ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

Predicting Heart Disease Using the XgBoost Algorithm and RandomizedSearch Optimizer

Prediksi Penyakit Jantung dengan Menggunakan Algoritma XgBoost dan RandomizedSearch Optimizer

Reo Sahobby[1], Dessyanto Boedi P[2], Mangaras Yanu F[3]

[1,2,3] Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

[1*]123170067@student.upnyk.ac.id, [2]dess@upnyk.ac.id, [3]mangaras.yanu@upnyk.ac.id
**: Penulis korenspondensi* (*corresponding author*)

Informasi Artikel
Received: November 2021
Revised: -
Accepted: -
Published: -

Abstract

Purpose: The purpose of this study was to identify heart disease based on tabular data or table data containing parameters from cardiac record data, and also to implements the XgBoost algorithm to predicting heart disease while reducing overfitting
Design/methodology/approach: Implements the XgBoost algorithm to build machine learning models and the RandomizedSearch Optimizer, then calculate the model's accuracy performance in prediciting heart disease
Findings/result: Machine learning models created with the XgBoost algorithm get 91% accuracy on training data and 83% on testing data. Tests carried out using other datasets get 78% accuracy, and the general model gets 90% accuracy.
Originality/value/state of the art: This research was conducted using the XgBoost algorithm combined with the RandomizedSearch Optimizer as a hyper parameter tuning for machine learning model making.

Keywords:machine learning; classification; heart diseases
Kata kunci: pembelajaran mesin; klasifikasi; penyakit jantung

Abstrak

Tujuan: Tujuan penelitian ini adalah untuk mengidentifikasi penyakit jantung berdasarkan data tabular atau data tabel yang berisi parameter dari data rekam jantung, dan juga dilakukan untuk menerapkan algoritma XgBoost untuk mengidentifikasi penyakit jantung sekaligus mengurangi *overfitting*.
Perancangan/metode/pendekatan: Menerapkan algoritma XgBoost untuk pembuatan model *machine learning* dan RandomizedSearch Optimizer, kemudian menghitung

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

performa akurasi model dalam mengidentifikasi penyakit jantung.
Hasil: Model *machine learning* yang dibuat dengan algoritma XgBoost mendapatkan akuasi 91% pada *data training* dan 83% pada *data testing*. Pengujian yang dilakukan menggunakan *dataset* lain mendapatkan akurasi 78%, dan model yang dibuat secara umum mendapatkan akuasi sebesar 90%.
Keaslian/ *state of the art*: Penelitian ini dilakukan menggunakan algoritma XgBoost yang dikombinasikan dengan RandomizedSearch Optimizer sebagai tuning *hyper parameter* untuk pembuatan model *machine learning*.

1. Pendahuluan

Jantung merupakan organ dalam manusia yang memiliki fungsi sangat penting yaitu untuk mengedarkan darah yang berisi oksigen dan nutrisi ke seluruh tubuh dan untuk mengangkut sisa hasil metabolisme tubuh, sehingga tubuh dapat bekerja dengan optimal. Sehingga akan sangat fatal apabila di dalam organ jantung terdapat gangguan seperti penyumbatan pembuluh darah, dan lain-lain. Penyakit jantung adalah penyakit yang menyerang organ jantung, contohnya adalah penyumbatan pembuluh darah pada jantung. Penyakit ini menyerang pembuluh darah arteri karena terjadi proses *arteosklerosis* pada dinding *arteri* yang menyebabkan penyempitan [1]**.**

Penyakit jantung dapat disebabkan oleh beberapa faktor seperti peningkatan kadar kolesterol karena dapat menyebabkan penumpukan lemak pada dinding arteri dan dapat menyebabkan *arteriosklerotik*. Selain itu dapat juga disebabkan oleh peningkatan tekanan darah atau *hipertensi*, karena saat tekanan darah meningkat maka dapat membebani kerja jantung dan juga menyebabkan *arteriosklerotik* karena saat tekanan darah meningkat akan menyebabkan gaya renggang yang dapat merobek lapisan *endotel arteri* dan *arteriol*. Kemudian penyakit jantung juga dapat disebabkan oleh kebiasaan merokok, orang dengan kebiasaan merokok mempunyai risiko 2,3 kali lebih besar untuk terkena penyakit jantung pada usia kurang dari 45 tahun.

