Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

Deteksi Penyakit Jantung Menggunakan SVM dan XGBoost dengan Interpretabilitas SHAP dan Integrasi LLM

Raihan Al Aziz, Egia Rosi Subhiyakto*

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1]111202214808@mhs.dinus.ac.id, [2,*]egia@dsn.dinus.ac.id
Email Penulis Korespondensi: egia@dsn.dinus.ac.id
Submitted: **22/04/2026**; Accepted: **02/06/2026**; Published**: 05/06/2026**

Abstrak−Penyakit kardiovaskular merupakan penyebab utama kematian secara global yang menuntut deteksi dini yang akurat, namun keterbatasan akses terhadap tenaga medis spesialis di negara berkembang sering menghambat proses diagnosis yang tepat waktu. Penelitian ini bertujuan untuk mengatasi kesenjangan kritis antara tingginya akurasi model machine learning dalam riset akademis dan minimnya adopsi aplikasi klinis praktis dengan mengembangkan sistem triage penyakit jantung berbasis kecerdasan buatan hibrida yang aman dan terpercaya. Metodologi yang diusulkan mengintegrasikan arsitektur dual-model di mana Support Vector Machine berperan sebagai model prediksi utama dan Extreme Gradient Boosting sebagai model second-opinion, keduanya dioptimalkan dengan teknik oversampling SMOTE untuk menangani ketidakseimbangan kelas data, serta menerapkan SHAP untuk memberikan transparansi terhadap keputusan model black-box. Sistem ini diperkaya dengan inovasi Dynamic Prompt Engineering pada Large Language Model Mistral-7B untuk menerjemahkan probabilitas numerik menjadi narasi medis yang aman, personal, dan empatik. Hasil eksperimen menunjukkan bahwa model Support Vector Machine dengan kernel RBF memberikan kinerja superior dengan akurasi mencapai 90.22% dan sensitivitas 94.12%, yang sangat krusial untuk meminimalkan kasus negatif palsu (false negative) dalam skrining medis, mengungguli model Extreme Gradient Boosting yang mencatat akurasi 88.04%. Analisis interpretabilitas mengidentifikasi tipe nyeri dada, kadar kolesterol, dan detak jantung maksimum sebagai indikator risiko utama, memvalidasi keselarasan model dengan pedoman kardiologi standar. Mekanisme validasi keamanan ganda melalui batasan risiko terprogram dan kontrol temperatur generasi bahasa memastikan sistem tidak menghasilkan halusinasi diagnostik yang berbahaya. Kesimpulannya, sistem yang diimplementasikan sebagai layanan mikro berbasis FastAPI ini terbukti layak secara teknis dengan latensi rendah, menawarkan solusi skrining awal yang akurat, transparan, dan komunikatif untuk mendukung efisiensi layanan kesehatan.

Kata Kunci: Explainable AI; Large Language Models; Machine Learning; Penyakit Jantung; Support Vector Machine

Abstract−Cardiovascular disease remains the leading cause of death globally, demanding accurate early detection, yet limited access to specialist medical personnel in developing countries often hinders timely diagnosis. This study aims to address the critical gap between the high accuracy of machine learning models in academic research and the minimal adoption of practical clinical applications by developing a safe and trustworthy hybrid artificial intelligence-based heart disease triage system. The proposed methodology integrates a dual-model architecture in which Support Vector Machine serves as the primary prediction model and Extreme Gradient Boosting as a second-opinion model, both optimized with SMOTE oversampling technique to handle class imbalance, and implements SHAP to provide transparency in black-box model decisions. The system is enriched with Dynamic Prompt Engineering innovation on the Mistral-7B Large Language Model to translate numerical probabilities into safe, personalized, and empathetic medical narratives. Experimental results show that the Support Vector Machine model with RBF kernel delivers superior performance with an accuracy of 90.22% and sensitivity of 94.12%, which is crucial for minimizing false negative cases in medical screening, outperforming the Extreme Gradient Boosting model which recorded 88.04% accuracy. Interpretability analysis identified chest pain type, cholesterol level, and maximum heart rate as the primary risk indicators, validating the model's alignment with standard cardiology guidelines. A dual safety validation mechanism through programmed risk thresholds and language generation temperature control ensures the system does not produce harmful diagnostic hallucinations. In conclusion, the system implemented as a FastAPI-based microservice is proven technically feasible with low latency, offering an accurate, transparent, and communicative early screening solution to support healthcare service efficiency.

Keywords: Explainable AI; Heart Disease; Large Language Models; Machine Learning; Support Vector Machine

1. PENDAHULUAN

Penyakit kardiovaskular (cardiovascular disease/CVD) merupakan penyebab utama kematian di seluruh dunia, dengan angka mortalitas mencapai 17.9 juta jiwa per tahun menurut World Health Organization (WHO). Di Indonesia, prevalensi penyakit jantung terus meningkat seiring dengan perubahan pola hidup dan keterbatasan akses ke tenaga medis spesialis. Deteksi dini sangat krusial untuk mencegah komplikasi fatal, namun proses diagnosis konvensional sering kali memerlukan waktu lama, melibatkan serangkaian tes medis kompleks, dan bergantung pada ketersediaan spesialis yang terbatas. Dalam dekade terakhir, machine learning (ML) telah menunjukkan potensi besar dalam memprediksi penyakit jantung dengan akurasi tinggi. Penelitian oleh Simatupang et al. menunjukkan bahwa Support Vector Machine (SVM) mencapai akurasi 92% pada dataset Cleveland Heart Disease [1], Bangun et al. mencapai akurasi 92% dengan recall 96% menggunakan SVM dengan K-Fold Cross Validation [2], sementara Roopa & Ramanjinappa mengimplementasikan XGBoost dengan SMOTE mencapai 90% akurasi [3], serta Rohmayani et al. yang memvalidasi efektivitas pendekatan ini dengan akurasi 92% pada studi terbaru [4]. Nagavalli et al. juga membuktikan efektivitas kombinasi SVM dan XGBoost dalam sistem deteksi penyakit jantung berbasis clinical decision support [5]. Namun, mayoritas penelitian ini berhenti pada pelaporan metrik akurasi tanpa deployment ke sistem praktis yang dapat digunakan oleh klinisi atau pasien-kesenjangan ini menghambat adopsi teknologi ML di setting klinis karena tidak tersedia interface user-friendly atau pathway integrasi dengan clinical workflows.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

Interpretabilitas model ML merupakan isu kritis dalam aplikasi medis karena model black-box sulit dipahami oleh klinisi, mengurangi kepercayaan terhadap prediksi AI. SHAP (SHapley Additive exPlanations) yang diperkenalkan oleh Lundberg & Lee telah menjadi standar emas untuk interpretasi model karena memberikan penjelasan berbasis teori game dengan jaminan teoritis yang kuat dibanding metode lainnya seperti LIME [6]. Ganie et al. mendemonstrasikan bahwa pendekatan ensemble yang dijelaskan dengan SHAP dapat meningkatkan transparansi model penyakit jantung secara signifikan [7]. Survey oleh Tjoa & Guan menekankan pentingnya explainable AI (XAI) dalam konteks medis untuk transparansi keputusan dan trustworthiness [8], namun Ghassemi et al. mengingatkan bahwa XAI saat ini memberikan "false hope" pada level keputusan individual pasien-sehingga mekanisme validasi tambahan diperlukan untuk memastikan keamanan deployment medis [9]. Di sisi lain, Large Language Models (LLMs) menunjukkan bahwa model berbasis transformer dapat mengkodekan pengetahuan klinis dan menghasilkan narasi medis yang koheren. Singhal et al. mengembangkan Med-PaLM yang mencapai 67.6% akurasi pada ujian lisensi medis melalui instruction prompt tuning, mendemonstrasikan bahwa LLMs dapat berkomunikasi dalam bahasa natural [10]. Untuk deployment efisien, Mistral 7B menawarkan alternatif dengan 7-billion parameters menggunakan grouped-query attention dan sliding window attention, outperforming Llama 2 13B sambil maintaining computational efficiency untuk setting resource-limited [11].

