UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 4 Nomor 2 Tahun 2026 Halaman 167-177.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Hybrid LBFA-Based Feature Selection for Improving Machine Learning Classification Performance in Heart Disease Prediction

Hana Siti Azizah, Eni Sumarminingsih*, dan Adji Achmad Rinaldo Fernandes
Departemen Statistika, Universitas Brawijaya, Malang, Indonesia
*Corresponding author: eni_stat@ub.ac.id

Submitted : 15 Maret 2026
Revised : 11 Mei 2026
Accepted : 18 Mei 2026

ABSTRACT

Feature selection and feature engineering are essential steps in developing accurate machine learning models, particularly when dealing with imbalanced datasets and redundant variables. However, many feature augmentation methods are often applied without a consistent preprocessing strategy, which can reduce model reliability and increase the risk of information leakage. To overcome this issue, this study proposes a hybrid classification framework that combines CatBoost-based feature selection with two feature augmentation techniques: LOGIT transformation and Log Density Ratio (LDR). A structured preprocessing pipeline was designed to ensure consistency throughout the modeling process. One-hot encoding was applied for the LOGIT transformation, while numerical standardization was used for LDR estimation. The generated features were then integrated with the selected original variables to produce richer feature representations for classification. The proposed framework was evaluated using the Heart Disease dataset with three gradient boosting algorithms, namely LightGBM, XGBoost, and CatBoost. Model performance was assessed using accuracy, precision, sensitivity, specificity, and F1-score. The results show that the proposed approach consistently improved classification performance across all models. Among the tested models, LightGBM combined with LOGIT and LDR achieved the best performance, obtaining an accuracy of 0.9618, precision of 0.9485, sensitivity of 0.9620, specificity of 0.9625, and F1-score of 0.9552. These findings suggest that combining feature selection with structured feature augmentation can significantly improve predictive performance in imbalanced classification tasks.

Keywords: *Feature Augmentation, Heart Disease Prediction, LOGIT Transformation, Log Density Ratio, LightGBM, XGBoost*

I. PENDAHULUAN

Ketidakseimbangan kelas (*class imbalance*) merupakan permasalahan yang sering muncul dalam berbagai aplikasi klasifikasi, khususnya pada domain kesehatan dan analisis risiko, di mana jumlah observasi pada kelas minoritas jauh lebih sedikit dibandingkan kelas mayoritas (El-sofany dkk., 2024). Kondisi ini menyebabkan model pembelajaran mesin cenderung bias terhadap kelas mayoritas sehingga kemampuan dalam mendeteksi kejadian penting yang jarang terjadi menjadi menurun. Dalam konteks klasifikasi biner, fenomena ini sering ditandai dengan nilai akurasi yang tinggi namun sensitivitas (*recall*) terhadap kelas minoritas yang rendah. Oleh karena itu, evaluasi kinerja model pada data tidak seimbang perlu mempertimbangkan metrik yang lebih representatif seperti *precision*, *recall*, dan *F1-score*.

Berbagai pendekatan telah dikembangkan untuk menangani permasalahan ketidakseimbangan data (*class imbalance*) dalam klasifikasi. Pendekatan yang umum digunakan meliputi manipulasi distribusi data melalui teknik *oversampling* dan *undersampling*, penerapan metode pembelajaran yang sensitif terhadap biaya kesalahan (*cost-sensitive learning*), serta penggunaan algoritma *ensemble* learning yang lebih adaptif terhadap distribusi kelas yang tidak seimbang. Salah satu teknik yang banyak digunakan adalah *Adaptive Synthetic* (ADASYN), yang mampu menghasilkan data sintetis pada kelas minoritas untuk meningkatkan kemampuan model dalam mendeteksi kelas tersebut (Susrifalah dkk., 2025). Selain itu, tahap prapemrosesan seperti *feature selection* dan reduksi dimensi juga berperan penting dalam meningkatkan kualitas informasi yang digunakan oleh model. Meskipun berbagai metode tersebut terbukti mampu meningkatkan performa klasifikasi, sebagian besar penelitian masih berfokus pada modifikasi

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

distribusi data atau optimasi algoritma pembelajaran, sementara eksplorasi terhadap strategi pembentukan ruang fitur yang lebih informatif masih relatif terbatas.

Salah satu pendekatan yang mulai mendapat perhatian adalah *feature augmentation*, yaitu proses memperkaya representasi data dengan membangun fitur baru yang berasal dari transformasi fitur asli. Pendekatan ini bertujuan untuk mengekstraksi pola tambahan yang tidak secara langsung terlihat pada ruang fitur awal sehingga model dapat menangkap hubungan yang lebih kompleks antara variabel prediktor dan variabel target. Dalam konteks klasifikasi biner, transformasi seperti logit dan *complementary log-log* digunakan untuk memodelkan probabilitas kejadian berdasarkan karakteristik data (I. Mawarni dkk., 2025), yang kemudian dapat dimanfaatkan sebagai representasi fitur tambahan. Selain itu, transformasi berbasis distribusi seperti *log-density ratio* dapat digunakan untuk merepresentasikan perbedaan kepadatan antar kelas. Integrasi berbagai transformasi tersebut memungkinkan pembentukan ruang fitur yang menggabungkan informasi linier dan nonlinier, sehingga berpotensi meningkatkan kemampuan model dalam mendeteksi pola pada data tidak seimbang. Namun demikian, proses augmentasi fitur yang dilakukan secara langsung pada seluruh variabel dapat menimbulkan redundansi informasi dan meningkatkan kompleksitas model. Oleh karena itu, tahap *feature selection* menjadi langkah penting sebelum proses augmentasi dilakukan agar ruang fitur yang dihasilkan lebih ringkas, relevan, dan tidak mengandung informasi yang bersifat redundan.

