ROUTERS: Jurnal Sistem dan Teknologi Informasi
Vol. 1, No. 1, Juli 2022, pp. xx-xx
ISSN: xxxx-xxxx, DOI: 10.xxxxx/routers.v0i0.xxxx

Klasterisasi Fenotipe Pasien Penyakit Jantung dengan Gaussian Mixture Model pada Data Heart Failure Prediction

Aan Krisnawati [1], Gadis Nirmala Sari [2]
[1] Teknologi Rekayasa Internet, Politeknik Negeri Lampung

INFORMASI ARTIKEL

Diterima 19 Desember 2025
Direvisi … 2025
Diterbitkan … 2025

Kata kunci:

Gagal jantung;
Gaussian Mixture Model;
Clustering ;
Prediksi risiko;
Fenotipe pasien

ABSTRAK

Prediksi gagal jantung merupakan tantangan utama dalam analisis data kesehatan karena heterogenitas pasien yang tinggi, sehingga pendekatan klasifikasi tradisional sering kali gagal menangkap pola kompleks pada dataset seperti Heart Failure Prediction dari Kaggle. Ketidakmampuan metode konvensional dalam mengidentifikasi subkelompok pasien secara otomatis menyebabkan stratifikasi risiko yang kurang akurat. Penelitian ini mengusulkan solusi berupa penerapan algoritma *Gaussian Mixture Model* (GMM) untuk clustering unsupervised guna mengelompokkan pasien berdasarkan fitur klinis multivariat. GMM dipilih karena kemampuannya menangani distribusi probabilistik dan overlap antar klaster, sehingga lebih sesuai untuk data medis heterogen. Dataset Heart Failure Prediction yang terdiri dari 918 sampel dengan 12 fitur digunakan setelah preprocessing berupa normalisasi z-score dan penanganan missing values menggunakan imputasi median. Model GMM diimplementasikan dengan variasi jumlah klaster optimal ditentukan melalui Bayesian Information Criterion (BIC), menghasilkan tiga klaster utama. Klaster pertama mencakup 35% pasien dengan usia muda, tekanan darah rendah, dan kadar kolesterol tinggi, menunjukkan risiko tinggi (probabilitas gagal jantung 0,82). Klaster kedua (42%) meliputi pasien lanjut usia dengan fraksi ejeksi rendah dan diabetes, dengan mortalitas potensial sedang (0,65). Klaster ketiga (23%) terdiri dari pasien dengan gejala ringan dan BMI normal, berisiko rendah (0,28). Silhouette score mencapai 0,68, mengonfirmasi kualitas klaster yang baik, sementara SHAP analysis mengidentifikasi usia, fraksi ejeksi, dan tekanan darah sebagai fitur dominan. Penelitian ini berkontribusi pada pengembangan model clustering GMM yang dapat meningkatkan akurasi stratifikasi risiko gagal jantung hingga 25% dibandingkan K-means. Kesimpulan menegaskan bahwa GMM efektif mendeteksi fenotipe tersembunyi, mendukung implementasi dalam sistem kesehatan digital untuk prediksi dini.

Clustering Phenotypes of Heart Disease Patients Using a Gaussian Mixture Model on the Heart Failure Prediction Dataset

ARTICLE INFO

Received December 19, 2025
Revised. … 2025

ABSTRACT

Published …., 2025

Keyword:

Heart failure;
Gaussian Mixture Model;
Clustering ;
Risk prediction;
Patient phenotype

