p-ISSN 2798-964X
e-ISSN 2798-9658

Jurnal Ilmiah Teknik dan Manajemen Industri
Jurnal Taguchi
Vol. 3, No. 1, Juli, 2023 hal. 80-91
DOI Article : 10.46306/tgc.v3i1.58

ANALISIS PENGARUH PENGURANGAN DIMENSI DATA PADA KEAKURATAN PREDIKSI PENYAKIT JANTUNG DENGAN MENGGUNAKAN SVM LINEAR

Akhmad Ghiffary Budianto[1], Akhmad Syarief[2]
[1]Rekayasa Elektro, Fakultas Teknik, Universitas Lambung Mangkurat
[2]Teknik Mesin, Fakultas Teknik, Universitas Lambung Mangkurat
*Email: ghiffaryb04@gmail.com, akhmad.syarief@ulm.ac.id

Abstract

Heart disease is a disorder in the form of plaque that occurs in large blood vessels. This disrupts the supply of oxygen to the organs of the body. Heart disease is 1 of the 3 most common causes of death worldwide. Therefore, early detection based on the examination of medical data is needed to prevent the impact. The method used for classification is Support vector machine (SVM) and dimension reduction is Principal component analysis (PCA). The dataset is from Kaggle, medical records of 299 patients with 12 features and 1 label. The results obtained are the level of accuracy of PCA 6 features and without PCA both produce 82.9% and a total of 51 misclassifications. The processing time required is slightly longer for PCA 6 features (0.69121 seconds) than without PCA (0.46173 seconds). Because it has the same level of accuracy, the f-score metric is used to assess the classification model. The SVM with PCA 6 features has an f-score of 0.879, this is slightly better than SVM without PCA, which is 0.878.

Keywords: machine learning, PCA, SVM linear, heart disease, classification

Abstrak

Penyakit jantung merupakan penumpukan berupa plak yang terjadi pada pembuluh darah besar. Hal ini menyebabkan terganggunya suplai oksigen ke organ tubuh. Penyakit jantung menjadi 1 dari 3 penyebab kematian terbanyak di dunia. Oleh karena itu, deteksi dini berdasarkan pemeriksaan data medis diperlukan untuk mencegah dampaknya. Metode yang digunakan untuk klasifikasi yaitu Support vector machine (SVM) dan pengurangan dimensi yaitu Principal component analysis (PCA). Dataset yang digunakan yaitu dari Kaggle, rekam medis sebanyak 299 pasien dengan 12 fitur dan 1 label. Hasil yang didapat yaitu tingkat akurasi PCA 6 fitur dan tanpa PCA sama-sama menghasilkan 82.9% dan total misklasifikasi sebanyak 51. Lama komputasi yang diperlukan sedikit lebih lama untuk PCA 6 fitur (0.69121 detik) dibanding tanpa PCA (0.46173 detik). Karena memiliki tingkat akurasi yang sama, metrik f-score digunakan untuk menilai model klasifikasi. SVM dengan PCA 6 fitur memiliki f-score sebesar 0.879, hal ini sedikit lebih baik dibanding SVM tanpa PCA yaitu 0.878.

Kata kunci: machine learning, PCA, SVM linier, penyakit jantung, klasifikasi

PENDAHULUAN

Penyakit jantung merupakan gangguan yang terjadi pada sistem pembuluh darah besar. Hal ini menyebabkan peredaran darah dan jantung tidak berfungsi sebagai mana mestinya (Sutanto & Hernita, 2010). Penyakit jantung dianggap penyakit yang paling mengancam kesehatan masyarakat. Penyakit jantung juga menjadi masalah Kesehatan yang kritis disebabkan karena tingkat penderita dalam suatu populasi, tingkat kematian

penderita dan biaya perawatan yang tinggi. Dalam data yang dirilis *World Health Organization* (WHO), mortalitas atau tingkat kematian yang disebabkan penyakit jantung mencapai angka 17,8 Juta kematian atau satu dari tiga kematian di dunia disebabkan oleh penyakit jantung. Sedangkan di Indonesia, Data Riset Kesehatan Dasar (Riskesdas) tahun 2013-2018 menunjukkan penyakit jantung memiliki tren meningkat dari 0,5% menjadi 1,5%. Untuk biaya perawatan dengan BPJS, penyakit jantung menjadi klaim pembiayaan terbesar pada tahun 2021 yaitu sebesar Rp. 7,7 Triliun (Kementrian Kesehatan, 2022)

Masalah-masalah yang muncul akibat penyakit jantung dapat dicegah dan diminimalkan dengan adanya deteksi dini. Deteksi dini dengan menggunakan data rekam medis terbukti secara positif mampu mencegah risiko yang muncul dikemudian hari (Wismarini, 2022). Masyarakat yang masuk dalam kategori berisiko terkena penyakit jantung sebaiknya rutin melakukan *medical check-up*. *Medical check-up* yang dilakukan menggunakan analisis darah memiliki beberapa variabel yang berkaitan erat dengan penyakit jantung. Pada beberapa variabel tersebut antara lain gula darah, kadar kolesterol dan tekanan darah. Indikator dari beberapa variabel tersebut yang nantinya digunakan untuk menilai apakah orang tersebut berisiko terkena penyakit jantung atau tidak. Hasil penelitian terbaru di Korea Selatan pada tahun 2004-2014, tren dari penyakit jantung meningkat pada usia tertentu. Selain itu, penyakit jantung juga disertai dengan penyakit utama seperti hipertensi, dislipedemia dan diabetes (Cho et al., 2022). Sedangkan menurut *the American Heart Association*, penyakit jantung memiliki kaitan erat terhadap kebiasaan hidup seperti merokok atau tidak, olahraga atau aktivitas fisik secara rutin, nutrisi, berat badan dan obesitas. Untuk faktor risiko penyakit jantung dapat dilihat dari tingkat kolesterol darah, tekanan darah tinggi, diabetes, sindrom metabolisme dan keteraturan waktu tidur (Tsao et al., 2023).

