Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 3, June 2024, pp. 821-830

DOI: https://doi.org/10.52436/1.jutif.2024.5.3.2047
p-ISSN: 2723-3863
e-ISSN: 2723-3871

IMPROVING HEART DISEASE PREDICTION ACCURACY USING PRINCIPAL COMPONENT ANALYSIS (PCA) IN MACHINE LEARNING ALGORITHMS

Zirji Jayidan[1], Amril Mutoi Siregar[*2], Sutan Faisal[3], Hanny Hikmayanti[4]

[1,2,3,4]Informatics Departement, Faculty of Computer Sciences, Universitas Buana Perjuangan Karawang, Indonesia
Email: [1]if20.zirjijayidan@mhs.ubpkarawang.ac.id, [2]amril.mutoi@ubpkarawang.ac.id, [3]sutanfaisal@ubpkarawang.ac.id, [4]hanny.hikmayanti@ubpkarawang.ac.id

(Article received: May 06, 2024; Revision: May 25, 2024; published: June 04, 2024)

Abstract

This study aims to improve the accuracy of heart disease prediction using Principal Component Analysis (PCA) for feature extraction and various machine learning algorithms. The dataset consists of 334 rows with 49 attributes, 5 classes and 31 target diagnoses. The five algorithms used were K-nearest neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). Results show that algorithms using PCA achieve high accuracy, especially RF, LR, and DT with accuracy up to 1.00. This research highlights the potential of PCA-based machine learning models in early diagnosis of heart disease.

Keywords: *Diagnostic Accuracy, Feature Extraction, Heart Disease Prediction, Machine Learning Algorithm, Principal Component Analysis (PCA).*

PENINGKATAN AKURASI PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS* (PCA) PADA ALGORITMA *MACHINE LEARNING*

Abstrak

Penelitian ini bertujuan meningkatkan akurasi prediksi penyakit jantung menggunakan Principal Component Analysis (PCA) untuk ekstraksi fitur dan berbagai algoritma machine learning. Dataset terdiri dari 334 baris dengan 49 atribut, 5 kelas dan 31 diagnosis target. Lima algoritma yang digunakan adalah *K-nearest neighbors* (KNN), *Logistic Regression* (LR), *Random Forest* (RF), *Support Vector Machine* (SVM), dan *Decision Tree* (DT). Hasil menunjukkan bahwa algoritma yang menggunakan PCA mencapai akurasi tinggi, terutama RF, LR, dan DT dengan akurasi hingga 1.00. Penelitian ini menyoroti potensi model machine learning berbasis PCA dalam diagnosis dini penyakit jantung.

Kata kunci: *Akurasi Diagnostik, Algoritma Machine Learning, Ekstraksi Fitur, Prediksi Penyakit Jantung, Principal Component Analysis (PCA).*

1. PENDAHULUAN

Menurut Organisasi Kesehatan Dunia (WHO) pada tahun 2021, diperkirakan pada tahun 2019 dari 17,9 juta kematian di seluruh dunia, 32% dari seluruh kematian di dunia, 85% disebabkan oleh penyakit jantung dan stroke. Dari 17 juta kematian dini (dibawah usia 70 tahun) diakibatkan penyakit tidak menular pada tahun 2019, 38% disebabkan oleh penyakit jantung dan pembuluh darah [1].

Menurut *World Heart Federation* (WHF), Lebih dari 500 juta orang di seluruh dunia masih terpengaruh oleh penyakit kardiovaskular, yang menyebabkan 20,5 juta kematian pada tahun 2021 [2].

Berdasarkan data dari *Global Burden of Disease dan Institute for Health Metrics and Evaluation* (IHME) tahun 2014-2019, penyakit jantung adalah penyebab utama kematian di Indonesia. Data dari Badan Riset Kesehatan Dasar (Riskesdas) tahun 2013 dan 2018 menunjukkan adanya peningkatan tren penyakit jantung, yaitu dari 0,5% pada tahun 2013 menjadi 1,5% pada tahun 2018 [3].

Oleh karena itu, jika penyakit ini dapat dideteksi sejak dini, dampak buruk yang mungkin timbul dapat segera dicegah. Algoritma klasifikasi *machine learning* memiliki kemampuan untuk menngetahui cara kerja dan kefektifan fitur *Principal Component Analysis* (PCA) dari masing-masing algoritma.

Principal component Analysis (PCA) adalah fitur pengurangan dimensi yang paling popular [4].

Pendekatan pada penelitian ini dalam klasifikasi adalah menggunakan pengklasifikasi pembelajaran terawasi. Kami menggunakan lima pengklasifikasi yang terkenal yaitu *Decision Tree* (DT), *Random Forest* (RF), *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), dan *Logistic Regression* (LR). Kinerjanya didasarkan pada fitur berbasis *Principal Component Analysis* (PCA).

Beberapa penelitian sebelumnya telah membahas tentang fitur PCA pada algoritma *machine learning* [5]. Hasil penelitian Huan-Jung Chiu, Tzuu-Hseng S. LI, dan Ping-Huan Kuo memverifikasi metode menggunakan SVM dengan nilai 10-fold, memvalidasi silang 10 kali lipat. *Principal Component Analysis* (PCA) digunakan untuk mengidentifikasi bagian data yang berharga dan selanjutnya mengurangi dimensi data. Untuk memprediksi kanker payudara berdasarkan sembilan

atribut individu, termasuk usia, indeks massa tubuh, glukosa, insulin, dan penilaian model homeostasis. menggunakan *dataset* Manuel Gomes dari Pusat Rumah Sakit Universitas Coimbra, dan mendapat nilai akurasi sebesar 86,97% [6].

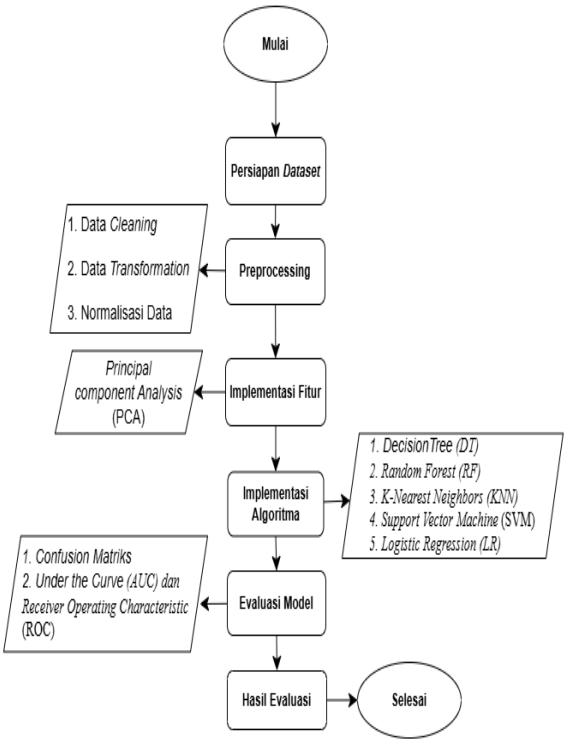
Pada penelitian tahun 2020, yang dilakukan oleh Subhash Waskle, Lokesh Parashar, dan Upendra Singh, Penelitian ini mengusulkan sebuah pendekatan untuk mengembangkan IDS yang efisien dengan menggunakan *Principal Component Analysis* (PCA) dan algoritma klasifikasi *random forest*. Di mana PCA membantu untuk mengorganisir *dataset* dengan mengurangi dimensi *dataset* dan *random forest* akan membantu dalam klasifikasi Hasil yang diperoleh dengan metode yang digunakan memiliki nilai untuk waktu kinerja (menit) adalah 3.24 menit, tingkat akurasi (%) adalah 96.78 dan tingkat kesalahan (%) adalah 0.21%. [7].

