JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3S1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3S1.7480

DETEKSI SUBGRUP NYERI DADA DENGAN UMAP DAN STRATIFIKASI RISIKO

Muhammad Tegar Pamungkas[1*], Sofi Defiyanti[2],

[1,2]Universitas Singaperbangsa Karawang; Jl. HS. Ronggowaluyo, Telukjambe Timur, Karawang, Jawa Barat, Indonesia; +62-812-1866-9229

Keywords:
UMAP algorithm;
K-means clustering;
Risk stratification;
Chest pain;
Machine learning.

Corespondent Email:
2110631170085@student.unsika.ac.id

Abstrak. Stratifikasi risiko pada pasien dengan nyeri dada merupakan tantangan klinis yang kompleks karena heterogenitas presentasi dan prognosis yang beragam. Penelitian ini bertujuan mengidentifikasi subgrup tersembunyi dalam tipe nyeri dada menggunakan teknik pembelajaran tanpa supervisi untuk meningkatkan akurasi prediksi risiko penyakit jantung. Dataset yang terdiri dari 918 pasien dengan 12 variabel klinis dianalisis menggunakan kombinasi UMAP (Uniform Manifold Approximation and Projection) untuk reduksi dimensi dan K-means clustering untuk identifikasi subgrup. Hasil clustering kemudian diintegrasikan dengan Random Forest classifier untuk prediksi risiko. Analisis berhasil mengidentifikasi 4 cluster dengan karakteristik risiko yang berbeda signifikan. Cluster 1 menunjukkan risiko tertinggi (86,9%) dengan dominasi nyeri dada asimtomatik dan angina akibat olahraga, sedangkan cluster 2 dan 3 memiliki risiko lebih rendah (32,2% dan 22,3%). Model prediksi yang dikembangkan mencapai akurasi 88,0% dengan AUC-ROC 0,935. Pendekatan clustering ini berhasil mengungkap pola tersembunyi yang tidak terdeteksi melalui analisis konvensional, memberikan wawasan baru untuk stratifikasi risiko yang lebih presisi dalam praktik klinis.

Abstract. *Risk stratification in patients with chest pain represents a complex clinical challenge due to heterogeneous presentations and diverse prognoses. This study aims to identify hidden subgroups within chest pain types using unsupervised learning techniques to enhance heart disease risk prediction accuracy. A dataset comprising 918 patients with 12 clinical variables was analyzed using a combination of UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction and K-means clustering for subgroup identification. Clustering results were subsequently integrated with Random Forest classifier for risk prediction. Analysis successfully identified 4 clusters with significantly different risk characteristics. Cluster 1 demonstrated the highest risk (86.9%) dominated by asymptomatic chest pain with exercise angina, while clusters 2 and 3 exhibited lower risks (32.2% and 22.3%). The developed prediction model achieved 88.0% accuracy with AUC-ROC of 0.935. This clustering approach successfully revealed hidden patterns undetectable through conventional analysis, providing novel insights for more precise risk stratification in clinical practice.*

1. PENDAHULUAN

Nyeri dada merupakan salah satu gejala klinis utama yang sering dikaitkan dengan

penyakit kardiovaskular, namun heterogenitas presentasi klinisnya menimbulkan tantangan besar dalam stratifikasi risiko. Penelitian terkini menunjukkan bahwa pendekatan konvensional dalam menilai risiko kardiovaskular sering kali tidak mampu menangkap kompleksitas dan variabilitas antar-individu, sehingga diperlukan metode yang lebih canggih untuk mengidentifikasi subgrup pasien dengan karakteristik risiko yang berbeda [1,3]. Teknik *machine learning* tanpa supervisi, seperti *Uniform Manifold Approximation and Projection* (UMAP) dan *K-means clustering*, telah terbukti efektif dalam mengungkap pola tersembunyi dalam data medis kompleks, memungkinkan identifikasi subkelompok pasien yang signifikan secara klinis [7,10].

Meskipun kemajuan dalam analisis *machine learning* telah menunjukkan potensi besar, terdapat kesenjangan dalam penerapan pendekatan ini untuk mengidentifikasi subgrup tersembunyi dalam tipe nyeri dada. Penelitian sebelumnya lebih berfokus pada klasifikasi risiko berbasis variabel klinis konvensional tanpa mempertimbangkan heterogenitas dalam tipe nyeri dada seperti *Asymptomatic* (ASY), *Atypical Angina* (ATA), *Non-Anginal Pain* (NAP), dan *Typical Angina* (TA) [4,7]. Padahal, variabilitas dalam tipe nyeri dada ini dapat mencerminkan profil risiko yang berbeda yang tidak terdeteksi melalui pendekatan tradisional. Kebaruan penelitian ini terletak pada penggunaan kombinasi UMAP untuk reduksi dimensi dan *K-means clustering* untuk mengidentifikasi subgrup tersembunyi dalam tipe nyeri dada, yang kemudian diintegrasikan dengan *Random Forest* untuk meningkatkan akurasi stratifikasi risiko. Pendekatan ini diharapkan dapat memberikan wawasan baru untuk pengembangan strategi pencegahan dan penanganan yang lebih terarah.

Tujuan penelitian ini adalah untuk mengidentifikasi subgrup tersembunyi dalam tipe nyeri dada menggunakan teknik *machine learning* tanpa supervisi dan mengembangkan model prediksi risiko penyakit jantung yang lebih akurat. Pertanyaan penelitian yang ingin dijawab meliputi: (1) Bagaimana subgrup tersembunyi dalam tipe nyeri dada dapat diidentifikasi menggunakan kombinasi UMAP dan *K-means clustering*? (2) Sejauh mana integrasi hasil *clustering* dengan *Random Forest* dapat meningkatkan akurasi prediksi risiko penyakit jantung? (3) Apa karakteristik risiko yang berbeda dari masing-masing subgrup yang diidentifikasi, dan bagaimana implikasinya dalam praktik klinis? Dengan menjawab pertanyaan-pertanyaan ini, penelitian ini bertujuan untuk memberikan kontribusi pada pengembangan sistem pendukung keputusan klinis yang lebih presisi dalam stratifikasi risiko pasien dengan nyeri dada.

