JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 268~273
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608

Perbandingan Performa Algoritma XGBoost, CatBoost Dan GBM Dalam Prediksi Penyakit Kardiovaskular

[1]Panwasto Samosir P, [2]Umniy Salamah
[1,2] Fakultas Ilmu Komputer, Universitas Mercu Buana, Indonesia
[1]41521010178@mercubuana.ac.id; [2]umniy.salamah@mercubuana.ac.id;

Article Info

Article history:

Received, 2025-01-07
Revised, 2025-01-25
Accepted, 2025-01-31

Kata Kunci:

kardiovaskular
xgboost
catboost
gradient boosting

ABSTRAK

Penyakit kardiovaskular adalah penyebab utama kematian secara global, yang mencakup gangguan pada jantung dan pembuluh darah, seperti hipertensi dan penyakit jantung koroner. Faktor risiko penyakit ini meliputi kebiasaan hidup yang tidak sehat serta faktor yang tidak dapat diubah, seperti usia dan riwayat keluarga. Untuk mengatasi tantangan dalam deteksi dini dan prediksi penyakit kardiovaskular, pendekatan machine learning, khususnya algoritma boosting, telah menunjukkan potensi yang signifikan. Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma boosting utama, yaitu XGBoost, CatBoost, dan Gradient Boosting, dalam memprediksi risiko penyakit kardiovaskular menggunakan dataset yang tersedia secara online. Hasil penelitian menunjukkan bahwa CatBoost memiliki performa terbaik dengan akurasi sebesar 75%, Precision 0.83, dan ROC AUC 0.81, yang mengindikasikan kemampuannya dalam menghasilkan prediksi yang lebih akurat. Gradient Boosting memiliki akurasi 70% dan menunjukkan keseimbangan yang baik antara Recall dan Precision, sementara XGBoost memiliki kinerja terendah dengan akurasi 63.3% di semua metrik evaluasi. Berdasarkan hasil ini, CatBoost adalah model yang paling efektif untuk memprediksi risiko penyakit kardiovaskular.

Keywords:

cardiovascular
xgboost
catboost
gradient boosting

ABSTRACT

Cardiovascular disease remains the primary cause of mortality globally, encompassing conditions affecting the heart and blood vessels, such as hypertension and coronary artery disease. Risk factors include unhealthy lifestyle habits and immutable factors like age and family history. To tackle the challenges in early detection and prediction of cardiovascular disease, machine learning techniques, especially boosting algorithms, have emerged as promising tools. This study evaluates the performance of three prominent boosting algorithms: XGBoost, CatBoost, and Gradient Boosting—using publicly available datasets to predict cardiovascular disease risk. The findings reveal that CatBoost surpasses the other models with an accuracy of 75%, a Precision of 0.83, and a ROC AUC of 0.81, highlighting its exceptional predictive capabilities. Gradient Boosting achieves 70% accuracy with a well-balanced Recall and Precision, whereas XGBoost records the lowest performance with 63.3% accuracy across all metrics. These results position CatBoost as the most effective model for cardiovascular disease risk prediction.

Penulis Korespondensi:

Panwasto Samosir P,
Program Studi Teknik Informatika,
Universitas Mercu Buana,
Email: 41521010178@mercubuana.ac.id

1. PENDAHULUAN

Penyakit kardiovaskular mencakup berbagai gangguan yang memengaruhi jantung dan pembuluh darah, seperti penyakit jantung koroner, hipertensi, penyakit arteri perifer, dan gagal jantung. Kondisi ini dipicu oleh

sejumlah faktor risiko yang terbagi menjadi dua kategori: faktor yang dapat dimodifikasi dan faktor yang tidak dapat dimodifikasi. Faktor risiko yang dapat dimodifikasi meliputi kebiasaan hidup tidak sehat, seperti merokok, konsumsi alkohol berlebihan, serta pola makan tinggi lemak dan kolesterol. Di sisi lain, faktor risiko yang tidak dapat dimodifikasi mencakup usia, jenis kelamin, dan riwayat keluarga [1][2][3].