Prediksi penyakit jantung berbasis data klinis merupakan komponen penting dalam pengembangan sistem pendukung keputusan medis. Sejumlah penelitian menunjukkan bahwa performa klasifikasi tidak hanya dipengaruhi oleh pemilihan algoritma, tetapi juga oleh kualitas representasi data sebelum proses pembelajaran model. Studi terbaru melaporkan bahwa algoritma seperti *Extreme Gradient Boosting, Random Forest, ensemble learning,* dan *neural network* cenderung menghasilkan performa yang lebih baik dibandingkan beberapa metode klasik, sementara teknik reduksi dimensi dan *feature selection* juga terbukti mampu meningkatkan kemampuan deteksi. Sebagai contoh, studi komparatif yang mengevaluasi berbagai model termasuk *Logistic Regression, Random Forest, Support Vector Machine* (SVM), XGBoost, dan *deep neural network* menunjukkan bahwa model hibrida XGBoost–SVM dapat mencapai akurasi 89,3% dengan *precision* sebesar 0,90, *sensitivity* sebesar 0,91, dan F1-*score* sebesar 0,905 (Almutairi & Dardouri, 2025).

Berbagai penelitian juga menegaskan bahwa kualitas ruang fitur merupakan faktor penting dalam keberhasilan model prediksi penyakit jantung. Noroozi dkk. (2023) menunjukkan bahwa penerapan berbagai metode *feature selection* dapat meningkatkan performa klasifikasi penyakit jantung, meskipun tingkat pengaruhnya berbeda pada setiap algoritma yang digunakan. Penelitian lain juga melaporkan bahwa kombinasi *feature selection* dengan model berbasis *boosting* seperti XGBoost mampu menghasilkan performa prediksi yang tinggi, sementara *Random Forest* tetap mempertahankan akurasi yang baik meskipun menggunakan jumlah fitur yang lebih terbatas (Aprianto & Anasanti, 2025; El-Sofany dkk., 2024). Temuan-temuan tersebut menunjukkan bahwa kualitas representasi fitur memiliki peran penting dalam meningkatkan kemampuan model klasifikasi. Namun demikian, sebagian besar penelitian sebelumnya masih berfokus pada pemilihan algoritma klasifikasi terbaik, optimasi model, atau penerapan teknik konvensional seperti *stacking*, *oversampling*, dan *feature selection*. Pendekatan yang secara khusus mengeksplorasi transformasi fitur dan pembentukan ruang fitur melalui strategi feature augmentation masih belum banyak dibahas.

Beberapa penelitian menunjukkan bahwa augmentasi dan transformasi ruang fitur dapat menghasilkan representasi data yang lebih informatif, meningkatkan kemampuan separasi antar kelas, serta memperbaiki generalisasi model klasifikasi (Li & Cui, 2023; Wang dkk., 2020). Selain itu, pembentukan ruang fitur baru melalui proses augmentasi juga berpotensi membantu model dalam menangkap pola kompleks yang tidak sepenuhnya direpresentasikan oleh fitur asli. Dengan demikian, diperlukan suatu pendekatan yang tidak hanya menekankan pada pemilihan *classifier*, tetapi juga memperhatikan konsistensi *preprocessing* dan pembentukan ruang fitur sebelum proses pembelajaran dilakukan. Berdasarkan kondisi tersebut, penelitian ini mengusulkan pendekatan yang mengintegrasikan *feature selection* berbasis CatBoost dengan augmentasi fitur menggunakan transformasi LOGIT dan *Log Density Ratio* (LDR) untuk membangun representasi fitur yang lebih informatif pada klasifikasi penyakit jantung. Berdasarkan kondisi tersebut, penelitian ini mengusulkan penataan pipeline analisis yang lebih konsisten dalam pembentukan ruang fitur. Prediktor asli terlebih dahulu diseleksi menggunakan *CatBoost-based feature selection* untuk memperoleh sepuluh fitur paling informatif, kemudian dilakukan augmentasi melalui dua pendekatan, yaitu transformasi logit dan *log-density ratio*. Dengan pendekatan ini, dibangun skenario fitur asli dengan kombinasi logit dan *log-density ratio* kemudian dievaluasi menggunakan tiga algoritma utama, yaitu XGBoost, LightGBM, dan CatBoost. Pendekatan ini diharapkan tidak hanya meningkatkan performa klasifikasi, tetapi juga memberikan *pipeline preprocessing* yang lebih jelas dan konsisten dalam memisahkan prediktor asli dari fitur hasil augmentasi.