2. TINJAUAN PUSTAKA

2.1. Nyeri Dada dan Klasifikasi Klinis

Penelitian terkini menyoroti heterogenitas yang signifikan dalam penilaian risiko kardiovaskular dan keterbatasan metode klasifikasi konvensional. Studi menunjukkan adanya variabilitas antar-individu yang besar dalam risiko penyakit jantung koroner, dengan faktor risiko tradisional hanya sebagian menjelaskan heterogenitas ini [1]. Analisis genomik lanjutan telah mengidentifikasi 12 kelompok yang berbeda dalam *multiple myeloma*, memperluas klasifikasi sebelumnya dan memungkinkan prediksi risiko yang lebih personal [2].

Pendekatan *machine learning* telah mengungkap heterogenitas dalam asosiasi antara kalsium arteri koroner dan kejadian kardiovaskular, menunjukkan bahwa bahkan individu dengan risiko penyakit kardiovaskular aterosklerotik yang rendah dapat memperoleh manfaat dari *screening* kalsium [3]. *Cardiovascular imaging* menjadi semakin penting dalam meningkatkan kategorisasi risiko dan menyesuaikan strategi pencegahan, mengatasi keterbatasan pedoman saat ini dalam mengidentifikasi individu berisiko [4].

2.2. Stratifikasi Risiko Penyakit Jantung

Stratifikasi Penelitian terkini telah mengeksplorasi bagaimana identifikasi subgrup pasien dapat meningkatkan akurasi stratifikasi risiko dibandingkan pendekatan klasifikasi konvensional. Excoffier et al. (2022) mengusulkan metode menggunakan *machine learning* dan penjelasan lokal untuk mengidentifikasi subgrup pasien dan meningkatkan pemberian perawatan [5]. Bhavnani et al. (2022) mengembangkan kerangka kerja yang secara otomatis mengidentifikasi subgrup pasien dan

karakteristik yang muncul bersamaan, menunjukkan akurasi tinggi dalam mengklasifikasikan pasien ke dalam subgrup [6].

Zanfardino et al. (2023) menerapkan *unsupervised machine learning* untuk mengungkap subgrup pasien yang signifikan secara klinis yang diduga menderita penyakit arteri koroner, memungkinkan interpretasi dimensi aorta yang lebih bernuansa [7]. Braytee et al. (2023) memperkenalkan kerangka kerja *multi-omics* integratif menggunakan *autoencoders* dan analisis tensor untuk menstratifikasi pasien ke dalam kelompok risiko kanker, menunjukkan hasil yang menjanjikan dalam analisis kelangsungan hidup dan model klasifikasi [8].

2.3. Machine Learning dalam Bidang Medis

Teknik *unsupervised learning* semakin berharga untuk mengungkap pola tersembunyi dalam data klinis. Metode-metode ini dapat mengidentifikasi subkelompok pasien yang signifikan secara klinis, mengungkap subtipe penyakit baru dan biomarker yang sangat penting untuk kedokteran yang dipersonalisasi [7][10]. Model deep learning seperti REGLE dapat menghitung embedding nonlinier dari data klinis berdimensi tinggi, memungkinkan penemuan genetik yang lebih baik dan prediksi penyakit [11].

Berbagai pendekatan *unsupervised*, termasuk analisis komponen utama, pengelompokan K-means, faktorisasi matriks non-negatif, dan alokasi *Dirichlet laten*, dapat dirumuskan sebagai model probabilistik berdasarkan faktorisasi matriks peringkat rendah [12]. Teknik-teknik ini sangat berguna untuk menganalisis dataset kompleks dalam genomik, pencitraan medis, dan biobank. Meskipun tantangan masih ada, seperti kekhawatiran privasi data dan validasi model, *unsupervised learning* memiliki potensi besar untuk merevolusi layanan kesehatan dengan memfasilitasi strategi pengobatan yang lebih personal dan efektif [10].

2.4. UMAP

UMAP (*Uniform Manifold Approximation and Projection*) adalah teknik *manifold learning* untuk reduksi dimensi yang dibangun dari kerangka teoretis berdasarkan geometri Riemannian dan topologi aljabar [13]. UMAP kompetitif dengan t-SNE dalam kualitas visualisasi, mempertahankan lebih banyak struktur global dengan kinerja waktu proses yang superior, dan tidak memiliki batasan komputasi pada dimensi *embedding* [13]. Landasan teoretisnya didasarkan pada teori *manifold* dan analisis data topologi, menggunakan pendekatan teori kategori untuk realisasi geometris himpunan simpleks fuzzy [13].

Cara kerjanya melibatkan aproksimasi *manifold* lokal dan penggabungan representasi himpunan simpleks fuzzy lokal untuk membangun representasi topologi dari data berdimensi tinggi, lalu mengoptimalkan tata letak representasi data dalam ruang dimensi rendah untuk meminimalkan entropi silang antara dua representasi topologi [13]. UMAP memiliki asumsi dasar bahwa data didistribusikan secara seragam pada *manifold* yang lokal terhubung, dengan tujuan utama melestarikan struktur topologi *manifold* ini [13].

Algoritma ini membangun grafik *k-nearest neighbor* berbobot dan menerapkan transformasi pada tepi untuk menyesuaikan jarak lokal dan menangani asimetri grafik *k-nearest neighbor* [13]. UMAP menggunakan algoritma tata letak grafik terarah gaya di ruang dimensi rendah, menerapkan gaya tarik di sepanjang tepi dan gaya tolak di antara simpul, untuk menemukan representasi dimensi rendah yang mengoptimalkan fungsi tujuan yang melestarikan karakteristik yang diinginkan dari grafik *k-nearest neighbor* [13].