Studi sebelumnya telah dilakukan oleh Amin Ul Haq, Jian Ping Li, Abdus Saboor, Jalaluddin Khan, Samad Wali, Sultan Ahmad, Amjad Ali, Ghufran Ahmad Khan, dan Wang Zhou pada tahun 2021. Dalam penelitian ini, mengusulkan metode identifikasi kanker payudara baru dengan menggunakan machine learning algoritma machine learning dan data klinis. Dalam metode yang diusulkan, metode *supervised* dan *unsupervised* dengan menggunakan fitur *Principal Component Analysis* (PCA) didalamnya, kemudian fitur-fitur yang dipilih ini telah digunakan untuk pelatihan dan pengujian SVM sebagi pendukung pengklasifikasi untuk mendeteksi kanker payudara secara akurat dan tepat waktu. Selain itu, dalam pendekatan yang diusulkan, metode validasi *K-fold cross*. Metode yang diusulkan telah mencapai nilai yang tinggi dalam hal akurasi pada fitur yang dipilih oleh algoritma SVM pada metode supervised dan mencapai akurasi 99,91% [8].

Berdasarkan permasalahn yang telah dijelaskan sebelumnya serta dukungan dari penelitian sebelumnya terkait implementasi fitur dalam meningkatkan kinerja algoritma *machine learning*, kami memperluas penelitian ini dengan menggunakan lima jenis pengklasifikasi yang berbeda, yaitu: Algoritma *Decision Tree* (DT), *Random Forest* (RF), *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), dan *Logistic Regression* (LR) dengan pembelajaran secara terawasi (*Supervised*). Kami menerapkan pendekatan berbasis PCA dengan teknik yang bervariasi pada setiap pengklasifikasi. Evaluasi dilakukan dengan membandingkan kinerja masing-masing pengklasifikasi, terutama melalui *confusion matrix* dan juga AUC-ROC untuk menentukan tingkat akurasi terbaik dari setiap algoritma [9]. Harapannya, model yang dikembangkan dapat membantu ahli dalam melakukan penilaian yang lebih konsisten dan efisien dalam memprediksi penyakit jantung secara dini.

2. METODE PENELITIAN

Proses penelitian dimulai dengan melakukan studi literatur. Tujuan dari studi literatur ini adalah untuk menemukan dasar teoritis yang relevan dan mencari referensi ilmiah yang mendukung penelitian. Selama penelitian, langkah-langkah atau tahapan dapat dilihat pada Gambar 1.



Gambar 1. Flowchart tahapan penelitian

2.1. Persiapan *Dataset*

Penelitian ini menggunakan *dataset* yang berasal dari internet melalui *Kaggle*:

www.kaggle.com/datasets/pratikshakya/heart-disease-details, *dataset* ini merupakan data tabular dan bersifat *supervised*.

Terdiri dari 122853 total data semua. Terdapat 334 baris 49 kolom jumlah fitur termasuk kolom name dengan 5 kelas dari 31 diagnosis target, yaitu: *Name, Gender, Age ,Chest pain, Shortness of breath, Fatigue, Systolic, Diastolic, Heart rate* (bpm), *Lung sounds, Cholesterol level* (mg/dL), LDL *level* (mg/dL), HDL *level* (mg/dL), *Diabetes, Atrial fibrillation, Mitral valve prolapse, Rheumatic fever, Mitral stenosis, Aortic stenosis, Tricuspid stenosis, Pulmonary stenosis, Dilated cardiomyopathy, Hypertrophic cardiomyopathy, Restrictive cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy, Takotsubo cardiomyopathy, Drug use, Fever,Chills, Joint pain, Alcoholism, Hypertension, Fainting, Dizziness, Smoking, High cholesterol, Echocardiogram, Blood culture* ECG, *Cardiac* CT, *Obesity , Murmur , Chest x-ray, Previous illnesses, Pulmonary function tests, Spirometry, Diagnosis, Medications, Treatment*.

2.2. *Preprocessing*

Dalam menghasilkan data yang berkualitas, digunakan teknik-teknik sebagai berikut:

2.2.1. *Data Cleaning*

Dataset terlebih dahulu dibersihkan dan nilai observasi disajikan dengan tidak mengulang awalan dari setiap fitur. Hal ini digunakan untuk mengidentifikasi dan menghapus kolom name, data yang tidak lengkap data yang tidak lengkap (*Missing Value*) dan data duplikat [10].

2.2.2. *Data Transformation*

Untuk meningkatkan akurasi dan efisiensi algoritma yang digunakan dalam penelitian ini dilakukan proses encoding dilakukan proses pengkodean untuk mengubah kategori menjadi numerik dan memisahkannya. Terdapat 12 variabel yang kami rubah menjadi *category* yaitu Diagnosis, *Gender, Echocardiogram, Blood culture, EKG, Cardiac CT, Chest x-ray, Previous illnesses, Pulmonary function tests, Spirometry, Medications, Treatment. Lalu, Chest pain, Shortness of breath, Fatigue, Systolic, Diastolic*, Diagnosis menjadi data numerik.

2.2.3. *Normalisasi Data*

Normalisasi data merupakan teknik yang digunakan untuk mengubah nilai atribut (fitur) numerik agar berada dalam skala yang sama, umumnya dalam rentang [0, 1] atau [-1, 1] [11]. Tujuan utama dari normalisasi adalah untuk mencegah fitur dengan skala besar mendominasi fitur dengan skala kecil. Teknik ini sangat berguna ketika fitur memiliki rentang yang sangat bervariasi atau saat menggunakan algoritma yang peka terhadap skala data, seperti K-*Nearest Neighbors* dan jaringan saraf.

2.3. *Principal Component Analysis* (PCA)

Principal component Analysis (PCA) adalah algoritma pengurangan dimensi yang paling populer. PCA sebagai teknik statistik digunakan untuk menganalisis keterkaitan. Data direduksi menjadi sejumlah kecil dimensi yang disebut komponen utama [12].

2.4. *K-Nearest Neighbors* (KNN)

Algoritma *K-nearest neighbors* merupakan sebuah teknik klasifikasi yang bersifat terawasi. Metode ini mengelompokkan objek berdasarkan pada tetangga terdekatnya. Ia termasuk dalam kategori pembelajaran berbasis contoh di mana atribut sebuah objek diukur jaraknya dari atribut tetangganya menggunakan metrik jarak Euclidean.

Algoritma ini memanfaatkan kumpulan titik yang disebut sebagai *neighbors* untuk belajar cara mengklasifikasikan titik lainnya. Data dikelompokkan berdasarkan kesamaan di antara mereka, dan KNN dapat digunakan untuk mengisi nilai yang hilang dalam data. Setelah nilai-nilai yang hilang diisi, berbagai teknik prediksi diterapkan pada *dataset* [13]. Ini memungkinkan mencapai tingkat akurasi yang tinggi dengan memanfaatkan berbagai kombinasi dari algoritma ini. Model KNN tidak melibatkan proses pelatihan yang nyata, tetapi dalam implementasi umumnya, data pelatihan disimpan dalam memori untuk digunakan dalam tahap prediksi.