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

II. METODE PENELITIAN

A. Jenis Penelitian dan Sumber Data

Penelitian ini merupakan penelitian terapan yang bertujuan menguji pengaruh pendekatan *Hybrid Logit-Based Feature Augmentation* (LBFA) dan *feature selection* terhadap performa klasifikasi pada kasus prediksi penyakit jantung. Pendekatan terapan dipilih karena penelitian ini berfokus pada implementasi metode, pengukuran peningkatan kinerja model, serta perbandingan performa beberapa algoritma *machine learning* secara eksperimental pada dataset riil. Data yang digunakan merupakan data sekunder berupa dataset *Heart Disease* yang diperoleh dari Kaggle, yang dapat diakses melalui https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

Dataset ini dipilih karena banyak digunakan sebagai *benchmark* dataset dalam penelitian klasifikasi medis serta relevan untuk mengevaluasi pengaruh rekayasa fitur (*feature engineering*) dan seleksi fitur terhadap performa model klasifikasi. Dataset ini terdiri dari 246.022 observasi dengan 35 variabel prediktor serta satu variabel target biner yang merepresentasikan status penyakit jantung yang dinotasikan sebagai berikut:

$$y_i = \begin{cases} 1, menyatakan\ pasien\ dengan\ penyakit\ jantung \\ 0, menyatakan\ pasien\ tanpa\ penyakit\ jantung \end{cases}$$

B. Teknik Analisis Data

Analisis data dalam penelitian ini menerapkan kerangka *Hybrid Logit-Based Feature Augmentation* (LBFA) dengan seleksi fitur berbasis CatBoost yang telah ditata ulang, kemudian mengevaluasikannya menggunakan tiga algoritma klasifikasi utama, yaitu XGBoost, LightGBM, dan CatBoost. Secara umum, tahapan penelitian dapat dijelaskan sebagai berikut:

1. Eksplorasi Data (*Data Exploration*)
   Tahap eksplorasi data dilakukan untuk memahami karakteristik dataset sebelum proses pemodelan. Pemeriksaan dilakukan terhadap struktur dataset dan tipe variabel, distribusi variabel target, *missing values*, duplikasi observasi. Tahap eksplorasi bertujuan (Han, J., 2012; He & Garcia, 2009).
2. Pra-pemrosesan Data (*Data Preprocessing*)
   Bagi variabel kategorikal diubah menjadi representasi numerik menggunakan teknik *encoding* menggunakan *one-hot encoding.* Sehingga jika suatu variabel memiliki k kategori, maka variabel tersebut direpresentasikan menjadi $k-1$ variabel *dummy* (Kotsiantis dkk., 2006). Selanjutnya, dilakukan proses standardisasi menggunakan transformasi pada persamaan (1).

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (1)$$

   dengan x_{ij} observasi ke-i pada fitur numerik ke-j, μ_j merupakan rata-rata fitur ke-j, serta σ_j simpangan baku fitur ke-j .

3. *Feature Selection* menggunakan CatBoost-FS
   Proses *feature selection* dilakukan menggunakan CatBoost *Feature Selection* (CatBoost-FS) untuk mengidentifikasi fitur yang paling informatif pada ruang fitur hasil augmentasi (Fan dkk., 2024). Prokhorenkova dkk. (2018) menyatakan bahwa CatBoost merupakan algoritma *gradient boosting* berbasis pohon keputusan yang membangun model secara iteratif menggunakan persamaan (2)

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \quad (2)$$

   dengan $f_t(x_i)$ merupakan fungsi pohon keputusan pada iterasi ke-t.
   Tingkat kepentingan setiap fitur diukur melalui nilai *feature importance* yang dihitung berdasarkan perubahan fungsi kerugian akibat penggunaan fitur tersebut pada model dalam persamaan (3)

$$FI_j = \sum_{t=1}^{T} \Delta L_{j,t} \quad (3)$$

   Dalam penelitian ini, model CatBoost dilatih pada data latih untuk menghitung nilai *feature importance* seluruh variabel pada ruang fitur hibrida. Fitur kemudian diurutkan berdasarkan nilai kepentingannya, dan subset fitur dengan kontribusi terbesar dipilih untuk digunakan pada tahap pemodelan klasifikasi.
4. Pemisahan Data Latih dan Data Uji

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Dataset kemudian dibagi menjadi dua subset yaitu data latih dan data uji menggunakan rasio 80:20 dengan metode *stratified sampling* untuk mempertahankan proporsi kelas. Pembagian ini memastikan bahwa evaluasi model dilakukan pada data yang tidak terlibat dalam proses pelatihan.

5. Pembentukan Fitur Logit

Pada tahap ini, pemodelan regresi logistik dilakukan untuk memperoleh probabilitas posterior kejadian penyakit jantung. Regresi logistik digunakan karena mampu memodelkan hubungan antara variabel prediktor dan probabilitas kejadian biner melalui fungsi logistik (Hosmer & Lemeshow, 2000). Probabilitas tersebut dinyatakan sebagai $p_i = P(Y_i = 1 \mid X_i)$ yang dimodelkan menggunakan fungsi logistik sebagaimana pada persamaan (4).

$$\pi(x_i) = P(Y_i = 1|X_i) = \frac{e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi})}}{1+e^{(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi})}} \quad (4)$$

dengan $\pi(x_i)$ menyatakan peluang terjadinya peristiwa "*heart attack*", yaitu $P(Y_i = 1 \mid X_i)$, pada observasi ke-i. Variabel Y_i merupakan variabel respon biner pada observasi ke-i dengan $i = 1,2,\dots,n$. Parameter β_0 menyatakan intersep atau konstanta model, sedangkan β_p merupakan koefisien regresi untuk prediktor ke-j dengan $j = 1,2,\dots,p$. Selanjutnya, X_{pi} menyatakan nilai variabel prediktor ke-j pada observasi ke-i.

