

Jenis artikel	Artikel Penelitian / Artikel Tinjauan / Technical Note / Studi Kasus	ID artikel	TEK-2026-0005
Tautan permanen	https://doi.org/10.xxxx/teknovia.xxxx	Riwayat artikel	Masuk: 10-07-2026 Revisi: 11-07-2026
Keputusan & terbit	Diterima: 31-07-2026 Terbit: 01-07-2026	Lisensi	Creative Commons Attribution 4.0 International

Analisis Perbedaan Prediksi Penyakit Liver Berdasarkan Gender Menggunakan Machine Learning

Analysis of Gender-Based Differences in Liver Disease Prediction Using Machine Learning

Sumiati^{1*}, Virda Mega Ayu², Harianto³, Hasby Ashshiddiqie Kusuma⁴, Mizan Faqih Fais Salim⁵

¹Ilmu Komputer, Universitas Pakuan, BOGOR, Kode Pos, 16143, INDONESIA

^{2,3,4,5}, Sistem dan Teknologi Informasi, Institut Teknologi dan Bisnis Dewantara ,Bogor, 16913 , Indonesia

ORCID: Penulis 1 0000-0000-0000-0000 | Penulis 2 0000-0000-0000-0000 | Penulis 3 0000-0000-0000-0000

***Penulis korespondensi: 1072925003@student.unpak.ac.id**

[Hapus petunjuk ini sebelum mengirim naskah. Gunakan nama penulis tanpa gelar. Pastikan urutan, afiliasi, surel, dan ORCID sama persis dengan metadata OJS.]

Abstract

The evaluation of machine learning (ML) models in clinical prediction typically relies on aggregate metrics, which can obscure performance disparities across subgroups. This study investigates gender-based disparities in liver disease prediction using the Indian Liver Patient Dataset. Three algorithms—Logistic Regression, Random Forest, and XGBoost—are evaluated under three experimental settings: (1) a global model with gender as a feature, (2) a global model without gender, and (3) gender-specific models trained separately for each subgroup. The results show that the best overall model is XGBoost without the gender feature, achieving an F1-score of 0.8402 and an AUC of 0.7782. However, all configurations consistently yield higher recall for male patients than for female patients, indicating a systematic bias in disease detection. Removing the gender feature does not eliminate this disparity, suggesting that bias is embedded within the distribution of clinical features. A subgroup-specific modeling approach using Random Forest substantially reduces this disparity, achieving the lowest female false negative rate (FNR) of 0.1111 and the smallest recall gap (+0.0649). These findings highlight that subgroup-based evaluation is essential for validating AI models in healthcare and demonstrate that population-aware modeling strategies can improve clinical fairness without significantly compromising overall performance

Keywords: machine learning, liver disease, fairness, gender bias, subgroup evaluation, false negative rate, XGBoost, Random Forest

Abstrak

Evaluasi model machine learning (ML) dalam prediksi klinis umumnya bergantung pada metrik agregat, sehingga berpotensi menyembunyikan ketimpangan kinerja antar subkelompok. Penelitian ini menginvestigasi disparitas prediksi penyakit liver berbasis gender menggunakan Indian Liver Patient Dataset. Tiga algoritma Logistic Regression, Random Forest, dan XGBoost dievaluasi melalui tiga skenario eksperimen: (1) model global dengan fitur gender, (2) model global tanpa fitur gender, dan (3) model terpisah per gender. Hasil menunjukkan bahwa model terbaik secara agregat adalah XGBoost tanpa fitur gender dengan F1-score sebesar 0,8402 dan AUC 0,7782. Namun, seluruh konfigurasi secara konsisten menghasilkan recall yang lebih tinggi pada pasien

laki-laki dibandingkan perempuan, mengindikasikan bias sistematis dalam deteksi penyakit. Penghapusan fitur gender tidak mengeliminasi disparitas tersebut, menunjukkan bahwa bias terinternalisasi dalam distribusi fitur klinis. Pendekatan model terpisah per gender menggunakan Random Forest secara signifikan mengurangi ketimpangan, dengan false negative rate (FNR) perempuan terendah sebesar 0,1111 dan selisih recall paling kecil (+0,0649). Temuan ini menegaskan bahwa evaluasi berbasis subkelompok merupakan komponen esensial dalam validasi model AI di bidang kesehatan, serta menunjukkan bahwa strategi pemodelan yang mempertimbangkan karakteristik subpopulasi dapat meningkatkan keadilan klinis tanpa mengorbankan performa secara signifikan.

Kata kunci: machine learning, penyakit liver, fairness, gender bias, subgroup evaluation, false negative rate, XGBoost, Random Forest

1. Pendahuluan

Morbiditas dan mortalitas global, dengan beban penyakit yang terus meningkat terutama di negara berkembang. Deteksi dini melalui analisis parameter biokimia darah seperti bilirubin, enzim hati, dan protein serum memiliki peran krusial dalam mencegah progresivitas penyakit dan komplikasi lanjutan. Dalam beberapa tahun terakhir, penerapan machine learning (ML) pada data rekam medis telah menunjukkan potensi signifikan dalam meningkatkan akurasi skrining dan diagnosis penyakit liver secara otomatis [1][2].

Sejumlah penelitian telah mengimplementasikan berbagai algoritma klasifikasi seperti Logistic Regression, Random Forest, dan XGBoost untuk memprediksi penyakit liver dengan performa yang relatif tinggi berdasarkan metrik agregat seperti akurasi, F1-score, dan area under the ROC curve (AUC) [3][4]. Namun demikian, pendekatan evaluasi berbasis metrik agregat ini memiliki keterbatasan mendasar, karena berpotensi menyembunyikan ketimpangan performa model pada subkelompok populasi tertentu.

Dalam konteks sistem kesehatan berbasis AI, isu fairness atau keadilan algoritmik menjadi perhatian yang semakin penting. Studi oleh [5] menunjukkan bahwa model prediksi kesehatan dapat menghasilkan bias sistematis terhadap kelompok tertentu akibat representasi data yang tidak seimbang. Penelitian yang lebih baru juga menegaskan bahwa bias dalam model ML medis dapat muncul bahkan ketika atribut sensitif seperti gender dihilangkan dari fitur input, karena informasi tersebut secara implisit terkandung dalam variabel klinis lainnya [6] [7].