2.5. *Logistic Regression* (LR)

Logistic regression adalah sebuah algoritma klasifikasi. Untuk masalah klasifikasi *biner*, untuk memprediksi nilai variabel prediktif y ketika y ∈ [0, 1], 0 adalah kelas negatif dan 1 adalah kelas positif. Ia juga menggunakan multi klasifikasi untuk memprediksi nilai y ketika y ∈ [0, 1, 2, 3]. Untuk mengklasifikasikan dua kelas 0 dan 1, sebuah hipotesis h(θ) = θTX akan dirancang dan ambang batas keluaran pengklasifikasi adalah hθ(x) sebesar 0.5. Jika nilai hipotesis hθ(x) ≥ 0.5, maka akan memprediksi y = 1 yang berarti orang tersebut memiliki penyakit jantung [14].

Model Regresi Logistik memiliki fungsi log-odds sebagai berikut:

$$\pi(x) = \frac{exp\ (\beta 0+\beta 1X1+\ldots+\beta nXn)}{1+exp(\beta 0+\beta 1X1+\ldots+\beta nXn)} \quad (1)$$

Dimana:

Exp adalah probabilitas kelas

$\pi(x)$ adalah log-odds dari kelas positif

$\beta 0, \beta 1, \ldots, \beta n$ adalah parameter selama pelatihan model

$X1, X2, \ldots, Xn$ adalah fitur dari data [15].

2.6. *Random Forest* (RF)

Pengklasifikasi *random forest* (RF) merupakan metode serbaguna yang diaplikasikan untuk tugas-tugas klasifikasi dan regresi. Metode ini membangun ansambel pohon keputusan, yang mampu menafsirkan dan membuat prediksi pada hasil kategorikal dengan probabilitas yang dapat disesuaikan. Namun, diperlukan kehati-hatian untuk mencegah *overfitting*, terutama ketika berhadapan dengan *dataset* kecil. Keuntungan dari algoritma *random forest* adalah kemampuannya untuk menangani data atribut yang tidak lengkap [16].

Model umum *random forest* dapat direpresentasikan sebagai sebuah fungsi:

$$f(x) = \sum_{n=1}^{N} ci.\, l(x \epsilon Ri) \quad (2)$$

Dimana:
N adalah jumlah simpul dalam pohon
Ci adalah nilai konstan yang diberikan untuk setiap daun (simpul terminal) dari pohon
Ri adalah kumpulan data yang diwakili oleh node
I (.) adalah fungsi indikator [17].

2.7. *Support Vector Machine* (SVM)

SVM digunakan untuk menyelesaikan masalah klasifikasi dan regresi, terutama dengan memetakan sampel data ke dalam ruang fitur oleh fungsi kernel dan kemudian mengklasifikasikannya dengan *hyperplane* [18]. Fungsi keputusan untuk klasifikasi SVM dapat direpresentasikan sebagai:

$$f(x) = sign\,(w.x + b) \quad (3)$$

Dimana :
F(x) adalah fungsi Keputusan
W adalah vektor bobot
X adalah vektor fitur data
B adalah bias Tanda adalah fungsi penandaan (positif atau negatif) [19].

2.8. *Decision Tree* (DT)

Pohon keputusan adalah algoritma pembelajaran mesin yang diawasi. Format pohon keputusan adalah pohon sederhana di mana setiap *node* adalah *node* daun atau *node* keputusan. Teknik pohon keputusan sederhana dan mudah dipahami untuk pengambilan keputusan, pohon keputusan berisi node internal dan eksternal yang saling terkait, *node* internal adalah bagian keputusan yang membuat keputusan dan merupakan *node* anak yang mengunjungi *node* berikutnya, node daun, di sisi lain, tidak memiliki anak dan terkait dengan label [20].

Pohon keputusan dapat direpresentasikan dengan sekumpulan aturan keputusan yang bersifat hirarkis [21]. Secara umum, model pohon keputusan dapat direpresentasikan sebagai sebuah fungsi:

$$f(x) = \sum_{n=1}^{N} ci.\ I(x \epsilon Ri) \quad (4)$$

Sama seperti *random forest*, kecuali bahwa pelatihan *decision tree* melibatkan pemilihan aturan-aturan kumpulan dan membagi data ke dalam dua kelompok berdasarkan aturan-aturan tersebut.

2.9. *Confusion Matrix*

Confusion matriks digunakan untuk mengevaluasi kinerja metode klasifikasi dan mencakup empat istilah sebagai ekspresi: *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN) [22]. Hasil klasifikasi. TN menunjukkan nilai kumpulan yang diidentifikasi dengan benar, tetapi FP menunjukkan nilai kumpulan yang salah diidentifikasi sebagai positif [23]. Model ini digunakan untuk menghitung nilai akurasi, presisi, *recall*, dan F1-*score*. Presisi mencerminkan perbandingan antara nilai positif yang diprediksi dengan benar dan kumpulan nilai positif, sedangkan akurasi mengukur kinerja model [24].

Evaluasi Model:

$$Akurasi = \frac{Jumlah\ Prediksi\ yang\ Benar}{Jumlah\ Total\ Data} \quad (5)$$

Confusion Matrix:

$$\begin{pmatrix} True\ Negative\ (TN) & False\ Positive\ (FP) \\ False\ Negative\ (FN) & True\ Positive\ (TP) \end{pmatrix} \quad (6)$$

2.10. *Under the Curve* (AUC) & *Receiver Operating Characteristic* (ROC)

Area di bawah kurva (AUC) adalah area di bawah karakteristik operasi penerima (ROC). Kurva Karakteristik Operasi Penerima (ROC) dihasilkan dari hubungan antara sensitivitas dan spesifisitas pada berbagai titik potong. Nilai AUC secara teori berkisar antara 0 dan 1, di mana AUC mengukur kecocokan keseluruhan model yang digunakan. Semakin besar area di bawah kurva, semakin baik kemampuan variabel yang diteliti dalam memprediksi kejadian [22].

Sebuah AUC yang memiliki nilai lebih tinggi mencerminkan kinerja yang lebih baik dalam memprediksi dengan akurat antara 0 dan 1 [25]. Sementara itu, nilai yang mendekati nol mengindikasikan model yang kurang efektif, sementara nilai yang mendekati satu menandakan model yang memiliki kinerja yang baik [26] AUC digunakan untuk menghitung perbedaan kinerja algoritme [27].

3. HASIL DAN PEMBAHASAN

Hasil setelah *preprocessing* data, kinerja pengklasifikasi divisualisasikan dengan metrik kinerja yang berbeda. preprocessing data termasuk penggantian nilai yang hilang. Duplikat data, *encoder* untuk data, normalisasi data dan pengurangan dimensi berbasis PCA diterapkan di semua algoritma

machine learning yang kami gunakan untuk penelitian ini. *Confusion matrix* dan AUC-ROC juga digunakan untuk evaluasi kinerja.

