik-operasional untuk merepresentasikan konsekuensi *false negative* yang lebih berat daripada *false positive* pada skrining awal kardiovaskular. Nilai ini tidak dimaksudkan sebagai estimasi biaya klinis universal, melainkan sebagai asumsi operasional yang diuji sensitivitasnya melalui perbandingan rasio 1:3, 1:5, dan 1:10 pada bagian hasil. Dalam konteks skrining awal, kasus positif yang tidak terdeteksi dapat menunda intervensi dini, sedangkan *false positive* masih dapat dikoreksi melalui pemeriksaan lanjutan. Setelahnya, *decision curve analysis* [23] dijalankan untuk membandingkan *net benefit* lintas *threshold probability* terhadap dua strategi naif sebagai referensi, yaitu *treat-all* dan *treat-none*. Lapisan terakhir adalah audit *subgroup fairness* yang menghitung *recall*, *precision*, F1, dan FN-rate secara terpisah pada sub-populasi berdasarkan jenis kelamin, kelompok usia, dan kategori ras-etnis untuk memeriksa kemungkinan bias algoritmik.

3.5 Interpretasi SHAP

Interpretasi model memakai *TreeExplainer* dari pustaka SHAP yang merupakan implementasi efisien untuk model berbasis pohon [6]. *SHAP value* memberi dekomposisi additif kontribusi setiap fitur terhadap selisih antara prediksi suatu instance dan rata-rata prediksi model di seluruh data. Tiga tingkat interpretasi digunakan. Pada tingkat global, *mean absolute SHAP value* berfungsi sebagai ukuran kepentingan fitur, ditampilkan melalui *beeswarm plot* dan *dependence plot*. Pada tingkat lokal, *waterfall plot* dipakai untuk satu kasus *false negative* representatif. Pada tingkat agregat, rata-rata *SHAP value* (bukan absolut) dihitung lintas seluruh kasus *false negative* untuk memperlihatkan pola fitur yang secara sistematis mendorong prediksi ke kelas Sehat padahal pasien sebenarnya Sakit. Komponen agregat FN tersebut yang melengkapi pembacaan SHAP global yang lazim diadopsi pada literatur sebelumnya.

4. Hasil dan Pembahasan

4.1 Performa Model pada Test set

Tabel 2 menampilkan empat metrik kunci beserta 95% *bootstrap confidence interval* untuk delapan konfigurasi pada *test set* 49205 responden, sementara distribusi kesalahan klasifikasinya tersaji pada Tabel 3 dan divisualisasikan pada Gambar 2. *Baseline* RF mencatat akurasi 94,9%, yang sebagian besar dijelaskan oleh kelas mayoritas, karena *recall* pada kelas positif hanya 21,2%. Artinya, hampir empat dari lima pasien dengan riwayat infark miokard luput dari deteksi. CSRF 1:3 sebagai model utama menukar sedikit *specificity* (0,968 dari 0,991) dengan kenaikan *recall* lebih dari dua kali lipat menjadi 49,4% (95% CI: 47,5–51,2%) dan menghasilkan F1 0,483 (0,466–0,499), nilai tertinggi pada kelompok *Random Forest* dan *resampling*. CS-XGBoost dan CS-LightGBM memperoleh F1 sedikit lebih tinggi, yakni 0,487 dan 0,488, sebagai pembanding berbasis *gradient boosting*; karena fokus studi ini adalah *Cost-sensitive Random Forest* yang lebih sederhana dan interpretabel, CSRF 1:3 tetap diposisikan sebagai model utama dengan trade-off operasional yang seimbang antara sensitivitas dan beban kerja klinis.

Catatan yang patut dicermati dari Tabel 2 berada pada kolom PR-AUC. Seluruh konfigurasi *Cost-sensitive* dan *resampling* berkumpul pada rentang 0,383–0,422, sangat dekat dengan nilai *Baseline* RF sebesar 0,414. Karena PR-AUC adalah ringkasan kapasitas *ranking* model di seluruh *threshold*, kedekatan nilai ini mengisyaratkan bahwa peningkatan *recall* yang teramati pada Tabel 2 sebagian besar bersumber dari pergeseran *operating point*, yaitu titik keputusan yang dipilih pada kurva *precision-recall*, bukan dari perubahan kurva itu sendiri. Pembacaan tersebut sejalan dengan analisis Fernández dkk. [22] yang menyatakan bahwa pada data besar dengan distribusi mendekati populasi sebenarnya, dampak intervensi *imbalance* handling terhadap *ranking* model relatif lebih kecil daripada ketika diuji pada *dataset* miniatur. Implikasi praktisnya tetap perlu dipertimbangkan. Meskipun *ranking* probabilitas relatif setara, pergeseran *operating point* yang dihasilkan *Cost-sensitive learning* menyediakan trade-off *recall-precision* yang lebih sesuai untuk konteks klinis dengan biaya FN dominan, dan trade-off tersebut terbukti stabil pada validasi LOSO sebagaimana dilaporkan di Subbab 4.2.

Tabel 2 Performa delapan konfigurasi model pada test set (95% Bootstrap CI)

Model	Recall	F1	PR-AUC	Specificity
Baseline RF	0,212 (0,197–0,227)	0,311 (0,292–0,330)	0,414 (0,395–0,436)	0,991 (0,990–0,992)
CSRF balanced	0,673 (0,656–0,691)	0,411 (0,398–0,424)	0,411 (0,391–0,433)	0,907 (0,905–0,910)
CSRF 1:3 (utama)	0,494 (0,475–0,512)	0,483 (0,466–0,499)	0,414 (0,395–0,436)	0,968 (0,967–0,970)
SMOTE + RF	0,586 (0,569–0,605)	0,437 (0,423–0,452)	0,385 (0,366–0,405)	0,937 (0,935–0,939)
ADASYN + RF	0,581 (0,564–0,599)	0,441 (0,427–0,455)	0,383 (0,364–0,404)	0,939 (0,937–0,941)
RUS + RF	0,787 (0,772–0,802)	0,319 (0,309–0,329)	0,412 (0,392–0,433)	0,818 (0,815–0,821)
CS-XGBoost 1:3	0,502 (0,483–0,520)	0,487 (0,471–0,502)	0,421 (0,401–0,445)	0,968 (0,966–0,969)
CS-LightGBM 1:3	0,504 (0,485–0,522)	0,488 (0,472–0,504)	0,422 (0,401–0,445)	0,968 (0,966–0,969)