Dengan melakukan transformasi logit, hubungan antara probabilitas dan kombinasi linier prediktor dapat dituliskan persamaan (5).

$$\hat{\eta}_i = logit\,(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \quad (5)$$

Secara matriks, model regresi logistik dapat dinyatakan sebagai

$$\eta = \begin{bmatrix} ln\left(\frac{\pi_1}{1-\pi_1}\right) \\ ln\left(\frac{\pi_2}{1-\pi_2}\right) \\ \vdots \\ ln\left(\frac{\pi_i}{1-\pi_i}\right) \end{bmatrix}_{(n\times 1)} = \boldsymbol{X\beta} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix}_{(n\times(p+1))} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(1\times(p+1))}$$

Nilai $\eta_i$kemudian digunakan sebagai fitur augmentasi logit yang ditambahkan ke dalam ruang fitur sebelum proses *feature selection* dan pemodelan klasifikasi dilakukan. Pendekatan ini bertujuan membentuk representasi fitur yang lebih informatif melalui transformasi logit terhadap probabilitas posterior hasil regresi logistik ((Hosmer & Lemeshow, 2000; Wang dkk., 2020)).

6. Pembentukan Fitur Log-Density Ratio (LDR)

Fitur LDR dibentuk untuk menangkap perbedaan distribusi fitur antar kelas berdasarkan rasio kepadatan peluang bersyarat. Misalkan $f_1(x)$ dan $f_0(x)$ masing-masing menyatakan fungsi kepadatan peluang dari kelas positif ($Y = 1$) dan kelas negatif ($Y = 0$). Rasio kepadatan didefinisikan sebagai persamaan (6) (Sugiyama, 2012).

$$r(x) = \frac{f_1(x)}{f_0(x)} \quad (6)$$

Selanjutnya, untuk memperoleh representasi yang lebih stabil, rasio tersebut ditransformasikan dalam bentuk logaritmik sehingga diperoleh *Log-Density Ratio* (LDR) pada persamaan (7)

$$LDR(x) = \log\left(\frac{f_1(x)}{f_0(x)}\right) \quad (7)$$

Dalam penelitian ini, fungsi kepadatan $f_1(x)$ dan $f_0(x)$ diestimasi menggunakan *Kernel Density Estimation* (KDE), pada persamaan (8)

$$\hat{f}_c(x) = \frac{1}{n_c h}\sum_{i=1}^{n_c} K\left(\frac{x-x_i}{h}\right), c \in \{0,1\}, \quad (8)$$

dengan $n_c$menyatakan jumlah observasi pada kelas c, $K(\cdot)$ fungsi kernel, dan h merupakan parameter *bandwidth*. Nilai LDR untuk observasi ke-i kemudian diperoleh sebagai

$$Z_i = LDR_i = \log\left(\frac{\hat{f}_1(x_i)}{\hat{f}_0(x_i)}\right), \quad (9)$$

dalam bentuk matriks,

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$\boldsymbol{Z} = \begin{bmatrix} \log\left(\frac{\hat{f}_1^{(1)}(x_{11}^{(c)}) + \varepsilon}{\hat{f}_1^{(0)}(x_{i11}^{(c)}) + \varepsilon}\right) & \dots & \log\left(\frac{\hat{f}_q^{(1)}(x_{1q}^{(c)}) + \varepsilon}{\hat{f}_1^{(0)}(x_{i11}^{(c)}) + \varepsilon}\right) \\ \log\left(\frac{\hat{f}_1^{(1)}(x_{21}^{(c)}) + \varepsilon}{\hat{f}_1^{(0)}(x_{21}^{(c)}) + \varepsilon}\right) & \dots & \log\left(\frac{\hat{f}_1^{(1)}(x_{21}^{(c)}) + \varepsilon}{\hat{f}_1^{(0)}(x_{21}^{(c)}) + \varepsilon}\right) \\ \vdots & & \vdots \\ \log\left(\frac{\hat{f}_1^{(1)}(x_{n1}^{(c)}) + \varepsilon}{\hat{f}_1^{(0)}(x_{n1}^{(c)}) + \varepsilon}\right) & \ddots \dots & \log\left(\frac{\hat{f}_q^{(1)}(x_{nq}^{(c)}) + \varepsilon}{\hat{f}_q^{(0)}(x_{nq}^{(c)}) + \varepsilon}\right) \end{bmatrix}_{n\times q}$$

yang selanjutnya digunakan sebagai fitur augmentasi tambahan dan digabungkan dengan fitur asli serta fitur logit pada ruang fitur hibrida sebelum tahap *feature selection* dan pemodelan klasifikasi.