3.1. Pengukuran Kinerja dengan *Confusion Matrix*

Kami menerapkan tujuh metode pada *dataset* dan mengukur kinerja masing-masing model. Rumusan kinerja tersebut adalah:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (7)$$

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$F1-Score = \frac{2\,x\,Recall\,x\,Precision}{Recall\,x\,Precision} \quad (10)$$

3.1.1. *K-Nearest Neighbors* (KNN)

Tabel 1. Pengukuran kinerja Algoritma *K-Nearest Neighbors* (KNN) dengan PCA

Performa *K-Nearest Neighbors* (KNN) PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
61%	69%	61%	60%

Tabel 2. Pengukuran kinerja Algoritma *K-Nearest Neighbors* (KNN) Non-PCA

Performa *K-Nearest Neighbors* (KNN) Non-PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
55%	68%	55%	55%

Dapat dilihat pada Tabel 1, menunjukkan hasil performa algoritma *K-Nearest Neighbors* (KNN) setelah diterapkannya *Principal Component Analysis* (PCA). PCA digunakan untuk mengurangi jumlah dimensi data dengan tujuan meningkatkan performa model. Hasil yang dicapai mencakup akurasi 61%, presisi 69%, *recall* 61%, dan F1-*Score* 60%.

Sementara itu, Tabel 2, menampilkan performa KNN tanpa PCA, dengan akurasi 55%, presisi 68%, *recall* 55%, dan F1-*Score* 55%.

Dari kedua tabel tersebut, terlihat bahwa akurasi tertinggi pada KNN diperoleh dengan menggunakan PCA, yaitu pada Tabel 1 dengan akurasi 61%.

Keuntungan utama KNN meliputi kemudahan pemahaman dan implementasi, serta tidak memerlukan asumsi khusus tentang distribusi data. Namun, KNN juga memiliki kelemahan seperti performa yang lambat pada data berukuran besar dan sensitivitas terhadap fitur yang tidak relevan atau berisik, yang dapat diatasi dengan menggunakan teknik seperti PCA untuk mengurangi dimensi data.

Gambar 2. Matriks evaluasi *K-Nearest Neighbors* (KNN) dengan PCA memperlihatkan matriks evaluasi hasil performa KNN dengan PCA, yang meliputi informasi mengenai *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN), yang membantu dalam analisis performa secara lebih mendetail.

	precision	recall	f1-score	support
Abdominal aortic aneurysm	1.00	0.56	0.71	9
Acute native valve endocarditis	0.10	1.00	0.19	10
Acute pericarditis	0.77	1.00	0.87	10
Aortic stenosis	1.00	1.00	1.00	9
Arrhythmogenic right ventricular cardiomyopathy	1.00	0.86	0.92	7
Candidal endocarditis	0.00	0.00	0.00	9
Chronic pericarditis	1.00	0.29	0.44	7
Constrictive pericarditis	1.00	0.67	0.80	9
Coronary artery disease (CAD)	0.00	0.00	0.00	6
Dilated cardiomyopathy	1.00	1.00	1.00	8
Hypertrophic cardiomyopathy	0.00	0.00	0.00	6
Infective endocarditis on prosthetic valve	0.00	0.00	0.00	6
Infrarenal aortic aneurysm	0.00	0.00	0.00	10
Mitral regurgitation	1.00	0.89	0.94	9
Mitral stenosis	0.90	1.00	0.95	9
Native valve endocarditis involving multiple valves	1.00	0.20	0.33	10
Pericardial effusion with tamponade	0.78	1.00	0.88	7
Pericarditis with myocarditis	1.00	1.00	1.00	4
Primary pulmonary hypertension	1.00	1.00	1.00	9
Proximal aortic aneurysm	0.00	0.00	0.00	9
Pulmonary stenosis	1.00	1.00	1.00	8
Restrictive cardiomyopathy	1.00	0.50	0.67	8
Secondary pulmonary hypertension due to COPD	1.00	0.70	0.82	10
Secondary pulmonary hypertension due to chronic thromboembolic disease	1.00	1.00	1.00	11
Secondary pulmonary hypertension due to obesity and sleep apnea	0.50	1.00	0.67	8
Secondary pulmonary hypertension due to scleroderma	1.00	0.50	0.67	10
Subacute native valve endocarditis	1.00	1.00	1.00	9
Takotsubo cardiomyopathy	1.00	0.38	0.55	8
Thoracic aortic aneurysm	0.00	0.00	0.00	6
Thoracoabdominal aortic aneurysm	0.00	0.00	0.00	9
Tricuspid stenosis	1.00	1.00	1.00	7
accuracy			0.61	257
macro avg	0.68	0.60	0.59	257
weighted avg	0.69	0.61	0.60	257

Gambar 2. Matrix Evaluasi *K-Nearest Neighbors* (KNN) dengan PCA

3.1.2. *Logistic Regression* (LR)

Tabel 3. Pengukuran *Logistic Regression* (LR) dengan PCA

Performa *Logistic Regression* (LR) PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
1.00	1.00	1.00	1.00

Tabel 4. Pengukuran *Logistic Regression* (LR) Non-PCA

Performa *Logistic Regression* (LR) Non-PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
91%	96%	91%	91%

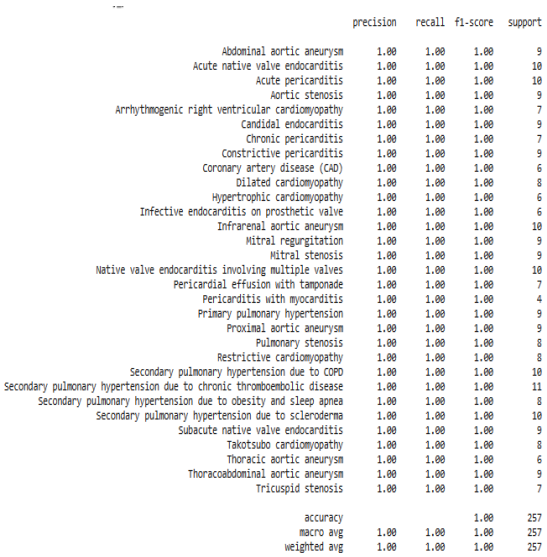
Dapat dilihat pada tabel 3 ini menunjukkan hasil evaluasi performa *Logistic Regression* yang menggunakan PCA. Dengan hasil Akurasi: 1.00, Presisi: 1.00, *Recall*: 1.00, dan F1-*Score*: 1.00.

Akan tetapi, pada Tabel 4. *Evaluasi Logistic Regression* (LR) Tanpa PCA. Tabel ini memperlihatkan evaluasi performa *Logistic Regression* tanpa menggunakan PCA. Dengan hasil nilai akurasi: 91%, Presisi: 96%, *Recall*: 91%, F1-*Score*: 92%.

Deskripsi pada Tabel 3 dan 4 menyatakan bahwa hasil terbaik dalam *Logistic Regression* (LR) diperoleh dari Tabel 3 dengan penerapan PCA, yang menghasilkan akurasi sebesar 1.00.

Dengan mengurangi dimensi, PCA dapat meningkatkan performa model, seperti yang terlihat pada perbandingan antara Tabel 3 dan Tabel 4. Berdasarkan hasil di atas, penerapan PCA pada *Logistic Regression* menunjukkan peningkatan signifikan dalam performa model.