Gejala klinis penyakit jantung antara lain adalah sering merasakan sesak napas yang ditandai dengan napas yang berat dan pendek sewaktu melakukan aktivitas berat, semakin lama rasa sesak napas akan semakin bertambah. Gejala lainnya adalah *klaudiokasi intermiten*, yaitu rasa nyeri pada daerah ekstrimitas bawah dan terjadi selama atau setelah melakukan olah raga. Gejala lainnya adalah mengalami perubahan warna kulit, dan kadar kolesterol yang meningkat biasanya di atas 180mg/dl untuk usia kurang dari 30 tahun, dan di atas 200mg/dl untuk yang berusia lebih dari 30 tahun.

Melihat dari tingkat bahaya penyakit jantung, maka penelitian yang membahas tentang penyakit jantung sudah banyak dilalukan menggunakan beberapa algoritma, namun dari beberapa penelitian yang ada dirasa masih memiliki kekurangan yaitu *overfitting* yang ada pada model yang dibuat. *Overfitting* adalah kondisi dimana model *machine learning* yang dibuat memiliki

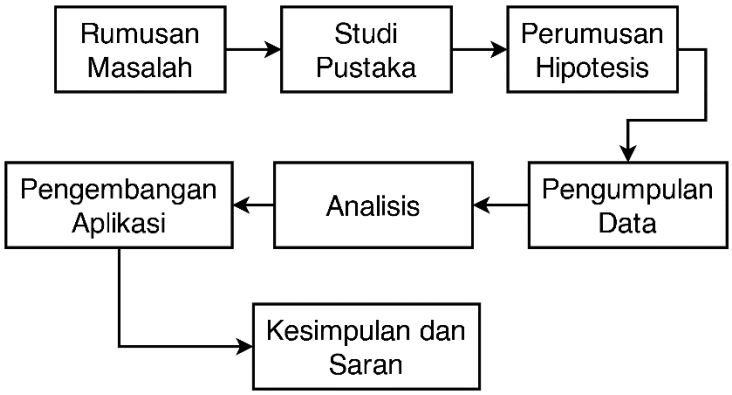
hasil akurasi pada *data training* yang sangat bagus, namun pada saat dilakukan pengujian menggunakan *data testing* didapatkan hasil akurasi model yang buruk [2].

Beberapa penelitian yang sudah dilakukan adalah penelitian dari Erwin Praseyo dan Budi Prasetro, penelitian tersebut dilakukan dengan algoritma C4.5 yang dikombinasikan dengan teknik *bagging*. Hasil dari penelitian tersebut ditampilkan dalam *confusion matrix* dengan akurasi algoritma C4.5 adalah 72,98% dan algoritma C4.5 yang dikombinasikan dengan *bagging* mendapat akurasi 81,84%. Didapat kesimpulan bahwa penerapan teknik *bagging* dapat meningkatkan akurasi sebesar 8,86% [3]. Kemudian penelitian yang dilakukan oleh Putra dan Rini yang dilakuakan untuk membandingkan beberapa algoritma seperti *Naive Bayes*, SVM, C4.5, *Logistic Regression*, dan *Back Propagation* dalam memprediksi penyakit jantung. Hasil dari penelitian tersebut adalah akurasi tertinggi didapatkan oleh algoritma *Naive Bayes* dengan 84,07% dan didapatkan kesimpulan bahwa algoritma *Naive Bayes* menjadi algoritma terbaik di dalam penelitian tersebut [4]. Kemudian penelitian yang dilakukan oleh Wibisono dan Ahmad Fahrurozi, penelitian tersebut dilakukan untuk membandingkan beberapa algoritma seperti algoritma *Naive Bayes*, KNN, *Decision Tree*, *Random Forest*, dan SVM pada kasus pengenalan penyakit jantung koroner. Hasil dari penelitian tersebut adalah algoritma *Random Forest* mendapat akurasi 85,67%, algoritma *Naive Bayes* dan *Decision Tree* mendapat akurasi yang sama, yaitu 80,33%. Algoritma KNN mendapatkan akurasi sebesar 69,67% . Dari penelitian tersebut didapatkan kesimpulan bahwa algoritma *Random Forest* menjadi algoritma terbaik dalam penelitian tersebut [5].

