Latent ECG Signatures for the Identification of High-Risk Phenotypes in Coronary Artery Disease

Josseline Madrid[1], Sara Artal[1], Patricia B Munroe[2], Ana Mincholé[1, 3], Julia Ramírez[1, 2, 3]

[1]IIS, I3A, University of Zaragoza, Zaragoza, Spain
[2] William Harvey Research Institute, Barts and the London Faculty of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom
[3]Centro de Investigación Biomédica en Red – Biomateriales, Bioingeniería y Nanomedicina

Abstract

ECG-derived deep learning embeddings via transfer learning have improved disease prediction, but their utility in identifying cardiovascular risk profiles remains unexplored. We aimed to extract latent features from a pre-trained model to refine coronary artery disease (CAD) risk phenotypes and study their clinical interpretability.

Latent features were extracted from 10-second, lead-I ECG recordings of 1,928 individuals with CAD from the UK Biobank using a Contrastive Predictive Coding model. Principal components (PCs) were calculated and used as input to an unsupervised clustering model. Then, their association with atrial fibrillation (AF) and heart failure (HF) and with traditional ECG parameters was analyzed.

The latent PCs identified two clusters, with Cluster 2 exhibiting higher incidence of events (AF:8.9% and HF:6.6%). Although cluster-based risk stratification was less pronounced than with traditional parameters, individual PCs demonstrated robust predictive ability. Furthermore, predictive PCs correlated with slower ventricular conduction and prolonged repolarization, established markers of electrical remodeling.

Our findings confirm that latent features hold discriminative value for cardiovascular outcomes in CAD populations. Future work should explore alternative clustering architectures to enhance the risk stratification.

1. Introduction

Coronary artery disease (CAD) represents a major health burden worldwide[1]. As a highly heterogeneous condition, CAD manifests through diverse clinical pathways across individuals, with atrial fibrillation[2] (AF) and heart failure[3] (HF) representing two of its most common and worsened outcomes. Therefore, developing accessible and non-invasive ECG-based risk stratification strategies is essential for enabling large-scale screening and targeted prevention[4].

In previous work[4], we developed and validated an unsupervised clustering approach that identified distinct subgroups among individuals with CAD. One of these subgroups exhibited specific ECG characteristics (i.e., QRS width, QT and Tpe intervals) that were strongly associated with HF risk. This model included time-interval and morphological advanced ECG-derived parameters. However, while these standard metrics are clinically robust, they may overlook subtle, latent interactions within the ECG signal that contain further prognostic information.

The emergence of deep learning has revolutionized signal processing by enabling the automated extraction of complex features that hold biological relevance[5] and prognostic value[6]. Recent studies have shown that utilizing large-scale pre-trained models, can generate ECG-derived embeddings that enhance the performance of both supervised[7–9] and unsupervised tasks[6,10–12]. These latent representations often outperform traditional metrics in detecting subclinical cardiac variations[11], offering a more nuanced approach to risk stratification.

We hypothesize that latent features extracted from a pre-trained deep learning model[7] can capture subclinical cardiac signatures that refine the CAD phenotypes identified in our previous work. Moreover, we aim to investigate whether these latent features hold information on the future incidence of cardiac outcomes. Finally, to ensure clinical interpretability, we aim to correlate these latent features with established ECG morphological parameters used on our previous work[4].

2. Methods

2.1. Study Population

Our study population included 1,928 individuals with a diagnosis of CAD from the UK Biobank (UKB) cohort[13], exclusion criteria and outcomes definition are previously described [4]. Available information included 10-second ECG signals recorded at rest, electronic health records and follow-up data (4-years). This work was conducted under application number 8256.

2.2. ECG processing and latent feature extraction

Table 1. Incidence of AF and HF diagnoses.

	Cluster1 (n=1533)		Cluster 2 (n=395)		Fisher's
Outcome	**n**	**%**	**n**	**%**	**P-value**
HF	54	3.5%	26	6.6%	**0.01**
AF	66	4.3%	35	8.9%	**< 0.001**

We performed minimal signal pre-processing on the I-lead 10-second ECG recordings, which included low-pass filtering and baseline wander removal, as described in [4]. The signals were subsequently segmented into 3-second windows. Signal processing analysis was performed using MATLAB (version R2023b).