Untuk mengarahkan penelitian kesehatan dan memperkuat upaya pencegahan serta pengobatan yang lebih efektif terhadap penyakit kardiovaskular, Amerika Serikat mendirikan National Heart, Lung, and Blood Institute (NHLBI) pada tahun 1948. NHLBI bertujuan untuk memajukan penelitian dalam bidang kesehatan jantung, paru-paru, dan darah. Salah satu inisiatif penting yang dilakukan oleh NHLBI adalah pendirian Cardiovascular Health Study (CHS), yang dimulai pada tahun 1988. Studi ini fokus pada populasi dewasa di atas usia 65 tahun dan bertujuan untuk mengidentifikasi faktor risiko penyakit jantung koroner serta faktor-faktor yang berkontribusi terhadap perkembangan penyakit kardiovaskular [4], [5]. Inisiatif ini telah menjadi tonggak penting dalam memahami penyakit kardiovaskular dan memberikan landasan bagi penelitian lebih lanjut serta upaya pencegahan yang lebih efektif [6].

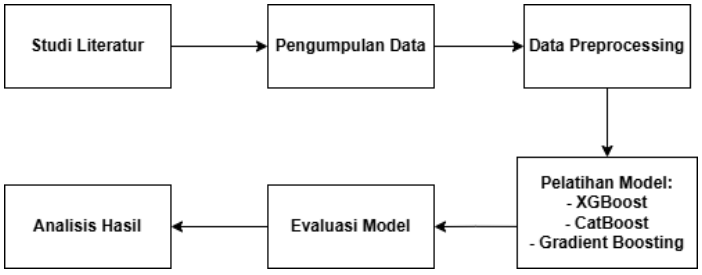
Berdasarkan data dari World Health Organization (WHO), penyakit kardiovaskular menjadi penyebab utama kematian di seluruh dunia, dengan lebih dari 17,9 juta kematian setiap tahun. Angka ini mencakup sekitar 31% dari total kematian global, di mana 85% di antaranya disebabkan oleh serangan jantung dan stroke [7][8][9]. Meskipun ada kemajuan signifikan dalam deteksi dan pengobatan penyakit kardiovaskular, banyak kasus masih terdiagnosis terlambat atau tidak terdeteksi sama sekali. Hal ini sering kali disebabkan oleh ketidakmampuan dalam mengidentifikasi faktor risiko individu secara tepat waktu dan efektif [10], [11].

Dalam konteks ini, perlu adanya pendekatan yang lebih proaktif dan akurat dalam melakukan prediksi risiko penyakit kardiovaskular. Pendekatan machine learning telah menjadi sorotan dalam upaya mengatasi tantangan deteksi dini dan prediksi penyakit kardiovaskular. Machine learning menawarkan potensi untuk menganalisis data medis dalam jumlah besar dan kompleks, mengidentifikasi pola-pola yang mungkin tidak terlihat oleh metode konvensional, dan memprediksi risiko penyakit secara lebih akurat [12], [13]. Meskipun terdapat kemajuan dalam penerapan machine learning untuk prediksi penyakit kardiovaskular, permasalahan muncul dalam menentukan algoritma yang paling efektif. Teknik boosting dalam machine learning, yang melibatkan penggabungan beberapa model sederhana untuk membentuk model yang lebih kuat, telah menunjukkan hasil yang menjanjikan dalam berbagai aplikasi, termasuk dalam prediksi penyakit. Namun, belum ada konsensus mengenai algoritma boosting mana yang paling efektif untuk prediksi penyakit kardiovaskular.