Logistic Regression dapat mengalami *overfitting* jika terlalu banyak atau terlalu kompleks. Penggunaan PCA membantu mengurangi dimensi fitur, mengurangi risiko *overfitting*, dan meningkatkan generalisasi.



	precision	recall	f1-score	support
Abdominal aortic aneurysm	1.00	1.00	1.00	9
Acute native valve endocarditis	1.00	1.00	1.00	10
Acute pericarditis	1.00	1.00	1.00	10
Aortic stenosis	1.00	1.00	1.00	9
Arrhythmogenic right ventricular cardiomyopathy	1.00	1.00	1.00	7
Candidal endocarditis	1.00	1.00	1.00	9
Chronic pericarditis	1.00	1.00	1.00	7
Constrictive pericarditis	1.00	1.00	1.00	9
Coronary artery disease (CAD)	1.00	1.00	1.00	6
Dilated cardiomyopathy	1.00	1.00	1.00	8
Hypertrophic cardiomyopathy	1.00	1.00	1.00	6
Infective endocarditis on prosthetic valve	1.00	1.00	1.00	6
Infrarenal aortic aneurysm	1.00	1.00	1.00	10
Mitral regurgitation	1.00	1.00	1.00	9
Mitral stenosis	1.00	1.00	1.00	9
Native valve endocarditis involving multiple valves	1.00	1.00	1.00	10
Pericardial effusion with tamponade	1.00	1.00	1.00	7
Pericarditis with myocarditis	1.00	1.00	1.00	4
Primary pulmonary hypertension	1.00	1.00	1.00	9
Proximal aortic aneurysm	1.00	1.00	1.00	9
Pulmonary stenosis	1.00	1.00	1.00	8
Restrictive cardiomyopathy	1.00	1.00	1.00	8
Secondary pulmonary hypertension due to COPD	1.00	1.00	1.00	10
Secondary pulmonary hypertension due to chronic thromboembolic disease	1.00	1.00	1.00	11
Secondary pulmonary hypertension due to obesity and sleep apnea	1.00	1.00	1.00	8
Secondary pulmonary hypertension due to scleroderma	1.00	1.00	1.00	10
Subacute native valve endocarditis	1.00	1.00	1.00	9
Takotsubo cardiomyopathy	1.00	1.00	1.00	8
Thoracic aortic aneurysm	1.00	1.00	1.00	6
Thoracoabdominal aortic aneurysm	1.00	1.00	1.00	9
Tricuspid stenosis	1.00	1.00	1.00	7
accuracy			1.00	257
macro avg	1.00	1.00	1.00	257
weighted avg	1.00	1.00	1.00	257

Gambar 3. Matrix Evaluasi *Logistic Regression* (LR) dengan PCA

Pada Gambar 3 ini menampilkan matriks evaluasi untuk *Logistic Regression* yang menggunakan PCA. *Matriks* ini berguna untuk memvisualisasikan performa algoritma dengan menunjukkan hubungan antara prediksi yang benar dan yang salah untuk setiap kelas.

3.1.3. *Random Forest* (RF)

Tabel 5. Pengukuran *Random Forest* (RF) dengan PCA

Performa *Random Forest* (RF) PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
1.00	1.00	1.00	1.00

Tabel 6. Pengukuran *Random Forest* (RF) Non-PCA

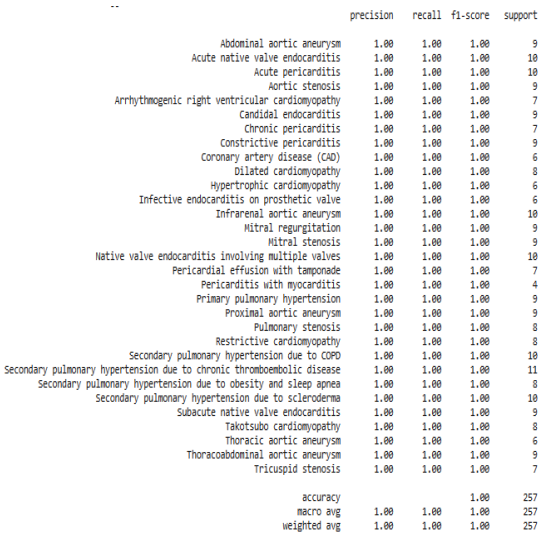
Performa *Random Forest* (RF) Non-PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
92%	93%	91%	91%

Dapat dilihat pada tabel 5 ini menunjukkan hasil evaluasi performa *Random Forest* yang menggunakan PCA. Dengan hasil Akurasi: 1.00, Presisi: 1.00, *Recall*: 1.00, dan F1-*Score*: 1.00.

Pada tabel 6. Tabel ini memperlihatkan evaluasi performa *Random Forest* tanpa menggunakan PCA, dengan hasil nilai akurasi: 91%, Presisi: 96%, *Recall*: 91%, F1-*Score*: 92%.

Deskripsi pada Tabel 5 dan 6 menyatakan bahwa hasil terbaik dalam *Random Forest* (RF) diperoleh dari Tabel 3 dengan penerapan PCA, yang menghasilkan akurasi sebesar 1.00

Dalam implementasi *Principal Component Analysis* (PCA), *Random Forest* dapat mengurangi jumlah fitur dengan menjaga fitur yang paling penting, yang dapat mengurangi waktu komputasi dan memori, membantu menghilangkan korelasi antar fitur, yang bisa memperbaiki performa model *Random Forest*, dan juga memilih komponen yang menjelaskan varians terbesar dalam data, yang bisa meningkatkan akurasi model.



	precision	recall	f1-score	support
Abdominal aortic aneurysm	1.00	1.00	1.00	9
Acute native valve endocarditis	1.00	1.00	1.00	10
Acute pericarditis	1.00	1.00	1.00	10
Aortic stenosis	1.00	1.00	1.00	9
Arrhythmogenic right ventricular cardiomyopathy	1.00	1.00	1.00	7
Candidal endocarditis	1.00	1.00	1.00	9
Chronic pericarditis	1.00	1.00	1.00	7
Constrictive pericarditis	1.00	1.00	1.00	9
Coronary artery disease (CAD)	1.00	1.00	1.00	6
Dilated cardiomyopathy	1.00	1.00	1.00	8
Hypertrophic cardiomyopathy	1.00	1.00	1.00	6
Infective endocarditis on prosthetic valve	1.00	1.00	1.00	6
Infrarenal aortic aneurysm	1.00	1.00	1.00	10
Mitral regurgitation	1.00	1.00	1.00	9
Mitral stenosis	1.00	1.00	1.00	9
Native valve endocarditis involving multiple valves	1.00	1.00	1.00	10
Pericardial effusion with tamponade	1.00	1.00	1.00	7
Pericarditis with myocarditis	1.00	1.00	1.00	4
Primary pulmonary hypertension	1.00	1.00	1.00	9
Proximal aortic aneurysm	1.00	1.00	1.00	9
Pulmonary stenosis	1.00	1.00	1.00	8
Restrictive cardiomyopathy	1.00	1.00	1.00	8
Secondary pulmonary hypertension due to COPD	1.00	1.00	1.00	10
Secondary pulmonary hypertension due to chronic thromboembolic disease	1.00	1.00	1.00	11
Secondary pulmonary hypertension due to obesity and sleep apnea	1.00	1.00	1.00	8
Secondary pulmonary hypertension due to scleroderma	1.00	1.00	1.00	10
Subacute native valve endocarditis	1.00	1.00	1.00	9
Takotsubo cardiomyopathy	1.00	1.00	1.00	8
Thoracic aortic aneurysm	1.00	1.00	1.00	6
Thoracoabdominal aortic aneurysm	1.00	1.00	1.00	9
Tricuspid stenosis	1.00	1.00	1.00	7
accuracy			1.00	257
macro avg	1.00	1.00	1.00	257
weighted avg	1.00	1.00	1.00	257

Gambar 4. Matrix Evaluasi *Random Forest* (RF) dengan PCA

Dapat dilihat dengan jelas pada Gambar 4 nilai akurasi dari *Random Forest* (RF) yang mengimplementasikan PCA mendapatkan nilai akurasi sebesar 100.