Beberapa upaya mengembangkan chatbot medis telah dilakukan, namun memiliki keterbatasan. Majeed et al. mengembangkan chatbot diagnostik untuk penyakit jantung menggunakan XGBoost pada Cleveland Heart Disease dataset dengan performa superior, namun tidak mengintegrasikan explainability (SHAP) dan LLM-based natural language generation (NLG), sehingga output hanya berupa classification result tanpa narrative explanation [12]. Antia et al. mengembangkan Healthy Heart Assistant berbasis WhatsApp untuk self-care management pasien hipertensi dengan satisfaction score 90%, namun hanya fokus pada edukasi tanpa mengintegrasikan model prediksi ML berbasis clinical features untuk risk assessment [13]. Nurlita & Munawaroh mengembangkan chatbot dengan NLP menggunakan Laravel+Gemini API yang menunjukkan feasibility real-time response, tetapi aplikasinya di luar domain medis dan tanpa safety validation yang diperlukan untuk medical decision making [14]. Meskipun terdapat kemajuan dalam prompt engineering untuk kesehatan mental [15][16], optimasi RAG dialog medis [17], dan arsitektur microservice data Kesehatan [18], belum ada sistem terpadu untuk kardiologi yang menggabungkan elemen-elemen tersebut. Kesenjangan riset yang teridentifikasi: pertama, penelitian ML berhenti pada metrik akurasi tanpa deployment praktis; kedua, chatbot medis yang ada tidak mengintegrasikan prediksi ML, explainability (SHAP), dan LLM-based NLG dalam satu framework terpadu; ketiga, belum ada Dynamic Prompt Engineering yang mengadaptasi prompt LLM berdasarkan probabilitas model dan SHAP values untuk menghasilkan narasi medis yang aman dan personal; keempat, dual-layer safety validation (hard-coded threshold + LLM temperature control) belum diterapkan untuk mencegah hallucination LLM dalam diagnosis medis. Di samping kesenjangan teknis tersebut, terdapat pula dimensi etis dan legal yang belum tertangani: penggunaan LLM dalam konteks medis menimbulkan risiko hallucination yang dapat membahayakan pasien jika tidak dikontrol, serta isu pertanggungjawaban hukum ketika keputusan klinis dipengaruhi oleh output AI generatif. Penelitian ini merespons risiko tersebut melalui mekanisme dual-layer safety validation yang dirancang secara eksplisit untuk mencegah output LLM yang tidak konsisten dengan probabilitas model prediksi.

Penelitian ini bertujuan mengembangkan sistem triage penyakit jantung berbasis hybrid AI yang mengintegrasikan: prediksi akurat menggunakan dual-model architecture (SVM dengan RBF kernel dan XGBoost sebagai second-opinion) dengan preprocessing StandardScaler dan SMOTE, serta pembagian data training (80%) dan testing (20%); explainability berbasis SHAP mengidentifikasi top-3 fitur kontributif (Chest Pain Type, Cholesterol, MaxHR) untuk validasi bahwa model belajar pola medis legitimate [6]; Dynamic Prompt Engineering yang mengadaptasi prompt Mistral-7B berdasarkan probabilitas prediksi dan SHAP values untuk risk stratification (Low: =33%, Moderate: 33-66%, High: >66%) dengan medical disclaimers dan actionable recommendations dual-layer safety validation melalui hard-coded logic dan LLM temperature (0.3) untuk mencegah hallucination yang tidak konsisten dengan model output; serta deployment production-ready dalam FastAPI microservice dengan endpoint /parse (Named Entity Recognition), /diagnosis (ML inference + LLM response, latency 2.5-4.0s), dan /sessions (conversation history). Urgensi penelitian terletak pada kebutuhan translasi riset akademik ML ke aplikasi klinis yang aman dan dapat dipercaya. Dengan meningkatnya beban penyakit kardiovaskular di Indonesia dan keterbatasan tenaga medis spesialis, sistem triage otomatis dapat berfungsi sebagai first-line screening tool yang mengurangi beban klinisi, meningkatkan efisiensi alokasi resources medis, dan memberikan akses risk assessment kepada populasi underserved. Kontribusi penelitian ini bukan hanya mencapai akurasi tinggi (target =90%), tetapi engineering sistem AI medis yang lengkap, aman, transparan dengan mempertimbangkan explainability, conversational interface, dan medical safety protocols-aspek-aspek yang menjadi gap utama antara riset akademis dan adoptability klinis.

Penelitian menggunakan Heart Disease Dataset standar dari repositori publik UCI/Kaggle dengan pembagian data training (80%) dan testing (20%) untuk evaluasi robust. Model ML fokus pada SVM (RBF kernel, soft-margin) dan XGBoost (gradient boosting dengan regularization) berdasarkan evidence bahwa keduanya memberikan performa terbaik pada dataset medis berukuran kecil hingga menengah. Explainability menggunakan SHAP karena theoretical guarantees yang kuat (consistency, local accuracy, missingness properties) dibanding metode XAI lainnya seperti LIME. LLM menggunakan Mistral-7B karena beberapa pertimbangan spesifik: (1) computational efficiency dengan 7 miliar parameter yang memungkinkan deployment di lingkungan resource-limited tanpa infrastruktur GPU kelas enterprise; (2) sifat open-source yang mendukung reproducibility riset dan menghindari ketergantungan vendor; (3)

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

kemampuan multilingual yang memadai untuk narasi Bahasa Indonesia; dan (4) performa yang telah terbukti melampaui Llama 2 13B pada berbagai benchmark umum [11]. Meskipun model domain-spesifik seperti Med-PaLM atau BioMedLM menawarkan kedalaman pengetahuan klinis yang lebih tinggi, kebutuhan infrastruktur dan lisensinya tidak sesuai dengan konteks deployment penelitian ini yang bersifat proof-of-concept. Sistem chatbot menggunakan Bahasa Indonesia dan Bahasa Inggris untuk narasi medis dengan prompt templates culturally appropriate. Penelitian merupakan proof-of-concept dengan evaluasi berbasis dataset offline-clinical validation prospektif dengan real patients akan menjadi future work memerlukan ethical clearance dan kolaborasi institusi kesehatan. Sistem dikembangkan sebagai research prototype belum memenuhi standar regulasi medical device (FDA approval, BPOM certification), sehingga deployment ke setting klinis nyata memerlukan proses sertifikasi tambahan. Sistem dirancang sebagai triage tool untuk risk assessment, bukan pengganti diagnosis profesional-output sistem harus diverifikasi oleh klinisi berlisensi sebelum keputusan klinis dibuat.

Secara teoritis, penelitian ini memberikan tiga kontribusi utama. Pertama, penelitian ini memperkenalkan Dynamic Prompt Engineering sebagai metodologi baru untuk menjembatani output probabilistik ML dengan natural language generation dalam konteks medis, yakni dengan menginjeksikan structured constraints dari prediksi model dan nilai SHAP ke dalam konteks prompt demi konsistensi antara bukti numerik dan narasi penjelasan. Kedua, penelitian ini mendemonstrasikan integrasi SHAP dengan LLM untuk menghasilkan explainability yang secara teknis akurat sekaligus mudah dipahami, dengan menerjemahkan nilai SHAP yang abstrak menjadi interpretasi medis berbahasa sederhana yang dapat dipahami pasien awam. Ketiga, penelitian ini memberikan proof-of-concept mekanisme validasi berlapis (dual-layer validation) sebagai solusi atas kritik bahwa XAI saja tidak cukup untuk safe deployment medis kombinasi rule-based safeguards dan fleksibilitas generatif menghasilkan sistem yang robust terhadap edge cases sekaligus tetap empatik dalam komunikasi dengan pasien.

Secara praktis: menyediakan tool triage otomatis yang mengurangi beban klinisi dalam screening awal pasien berisiko tinggi; memberikan akses risk assessment berbasis AI sebagai early warning system untuk screening pre-consultation; meningkatkan efisiensi operasional institusi kesehatan sebagai first-line screening tool yang terintegrasi dengan intake processes; menyediakan blueprint untuk pengembangan sistem hybrid AI (ML + XAI + LLM) di domain medis lainnya seperti diabetes atau cancer screening, mempercepat translasi riset akademik ke aplikasi klinis dengan reusable architectural patterns dan safety frameworks.

Penelitian ini disusun dalam empat bagian utama untuk memudahkan pemahaman alur pengembangan sistem secara menyeluruh. Bagian 2 menjelaskan metodologi yang digunakan, mencakup desain arsitektur sistem, pemilihan dan konfigurasi algoritma machine learning, serta strategi preprocessing dan evaluasi model. Bagian 3 menyajikan hasil eksperimen secara kuantitatif dan kualitatif, meliputi analisis performa model klasifikasi, interpretasi SHAP, serta evaluasi kualitas narasi yang dihasilkan oleh LLM. Bagian 4 menutup penelitian dengan kesimpulan yang merangkum temuan utama serta rekomendasi konkret untuk pengembangan lanjutan, termasuk validasi klinis prospektif dan integrasi dengan sistem rekam medis elektronik. Dengan demikian, penelitian ini diharapkan tidak hanya berkontribusi pada pengembangan ilmu pengetahuan di bidang kecerdasan buatan dan informatika kesehatan, tetapi juga memberikan dampak praktis yang nyata bagi peningkatan kualitas layanan kesehatan, khususnya dalam upaya deteksi dini penyakit jantung yang lebih akurat, transparan, dan dapat diakses oleh seluruh lapisan masyarakat.