7. Pembentukan Fitur Augmentasi LBFA
   Pendekatan *Logit-Based Feature Augmentation* (LBFA) digunakan untuk membentuk fitur tambahan dari probabilitas regresi logistik. Berdasarkan teorema Bayes, *logit* merepresentasikan *log-odds* probabilitas posterior kelas, sedangkan *Log-Density Ratio* (LDR) merepresentasikan log rasio kepadatan *likelihood* antar kelas.
   Fitur augmentasi dibentuk melalui dua komponen utama, yaitu skor logit $(\hat{\eta})$ dari model regresi logistik dan matriks LDR (Z) yang diperoleh dari estimasi rasio kepadatan antar kelas. Kedua komponen tersebut kemudian digabungkan dengan matriks fitur asli X untuk membentuk ruang fitur *Hybrid* LBFA sebagai berikut:
   $$\boldsymbol{X}^{(hybri\ )} = [\boldsymbol{X} \mid \boldsymbol{\eta} \mid \boldsymbol{Z}] \in \mathbb{R}^{n\times(p+1+q)}$$
   Ruang fitur ini mengintegrasikan fitur asli dengan informasi probabilistik dan distribusional antar kelas yang dihasilkan dari logit dan LDR.
8. Pemodelan Klasifikasi: *Light Gradient Boosting Machine* (LightGBM)
   LightGBM sebagai algoritma utama merupakan algoritma *gradient boosting decision tree* yang membangun model secara iteratif melalui penambahan pohon keputusan untuk meminimalkan fungsi kerugian (Ke dkk., 2017). Pada setiap iterasi, model baru dibangun untuk mempelajari residual dari prediksi model sebelumnya.
   Prediksi model pada observasi ke-i dinyatakan sebagai persamaan (10)
   $$\hat{y}_i^{LightGBM} = \sum_{t=1}^{T} f_t\,(x_i) \tag{10}$$
   dengan $f_t(x_i)$ merupakan fungsi prediksi dari pohon keputusan pada iterasi ke-t, dan T menyatakan jumlah total pohon.
   Fungsi objektif yang diminimalkan pada LightGBM terdiri dari fungsi kerugian dan komponen regularisasi, yaitu
   $$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{11}$$
   dengan
   $$\Omega(f_t) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2 \tag{12}$$
   di mana $l(\cdot)$ adalah fungsi kerugian, J jumlah daun pada pohon, w_j bobot daun ke-j, serta γ dan λ merupakan parameter regularisasi.
   Pada setiap iterasi, LightGBM memperbarui model menggunakan informasi Gradien dan Hessian dari fungsi kerugian, yang dinyatakan sebagai persamaan (13).
   $$g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \qquad h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{13}$$
   Berbeda dengan metode boosting konvensional, LightGBM menggunakan strategi *leaf-wise tree growth* yang memilih *node* dengan penurunan kerugian terbesar untuk diperluas sehingga proses pembelajaran menjadi lebih efisien dan akurat pada *dataset* berukuran besar. Secara umum, proses pemodelan dimulai dengan menginisialisasi prediksi awal model, kemudian menghitung nilai gradien dan Hessian dari fungsi kerugian.
9. Pemodelan Klasifikasi: *Extreme Gradient Boosting* (XGBoost)

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

XGBoost merupakan algoritma *gradient boosting* yang membangun model secara iteratif melalui penambahan pohon keputusan untuk meminimalkan fungsi kerugian (Chen & Guestrin, 2016). Prediksi model dinyatakan sebagai

$$\hat{y}_i^{XGBoost} = \sum_{t=1}^{T} f_t(x_i) \tag{14}$$

dengan $f_t(x_i)$ merupakan fungsi prediksi dari pohon keputusan pada iterasi ke-t, dan T menyatakan jumlah total pohon.

Seperti pada LightGBM, proses optimisasi dilakukan menggunakan pendekatan Gradien dan Hessian dari fungsi kerugian sebagaimana pada persamaan (11)–(13). Namun, berbeda dengan LightGBM yang menggunakan strategi *leaf-wise tree growth*, XGBoost membangun pohon keputusan menggunakan pendekatan *level-wise tree growth*, di mana semua *node* pada level yang sama dikembangkan secara bersamaan. Pendekatan ini menghasilkan struktur pohon yang lebih seimbang dan stabil, meskipun umumnya membutuhkan waktu komputasi yang lebih besar dibandingkan LightGBM.

10. Pemodelan Klasifikasi: *Categorical Boosting* (CatBoost)

CatBoost merupakan algoritma *gradient boosting decision tree* yang dirancang untuk meningkatkan stabilitas pembelajaran serta mengurangi *prediction shift* pada proses *boosting* (Prokhorenkova dkk., 2018). Prediksi model diperbarui secara iteratif melalui penambahan pohon keputusan yang memodelkan kesalahan prediksi sebelumnya.

Proses pembaruan prediksi pada iterasi ke-t dapat dinyatakan sebagai

$$F^{(t)}(x_i) = F^{(t-1)}(x_i) + \alpha_t h_t(x_i) \tag{15}$$

dengan $F^{(t)}(x_i)$ merupakan fungsi prediksi model pada iterasi ke-t, $h_t(x_i)$ fungsi pohon keputusan yang dibangun pada iterasi tersebut, dan $\alpha_t$adalah parameter *learning rate* yang mengontrol kontribusi pohon baru.

Berbeda dengan pendekatan *boosting* konvensional, CatBoost menggunakan strategi *ordered boosting*, di mana estimasi gradien dihitung berdasarkan permutasi data sehingga setiap observasi hanya menggunakan informasi dari observasi sebelumnya dalam urutan tersebut. Pendekatan ini bertujuan mengurangi bias pada proses pembelajaran serta meningkatkan stabilitas prediksi model.

11. Evaluasi Kinerja Model

Evaluasi dilakukan pada data uji menggunakan beberapa metrik klasifikasi. Menurut Powers (2011), metrik evaluasi tersebut dihitung berdasarkan persamaan berikut.