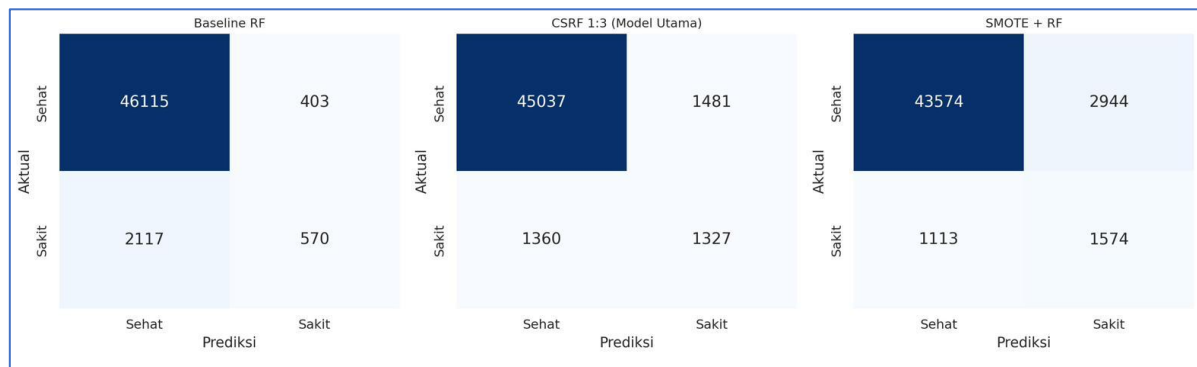
Implikasi praktis dari pola tersebut adalah bahwa pemilihan konfigurasi pada konteks ini lebih akurat dibingkai sebagai pemilihan *operating point* yang sesuai dengan kapasitas rujukan dan toleransi terhadap *false positive*. RUS+RF berada di ujung agresif dengan *recall* 78,7%, tetapi *precision* rendah dan FP membengkak ke 8468 kasus pada *test set*. CSRF balanced berada di posisi tengah-tinggi dengan *recall* 67,3% dan FP 4316. CSRF 1:3 dipilih sebagai *operating point* utama karena menghasilkan F1 tertinggi pada kelompok *Random Forest* dan *resampling* sambil menjaga FP pada level yang lebih realistis (1481) untuk fasilitas dengan kapasitas rujukan terbatas. Pilihan ini secara sadar menukar sebagian *recall* demi pengurangan beban FP dibandingkan CSRF balanced, sebuah trade-off yang dapat dievaluasi ulang sesuai prioritas operasional pada masing-masing lingkungan klinis.

Tabel 3 Confusion matrix counts delapan konfigurasi model pada test set

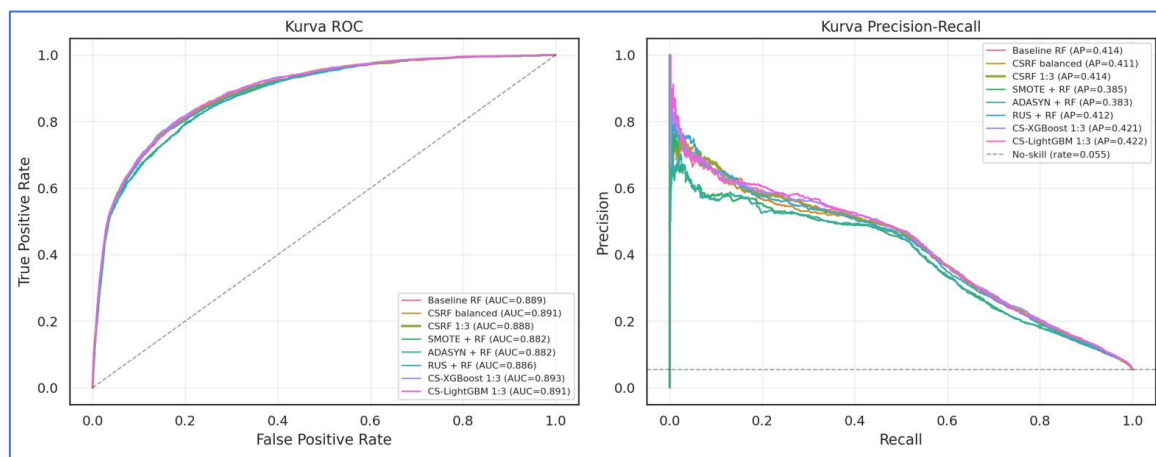
Model	TP	TN	FP	FN
Baseline RF	570	46115	403	2117
CSRF balanced	1809	42202	4316	878
CSRF 1:3 (utama)	1327	45037	1481	1360
SMOTE + RF	1574	43574	2944	1113

Model	TP	TN	FP	FN
ADASYN + RF	1561	43687	2831	1126
RUS + RF	2116	38050	8468	571
CS-XGBoost 1:3	1348	45014	1504	1339
CS-LightGBM 1:3	1355	45007	1511	1332

Visualisasi confusion matrix delapan konfigurasi pada Gambar 2 memperlihatkan pergeseran TP-FN secara langsung. Kurva ROC dan PR pada Gambar 3 menegaskan pengamatan PR-AUC, yaitu kurva-kurva tersebut sebagian besar berimpit, sehingga pergeseran *operating point* tampak sebagai pergerakan titik di sepanjang kurva yang sekaligus mengonfirmasi konsistensi mekanisme antarkonfigurasi.



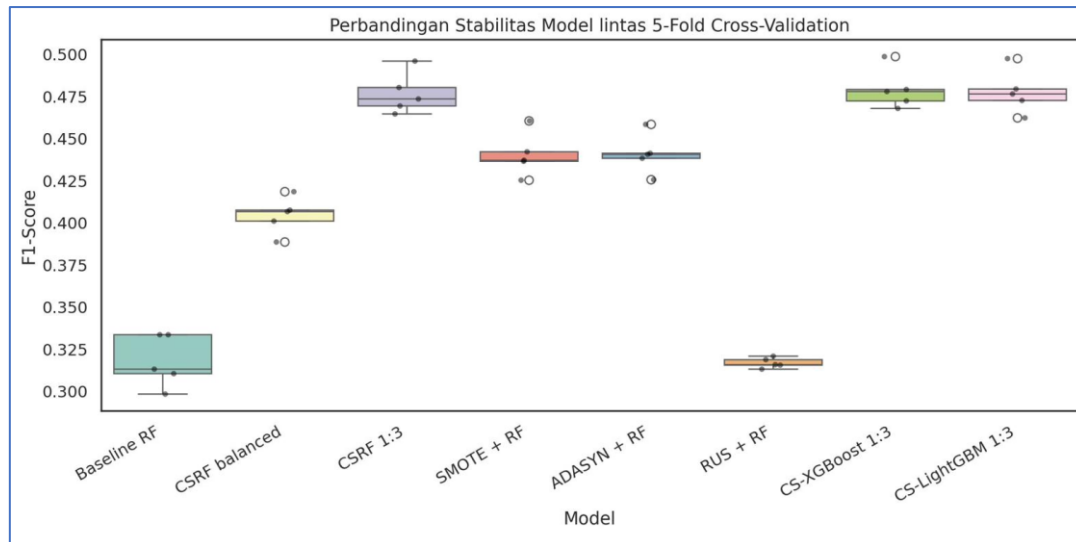
Gambar 2 Confusion matrix delapan konfigurasi model pada *test set*