Dalam kasus penyakit liver, perbedaan biologis dan fisiologis antara laki-laki dan perempuan dapat memengaruhi distribusi biomarker hati, seperti kadar enzim ALT, AST, dan alkaline phosphatase [8]. Selain itu, ketidakseimbangan representasi gender dalam dataset medis seringkali menyebabkan model lebih optimal dalam memprediksi kelompok mayoritas, sehingga meningkatkan risiko kesalahan diagnosis pada kelompok minoritas, khususnya perempuan. Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis apakah model machine learning untuk prediksi penyakit liver menunjukkan perbedaan performa yang signifikan antara pasien laki-laki dan perempuan. Untuk menjawab pertanyaan ini, penelitian mengadopsi pendekatan evaluasi berbasis subkelompok (subgroup evaluation) yang secara eksplisit mengukur dimensi fairness dalam kinerja model.

Secara spesifik, penelitian ini mengkaji tiga pertanyaan utama:

1. Apakah algoritma Logistic Regression, Random Forest, dan XGBoost menghasilkan performa prediksi yang berbeda antara pasien laki-laki dan perempuan?
2. Apakah evaluasi berbasis subkelompok mampu mengungkap masalah fairness yang tidak terdeteksi melalui metrik agregat?
3. Apakah pendekatan model terpisah per gender dapat meningkatkan keadilan prediksi dibandingkan model global?

Kontribusi utama penelitian ini meliputi:

1. Perancangan tiga skenario eksperimen yang sistematis untuk mengevaluasi pengaruh gender dalam prediksi penyakit liver

2. Evaluasi komprehensif metrik fairness berbasis subkelompok, termasuk selisih recall, selisih F1-score, dan false negative rate (FNR) per gender
3. Identifikasi konfigurasi model yang mampu menyeimbangkan performa agregat dan keadilan subkelompok dalam konteks klinis

Temuan dari penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan sistem AI kesehatan yang lebih adil, transparan, dan bertanggung jawab, khususnya dalam aplikasi prediksi penyakit berbasis data klinis.

2. Kajian Terkait dan Landasan Teori

2.1 Prediksi Penyakit Liver dengan Machine Learning

Penerapan machine learning (ML) dalam prediksi penyakit liver telah berkembang pesat dalam beberapa tahun terakhir, terutama dengan memanfaatkan dataset klinis seperti Indian Liver Patient Dataset dan dataset rekam medis serupa. Studi-studi terkini menunjukkan bahwa algoritma klasifikasi mampu mengidentifikasi pola kompleks dari parameter biokimia darah untuk mendukung diagnosis penyakit liver secara otomatis. Penelitian oleh [1][2] menunjukkan bahwa algoritma berbasis ensemble learning seperti Random Forest dan Gradient Boosting secara konsisten memberikan performa yang lebih tinggi dibandingkan model klasik seperti Logistic Regression dan Naive Bayes, terutama dalam hal akurasi dan area under the ROC curve (AUC). Selain itu, [4] menekankan pentingnya pendekatan explainable AI untuk meningkatkan interpretabilitas model dalam konteks klinis. Dari sisi fitur, parameter enzim hati seperti Alanine Aminotransferase (ALT/SGPT) dan Aspartate Aminotransferase (AST/SGOT), serta indikator bilirubin (total dan direct), secara konsisten diidentifikasi sebagai prediktor utama dalam berbagai studi [3],[7]. Penggunaan teknik rekayasa fitur dan transformasi distribusi juga terbukti dapat meningkatkan stabilitas dan performa model, khususnya pada dataset dengan distribusi yang tidak seimbang. Meskipun demikian, sebagian besar penelitian masih berfokus pada optimasi performa global model tanpa mempertimbangkan variasi performa antar subkelompok populasi, seperti perbedaan berdasarkan gender atau karakteristik demografis lainnya.

2.2 Fairness dalam Machine Learning untuk Kesehatan

Fairness atau keadilan dalam sistem machine learning, khususnya di bidang kesehatan, telah menjadi isu kritis dalam pengembangan sistem AI yang bertanggung jawab. Studi klasik oleh [5] menunjukkan bahwa algoritma prediksi kesehatan dapat menghasilkan bias sistematis terhadap kelompok tertentu akibat penggunaan proxy variabel yang tidak merepresentasikan kebutuhan klinis secara adil. Penelitian lebih mutakhir menunjukkan bahwa bias dalam model ML medis dapat tetap muncul meskipun atribut sensitif seperti gender atau ras tidak secara eksplisit digunakan dalam pelatihan model. Hal ini disebabkan oleh adanya korelasi laten antara variabel klinis dengan atribut sensitif tersebut [6][7]. Fenomena ini dikenal sebagai fairness through unawareness fallacy, di mana penghapusan atribut sensitif tidak secara otomatis menghilangkan bias dalam model. Dalam literatur fairness, beberapa definisi formal telah dikembangkan, seperti demographic parity, equalized odds, dan equal opportunity, yang masing-masing memiliki implikasi berbeda dalam konteks aplikasi kesehatan. Konsep equal opportunity, yang menekankan kesetaraan recall antar subkelompok, dianggap paling relevan dalam kasus skrining penyakit karena berfokus pada kemampuan model dalam mendeteksi pasien yang benar-benar sakit [9][10]. Dalam konteks klinis, metrik seperti recall (sensitivitas) dan false negative rate (FNR) menjadi sangat penting. Tingginya FNR pada suatu subkelompok menunjukkan tingginya risiko missed diagnosis, yaitu kondisi di mana pasien yang sebenarnya sakit tidak terdeteksi oleh sistem, yang dapat berakibat fatal [11]. Oleh karena itu, evaluasi model berbasis subkelompok menjadi komponen penting dalam validasi sistem AI kesehatan.