Banyaknya variabel dan besarnya ukuran data dalam deteksi penyakit jantung menjadi tantangan dalam proses prediksi menggunakan *machine learning*. Kedua hal tersebut berkaitan dengan lamanya proses *training data* dan tingkat keakurasian dari hasil prediksi. Reduksi data digunakan untuk mengatasi masalah tersebut, dimana variabel data tidak perlu dihilangkan namun hanya perlu dikombinasikan. Sehingga tidak menghilangkan informasi yang dianggap penting dari sebuah data (Eliyanto & Suparman, 2020). Salah satu metode reduksi data yaitu PCA (*Principal component analysis*). Metode PCA ini umum digunakan untuk mereduksi data pada bidang kesehatan diantaranya reduksi data untuk penyakit kanker (Sirait, Adiwijaya, & Astuti, 2019)(Wibawa & Maysanjaya, 2018), penyakit diabetes (Dinanti & Purwadi, 2023), dan penyakit jantung (Utomo & Mesran, 2020).

Selain dari ukuran data dan banyaknya variabel, pemilihan algoritma yang tepat dalam klasifikasi juga mempengaruhi keakuratan dan lama waktu komputasi pada *machine learning*. Salah satu teknik yang banyak digunakan dalam klasifikasi pada bidang kesehatan adalah SVM (*Support vector machine*). Penggunaan SVM untuk klasifikasi penyakit aritmia melalui sinyal EKG (Ramadhani, Adiwijaya, & Utama, 2018) dan klasifikasi pada tumor otak pada citra MRI (Febrianti, Sardjono, & Babgei, 2020). Dari kedua penelitian tersebut, akurasi pada klasifikasi penyakit memiliki keakuratan diatas 80%. Penelitian terbaru menunjukkan reduksi dimensi data menggunakan PCA dan klasifikasi menggunakan SVM menghasilkan akurasi rata-rata 81.57% pada deteksi kanker (Sirait et al., 2019). Namun penelitian tersebut belum menunjukkan perbedaan lama waktu komputasi.

Penelitian ini bertujuan untuk membuat *machine learning* untuk klasifikasi penyakit jantung dengan mereduksi dimensi variabel data yang digunakan sehingga dapat mempersingkat waktu komputasi tetapi tanpa mengurangi keakurasian dari hasil model klasifikasi. Oleh karena itu, penggunaan PCA sebagai teknik reduksi data dan SVM linear diharapkan mampu menjawab permasalahan diatas.

p-ISSN 2798-964X
e-ISSN 2798-9658

STUDI KEPUSTAKAAN

Penyakit Jantung

Jantung merupakan organ vital dalam tubuh yang berfungsi untuk memompa darah ke seluruh tubuh. Gangguan yang terjadi pada jantung akan menyebabkan masalah-masalah terhadap organ-organ lain. Penyakit jantung memiliki berbagai macam jenis yaitu:

a. Penyakit jantung koroner yaitu jenis penyakit yang terjadi karena adanya penyumbatan pada pembuluh darah arteri disebabkan penumpukan plak. Penumpukan plak pada pembuluh darah arteri ini menyebabkan terganggunya suplai oksigen ke seluruh tubuh (Ghani, Susilawati, & Novriani, 2016).
b. Penyakit jantung bawaan yaitu jenis penyakit yang terjadi karena ada kelainan pada fungsi atau struktur jantung yang dimiliki sejak lahir. Salah satu jenis penyakit ini adalah aritmia. Aritmia merupakan kelainan pada irama jantung. Aritmia bisa dideteksi melalui HRV (*Heart Rate Variability*). Sistem deteksi dini untuk aritmia juga bisa dilakukan dengan cara mengkaslifikasikan denyut jantung ke dalam pola yang normal dan aritmia (Bazudewa, Satwika, & Juliharta, 2020). Data yang digunakan yaitu dengan merekam EKG (Electrocardiogram) dari pasien.

Penelitian ini berfokus pada penyakit jantung koroner, karena penyakit jantung koroner umumnya disebabkan oleh gaya hidup yang tidak sehat. Adapun beberapa faktor risiko dominan yang menyebabkan penyakit jantung koroner ini adalah hipertensi, gangguan mental emosional, diabetes melitus, stroke, usia ≥40 tahun, kebiasaan merokok, jenis kelamin perempuan, tingkat pendidikan rendah, obesitas sentral, dan status sosial ekonomi rendah (Ghani et al., 2016).