- Akurasi (*Accuracy*)

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{16}$$

- Presisi (*Precision)*

$$Precision = \frac{TP}{(TP + FP)} \tag{17}$$

- Sensitivitas ($Sensitivity$)

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{18}$$

- Spesifisitas ($Specificity$)

$$Specificity = \frac{TN}{(TN + FP)} \tag{19}$$

- F1-*Score*

$$F1 - Score = 2 \times \frac{(Precision \times Sensitivity)}{(Precision + Sensitivity)} \tag{20}$$

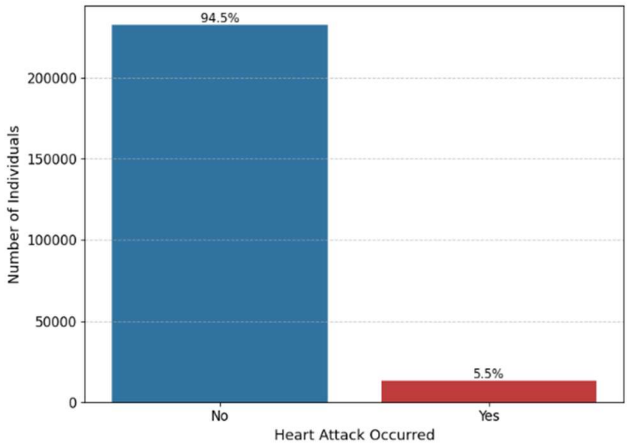
DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

III. HASIL DAN PEMBAHASAN

A. Jenis Penelitian dan Sumber Data

Sebelum dilakukan tahap pemodelan klasifikasi, terlebih dahulu dilakukan identifikasi terhadap distribusi kelas pada data penyakit jantung. Langkah ini penting untuk memahami keseimbangan data serta memberikan gambaran awal mengenai karakteristik variabel target yang akan diprediksi. Distribusi kejadian serangan jantung pada dataset yang digunakan disajikan pada Gambar 1.



Gambar 1. Distribusi Kejadian Serangan Jantung

Berdasarkan Gambar 1 terlihat bahwa distribusi kelas pada dataset menunjukkan ketidakseimbangan yang cukup signifikan. Sebagian besar observasi termasuk dalam kategori tidak mengalami serangan jantung (*No*) dengan proporsi sekitar 94,5%, sedangkan kategori mengalami serangan jantung (*Yes*) hanya sekitar 5,5% dari total data. Kondisi ini mengindikasikan bahwa *dataset* termasuk dalam kategori *imbalanced dataset*, di mana jumlah observasi pada kelas minoritas jauh lebih kecil dibandingkan kelas mayoritas. Ketidakseimbangan ini berpotensi menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas dan kurang optimal dalam mendeteksi kejadian pada kelas minoritas. Oleh karena itu, diperlukan pendekatan pemodelan yang mampu menangkap pola pada kelas minoritas secara lebih efektif, serta penggunaan metrik evaluasi yang tidak hanya bergantung pada akurasi, tetapi juga mempertimbangkan *precision*, *recall*, dan *F1-score*.

B. *Feature Selection*

Feature selection dilakukan menggunakan CatBoost *Feature Importance* untuk mengidentifikasi variabel yang paling informatif terhadap prediksi kejadian serangan jantung. Metode ini mengukur kontribusi setiap fitur terhadap pembentukan pohon keputusan selama proses pelatihan model.

Berdasarkan hasil seleksi, diperoleh sepuluh fitur dengan nilai importance tertinggi yang kemudian digunakan sebagai ruang fitur utama pada tahap analisis selanjutnya. Peringkat fitur tersebut disajikan pada Tabel 1.

Tabel 1. *Top 10* Variabel Terpenting Berdasarkan CatBoost *Feature Importance*

Variabel	*Importance*
AgeCategory	22,0960
HadAngina	14,7299
ChestScan	9,0397
GeneralHealth	7,5417
Sex	4,7817
RemovedTeeth	4,4268
WeightInKilograms	3,9670
SmokerStatus	3,2837
HadDiabetes	3,1899
BMI	2,8099

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Fitur *AgeCategory, HadAngina,* dan *ChestScan* menunjukkan kontribusi terbesar dalam membedakan kelas target. Sepuluh fitur terpilih ini kemudian digunakan sebagai ruang fitur asli sebelum dilakukan proses *feature augmentation* melalui pembentukan fitur Logit dan LDR.

C. *Log Based Feature Augmentation* (LBFA)

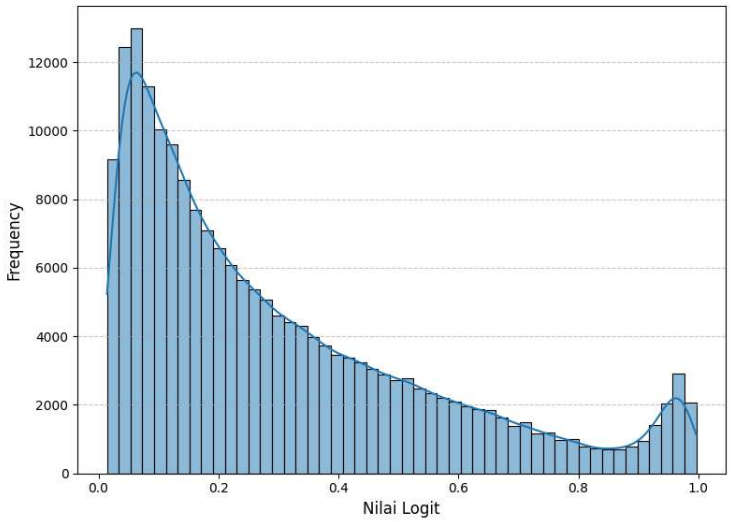
1. Fitur Logit

   Nilai fitur Logit yang dihasilkan untuk setiap observasi ditampilkan pada Tabel 2. Nilai tersebut merepresentasikan transformasi logit dari probabilitas kejadian kelas positif sehingga dapat digunakan sebagai fitur tambahan dalam proses augmentasi.