2.3 Gender sebagai Dimensi Fairness dalam Prediksi Penyakit

Gender merupakan salah satu dimensi penting dalam analisis fairness pada sistem AI kesehatan. Perbedaan biologis dan fisiologis antara laki-laki dan perempuan dapat memengaruhi profil biomarker, respons terhadap penyakit, serta progresivitas penyakit liver. Studi oleh [8] menunjukkan adanya perbedaan signifikan dalam kadar enzim hati dan respons farmakologis antara pasien laki-laki dan perempuan. Selain faktor biologis, ketidakseimbangan representasi gender dalam dataset medis juga menjadi sumber utama bias dalam model machine learning. Dataset yang didominasi oleh satu kelompok tertentu cenderung menghasilkan model yang lebih optimal untuk kelompok tersebut, sementara performa pada kelompok minoritas menjadi kurang akurat [12][6]. Penelitian terbaru juga menunjukkan bahwa perbedaan distribusi fitur antar gender dapat menyebabkan pergeseran performa model (distribution shift) yang signifikan, bahkan ketika model dilatih tanpa memasukkan variabel gender secara eksplisit [7]. Hal ini menunjukkan bahwa bias dalam model tidak hanya berasal dari atribut sensitif, tetapi juga dari struktur data itu sendiri. Berdasarkan celah penelitian tersebut, studi ini berkontribusi dengan melakukan evaluasi eksplisit terhadap perbedaan performa prediksi penyakit liver antara pasien laki-laki dan perempuan, serta mengkaji efektivitas pendekatan pemodelan berbasis

subkelompok dalam mengurangi ketimpangan tersebut.

3. Metode Penelitian

Penelitian ini mengadopsi kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining), yang merupakan metodologi standar dalam pengembangan model data mining dan machine learning. CRISP-DM terdiri dari enam tahapan utama, yaitu: (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation, dan (6) deployment.

Dalam konteks penelitian ini, tahapan deployment diadaptasi menjadi pengemasan hasil penelitian dalam bentuk analisis komprehensif terhadap performa model dan aspek fairness berbasis subkelompok. Pendekatan ini memungkinkan evaluasi tidak hanya dari sisi performa agregat, tetapi juga dari sisi keadilan model dalam konteks klinis.

3.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah Indian Liver Patient Dataset, yang diperoleh dari UCI Machine Learning Repository [13] Dataset ini terdiri dari 583 rekam medis pasien dari wilayah India Utara, yang mencakup parameter biokimia darah yang relevan untuk diagnosis penyakit liver.

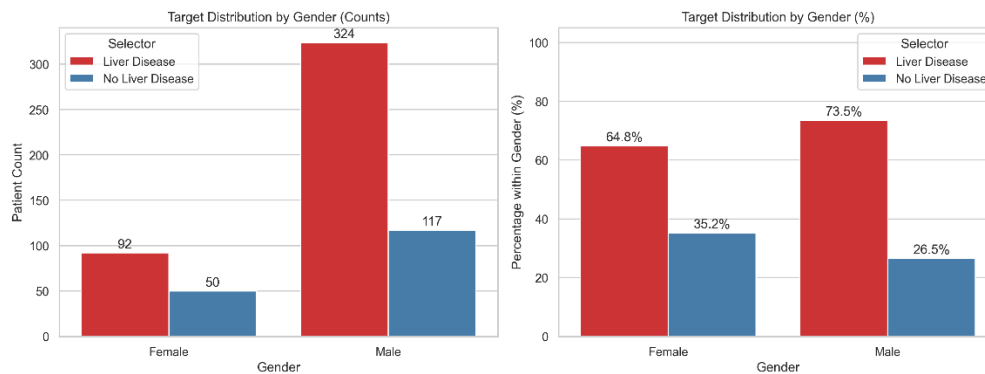
Tabel 1. Ringkasan Dataset ILPD

No.	Karakteristik	Nilai
1	Jumlah sampel	583
2	Jumlah fitur	11 (9 numerik + gender+target)
3	Jumlah kelas target	2 (Penyakit liver /tidak)
4	Sampel penyakit liver	416 (71,36%)
5	Sampel tanpa penyakit liver	167 (28,64%)
6	Pasien laki-laki	441 (75,64%)
7	Pasien perempuan	142 (24,36%)