3.1.4. *Support Vector Machine* (SVM)

Tabel 7. Pengukuran Kinerja Algoritma *Support Vector Machine* (SVM) dengan PCA

Performa *Support Vector Machine* (SVM) PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
18%	14%	18%	14%

Tabel 8. Pengukuran Kinerja Algoritma *Support Vector Machine* (SVM) Non-PCA

Performa *Support Vector Machine* (SVM) Non-PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
81%	90%	82%	80%

Dapat dilihat pada tabel.7. Menunjukkan persentase prediksi yang benar dari total prediksi yang dilakukan. SVM tanpa PCA menunjukkan akurasi yang jauh lebih tinggi (81%) dibandingkan table 8. SVM dengan PCA (18%). Presisi: Mengukur ketepatan dari prediksi positif yang benar. Nilai presisi untuk SVM tanpa PCA adalah 90%, sedangkan untuk SVM dengan PCA adalah 14%. *Recall*: Mengukur seberapa baik model mendeteksi semua contoh positif. *Recall* untuk SVM tanpa PCA adalah 82%, dan untuk SVM dengan PCA adalah 18%. F1-*Score*: Merupakan rata-rata harmonis dari presisi dan *recall*. F1-*Score* untuk SVM tanpa PCA adalah 80%, sementara untuk SVM dengan PCA adalah 14%.

Dari tabel-tabel tersebut, dapat disimpulkan bahwa pada *dataset* ini, SVM tanpa PCA menunjukkan kinerja yang jauh lebih baik dibandingkan dengan SVM yang menggunakan PCA. Hal ini mungkin disebabkan oleh hilangnya informasi selama proses pengurangan dimensi dengan PCA atau karakteristik *dataset* yang lebih cocok untuk SVM tanpa pengurangan dimensi.

	precision	recall	f1-score	support
Abdominal aortic aneurysm	0.50	1.00	0.67	1
Acute native valve endocarditis	0.50	1.00	0.67	2
Acute pericarditis	0.50	1.00	0.67	1
Aortic stenosis	1.00	1.00	1.00	1
Arrhythmogenic right ventricular cardiomyopathy	1.00	1.00	1.00	3
Candidal endocarditis	1.00	0.33	0.50	3
Chronic pericarditis	1.00	0.67	0.80	3
Constrictive pericarditis	1.00	1.00	1.00	2
Coronary artery disease (CAD)	1.00	0.25	0.40	4
Dilated cardiomyopathy	1.00	1.00	1.00	2
Hypertrophic cardiomyopathy	1.00	1.00	1.00	4
Infective endocarditis on prosthetic valve	1.00	1.00	1.00	5
Infrarenal aortic aneurysm	0.00	0.00	0.00	1
Mitral regurgitation	1.00	1.00	1.00	1
Mitral stenosis	1.00	1.00	1.00	1
Native valve endocarditis involving multiple valves	1.00	1.00	1.00	1
Pericardial effusion with tamponade	1.00	1.00	1.00	1
Pericarditis with myocarditis	1.00	1.00	1.00	2
Primary pulmonary hypertension	1.00	0.50	0.67	2
Proximal aortic aneurysm	0.50	1.00	0.67	1
Pulmonary stenosis	1.00	1.00	1.00	2
Restrictive cardiomyopathy	1.00	1.00	1.00	2
Secondary pulmonary hypertension due to COPD	1.00	1.00	1.00	1
ry pulmonary hypertension due to chronic thromboembolic disease	0.00	0.00	0.00	1
Secondary pulmonary hypertension due to obesity and sleep apnea	0.75	1.00	0.86	3
Secondary pulmonary hypertension due to scleroderma	0.67	1.00	0.80	2
Subacute native valve endocarditis	1.00	1.00	1.00	3
Takotsubo cardiomyopathy	1.00	1.00	1.00	2
Thoracic aortic aneurysm	1.00	0.25	0.40	4
Thoracoabdominal aortic aneurysm	0.20	1.00	0.33	1
Tricuspid stenosis	1.00	1.00	1.00	3
accuracy			0.82	65
macro avg	0.83	0.84	0.79	65
weighted avg	0.90	0.82	0.80	65

Gambar 5. Matrix Evaluasi *Support Vector Machine* (SVM) Non-PCA

Dapat dilihat pada Gambar 5. Nilai akurasi dari SVM non-PCA dengan nilai akurasi sebesar 82%, *Precission* 90%, *Recall* 82%, dan F1- *Score* 80%.

3.1.5. *Decision Tree* (DT)

Tabel 9. Pengukuran *Decision Tree* (DT) dengan PCA

Performa *Decision Tree* (DT) PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
1.00	1.00	1.00	1.00

Tabel 10. Pengukuran *Decision Tree* (DT) Non-PCA

Performa *Decision Tree* (DT) Non- PCA			
Accuracy	***Precission***	***Recall***	***F1- Score***
89%	81%	89%	85%

Tabel 9. Kinerja *Decision Tree* (DT) dengan PCA, menampilkan hasil performa model *Decision Tree* yang menggunakan *Principal Component Analysis* (PCA). Hasilnya menunjukkan semua metrik performa seperti akurasi, precision, *recall*, dan F1-*Score* masing-masing bernilai 1.00.

Namun, pada Tabel 10. Kinerja *Decision Tree* (DT) tanpa PCA, menunjukkan performa model *Decision Tree* tanpa penerapan PCA dengan nilai akurasi yang dicapai adalah 89%, *precision* 81%, *recall* 97%, dan F1-*Score* 89%.

Akurasi tertinggi untuk model *Decision Tree* (DT) terlihat pada Tabel 9. Algoritma DT dengan penerapan PCA, di mana semua metrik performa mencapai 1.00. Hasil ini setara dengan performa model Logistic Regression (LR) dan Random Forest (RF).

Penggunaan PCA terbukti sangat efektif dalam meningkatkan performa model *Decision Tree*, yang terlihat dari perbandingan metrik performa antara model menggunakan PCA dengan dan tanpa PCA.

Model *Decision Tree* yang menggunakan PCA tidak hanya memiliki akurasi yang tinggi, tetapi juga menunjukkan kemampuan klasifikasi yang sempurna, berdasarkan hasil confusion matrix.