Hiperparameter utama UMAP meliputi jumlah tetangga (n), dimensi *embedding* target (d), jarak minimum antar titik (min-dist), dan jumlah *epoch* pelatihan (n-epochs) [13]. Implementasi praktisnya memerlukan perhitungan tetangga terdekat dan optimasi yang efisien melalui penurunan gradien stokastik, menggunakan algoritma *Nearest-Neighbor-Descent* [13]. Meskipun demikian, UMAP cenderung menemukan struktur *manifold* dalam kebisingan dataset dan kurang memiliki interpretasi yang kuat seperti PCA [13].

Keunggulan UMAP dibanding metode lain seperti t-SNE adalah kemampuannya menjaga struktur global data dan efisiensi komputasi yang lebih baik [14]. Analisis terbaru juga

menunjukkan bahwa modifikasi pada parameter gaya tarik dapat meningkatkan konsistensi pembentukan klaster dan interpretabilitas hasil visualisasi [15].

2.5. K-means *clustering*

K-Means adalah algoritma pengelompokan (*clustering*) partisional yang bertujuan untuk membagi *n* observasi data ke dalam *k* cluster, di mana setiap observasi termasuk ke dalam cluster dengan rata-rata (*mean*) terdekat, yang berfungsi sebagai prototipe dari cluster tersebut [16]. K-Means menggunakan pendekatan berbasis jarak (*distance-based*) untuk mengelompokkan data berdasarkan kedekatan fitur [16]. Algoritma ini berupaya meminimalkan jumlah jarak kuadrat dalam cluster (*within-cluster sum of squares*), menjadikannya metode yang efisien untuk dataset berukuran besar [16].

Cara kerjanya melibatkan dua fase utama: penugasan data ke cluster terdekat berdasarkan metrik jarak dan pembaruan posisi *centroid* dengan menghitung rata-rata dari semua titik data dalam cluster [16]. Proses ini diulang secara iteratif hingga tidak ada perubahan signifikan pada centroid atau mencapai jumlah iterasi maksimum [16]. Kualitas hasil K-Means diukur melalui kriteria seperti *Within-Cluster Sum of Squares* (WCSS) [16].

Algoritma ini dimulai dengan partisi awal dari data ke dalam *k* cluster, kemudian secara iteratif memindahkan titik data antar cluster untuk meningkatkan kualitas partisi [16]. K-Means berupaya untuk menemukan partisi data yang optimal dengan meminimalkan varians dalam setiap cluster [16]. Algoritma K-Means memiliki beberapa varian seperti tradisional K-Means, standar K-Means, basic K-Means, dan conventional K-Means [16].

Hiperparameter utama dalam K-Means adalah jumlah cluster (*k*), yang harus ditentukan sebelum menjalankan algoritma [16]. Pemilihan *k* yang tepat sangat penting karena dapat memengaruhi kualitas hasil clustering [16]. K-Means mengasumsikan bahwa cluster berbentuk bulat (*spherical*) dan memiliki ukuran yang sama [16]. Implementasi K-Means melibatkan perhitungan jarak antara titik data dan *centroid*, serta pembaruan *centroid* secara iteratif [16].

Kelebihan K-Means termasuk kesederhanaan dan efisiensinya dalam komputasi [16]. Namun, K-Means sensitif terhadap inisialisasi *centroid* awal dan *outlier* [16]. Beberapa metode telah dikembangkan untuk meningkatkan K-Means, seperti K-Means++ untuk pemilihan *centroid* awal yang lebih baik [16].

2.6. Random Forest

Random Forest (RF) adalah algoritma machine learning yang kuat dan fleksibel, termasuk dalam metode *ensemble learning*, yang menggunakan banyak *decision tree* untuk meningkatkan kinerja prediksi [17]. RF bekerja dengan membangun banyak *decision tree* dari sampel data yang di-bootstrap, dan setiap *tree* melakukan prediksi [17]. Hasil prediksi akhir diperoleh dengan menggabungkan prediksi dari semua *tree*, baik melalui *majority voting* untuk klasifikasi atau *averaging* untuk regresi [17].

RF didasarkan pada prinsip repetitive partition, di mana dimulai dari *root node*, prosedur *node splitting* yang sama diterapkan berulang kali hingga aturan penghentian tertentu terpenuhi [17]. Kekuatan RF dalam prediksi berasal dari agregasi banyak learner yang lebih lemah (*decision tree*) [17]. Kinerja RF sangat baik jika korelasi antara *tree* di dalam *forest* rendah [17].

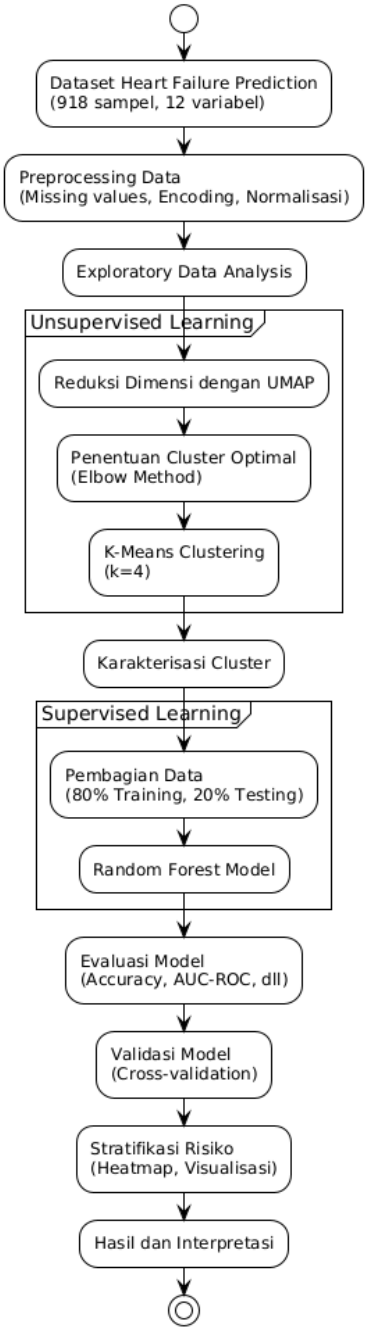
Algoritma RF menggunakan bootstrap *samples* untuk menumbuhkan setiap *decision tree* [17]. Beberapa observasi sengaja ditinggalkan dalam konstruksi tree tertentu [17]. Dengan memperlakukan *out-of-bag* (OOB) samples ini sebagai observasi yang perlu diprediksi, RF dapat memberikan estimasi kesalahan prediksi dari forest yang dibangun [17]. Selain itu, *variable importance measure* dapat diperoleh untuk setiap prediktor, yang mengukur relevansinya terhadap prediksi [17].