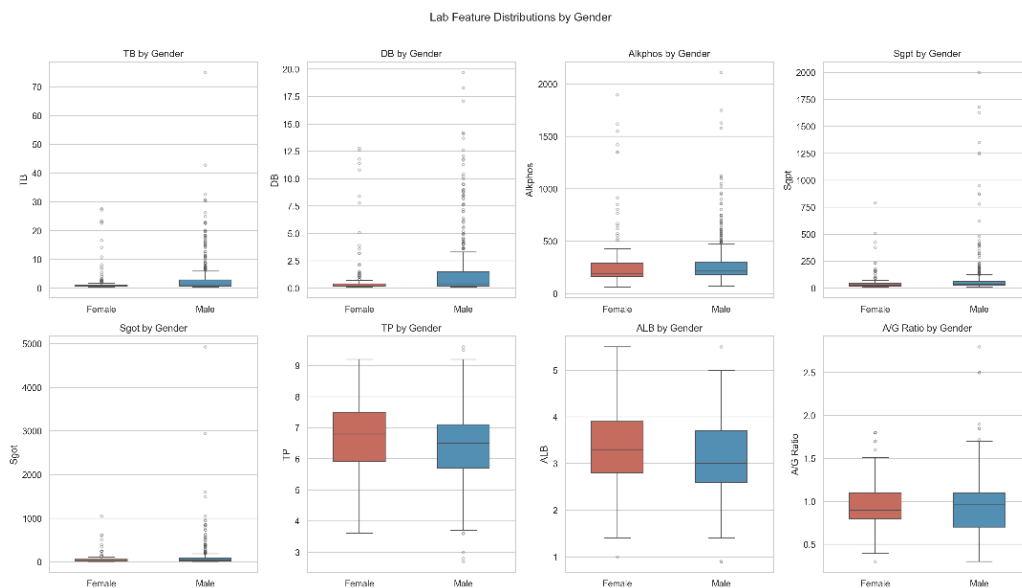
Tabel 2. Fitur-fitur dalam Dataset

No.	Fitur	Keterangan
1	Age	Usia pasien (tahun)
2	Gender	Jenis kelamin (Male/Female)
3	TB (Total Bilirubin)	Kadar total bilirubin (mg/dL)
4	DB (Direct Bilirubin)	Kadar bilirubin langsung (mg/dL)
5	Alkphos (Alkaline Phosphatase)	Kadar alkalin fosfatase (UI/L)
6	Sgpt (alanine Aminotransferase)	Kadar SGPT/ALT (IU/L)
7	Sgot (Aspartate Aminotransferase)	Kadar SGOT/AST (IU/L)
8	TP (Total Proteins)	Kadar total protein (g/dL)
9	ALB (Albumin)	Kadar albumin (g/dL)
10	A/G Ratio	Rasio albumin terhadap globulin
11	Target	Penyakit Liver = 1 / Tidak = 0

Analisis eksplorasi menunjukkan adanya dua lapisan ketidakseimbangan pada dataset: ketidakseimbangan kelas target (71,36% positif) dan ketidakseimbangan gender (75,64% laki-laki). Prevalensi penyakit liver per subkelompok juga berbeda: 73,47% pada pasien laki-laki versus 64,79% pada pasien perempuan, yang mengindikasikan adanya pergeseran distribusi subkelompok (subgroup distribution shift) yang harus dipertimbangkan dalam evaluasi.



Gambar 1. Distribusi target berdasarkan gender.



Gambar 2. Distribusi biomarker antar gender.

3.2 Persiapan Data

3.2.1 Rekayasa Fitur

Tahapan persiapan data meliputi langkah-langkah berikut:

- Pengkodean target biner: Penyakit Liver = 1, Tidak = 0
- Pengkodean gender biner: Laki-laki = 1, Perempuan = 0; kolom Gender asli dipertahankan untuk evaluasi subkelompok
- Rekayasa fitur klinis: Ditambahkan fitur $\text{fractional_bilirubin} = \text{DB} / (\text{TB} + 1 \times 10^{-6})$, yaitu rasio bilirubin langsung terhadap bilirubin total, yang merupakan indikator klinis fungsi ekskresi hati
- Transformasi log1p: Diterapkan pada fitur TB, DB, Alkphos, Sgpt, dan Sgot untuk mengompres distribusi yang sangat miring ke kanan (right-skewed) akibat adanya outlier ekstrem
- Penskalaan dengan RobustScaler: Menggunakan median dan IQR sebagai basis penskalaan, lebih tahan terhadap outlier dibandingkan StandardScaler.

3.2.2 Pemisahan Data dengan Stratifikasi Empat Kelas

Mengingat adanya kelompok yang sangat kecil. Perempuan tanpa penyakit liver berjumlah hanya 50 sampel dari total 583 (8,6%), dilakukan stratifikasi empat kelas berdasarkan kombinasi Gender $\times$ Target. Hal ini memastikan proporsi setiap subkelompok terjaga secara proporsional antara data latih dan data uji.

Tabel 1. Distribusi Data Latih dan Uji

No.	Kelompok	Data Latih	Data Uji	Proporsi Latih	Proporsi Uji
1	Laki-laki + Penyakit Liver	324	81	-	-
2	Laki-laki + Tidak Sakit	49	8	-	-
3	Perempuan + Penyakit Liver	53	18	-	-
4	Perempuan + Tidak Sakit	40	10	-	-
5	Total	466	117	80%	20%

Proporsi "Perempuan tanpa Penyakit Liver" pada data latih adalah 8,58% (40/466) dan pada data uji 8,55% (10/117), membuktikan keberhasilan stratifikasi empat kelas.

3.3 Desain Eksperimen

Tiga skenario eksperimen dirancang untuk menganalisis pengaruh gender pada performa prediksi:

Tabel 4. Ringkasan Desain Eksperimen

Eksperimen	Deskripsi	Fitur Input
Exp1 (Global + Gender)	Satu model untuk semua pasien, gender dimasukkan sebagai fitur	11 fitur (termasuk gender)
Exp2 (Global – Gender)	Satu model untuk semua pasien, gender dikecualikan dari fitur	10 fitur (tanpa gender eksplisit)
Exp3 (Subkelompok Terpisah)	Model terpisah untuk laki-laki dan perempuan	10 fitur per model

Tiga algoritma dilatih pada setiap eksperimen:

- Logistic Regression (dengan pembobotan kelas `class_weight='balanced'`)
- Random Forest (200 pohon, `class_weight='balanced'`)
- XGBoost (200 estimator)

Seluruh model menggunakan `random_state=42` untuk reproduktibilitas.

3.4 Metrik Evaluasi

Evaluasi dilakukan pada dua tingkat:

Metrik Kinerja Global: Akurasi, Presisi, Recall (Sensitivitas), F1-score, AUC-ROC

Metrik Fairness Subkelompok:

- $\Delta Accuracy = Accuracy_{male} - Accuracy_{female}$
- $\Delta Recall = Recall_{laki} - Recall_{perempuan}$
- $FNR_g = TP_g + FN_g$
- $\Delta FNR = FNR_{perempuan} - FNR_{laki} - laki$

Dalam konteks skrining penyakit, FNR yang tinggi pada suatu subkelompok berarti tingkat *missed diagnosis* yang tinggi pada kelompok tersebut, yang merupakan risiko klinis paling kritis. Nilai bootstrap 95% CI dihitung untuk *recall* perempuan (n=28) menggunakan 1.000 iterasi *re-sampling*.