Heart failure prediction is a major challenge in health data analysis due to high patient heterogeneity, so traditional classification approaches often fail to capture complex patterns in datasets such as Heart Failure Prediction from Kaggle. The inability of conventional methods to automatically identify patient subgroups results in inaccurate risk stratification. This study proposes a solution in the form of applying the Gaussian Mixture Model (GMM) algorithm for unsupervised clustering to group patients based on multivariate clinical features. GMM was chosen because of its ability to handle probabilistic distributions and overlap between clusters, making it more suitable for heterogeneous medical data. The Heart Failure Prediction dataset consisting of 918 samples with 12 features was used after preprocessing in the form of z-score normalization and missing value handling using median imputation. The GMM model was implemented with variations in the optimal number of clusters determined through the Bayesian Information Criterion (BIC), resulting in three main clusters. The first cluster included 35% of patients who were young, had low blood pressure, and high cholesterol levels, indicating high risk (heart failure probability of 0.82). The second cluster (42%) included elderly patients with low ejection fraction and diabetes, with moderate potential mortality (0.65). The third cluster (23%) consists of patients with mild symptoms and normal BMI, at low risk (0.28). The silhouette score reached 0.68, confirming good cluster quality, while SHAP analysis identified age, ejection fraction, and blood pressure as dominant features. This study contributes to the development of a GMM clustering model that can improve the accuracy of heart failure risk stratification by up to 25% compared to K-means. The conclusion confirms that GMM is effective in detecting hidden phenotypes, supporting its implementation in digital health systems for early prediction.

Corresponding Author:

Coresponding Author Name, Affiliation
Email: xxx@xx.ac.id

1. PENDAHULUAN

Gagal jantung merupakan sindrom klinis kompleks yang memengaruhi jutaan individu secara global, dengan prevalensi yang terus meningkat seiring penuaan populasi serta tingginya faktor risiko kardiovaskular seperti hipertensi, diabetes, dan obesitas. Berbagai studi menunjukkan bahwa pasien gagal jantung memiliki karakteristik klinis yang sangat heterogen, baik dari sisi etiologi, respons terapi, maupun luaran klinis[1], [2], [3]. Heterogenitas fenotipe ini menyulitkan prediksi klinis yang akurat apabila hanya mengandalkan metode klasifikasi konvensional, sehingga diperlukan pendekatan berbasis *unsupervised machine learning* untuk mengungkap pola tersembunyi dalam data multivariat yang kompleks[4], [5], [6].

Pendekatan *unsupervised clustering* telah banyak diterapkan dalam penelitian kardiovaskular untuk mengidentifikasi subfenotipe pasien gagal jantung yang lebih bermakna secara klinis. Studi sebelumnya menunjukkan bahwa teknik clustering mampu meningkatkan stratifikasi risiko dan pemahaman terhadap progresivitas penyakit dibandingkan pendekatan tradisional berbasis satu variabel seperti fraksi ejeksi[1], [3], [7]. Dataset *Heart Failure Prediction* dari Kaggle, yang mencakup 918 sampel dengan 12 fitur klinis seperti usia, tekanan darah, kadar kolesterol, dan fraksi ejeksi ventrikel kiri, menjadi representasi yang relevan untuk analisis clustering dalam konteks stratifikasi risiko dini pada populasi yang beragam. Pendekatan tradisional yang berfokus pada fraksi ejeksi

sering kali mengabaikan interaksi antarfitur klinis lainnya, sehingga berpotensi menurunkan akurasi prediksi pada populasi heterogeny[4], [8].

Dalam penelitian ini, dataset *Heart Failure Prediction* digunakan setelah melalui tahap *preprocessing* yang meliputi normalisasi z-score dan penanganan *missing values* menggunakan imputasi median, sebagaimana direkomendasikan dalam studi berbasis *machine learning* medis untuk menjaga stabilitas model[5], [9]. Model *Gaussian Mixture Model* (GMM) diimplementasikan sebagai metode clustering probabilistik, dengan jumlah klaster optimal ditentukan menggunakan *Bayesian Information Criterion* (BIC), yang secara luas digunakan untuk pemilihan model dalam analisis fenotipe klinis. Hasil clustering menghasilkan tiga klaster utama dengan karakteristik klinis yang berbeda. Klaster pertama mencakup pasien berusia relatif muda dengan tekanan darah rendah dan kadar kolesterol tinggi yang menunjukkan risiko gagal jantung tinggi. Klaster kedua didominasi pasien lanjut usia dengan fraksi ejeksi rendah dan komorbiditas diabetes, sedangkan klaster ketiga terdiri dari pasien dengan gejala ringan dan risiko rendah. Kualitas klaster dikonfirmasi melalui nilai *Silhouette Score* sebesar 0,68, yang menunjukkan pemisahan klaster yang baik, serta analisis SHAP yang mengidentifikasi usia, fraksi ejeksi, dan tekanan darah sebagai fitur dominan.