Dari penjelasan dan penelitian sebelumnya yang sudah ada, penelitian ini dilakukan untuk memprediksi penyakit jantung menggunakan algoritma XgBoost sekaligus mengurangi *overfitting* karena algoritma XgBoost memiliki *regularization* sebagai teknik untuk mencegah terjadinya *overfitting*. Hasil dari penelitian ini adalah model *machine learning* dan *confusion matrix* untuk mengetahui akurasi algoritma XgBoost dalam membuat model prediksi penyakit jantung.

2. Metode/Perancangan

Metode penelitian yang dilakukan ini adalah penelitian kuantitatif, untuk tahapan penelitian ini dapat dilihat pada **Gambar 1.** di bawah ini.



Gambar 1. Tahapan Penelitian

Seperti yang dapat dilihat pada **Gambar 1.** di atas, penelitian ini terdari dari beberapa tahapan seperti perumusan masalah, studi pustaka, hingga pengembangan aplikasi. Namun dari tahapan yang ada, akan dilakukan pembahasan lebih pada tahapan analisis dan pengembangan aplikasi.

2.1. Pengumpulan Data

Data yang digunakan dalam pelenitian ini adalah data sekunder, data sekunder adalah data yang bisa didapatkan secara tidak langsung misalnya melalui orang lain, sumber dokumen, *website*, dan sumber lainnya [6]. Data yang digunakan adalah *cleveland dataset* yang berjumlah 303 data dengan pembagian target 138 data tergolong dalam klasifikasi tidak memiliki penyakit jantung, dan 165 data tergolong dalam klasifikasi memiliki penyakit jantung. Data *cleveland dataset* memiliki 13 parameter yang dapat dilihat pada **Tabel 1.** di bawah ini.

Tabel 1. Parameter Dataset

No.	Parameter	Deskripsi	Keterangan
1.	Age	Umur pasien	numerik
2.	Sex	Jenis kelamin pasien	0: wanita, 1: pria
3.	Cp	Chest pain type	1: typical angina, 2: atypical angina 3: non-angina pain, 4: asympotomatic
4.	Trestbps	Tekanan darah pasien	Numerik
5.	Chol	Kadar serum kolesterol	Numerik
6.	Fbs	Kadar gula darah apakah > 120mg/dl	0: false, 1: True
7.	Restecg	Hasil ECG selama istirahat	0: normal, 1: abnormal (memiliki kelainan pada gelombang ST-T) 2: hipertrofil ventrikel
8.	Thalach	Detak jantung maksimal yang dicapai	numerik
9.	Exang	Ukuran boolean yang menunjukan apakah latihan angina terjadi	0: no 1: yes
10.	Oldpeak	Segment ST yang diperoleh dari hasil ECG	Numerik
11.	Slope	Jenis kemiringan segment ST untuk latihan maksimal (puncak)	1: upsloping 2: flat, 3: downsloping
12.	Ca	Jumlah vessel yang diwarnai oleh fluroskopi	0, 1, 2, 3
14.	Thal	Parameter thalasemia	1: normal, 2: cacat tetap, 3: reversible
14.	Target	Target kelas klasifikasi	0: tidak terkena penyakit jantung 1: terkena penyakit jantung

Seperti yang dapat dilihat pada **Tabel 1.** di atas, jumlah parameter adalah 13 dan 1 kolom target yang terdiri dari 0 dan 1. Untuk penjelas lebih mengenai parameter adalah sebagai berikut.