Karena itu, penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma boosting dalam prediksi penyakit kardiovaskular, yaitu XGBoost, CatBoost, dan Gradient Boosting Machine (GBM) [14]. XGBoost (Extreme Gradient Boosting) dikenal dengan efisiensinya dan kemampuan menangani data yang hilang [15], [16], [17]. CatBoost (Categorical Boosting) dirancang untuk menangani data kategorikal tanpa perlu banyak pra-pemrosesan. Sementara itu, GBM (Gradient Boosting Machine) adalah salah satu algoritma boosting yang paling awal dan banyak digunakan, terkenal dengan kemampuannya dalam menangani data yang bervariasi [18]. Ketiga algoritma ini akan diuji menggunakan dataset yang tersedia secara online dan terbuka (open-source). Dengan membandingkan kinerja ketiga algoritma ini pada dataset yang sama, diharapkan dapat ditemukan algoritma yang paling sesuai untuk prediksi penyakit kardiovaskular. Hasil dari penelitian ini diharapkan mampu memberikan kontribusi yang signifikan dalam upaya deteksi dini dan pencegahan penyakit kardiovaskular, sehingga dapat mengurangi angka kematian dan meningkatkan kualitas hidup pasien.

2. METODE PENELITIAN

Penelitian ini dilakukan dengan mengikuti tahapan yang telah dirancang oleh peneliti. Proses penelitian tersebut digambarkan melalui diagram alir berikut:



Gambar 1 Tahapan Penelitian

Penelitian ini dilakukan secara sistematis melalui beberapa tahapan, dimulai dari pengumpulan data hingga analisis hasil. Data yang digunakan diperoleh dari platform Kaggle dengan nama Heart Failure Prediction Dataset https://www.kaggle.com/andrewmvd/heartfailure-clinical-data, yang terdiri atas 299 sample dan 12 atribut, yaitu usia, anemia, kadar creatinine phosphokinase, diabetes, fraksi ejeksi, tekanan darah tinggi, jumlah trombosit, kadar serum creatinine, kadar serum sodium, jenis kelamin, status merokok, dan status kematian. Tahapan awal melibatkan data preprocessing untuk memastikan kualitas data yang optimal sebelum analisis lebih lanjut. Tahapan ini meliputi pengelolaan nilai yang hilang, normalisasi variabel numerik, serta pemisahan data menjadi data latih dan data uji dengan rasio 80:20.

Tabel 1 Data Frame

age	anaemia	creatinine phosphokinase	diabetes	ejection fraction	high blood pressure	platelets	serum creatinine	serum sodium	sex	smoking	death event
75	0	582	0	20	1	265000	1.9	130	1	0	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	1
65	0	146	0	20	0	162000	1.3	129	1	1	1
50	1	111	0	20	0	210000	1.9	137	1	0	1
65	1	160	1	20	0	327000	2.7	116	0	0	1

Selanjutnya, model klasifikasi dikembangkan menggunakan tiga algoritma, yaitu XGBoost, CatBoost, dan Gradient Boosting. Ketiga algoritma ini digunakan untuk membangun model prediksi risiko penyakit kardiovaskular. Evaluasi model dilakukan dengan menggunakan metrik performa seperti akurasi, precision, recall, dan F1-score untuk mengidentifikasi algoritma dengan kinerja terbaik. Tahapan akhir berupa analisis hasil dilakukan untuk menggali wawasan dari evaluasi model, sehingga dapat diidentifikasi pola-pola penting dalam data yang berkontribusi terhadap risiko penyakit kardiovaskular. Analisis ini juga bertujuan untuk memvalidasi efektivitas pendekatan penelitian yang diterapkan.

3. HASIL DAN ANALISIS

Tahapan dalam penelitian ini mencakup beberapa langkah utama, yaitu data preprocessing, visualisasi data, pelatihan model, dan evaluasi hasil menggunakan confusion matrix untuk menentukan model yang paling optimal. Langkah pertama adalah mengimpor pustaka yang dibutuhkan. Penelitian ini memanfaatkan Python dengan pustaka seperti sklearn, numpy, pandas, matplotlib, seaborn dan plotly. Dataset dalam format CSV dimuat menggunakan pustaka pandas.