3.5 Interpretasi Model

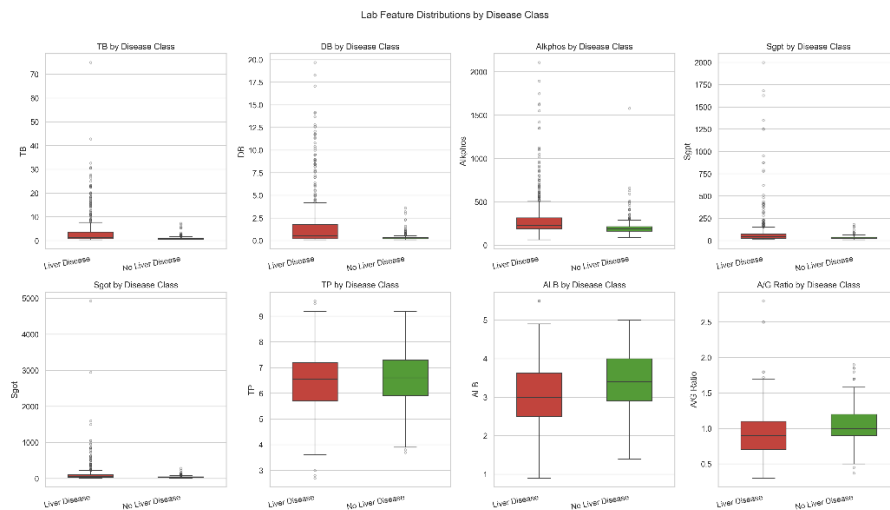
Interpretabilitas model dievaluasi menggunakan dua pendekatan:

- Feature Importance: Berdasarkan *impurity gain* untuk Random Forest dan XGBoost, dan koefisien model untuk Logistic Regression
- SHAP (*SHapley Additive exPlanations*): Nilai SHAP dihitung menggunakan TreeExplainer untuk Random Forest dan XGBoost, serta LinearExplainer untuk Logistic Regression

3.6 Analisis Eksploratif Data (EDA)

Analisis data eksplorasi mengungkap sejumlah karakteristik penting:

Distribusi Biomarker: Fitur-fitur biokimia seperti Sgpt, Sgot, Alkphos, TB, dan DB menunjukkan distribusi sangat miring ke kanan dengan outlier ekstrem. Pasien dengan penyakit liver memiliki median kadar enzim hati yang jauh lebih tinggi dibandingkan pasien sehat (Gambar 3).

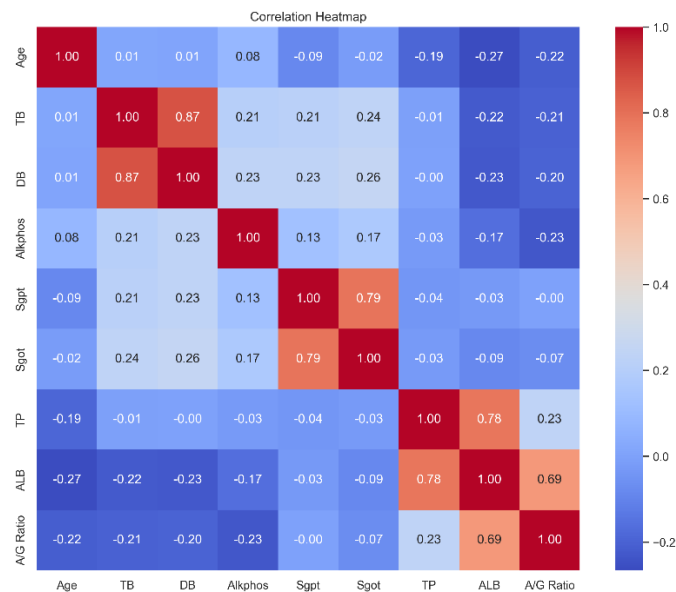


Gambar 3. Boxplot biomarker berdasarkan status penyakit liver.

Perbedaan Profil Biomarker per Gender: Analisis korelasi per subkelompok gender (Gambar 3) menunjukkan bahwa:

1. Alkphos merupakan fitur berbeda secara signifikan antara laki-laki dan perempuan
 2. DB memiliki separabilitas yang lebih lemah pada pasien perempuan dibandingkan laki-laki
- Ini mengindikasikan bahwa fitur-fitur prediktif tidak berperilaku identik across gender

Korelasi Fitur: Matriks korelasi (Gambar 4) menunjukkan korelasi tinggi antara TB dan DB ($r \approx 0,87$), serta antara Sgpt dan Sgot ($r \approx 0,79$), yang mendukung keputusan penambahan fitur rasio fractional_bilirubin.



Gambar 4. Matriks korelasi antar fitur.

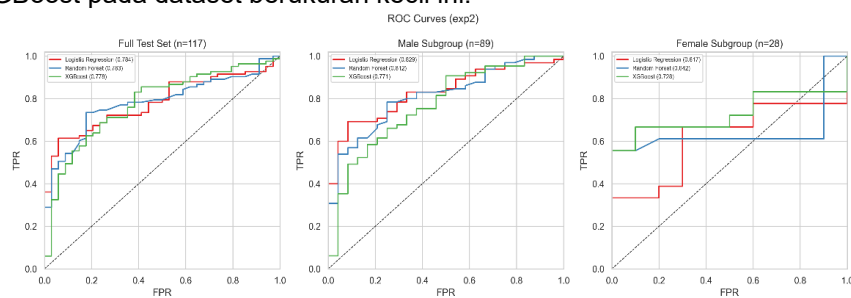
3.7 Performa Model Global

Tabel berikut merangkum performa model pada data uji keseluruhan (n=117):

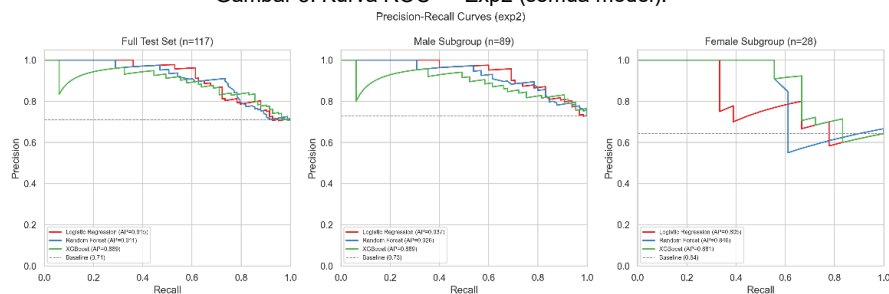
Tabel 5. Performa Model Global — Data Uji Keseluruhan