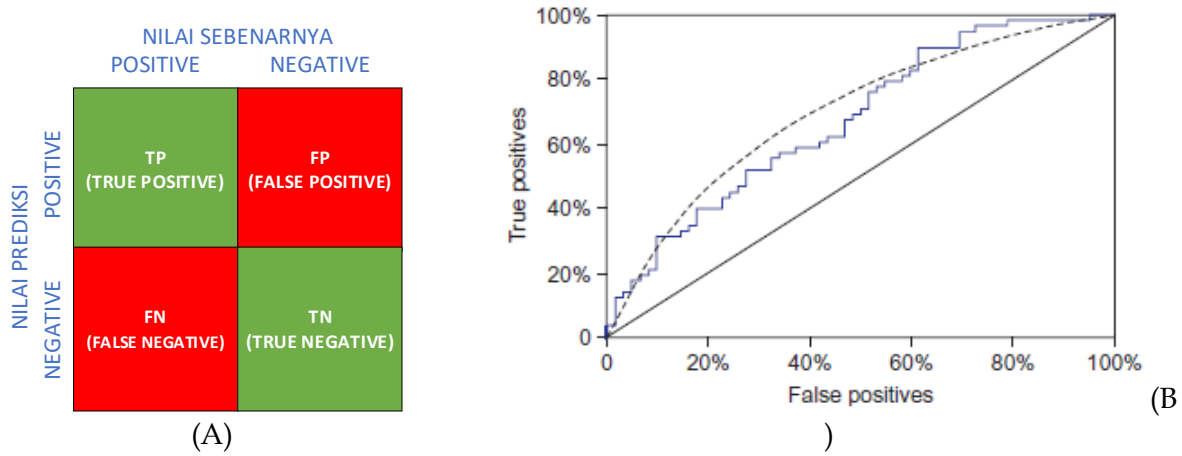
Machine learning untuk Klasifikasi

LINGO *Machine learning* adalah suatu area dalam Artificial Intelligence (AI) yang berhubungan dengan pengembangan teknik yang bisa diprogramkan dan belajar dari data masa lalu. *Machine learning* identik dengan pengenalan pola dan digunakan sebagai alat analisis dalam *data mining* (Santosa & Umam, n.d.). *Machine learning* terbagi atas dua pendekatan yaitu *Unsupervised* dan *Supervised Machine learning*. *Supervised machine learning* memiliki label dalam proses training dan *testing data* sedangkan *Unsupervised* sebaliknya tidak memiliki label dalam proses training.

Klasifikasi merupakan bagian dari *Supervised Machine learning*. Klasifikasi melakukan pengelompokan obyek berdasarkan kelompok yang sudah ada (sesuai dengan label) saat data dilakukan proses training. Klasifikasi memiliki nilai output dalam bentuk nilai diskrit (Santosa & Umam, n.d.).

Model klasifikasi pada kasus yang terbentuk dapat diukur performansinya dengan menggunakan *confusion matrix* dan ROC (*Receiver Operating Curve*) seperti pada Gambar 1 berikut:

p-ISSN 2798-964X
e-ISSN 2798-9658



(A) (B)

Gambar 1. Contoh *confusion matrix* (A) dan ROC *Curve* (B) (Witten Ian, Eibe, Hall Mark, & Pal Christopher, 2017)

Confusion matrix dapat dibuat dengan membandingkan antara nilai prediksi dengan nilai sebenarnya. Hasil yang baik tentunya yang memiliki nilai besar pada TP dan TN dan nilai kecil pada FP dan FN. *Confusion matrix* juga dapat digunakan untuk evaluasi dari klasifikasi (Santosa & Umam, n.d.) pada Tabel 1 sebagai berikut:

Tabel 1. Karakteristik dataset penyakit jantung

evaluasi	persamaan	deskripsi
Akurasi	(TN/TP)/jumlah data	Tingkat akurasi dari model klasifikasi
Miss classification rate/ error rate	(FP/FN)/jumlah data	Tingkat kesalahan dari model klasifikasi
Recall	TP/jumlah aktual positif	Proporsi positif aktual yang diidentifikasi dengan benar
Presisi	TP/jumlah prediksi positif	Proporsi identifikasi positif yang sebenarnya benar dari prediksi
False alarm rate	FP/jumlah aktual negatif	Proporsi positif palsu dari nilai sebenarnya pada model
Specificity	TN/jumlah aktual negatif	Proporsi nilai negatif sebenarnya dari model

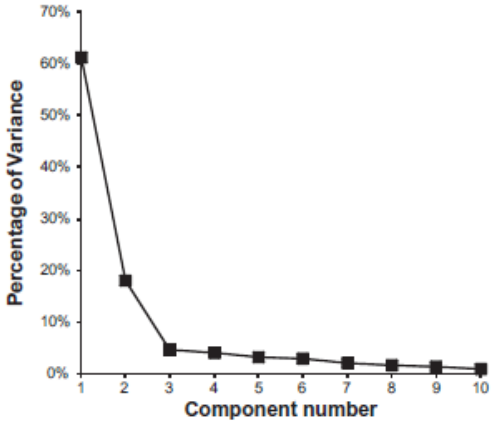
Kurva ROC menunjukkan *trade-off* antara *recall* dan *false alarm rate*, dimana untuk sumbu vertikal memplot nilai *True positive* dan sumbu horizontal memplot nilai *False positive*. Nilai *recall* yang tinggi (mendekati 1) adalah hal yang bagus, tetapi nilai *false positive rate* yang tinggi adalah hal yang tidak bagus. Idealnya model klasifikasi yang terbentuk yaitu nilai *recall* yang sebesar mungkin (mendekati 1) dan nilai *false positive rate* yang sekecil mungkin (mendekati 0) (Santosa & Umam, n.d.). Kurva ROC memiliki nilai area dibawah

p-ISSN 2798-964X
e-ISSN 2798-9658

kurva / *Area under curve* (AUC). AUC adalah ukuran kualitas keseluruhan pengklasifikasi. Nilai AUC yang lebih besar menunjukkan kinerja model klasifikasi yang lebih baik.