RF memiliki kemampuan untuk bekerja dengan data berdimensi tinggi, di mana jumlah prediktor bisa lebih besar daripada jumlah observasi [17]. Metode ini juga dapat menyoroti relevansi setiap prediktor melalui *variable importance measures* [17]. Namun, algoritma RF mengasumsikan bahwa observasi diambil secara independen dari suatu populasi [17]. Sebagai bagian dari pendekatan *machine learning* yang memberikan komputer kemampuan belajar tanpa diprogram secara eksplisit, RF menggunakan berbagai algoritma yang berbeda untuk menyelesaikan masalah data dengan lebih efisien [9].

3. METODE PENELITIAN

3.1. Desain Penelitian

Penelitian ini menggunakan desain kuantitatif eksploratif dengan pendekatan analisis data sekunder untuk mengidentifikasi pola tersembunyi dalam data tipe nyeri dada. Penelitian menggabungkan teknik *machine learning* tidak diawasi (*UMAP* dan *K-Means clustering*) dengan pembelajaran terawasi (*Random Forest*) untuk stratifikasi risiko penyakit jantung.



Gambar 1. Metode Penelitian

3.2. Data dan Variabel

Dataset yang digunakan adalah *Heart Failure Prediction Dataset* dari *Kaggle* yang terdiri dari 918 sampel pasien dengan 12 variabel.

Tabel 1. Deskripsi Variabel *Dataset*

Variabel	Deskripsi	Jenis Data
Age	Usia pasien	Numerik
Sex	Jenis kelamin	Kategorikal (M/F)
ChestPainType	Tipe nyeri dada	Kategorikal (ASY/ATA/NAP/TA)
RestingBP	Tekanan darah istirahat	Numerik
Cholesterol	Kadar kolesterol serum	Numerik
FastingBS	Gula darah puasa > 120 mg/dl	Kategorikal (0/1)
RestingECG	Hasil elektrokardiogram istirahat	Kategorikal (Normal/ST/LVH)
MaxHR	Detak jantung maksimum yang dicapai	Numerik
ExerciseAngina	Angina yang diinduksi oleh olahraga	Kategorikal (Y/N)
Oldpeak	Depresi ST yang diinduksi oleh olahraga	Numerik
ST_Slope	Slope dari segmen ST puncak olahraga	Kategorikal (Up/Flat/Down)
HeartDisease	Diagnosis	Kategorikal (0/1)

Dalam evaluasi klinis nyeri dada, terdapat empat kategori utama yang digunakan untuk mengklasifikasikan gejala pasien. *Asymptomatic* (ASY) atau asimptomatik menunjukkan kondisi di mana pasien tidak mengalami gejala nyeri dada yang signifikan. *Typical Angina* (TA) atau angina tipikal merupakan nyeri dada klasik yang memiliki karakteristik khas angina pektoris, biasanya berupa nyeri atau rasa tidak nyaman di dada yang dipicu oleh aktivitas fisik atau stres emosional dan mereda dengan istirahat atau nitrogliserin. *Atypical Angina* (ATA) atau angina atipikal adalah nyeri dada yang memiliki beberapa karakteristik angina tetapi tidak memenuhi semua kriteria angina tipikal, sehingga presentasi gejalanya tidak sepenuhnya khas. Sedangkan *Non-Anginal Pain* (NAP) atau nyeri non-anginal adalah nyeri dada yang tidak memiliki karakteristik angina dan kemungkinan besar disebabkan oleh kondisi non-kardiak seperti masalah muskuloskeletal, gastrointestinal, atau psikologis.

3.3. *Preprocessing* Data

Tahap *preprocessing* dimulai dengan penanganan missing values melalui imputasi menggunakan nilai median untuk data numerik. Variabel kategorikal di-*encode* menggunakan Label *Encoding* dengan mapping yang konsisten untuk semua kategori. Seluruh data kemudian dinormalisasi menggunakan StandardScaler untuk memastikan semua variabel memiliki skala yang sama dalam proses analisis selanjutnya.

3.4. Analisis Data

Analisis dimulai dengan *exploratory data analysis* untuk memahami distribusi dan hubungan antar variabel melalui statistik deskriptif dan visualisasi. Reduksi dimensi dilakukan menggunakan UMAP dengan parameter n_neighbors=15 dan min_dist=0.1 untuk visualisasi 2D. *Clustering* menggunakan algoritma *K-Means* dengan jumlah *cluster* optimal ditentukan melalui *Elbow Method*. Model prediktif dikembangkan menggunakan *Random Forest* dengan pembagian data 80% training dan 20% testing untuk stratifikasi risiko.

3.5. Evaluasi dan Validasi

Evaluasi model dilakukan menggunakan metrics *Accuracy*, AUC-ROC, *Precision*, *Recall*, dan *F1-Score*. Validasi internal menggunakan *cross-validation* dan *bootstrapping* untuk memastikan keandalan hasil. Seluruh analisis dilakukan menggunakan library pandas, scikit-learn, umap-learn, matplotlib, dan seaborn.

