Reviewer #1: This study aimed to identify clinical phenotypes of heart failure patients who respond favorably to cardiac resynchronization therapy (CRT) using unsupervised machine learning (hierarchical clustering) combined with gated SPECT imaging. A total of 217 patients were analyzed using 45 pre-treatment variables, including clinical, demographic, and imaging-based measures of dyssynchrony, fibrosis, and cardiac remodeling. Three phenotypes were identified with differing CRT response rates, suggesting the potential value of SPECT phase analysis for individualized patient selection.

1. The dataset was derived from three sources, yet no description of inter-dataset differences is provided. Variations in patient characteristics, imaging protocols, or follow-up procedures could impact clustering results. It also remains unclear whether data harmonization was performed prior to integration.

Thank you for the valuable comment regarding dataset characterization.

To address the inter-dataset difference, we included a new table (Table 1) which provides comprehensive summary of inter-dataset differences. We added to discussion an analysis of the resulting table. The added paragraphs are in red in the revised manuscript and I am also pasting here: “An analysis of inter-dataset differences, as presented in Table 1, reveals significant variations (p<0.05) in patient age and racial distribution across the three sources. Furthermore, differences were observed in several SPECT features, including geometric parameters, SRScore, stroke volume, PBW and scar quantification. Conversely, some characteristics demonstrated no significant differences, such as LVEF, QRS duration, PSD and the proportion of NYHA III patients. These findings indicate a degree of heterogeneity within our final dataset, with some shared features.”

Regarding your question about data harmonization, we state that "Data harmonization techniques were not applied to the supplied dataset. All the values were already standardized."

The imaging protocols for each source of our dataset was included in the Methods section: “All patients from the three sources that comprise our dataset underwent a gated-SPECT Myocardial Perfusion Imaging (MPI) within 7 days before CRT implantation. For VISION-CRT the protocol involved a resting gated-SPECT scan 30 minutes after the injection of 20-30 mCi of 99mTc-sestamibi using dual-headed cameras with high-resolution low-energy collimators, with a 180° orbit, with 8 or 16 frames ECG-gated. (21). For Taiwan-CRT the protocol consisted of a resting gated-SPECT 60-90 minutes after the injection of 20-30 mCi 99mTc-sestamibi using a dual-headed camera, with a 180° orbit, with 8 frames ECG-gated (17). For GUIDE-CRT the protocol was a resting gated-SPECT 60-90 minutes after the injection of 25-30 mCi of 99mTc-MIBI using a dual-headed or triple-headed camera with high-resolution low-energy collimators, with pixel size of 64x64 and 8 frames ECG-gated (16,22).”

The imaging processing method and tools was included in the Methods:” All the SPECT image processing was performed by the same operator using Emory Cardiac Toolbox (ECTb4, Atlanta, GA). Phase analysis has been shown to have sufficient temporal resolution (26), high reproducibility and repeatability (27,28), and robustness with low counts (29).”

We believe these additions and clarifications enhance the transparency and comprehensibility of our dataset and its potential impacts on the clustering results.

2. No internal validation metrics or assessments of cluster stability are reported. Given the multi-cohort design, such evaluations would strengthen the robustness and generalizability of the findings.

Thank you for your comment regarding the need for internal validation and assessment of cluster stability. We have incorporated additional stability and cluster validity assessments into the Methods and Results sections.

In the Methods section, we now specify that "Cluster validity was assessed using the silhouette score, and cluster stability was evaluated by rerunning the algorithm on rearranged datasets to confirm consistent dendrogram structures."

In the Results, in addition to the initial analysis, we now provide the silhouette scores for the different cluster arrangements (3, 4, and 5 clusters) in Table 2. Although the four-cluster arrangement did not present the best silhouette score, it was chosen because it provided very differing response rates, which is aligned with our objective of identifying phenotypes with distinct CRT responses.

Additionally, for dendrogram stability, we state that "We also validated this dendrogram by shifting and rearranging the observations in the dataset, all resulting in the same dendrogram." This verification confirms the consistency of the dendrogram structure and, consequently, the stability of the identified clusters.

With these additions we expect to state that internal validation and cluster stability were considered and reported.