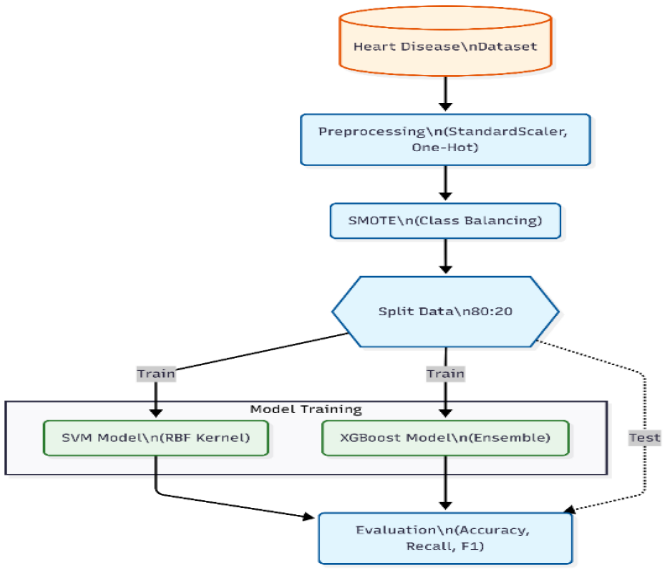
2. METODOLOGI PENELITIAN

Untuk menerapkan sistem hibrida klinis dalam konteks prediksi penyakit jantung dan komunikasi pasien, terdapat beberapa tahapan sebagai berikut:

a. Mengumpulkan data klinis yang relevan menggunakan Heart Disease Dataset standar yang mencakup parameter medis kritis dari repositori publik (Kaggle).
b. Melakukan analisis dan pra-pemrosesan data yang meliputi: pembersihan data, normalisasi numerik menggunakan StandardScaler untuk menyamaratakan rentang nilai variabel (seperti usia dan tekanan darah), pengkodean variabel kategorikal (One-Hot Encoding). Diketahui bahwa data medis sering mengalami ketidakseimbangan kelas (imbalanced class), dilakukan teknik over-sampling menggunakan SMOTE (Synthetic Minority Over-sampling Technique). Setelah data seimbang, data dibagi menjadi dua yaitu data training untuk pelatihan model (80%), dan data testing untuk evaluasi (20%).
c. Data yang telah disiapkan digunakan untuk pemodelan menggunakan dua algoritma: Model SVM dengan kernel RBF untuk klasifikasi non-linear dan XGBoost sebagai model pembanding yang memiliki efisiensi tinggi pada data tabular.
d. Mengevaluasi model dari hasil prediksi menggunakan evaluasi metrik klasifikasi (accuracy, precision, recall, f1-score). Selain itu, diimplementasikan lapisan interpretabilitas menggunakan SHAP yang diintegrasikan ke dalam antarmuka Generative AI (Mistral-7B) untuk menghasilkan diagnosis naratif yang empatik.
e. Mengintegrasikan model prediksi ke dalam sistem chatbot berbasis FastAPI dan LLM, memungkinkan komunikasi interaktif dengan pasien melalui antarmuka konversasional

Arsitektur sistem hibrida yang diusulkan, mencakup seluruh alur dari pra-pemrosesan data hingga deployment chatbot, ditunjukkan secara visual pada Gambar 1.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722



Gambar 1. Arsitektur Sistem Hibrida

2.1. Pra-pemrosesan Data dan SMOTE

Kualitas data sangat mempengaruhi kinerja model. Proses dimulai dengan normalisasi fitur numerik x menggunakan rumus skalasi standar:

$$x_{\text{norm}} = \frac{x-\mu}{\sigma} \tag{1}$$

Dimana x adalah nilai asli, μ adalah rata-rata fitur, dan σ adalah standar deviasi. Normalisasi ini penting terutama untuk algoritma berbasis jarak seperti SVM. Selanjutnya, untuk menangani ketidakseimbangan kelas yang umum dalam dataset medis, teknik SMOTE (Synthetic Minority Over-sampling Technique) diterapkan. SMOTE bekerja dengan mensintesis sampel baru untuk kelas minoritas berdasarkan k-nearest neighbors. Setiap sampel sintetis dihasilkan dengan persamaan:

$$x_{\text{new}} = x_i + \lambda \cdot (x_{\text{nn}} - x_i) \tag{2}$$

Dimana x_i adalah sampel minoritas yang dipilih, x_nn adalah salah satu dari k tetangga terdekat yang dipilih secara acak, dan $\lambda \in [0,1]$ adalah bilangan acak. Teknik ini memastikan bahwa kedua kelas memiliki representasi yang seimbang dalam data training, sehingga model tidak bias terhadap kelas mayoritas [19].

2.2. Support Vector Machine (SVM)

SVM digunakan sebagai classifier utama karena kemampuannya yang kuat dalam menangani data berdimensi tinggi dan non-linear. SVM bekerja dengan mencari hyperplane optimal yang memisahkan kelas dengan margin maksimum. Karena data medis jarang terpisahkan secara sempurna secara linear, kami menggunakan soft-margin SVM yang mengizinkan beberapa kesalahan klasifikasi dengan penalti. Masalah optimasi ini dirumuskan sebagai:

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \tag{3}$$

Dengan kendala:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Dimana w adalah vektor bobot hyperplane, b adalah bias, ξ_i adalah variabel slack yang mengizinkan klasifikasi yang salah, dan C adalah hyperparameter penalti yang mengontrol keseimbangan antara margin maksimum dan minimasi error. Nilai C yang besar mendorong model untuk fit lebih ketat terhadap data training (risiko overfitting), sedangkan nilai C kecil menghasilkan margin yang lebih lebar tetapi dengan toleransi error lebih besar. Untuk menangani data yang tidak terpisahkan secara linear, kami menggunakan Kernel Radial Basis Function (RBF):

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \tag{4}$$

Kernel ini memetakan fitur dari dimensi asli ke dimensi yang lebih tinggi dimana pemisahan linear dimungkinkan. Parameter γ mengontrol pengaruh setiap sampel pelatihan: nilai γ yang besar membuat model lebih sensitif terhadap fluktuasi data lokal (meningkatkan risiko overfitting), sedangkan nilai γ yang kecil menghasilkan batas keputusan yang lebih halus dan generalisable. Dalam implementasi kami, digunakan nilai default γ = 1/n_features yang

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

ditetapkan oleh scikit-learn. Dalam penelitian ini, parameter SVM menggunakan nilai default scikit-learn (C=1.0, γ='scale', kernel='rbf') tanpa optimasi hyperparameter lebih lanjut. Penggunaan nilai default ini merupakan keterbatasan yang diakui; penelitian lanjutan disarankan untuk menerapkan Grid Search atau Random Search dengan K-Fold Cross Validation (k=5 atau k=10) guna memperoleh klaim akurasi yang lebih dapat digeneralisasi, terutama mengingat ukuran dataset yang tergolong kecil hingga menengah.

2.3. XGBoost (Extreme Gradient Boosting)

XGBoost adalah algoritma ensemble berbasis gradient boosting yang membangun model prediksi secara iteratif. Model ini dipilih sebagai pembanding karena performanya yang superior pada data tabular dan efisiensi komputasinya yang tinggi. Pada setiap iterasi t, model baru f_t ditambahkan untuk meminimalkan fungsi loss $\mathcal{L}^{(t)}$:

$$\mathcal{L}^{(t)} = \mathrm{l} \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{5}$$

Dimana *l* adalah fungsi *loss* berupa *binary cross-entropy* untuk klasifikasi biner dengan formulasi l = -[y log(p̂) + (1-y) log(1-p̂)]; $\hat{y}_i^{(t-1)}$ adalah prediksi kumulatif dari semua pohon sebelumnya hingga iterasi t-1; $f_t(x_i)$ adalah kontribusi pohon keputusan baru pada iterasi t; dan $\Omega(f_t) = \gamma T + (1/2)\lambda\|w\|^2$ adalah term regularisasi yang mencegah *overfitting*, dengan T menyatakan jumlah daun dan w menyatakan bobot daun pada pohon tersebut. XGBoost menggunakan aproksimasi Taylor orde kedua untuk mempercepat optimasi dan meningkatkan stabilitas. Algoritma ini juga mengimplementasikan shrinkage (learning rate) yang membuat update setiap pohon lebih konservatif, meningkatkan generalisasi model.