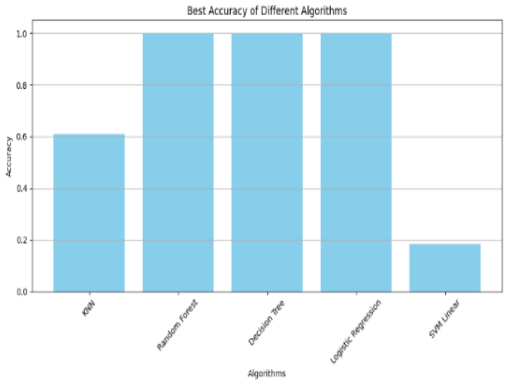
	precision	recall	f1-score	support
Abdominal aortic aneurysm	1.00	1.00	1.00	9
Acute native valve endocarditis	1.00	1.00	1.00	10
Acute pericarditis	1.00	1.00	1.00	10
Aortic stenosis	1.00	1.00	1.00	9
Arrhythmogenic right ventricular cardiomyopathy	1.00	1.00	1.00	7
Candidal endocarditis	1.00	1.00	1.00	9
Chronic pericarditis	1.00	1.00	1.00	7
Constrictive pericarditis	1.00	1.00	1.00	9
Coronary artery disease (CAD)	1.00	1.00	1.00	6
Dilated cardiomyopathy	1.00	1.00	1.00	8
Hypertrophic cardiomyopathy	1.00	1.00	1.00	6
Infective endocarditis on prosthetic valve	1.00	1.00	1.00	6
Infrarenal aortic aneurysm	1.00	1.00	1.00	10
Mitral regurgitation	1.00	1.00	1.00	9
Mitral stenosis	1.00	1.00	1.00	9
Native valve endocarditis involving multiple valves	1.00	1.00	1.00	10
Pericardial effusion with tamponade	1.00	1.00	1.00	7
Pericarditis with myocarditis	1.00	1.00	1.00	4
Primary pulmonary hypertension	1.00	1.00	1.00	9
Proximal aortic aneurysm	1.00	1.00	1.00	9
Pulmonary stenosis	1.00	1.00	1.00	8
Restrictive cardiomyopathy	1.00	1.00	1.00	8
Secondary pulmonary hypertension due to COPD	1.00	1.00	1.00	10
Secondary pulmonary hypertension due to chronic thromboembolic disease	1.00	1.00	1.00	11
Secondary pulmonary hypertension due to obesity and sleep apnea	1.00	1.00	1.00	8
Secondary pulmonary hypertension due to scleroderma	1.00	1.00	1.00	10
Subacute native valve endocarditis	1.00	1.00	1.00	9
Takotsubo cardiomyopathy	1.00	1.00	1.00	8
Thoracic aortic aneurysm	1.00	1.00	1.00	6
Thoracoabdominal aortic aneurysm	1.00	1.00	1.00	9
Tricuspid stenosis	1.00	1.00	1.00	7
accuracy			1.00	257
macro avg	1.00	1.00	1.00	257
weighted avg	1.00	1.00	1.00	257

Gambar 6. Matrix Evaluasi *Decision Tree* (DT) dengan PCA

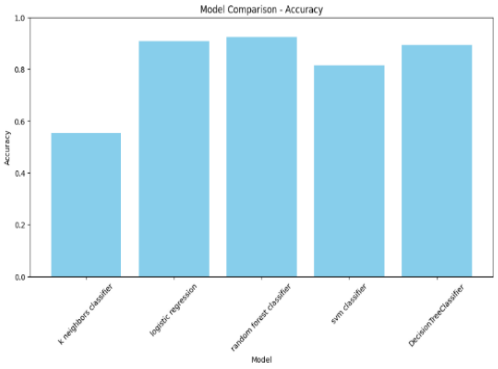
Terlihat jelas nilai akurasi dari DT yang memakai PCA dengan nilai akurasi sebesar 100, *Precission* 100, *Recall* 100, dan F1- *Score* 100.

Ini menunjukkan bahwa pengurangan dimensi menggunakan PCA sebelum penerapan algoritma klasifikasi dapat secara signifikan meningkatkan performa model.

Dari hasil laporan confusion matrix di atas, dapat disimpulkan bahwa:



Gambar 7. Diagram Hasil Akurasi Terbaik Menggunakan PCA



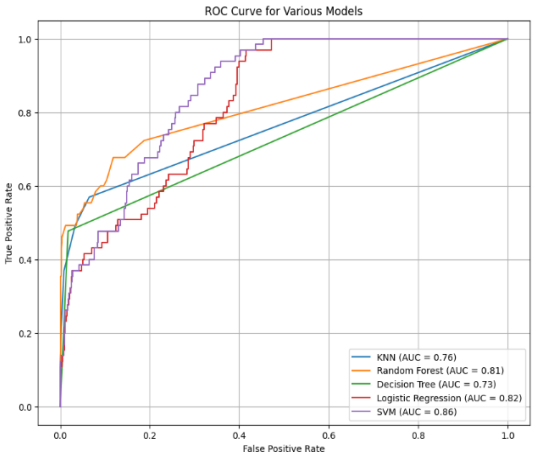
Gambar 8. Diagram Hasil Akurasi Terbaik non-PCA

Ada beberapa faktor yang mempengaruhi nilai atau tingkat akurasi diantaranya, yaitu:

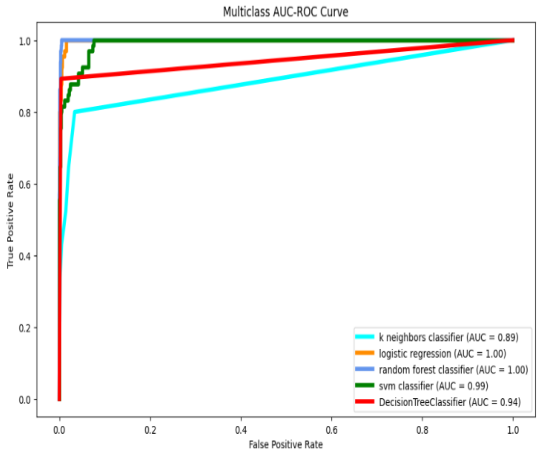
1. Informasi yang hilang: PCA bekerja dengan mengurangi dimensi data sambil mempertahankan sebanyak mungkin variasi.
2. Terjadinya *Overfitting*: Penggunaan PCA untuk mengurangi dimensi data berpotensi menimbulkan risiko *overfitting*, terutama jika hanya komponen utama yang dipertahankan dan komponen kecil yang mungkin penting untuk pemisahan kelas diabaikan [28].
3. Non-*linear Separability*: Jika setelah menggunakan PCA, data tidak terpisah secara linear dalam dimensi yang lebih rendah, contoh SVM mungkin tidak dapat memisahkan kelas secara efektif. SVM beroperasi dengan mencari hyperplane terbaik untuk pemisahan dalam ruang fitur, sehingga jika setelah reduksi dimensi, data tidak cukup terpisah secara linear, performa SVM dapat menurun [29].

3.2. Pengukuran kinerja *dengan Area under the curve* (AUC) & *Receiver operating characteristic* (ROC)

Setelah hasil evaluasi *confusion matrix* diketahui, dilanjutkan dengan evaluasi dengan AUC - ROC seperti yang diilustrasikan pada Gambar 2**.** *Multiclass AUC-ROC* dengan PCA dan Gambar 3. *Multiclass* AUC-ROC Non-PCA.



Gambar 9**.** *Multiclass* AUC-ROC dengan PCA



Gambar 10. *Multiclass* AUC-ROC Non-PCA

Dapat dilihat konvergensi antara semua model pengklasifikasi *machine learning*. Akurasi tertinggi akurasi tertinggi dengan PCA dimiliki oleh SVM dengan 86%, sedangkan akurasi terendah adalah *decision tree* yaitu 73%. Sementara akurasi tertinggi pada Non-PCA dimiliki oleh *logistic regression* dan *random forest* dengan hasil 1.00, sedangkan akurasi terendah dimiliki oleh KNN dengan nilai akurasi sebesar 89%.