Gambar 3 Kurva ROC dan *precision-recall* delapan konfigurasi model

4.2 Pengujian Statistik dan Validasi Geografis

Distribusi F1 lintas 5-fold *cross-validation* disajikan pada Gambar 4, dan hasil uji statistik perbandingan setiap konfigurasi terhadap *Baseline* RF tersaji pada Tabel 4. Setelah koreksi Holm dan Benjamini-Hochberg, seluruh konfigurasi *Cost-sensitive* dan *resampling*, kecuali RUS+RF, terbukti secara statistik lebih baik daripada *Baseline* pada metrik F1 lintas fold dengan p-value yang sangat kecil. CSRF 1:3 bersama dua varian *gradient boosting Cost-sensitive* menempati posisi teratas dengan mean difference F1 yang konsisten pada kisaran 0,159–0,161.



Gambar 4 Boxplot distribusi F1 lintas 5-Fold cross-validation

Nilai Cohen's d yang berkisar 4,39–10,39 pada konfigurasi yang signifikan perlu ditempatkan pada konteks ukuran sampelnya. Karena training fold berukuran sekitar 157 ribu instance, variansi F1 antarfold sangat kecil, sehingga effect size yang dihitung sebagai rasio mean difference terhadap standar deviasi antarfold cenderung besar bahkan untuk perbedaan moderat. Indikator yang lebih informatif untuk membaca dampak praktis pada konteks klinis adalah selisih FN absolut dan *recall* pada *test set*, sebagaimana tersaji pada Tabel 3 dan Tabel 2. Selisih FN absolut sebesar 757 kasus (2117 → 1360) pada *test set* 49205 responden menjadi ukuran kebermaknaan operasional utama dari kerangka ini.

Tabel 4 Uji statistik perbandingan F1 lintas fold setiap model terhadap *baseline RF*

Model	Mean Diff	p_raw	p_Holm	p_BH	Cohen's d	Sig
CSRF balanced	0,087	0,001	0,001	0,001	4,39	Ya
CSRF 1:3	0,159	<0,001	<0,001	<0,001	9,95	Ya
SMOTE + RF	0,123	<0,001	0,001	<0,001	6,43	Ya
ADASYN + RF	0,123	<0,001	0,001	<0,001	6,60	Ya
RUS + RF	−0,001	0,883	0,883	0,883	−0,07	Tidak
CS-XGBoost 1:3	0,161	<0,001	<0,001	<0,001	10,39	Ya
CS-LightGBM 1:3	0,160	<0,001	<0,001	<0,001	8,91	Ya

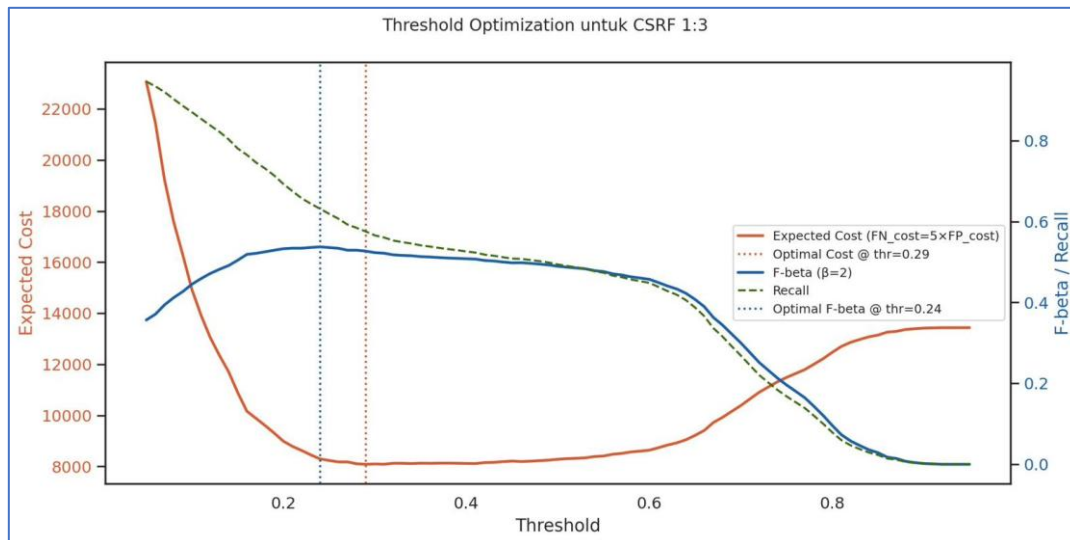
Leave-one-state-out validation lintas 53 yurisdiksi menyediakan bukti generalisasi yang lebih ketat sebagaimana ditampilkan pada Tabel 5. CSRF 1:3 mempertahankan *recall* rata-rata 0,495 dengan deviasi standar 0,047, nyaris identik dengan hasil pada *test set* independen. *Baseline RF* tetap berada di sekitar 0,22 dengan deviasi yang lebih lebar. Hal ini memperlihatkan bahwa *operating point* yang dihasilkan pembobotan kelas pada CSRF 1:3 stabil meskipun yurisdiksi penguji berganti, sehingga keunggulan operasional yang teramati di Subbab 4.1 dapat dipertahankan lintas distribusi geografis.