2.4. Evaluasi Model dan Metrik Klasifikasi

Model SVM dan XGBoost dievaluasi menggunakan metrik standar klasifikasi pada test set (20% dari data):

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{6}$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{7}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{(TP+FN)} \tag{8}$$

$$\text{F1-Score} = \frac{(Precision \cdot Recall)}{(Precision + Recall)} \tag{9}$$

Dimana TP (True Positive), TN (True Negative), FP (False Positive), dan FN (False Negative) adalah hasil dari confusion matrix. Selain itu, kami menggunakan Confusion Matrix untuk secara visual menganalisis jenis kesalahan (false positives vs false negatives) yang dibuat oleh setiap model. Dalam konteks medis, recall yang tinggi lebih diprioritaskan daripada precision pure, karena melewatkan pasien dengan penyakit jantung (false negative) lebih berbahaya daripada memberikan alarm palsu. Perlu dicatat bahwa evaluasi dilakukan menggunakan skema *hold-out* 80:20 tanpa *K-Fold Cross Validation*. Pada dataset berukuran kecil seperti yang digunakan dalam penelitian ini, pendekatan ini memiliki keterbatasan dalam hal stabilitas estimasi performa; hasil akurasi yang dilaporkan karenanya perlu diinterpretasikan sebagai indikasi performa pada partisi data tertentu, bukan sebagai estimasi generalisasi yang definitif.

2.5. Interpretabilitas dengan SHAP dan Integrasi Generative AI

Untuk menjelaskan prediksi model kepada pengguna dan profesional medis, kami mengimplementasikan SHAP (SHapley Additive exPlanations), yang mengukur kontribusi setiap fitur terhadap prediksi berdasarkan teori permainan kooperatif. Nilai SHAP φ_i untuk fitur i dihitung dengan mempertimbangkan semua kemungkinan kombinasi fitur dalam himpunan F:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)] \tag{10}$$

Dalam rumus tersebut, F merupakan himpunan lengkap dari semua fitur yang tersedia, sedangkan S adalah subset dari fitur yang tidak menyertakan fitur i. Fungsi $f_x(S)$ merepresentasikan output model ketika hanya fitur-fitur dalam subset S yang digunakan, sementara fitur-fitur lainnya dirata-ratakan. Adapun perbedaan $[f_x(S \cup \{i\}) - f_x(S)]$ mengukur kontribusi marginal fitur i terhadap prediksi model ketika fitur tersebut ditambahkan ke dalam subset S. Nilai SHAP ini, bersama dengan probabilitas prediksi dari model, dimasukkan ke dalam Large Language Model (Mistral-7B) melalui sebuah prompt template yang dirancang secara khusus untuk menghasilkan respons yang aman secara medis:

$$R = \mathrm{LLM}(P(y \mid x), \Phi, C) \tag{11}$$

Dalam rumus tersebut, P(y|x) adalah probabilitas kelas positif (risiko penyakit jantung) yang dihasilkan dari SVM atau XGBoost, sedangkan $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_n\}$ adalah vektor nilai SHAP untuk semua fitur yang menunjukkan

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

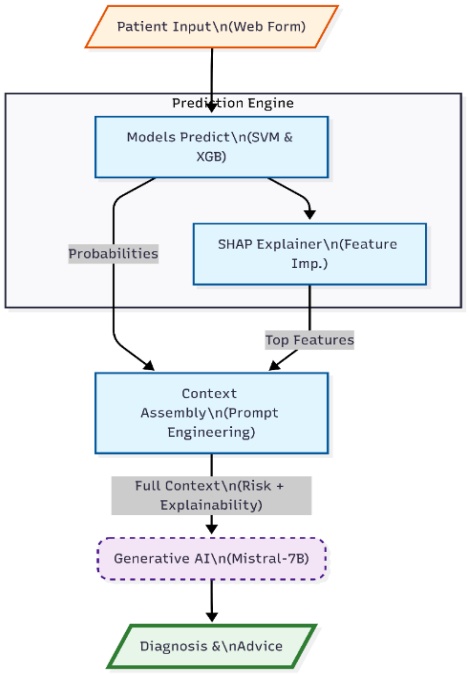
kontribusi relatif setiap fitur. C merepresentasikan konteks medis sistem, mencakup template prompt yang memastikan respons mematuhi protokol keselamatan, seperti larangan menyatakan diagnosis definitif dan kewajiban merekomendasikan konsultasi dokter. Adapun R adalah respons teks yang dihasilkan, yang menggabungkan temuan numerik dengan saran medis yang empatik dan aman. Pendekatan hybrid ini menjembatani kesenjangan antara akurasi algoritmik model machine learning dan kemampuan pemahaman/interpretasi yang diperlukan untuk komunikasi pasien yang efektif. Pendekatan hybrid ini menjembatani kesenjangan antara akurasi algoritmik model machine learning dan kemampuan pemahaman/interpretasi yang diperlukan untuk komunikasi pasien yang efektif

2.6. Integrasi Chatbot dan Sistem Deployment

Model prediksi yang telah dilatih diintegrasikan ke dalam sistem chatbot berbasis FastAPI dengan alur berikut:

a. Input Collection: Chatbot mengajukan pertanyaan terstruktur kepada pengguna untuk mengumpulkan parameter klinis (usia, jenis kelamin, jenis nyeri dada, tekanan darah, dll.)
b. Feature Engineering: Jawaban pengguna dikonversi ke vektor fitur numerik sesuai dengan skema encoding yang digunakan saat training
c. Model Inference: Vektor fitur dikirim ke model SVM dan XGBoost yang tersimpan untuk menghasilkan probabilitas prediksi
d. SHAP Explanation: Nilai SHAP dihitung untuk menunjukkan kontribusi setiap fitur terhadap prediksi
e. LLM Response Generation: Probabilitas, nilai SHAP, dan konteks medis diumpankan ke Mistral-7B untuk menghasilkan respons naratif
f. Safety Filter: Output LLM melalui filter keselamatan yang memastikan tidak ada klaim diagnostik definitif dan selalu merekomendasikan konsultasi profesional medis

Mekanisme Safety Filter bekerja dengan dua lapis validasi: pada lapis pertama, logika hard-coded Python memeriksa apakah kategori risiko dalam teks output LLM (Low/Moderate/High) konsisten dengan rata-rata probabilitas model (SVM dan XGBoost); jika output LLM mengklaim risiko rendah sementara probabilitas rata-rata >0.33, respons tersebut ditolak dan digantikan oleh template respons aman yang telah ditetapkan. Pada lapis kedua, parameter temperature LLM diatur pada nilai 0.3 untuk membatasi variabilitas generasi teks dan memastikan konsistensi saran di seluruh sesi. Arsitektur ini memungkinkan sistem untuk memberikan penilaian risiko yang didukung oleh matematika sambil tetap dapat dikomunikasikan dengan cara yang mudah dipahami dan aman secara klinis oleh pengguna akhir. Alur lengkap dari input pengguna hingga respons sistem, mencakup integrasi antara komponen ML inference, SHAP engine, dan LLM, ditunjukkan pada Gambar 2.



Gambar 2. Alur Inferensi dan Integrasi Chatbot

3. HASIL DAN PEMBAHASAN

Evaluasi kinerja model klasifikasi penyakit jantung dilakukan dengan membandingkan algoritma Support Vector Machine (SVM) dan XGBoost. Pengujian menggunakan data uji sebesar 20% dari total dataset untuk mengukur metrik Akurasi, Presisi, Recall, dan F1-Score.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

3.1. Kinerja Model Klasifikasi

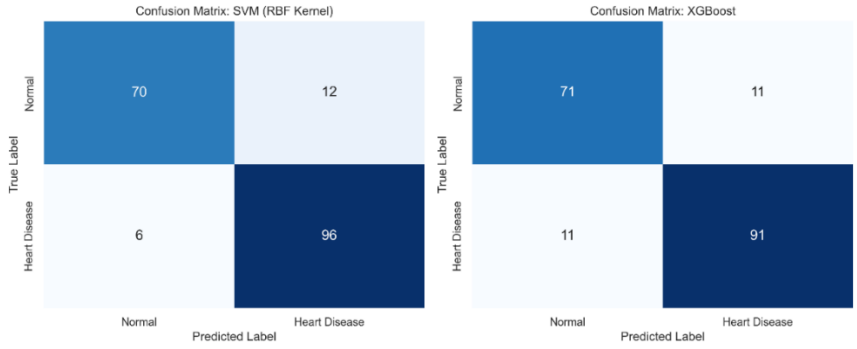
Hasil pengujian kuantitatif kedua model disajikan pada Tabel 1. Model SVM dengan kernel RBF menunjukkan keunggulan performa keseluruhan dengan akurasi mencapai 90.22%, lebih tinggi dibandingkan XGBoost yang mencapai 88.04%. Perbedaan ini signifikan mengingat kedua model dilatih pada dataset yang sama dengan konfigurasi preprocessing identik (StandardScaler + SMOTE).