- Cp (*Chest Pain Type*)
  Cp adalah nyeri dada yang disebabkan karena otot jantung tidak mendapatkan cukup darah yang kaya dengan oksigen. Nyeri dada dapat menjadi kekhawatiran pada masalah jantung, sehingga sangan umum ketika gejala ini muncul orang akan mengira terjadi serangan jantung.
- Trestbps
  Tekanan darah adalah ukuran kekuatan yang digunakan jantung untuk memompa darah ke seluruh tubuh. Pengaruh dalam penyakit jantung adalah apabila tekanan darah

mengalami pengingkatan. Pengingkatan tekanan darah merupakan beban yang berat sehingga akan menyebabkan *hipertropi ventrikel* atau pembesaran *ventrikel*.

- Chol
  Serum kolesterol adalah senyawa lipid yang mempunyai inti *siklopenta perhidrofenanta* [7]. Permasalahan dalam penyakit jantung adalah apabila terjadi penginkatan kadar serum kolesterol. Ukuran kolesterol normal sebaiknya kurang dari 200mg/dl.
- Fbs
  Fbs adalah ukuran kadar gula yang dilakukan pasien setelah berpuasa selama 8 jam. Ukuran fbs apabila melebihi 120mg/dl maka dapat dikatakan bahawa pasien tersebut memiliki potensi untuk *diabetes*. Pengaruh *diabetes militus* dalam kesehatan jantung adalah pasien yang memiliki *diabetes militus* mempunyai peluang 10,25 kali lebih besar untuk terkena penyakit jantung daripada pasien yang tidak memiliki *diabetes militus*.
- Restecg
  Restecg adalah hasil pengukuran ECG pada jantung. ECG dilakukan dengan cara menempatkan sepuluh elektroda pada titik-titik tertentu, enam elektroda dipasangkan di area dada, dan selebihnya dipasangkan pada area ekstrimitas. Pemeriksaan ECG merupakan hal yang wajib dilakukan kepada pasien yang memiliki tanda-tanda atau gejala penyakit jantung [8].
- Thalach
  Thalach adalah ukuran detak jantung maksimal yang dapat dicapai. Detak jantung manusia normal berkisar antara 60 – 100 kali per menit [9].
- Exang
  Exang adalah informasi mengenai latihan angina (nyeri dada) yang dilakukan, apabila selama aktivitas latihan angina dilakukan pasien merasakan rasa sakit maka *exang* akan bernilai 1, dan apabila tidak merasakan rasa sakit maka *exang* akan bernilai 0 [10].
- Oldpeak
  Oldpeak adalah ukuran gelombang segment ST yang terjadi diakibatkan oleh latihan *relative* terhadap kondisi jantung saat beristirahat atau tidak bekerja keras. Ukuran oldpeak biasanya didefinisikan dengan nilai 0 sampai 3mm [11].
- Slope
  Slope adalah jenis kemiringan segment ST yang terdapat pada gambar grafik hasil pemerikasaan ECG, segment ST yang dihasilkan dari pemeriksaan nantinya dapat dihubungkan dengan ketidaknormalan pada dinding jantung [12].
- Ca
  Ca adalah jumlah dari *vessel* yang diwarnai oleh *fluroskopi*. Teknik *fluroskopi* adalah teknik yang memanfaatkan salah satu sinar X yaitu jika sinar tersebut terkena bahan maka akan berpenda menjadi warna tertentu [13]. *Fluroskopi* juga dapat digunakan untuk menunjang prosedur medis tertentu contohnya pada prosedur pemasangan *ring* jantung [14].
- Thal
  Thal adalah parameter *thalasemia* yang dimiliki pasien. *Thalasemia* adalah sebuah penyakit keturunan yang diakibatkan oleh gagalnya pembentukan satu dari empat asam

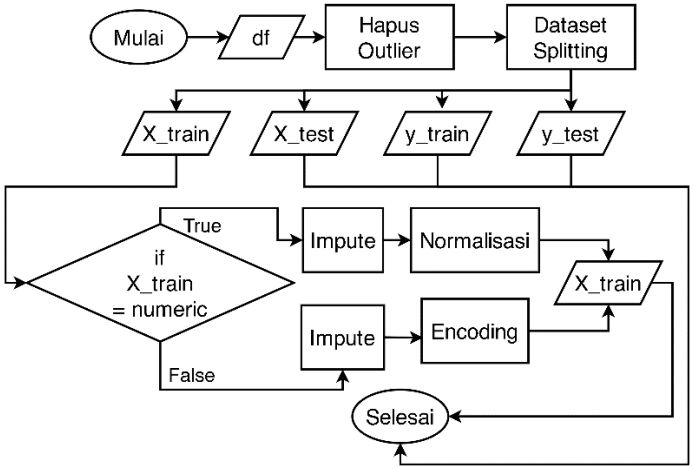
ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

amino yang membentuk *hemoglobin*, sehingga *hemoglobin* tidak terbentuk secara sempurna [15].