Eksperimen	Model	Akurasi	Presisi	Recall	F1-score	AUC-ROC
Exp1	Logistic Regression	0,6752	0,9592	0,5663	0,7121	0,7835
Exp1	Random Forest	0,7179	0,7604	0,8795	0,8156	0,7849
Exp1	XGBoost	0,7350	0,8023	0,8313	0,8166	0,7860
Exp2	Logistic Regression	0,6752	0,9592	0,5663	0,7121	0,7838
Exp2	Random Forest	0,7179	0,7604	0,8795	0,8156	0,7826
Exp2	XGBoost	0,7692	0,8256	0,8554	0,8402	0,7782
Exp3 (♂)	Logistic Regression	0,7079	0,9333	0,6462	0,7636	0,8179
Exp3 (♂)	Random Forest	0,7753	0,7848	0,9538	0,8611	0,7917
Exp3 (♂)	XGBoost	0,7303	0,7662	0,9077	0,8310	0,7981
Exp3 (♀)	Logistic Regression	0,5000	0,8333	0,2778	0,4167	0,5611
Exp3 (♀)	Random Forest	0,7500	0,7619	0,8889	0,8205	0,7556
Exp3 (♀)	XGBoost	0,6786	0,7368	0,7778	0,7568	0,7778

Model terbaik secara keseluruhan adalah XGBoost pada Eksperimen 2 (global tanpa fitur gender), dengan F1 = 0,8402, akurasi = 0,7692, dan AUC = 0,7782. Fakta bahwa menghilangkan fitur gender eksplisit justru meningkatkan performa global (F1 dari 0,8166 → 0,8402 untuk XGBoost dari Exp1 ke Exp2) mengindikasikan bahwa fitur gender yang terkode secara biner dapat menambahkan noise ke dalam model XGBoost pada dataset berukuran kecil ini.



Gambar 5. Kurva ROC — Exp2 (semua model).



Gambar 6. Kurva Precision-Recall — Exp2 (semua model).

3.8 Analisis Fairness Berbasis Subkelompok Gender

Tabel berikut merangkum metrik fairness seluruh konfigurasi:

Tabel 6. Metrik Fairness Antar Subkelompok Gender

Eksperimen	Model	Acc♂	Acc♀	ΔAcc	Recall♂	Recall♀	ΔRecall	FNR♀	FNR♂	ΔFNR
Exp1	Logistic Regression	0,7079	0,5714	+0,1365	0,6308	0,3333	+0,2975	0,6667	0,3692	+0,2975
Exp1	Random Forest	0,7753	0,5357	+0,2396	0,9538	0,6111	+0,3427	0,3889	0,0462	+0,3427
Exp1	XGBoost	0,7528	0,6786	+0,0742	0,8615	0,7222	+0,1393	0,2778	0,1385	+0,1393
Exp2	Logistic Regression	0,7079	0,5714	+0,1365	0,6308	0,3333	+0,2975	0,6667	0,3692	+0,2975
Exp2	Random Forest	0,7753	0,5357	+0,2396	0,9538	0,6111	+0,3427	0,3889	0,0462	+0,3427
Exp2	XGBoost	0,7978	0,6786	+0,1192	0,9077	0,6667	+0,2410	0,3333	0,0923	+0,2410
Exp3	Logistic Regression	0,7079	0,5000	+0,2079	0,6462	0,2778	+0,3684	0,7222	0,3538	+0,3684
Exp3	Random Forest	0,7753	0,7500	+0,0253	0,9538	0,8889	+0,0649	0,1111	0,0462	+0,0649
Exp3	XGBoost	0,7303	0,6786	+0,0517	0,9077	0,7778	+0,1299	0,2222	0,0923	+0,1299

Keterangan: ΔRecall dan ΔFNR = nilai laki-laki dikurangi perempuan / FNR perempuan dikurangi FNR laki-laki (positif = laki-laki lebih diuntungkan)

3.9 Analisis Confusion Matrix per Subkelompok

Tabel berikut merinci distribusi TP, TN, FP, dan FN untuk setiap subkelompok gender pada konfigurasi yang relevan:

Tabel 7. Detail Confusion Matrix per Subkelompok Gender (Konfigurasi Terpilih)

Konfigurasi	Kelompok	TP	TN	FP	FN	FPR	FNR
Exp1 / LogReg	Perempuan	6	10	0	12	0,0000	0,6667
Exp1 / LogReg	Laki-laki	41	22	2	24	0,0833	0,3692
Exp1 / XGBoost	Perempuan	13	6	4	5	0,4000	0,2778
Exp1 / XGBoost	Laki-laki	56	11	13	9	0,5417	0,1385
Exp2 / XGBoost	Perempuan	12	7	3	6	0,3000	0,3333
Exp2 / XGBoost	Laki-laki	59	12	12	6	0,5000	0,0923
Exp3 / RF (♀)	Perempuan	16	5	5	2	0,5000	0,1111
Exp3 / RF (♂)	Laki-laki	62	7	17	3	0,7083	0,0462

Data uji perempuan berjumlah 28 sampel (18 penyakit liver, 10 tanpa penyakit liver); data uji laki-laki berjumlah 89 sampel (65 penyakit liver, 24 tanpa penyakit liver).

Pada konfigurasi Logistic Regression Exp1, meskipun FPR perempuan = 0 (tidak ada prediksi salah untuk kelas negatif), FNR perempuan mencapai 0,6667, menunjukkan *over-konservatisme* terhadap prediksi negatif yang mengakibatkan banyak pasien sakit yang terlewatkan.