Tabel 1. Perbandingan Kinerja Model SVM dan XGBoost

Model	Akurasi	Presisi	Recall	F1-Score
SVM	90.22%	88.89%	94.12%	91.43%
XGboost	88.04%	89.22%	89.22%	89.22%

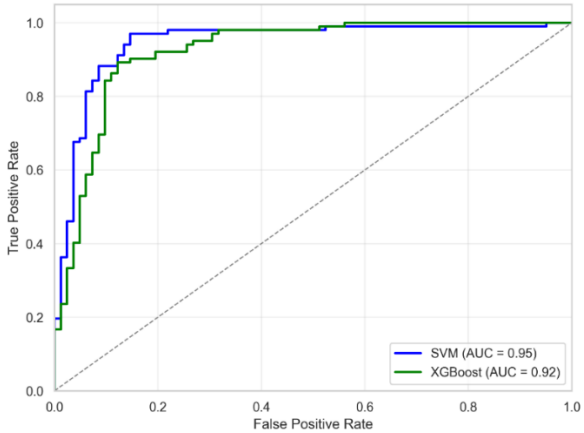
Keunggulan SVM terutama terlihat pada nilai Recall (Sensitivity) sebesar 94.12%, yang merupakan metrik paling kritis dalam konteks deteksi medis. Recall mengukur proporsi kasus positif aktual yang berhasil diidentifikasi: dari 102 pasien yang sebenarnya menderita penyakit jantung dalam test set, model SVM berhasil mengidentifikasi 96 kasus dan hanya melewatkan 6 kasus (False Negative). Dalam konteks klinis, False Negative adalah kesalahan paling berbahaya karena dapat menyebabkan pasien berisiko tinggi tidak mendapatkan intervensi medis yang diperlukan.

Model SVM mengorbankan sedikit Presisi (88.89% vs 89.22%) untuk mendapatkan Recall yang jauh lebih tinggi, menghasilkan 12 False Positives dibandingkan 11 pada XGBoost (selisih minimal). Dalam sistem triage medis berbasis AI, trade-off ini dianggap acceptable karena: (1) False Positive dapat dikoreksi melalui pemeriksaan lanjutan oleh dokter, (2) False Negative dapat berakibat fatal jika pasien tidak kembali untuk pemeriksaan, dan (3) nilai F1-Score SVM (91.43%) yang lebih tinggi menunjukkan keseimbangan precision-recall secara keseluruhan tetap optimal



Gambar 3. Confusion Matrix: Perbandingan Model SVM dan XGBoost

Visualisasi Confusion Matrix pada Gambar 3 memperlihatkan detail distribusi kelas prediksi versus aktual. Gambar 4, Kurva ROC (Receiver Operating Characteristic), mengonfirmasi reliabilitas kedua model dengan Area Under Curve (AUC) yang mendekati 1.0. Selain itu, Kurva Pembelajaran (Learning Curve) pada Gambar 5 menunjukkan bahwa model XGBoost mampu belajar secara stabil dari data latih tanpa indikasi overfitting yang signifikan, menjadikannya kandidat yang kuat untuk integrasi dalam sistem yang membutuhkan generalisasi baik.

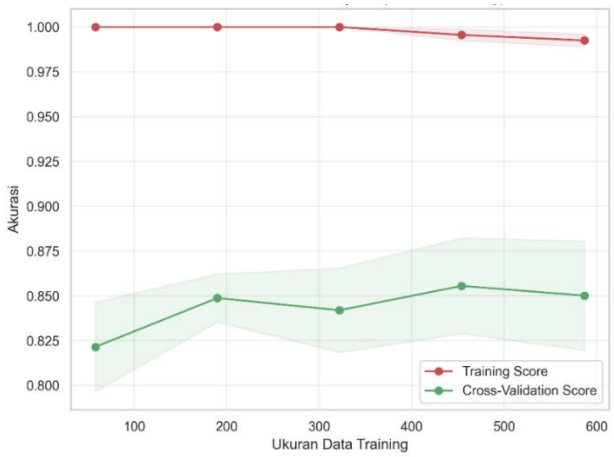


Gambar 4. Kurva ROC-AUC untuk Evaluasi Model

Kurva ROC (Receiver Operating Characteristic) pada Gambar 4 mengonfirmasi reliabilitas kedua model dengan Area Under Curve (AUC) yang mendekati nilai ideal 1.0. Nilai AUC yang tinggi mengindikasikan bahwa kedua model memiliki kemampuan diskriminasi yang sangat baik di berbagai threshold probabilitas, dengan kemampuan pemisahan kelas yang kuat. Kurva ROC yang mendekati sudut kiri atas menunjukkan bahwa model dapat

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

mencapai True Positive Rate (TPR) yang tinggi dengan False Positive Rate (FPR) yang rendah, merupakan karakteristik ideal untuk sistem diagnostic.

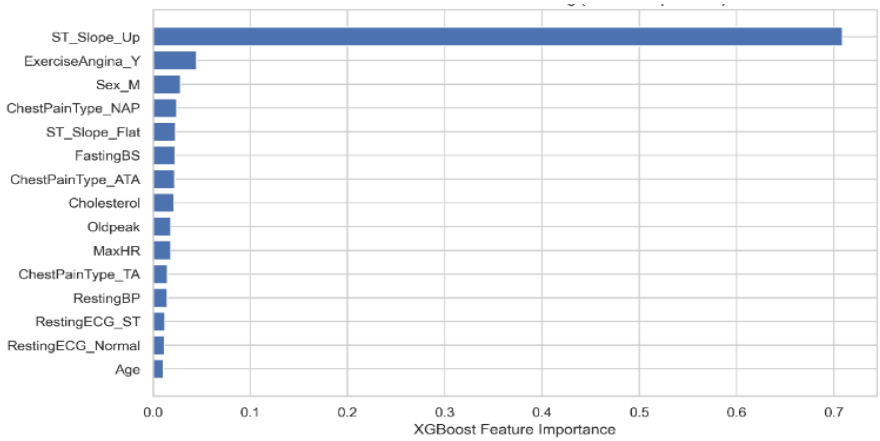


Gambar 5. Kurva Pembelajaran (XGBoost Stability)

SVM dipilih sebagai model utama untuk sistem chatbot karena Recall tertinggi (94.12%) yang menjadi prioritas dalam deteksi medis, F1-Score tertinggi (91.43%), dan akurasi keseluruhan tertinggi (90.22%). XGBoost dipertahankan sebagai model pembanding karena memberikan "second opinion" dalam sistem dual-model, menunjukkan stabilitas generalisasi yang sangat baik, dan mempertahankan performa yang kuat (88.04% accuracy, F1 89.22%). Hasil ini konsisten dengan literatur yang menunjukkan bahwa SVM dengan kernel RBF sangat efektif untuk dataset medis berukuran kecil hingga menengah (n < 1000), sementara XGBoost unggul dalam hal robustness dan interpretabilitas pada data tabular [20].

3.2. Analisis Fitur (Explainability)

Transparansi keputusan model adalah komponen kritis dalam sistem diagnostik berbasis AI. Kami menganalisis Feature Importance menggunakan dua pendekatan: Feature Importance bawaan XGBoost dan SHAP (SHapley Additive exPlanations) values untuk interpretasi yang lebih komprehensif dan dapat diandalkan secara statistik.



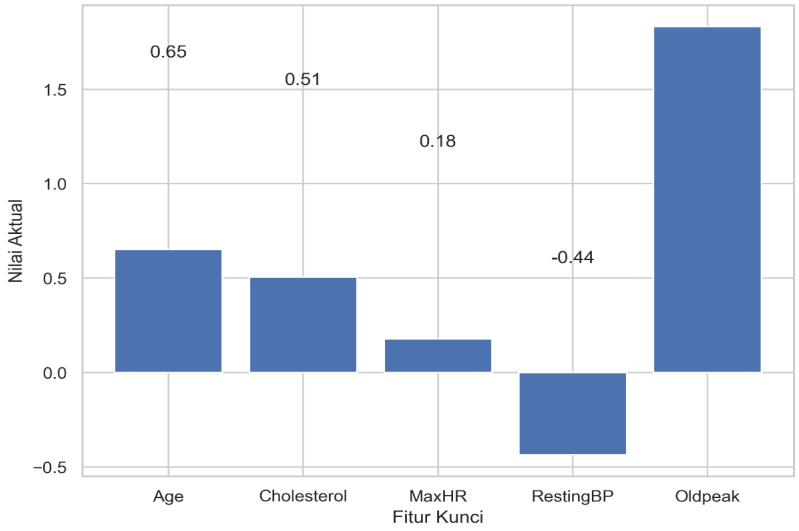
Gambar 6. Analisis Global Fitur Penting (Feature Importance)

Hasil analisis Feature Importance dari model XGBoost ditampilkan pada Gambar 6. Analisis menunjukkan bahwa fitur 'Chest Pain Type' (Jenis Nyeri Dada), 'Cholesterol' (Kadar Kolesterol), dan 'MaxHR' (Detak Jantung Maksimum) adalah kontributor terbesar terhadap prediksi risiko penyakit jantung, dengan importance scores mencapai nilai teratas dalam model. 'Chest Pain Type' merupakan fitur paling dominan karena hubungannya yang langsung dengan presentasi klinis penyakit jantung; 'Cholesterol' mencerminkan faktor risiko kardiovaskular yang terkenal luas dalam literatur medis; dan 'MaxHR' menggambarkan respons kardiopulmoner terhadap stress, yang merupakan indikator penting kesehatan jantung. Temuan ini selaras dengan konsensus literatur medis kardiologi dan guideline klinis internasional, sehingga memvalidasi bahwa model telah belajar fitur-fitur yang secara medis relevan dan tidak hanya menggali pola statisitik noise dalam data.