Tabel 5 Rangkuman *leave-one-state-out* validation lintas 53 yurisdiksi

Model	n_states	Recall mean	Recall std	F1 mean	F1 std	PR-AUC mean
Baseline RF	53	0,2185	0,0387	0,3178	0,0451	0,4176
CSRF 1:3	53	0,4946	0,0474	0,4772	0,0393	0,4192

4.3 Trade-off Operasional: *Threshold*, *Decision Curve*, dan Kalibrasi

Pencarian *threshold* optimum berbasis *cost matrix* dengan CFN = 5 dan CFP = 1 menemukan nilai 0,29 sebagai titik yang meminimalkan *expected cost* pada CSRF 1:3. Apabila yang dimaksimalkan adalah F-beta ($\beta = 2$), *threshold* optimum bergeser ke 0,24, lebih agresif menukar *precision* dengan *recall* daripada minimisasi *expected cost*. Kurva total cost terhadap *threshold* (Gambar 5) memperlihatkan dataran landai antara 0,25–0,35, sehingga kebijakan operasional pada rentang tersebut praktis ekuivalen.



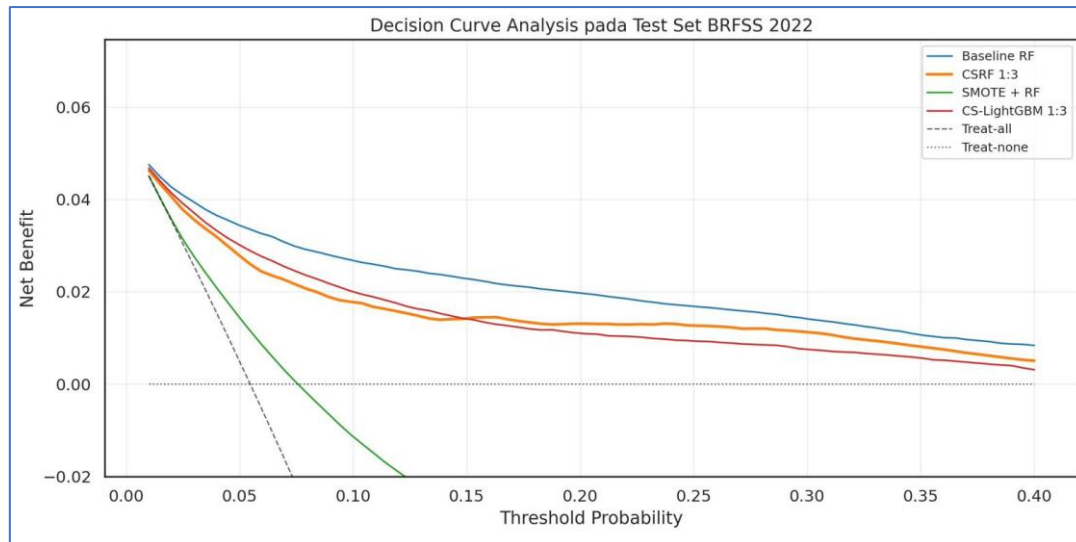
Gambar 5 Kurva *threshold* vs total cost (C_FN = 5, C_FP = 1) untuk CSRF 1:3

Sensitivitas terhadap rasio biaya disajikan pada Tabel 6. Peningkatan bobot dari 1:3 ke 1:10 menaikkan *recall* dari 0,494 ke 0,619 dan menurunkan FN dari 1360 ke 1025, sementara FP membengkak dari 1481 ke 3238. F1 mencapai puncak pada rasio 1:3 (0,483) dan menurun pada rasio 1:10 (0,438). Implikasi praktisnya bergantung pada tujuan operasional. Pada fasilitas dengan kapasitas rujukan terbatas, rasio 1:3 lebih realistis. Pada konteks ketika keselamatan pasien diprioritaskan secara mutlak dan biaya rujukan rendah, rasio 1:10 lebih sesuai. Hasil tersebut sekaligus memberi justifikasi empiris bagi pemilihan rasio 1:3 sebagai *operating point* utama studi ini.

Tabel 6 *Sensitivity analysis* rasio biaya pada CSRF 1:3

Cost Ratio	Recall	Precision	F1	PR-AUC	FN	FP
1:3 (utama)	0,494	0,473	0,483	0,414	1360	1481
1:5	0,528	0,442	0,481	0,413	1268	1793
1:10	0,619	0,339	0,438	0,409	1025	3238

Decision curve analysis (Gambar 6) memperluas perspektif tersebut dengan menyajikan *net benefit* lintas *threshold probability* dibandingkan dua referensi naif, yaitu *treat-all* dan *treat-none*. CSRF 1:3 unggul atas *Baseline* RF pada hampir seluruh rentang *threshold* yang relevan secara klinis (0,05–0,30) dan menghasilkan *net benefit* positif yang lebih tinggi daripada *treat-all* pada region *threshold* $\geq 0,06$. Hasil DCA mempertegas keunggulan CSRF 1:3 di luar metrik agregat, dengan utilitas klinis yang nyata pada region keputusan yang masuk akal.

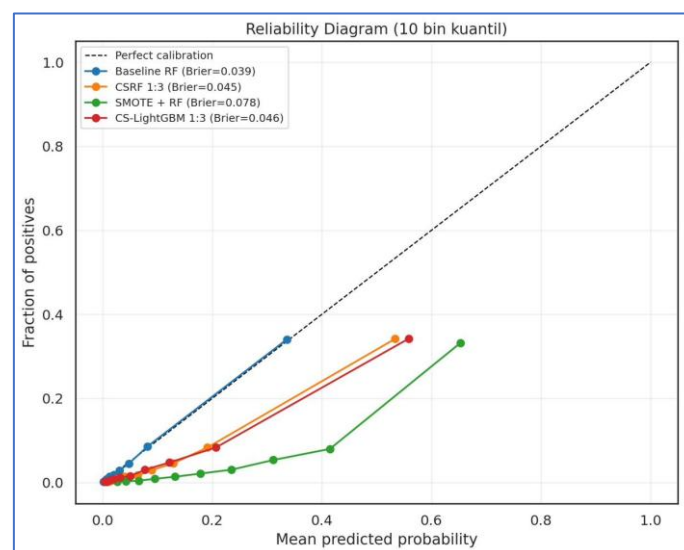


Gambar 6 Decision curve analysis: net benefit lintas threshold probability untuk CSRF 1:3, Baseline RF, dan dua strategi naif

Aspek terakhir dari trade-off operasional menyangkut kalibrasi probabilitas. Model *Cost-sensitive* cenderung menghasilkan estimasi probabilitas yang bias ke atas pada kelas minoritas karena pembobotan eksplisit di dalam fungsi *impurity*. Tabel 7 memperlihatkan dampak kalibrasi *isotonic* pada *threshold default* 0,5, yaitu Brier score membaik dari 0,0454 menjadi 0,0390 namun *recall* menurun ke 19,8% karena distribusi probabilitas yang tergeser. Pola ini bukanlah kelemahan kalibrasi, melainkan konsekuensi natural bahwa *threshold* operasional perlu ditentukan ulang sesudah kalibrasi. Untuk implementasi lapangan, kalibrasi dan pemilihan *threshold* sebaiknya digabungkan dalam satu rangkaian optimasi, yakni kalibrasi terlebih dahulu, kemudian *threshold* dicari ulang pada distribusi probabilitas yang sudah terkalibrasi mengikuti *cost matrix* klinis yang sama. *Joint optimization* tersebut menjadi salah satu agenda pengembangan teknis pada penerapan lanjutan model ini.