PCA (*Principal component analysis*)

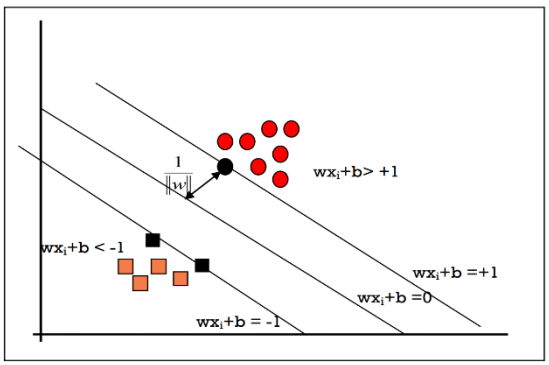
Principal component analysis (PCA) melakukan tranformasi data secara linier menjadi ruang dimensi yang lebih sedikit. Karena proses komputasi untuk transformasi data dengan menggunakan *matrix* yang terdiri dari *eigen vector* dari matriks *covariace* tentu memerlukan waktu yang lebih lama. Pada kumpulan data yang bersifat numerik, biasanya PCA digunakan sebelum *data mining* sebagai bentuk pembersihan data dan pengurangan dimensi (Witten Ian et al., 2017). Sebagai contoh pengaplikasian pengurangan dimensi menggunakan proporsi 95% varians dari data. Sehingga pada Gambar 2, data pada 10 dimensi dapat terwakili pada 7 dimensi sesuai dengan 95% kumulatif varians yang digunakan.



Gambar 2. Penggunaan PCA sesuai dengan nilai varians dari data (Witten Ian et al., 2017)

SVM (*Support vector machine*) Linier

Support vector machine (SVM) merupakan salah satu metode untuk *Supervised Machine learning*. SVM bekerja dengan cara menemukan fungsi pemisah (klasifier/hyperplane) yang optimal agar bisa memisahkan data dari dua jenis kelas yang berbeda (Vapnik, 1995). SVM dapat memberikan hasil klasifikasi yang lebih baik dari ANN. Hal ini disebabkan ANN menemukan solusi local optimal sedangkan SVM dapat menemukan solusi yang global optimal (Santosa & Umam, n.d.). Gambar 3 menunjukkan bagaimana cara SVM melakukan klasifikasi terhadap dua data berbeda dengan memaksimalkan hyperplane/klasifier yang terbentuk linier.

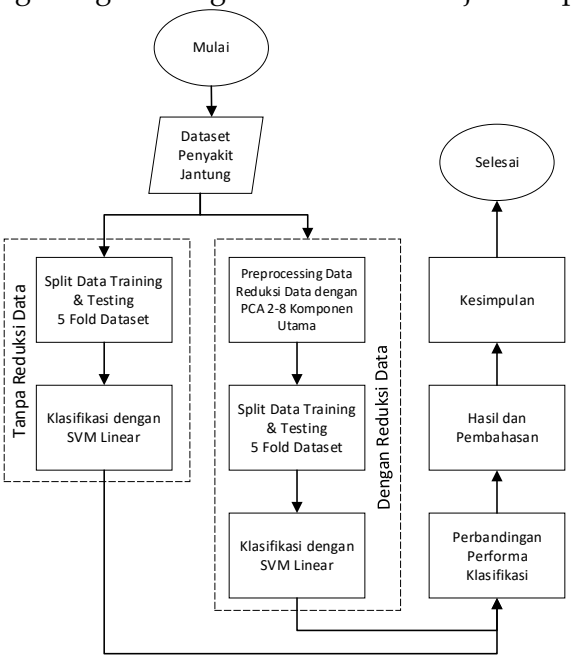


Gambar 3. Pencarian fungsi pemisah yang optimal secara linier (Santosa & Umam, n.d.)

p-ISSN 2798-964X
e-ISSN 2798-9658

METODE PENELITIAN

Dalam penelitian ini, ada beberapa langkah yang dilakukan untuk mengetahui pengaruh dari pengurangan dimensi data terhadap akurasi dan waktu pemrosesan dari klasifikasi penyakit jantung. Langkah-langkah tersebut ditunjukkan pada Gambar 4.



Gambar 4. Diagram Alir Penelitian

Dataset Penyakit Jantung

Model Proses awal yang dilakukan pada penelitian ini adalah pengumpulan dataset penyakit jantung. Dataset penyakit jantung yang digunakan dataset sekunder dari Kaggle (Chicco & Jurman, 2020). Dataset yang dikumpulkan berasal dari 299 pasien penyakit jantung di *the Allied Hospital in Faisalabad* (Punjab, Pakistan) selama April – Desember 2015. Adapun karakteristik dari dateset yang digunakan dapat dilihat pada Tabel 2 dan 3 berikut:

Tabel 2. Karakteristik dataset penyakit jantung

No.	Atribut	Tipe Data	Range Data
1	*age*	Numerik	[40, ..., 95]
2	*anaemia*	Biner	0, 1
3	*High blood pressure*	Biner	0, 1
4	*Creatinine phosphokinase (CPK)*	Numerik	[23, ..., 7861]
5	*Diabetes*	Biner	0, 1
6	*Ejection fraction*	Numerik	[14, ..., 80]
7	*Sex*	Biner	0, 1
8	*Platelets*	Numerik	[25.01, ..., 850.00]
9	*Serum creatinine*	Numerik	[0.50, ..., 9.40]

p-ISSN 2798-964X
e-ISSN 2798-9658

No.	Atribut	Tipe Data	Range Data
10	*Serum sodium*	Numerik	[114, ..., 148]
11	*Smoking*	Biner	0, 1
12	*Time*	Numerik	[4, ..., 285]
13	*Target (death event)*	Biner	0, 1

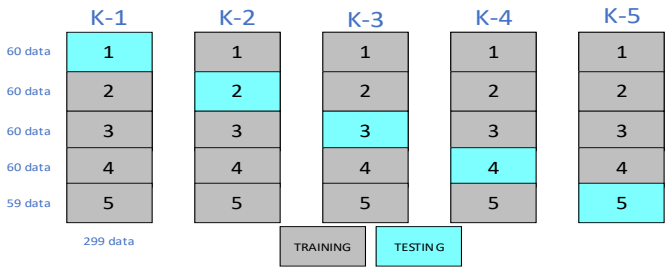
Tabel 3. Deskripsi dataset penyakit jantung

No.	Atribut	Deskripsi
1	*age*	Umur dari pasien
2	*anaemia*	Kurangnya sel darah merah atau hemoglobin, 0 jika kurang sel darah merah dan 1 jika sebaliknya
3	*High blood pressure*	1 jika pasien memiliki darah tinggi dan 0 jika sebaliknya
4	*Creatinine phosphokinase (CPK)*	Tingkat enzim CPK dalam darah
5	*Diabetes*	1 jika pasien memiliki diabetes dan 0 jika sebaliknya
6	*Ejection fraction*	Persentase darah meninggalkan jantung setiap kali kontraksi
7	*Sex*	1 jika laki-laki dan 0 jika perempuan
8	*Platelets*	Platelets dalam darah
9	*Serum creatinine*	Tingkat creatinine dalam darah
10	*Serum sodium*	Tingkat sodium dalam darah
11	*Smoking*	1 jika pasien merokok dan 0 jika tidak merokok
12	*Time*	Waktu pendataan terhadap pasien
13	*Target (death event)*	1 jika pasien meninggal selama waktu pengamatan dan 0 jika sebaliknya

Pengolahan Data dan Validasi Model Klasifikasi

Proses pengolahan data dilakukan dengan menggunakan software MATLAB R2020a pada PC dengan prosesor Intel Core i7-7700HQ @2.8GHz, RAM 16 GB DDR4 dan sistem operasi Windows 64-bit. Algoritma yang digunakan untuk klasifikasi yaitu dengan *Support vector machine* (SVM) kernel linier. Dan untuk pengurangan dimensi dengan PCA (*Principal component analysis*) yang mewakili 95% variansi dari data.

p-ISSN 2798-964X
e-ISSN 2798-9658



Gambar 5. k-cross validation (k=5)

Gambar 5 menunjukkan metode untuk validasi pada model klasifikasi. K-cross validation dimana nilai k=5, hal ini berarti membagi data menjadi 5 bagian data menjadi 4 bagian untuk data training dan 1 data testing. Sehingga jika data berjumlah 299, maka akan ada 1 bagian data yang berisikan 59 data dan 4 bagian data lainnya berisikan masing-masing 60 data. Pembagian data (*data splitting*) mengikuti aturan 80% untuk *data training* dan 20% untuk *data testing*. Setelah melalui proses validasi, hasil model klasifikasi selanjutnya akan diukur kinerjanya dengan metrik evaluasi yaitu *confusion matrix* dan ROC serta membandingkan nilai akurasi dan waktu *training time*.

HASIL DAN PEMBAHASAN

Hasil kinerja dari model klasifikasi yang dibuat dengan menggunakan SVM linier pada dataset yang dilakukan pengurangan dimensi menggunakan PCA dan tanpa PCA dapat dilihat pada Tabel 4.

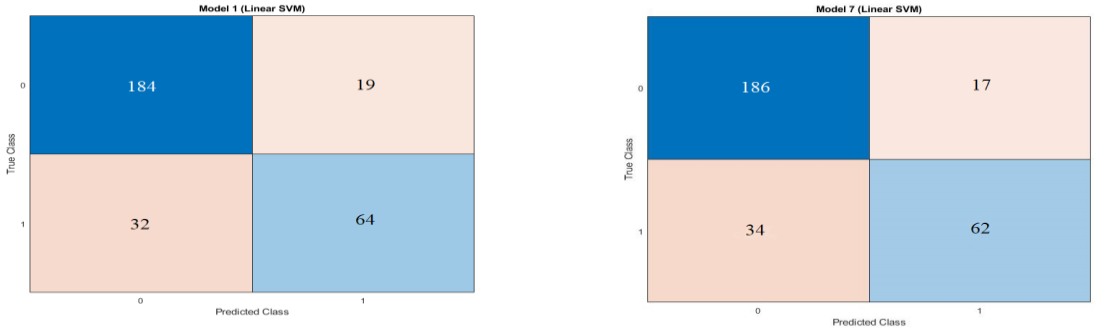
Tabel 4. Kinerja klasifikasi menggunakan PCA dan tanpa PCA