2.2. Analisis

2.2.1. Preprocessing

Tahapan proses *preprocessing* yang dilakukan pada penelitian ini dapat dilihat pada *Flowchart* yang ditampilkan pada **Gambar 2.** di bawah ini.



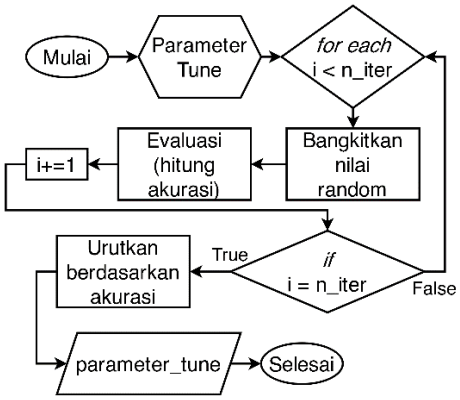
Gambar 2. Flowchart Preprocessing

Seperti yang dapa dilihat pada Gambar 1.2 di atas, proses *preprocessing* yang dilakukan terdiri dari beberapa proses seperti hapus *outlier*, kemudian melakukan *dataset splitting* untuk membagi *dataframe* mejadi empat. Kemudian pada data dengan tipe data numerik akan dilakukan proses normalisasi, proses normalisasi dilakukan supaya proses *training* menjadi lebih cepat karena dapat memudahkan model untuk memahami data [16]. Sedangkan pada data dengan tipe data kategorik akan dilakukan proses *encoding*.

2.2.2. Training

Pada penelitian ini proses *training* dilakuakan dengan bantuan RandomizedSearch Optimizer untuk penentuan *hyper parameter* terbaik. *Flowchart* RandomizedSearch Optimizer dapat dilihat pada **Gambar 3.** di bawah ini.

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx



Gambar 3. Flowchart RandomizedSearch Optimizer

Seperti yang dapat dilihat pada **Gambar 3.** di atas, *hyper parameter* yang akan dituning dituliskan dalam bentuk list dan menuliskan jumlah iterasi percobaan yang diminta. RandomizedSearch Optimizer akan menemukan parameter terbaik dari parameter yang diinputkan. RandomizedSearch Optimizer juga memiliki *cross validation*. Algoritma XgBoost memiliki rumus utama untuk menentukan nilai *similarity*, nilai *gain*, dan *output value* yang dapat dilihat pada **persamaan 1**, **2**, dan **3**di bawah ini.

$$similarity = \frac{\sum(residual)^2}{\sum[prev\ probabillity_i \times (1 - prev\ probabillity_i)] + \lambda} \quad (1)$$

$$gain = left\ similarity + right\ similarity - root\ similarity \quad (2)$$

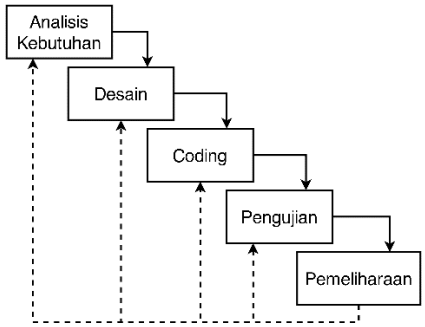
$$O_{value} = \frac{\sum(residual_i)}{\sum[prev\ probabillity_i \times (prev\ probability_i)] + \lambda} \quad (3)$$

2.2.3. Evaluasi

Evalusai yang dilakukan bertujuan untuk melihat performa model *machine learning* yang dibuat. Untun melakukan evaluasi dilakukan dengan cara menguji model yang ada menggunakan *data train* dan *data testing* yang ada, dan juga dilakukan evaluasi dengan *dataset* lain yaitu *statelog dataset*. Hasil evaluasi ditampilkan dalam bentuk *confussion matrix*, sehingga dapat menghitung akurasi model dari *confussion matrix* tersebut.