Untuk memberikan penjelasan pada tingkat individu pasien, kami menggunakan nilai SHAP (Gambar 7) yang mengukur kontribusi marginal setiap fitur terhadap prediksi akhir untuk kasus spesifik. Berbeda dengan Feature Importance global yang menunjukkan pentingnya rata-rata fitur di seluruh dataset, SHAP values memberikan penjelasan lokal yang dapat disesuaikan untuk setiap pasien individual. Visualisasi SHAP Dependence Plot pada

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

Gambar 7 memetakan nilai-nilai input pasien (misalnya Usia, Tekanan Darah Istirahat, Kadar Kolesterol) terhadap dampak SHAP-nya (kontribusi terhadap prediksi risiko), memberikan pemahaman visual tentang bagaimana perubahan nilai fitur mempengaruhi estimasi risiko. Setiap titik dalam plot mewakili satu pasien, memungkinkan observasi hubungan non-linear antara fitur dan prediksi. Misalnya, jika plot menunjukkan bahwa nilai Cholesterol tinggi secara konsisten berkorelasi dengan nilai SHAP positif besar (meningkatkan prediksi risiko), ini memberikan interpretasi yang dapat dipertanggungjawabkan secara medis.



Gambar 7. Profil Risiko Pasien #2 (Visualisasi Lokal)

3.3. Implementasi Antarmuka Chatbot

Berbeda dengan penelitian sejenis yang hanya berhenti pada perhitungan metrik akurasi, penelitian ini mengintegrasikan model Machine Learning ke dalam arsitektur microservice berbasis FastAPI yang siap untuk produksi. Sistem ini dirancang untuk menjembatani kesenjangan kritis antara output numerik mentah (probabilitas prediksi) dan pemahaman pasien awam melalui lapisan Natural Language Processing (NLP) dan Generative AI. Inovasi utama terletak pada tidak hanya menyajikan hasil diagnostik, tetapi juga menjelaskannya secara transparan dan kontekstual menggunakan nilai SHAP dan LLM yang terintegrasi secara ketat dengan logika validasi medis.

Backend dibangun menggunakan framework FastAPI (Python 3.10) yang mendukung pemrosesan asinkron (async/await) untuk menangani permintaan inferensi tanpa memblokir main thread, memungkinkan sistem untuk menangani multiple concurrent requests dengan latensi rendah. Arsitektur ini terdiri dari tiga endpoint utama yang menangani siklus hidup diagnosis, seperti dirincikan pada Tabel 2.

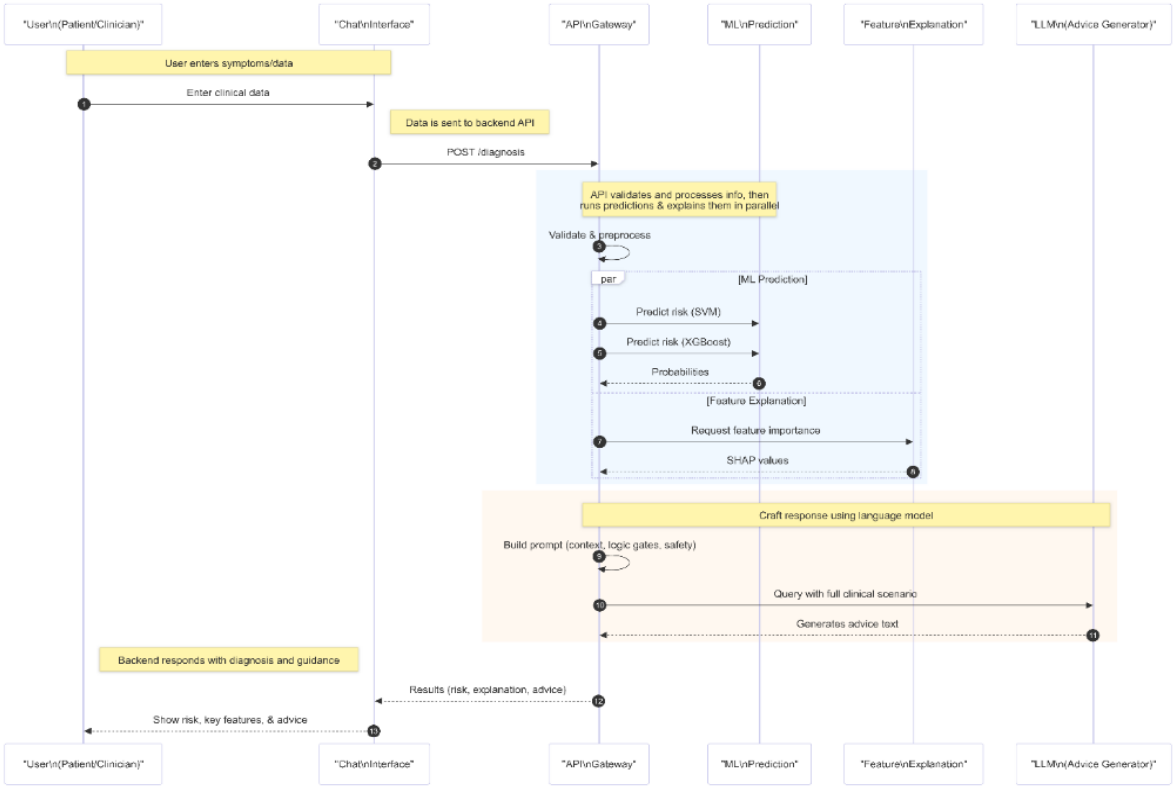
Tabel 2. Spesifikasi Teknis Endpoint API Sistem

Endpoint	Method	Fungsi Utama	Latency (avg)
/parse	POST	Ekstraksi entitas medis (NER) dari teks natural user	< 100ms
/diagnosis	POST	Orkestrasi ML Inference & Generasi Respons LLM	2.5s - 4.0s
/sessions	GET/POST	Manajemen riwayat percakapan & penyimpanan state	< 50ms

Proses dimulai ketika frontend mengirimkan payload data klinis ke endpoint /diagnosis. Server melakukan pra-pemrosesan (StandardScaler) dan inferensi paralel pada model SVM dan XGBoost yang telah diserialisasi (pickled), memungkinkan sistem untuk mendapatkan "second opinion" dari kedua algoritma secara bersamaan. Secara bersamaan dengan inferensi model, SHAP TreeExplainer menghitung nilai kontribusi marginal fitur lokal untuk instance tersebut, mengidentifikasi faktor-faktor yang paling mempengaruhi prediksi. Alur data lengkap dari input pengguna hingga respons sistem divisualisasikan pada diagram sekuensial di bawah ini, menunjukkan bagaimana setiap komponen sistem berinteraksi secara real-time.

Setiap tahapan dalam alur data tersebut dirancang dengan mempertimbangkan efisiensi komputasi dan keandalan sistem secara bersamaan. Pada tahap pra-pemrosesan, StandardScaler diterapkan secara konsisten menggunakan parameter yang telah dikalibrasi dari data training, sehingga distribusi fitur input pengguna selalu selaras dengan distribusi yang dipelajari model selama pelatihan. Inferensi paralel pada dua model secara simultan tidak hanya meningkatkan throughput sistem, tetapi juga memberikan mekanisme cross-validation implisit yang memperkuat kepercayaan terhadap hasil prediksi akhir. Apabila kedua model menunjukkan konsensus yang kuat, sistem akan memberikan respons dengan tingkat kepercayaan yang lebih tinggi, sedangkan perbedaan signifikan antara keduanya akan memicu mekanisme peringatan tambahan kepada pengguna. Desain ini mencerminkan filosofi utama penelitian bahwa transparansi dan kehati-hatian dalam pengambilan keputusan medis berbasis AI harus selalu diutamakan di atas kecepatan maupun kesederhanaan implementasi.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722



Gambar 8. Diagram Sekuensial Alur Data (User → API Gateway → ML Inference → SHAP Engine → LLM → Response)

Novelty utama dari sistem ini terletak pada mekanisme Dynamic Prompt Engineering yang memastikan validitas medis dan konsistensi output. Hasil prediksi numerik dan nilai SHAP tidak langsung disajikan kepada pengguna dalam bentuk mentah, melainkan diinjeksikan ke dalam context window LLM (Mistral-7B Instruct) melalui sebuah mekanisme orkestrasi yang ketat. Algoritma kami membangun system prompt secara real-time dengan struktur logika yang imperativ untuk memastikan bahwa setiap respons yang dihasilkan tetap valid dan aman secara medis, sebagaimana dijelaskan pada Tabel 3. Pendekatan ini membedakan sistem kami dari chatbot medis generik yang mungkin menghasilkan respons yang mengabaikan thresholds risiko atau memberikan saran yang tidak didukung oleh probabilitas model.