Setelah data dimuat, langkah berikutnya adalah membangun feature matrix X sebagai variabel independen dan target vector Y sebagai variabel dependen. Tahap data preprocessing dilakukan dengan mengidentifikasi dan

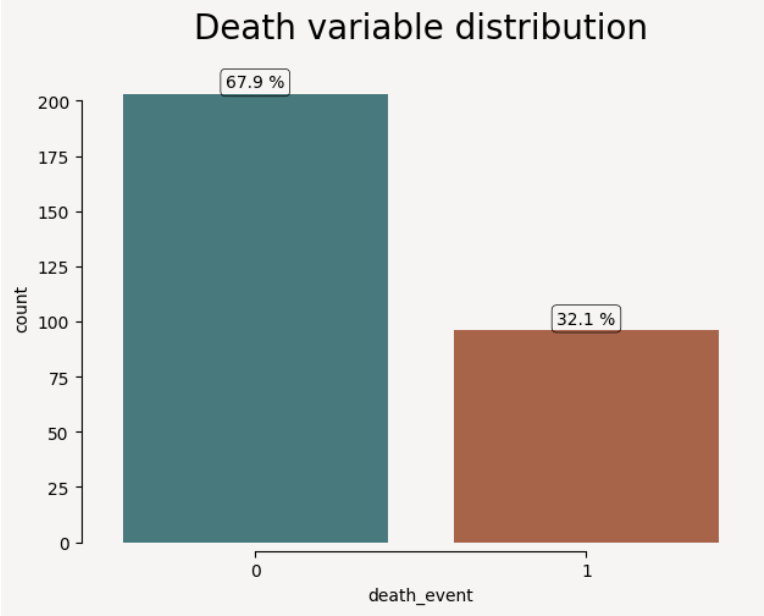
	count	mean	std	min	25%	50%	75%	max
age	299.0	60.833893	11.894809	40.0	51.0	60.0	70.0	95.0
anaemia	299.0	0.431438	0.496107	0.0	0.0	0.0	1.0	1.0
creatinine_phosphokinase	299.0	581.839465	970.287881	23.0	116.5	250.0	582.0	7861.0
diabetes	299.0	0.418060	0.494067	0.0	0.0	0.0	1.0	1.0
ejection_fraction	299.0	38.083612	11.834841	14.0	30.0	38.0	45.0	80.0
high_blood_pressure	299.0	0.351171	0.478136	0.0	0.0	0.0	1.0	1.0
platelets	299.0	263358.029264	97804.236869	25100.0	212500.0	262000.0	303500.0	850000.0
serum_creatinine	299.0	1.393880	1.034510	0.5	0.9	1.1	1.4	9.4
serum_sodium	299.0	136.625418	4.412477	113.0	134.0	137.0	140.0	148.0
sex	299.0	0.648829	0.478136	0.0	0.0	1.0	1.0	1.0
smoking	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0
death_event	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0

Gambar 2 Statistik Data

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 268~273
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608

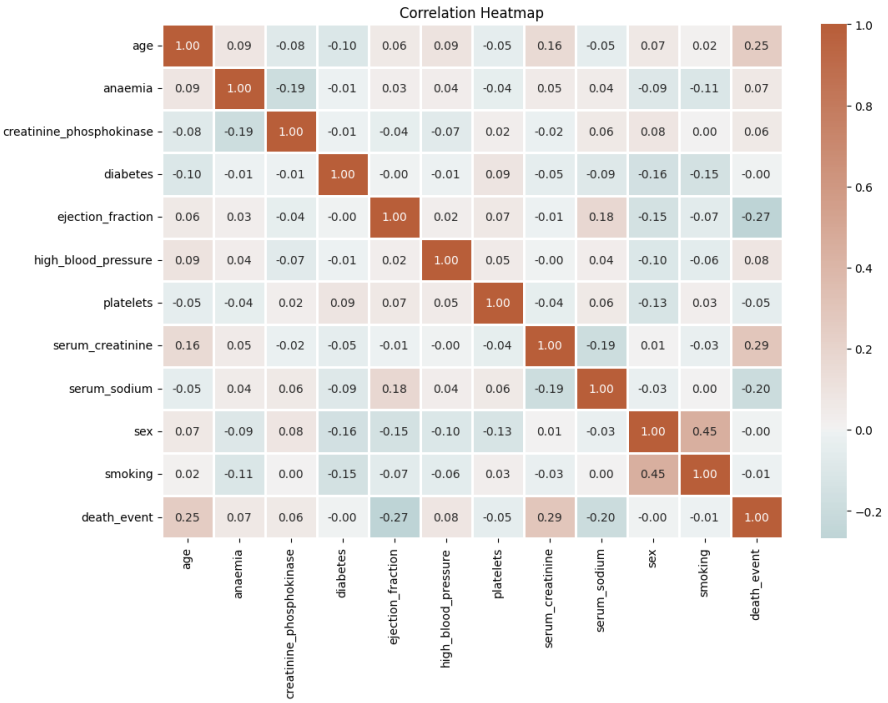
menangani nilai yang hilang untuk mencegah gangguan selama proses pelatihan. Statistik deskriptif, seperti nilai percentile, mean, std dan lainnya dari data frame.