Tabel 7 Dampak kalibrasi isotonic terhadap metrik operasional CSRF 1:3 pada *threshold* 0,5

Stage	Brier	Recall	F1	PR-AUC	FN	FP
Sebelum kalibrasi	0,0454	0,494	0,483	0,414	1360	1481
Sesudah isotonic	0,0390	0,198	0,295	0,414	2156	384

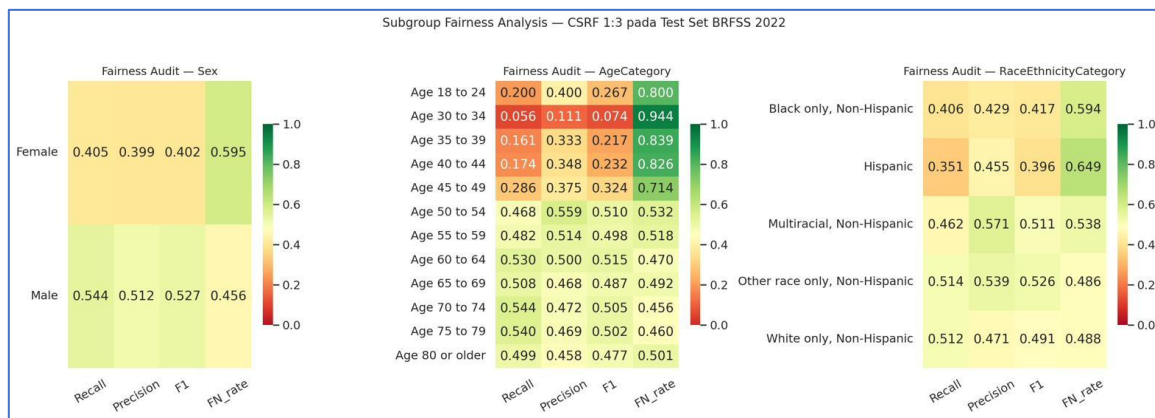


Gambar 7 Kurva kalibrasi sebelum dan sesudah penyesuaian isotonic untuk CSRF 1:3

4.4 Audit Subgroup fairness

Audit *fairness* pada Gambar 8 menampakkan pola disparitas yang penting untuk diantisipasi sebelum penerapan. Pada dimensi jenis kelamin, *recall* responden perempuan sebesar 0,405 sedangkan laki-laki 0,544, sebuah kesenjangan sekitar 14 poin persentase. Pada dimensi usia, *recall* pada kelompok muda (18–24, 30–34, 35–39 tahun) berada di bawah 0,21, sedangkan pada kelompok 50 tahun ke atas konsisten di kisaran 0,47–0,55. Pola tersebut sebagian mencerminkan prevalensi yang sebenarnya, karena kasus positif pada usia muda hanya sekitar 0,4% sampel, sehingga model memiliki sinyal yang sangat tipis untuk dipelajari. Sebagian lainnya berkaitan dengan struktur fitur, ketika variabel klinis dominan seperti HadAngina dan ChestScan memang lebih sering muncul pada usia lanjut.

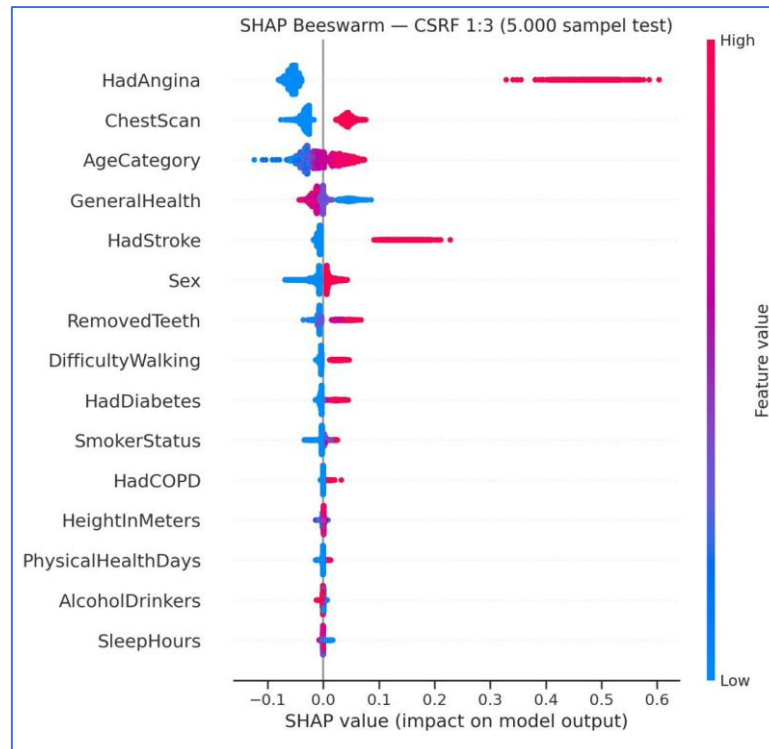
Pada dimensi ras-etnis, *recall* subgroup White (0,512) dan Other (0,514) lebih tinggi daripada Hispanic (0,351) dan Black (0,406). Disparitas semacam ini menjadi sinyal bahwa kalibrasi *fairness-aware* perlu dipertimbangkan ketika model diterapkan pada populasi heterogen, misalnya lewat *threshold* per subgroup atau *equalized odds post-processing*. Hasil audit dilaporkan secara transparan sebagai masukan untuk pengembangan mitigasi pada studi lanjutan.