2.3. Pengembangan Aplikasi

Pengembangan aplikasi yang dilakukan dalam penelitian ini dilakukan dengan model pengembangan *waterfall*. Untuk tahapan *waterfall* dapat dilihat pada **Gambar 4.** di bawah ini.



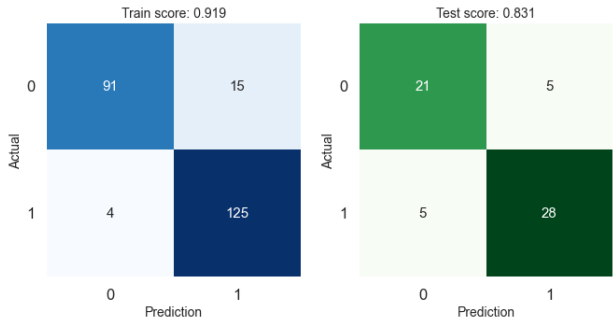
Gambar 4. Tahapan Waterfall

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

Seperti yang dapat dilihat pada **Gambar 4.** di atas, tahapan *waterfall* diawali dari analisis kebutuhan baik kebutuhan fungsional maupun non fungsional, kemudian desain, coding, pengujian aplikasi, hingga pemeliharaan aplikasi setelah selesai dibuat.

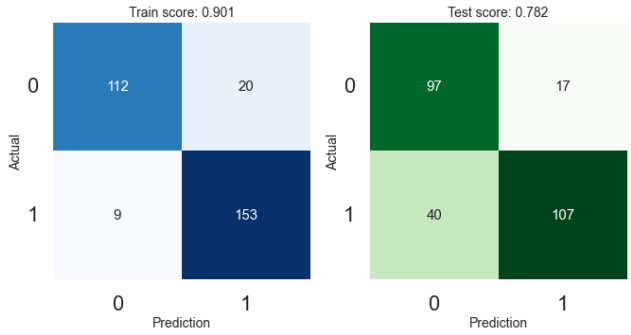
3. Hasil dan Pembahasan

Hasil dari penelitian ini adalah model *machine learning* yang dapat memprediksi penyakit jantung. Hasil evaluasi dan pengujian model yang dibuat dapat dilihat pada **Gambar 5.** di bawah ini.



Gambar 5. Confusion Matrix Model

Pada **Gambar 5.** di atas dapat dilihat bahwa hasil mode yang dibuat mendapatkan akurasi sebesar 91% pada *data training* dan akurasi sebesar 83% pada *data testing*. Sedangkan untuk hasil pengujian model dengan *dataset* lain dapat dilihat pada **Gambar 6.** di bawah ini.



Gambar 6. Confusion Matrix Pada Dataset Lain

Seperti yang dapat dilihat pada **Gambar 6.** di atas, akurasi model secara keseluruhan mendapatkan akurasi sebesar 90% dan apabila model dilakukan pengujian dengan *dataset* lain yaitu *dataset statelog*, maka model mendapatkan akurasi sebesar 78%.

4. Kesimpulan dan Saran

Berdasarkan penelitian yang dilakukan dapat disimpulkan bahwa algoritma XgBoost dapat digunakan untuk menyelesaikan permasalahan prediksi penyakit jantung. Untuk akurasi model *machine learning* yang dibuat mendapatkan akurasi sebesar 90%, akurasi tersebut dinilai cukup tinggi apabila dibandingkan dengan penelitian yang dilakukan dengan algoritma lain. Pengujian

yang dilakukan berdasarkan *data training* dan *data testing*, model mendapatkan akurasi 91% pada *data training* dan akurasi 83% pada *data testing*.