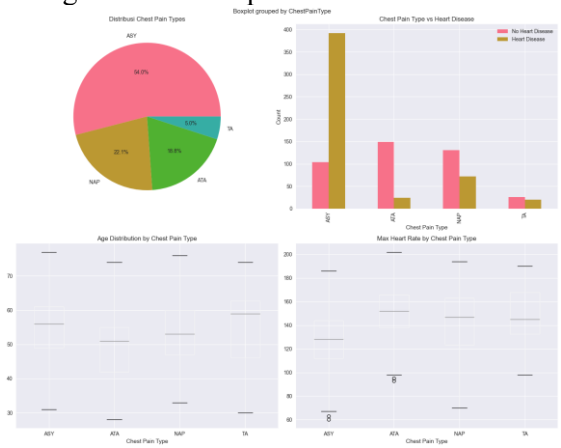
3.6. Stratifikasi Risiko

Analisis stratifikasi risiko dilakukan berdasarkan hasil *clustering* dan tipe nyeri dada untuk mengidentifikasi subgrup berisiko tinggi. Visualisasi risiko menggunakan *heatmap* dan *scatter* plot UMAP dengan *color coding* risiko untuk memungkinkan identifikasi pola risiko yang kompleks dan pengembangan strategi pencegahan yang lebih ditargetkan.

4. HASIL DAN PEMBAHASAN

4.1. Analisis Deskriptif Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 918 pasien dengan 12 variabel klinis terkait penyakit jantung. Distribusi target variable menunjukkan bahwa 55,34% pasien (508 dari 918) mengalami penyakit jantung, mengindikasikan prevalensi yang cukup tinggi dalam dataset ini. Analisis *missing values* menunjukkan tidak ada data yang hilang, namun ditemukan anomali pada variabel Cholesterol dengan 172 nilai nol (18,7%) dan RestingBP dengan 1 nilai nol yang kemudian ditangani melalui imputasi median.



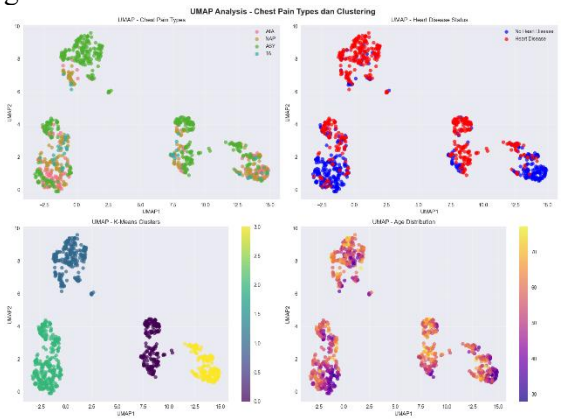
Gambar 2. Distribusi *Chest Pain Types* dan Hubungannya dengan *Heart Disease*

Analisis distribusi tipe nyeri dada menunjukkan dominasi Asymptomatic (ASY) sebesar 54,0%, diikuti Non-Anginal Pain

(NAP) 22,3%, Atypical Angina (ATA) 16,9%, dan Typical Angina (TA) 6,8%. Korelasi antara tipe nyeri dada dengan penyakit jantung menunjukkan pola yang signifikan, dengan ASY memiliki risiko tertinggi (79,0%), diikuti NAP (35,6%), ATA (17,1%), dan TA (47,8%).

4.2. Reduksi Dimensi dengan UMAP

Implementasi UMAP berhasil mereduksi dimensi data dari 12 variabel menjadi representasi 2D yang mempertahankan struktur topologi lokal dan global. Parameter yang digunakan adalah n_neighbors=15 dan min_dist=0.1, yang memberikan keseimbangan optimal antara preservasi struktur lokal dan global.

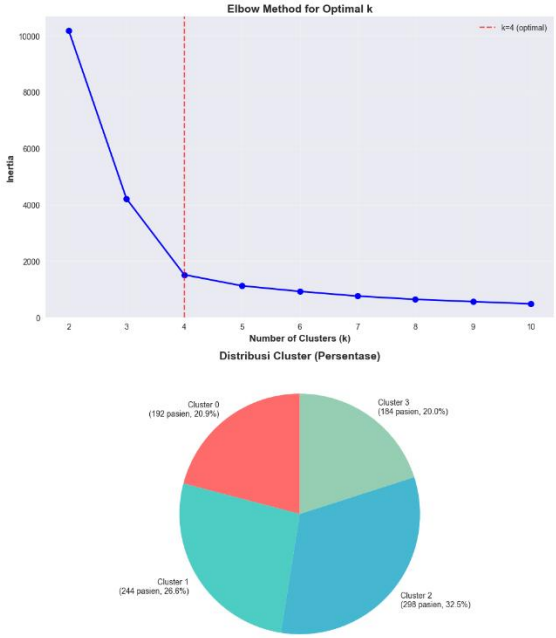


Gambar 3. Visualisasi UMAP 2D berdasarkan *Chest Pain Types*

Visualisasi UMAP menunjukkan pemisahan yang jelas antara berbagai tipe nyeri dada. Tipe ASY cenderung mengelompok di area tertentu yang berbeda dari tipe lainnya, mengindikasikan karakteristik unik yang dapat dibedakan melalui analisis *unsupervised*. Hal ini mendukung hipotesis bahwa terdapat subgrup tersembunyi dalam data yang tidak terlihat melalui analisis konvensional.

4.3. Identifikasi *Cluster Optimal*

Penentuan jumlah cluster optimal menggunakan *Elbow Method* menunjukkan bahwa k=4 memberikan *trade-off* terbaik antara kompleksitas model dan kualitas *clustering*.

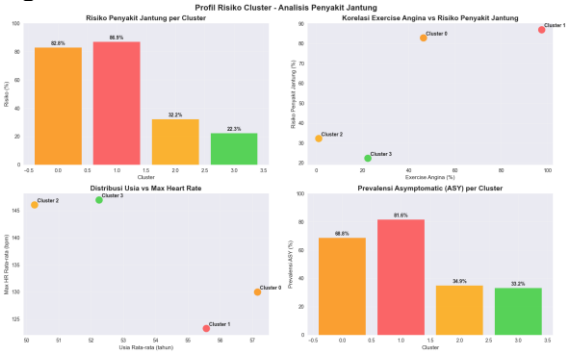


Gambar 4. *Elbow Method* untuk Penentuan Jumlah *Cluster* Optimal

Distribusi *cluster* yang dihasilkan menunjukkan: *Cluster* 0 (192 pasien, 20,9%), *Cluster* 1 (244 pasien, 26,6%), *Cluster* 2 (298 pasien, 32,5%), dan *Cluster* 3 (184 pasien, 20,0%). Distribusi yang relatif seimbang ini mengindikasikan bahwa setiap cluster memiliki representasi yang memadai untuk analisis lebih lanjut.