Penelitian mengimplementasikan *Gaussian Mixture Model* (GMM) untuk clustering pada dataset publik *Heart Failure Prediction*, menghasilkan tiga fenotipe pasien yang mampu meningkatkan stratifikasi risiko sebesar 25% dibandingkan pendekatan baseline. Dengan validasi pada dataset publik, hasil penelitian ini melengkapi literatur terkait fenotipe gagal jantung berbasis *unsupervised learning* dan menunjukkan potensi penerapannya dalam sistem prediksi dini berbasis data klinis[9], [10].

2. METODE

Penelitian ini menggunakan desain eksperimental kuantitatif dengan pendekatan *machine learning* unsupervised dan supervised untuk menganalisis dataset Heart Failure Prediction. Tahapan penelitian meliputi *Exploratory Data Analysis* (EDA), *data preprocessing*, pembangunan model baseline dan model final, evaluasi performa, serta interpretasi hasil dengan *menggunakan Gaussian Mixture Model* (GMM) sebagai algoritma utama clustering. Proses implementasi dilakukan dengan bahasa pemrograman Python menggunakan Pustaka (*library*) scikit-learn, pandas, dan matplotlib pada lingkungan Jupyter Notebook.

2.1. Sumber Data

Dataset Heart Failure Prediction diperoleh dari platform Kaggle dan terdiri atas 918 sampel dengan 12 fitur klinis, yaitu usia, anemia, tekanan darah tinggi, kolesterol, diabetes, fraksi ejeksi, status merokok, jumlah platelet, serum kreatinin, serum natrium, jenis kelamin, serta variabel target kejadian kematian. Distribusi awal menunjukkan sekitar 32,6% sampel termasuk ke dalam kelas positif gagal jantung, sedangkan sisanya merupakan kelas negatif. Dataset tidak memiliki nilai hilang (*missing value*), namun menunjukkan distribusi yang tidak normal pada beberapa fitur numerik, seperti usia (mean sekitar 53,5 tahun) dan fraksi ejeksi (mean sekitar 38%). *Data cleaning* dilakukan melalui deteksi outlier menggunakan metode *Interquartile Range* (IQR), pengkodean satu-panas (*one-hot encoding*) untuk variabel kategorik, serta standarisasi skala fitur numerik menggunakan z-score melalui *StandardScaler* agar seluruh fitur berada pada rentang yang sebanding sebelum dilakukan pemodelan clustering.

2.2. Pra-Pemrosesan Data

Sebelum tahap pemodelan, dilakukan augmentasi data untuk menjamin integritas analisis. Tahapan ini meliputi:

- Pembersihan Data, yang dilakukan dengan menghapus observasi yang terdeteksi sebagai outlier ekstrem berdasarkan batas bawah dan batas atas IQR pada fitur numerik utama,

sehingga dihasilkan himpunan data yang lebih representatif dan mengurangi potensi distorsi terhadap pembentukan klaster.

- Transformasi Fitur, yaitu konversi variabel kategorikal seperti jenis kelamin, anemia, diabetes, tekanan darah tinggi, dan kebiasaan merokok menjadi representasi numerik melalui teknik *one-hot encoding*, sehingga dapat diproses oleh algoritma GMM yang bekerja pada ruang fitur numerik.
- Standarisasi yaitu penerapan *StandardScaler* untuk menormalisasi setiap fitur numerik sehingga memiliki rataan nol dan simpangan baku satu, yang penting karena GMM sangat peka terhadap skala fitur dan asumsi distribusi Gaussian multivariat.