3.10 Bootstrap Confidence Interval pada Recall Perempuan

Mengingat ukuran sampel perempuan yang kecil di data uji (n=28), nilai recall untuk subkelompok perempuan memiliki ketidakpastian yang lebih besar. Tabel berikut menyajikan interval kepercayaan 95% berbasis bootstrap untuk recall perempuan:

Tabel 8. Bootstrap 95% CI untuk Recall Perempuan

Eksperimen	Model	Recall♀	CI Bawah	CI Atas
Exp1	Logistic Regression	0,3333	0,1176	0,5716
Exp1	Random Forest	0,6111	0,4373	0,8824
Exp1	XGBoost	0,7222	0,5000	0,9333
Exp2	Logistic Regression	0,3333	0,1175	0,5714
Exp2	Random Forest	0,6111	0,3571	0,8421
Exp2	XGBoost	0,6667	0,4444	0,8889
Exp3	Logistic Regression	0,2778	0,0768	0,5000
Exp3	Random Forest	0,8889	0,8235	1,0000
Exp3	XGBoost	0,7778	0,5714	0,9444

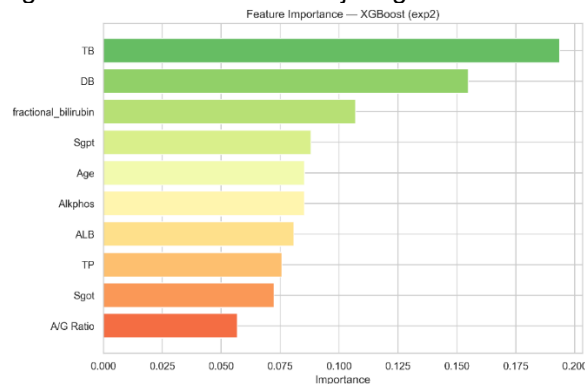
CI yang lebih sempit pada Exp3/Random Forest (0,8235–1,0000) mengindikasikan estimasi recall perempuan yang relatif stabil dan konsisten pada konfigurasi terbaik dari perspektif *fairness*.

3.11 Interpretabilitas: Feature Importance dan SHAP

3.11.1 Feature Importance Global (Exp2 / XGBoost)

Analisis feature importance pada model terbaik global (Exp2 / XGBoost) menunjukkan bahwa fitur-fitur dengan kontribusi tertinggi adalah:

1. DB (Direct Bilirubin), fitur paling penting secara konsisten
2. TB (Total Bilirubin), korelasi tinggi dengan DB
3. fractional_bilirubin, fitur rekayasa yang berhasil merangkum rasio DB/TB
4. Sgpt (ALT), enzim yang khas meningkat pada kerusakan sel hati
5. Sgot (AST), enzim yang relevan untuk kerusakan jaringan hati.

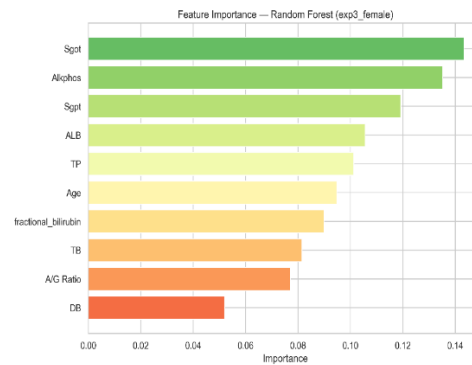


Gambar 7. Feature importance — Exp2 / XGBoost.

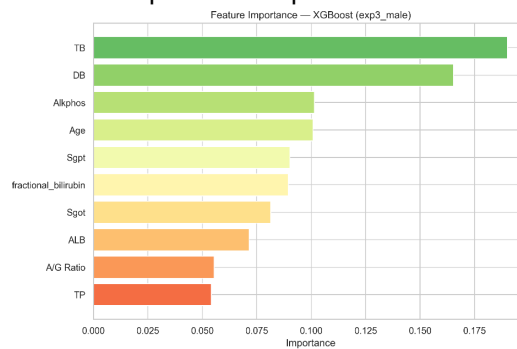
3.11.2 Perbedaan Feature Importance per Gender (Exp3)

Perbandingan feature importance antara model laki-laki dan perempuan pada Exp3 menunjukkan perbedaan yang signifikan:

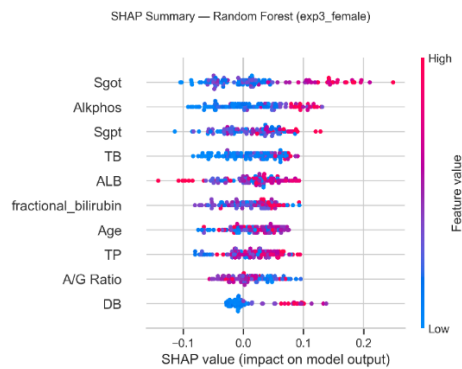
- a) Model Khusus Laki-laki (Exp3 / XGBoost):
 - DB dan Sgot mendominasi
 - TB dan Sgpt berada di posisi menengah
 - Usia (Age) memiliki kontribusi yang lebih kecil
- b) Model Khusus Perempuan (Exp3/Random Forest):
 - Alkphos muncul sebagai fitur yang lebih dominan dibandingkan pada model laki-laki
 - DB tetap relevan namun dengan bobot berbeda
 - fractional_bilirubin memiliki kepentingan yang lebih tinggi relatif terhadap model laki-laki



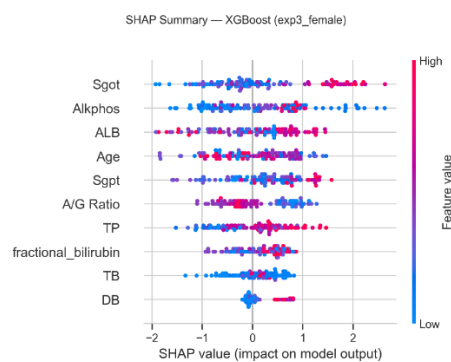
Gambar 8 . Feature importance Exp3 / Random Forest (model khusus perempuan).