A Contrastive Predicting Coding (CPC) unsupervised model, pretrained on approximately 1.6 million ECG recordings from various public access datasets, was employed to extract latent features on each ECG segment. The model architecture consisted of 6 convolutional layers (kernel size 3, stride 2, and 1,024 channels), a context aggregator consisting of a Gated Recurrent Unit (GRU) layer, and a prediction head that included two convolutional layers interleaved with a ReLU activation layer. Specifically, this contrastive approach learns representations by predicting the latent embeddings of the future eight-time steps of ECG signals from past context. The methodology of this model is described thoroughly in [7], our implementation specifically utilized only lead I for training and feature extraction. We obtained the latent context features from the output of the GRU layer and averaged them across the temporal dimension. These features were then standardized using z-scores and used for all subsequent analyses.

2.3. Unsupervised clustering of latent space

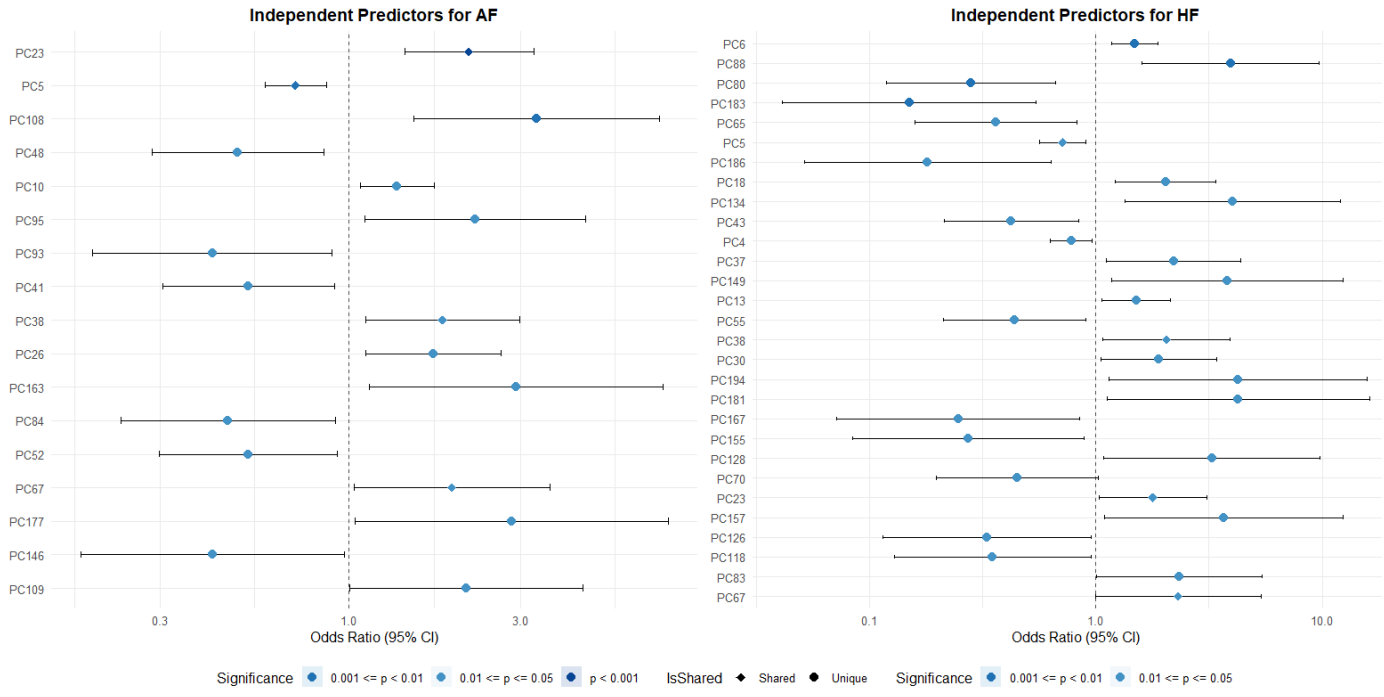
To reduce signal noise and intra-subject variability, we calculated mean latent features of the three segments for each subject. Dimensionality reduction was then performed using principal component analysis, retaining the number of components required to explain 90% of the total variance. The resulting feature space was used as input for a k-means clustering algorithm, where k was the optimal number of clusters determined using the Silhouette method, replicating the design of our previous work [4]. Clustering analysis was conducted using MATLAB (version R2023b).

2.4. Association with future risk and clinical interpretability

The association between the identified clusters and incident heart failure (HF) and atrial fibrillation (AF) was evaluated using Fisher's exact test. Differences in clinical outcomes are reported as absolute numbers and percentages. Statistical significance was defined as $P<0.05$. Next, we evaluated the combined ability of the latent features on predicting incident outcomes by constructing multivariable logistic regression models using the retained PCs as predictors.