Gambar 8 Heatmap subgroup fairness CSRF 1:3 berdasarkan jenis kelamin, kelompok usia, dan kategori ras-etnis

4.5 Interpretasi SHAP dan Diseksi False negative

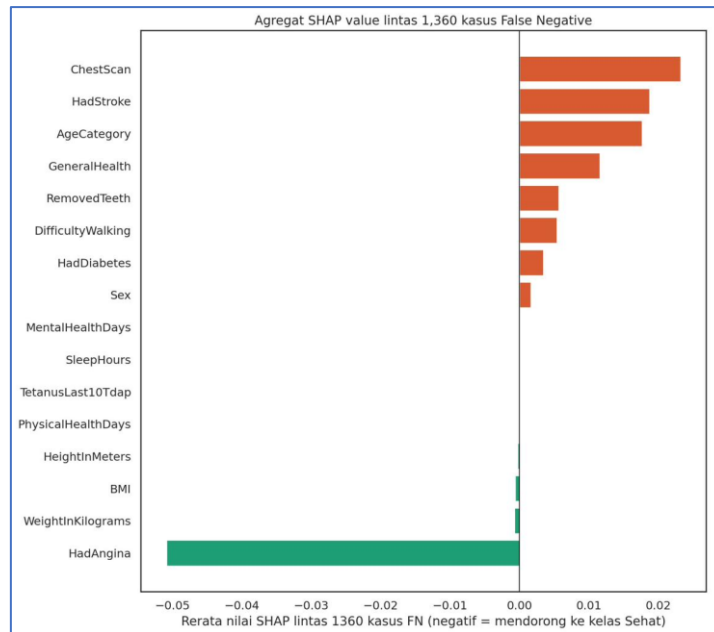
Mean absolute *SHAP value* lintas *test set* memunculkan tujuh fitur dominan, yaitu HadAngina (0,082), ChestScan (0,037), AgeCategory (0,028), GeneralHealth (0,017), HadStroke (0,014), Sex (0,014), dan RemovedTeeth (0,012). *Beeswarm plot* pada Gambar 9 memperlihatkan pola kontribusi yang sejalan dengan pemahaman klinis. Riwayat angina memberi kontribusi positif besar pada kelas Sakit, ChestScan positif menggeser prediksi ke arah Sakit, dan usia yang lebih tua secara monoton menaikkan probabilitas Sakit. Konsistensi semacam ini memberi validasi muka (*face validity*) yang penting untuk penerimaan model di lingkungan klinis [17], [19].



Gambar 9 Beeswarm plot SHAP untuk sepuluh fitur teratas pada CSRF 1:3

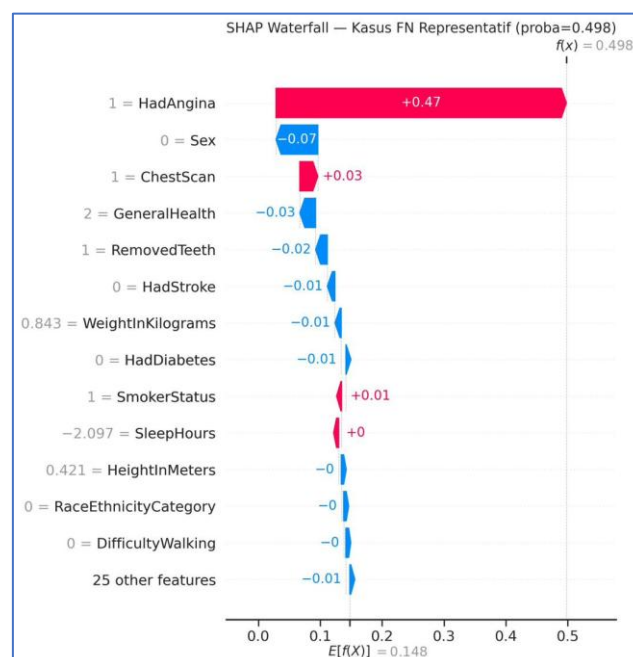
Posisi HadAngina dan ChestScan yang dominan pada SHAP global perlu dibaca dengan cermat. Angina pectoris secara klinis adalah manifestasi simptomatik dari iskemia miokard, proses patologis yang sama yang dapat berujung pada infark miokard. ChestScan dalam instrumen BRFSS perlu dibaca sebagai riwayat CT/CAT scan area dada yang dapat mencerminkan paparan layanan kesehatan atau proses skrining tertentu, bukan sebagai indikator langsung diagnosis kardiovaskular. Karena itu, kontribusi SHAP yang tinggi pada ChestScan lebih tepat dipahami sebagai sinyal asosiatif dalam data survei *cross-sectional*, bukan bukti mekanisme kausal terhadap infark miokard. Dengan pembacaan tersebut, kedua variabel ini lebih tepat dipahami sebagai sinyal yang berkaitan erat dengan target, dan capaian performa model perlu dibingkai sebagai klasifikasi *cross-sectional* terhadap asosiasi gejala. Pembacaan ini sejalan dengan karakter *dataset* BRFSS sebagai survei *self-report* yang merekam riwayat klinis pada saat survei dilakukan, dan secara alamiah menempatkan model pada peran *risk stratification* operasional, bukan prediksi prospektif. Studi lanjutan yang mengeksplorasi performa model tanpa kedua fitur tersebut akan memberi estimasi kapasitas prediktif fitur non-klinis (gaya hidup, demografi, indikator kesehatan umum) secara terpisah, sebuah arah pengembangan yang dibahas pada Subbab 4.6. *Dependence plot* memperlihatkan beberapa interaksi yang bermakna, antara lain AgeCategory dengan PneumoVaxEver ($|corr| = 0,445$), GeneralHealth dengan PhysicalHealthDays ($|corr| = 0,514$), dan ChestScan dengan GeneralHealth. Pola interaksi tersebut sejalan dengan logika klinis. Pasien dengan riwayat ChestScan umumnya juga memiliki GeneralHealth yang lebih rendah, dan pasien usia lanjut cenderung lebih sering mendapat konsultasi PneumoVax.

Investigasi terhadap 1360 kasus *false negative* adalah lapisan interpretabilitas yang membedakan studi ini dari literatur sebelumnya. Gambar 10 menampilkan agregat *SHAP value* (bukan absolut) pada seluruh FN, sebuah orientasi yang penting karena memperlihatkan arah kontribusi, bukan sekadar magnitudo. Pola dominan terlihat pada HadAngina dengan rata-rata *SHAP value* $-0,051$. Artinya, pada kasus FN, fitur HadAngina umumnya bernilai negatif (tanpa riwayat angina), dan ini yang paling kuat mendorong model menuju kelas Sehat. Fitur lain seperti ChestScan, HadStroke, AgeCategory, dan GeneralHealth memberi dorongan kecil ke arah Sakit, tetapi magnitudonya belum cukup untuk mengompensasi absennya sinyal angina.