3. The analysis relies solely on hierarchical clustering. A comparison with alternative unsupervised methods could help assess the consistency and reproducibility of the identified phenotypes.

Thank you for your comment regarding the use of only hierarchical clustering.

In the Methods section, we revised the rationale for choosing hierarchical clustering (HC) and added the following explanation: "Among unsupervised learning methods, HC offers distinct advantages, including the ability to visualize data relationships with a dendrogram, the flexibility to explore data without predefined cluster numbers, and the potential to identify nested clusters within larger groups. It is particularly useful when the data structure is hierarchical or when the number of clusters is unknown beforehand. This approach directly aligns with our study's context, as current guidelines reveal a clear hierarchical structure in CRT patient selection, and the potential number of patient subgroups is undefined. Accordingly, only hierarchical clustering was applied in this manuscript."

In the Discussion section, we compared our approach with other studies that used unsupervised learning methods, such as Feeny et al. and He et al., who employed PCA and K-means, and an unsupervised deep learning autoencoder, respectively. Furthermore, we established a detailed comparison with the work of Cikes et al. , who also used K-means, highlighting the similarities and differences in our findings and methodologies.

While our study focused exclusively on hierarchical clustering due to its advantages and alignment with the structure of our problem, the expanded discussion and comparisons with other unsupervised methods show that other approaches generate coherent results in the context of CRT. We believe that including a comparison of various unsupervised learning methods is beyond the objectives of this work, which aimed to correlate the findings of the chosen method with clinical implications. A comparative study of diverse methods would be more extensive and is not in the scope of this work, being considered for future research.

We believe that explanation could clarify our option for employing only HC and address this comment.

4. While the phenotypes are described, the clinical implications remain limited. Further elaboration in the discussion section on how these clusters could guide prospective CRT decision-making would enhance the translational relevance of the findings.

Thank you for your comment regarding the need to deepen the clinical implications of the identified phenotypes. We agree that a elaboration in the Discussion section on how these clusters could guide prospective CRT decision-making will enhance the translational relevance of our findings.

In the Discussion section, we have expanded the interpretation of their clinical implications:

"The results obtained from this study are relevant for a more precise characterization of patient phenotypes that have the greatest potential to benefit from CRT.".

"If treatment decisions were based purely on strong guideline recommendations, many patients with QRS > 150ms and LVEF < 35% would be referred for CRT, however a positive response rate to treatment could be lower than expected for a specific group.".

"The results suggest that assessing cardiac dyssynchrony, the presence of fibrotic areas and the degree of cardiac remodeling is important to predict the potential benefit of CRT.".

We highlighted the multifactorial approach of our study: "Our study combines clinical features with data on mechanical dyssynchrony, cardiac remodeling, and myocardial fibrosis, offering a more comprehensive framework than traditional isolated metrics like QRS duration and LVEF.".

We provided examples of the identified phenotypes and their implications: "For example, cluster 2 that presented only borderline dyssynchrony and non-dilated hearts showed a high CRT response rate (86.2%).". "In contrast, cluster 4, characterized by extensive fibrosis and moderate dyssynchrony, had lower response rates, despite meeting standard CRT eligibility criteria.".

Finally, we elaborated on how our phenotype-based selection strategy can impact clinical practice and cost-effectiveness: "This suggests that phenotyping can identify patients likely to benefit most from CRT and those for whom the procedure may be futile.". "The machine learning–derived clusters identified in our study may enhance CRT decision-making by integrating clinical data with gated SPECT-derived measures of mechanical dyssynchrony, fibrosis, and cardiac remodeling.". "These clusters can help to differentiate patients with a high likelihood of CRT response, like patients in cluster 2, from those with limited benefit, like patients in cluster 4. This refined, phenotype-based selection strategy has potential implications for cost-effectiveness, especially in middle- and low-income countries where CRT devices represent a major financial investment as demonstrated by Bertoldi et al .". "By reducing non-responder rates and guiding therapy to prioritize patients most likely to benefit could support an efficient use of limited healthcare resources.".

We believe these additions strengthen the clinical implications of our findings, highlighting how the identified phenotypes could be used for more precise and personalized decision-making in CRT indication.