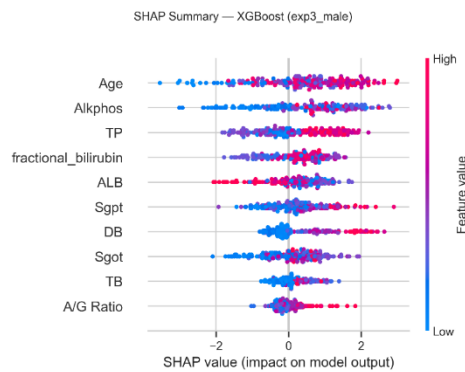
Gambar 9. Feature importance — Exp3 / XGBoost (model khusus laki-laki).



Gambar 10. SHAP summary plot — Exp3 / Random Forest (perempuan).



Gambar 11. SHAP summary plot — Exp3 / XGBoost (perempuan).



Gambar 12. SHAP summary plot — Exp3 / XGBoost (laki-laki).

Perbedaan urutan kepentingan fitur antar gender ini memperkuat argumen bahwa biomarker hati berperilaku berbeda pada populasi laki-laki dan perempuan, dan mendukung alasan mengapa model terpisah per gender memberikan hasil yang lebih adil untuk subkelompok perempuan.

4. Kesimpulan

Penelitian ini menginvestigasi apakah model ML untuk prediksi penyakit liver menunjukkan perbedaan kinerja yang signifikan antara subkelompok laki-laki dan perempuan. Berdasarkan eksperimen terhadap Indian Liver Patient Dataset dengan tiga algoritma dan tiga skenario desain eksperimen, empat simpulan utama dapat diambil: Model terbaik secara agregat adalah XGBoost pada Eksperimen 2 (tanpa fitur gender eksplisit), dengan $F1 = 0,8402$, akurasi = $0,7692$, dan $AUC = 0,7782$. Disparitas kinerja berbasis gender bersifat konsisten dan sistematis. Seluruh sembilan konfigurasi model menunjukkan recall laki-laki yang lebih tinggi dari perempuan tanpa pengecualian. Disparitas ini bervariasi dari $\Delta\text{Recall} = +0,0649$ hingga $+0,3684$ tergantung konfigurasi. Menghilangkan fitur gender sebagai input tidak mengeliminasi disparitas. Ketimpangan kinerja melekat pada distribusi fitur klinis, bukan hanya pada representasi eksplisit gender dalam model. Model terpisah per gender dengan Random Forest memberikan kesetaraan terbaik, dengan FNR perempuan = $0,1111$ dan selisih recall hanya $+0,0649$. Konfigurasi ini direkomendasikan untuk skenario klinis yang mengutamakan minimisasi missed diagnosis pada pasien perempuan. Temuan-temuan ini menegaskan bahwa evaluasi subkelompok harus dianggap wajib dalam validasi model AI untuk aplikasi medis, dan metrik agregat saja tidak mencukupi untuk penilaian yang bertanggung jawab.

5. Referensi

- [1] A. Kumar *et al.*, "Advances in Water Resources Multi-model evaluation of catchment- and global-scale hydrological model simulations of drought characteristics across eight large river catchments," *Adv. Water Resour.*, vol. 165, no. October 2021, p. 104212, 2022, doi: 10.1016/j.advwatres.2022.104212.
- [2] S. Singh, S. Rangarajan, D. Thattai, and S. Vaithilingam, "A STANDARD FOR ASSESSMENT OF INDIAN MANGROVES RESPONSE TO SEA-LEVEL RISE USING RSET-MH TECHNIQUE A STANDARD FOR ASSESSMENT OF INDIAN MANGROVES RESPONSE TO SEA-LEVEL RISE USING RSET-MH TECHNIQUE," vol. 35, no. 4, pp. 429–439, 2023.
- [3] T. Patel and M. Shah, "Study of Economic Feasibility for a Decentralized Small off Grid Geothermal Power Plant Using Slim Boreholes," pp. 2–7, 2021.
- [4] M. Mosallam, "International Journal of Data and Network Science Artificial intelligence and entrepreneurship education : A paradigm in Qatari higher education institutions after covid-19 pandemic," vol. 7, pp. 695–706, 2023, doi: 10.5267/j.ijdns.2023.3.002.
- [5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an

- algorithm used to manage the health of populations,” vol. 453, no. October, pp. 447–453, 2019.
- [6] L. Seyyed-kalantari, H. Zhang, M. B. A. Mcdermott, I. Y. Chen, and M. Ghassemi, “algorithms applied to chest radiographs in under-served patient populations,” vol. 27, no. December, 2021, doi: 10.1038/s41591-021-01595-0.
 - [7] P. Chen, W. Shi, Y. Liu, and X. Cao, “Slip rate deficit partitioned by fault-fold system on the active Haiyuan fault zone , Northeastern Tibetan Plateau,” *J. Struct. Geol.*, vol. 155, no. January, p. 104516, 2022, doi: 10.1016/j.jsg.2022.104516.
 - [8] L. Albarqouni, H. Greenwood, C. Dowsett, T. Hoffmann, R. Thomas, and P. Glasziou, “Attitudes , beliefs , behaviours and perspectives on barriers and enablers of Australian general practitioners towards drug interventions : a national survey,” pp. 1–8, 2024, doi: 10.1136/fmch-2023-002457.
 - [9] M. Hardt and L. G. El Ghadi, “Equality of Opportunity in Supervised Learning,” pp. 1–22, 2020.
 - [10] S. R. Pfohl, A. Foryciarz, and N. H. Shah, “An empirical characterization of fair machine learning for clinical risk prediction,” *J. Biomed. Inform.*, vol. 113, no. October 2020, p. 103621, 2021, doi: 10.1016/j.jbi.2020.103621.
 - [11] R. Article, “Machine Learning in Medicine,” pp. 1347–1358, 2020, doi: 10.1056/NEJMra1814259.
 - [12] V. Peterson, D. H. Milone, and E. Ferrante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer- aided diagnosis,” pp. 1–3, 2020, doi: 10.1073/pnas.1919012117.
 - [13] N. Ramana, B. & Venkateswarlu, “UCI Machine Learning Repository. ILPD (Indian Liver Patient Dataset).,” 2022, 2022. <https://doi.org/10.24432/C5D02C>.