Finally, to enhance clinical interpretability of the

Figure 1. Multivariable logistic regression models of PCs latent features for AF and HF.



derived latent space we analyzed the correlation between the latent PCs and the ECG parameters used in [4]. To further elucidate the physiological characteristics captured by the latent PCs, we generated representative median heartbeats for individuals at the distributional extremes (5th vs 95th percentiles) of each significant PC. To ensure a robust morphological comparison the heartbeats from both groups were aligned by their R-peaks.

3. Results

The study population had a median age of 70 years and was 69.7% male. During the follow-up time, 101 cases of AF and 80 cases of HF were diagnosed.

From each 3-second ECG segment, 1024 latent context features were extracted. After dimensionality reduction, 194 PCs were retained and used as input for the clustering model, which identified two clusters with different sizes: Cluster 1 (n=1,533) and cluster 2: (n=395). Baseline demographic and clinical characteristics (i.e., age, body mass index, blood pressure, sex distribution, diabetes diagnosis, alcohol/tobacco consumption) were similar between clusters. However, these clusters showed significant differences in the association with incident outcomes. Cluster 2 exhibited significantly higher incidence of both AF (8.9%, $P < 0.001$) and HF (6.6%, $P=0.01$) compared to Cluster 1 (**Table 1**).

Multivariable logistic regression models assessed the combined predictive power of the latent PCs. The most significant features for AF included PC23, PC5, PC108, PC48, and PC10, while for HF we identified PC6, PC88, PC183, PC80, and PC5. Interestingly, we observed several PCs were significantly associated for both outcomes, such as PC5 exhibiting a negative effect and PC23, PC38 and PC67 showed positive effects. These findings suggest a shared latent cardiovascular risk profile. Estimates and P-values are described in **Figure 1**.

Figure 2 illustrates the correlation between the most significant latent features and the ECG parameters calculated in [4]. PC5 correlated with QRS and T-wave amplitude and morphology, and ST-segment amplitude, while PC23 showed association with QT and RR interval. Weaker correlations were observed for the PC38 with QRS width and PC67 with Tpe interval and QRS amplitude. This correlation results are consistent with the morphological features observed in **Figure 3**, where the median heartbeats for the 5th and 95th percentile for the PC6 and PC23 are illustrated.

4. Discussion and Conclusions

Our main finding is that latent features extracted from a CPC model can identify two distinct clusters with significantly different associations with incident AF and HF conditions. Despite both clusters having similar demographical and clinical profiles, individuals in cluster 2 exhibited almost a two-fold incidence of events.

However, we observed the risk stratification achieved in our previous work with ECG parameters was more pronounced[4]. This might suggest that for the specific k-means clustering algorithm, ECG parameters may more directly capture the variance required to define distinct clinical subgroups. This discrepancy in clustering performance may stem from the intrinsic nature of the latent features. Unlike the morphological and time-interval parameters used in our previous work, the CPC-derived embeddings are based on correlation on future steps[7]. Consequently, these features may lack the specific type of morphological information captured by our ECG features.

Our analysis revealed a complex risk architecture where specific PCs were uniquely associated with AF or HF, while others formed a shared latent risk profile. The

Figure 2. Correlation of PCs latent features and ECG parameters.