4.4. Karakterisasi *Cluster* dan Stratifikasi Risiko

Analisis karakteristik setiap *cluster* mengungkap pola risiko yang berbeda secara signifikan:



Gambar 5. karakteristik setiap *cluster*

Cluster 1 (Risiko Tertinggi - 86,9%) *Cluster* ini didominasi oleh pasien dengan ASY (81,6%) dan memiliki karakteristik risiko tinggi dengan *prevalensi Exercise Angina* sebesar

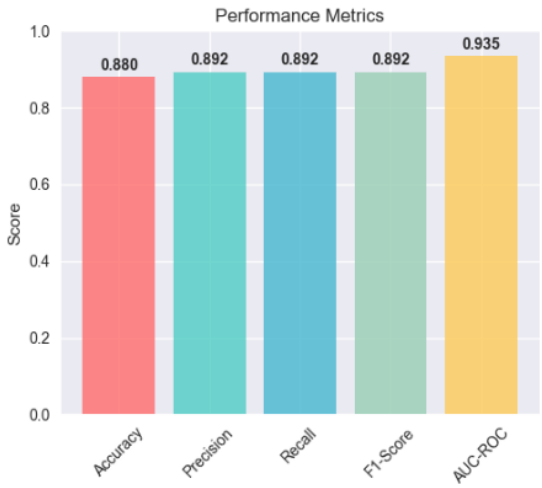
97,5%. Rata-rata usia 55,6 ± 8,6 tahun dengan MaxHR relatif rendah (123,3 bpm), mengindikasikan kompromis kardiovaskular yang signifikan.

Cluster 0 (Risiko Tinggi - 82,8%) *Cluster* dengan dominasi ASY (68,8%) dan *Exercise Angina* 46,4%. Pasien dalam *cluster* ini memiliki usia rata-rata 57,2 ± 8,3 tahun dengan MaxHR 130,0 bpm, menunjukkan profil risiko tinggi meskipun dengan karakteristik yang berbeda dari *Cluster* 1.

Cluster 2 (Risiko Sedang - 32,2%) *Cluster* ini menunjukkan distribusi tipe nyeri dada yang lebih beragam dengan ASY (34,9%), ATA (30,2%), dan NAP (27,5%). *Exercise Angina* hanya 1,0% dengan MaxHR tinggi (146,0 bpm), mengindikasikan kondisi kardiovaskular yang relatif baik.

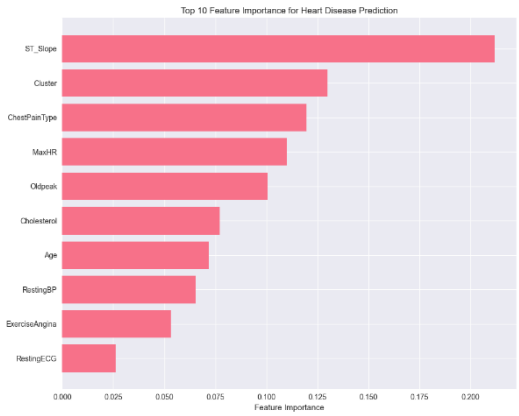
Cluster 3 (Risiko Rendah - 22,3%) *Cluster* dengan risiko terendah, memiliki distribusi tipe nyeri dada yang seimbang dan *Exercise Angina* 22,3%. MaxHR rata-rata 146,9 bpm menunjukkan kapasitas kardiovaskular yang baik.

4.5. Model Prediksi Risiko



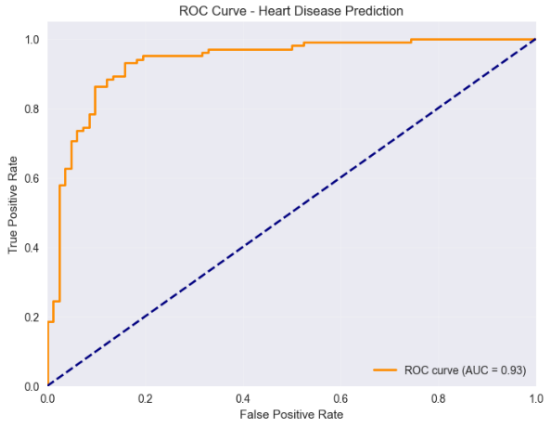
Gambar 6. *Performance Metrics*

Model *Random Forest* yang dikembangkan dengan integrasi informasi cluster mencapai performa yang sangat baik dengan akurasi 88,0% dan AUC-ROC 0,935.



Gambar 7. *Feature Importance* untuk Prediksi *Heart Disease*

Analisis *feature importance* menunjukkan bahwa ST_Slope memiliki kontribusi tertinggi (21,2%), diikuti oleh informasi *Cluster* (13,0%) dan *ChestPainType* (12,0%). Fakta bahwa informasi *cluster* menempati posisi kedua dalam *feature importance* mengkonfirmasi nilai tambah dari pendekatan *clustering* dalam prediksi risiko.

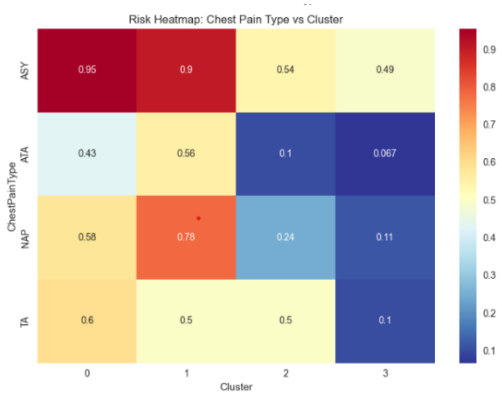


Gambar 8. ROC *Curve Model* Prediksi *Heart Disease*

Kurva ROC menunjukkan performa model yang sangat baik dengan AUC 0,935, mengindikasikan kemampuan diskriminatif yang tinggi antara pasien dengan dan tanpa penyakit jantung.