Gambar 10 Agregat SHAP value pada 1360 kasus false negative CSRF 1:3

Profil klinis kasus FN memperkuat pembacaan tersebut. Sebanyak 57,7% di antaranya berjenis kelamin laki-laki, 33,9% berusia 70 tahun ke atas (terdiri dari 17,9% pada kategori 80+ dan 16,2% pada kategori 70–74), namun 80,5% tidak memiliki riwayat COPD, 83,3% tidak memiliki riwayat stroke, dan 67,4% tidak memiliki riwayat diabetes. GeneralHealth pada FN bermodus Good (37,0%) atau Fair (25,4%). Profil seperti ini menggambarkan pasien yang secara superfisial tampak baik-baik saja, tidak memiliki kondisi komorbid yang biasanya menjadi sinyal kuat, tetapi nyatanya pernah mengalami infark miokard. Untuk populasi semacam itu, model memerlukan fitur tambahan di luar checklist riwayat penyakit, misalnya biomarker laboratoris, parameter EKG, atau riwayat keluarga, yang memang tidak tersedia pada BRFSS sebagai survei berbasis populasi. Hasil diseksi forensik kasus FN ini memberi arah konkret untuk pengembangan *dataset* dan model selanjutnya.



Gambar 11 Waterfall plot SHAP pada salah satu kasus false negative representatif (probabilitas prediksi = 0,498)

Dekomposisi prediksi pada satu kasus FN representatif (Gambar 11) memperlihatkan dinamika tersebut secara mikroskopis. Probabilitas prediksi mencapai 0,498, nyaris menyentuh *threshold* 0,5, sehingga kasus ini tergelincir tipis ke kelas Sehat. HadAngina = 1 memberi dorongan +0,47, ChestScan = 1 memberi dorongan +0,03, sedangkan Sex = 0 (perempuan) menariknya -0,07 dan kombinasi GeneralHealth = 2 (Good), RemovedTeeth = 1, serta beberapa fitur lain memberi kontribusi negatif kumulatif yang cukup menahan prediksi di bawah *threshold*. Perbedaan antara pola agregat *false negative* dan contoh lokal pada *waterfall plot* menunjukkan bahwa SHAP perlu dibaca pada dua level berbeda. Analisis agregat menggambarkan kecenderungan umum seluruh kasus *false negative*, sedangkan *waterfall plot* menjelaskan dinamika prediksi pada satu kasus borderline tertentu. Kasus borderline semacam ini secara klinis dapat tertangkap dengan *threshold* operasional yang lebih rendah, dan hasil sensitivitas pada Tabel 6 menunjukkan rasio cost yang lebih agresif memang menyelamatkan sebagian kasus seperti ini dengan beban FP yang lebih besar.

4.6 Posisi Penelitian dan Keterbatasan

Kerangka evaluasi yang disusun di studi ini menjawab tiga keterbatasan studi terdahulu sekaligus pada skala populasi yang representatif. Anggaran *tuning* yang seragam menekan bias yang umum muncul ketika model utama disesuaikan secara khusus sementara *baseline* berjalan dengan parameter *default*. *Bootstrap* CI menyatakan ketidakpastian sampling secara eksplisit pada setiap metrik. Validasi LOSO lintas 53 yurisdiksi memperlihatkan bahwa keunggulan operasional CSRF 1:3 stabil di seluruh distribusi geografis. *Threshold* optimization menerjemahkan asimetri biaya klinis menjadi level keputusan operasional, dilengkapi *decision curve analysis* yang menyediakan perspektif *net benefit* lintas preferensi. Audit *fairness* dan SHAP agregat pada kasus FN adalah komponen yang jarang dilaporkan pada literatur prediksi penyakit jantung berbahasa Indonesia, dan keduanya memberi arah praktis untuk pengembangan iteratif.

Sejumlah ancaman validitas perlu disampaikan secara terbuka sebagai bahan pertimbangan untuk replikasi. Sebagai survei *self-report*, BRFSS rentan terhadap *recall bias* dan *underreporting*, khususnya pada variabel sensitif. Fitur klinis yang tersedia juga terbatas pada variabel survei dan tidak mencakup biomarker laboratoris seperti kolesterol atau troponin. Audit *fairness* memperlihatkan disparitas yang membutuhkan mitigasi formal pada tahap *pre-deployment*. Kalibrasi probabilitas memerlukan *joint optimization* dengan *threshold* sebagaimana telah diuraikan pada Subbab 4.3. Anggaran *tuning* sebesar 15 iterasi RandomizedSearchCV ditetapkan agar setara antarkonfigurasi, sehingga karakter “anggaran seragam” lebih tepat dipakai daripada karakter “pencarian menyeluruh”. *Encoding* label sederhana pada variabel nominal mengikuti praktik umum pada literatur *Random Forest* dengan BRFSS, dan perbandingan dengan skema *encoding* alternatif menjadi salah satu agenda *ablation*. Konteks geografis *dataset* adalah Amerika Serikat, sehingga transferabilitas ke populasi Indonesia menuntut validasi eksternal pada data lokal.

Pengembangan ke depan diarahkan ke empat agenda yang saling berkait. Pertama, *ablation* tanpa fitur klinis dominan (HadAngina, ChestScan, HadStroke) untuk memetakan kapasitas prediktif fitur gaya hidup dan demografi secara terpisah, sehingga dapat dinilai sejauh mana model tetap mampu mengklasifikasikan profil risiko ketika sinyal klinis yang paling dekat dengan target dihilangkan. Kedua, *joint optimization* antara kalibrasi dan *threshold* pada satu *pipeline* operasional. Ketiga, mitigasi *fairness* berbasis *threshold* per subgroup atau *equalized odds post-processing*. Keempat, integrasi biomarker laboratoris serta validasi eksternal pada populasi Indonesia. Kerangka evaluasi yang disusun pada studi ini dirancang agar mudah direplikasi untuk keempat agenda tersebut.