Jumlah Fitur PCA	TP	FP	TN	FN	Akurasi	*precision*	*recall*	*f-score*
1	203	0	0	96	67.9%	100.00%	67.89%	0.809
2	203	0	0	96	67.9%	100.00%	67.89%	0.809
3	184	19	57	39	80.6%	90.64%	82.51%	0.864
4	180	23	58	38	79.6%	88.67%	82.57%	0.855
5	184	18	62	34	82.3%	91.09%	84.40%	0.876
6	186	17	62	34	82.9%	91.63%	84.55%	0.879
7	184	19	63	33	82.6%	90.64%	84.79%	0.876
Tanpa PCA	184	19	64	32	82.9%	90.64%	85.19%	0.878

Dari Tabel 4 diatas, penggunaan PCA 6 fitur dan SVM linier serta tanpa PCA dan SVM linier sama-sama menghasilkan tingkat akurasi klasifikasi sebesar 82.9%. Hanya saja yang menjadi perbedaan pada penggunaan PCA 6 fitur metrik *precision* dan *f-score* sedikit lebih tinggi daripada tanpa PCA. Hal ini berarti penggunaan PCA 6 fitur dapat

p-ISSN 2798-964X
e-ISSN 2798-9658

memprediksi penderita penyakit jantung yang benar-benar memiliki penyakit jantung sebesar 91.63%, sedangkan tanpa PCA hanya 90.64%. Nilai metrik *precision* mendekati 1 menandakan model klasifikasi sangat baik karna memiliki sedikit positif palsu (*False positive*) saat dilakukan pengujian dataset. Untuk metrik *recall*, pengolahan data tanpa PCA (85.19%) menghasilkan nilai sedikit lebih baik daripada dengan PCA 6 fitur (84.55%). Hal ini menandakan model klasifikasi yang dibentuk dari tanpa PCA menghasilkan model klasifikasi yang baik karna bisa meminimalkan prediksi negatif palsu (*False negative*) dibandingkan dengan PCA 6 fitur. Hubungan antara metrik *precision* dan *recall* memiliki sifat "jungkat-jungkit". Saat nilai metriks *precision* tinggi (100%) tapi nilai metriks *recall* rendah (67.89%), sedangkan saat nilai metiks *precision* rendah (90.64%) tapi nilai metriks *recall* tinggi (85.19%). Oleh karena itu, metrik *f-score* digunakan untuk mencari penengah antara kedua metrik tersebut. Dalam penelitian ini, nilai metrik *f-score* tertinggi yaitu 0.879 pada PCA 6 fitur. Hal ini hanya sedikit lebih baik daripada *f-score* tanpa PCA yaitu 0.878.



(a) Tanpa PCA

(b) 6 PCA

Gambar 6. Perbandingan *confusion matrix* dengan 6 PCA dan tanpa PCA

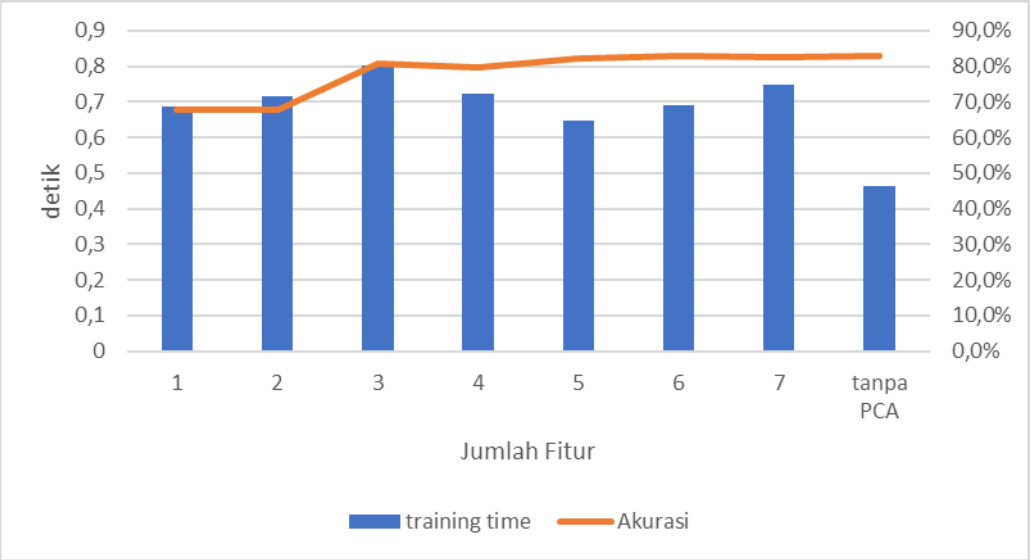
Gambar 6 menunjukkan perbandingan kinerja model klasifikasi SVM linier tanpa PCA dan dengan PCA 6 fitur. Penggunaan PCA 6 fitur dipilih karena memiliki nilai *f-score* yang sedikit lebih baik daripada tanpa PCA. Dari *confusion matrix*, jumlah *False positive* 6 PCA (17) lebih sedikit dibandingkan tanpa PCA (19) dan *True positive* 6 PCA (186) lebih banyak dibandingkan tanpa PCA (184). Sehingga menghasilkan nilai *precision* 6 PCA lebih tinggi dibandingkan tanpa PCA. Sedangkan jumlah *False negative* tanpa PCA (32) lebih sedikit dibandingkan PCA 6 fitur (34). Ini yang menyebabkan nilai metrik *recall* tanpa PCA sedikit lebih baik daripada PCA 6 fitur.