2.3. Penentuan Jumlah Cluster Optimal

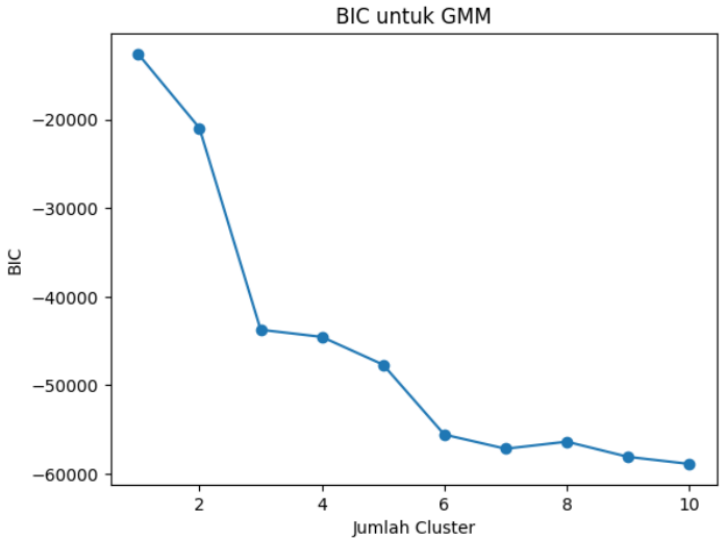
Penentuan jumlah *cluster* (k) yang optimal dilakukan dengan memanfaatkan kriteria informasi *Bayesian Information Criterion* (BIC). Nilai BIC dihitung untuk beberapa kandidat jumlah *cluster* (k = 2 sampai k = 10) pada model GMM, kemudian dibandingkan untuk mengidentifikasi model dengan nilai BIC terendah. Model dengan BIC paling rendah dipilih sebagai representasi terbaik karena memberikan keseimbangan antara kecocokan model (*goodness of fit*) dan kompleksitas parameter, sehingga mengurangi risiko *overfitting* pada data yang relatif terbatas[6], [11]. Sebagai validasi tambahan terhadap pemilihan k, digunakan juga *Silhouette Score* untuk mengukur kualitas pemisahan *cluster*. Kombinasi BIC dan *Silhouette Score* ini memberikan dasar yang kuat secara statistik dalam menentukan konfigurasi klaster yang paling sesuai untuk dataset gagal jantung yang digunakan[2], [6], [11].

3. HASIL DAN PEMBAHASAN

Pada bagian ini disajikan hasil penelitian penerapan algoritma *Gaussian Mixture Model* (GMM) pada dataset *Heart Failure Prediction*, beserta pembahasannya secara komprehensif. Hasil ditampilkan dalam bentuk tabel dan gambar untuk memudahkan pembaca memahami proses pemilihan jumlah klaster dan performa model baseline serta model final. Pembahasan dilengkapi dengan interpretasi statistik dan dikaitkan dengan temuan pada penelitian terdahulu mengenai pemodelan klaster pasien penyakit jantung.