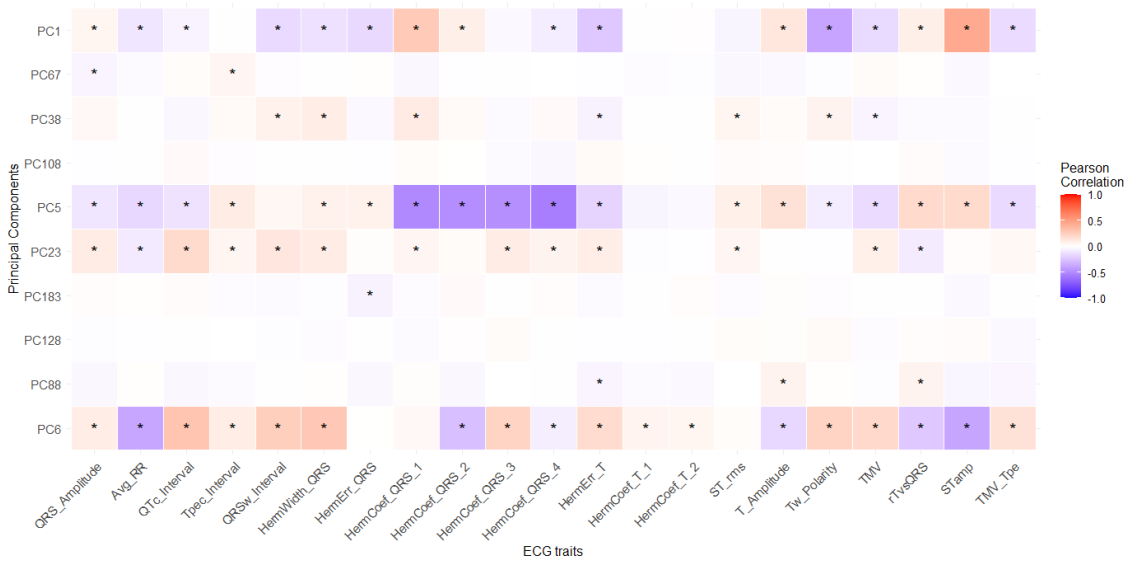
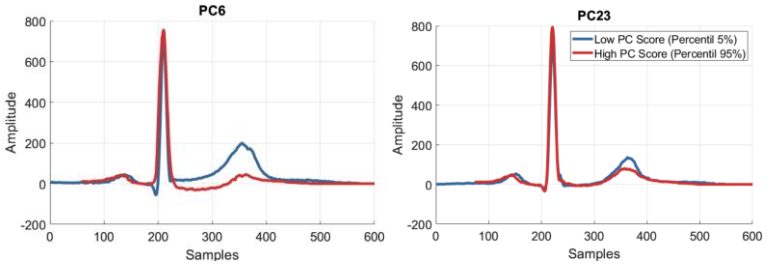


Figure 3. Median heartbeat of PCs 23 and 6.



correlation analysis revealed that several latent features mapped directly with ECG parameters. The PCs that were most significant in the association with risk of HF were correlated with shortened RR intervals, inverted T-wave polarity, prolonged QT intervals, and widened QRS duration. These associations indicate a slower ventricular conduction and prolonged repolarization pattern, established markers of electrical remodeling and progression to HF[3,4,14,15]. Similarly for AF, strong correlation was observed with prolonged QT intervals, widened QRS duration and shortened RR intervals, consistent with a pattern of electrical instability which may lead to AF[16]. Notably, some of the PCs significant for AF did not show correlation with ECG parameters which might be due to the absence of atrial-specific parameters.

As a limitation of this study, we mention the pre-training objective of the CPC model. Unlike autoencoders[12] or supervised models trained on specific diagnostic labels/diseases[6,8], the CPC task focuses on predicting the latent representations of the next eight temporal steps of the signal. While this approach yields a robust representation of signal dynamics, it may not focus on the morphological features required for clinical stratification as effectively as a task-specific model.

Given the configuration of k-means, which seeks to create spheric clusters[17], this may not be the optimal algorithm for the high-dimensional, non-linear structure of the latent space. Other studies have used density-based[11] or Gaussian Mixture Models[10] for the clustering task which might offer better partition this complex feature space. Future studies should explore whether alternative clustering architectures can improve the risk stratification. Additionally, to enhance the clinical interpretability future work should investigate the correlation with atrial metrics.

Acknowledgments

This work was supported by projects PID 2021-128972OA-I00, CNS2022-135899, CNS2023-143599, PID2023-148975OB-I00 and TED2021-130459B-I00, funded by the Spanish Ministry of Science and Innovation (MCIN) and by Gobierno de Aragón, and by fellowships RYC2019-027420-I and RYC2021-031413-I from MCIN. PBM acknowledges the support of the National Institute for Health and Care Research Barts Biomedical Research Centre (NIHR203330).

References

1. Martin SS, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, et al. 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. Circulation. American Heart Association; 2024;149:e347–913. https://doi.org/10.1161/CIR.0000000000001209

2. Batta A, Hatwal J, Batta A, Verma S, Sharma YP. Atrial fibrillation and coronary artery disease: An integrative review focusing on therapeutic implications of this relationship. World J. Cardiol. Baishideng Publishing Group Inc; 2023. p. 229–43. https://doi.org/10.4330/wjc.v15.i5.229

3. Gheorghiade M, Sopko G, Luca L De, Velazquez EJ, Parker JD, Binkley PF, et al. Navigating the crossroads of coronary artery disease and heart failure. Circulation. 2006. p. 1202–13. https://doi.org/10.1161/CIRCULATIONAHA.106.623199