Analisis visualisasi data pada tahap ini bertujuan untuk memahami distribusi variabel dan menggali pola yang berhubungan dengan faktor risiko penyakit jantung. Visualisasi pada gambar 3 menunjukkan bahwa 67,9% sampel atau sebanyak 203 pasien pada variabel *death_event* merupakan pasien yang tidak meninggal, sementara 32,1% atau sebanyak 96 pasien meninggal.

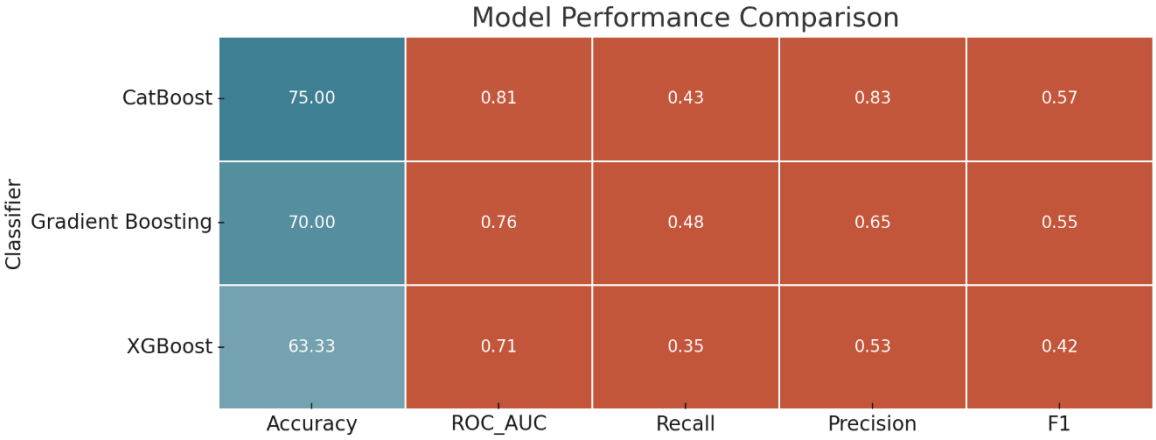


Gambar 3 Distribusi Variabel Status Kematian

Visualisasi heatmap pada gambar 4 menunjukkan korelasi antar variabel dalam dataset. Variabel *death_event* memiliki korelasi negatif dengan *age* (-0,25), *ejection_fraction* (-0,27), dan *serum_creatinine* (-0,20), serta korelasi positif dengan *smoking* (0,45). Hubungan ini menunjukkan bahwa faktor-faktor seperti usia, *ejection fraction*, dan merokok berpengaruh terhadap risiko kematian akibat penyakit jantung.