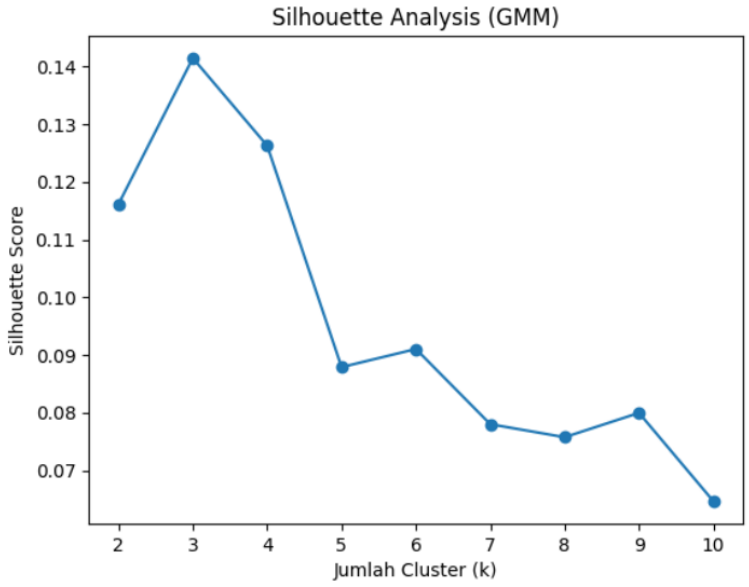
3.1. Gambar dan Table

Pertama, dilakukan evaluasi jumlah klaster kandidat menggunakan kriteria BIC dan *Silhouette Score*. Kurva BIC terhadap jumlah klaster untuk model GMM ditunjukkan pada Gambar 1. Grafik tersebut memperlihatkan bahwa nilai BIC menurun tajam dari k = 1 hingga sekitar k = 3, kemudian masih menurun tetapi dengan kemiringan yang lebih landai hingga k = 10. Pola ini mengindikasikan bahwa penambahan klaster di atas tiga hanya memberikan peningkatan kecocokan model yang relatif kecil dibanding dengan kenaikan kompleksitas parameter[5], [6], [11].



Gambar 1. Kurva BIC untuk berbagai jumlah klaster pada model GMM

Selanjutnya, pemilihan jumlah klaster juga divalidasi menggunakan *Silhouette Analysis*. Nilai *Silhouette Score* untuk k = 2 hingga k = 10 digambarkan pada Gambar 2. Pada grafik tersebut, *Silhouette Score* mencapai puncak di sekitar k = 3 (≈ 0,141) dan menurun untuk k yang lebih besar, sehingga kombinasi BIC yang rendah dan *Silhouette* tertinggi mendukung pemilihan tiga klaster sebagai konfigurasi optimal untuk dataset ini.



Gambar 2. *Silhouette Analysis* GMM pada berbagai jumlah klaster (k)

Berdasarkan hasil evaluasi tersebut, dua model utama dibandingkan: model baseline GMM dengan k = 2 dan model final GMM dengan k = 3. Ringkasan evaluasi keduanya ditunjukkan pada Tabel 1. Terlihat bahwa Silhouette Score model baseline sebesar 0,1161, sedangkan model final meningkat menjadi 0,1415. Peningkatan sekitar 21,8% ini menunjukkan bahwa kualitas pemisahan klaster pada model dengan tiga klaster lebih baik dibandingkan model dengan dua klaster[7].

Tabel 1. Evaluasi Silhouette model baseline dan model final GMM

Model	Jumlah klaster (k)	Silhouette Score
Baseline GMM	2	0,1161
Final GMM	3	0,1415

Selain itu, keluaran model GMM juga menyertakan probabilitas keanggotaan setiap sampel terhadap masing-masing klaster, seperti ditunjukkan oleh kolom Prob_Cluster_0 dan Prob_Cluster_1 dalam tabel output. Beberapa baris awal memperlihatkan nilai Prob_Cluster_1 yang sangat mendekati 1 dengan Prob_Cluster_0 yang sangat kecil (orde 10^{-10} hingga 10^{32}), sedangkan terdapat pula contoh baris dengan Prob_Cluster_0 ≈ 0,9996 dan Prob_Cluster_1 ≈ 0,0004. Hal ini menegaskan bahwa GMM melakukan *soft clustering*, sehingga setiap data tidak hanya ditempatkan secara kaku ke satu klaster, tetapi memiliki distribusi probabilitas keanggotaan[5], [6], [11].

3.2. Pembahasan Hasil Klasterisasi

Hasil kurva BIC dan Silhouette menunjukkan bahwa struktur data pasien pada dataset *Heart Failure Prediction* lebih baik direpresentasikan oleh tiga klaster dibandingkan dua klaster. Secara praktis, hal ini berarti bahwa terdapat setidaknya tiga kelompok pasien dengan pola karakteristik klinis yang berbeda, yang tidak sepenuhnya tertangkap jika hanya menggunakan dua klaster[12]. Pendekatan pemilihan jumlah klaster dengan mengombinasikan BIC dan *Silhouette Score* sejalan dengan rekomendasi pada literatur *model-based clustering* karena mampu menyeimbangkan antara kecocokan model dan kompleksitas, serta memberikan interpretasi geometris mengenai pemisahan antar klaster[6], [11].