Tabel 2. Nilai Fitur Logit

i	Logit (η_i)
1	0,9851
2	0,4123
3	0,4701
4	0,6513
5	0,9443
⋮	⋮
19.6817	0,3686

Untuk memahami karakteristik fitur yang dihasilkan, distribusi nilai LOGIT divisualisasikan menggunakan histogram sebagaimana ditunjukkan pada Gambar 2.



Gambar 2. Distribusi Nilai Logit

Berdasarkan Gambar 2, distribusi nilai *logit* menunjukkan pola yang tidak simetris dengan konsentrasi observasi yang dominan pada rentang nilai rendah dan frekuensi yang menurun secara bertahap seiring meningkatnya nilai *logit*. Pola ini mengindikasikan bahwa sebagian besar observasi memiliki probabilitas yang relatif kecil terhadap terjadinya serangan jantung (Y = 1), sementara hanya sebagian kecil observasi yang memiliki probabilitas tinggi. Selain itu, terlihat adanya ekor distribusi pada nilai tinggi yang menunjukkan keberadaan sejumlah kecil observasi dengan tingkat keyakinan model yang kuat terhadap kelas positif. Karakteristik distribusi ini mencerminkan bahwa transformasi *logit* mampu memperluas skala probabilitas menjadi representasi yang lebih sensitif terhadap perbedaan antar observasi. Dengan demikian, fitur *logit* tidak hanya merepresentasikan peluang kejadian, tetapi juga mempertegas variasi tingkat risiko antar individu, sehingga berpotensi meningkatkan kemampuan model dalam membedakan pola pada data yang tidak seimbang melalui proses *feature augmentation*.

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

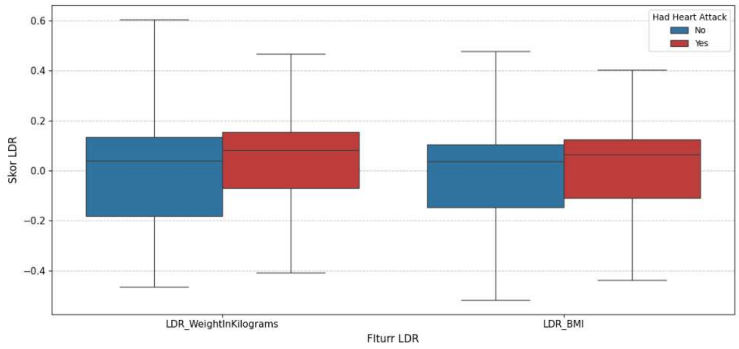
2. Fitur LDR

Nilai fitur LDR yang dihasilkan untuk setiap observasi disajikan pada Tabel 3. Nilai tersebut merepresentasikan log rasio kepadatan antara dua kelas sehingga dapat digunakan sebagai fitur tambahan dalam proses augmentasi.

Tabel 3. Nilai Fitur LDR

i	LDR *Weight (kg)*	LDR BMI
1	-0,2757	-0,3568
2	-0,3017	-0,3388
3	-0,2757	-0,1537
4	0,2279	0,1044
5	0,0845	0,0478
⋮	⋮	⋮
19.6817	-0,3017	-0,3388

Untuk memahami karakteristik fitur yang dihasilkan, distribusi nilai LDR divisualisasikan menggunakan boxplot sebagaimana ditunjukkan pada Gambar 3.



Gambar 3. Distribusi Nilai LDR

Berdasarkan Gambar 3, distribusi skor *LDR_WeightInKilograms* dan *LDR_BMI* menunjukkan adanya perbedaan kecenderungan nilai antara kelas *Had Heart Attack = Yes* dan *No*. Pada kedua fitur, kelas *Yes* cenderung memiliki median skor yang lebih tinggi dibandingkan kelas *No*, yang mengindikasikan adanya perbedaan karakteristik distribusi antar kelas target. Meskipun masih terdapat tumpang tindih antar distribusi, pergeseran median dan rentang interkuartil menunjukkan bahwa fitur hasil transformasi *Log Density Ratio* (LDR) memiliki kemampuan diskriminatif dalam membedakan pola antar kelas. Hal ini menunjukkan bahwa fitur LDR berpotensi membantu model klasifikasi dalam meningkatkan separasi antara kelas mayoritas dan minoritas pada data yang tidak seimbang.

D. Hasil Evaluasi Model

Kinerja model klasifikasi dievaluasi menggunakan beberapa metrik evaluasi yang dihitung dari matriks konfusi, yaitu *accuracy, precision, sensitivity, specificity*, dan F1-*score* sebagaimana didefinisikan pada Persamaan (16)–(20). Penggunaan beberapa metrik evaluasi ini bertujuan untuk memberikan gambaran performa model secara lebih komprehensif, terutama pada kasus ketidakseimbangan kelas (imbalanced class), di mana ukuran akurasi saja sering kali tidak cukup merepresentasikan kemampuan model dalam mendeteksi kelas minoritas.