4. DISKUSI

4.1. Penelitian Sebelumnya

Dari studi sebelumnya mengenai prediksi penyakit jantung menggunakan algoritma *machine learning*, sudah terbukti bahwa penggunaan model tersebut dapat memberikan prediksi yang akurat.. pada penelitian yang dilakukan oleh *Institute of Electrical and Electronics Engineers and Hindusthan Institute of Technology*, tahun 2020 menunjukan bahwa PCA dapat membantu untuk mengorganisir *dataset* dengan mengurangi dimensi *dataset* dan *random forest* akan membantu dalam klasifikasi nilai, tingkat akurasi sebesar 96.78%.

4.2. Interpretasi Hasil

Dengan demikian, dalam konteks penelitian ini, algoritma *machine learning* dapat efektif mengkasifikasikan penyakit jantung dengan tepat dan memiliki nilai akurasi yang memuaskan dengan membandingkan lima algoritma *machine learning* baik menggunakan PCA maupun tidak, hanya saja terjadi penurunan akurasi pada evaluasi *confusion matrix* algoritma SVM menggunakan PCA dikarenakan beberapa faktor, yaitu: informasi yang hilang, terjadinya *overfitting*, dan Non-*linear Separability*.

5. SARAN

Berdasarkan temuan penelitian ini, ada beberapa rekomendasi untuk pengembangan lebih lanjut dalam upaya prediksi dini penyakit jantung dengan menggunakan model *machine learning*, antara lain cara pengolahan data yang kurang maksimal dalam pembagian data training dan data testing. Butuh perbandingan reduksi data yang cara kerjanya hampir sama seperti PCA agar dapat dilihat seberapa baikkah nilai akurasi PCA dengan fitur reduksi data yang lain.

6. KESIMPULAN

Penelitian ini berhasil mengembangkan model *machine learning* yang mampu memprediksi penyakit jantung dengan akurasi tinggi berdasarkan gejala-gejala yang ada dalam *dataset*.. Model ini mencapai tingkat akurasi yang tinggi menggunakan fitur PCA pada algoritma DT,RF,dan LR sebesar 1.00. namun dalam evaluasi AUC akurasi terbaik diperoleh oleh algoritma SVM dengan nilai akuarasi mencapai 86%. Diharapkan dengan adanya model ini,

efisiensi dalam mendeteksi dan mencegah penyakit jantung sejak dini dapat meningkat.

DAFTAR PUSTAKA

[1] World Health Organization (WHO), "Cardiovascular diseases (CVDs)," World Health Organization (WHO).

[2] World Heart Federation (WHF), "World-Heart-Report-2023," *World Heart Federation (WHF)*, pp. 3–4, 2023.

[3] Kementerian Kesehatan RI, "Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer," Kementerian Kesehatan RI.

[4] D. Speyer, "Good PCA examples for teaching," Stack Exchange Inc.

[5] L. G. Kabari and B. B. Nwamae, "Principal Component Analysis (PCA) - An Effective Tool in Machine Learning," 2019. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

[6] H. J. Chiu, T. H. S. Li, and P. H. Kuo, "Breast cancer–detection system using PCA, multilayer perceptron, transfer learning, and support vector machine," *IEEE Access*, vol. 8, pp. 204309–204324, 2020, doi: 10.1109/ACCESS.2020.3036912.

[7] Institute of Electrical and Electronics Engineers and Hindusthan Institute of Technology, *Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) : 02-04, July 2020*.

[8] A. U. Haq *et al.*, "Detection of Breast Cancer through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques," *IEEE Access*, vol. 9, pp. 22090–22105, 2021, doi: 10.1109/ACCESS.2021.3055806.

[9] A. Gron, "Hands-on Machine Learning with Scikit-learn, Keras, and Tensorflow," [s.l.]: O'Reilly Media.

[10] J. Resti, N. Salsabilla Basuni, and A. Mutoi Siregar, "Comparison of the Accuracy of Drug User Classification Models Using Machine Learning Methods," vol. 5, p. 2026, doi: 10.29207/resti.v7ix.xxx.

[11] R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart Disease Prediction using Exploratory Data Analysis," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 130–139. doi: 10.1016/j.procs.2020.06.017.

[12] E. Elhaik, "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-14395-4.

[13] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput Sci*, vol. 1, no. 6, Nov. 2020, doi: 10.1007/s42979-020-00365-y.

[14] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[15] unud, "analisis data kategori menggunakan regresi logistik," sinta.unud.ac.id.

[16] M. Mia, A. F. N. Masruriyah, and A. R. Pratama, "The Utilization of Decision Tree Algorithm In Order to Predict Heart Disease," *JURNAL SISFOTEK GLOBAL*, vol. 12, no. 2, p. 138, Sep. 2022, doi: 10.38101/sisfotek.v12i2.551.

[17] Luthfiana Ratnawati dan Dwi Ratna Sulistyaningrum, "Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel," *JURNAL SAINS DAN SENI ITS Vol. 8, No. 2* , 2019.

[18] D. Cheng, Y. Shi, T. Lin, B. H. Gwee, and K. A. Toh, "Hybrid K-means clustering and support vector machine method for via and metal line detections in delayered IC images," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 12, pp. 1849–1853, Dec. 2018, doi: 10.1109/TCSII.2018.2827044.

[19] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-76635-9.

[20] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, 2018, doi: 10.1155/2018/3860146.

[21] Binus University School of Information System, "DECISION TREE ALGORITMA BESERTA CONTOHNYA PADA DATA MINING," Binus University School of Information System.

[22] T. T. Maskoen and D. Purnama, "Area Under the Curve dan Akurasi Cystatin C untuk Diagnosis Acute Kidney Injury pada Pasien Politrauma," *Majalah Kedokteran Bandung*, vol. 50, no. 4, pp. 259–264, Dec. 2018, doi: 10.15395/mkb.v50n4.1342.

[23] H. Yun, "Prediction model of algal blooms

using logistic regression and confusion matrix," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2407–2413, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2407-2413.

[24] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput Oper Res*, vol. 152, Apr. 2023, doi: 10.1016/j.cor.2022.106131.

[25] M. H. Z. Al Faroby, M. I. Irawan, and N. N. T. Puspaningsih, "XGBoost and Network Analysis for Prediction of Proteins Affecting Insulin based on Protein Protein Interactions," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 253–262, Nov. 2020, doi: 10.22219/kinetik.v5i4.1076.

[26] S. Narkhede, "Understanding AUC - ROC Curve," Understanding AUC - ROC Curve.

[27] A. J. Bowers and X. Zhou, "Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes," *J Educ Stud Placed Risk*, vol. 24, no. 1, pp. 20–46, Jan. 2019, doi: 10.1080/10824669.2018.1523734.

[28] T. Milo and A. Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, Jun. 2020, pp. 2617–2622. doi: 10.1145/3318464.3383126.

[29] Z. Mushtaq, A. Yaqub, A. Hassan, and S. Feng Su, "Performance Analysis of Supervised Classifiers using PCA based Techniques on Breast Cancer.".