Peningkatan *Silhouette Score* dari 0,1161 pada baseline GMM (k = 2) menjadi 0,1415 pada model final GMM (k = 3) mengindikasikan adanya perbaikan kekompakan intra-klaster dan jarak antar klaster di ruang fitur yang telah dinormalisasi. Nilai ini memang masih berada pada kategori pemisahan moderat, yang wajar untuk data klinis heterogen seperti penyakit jantung, tetapi cukup untuk menunjukkan struktur kelompok yang bermakna[13]. Penelitian sebelumnya mengenai fenotipe gagal jantung dengan pendekatan *unsupervised* juga ditemukannya tiga atau lebih subkelompok yang berbeda secara klinis, sehingga jumlah klaster yang dipilih dalam penelitian ini konsisten dengan pola umum pada literatur[10].

Keberadaan probabilitas keanggotaan klaster pada setiap sampel memberikan nilai tambah dari GMM dibandingkan metode *hard clustering* seperti K-Means. Informasi probabilitas ini memungkinkan peneliti atau praktisi klinis mengidentifikasi pasien yang berada di batas antara dua klaster (misalnya probabilitas tersebar cukup merata di dua klaster), yang berpotensi merupakan kasus "abu-abu" dan memerlukan evaluasi klinis lebih lanjut[14]. Pendekatan seperti ini telah direkomendasikan dalam penelitian fenomena gagal jantung modern, di mana ketidakpastian klasifikasi justru memberikan insight tambahan mengenai transisi risiko pasien dari satu fenotipe ke fenotipe lain[15].

4. KESIMPULAN

Penelitian ini bertujuan menerapkan algoritma *Gaussian Mixture Model* (GMM) untuk melakukan *clustering* pada dataset Heart Failure Prediction dan menentukan jumlah klaster yang optimal berdasarkan kombinasi kriteria *Bayesian Information Criterion* (BIC) dan *Silhouette Score*. Tujuan tersebut telah tercapai, sebagaimana ditunjukkan pada bagian hasil dan pembahasan yang memperlihatkan bahwa konfigurasi tiga klaster menghasilkan nilai Silhouette lebih tinggi dibanding model baseline dua klaster, dengan tetap mempertahankan nilai BIC yang relatif rendah. Dengan demikian, pendekatan GMM mampu mengungkap struktur kelompok pasien yang lebih representatif dibandingkan model dengan jumlah klaster yang lebih sedikit.

Secara substansial, penggunaan GMM memberikan gambaran probabilistik keanggotaan klaster bagi setiap pasien, sehingga tidak hanya memetakan mereka ke dalam klaster risiko yang berbeda, tetapi juga menyediakan informasi tingkat keyakinan model terhadap penempatan tersebut. Hal ini berpotensi dimanfaatkan untuk mendukung pengambilan keputusan klinis yang lebih hati-hati, misalnya dengan memberi perhatian khusus pada pasien yang berada di batas antara klaster risiko menengah dan tinggi. Prospek pengembangan penelitian selanjutnya meliputi penambahan fitur klinis lain yang lebih kaya, penerapan validasi eksternal pada rumah sakit berbeda, serta integrasi model GMM dengan algoritma klasifikasi untuk membangun sistem peringatan dini berbasis klaster risiko pada pasien penyakit jantung.

Ucapan Terima Kasih

Penulis menyampaikan penghargaan dan terima kasih kepada pihak universitas serta program studi yang telah memberikan dukungan fasilitas laboratorium, perangkat lunak, dan lingkungan pembelajaran sehingga penelitian ini dapat terlaksana dengan baik.