5. Kesimpulan

Penelitian ini menyusun dan menguji kerangka evaluasi metodologis untuk klasifikasi profil risiko kardiovaskular *self-report* berbasis *Cost-sensitive Random Forest* yang dipadukan SHAP, dengan BRFSS 2022 berisi 246022 responden dan rasio kelas 1:17,31 sebagai data uji. CSRF 1:3 sebagai *operating point* penyeimbang menurunkan *false negative* dari 2117 menjadi 1360 pada *test set* 49205

responden, setara perbaikan 757 kasus absolut atau 35,8% secara relatif dibandingkan *Baseline* RF, dengan *recall* 49,4% (95% CI: 47,5–51,2%) dan F1 0,483 (0,466–0,499) yang menjadi nilai tertinggi pada kelompok *Random Forest* dan *resampling*. Perbedaan tersebut signifikan secara statistik sesudah koreksi Holm dan Benjamini-Hochberg ($p < 0,001$) dan stabil pada validasi *leave-one-state-out* lintas 53 yurisdiksi. Konsistensi mekanisme *Cost-sensitive* juga ditegaskan oleh CS-XGBoost dan CS-LightGBM yang menghasilkan F1 berimpit pada 0,487 dan 0,488.

Tiga kontribusi penelitian ini saling melengkapi. Pemetaan empiris memperlihatkan bahwa pada skala populasi, kenaikan *recall* yang diperoleh dari intervensi *Cost-sensitive* dan *resampling* sebagian besar dijelaskan oleh pergeseran *operating point* pada kurva *precision-recall*, sehingga pemilihan konfigurasi paling akurat dibingkai sebagai pemilihan *operating point* yang sesuai dengan kapasitas rujukan klinis. Diseksi forensik terhadap 1360 kasus *false negative* memperlihatkan profil pasien yang secara superfisial tampak sehat, yaitu tidak memiliki riwayat angina, COPD, stroke, maupun diabetes secara dominan, sehingga mengindikasikan arah pengembangan menuju integrasi biomarker laboratoris yang tidak tersedia pada survei populasi. Pembacaan kritis terhadap posisi dominan HadAngina dan ChestScan pada SHAP global menempatkan capaian performa model pada konteks *self-report cross-sectional* yang lebih sesuai, dengan model diposisikan sebagai *decision-support* untuk *risk stratification* operasional. Audit *fairness* memperlihatkan disparitas yang dilaporkan secara transparan sebagai masukan untuk mitigasi pada tahap *pra-deployment*.

Kontribusi metodologis penelitian ini terletak pada perpaduan kerangka evaluasi berlapis dan analisis forensik kelompok *false negative* yang dapat direplikasi pada domain medis lain dengan karakter data populasi serupa. Pengembangan selanjutnya akan diarahkan ke empat agenda yang saling melengkapi sebagaimana diuraikan pada Subbab 4.6, yakni *ablation* fitur klinis dominan, *joint optimization* kalibrasi dan *threshold*, mitigasi *fairness post-processing*, serta integrasi biomarker laboratoris dan validasi eksternal pada populasi Indonesia.

Referensi

- [1] Global Burden of Cardiovascular Diseases and Risks 2023 Collaborators, "Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990–2023," *J. Am. Coll. Cardiol.*, vol. 86, no. 22, pp. 2167–2243, 2025, doi: 10.1016/j.jacc.2025.08.015.
- [2] K. Pytlak, "Indicators of heart disease (2022 update)," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>.
- [3] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [4] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced *datasets* in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.
- [5] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artif. Intell. Rev.*, vol. 57, no. 4, art. 80, 2024, doi: 10.1007/s10462-023-10652-8.
- [6] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
- [7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [8] M. Zhu, J. Xia, X. Jin, M. Yan, G. Cai, J. Yan, and G. Ning, "Class weights *Random Forest* algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018, doi: 10.1109/ACCESS.2018.2789428.
- [9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017, pp. 4768–4777.

- [10] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced *datasets*," *PLoS ONE*, vol. 10, no. 3, art. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [11] A. A. Noor, A. Manzoor, M. D. M. Qureshi, M. A. Qureshi, and W. Rashwan, "Unveiling explainable AI in healthcare: current trends, challenges, and future directions," *WIREs Data Min. Knowl. Discov.*, vol. 15, no. 2, art. e70018, 2025, doi: 10.1002/widm.70018.
- [12] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998, doi: 10.1162/089976698300017197.
- [13] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through *machine learning* techniques and optimization," *Processes*, vol. 11, no. 4, art. 1210, 2023, doi: 10.3390/pr11041210.
- [14] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization," *Algorithms*, vol. 16, no. 6, art. 308, 2023, doi: 10.3390/a16060308.
- [15] A. Naik, G. G. Tejani, and S. J. Mousavirad, "SGO enhanced *Random Forest* and extreme gradient boosting framework for heart disease prediction," *Sci. Rep.*, vol. 15, art. 18145, 2025, doi: 10.1038/s41598-025-02525-7.
- [16] H. Sadr, A. Salari, M. T. Ashoobi, and M. Nazari, "Cardiovascular disease diagnosis: a holistic approach using the integration of *machine learning* and deep learning models," *Eur. J. Med. Res.*, vol. 29, no. 1, art. 455, 2024, doi: 10.1186/s40001-024-02044-7.
- [17] C.-f. Chen, Z.-y. Ren, H.-h. Zong, Y.-t. Xiong, and Y. Hong, "Development and validation of explainable *machine learning* models for predicting 3-month functional outcomes in acute ischemic stroke: a SHAP-based approach," *Front. Neurol.*, vol. 16, art. 1678815, 2025, doi: 10.3389/fneur.2025.1678815.
- [18] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine learning-based predictive models for detection of cardiovascular diseases," *Diagnostics*, vol. 14, no. 2, art. 144, 2024, doi: 10.3390/diagnostics14020144.
- [19] B. Chulde-Fernández, D. Enríquez-Ortega, C. Guevara, P. Navas, A. Tirado-Espín, P. Vizcaíno-Imacaña, F. Villalba-Meneses, C. Cadena-Morejón, D. Almeida-Galárraga, and P. Acosta-Vargas, "Classification of heart failure using machine learning: a comparative study," *Life*, vol. 15, no. 3, art. 496, 2025, doi: 10.3390/life15030496.
- [20] N. Selayanti, S. A. Putri, M. Kristanaya, M. P. Azzahra, M. G. Navsih, and K. M. Hindrayani, "Penerapan *machine learning* algoritma *Random Forest* untuk prediksi penyakit jantung," *Prosiding Semin. Nas. Sains Data*, vol. 4, no. 1, pp. 895–906, 2024, doi: 10.33005/senada.v4i1.376.
- [21] C. Elkan, "The foundations of *Cost-sensitive learning*," in *Proc. 17th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Seattle, WA, USA, 2001, vol. 2, pp. 973–978.
- [22] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-98074-4.
- [23] A. J. Vickers and E. B. Elkin, "*Decision curve analysis*: a novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.
- [24] T. Chen and C. Guestrin, "*XGBoost*: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.