Tabel 5. Perbandingan misklasifikasi menggunakan PCA dan tanpa PCA dengan *training time*

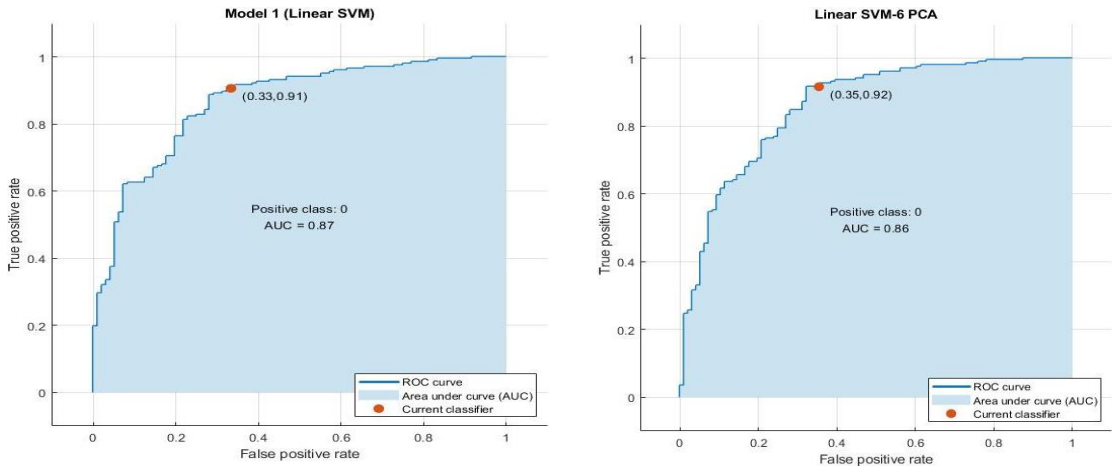
Jumlah Fitur PCA	total misklasifikasi	*training time*	AUC
1	96	0.68786	0.49
2	96	0.71714	0.48
3	58	0.80408	0.82
4	61	0.72265	0.81
5	53	0.64809	0.86
6	51	0.69121	0.86
7	52	0.7483	0.87
tanpa PCA	51	0.46173	0.87

p-ISSN 2798-964X
e-ISSN 2798-9658

Pada Tabel 5 menegaskan terkait akurasi yang dihasilkan tanpa PCA dan PCA 6 fitur sama-sama menghasilkan jumlah kesalahan dalam klasifikasi (misklasifikasi) sebanyak 51. Tetapi yang membedakan tentunya terkait waktu pelatihan model klasifikasi. PCA 6 fitur karena perlu sedikit komputasi terhadap *principal component* yang dianggap mampu mewakili 95% variansi dari total data yang digunakan, tentunya memerlukan waktu komputasi yang sedikit lebih lama dibandingkan dengan model klasifikasi tanpa PCA. Perbandingan antara waktu *training* model klasifikasi dengan tingkat akurasi dapat dilihat pada Gambar 7.



Gambar 7. Perbandingan *training time* dan akurasi dengan PCA dan tanpa PCA



(a) Tanpa PCA (b) 6 PCA

Gambar 8. Perbandingan ROC *Curve* dengan 6 PCA dan tanpa PCA

Indikator terakhir yang digunakan yaitu ROC *Curve* dan nilai dari AUC. Pada Gambar 8, Kurva ROC menunjukkan *trade-off* antara *recall* dan *false alarm rate*, dimana untuk sumbu vertikal memplot nilai *True positive* dan sumbu horizontal memplot nilai *False positive*. Tanpa PCA menghasilkan nilai *false positive rate* 0.33 dan *true positive rate* 0.91 sedangkan PCA 6 fitur menghasilkan nilai *false positive rate* 0.35 dan *true positive rate* 0.92.

Hal ini berarti PCA 6 fitur mampu memprediksi penderita penyakit jantung sedikit lebih baik daripada tanpa PCA, tetapi dengan catatan kesalahan dalam memprediksi nilai positif palsu yang juga lebih besar daripada tanpa PCA. Untuk nilai AUC, tanpa PCA (0.87) dan PCA 6 fitur (0.86) hanya beda tipis. Tentunya nilai AUC yang mendekati 1 menunjukkan bahwa model klasifikasi yang dibentuk lebih baik dalam melakukan klasifikasi saat pengujian dataset.

KESIMPULAN

Dari hasil penelitian yang telah dilakukan dapat diambil kesimpulan sebagai berikut:

1. Model klasifikasi dengan menggunakan SVM linier tanpa PCA dan PCA 6 fitur sama-sama menghasilkan tingkat akurasi sebesar 82.9%.
2. Model klasifikasi tanpa PCA mampu memprediksi penderita penyakit jantung dengan meminimalkan prediksi negatif palsu (*False negative*) dibandingkan dengan PCA 6 fitur. Hal ini dapat dilihat dari nilai metrik *recall* yaitu dengan PCA 6 fitur (84.55%) dan tanpa PCA (85.19%).
3. Model klasifikasi dengan PCA 6 fitur dapat memprediksi penderita penyakit jantung yang benar-benar memiliki penyakit jantung sebesar 91.63%, sedangkan tanpa PCA hanya 90.64%. Nilai metrik *precision* mendekati 1 menandakan model klasifikasi sangat baik karna memiliki sedikit positif palsu (*False positive*) saat dilakukan pengujian dataset
4. Hubungan antara metrik *precision* dan *recall* memiliki sifat "jungkat-jungkit", sehingga model klasifikasi yang ideal perlu mencari nilai yang sama-sama tinggi. Pada model klasifikasi ini, *f-score* menjadi nilai yang menjadi indikator penengah antara metrik *precision* dan *recall*. Nilai metrik *f-score* tertinggi yaitu 0.879 pada PCA 6 fitur. Hal ini hanya sedikit lebih baik daripada *f-score* tanpa PCA yaitu 0.878.
5. Tingkat akurasi yang dihasilkan sama, sehingga jumlah misklasifikasi juga sama-sama berjumlah 51. Tetapi waktu yang dilakukan untuk proses komputasi PCA 6 fitur (0.69121 detik) sedikit lebih lama dibandingkan dengan komputasi tanpa PCA (0.46173 detik). Hal ini disebabkan karena pengurangan jumlah fitur yang digunakan tidak terlalu banyak dalam fitur yang digunakan. Pada penelitian ini fitur yang digunakan ada sebanyak 12 dan 1 untuk label klasifikasi. Sedangkan pengurangan fitur yang terbaik yaitu sebanyak 6 PCA.

DAFTAR PUSTAKA

Bazudewa, W. R., Satwika, I. P., & Juliharta, I. G. P. K. (2020). KLASIFIKASI ARITMIA DENGAN HEART RATE VARIABILITY ANALISIS MENGGUNAKAN METODE BACKPROPAGATION. *Jurnal Informatika Dan Rekayasa Elektronik*, *3*(1), 1–10.

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, *20*(1), 1–16.

Cho, D.-H., Lee, C. J., Son, J.-W., Choi, J., Hwang, J., & Yoo, B.-S. (2022). Temporal trends in heart failure over 11 years in the aging Korean population: A retrospective study using the national health insurance database. *Plos One*, *17*(12), e0279541.

Dinanti, A., & Purwadi, J. (2023). Analisis Performa Algoritma K-Nearest Neighbor dan Reduksi Dimensi Menggunakan Principal Component Analysis. *Jambura Journal of Mathematics*, *5*(1), 155–165.

Eliyanto, J., & Suparman, S. (2020). Reduksi Dimensi untuk Meningkatkan Performa Metode Fuzzy Klastering pada Big Data. *Science, Technology, Engineering, Economics, Education, and Mathematics*, *1*(1).

Febrianti, A. S., Sardjono, T. A., & Babgei, A. F. (2020). Klasifikasi Tumor Otak pada Citra

Magnetic Resonance Image dengan Menggunakan Metode Support Vector Machine. *Jurnal Teknik ITS*, *9*(1), A118–A123.

Ghani, L., Susilawati, M. D., & Novriani, H. (2016). Faktor risiko dominan penyakit jantung koroner di Indonesia. *Buletin Penelitian Kesehatan*, *44*(3), 153–164.

Kementrian Kesehatan. (2022). Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer. Retrieved May 22, 2023, from https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220929/0541166/penyakit-jantung-penyebab-utama-kematian-kemenkes-perkuat-layanan-primer/

Ramadhani, G. T., Adiwijaya, A., & Utama, D. Q. (2018). Klasifikasi Penyakit Aritmia Melalui Sinyal Elektrokardiogram (ekg) Menggunakan Metode Local Features Dan Support Vector Machine. *EProceedings of Engineering*, *5*(1).

Santosa, B., & Umam, A. (n.d.). *Data Mining dan Big Data Analytics: Teori dan Implementasi Menggunakan Python & Apache Spark*. Penebar Media Pustaka.

Sirait, D. T. C., Adiwijaya, A., & Astuti, W. (2019). Analisis Perbandingan Reduksi Dimensi Principal Component Analysis (pca) Dan Partial Least Square (pls) Untuk Deteksi Kanker Menggunakan Data Microarray. *EProceedings of Engineering*, *6*(2).

Sutanto, & Hernita, P. (2010). *Cekal (Cegah dan Tangkal) Penyakit Modern (Hipertensi, Stroke, Jantung, Kolesterol, dan Diabetes)* (Edisi 1). Yogyakarta: ANDI.

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., … Buxton, A. E. (2023). Heart disease and stroke statistics—2023 update: a report from the American Heart Association. *Circulation*, *147*(8), e93–e621.

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 437–444.

Vapnik, V. N. (1995). The nature of statistical learning. *Theory*.

Wibawa, M. S., & Maysanjaya, I. M. D. (2018). Multi Layer Perceptron Dan Principal Component Analysis Untuk Diagnosa Kanker Payudara. *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, *7*(1), 90–99.

Wismarini, D. (2022). PREDIKSI PROSES PERSALINAN MENGGUNAKAN ALGORITMA KNN BERBOBOT PADA MONITORING ELEKTRONIK PERSONAL HEALTH RECORD IBU HAMIL. *Jurnal Manajemen Informatika Dan Sistem Informasi*, *5*(1), 65–76.

Witten Ian, H., Eibe, F., Hall Mark, A., & Pal Christopher, J. (2017). Data Mining: Practical Machine Learning Tools and Techniques. 4e éd. Amsterdam: Morgan Kaufmann.