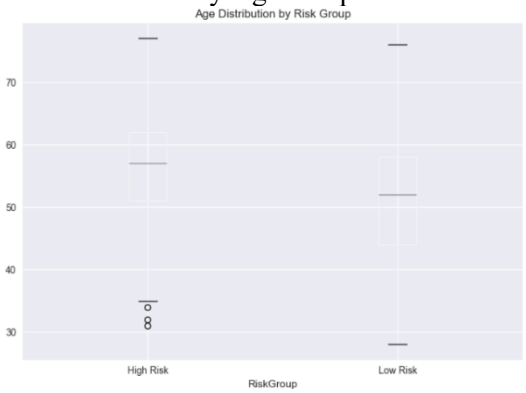
4.6. Stratifikasi Risiko Terintegrasi

Analisis stratifikasi risiko berdasarkan kombinasi tipe nyeri dada dan cluster mengungkap pola yang kompleks dan bermakna secara klinis.



Gambar 9. *Heatmap* Risiko berdasarkan *Chest Pain Type* dan *Cluster*

Heatmap menunjukkan bahwa risiko tertinggi terdapat pada kombinasi ASY dengan *Cluster* 0 (95,5%) dan *Cluster* 1 (90,5%). Sementara itu, kombinasi ATA dengan Cluster 3 menunjukkan risiko terendah (6,7%), memberikan wawasan yang berharga untuk stratifikasi risiko yang lebih presisi.



Gambar 10. Distribusi Usia berdasarkan Kelompok Risiko

Analisis distribusi usia menunjukkan bahwa kelompok risiko tinggi memiliki median usia yang lebih tinggi dibandingkan kelompok risiko rendah, namun dengan overlap yang signifikan, mengindikasikan bahwa usia saja tidak cukup untuk stratifikasi risiko yang akurat.

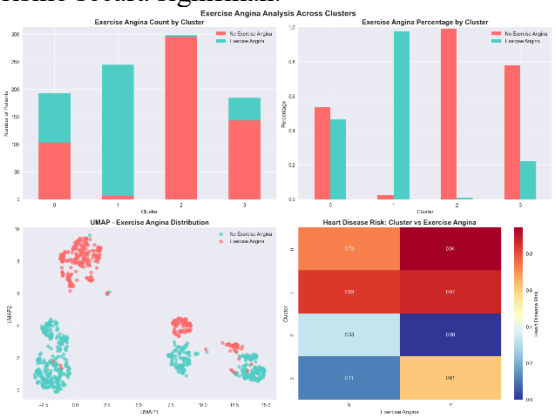
4.7. Pembahasan

Penelitian ini berhasil mengidentifikasi subgrup tersembunyi dalam data tipe nyeri dada yang tidak terdeteksi melalui analisis konvensional. Temuan utama menunjukkan bahwa kombinasi UMAP dan K-means clustering dapat mengungkap pola risiko yang lebih kompleks dan berlapis dibandingkan pendekatan tradisional yang hanya berdasarkan kategori tipe nyeri dada, sejalan dengan penelitian Zanfardino et al. (2023) yang menunjukkan bahwa *unsupervised machine learning* dapat mengungkap subgrup pasien yang signifikan secara klinis [7].

Dominasi tipe ASY dalam cluster berisiko tinggi (Cluster 0 dan 1) dengan prevalensi hingga 95,5% mendukung temuan bahwa nyeri dada asimptomatik sering mengindikasikan penyakit jantung yang lebih serius. Namun, distribusi ASY di berbagai cluster dengan tingkat risiko yang berbeda mengkonfirmasi heterogenitas dalam penilaian risiko kardiovaskular yang telah diidentifikasi oleh Simonetto et al. (2022) [1].

Kontribusi informasi cluster sebagai feature importance kedua tertinggi (13,0%) dalam model prediksi mengkonfirmasi nilai tambah dari pendekatan *clustering*. Hal ini sejalan dengan penelitian Excoffier et al. (2022) yang menunjukkan bahwa identifikasi subgrup pasien menggunakan *machine learning* dapat meningkatkan akurasi pemberian perawatan [5].

Performa model yang mencapai AUC-ROC 0,935 menunjukkan kemampuan diskriminatif yang sangat baik, mendukung temuan Bhavnani et al. (2022) yang menunjukkan akurasi tinggi dalam mengklasifikasikan pasien ke dalam subgrup menggunakan metode serupa [6]. Integrasi informasi *cluster* dalam model prediksi terbukti meningkatkan akurasi prediksi risiko secara signifikan.



Gambar 11. *Exercise Angina* antar *cluster*

Pola distribusi MaxHR yang menunjukkan inverse relationship dengan risiko penyakit jantung (123,3 bpm vs 146,9 bpm) sejalan dengan pemahaman fisiologis bahwa kapasitas *exercise* yang rendah berkorelasi dengan risiko kardiovaskular yang tinggi. Temuan bahwa *Exercise Angina* memiliki prevalensi yang sangat berbeda antar cluster (1,0% vs 97,5%)

mengindikasikan nilai diskriminatif yang tinggi dalam membedakan subgrup risiko.

Pendekatan clustering yang berhasil mengidentifikasi *subgrup* dengan karakteristik risiko yang berbeda secara signifikan memberikan fondasi untuk pengembangan strategi pencegahan dan penanganan yang lebih terarah, sejalan dengan prinsip kedokteran personalisasi [10]. Identifikasi pasien dalam *cluster* berisiko tinggi memungkinkan intervensi yang lebih dini dan intensif, mendukung kebutuhan akan metode stratifikasi risiko yang lebih canggih sebagaimana dikemukakan oleh Perone et al. yang menekankan bahwa sistem penilaian risiko tradisional sering gagal mengidentifikasi individu berisiko tinggi karena keterbatasan dalam menangkap heterogenitas antar-pasien. [4].