4. Madrid J, Young WJ, van Duijvenboden S, Orini M, Munroe PB, Ramírez J, et al. Unsupervised clustering of single-lead electrocardiograms associates with prevalent and incident heart failure in coronary artery disease. European Heart Journal - Digital Health [Internet]. 2025;ztaf013. https://doi.org/10.1093/ehjdh/ztaf013

5. Yeung MW, van de Leur RR, Benjamins JW, Vessies MB, Ruijsink B, Puyol-Antón E, et al. Deep representation learning of electrocardiogram reveals biological insights in cardiac phenotypes and cardiovascular diseases. iScience [Internet]. Elsevier Inc.; 2025 [cited 2026 Apr 11];28:113226. https://doi.org/10.1016/J.ISCI.2025.113226

6. Sau A, Ribeiro AH, McGurk KA, Pastika L, Bajaj N, Gurnani M, et al. Prognostic Significance and Associations of Neural Network-Derived Electrocardiographic Features. Circ Cardiovasc Qual Outcomes [Internet]. Lippincott Williams and Wilkins; 2024 [cited 2025 Jul 4];17:e010602. https://doi.org/10.1161/CIRCOUTCOMES.123.010602/SUPPL_FILE/CIRCCVQO_CIRCCQO-2023-010602-T_SUPP3.PDF

7. Artal S, Ramírez J, Miguel A, Martínez JP. Mejora de la clasificación de latidos cardíacos en señales ECG mediante preentrenamiento autosupervisado a gran escala.

8. Weimann K, Conrad TOF. Transfer learning for ECG classification. Scientific Reports 2021 11:1 [Internet]. Nature Publishing Group; 2021 [cited 2026 Apr 11];11:5251-. https://doi.org/10.1038/s41598-021-84374-8

9. Nguyen C V., Do CD. Transfer learning in ECG diagnosis: Is it effective? PLoS One [Internet]. Public Library of Science; 2025 [cited 2026 Apr 11];20:e0316043. https://doi.org/10.1371/JOURNAL.PONE.0316043

10. Nezamabadi K, Sardaripour N, Haghi B, Forouzanfar M. Unsupervised ECG Analysis: A Review. IEEE Rev. Biomed. Eng. Institute of Electrical and Electronics Engineers Inc.; 2023. p. 208–24. https://doi.org/10.1109/RBME.2022.3154893

11. Oskooei A, Chau SM, Weiss J, Sridhar A, Martínez MR, Michel B. DeStress: Deep Learning for Unsupervised Identification of Mental Stress in Firefighters from Heart-Rate Variability (HRV) Data. Studies in Computational Intelligence [Internet]. Springer, Cham; 2021 [cited 2026 Apr 11];914:93–105. https://doi.org/10.1007/978-3-030-53352-6_9

12. Porumb M, Griffen C, Hattersley J, Pecchia L. Nocturnal low glucose detection in healthy elderly from one-lead ECG using convolutional denoising autoencoders. Biomed Signal Process Control [Internet]. Elsevier; 2020 [cited 2026 Apr 11];62:102054. https://doi.org/10.1016/J.BSPC.2020.102054

13. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med [Internet]. Public Library of Science; 2015 [cited 2023 May 22];12:e1001779. https://doi.org/10.1371/JOURNAL.PMED.1001779

14. Iuliano S, Fisher SG, Karasik PE, Fletcher RD, Singh SN. QRS duration and mortality in patients with congestive heart failure. Am Heart J [Internet]. 2002;143:1085–91. https://doi.org/https://doi.org/10.1067/mhj.2002.122516

15. Zègre-Hemsey JK, Hogg M, Crandell J, Pelter MM, Gettes L, Chung EH, et al. Prehospital ECG with ST-depression and T-wave inversion are associated with new onset heart failure in individuals transported by ambulance for suspected acute coronary syndrome. J Electrocardiol [Internet]. 2021;69:23–8. https://doi.org/https://doi.org/10.1016/j.jelectrocard.2021.08.005

16. Niu J, Zhang M, Liu P, Hua C, Zhong G. Research progress on predicting atrial fibrillation recurrence after radiofrequency ablation based on electrocardiogram-related parameters. J Electrocardiol [Internet]. 2023;81:146–52. https://doi.org/https://doi.org/10.1016/j.jelectrocard.2023.08.015

17. Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Inf Sci (N Y) [Internet]. Elsevier; 2023 [cited 2026 Apr 12];622:178–210. https://doi.org/10.1016/J.INS.2022.11.139

Address for correspondence:
Josseline Madrid, jmadrid@unizar.es
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.