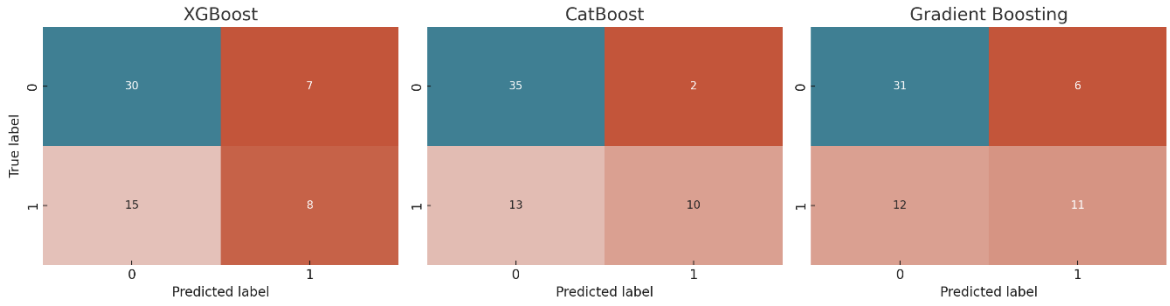
JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 268~273
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.3608



Gambar 4 Nilai Akurasi Model

Evaluasi performa model klasifikasi pada eksperimen ini dilakukan dengan membandingkan tiga algoritma pembelajaran mesin, yaitu XGBoost, CatBoost, dan Gradient Boosting. Visualisasi pada gambar 5 menunjukkan perbandingan kinerja tiga algoritma machine learning: CatBoost, Gradient Boosting, dan XGBoost, berdasarkan lima metrik evaluasi, yaitu Accuracy, ROC AUC, Recall, Precision, dan F1 Score. CatBoost memiliki performa terbaik dengan Accuracy tertinggi (75%), Precision tinggi (0.83), dan ROC AUC sebesar 0.81, meskipun Recall-nya relatif rendah (0.43), yang menunjukkan kelemahan dalam mendeteksi semua kasus positif. Gradient Boosting menyeimbangkan Recall (0.48) dan Precision (0.65), menghasilkan F1 Score yang cukup baik (0.55), namun dengan Accuracy yang sedikit lebih rendah (70%). Di sisi lain, XGBoost menunjukkan performa paling rendah pada semua metrik, dengan Accuracy 63.33%, ROC AUC 0.71, dan F1 Score 0.42, menandakan bahwa model ini kurang efektif dalam klasifikasi dibandingkan dengan dua model lainnya.

Confusion matrix pada gambar 6 menunjukkan bahwa CatBoost memiliki performa terbaik dengan jumlah kesalahan prediksi terendah, yaitu 2 False Positive dan 13 False Negative, serta 35 True Negative dan 10 True Positive. Gradient Boosting memiliki keseimbangan yang cukup baik dengan 6 False Positive dan 12 False Negative. XGBoost menunjukkan performa terlemah dengan jumlah kesalahan prediksi yang lebih tinggi, yaitu 7 False Positive dan 15 False Negative, menunjukkan bahwa CatBoost adalah model paling akurat dalam klasifikasi ini.



Gambar 5 Confusion Matrix

4. KESIMPULAN

Hasil dari eksperimen berdasarkan dataset Kaggle (Heart Failure Prediction) menunjukkan bahwa model klasifikasi CatBoost memiliki kinerja terbaik dengan skor tertinggi pada metrik akurasi sebesar 75% yang menandakan kemampuan prediksi yang sangat baik. Gradient Boosting memberikan hasil yang cukup seimbang antara Recall dan Precision, sehingga mampu mencapai performa yang konsisten di beberapa metrik. Sementara itu, model XGBoost memiliki kinerja terendah dengan tingkat kesalahan lebih tinggi, terutama dalam mendeteksi kelas positif, dibandingkan kedua model lainnya.

UCAPAN TERIMA KASIH

Terima kasih kepada Program Studi Teknik Informatika, Universitas Mercu Buana yang telah mendukung penelitian ini.

REFERENSI

[1] S. M. Ganie, P. K. D. Pramanik, M. B. Malik, A. Nayyar, and K. S. Kwak, "An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms," *Computer Systems Science and Engineering*, vol. 46, no. 3, pp. 3993–4006, 2023, doi: 10.32604/csse.2023.035244.

[2] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.