Tabel 3. Struktur Logika Prompt Engineering untuk Validasi Medis

Komponen Prompt	Deskripsi Logika & Instruksi Sistem
Context Injection	Menyisipkan probabilitas prediksi dari kedua model: "SVM Risk: 92%, XGB Risk: 89%, Mean Risk: 90.5%" dengan penjelasan bahwa rata-rata digunakan sebagai estimasi final
Explainability Layer	Mengkonversi top-3 fitur SHAP menjadi narasi medis yang mudah dipahami: "Faktor utama: Kolesterol (SHAP: +0.45), Jenis Nyeri Dada (SHAP: +0.38), Detak Jantung Maksimum (SHAP: -0.12)" dengan interpretasi klinis untuk setiap faktor
Risk Stratification	Safety rule imperatif berbasis threshold: "IF risk_score > 0.66 THEN risk_level = 'HIGH' AND recommendation = 'Immediate Cardiology Consultation'. IF 0.33 < risk_score ≤ 0.66 THEN risk_level = 'MODERATE' AND recommendation = 'Schedule Cardiology Appointment'. IF risk_score ≤ 0.33 THEN risk_level = 'LOW' AND recommendation = 'Maintain Healthy Lifestyle'."
Safety Guardrails	Instruksi disclaimer imperatif: "ALWAYS end response with: 'Ini adalah penilaian berbasis AI, bukan diagnosis medis. Konsultasikan hasil ini dengan dokter profesional untuk evaluasi lebih lanjut.' NEVER provide specific medications. NEVER claim 100% certainty. DO use conditional language like 'may indicate' atau 'suggests'."

Mekanisme Dynamic Prompt Engineering ini memungkinkan LLM untuk bertindak sebagai "penerjemah medis" yang cerdas namun terkontrol. Sebagai contoh, jika model mendeteksi risiko tinggi (probability > 0.66) karena faktor usia, kolesterol, dan jenis nyeri dada tertentu, LLM akan menyusun respons dengan struktur: "Berdasarkan analisis Anda, sistem mengestimasi risiko penyakit jantung sebesar 88% (SVM: 90%, XGBoost: 86%). Beberapa faktor berkontribusi, pada hasil ini: (1) Kadar kolesterol Anda melebihi batas normal untuk kelompok usia Anda, meningkatkan risiko; (2) Pola nyeri dada yang Anda laporkan selaras dengan presentasi klinis yang kami pelajari; (3) Detak jantung maksimum yang relatif rendah sedikit menurunkan perkiraan risiko. Mengingat estimasi risiko yang

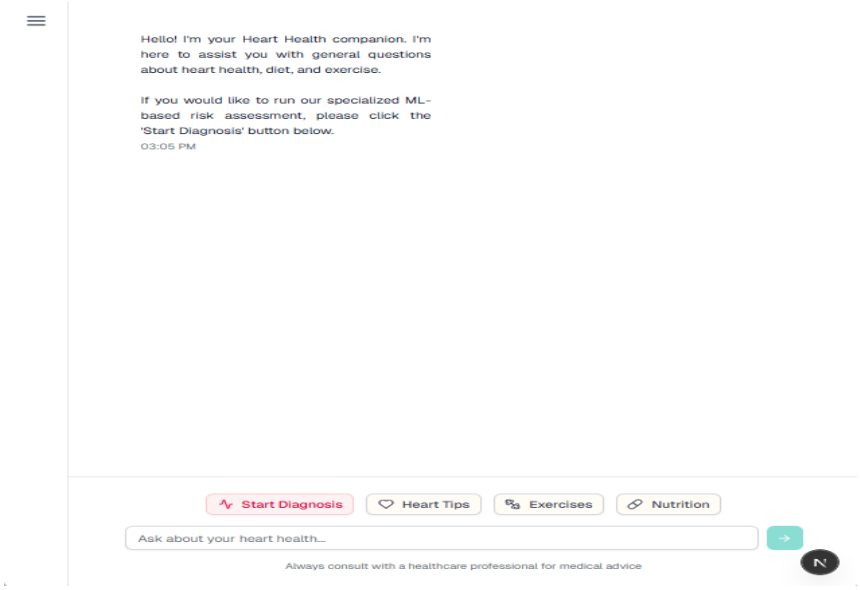
Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

tinggi, kami sangat merekomendasikan Anda untuk berkonsultasi dengan dokter spesialis jantung dalam waktu dekat. Ini adalah penilaian berbasis AI, bukan diagnosis medis definitif." Pendekatan terstruktur ini memastikan bahwa respons selalu mematuhi protokol keselamatan medis tanpa menghilangkan nuansa dan empati dalam komunikasi pasien. Format respons dari API dikirimkan dalam bentuk JSON terstruktur yang memisahkan teks percakapan (narasi dari LLM), skor probabilitas mentah dari kedua model, dan faktor-faktor penentu utama (SHAP factors dengan nilai kontribusi masing-masing) untuk keperluan debugging, visualisasi frontend, atau audit trail medis, seperti terlihat pada contoh struktur data di Gambar 10. Pemisahan logis ini memungkinkan frontend untuk me-render berbagai komponen secara independen: narasi teks untuk pengguna, grafik probabilitas untuk visualisasi risiko, dan daftar SHAP factors untuk transparansi algoritmik.


```
{
  "status": "success",
  "data": {
    "diagnosis_id": "DX-20240121-001",
    "timestamp": "2024-01-21T14:30:00Z",
    "clinical_metrics": {
      "svm_probability": 0.88,
      "xgboost_probability": 0.91,
      "risk_category": "ALARM_HIGH"
    },
    "explainability": {
      "shap_factors": [
        "Cholesterol (+0.45)",
        "Max Heart Rate (-0.22)",
        "Chest Pain Type (Non-Anginal)"
      ]
    },
    "llm_response": {
      "text": "Berdasarkan analisis, Anda memiliki risiko tinggi...",
      "safety_flags": {
        "disclaimer_present": true,
        "emergency_keywords": false
      }
    }
  }
}
```


Gambar 9. Diagram Sekuensial Alur Data (User → API Gateway → ML Inference → SHAP Engine → LLM → Response)

Mengingat sensitivitas domain kesehatan dan kebutuhan akan kepercayaan pengguna, sistem dilengkapi dengan validasi keamanan lapis ganda (dual-layer safety validation). Pertama, logika hard-coded pada Python memastikan bahwa model output dipetakan ke kategori risiko yang transparan (Low/Moderate/High) dan tidak ada "halusinasi" LLM yang dapat mengklaim pasien sehat ketika probabilitas rata-rata kedua model menunjukkan risiko moderat atau tinggi. Jika output LLM bertentangan dengan threshold risiko yang telah ditetapkan, respons tersebut akan ditolak dan diganti dengan template yang lebih aman. Kedua, parameter temperature pada LLM diatur rendah (0.3) untuk mengurangi variabilitas dan kreativitas jawaban yang tidak perlu, memastikan konsistensi saran medis di seluruh sesi pengguna berbeda. Kombinasi kedua layer ini menghasilkan sistem yang menggabungkan kecanggihan prediksi statistik dari SVM/XGBoost dengan komunikasi yang empatik namun terkontrol dari LLM, memberikan pengalaman pengguna yang aman dan dapat dipercaya (trustworthy) dalam konteks kesehatan.
Antarmuka awal chatbot yang ditampilkan kepada pengguna ditunjukkan pada Gambar 10, memperlihatkan tampilan pembuka dan instruksi awal sistem.