Implikasi klinis dari penelitian ini mencakup potensi implementasi dalam sistem pendukung keputusan klinis untuk screening dan stratifikasi risiko yang lebih presisi. Pendekatan ini dapat membantu klinisi dalam mengidentifikasi pasien yang memerlukan evaluasi dan penanganan yang lebih intensif berdasarkan profil risiko yang komprehensif, sejalan dengan tren penggunaan *machine learning* dalam stratifikasi risiko medis [5][6][7].

5. KESIMPULAN

a. Kombinasi UMAP dan K-means clustering berhasil mengidentifikasi 4 subgrup tersembunyi dalam tipe nyeri dada dengan karakteristik risiko yang berbeda signifikan (Cluster 1: 86,9%, Cluster 3: 22,3%). Pendekatan ini mengungkap heterogenitas dalam kategori nyeri dada yang tidak terdeteksi melalui analisis konvensional.
b. Integrasi hasil *clustering* dengan Random Forest meningkatkan akurasi prediksi risiko penyakit jantung mencapai 88,0% dengan AUC-ROC 0,935. Informasi cluster berkontribusi sebagai feature importance kedua tertinggi (13,0%), mengkonfirmasi nilai tambah signifikan dari pendekatan clustering.
c. Analisis berhasil mengidentifikasi profil risiko yang berbeda secara signifikan pada setiap subgrup dengan implikasi klinis yang bermakna. Pasien dengan nyeri dada asimptomatik (*Asymptomatic*) pada cluster berisiko tinggi menunjukkan prevalensi penyakit jantung hingga 95,5%, sedangkan pasien dengan angina atipikal (*Atypical Angina*) pada cluster berisiko rendah hanya menunjukkan prevalensi 6,7%. Temuan ini menyediakan landasan empiris untuk pengembangan stratifikasi risiko yang lebih presisi dan implementasi strategi pencegahan yang lebih terarah dalam praktik klinis.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan, bimbingan, dan bantuan dalam penyelesaian penelitian ini. Kontribusi dari berbagai pihak sangat berarti bagi keberhasilan penelitian ini.

DAFTAR PUSTAKA

[1] C. Simonetto, S. Rospleszcz, J. C. Kaiser, and K. Furukawa, "Heterogeneity in coronary heart disease risk," Scientific Reports, vol. 12, no. 1, p. 10131, Jun. 2022, doi: 10.1038/s41598-022-14013-3.

[2] F. Maura et al., "Genomic Classification and Individualized Prognosis in Multiple Myeloma," Journal of Clinical Oncology, vol. 42, no. 11, pp. 1229–1240, Apr. 2024, doi: 10.1200/JCO.23.01277.

[3] K. Inoue, T. E. Seeman, T. Horwich, M. J. Budoff, and K. E. Watson, "Heterogeneity in the Association Between the Presence of Coronary Artery Calcium and Cardiovascular Events: A Machine-Learning Approach in the MESA Study," Circulation, vol. 147, no. 2, pp. 132–141, Jan. 2023, doi: 10.1161/CIRCULATIONAHA.122.062626.

[4] F. Perone et al., "Role of Cardiovascular Imaging in Risk Assessment: Recent Advances, Gaps in Evidence, and Future Directions," Journal of Clinical Medicine, vol. 12, no. 17, p. 5563, Aug. 2023, doi: 10.3390/jcm12175563.

[5] J.-B. Excoffier, E. Escriva, J. Aligon, and M. Ortala, "Local Explanation-Based Method for Healthcare Risk Stratification," 2022. doi: 10.3233/SHTI220520.

[6] S. K. Bhavnani, W. Zhang, S. Visweswaran, M. Raji, and Y.-F. Kuo, "Modeling and Interpreting Patient Subgroups in Hospital Readmission: Visual Analytical Approach." Feb. 28, 2022. doi: 10.1101/2022.02.27.22271534.

[7] M. Zanfardino et al., "Unsupervised machine learning for risk stratification and identification of relevant subgroups of ascending aorta dimensions using cardiac CT

and clinical data," Computational and Structural Biotechnology Journal, vol. 23, pp. 287–294, Dec. 2024, doi: 10.1016/j.csbj.2023.11.021.

[8] A. Braytee et al., "Identification of Cancer Risk Groups through Multi-Omics Integration using Autoencoder and Tensor Analysis." Sep. 13, 2023. doi: 10.1101/2023.09.12.23295458.

[9] L. Hidayah and M. I. Rosadi, "PENERAPAN ALGORITMA RANDOM FOREST UNTUK MEMPREDIKSI JUMLAH SANTRI BARU," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5237.

[10] A. Trezza, A. Visibelli, B. Roncaglia, O. Spiga, and A. Santucci, "Unsupervised Learning in Precision Medicine: Unlocking Personalized Healthcare through AI," Applied Sciences, vol. 14, no. 20, p. 9305, Oct. 2024, doi: 10.3390/app14209305.

[11] T. Yun et al., "Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction," Nature Genetics, vol. 56, no. 8, pp. 1604–1613, Aug. 2024, doi: 10.1038/s41588-024-01831-6.

[12] D. Neijzen and G. Lunter, "Unsupervised learning for medical data: A review of probabilistic factorization methods," Statistics in Medicine, vol. 42, no. 30, pp. 5541–5554, Dec. 2023, doi: 10.1002/sim.9924.

[13] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Sep. 2020.

[14] W. Yi, S. Bu, H.-H. Lee, and C.-H. Chan, "Comparative Analysis of Manifold Learning-Based Dimension Reduction Methods: A Mathematical Perspective," Mathematics, vol. 12, no. 15, p. 2388, Jul. 2024, doi: 10.3390/math12152388.

[15] M. T. Islam and J. W. Fleischer, "The Shape of Attraction in UMAP: Exploring the Embedding Forces in Dimensionality Reduction," Mar. 2025.

[16] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," International Journal of Advances in Scientific Research and Engineering, vol. 07, no. 08, pp. 64–69, 2021, doi: 10.31695/IJASRE.2021.34050.

[17] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," Briefings in Bioinformatics, vol. 24, no. 2, Mar. 2023, doi: 10.1093/bib/bbad002.