[3] K. V. Tompra, G. Papageorgiou, and C. Tjortjis, "Strategic Machine Learning Optimization for Cardiovascular Disease Prediction and High-Risk Patient Identification," *Algorithms*, vol. 17, no. 5, May 2024, doi: 10.3390/a17050178.

[4] H. H. Alalawi and M. S. Alsuwat, "Detection of Cardiovascular Disease using Machine Learning Classification Models." [Online]. Available: www.ijert.org

[5] T. Shen, D. Liu, Z. Lin, C. Ren, W. Zhao, and W. Gao, "A Machine Learning Model to Predict Cardiovascular Events during Exercise Evaluation in Patients with Coronary Heart Disease," *J Clin Med*, vol. 11, no. 20, Oct. 2022, doi: 10.3390/jcm11206061.

[6] C. Author *et al.*, "Brain Tumor Detection and Classification Using Fine-Tuned CNN with ResNet50 and EfficientNet," *International Journal of Informatics and Computation (IJICOM)*, vol. 6, no. 1, 2024, doi: 10.35842/ijicom.

[7] Z. Hu, H. Qiu, Z. Su, M. Shen, and Z. Chen, "A Stacking Ensemble Model to Predict Daily Number of Hospital Admissions for Cardiovascular Diseases," *IEEE Access*, vol. 8, pp. 138719–138729, 2020, doi: 10.1109/ACCESS.2020.3012143.

[8] Y. Jiang *et al.*, "Cardiovascular disease prediction by machine learning algorithms based on cytokines in kazakhs of china," *Clin Epidemiol*, vol. 13, pp. 417–428, 2021, doi: 10.2147/CLEP.S313343.

[9] J. Yang and J. Guan, "A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm," *Information (Switzerland)*, vol. 13, no. 10, Oct. 2022, doi: 10.3390/info13100475.

[10] Y. Jiang *et al.*, "Cardiovascular disease prediction by machine learning algorithms based on cytokines in kazakhs of china," *Clin Epidemiol*, vol. 13, pp. 417–428, 2021, doi: 10.2147/CLEP.S313343.

[11] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[12] J. O. Kim, Y. S. Jeong, J. H. Kim, J. W. Lee, D. Park, and H. S. Kim, "Machine learning-based cardiovascular disease prediction model: A cohort study on the korean national health insurance service health screening database," *Diagnostics*, vol. 11, no. 6, Jun. 2021, doi: 10.3390/diagnostics11060943.

[13] Y. Amelia, "PERBANDINGAN METODE MACHINE LEARNING UNTUK MENDETEKSI PENYAKIT JANTUNG," 2023. [Online]. Available: http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

[14] W. Sardjono, A. Retnowardhani, R. Emil Kaburuan, and A. Rahmasari, "Artificial intelligence and big data analysis implementation in electronic medical records," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Dec. 2021, pp. 231–237. doi: 10.1145/3512576.3512618.

[15] M. Rizky Mubarok, R. Herteno, I. Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lambung Mangkurat Jalan Ahmad Yani Km, and K. Selatan, "HYPER-PARAMETER TUNING PADA XGBOOST UNTUK PREDIKSI KEBERLANGSUNGAN HIDUP PASIEN GAGAL JANTUNG."

[16] M. Ravly Andryan *et al.*, "KOMPARASI KINERJA ALGORITMA XGBOOST DAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK DIAGNOSA PENYAKIT KANKER PAYUDARA," *Jurnal Informatika dan Komputer)*, vol. 6, no. 1, pp. 1–5, 2022.

[17] G. Abdurrahman, H. Oktavianto, and M. Sintawati, "Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada Klasifikasi Penyakit Diabetes," 2022.

[18] F. V. Ongkosianbhadra and C. C. Lestari, "Pengembangan Model Prediksi Risiko Hipertensi Menggunakan Algoritma Gradient Boosting Decision Tree Yang Dioptimalkan," *Jurnal Informatika dan Sistem Informasi*, vol. 9, no. 2, pp. 35–44, Dec. 2023, doi: 10.37715/juisi.v9i2.4403