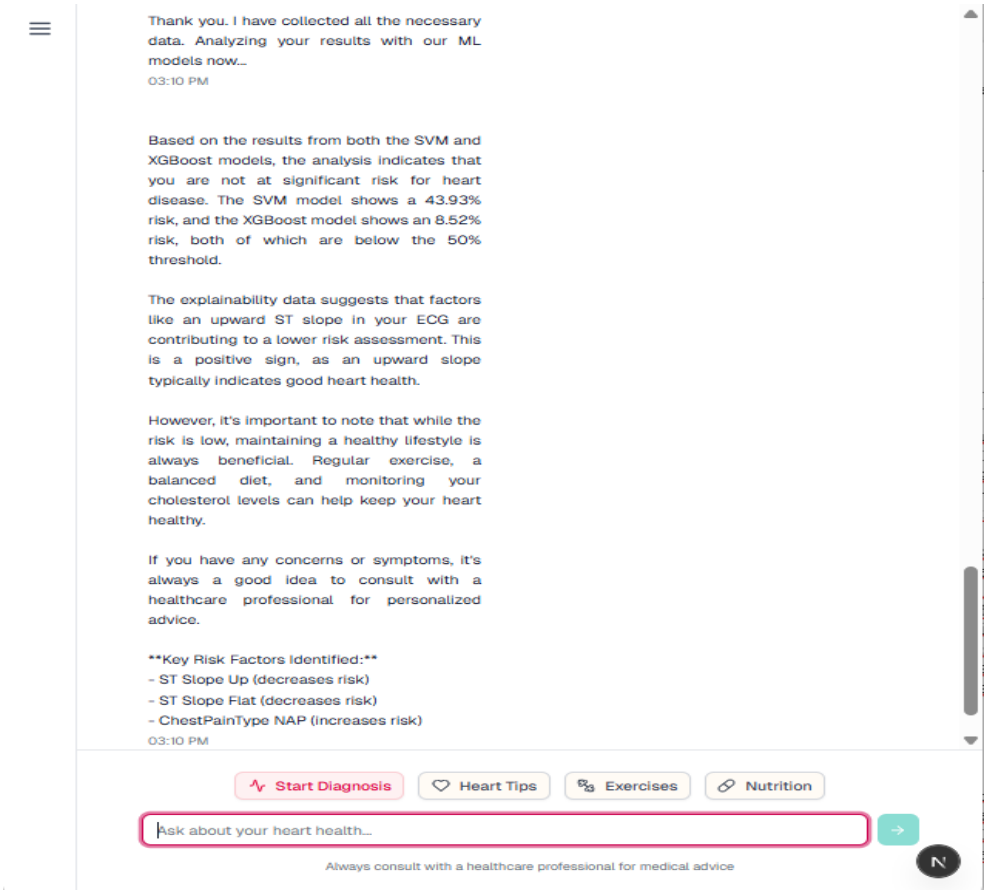


Gambar 10. Antarmuka Awal Chatbot "Heart Health Companion"

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

Antarmuka akhir yang diterima pengguna (Gambar 10 & 11) menggabungkan kecanggihan prediksi SVM/XGBoost dengan empati komunikasi manusia, memberikan pengalaman yang intuitif dan dapat dipercaya. Sistem menampilkan dialog interaktif yang mengumpulkan data klinis pengguna secara bertahap , memberikan respons risiko berdasarkan analisis model dengan transparansi penuh terhadap faktor-faktor yang mempengaruhi prediksi, dan menyertakan saran medis yang aman dan kontekstual, merealisasikan visi sistem triage berbasis AI yang dapat diakses oleh masyarakat umum.

Gambar 11 menampilkan contoh antarmuka percakapan aktif chatbot, termasuk respons naratif yang dihasilkan sistem berdasarkan data klinis pengguna dan hasil analisis SHAP.



Gambar 11. Antarmuka Awal Chatbot "Heart Health Companion"

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem chatbot diagnostik penyakit jantung berbasis hybrid AI yang mengintegrasikan machine learning, explainability, dan natural language generation dengan protokol keselamatan medis. Model SVM dengan kernel RBF menunjukkan performa superior dengan akurasi 90.22% dan recall 94.12%, menjadikannya model utama dalam arsitektur dual-model, sementara XGBoost (88.04%) dipertahankan sebagai second-opinion. Integrasi SHAP mengidentifikasi Chest Pain Type, Cholesterol, dan MaxHR sebagai fitur kontributor utama, selaras dengan guideline kardiologi internasional dan memvalidasi bahwa model belajar pola yang secara klinis bermakna. Novelty utama penelitian adalah mekanisme Dynamic Prompt Engineering yang menerjemahkan output probabilistik ML menjadi narasi medis yang aman dan empatik melalui LLM Mistral-7B, diperkuat dual-layer safety validation untuk mencegah hallucination diagnostik. Sistem yang diimplementasikan sebagai FastAPI microservice terbukti fungsional sebagai proof-of-concept dengan latency 2.5–4.0 detik per kueri, meskipun belum memenuhi standar deployment klinis nyata karena belum adanya validasi prospektif dengan pasien nyata. Untuk penelitian lanjutan, direkomendasikan: (1) user study formal melibatkan tenaga medis dan pasien, (2) evaluasi halusinasi LLM secara sistematis, (3) validasi klinis prospektif dengan ethical clearance, dan (4) integrasi dengan rekam medis elektronik (EHR).

REFERENCES

[1] S. Simatupang, R. Ramadhansyah, R. Tumanggor, E. P. Tan, and S. A. Fajar, “Prediction of Heart Disease Risk Based on Patient Health History Using the Support Vector Machine (SVM) Algorithm,” *ZERO: Jurnal Sains, Matematika dan Terapan*, vol. 9, no. 2, p. 612, Nov. 2025, doi: 10.30829/zero.v9i2.26087.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 14−26
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.9722

[2] S. A. Bangun, E. S. Ompusunggu, W. Wilson, and E. K. Harefa, "Support Vector Machine for Classifying Heart Failure, Hypertension, and Normal Heart Condition," *JUSIFO (Jurnal Sistem Informasi)*, vol. 11, no. 1, pp. 53–60, Jun. 2025, doi: 10.19109/jusifo.v11i1.28113.
[3] T. Roopa and G. D. Ramanjinappa, "Heart Disease Predictive Modeling with XGBoost and SMOTE-Driven Class Imbalance Mitigation," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 29914–29918, Dec. 2025, doi: 10.48084/etasr.14301.
[4] D. Rohmayani, C. A. Sugianto, R. S. Perdana, and M. M. Nafea, "Improving Extreme Gradient Boosting Model for Heart Disease Prediction Using SMOTE for Class Imbalance," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 4, pp. 1717–1728, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4753.
[5] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," *J. Healthc. Eng.*, vol. 2022, pp. 1–9, 2022, doi: 10.1155/2022/7351061.
[6] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Nov. 2017, pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
[7] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Sci. Rep.*, vol. 15, no. 1, p. 13912, Apr. 2025, doi: 10.1038/s41598-025-97547-6.
[8] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.
[9] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.
[10] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.
[11] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv preprint*, arXiv:2310.06825, Oct. 2023, doi: 10.48550/arXiv.2310.06825.
[12] B. A. Majeed, A. Y. Hardan, B. Y. Hardan, and D. F. Munaf, "Accurate AI-Based Chatbot to Diagnose Heart Diseases Pre-Human Doctor Consultation," *Revue d'Intelligence Artificielle*, vol. 38, no. 1, pp. 213–220, Feb. 2024, doi: 10.18280/ria.380121.
[13] S. E. Antia *et al.*, "Healthy Heart Assistant, a WhatsApp-Based Generative Pretrained Transformer Technology, for Self-Care in Hypertensive Patients," *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 3, p. 100243, Sep. 2025, doi: 10.1016/j.mcpdig.2025.100243.
[14] A. Nurlita and M. Munawaroh, "Pengembangan Chatbot Dengan Metode Natural Language Processing Untuk Layanan Pelanggan (Studi Kasus PT Masterlink Internet Solution)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3S1, Oct. 2025, doi: 10.23960/jitet.v13i3S1.8176.
[15] S. Boit and R. Patil, "A Prompt Engineering Framework for Large Language Model–Based Mental Health Chatbots: Conceptual Framework," *JMIR Ment. Health*, vol. 12, pp. e75078–e75078, Nov. 2025, doi: 10.2196/75078.
[16] B. Meskó, "Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial," *J. Med. Internet Res.*, vol. 25, p. e50638, Oct. 2023, doi: 10.2196/50638.
[17] M. Muhetaer, A. Yusupu, W. Yifan, M. Mutalipu, and F. Hao, "Medical QA dialogue datasets in RAG systems performance evaluation and ChatGPT optimization," *Sci. Rep.*, vol. 15, no. 1, p. 44467, Dec. 2025, doi: 10.1038/s41598-025-28015-4.
[18] O. T. Odofin, B. I. Adekunle, E. Ogbuefi, J. C. Ogeawuchi, O. S. Adanigbo, and T. P. Gbenle, "Improving Healthcare Data Intelligence through Custom NLP Pipelines and Fast API Micro services," *Journal of Frontiers in Multidisciplinary Research*, vol. 4, no. 1, pp. 390–397, 2023, doi: 10.54660/.JFMR.2023.4.1.390-397.
[19] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information*, vol. 13, no. 10, p. 475, Oct. 2022, doi: 10.3390/info13100475.
[20] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.