Untuk mengatasi *overfitting*, sayangnya pada penelitian ini *overfitting* yang diatasi tidak terlalu maksimal karena parameter lambda tidak dilakukan *tuning* dan menggunakan nilai *default* yaitu 1, karena jika nilai lambda terlalu tinggi akan menurunkan akurasi pada *data testing*. Maka saran yang dapat diberikan adalah melakukan *tuning* pada *hyper parameter* lambda dan bisa juga menggunakan *dataset* lain dengan jumlah yang lebih banyak.

Daftar Pustaka

[1] L. Marleni and A. Alhabib, "Faktor Risiko Penyakit Jantung Koroner di RSI SITI Khadijah Palembang," *J. Kesehat.*, vol. 8, no. 3, p. 478, 2017, doi: 10.26630/jk.v8i3.663.

[2] A. Septadaya, C. Dewi, and B. Rahayudi, "Implementasi Extreme Learning Machine dan Fast Independent Component Analysis untuk Klasifikasi Aritmia Berdasarkan Rekaman Elektrokardiogram," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 2548, no. 5, p. 964X, 2019.

[3] E. Prasetyo and B. Prasetiyo, "Peningkatan Akurasi Klasifikasi Algoritma C4.5 Menggunakan Teknik Bagging Pada Diagnosis Penyakit Jantung," vol. 7, no. 5, pp. 1035–1040, 2020, doi: 10.25126/jtiik.202072379.

[4] P. D. Putra and D. P. Rini, "Prediksi Penyakit Jantung dengan Algoritma Klasifikasi," *Pros. Annu. Res. Semin. 2019*, vol. 5, no. 1, pp. 978–979, 2019.

[5] A. B. Wibisono and A. Fahrurozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 3, pp. 161–170, 2019, doi: 10.35760/tr.2019.v24i3.2393.

[6] Z. A. Haqie, R. E. Nadiah, and O. P. Ariyani, "Inovasi Pelayanan Publik Suroboyo Bis Di Kota Surabaya," *JPSI (Journal Public Sect. Innov.*, vol. 5, no. 1, p. 23, 2020, doi: 10.26740/jpsi.v5n1.p23-30.

[7] M. S. Dr. Bernatal Saragih, S.P., *Kolesterol dan Usaha-Usaha Penurunannya*, 1st ed., no. September. Yogyakarta: Penerbit Bimotry Yogyakarta, 2011.

[8] Rosmalinda, D. Karim, and A. P. Dewi, "Gambaran Tingkat Pengetahuan Perawat Irna Medikal Dalam Menginterpretasi Hasil EKG," no. 1, 2014.

[9] N. Nahdliyah, "Penelitian Tentang Detak Jantung," *Jur. Sist. Komput. Univ. Sriwij.*, vol. 52, no. 1, pp. 1–5, 2019.

[10] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. November 2018, 2019, doi: 10.1016/j.imu.2019.100203.

[11] O. W. Purbo and P. Sudiarta, "Inovasi Teknologi Informasi dan Komunikasi Dalam Menunjang Technopreneurship," *Angew. Chemie Int. Ed. 6(11), 951–952.*, pp. 5–24, 2015.

[12] I. D. G. H. Wisana, "Identifikasi Isyarat Elektrokardiogram Segmen ST dan Kontraksi Ventrikel Prematur Berbasis Gelombang Singkat," *Univ. Gadjah Mada Yogyakarta*,

ISSN: 1829-667X / E-ISSN: 2460-9021
DOI:10.31515/telematika.v18i1.xxxx

2013.

[13] M. S. K. Ayu, "Proteksi Radiasi Pada Pasien, Pekerja, dan Lingkungan di Dalam Instalasi Radiologi," *Inst. Ilmu Kesehat. Str. Indones.*, 2019.

[14] W. A. Mustofa, "Manfaat Foto Rongten dan Dampaknya," *Intitut Ilmu Kesehat. Str. Indones.*, 2021.

[15] Wahyu Kusuma, "Self Acceptance Pada Remaja Penderita Thalasemia," pp. 8–27, 2016.

[16] T. T. Hanifa, S. Al-faraby, and Adiwijaya, "Analisis Churn Prediction pada Data Pelanggan PT . Telekomunikasi dengan Logistic Regression dan Underbagging," *Univ. Telkom*, vol. 4, no. 2, p. 78, 2017.