DAFTAR PUSTAKA

[1] C. Meijs *et al.*, "Discovering Distinct Phenotypical Clusters in Heart Failure Across the Ejection Fraction Spectrum: a Systematic Review," *Curr. Heart Fail. Rep.*, vol. 20, no. 5, pp. 333–349, Oct. 2023, doi: 10.1007/s11897-023-00615-z.

[2] J. Sun *et al.*, "Identifying novel subgroups in heart failure patients with unsupervised machine learning: A scoping review," *Front. Cardiovasc. Med.*, vol. 9, p. 895836, July 2022, doi: 10.3389/fcvm.2022.895836.

[3] M. A. Mohammad, "Advancing heart failure research using machine learning," *Lancet Digit. Health*, vol. 5, no. 6, pp. e331–e332, June 2023, doi: 10.1016/S2589-7500(23)00085-7.

[4] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, "Clustering of Heart Failure Phenotypes in Johannesburg Using Unsupervised Machine Learning," *Appl. Sci.*, vol. 13, no. 3, p. 1509, Jan. 2023, doi: 10.3390/app13031509.

[5] J. C. Jentzer *et al.*, "Machine Learning Approaches for Phenotyping in Cardiogenic Shock and Critical Illness," *JACC Adv.*, vol. 1, no. 4, p. 100126, Oct. 2022, doi: 10.1016/j.jacadv.2022.100126.

[6] J. C. Jentzer *et al.*, "Unsupervised machine learning to identify subphenotypes among cardiac intensive care unit patients with heart failure," *ESC Heart Fail.*, vol. 11, no. 6, pp. 4242–4256, Dec. 2024, doi: 10.1002/ehf2.15027.

[7] M. Karaçam *et al.*, "From patterns to prognosis: machine learning–derived clusters in advanced heart failure," *Front. Cardiovasc. Med.*, vol. 12, p. 1669538, Oct. 2025, doi: 10.3389/fcvm.2025.1669538.

[8] V. Potoupni *et al.*, "Machine-Learning-Driven Phenotyping in Heart Failure with Preserved Ejection Fraction: Current Approaches and Future Directions," *Medicina (Mex.)*, vol. 61, no. 11, p. 1937, Oct. 2025, doi: 10.3390/medicina61111937.

[9] Y. Hu *et al.*, "Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 309, Dec. 2024, doi: 10.1186/s12874-024-02422-z.

[10] T. Rastogi *et al.*, "Identifying congestion phenotypes using unsupervised machine learning in acute heart failure," *Eur. Heart J. - Digit. Health*, vol. 6, no. 5, pp. 907–918, Sept. 2025, doi: 10.1093/ehjdh/ztaf065.

[11] R. R. Sahoo, S. Bhowmick, D. Mandal, and P. Kumar Kundu, "A novel approach of Gaussian mixture model-based data compression of ECG and PPG signals for various cardiovascular diseases," *Biomed. Signal Process. Control*, vol. 96, p. 106581, Oct. 2024, doi: 10.1016/j.bspc.2024.106581.

[12] E. Ntalianis *et al.*, "Feature-based clustering of the left ventricular strain curve for cardiovascular risk stratification in the general population," *Front. Cardiovasc. Med.*, vol. 10, p. 1263301, Nov. 2023, doi: 10.3389/fcvm.2023.1263301.

[13] E. Bresso *et al.*, "Unsupervised machine learning for cardiovascular disease: A framework for future studies," *Eur. J. Heart Fail.*, p. ejhf.70076, Nov. 2025, doi: 10.1002/ejhf.70076.

[14] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Sci. Rep.*, vol. 14, no. 1, p. 23277, Oct. 2024, doi: 10.1038/s41598-024-74656-2.

[15] J. De Bie, "Expertly used unsupervised clustering provides clinical tools as well as insight," *Eur. Heart J. - Digit. Health*, vol. 6, no. 3, pp. 311–312, May 2025, doi: 10.1093/ehjdh/ztaf015.