Tabel 4. Evaluasi Model

Model	*Accuracy*	*Precision*	*Sensitivity*	*Specificity*	*F1-score*
LightGBM + LOGIT + LDR	0,9618	0,9485	0,9620	0,9625	0,9552
XGBoost + LOGIT + LDR	0,9187	0,8920	0,9100	0,9195	0,9009
CatBoost + LOGIT + LDR	0,8724	0,7810	0,8350	0,8756	0,8071

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Berdasarkan Tabel 4, kinerja model klasifikasi menunjukkan bahwa pendekatan *Hybrid Logit-Based Feature Augmentation* (LBFA) yang dikombinasikan dengan algoritma *boosting* mampu menghasilkan performa yang cukup baik dalam menangani data yang tidak seimbang. Model LightGBM + LOGIT + LDR memperoleh kinerja terbaik dengan nilai *accuracy* sebesar 0,9618, *precision* sebesar 0,9485, *recall* sebesar 0,9620, *specificity* sebesar 0,9625, serta F1-*score* sebesar 0,9552. Hasil ini menunjukkan bahwa model tidak hanya mampu mempertahankan tingkat akurasi yang tinggi, tetapi juga memiliki keseimbangan yang baik antara kemampuan mendeteksi kelas positif dan meminimalkan kesalahan klasifikasi.

Model XGBoost + LOGIT + LDR menunjukkan performa yang sedikit lebih rendah dengan F1-*score* sebesar 0,9009, meskipun masih mampu memberikan keseimbangan yang cukup baik antara *precision* sebesar 0,8920 dan *recall* sebesar 0,9100. Sementara itu, model CatBoost + LOGIT + LDR menghasilkan F1-*score* sebesar 0,8071, dengan nilai *precision* yang relatif lebih rendah dibandingkan model lainnya. Hal ini mengindikasikan bahwa model tersebut cenderung menghasilkan *false positive* yang lebih tinggi.

Secara keseluruhan, hasil ini menunjukkan bahwa integrasi fitur berbasis logit dan *log-density ratio* dalam kerangka LBFA mampu memperkaya representasi fitur dengan menggabungkan informasi probabilistik dan distribusional antar kelas. Pendekatan ini terbukti meningkatkan kemampuan model dalam membedakan pola pada data yang tidak seimbang, khususnya ketika dikombinasikan dengan algoritma LightGBM yang memiliki mekanisme pembelajaran yang efisien dalam menangkap hubungan kompleks antar variabel.

IV. KESIMPULAN

Berdasarkan hasil penelitian, penerapan pendekatan *Hybrid Logit-Based Feature Augmentation* (LBFA) yang mengintegrasikan transformasi *logit* dan *Log-Density Ratio* (LDR) terbukti mampu meningkatkan kualitas representasi fitur dalam proses klasifikasi data penyakit jantung yang tidak seimbang (*imbalanced data*). Hasil evaluasi menunjukkan bahwa kombinasi LBFA dengan algoritma LightGBM memberikan kinerja terbaik dibandingkan XGBoost dan CatBoost, dengan nilai *F1-score* tertinggi sebesar 0,9552 yang mencerminkan keseimbangan yang baik antara *precision* dan *recall*. Hal ini mengindikasikan bahwa pengayaan fitur berbasis informasi probabilistik dan distribusional dapat membantu model dalam membedakan pola antar kelas secara lebih efektif. Secara keseluruhan, pendekatan yang diusulkan tidak hanya meningkatkan performa model, tetapi juga memberikan alternatif strategi dalam menangani permasalahan klasifikasi pada *imbalanced data* melalui optimalisasi ruang fitur.

DAFTAR PUSTAKA

Almutairi, M., & Dardouri, S. (2025). Intelligent hybrid modeling for heart disease prediction. *Information (Switzerland), 16*(10), 869. https://doi.org/10.3390/info16100869

Aprianto, K., & Anasanti, M. D. (2025). Classifying heart disease through fusion of multi-source datasets: Integration of feature selection and explainable machine learning techniques. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 19*(3), 247–258. https://doi.org/10.22146/ijccs.92395

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

El-Sofany, H., Bouallegue, B., & El-Latif, Y. M. A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Scientific Reports, 14*(1), 23277. https://doi.org/10.1038/s41598-024-74656-2

Fan, Z., Gou, J., & Weng, S. (2024). A feature importance-based multi-layer CatBoost for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering, 36*(11), 5495–5507. https://doi.org/10.1109/TKDE.2024.3393472

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/0471722146

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146–3154).

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *International Journal of Computer Science, 1*(2), 111–117.

Li, J., & Cui, W. (2023). A new classifier for imbalanced data based on a generalized density ratio model. *Communications in Mathematics and Statistics, 11*(2), 327–347. https://doi.org/10.1007/s40304-021-00254-7

Mawarni, I., Ayshah, A. D., Yafe, D. F., & Fitri, F. (2025). Logit and complementary log-log modeling in the case of factors affecting heart failure disease. *UNP Journal of Statistics and Data Science, 3*(4), 430–436. https://doi.org/10.24036/ujsds/vol3-iss4/421

Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Scientific Reports, 13*(1), 22588. https://doi.org/10.1038/s41598-023-49962-w

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 6638–6648).

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Susrifalah, A., Vionanda, D., Kurniawati, Y., & Sulistiowati, D. (2025). Penerapan algoritma extreme gradient boosting dengan ADASYN untuk klasifikasi rumah tangga penerima Program Keluarga Harapan di Provinsi Sumatera Barat. *UNP Journal of Statistics and Data Science, 3*(2), 232–239. https://doi.org/10.24036/ujsds/vol3-iss2/369

Wang, H., Chen, C., Liu, W., Chen, K., Hu, T., & Chen, G. (2020). Incorporating label embedding and feature augmentation for multi-dimensional classification. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(4), 6178–6185. https://doi.org/10.1609/aaai.v34i04.6083

DOI: https://doi.org/10.24036/ujsds/vol4-iss2/478