Dear Reviewers,

We are grateful for the comments and suggestions that you have provided.

We have revised the manuscript to address the reviewers’ comments. Our reply to the comments is summarized below. In this document, the original reviewers’ comments are in black; our responses are in blue, and the revised manuscript is red. Old unchanged portions of the original manuscript are in orange.

**Editor:**

Comments to the author (if any):  
Additional points to address:  
1)      Intro needs lit review noting prior study limitations to provide rationale for your approach.

For the previous studies noted in the introduction, we have added their limitation

ns to provide rationale for our methodology. See below for the addition:

Machine learning (ML) encompasses a group of techniques which empower computers to learn tasks without explicit commands by drawing upon data to automatically discover meaningful patterns which can generate predictions at an individual level [8]. Previous studies have utilized ML to predict CRT outcomes using a variety of different features. Tokodi et al. [9] applied ML to predict all-cause mortality of CRT patients; Feeny et al. [10] and Kalscheur et al. [11] applied ML to predict CRT response using clinical features and QRS morphology [12]; However, these models are limited in that they rely on data from a single time point and incorporate only a narrow subset of features derived from various medical assessments. For a more accurate and clinically relevant modeling of CRT response within a medical decision-support framework, it is important to account for both the sequential nature in which diagnostic information is obtained and the associated costs of acquiring each test.

2)      Verify references conform to journal format requirements  
The references have been reformatted to conform to the journal requirements. Most notably references to the corresponding journal have been abbreviated according to the National Library of Medicine abbreviation style. Moreover, the format corresponding to the volume and page number have been changed. See the reference sections for the relevant changes.

**Reviewer 1:**  
Reviewer #1: This research is novel in its multi-stage machine learning framework for CRT response prediction, leveraging uncertainty quantification to optimize data acquisition. Unlike traditional models that use all available patient data, this approach selectively determines whether additional SPECT MPI imaging is needed, reducing unnecessary costs while maintaining predictive performance. The model achieves accuracy comparable to a fully data-driven approach while requiring SPECT MPI for only 52.7% of patients, demonstrating clinical efficiency. While similar adaptive ML methods exist in other fields, this is innovative in CRT prediction. Further validation with multi-center datasets would strengthen its impact.  
  
Specific comments:  
1. Comparing this work against state-of-the-art CRT response prediction models (especially recent deep learning-based approaches) would further validate its novelty.

We appreciate this valuable suggestion. To further validate the novelty of our approach, we have introduced another comparison multi-stage model that utilizes a multi-layered perceptron model neural network as its base model is introduced in the methods section and its results presented in the results section. A section has also been added to the discussion for comparison against the elastic-net logistic regression model. Table 2 and Table 3 have been updated to include the performance metrics.

Methods:

Considering the growing attention and demonstrated capabilities of neural networks in clinical prediction tasks, a multilayer perceptron (MLP) model was also incorporated as a base learner model with corresponding Ensemble models and a multi-stage model.

Results:

The multi-stage (MLP) model had AUC 0.70, accuracy 0.64, sensitivity 0.62, and specificity 0.67. Its Ensemble 1 had AUC 0.68, accuracy 0.64, sensitivity 0.68, and specificity 0.60 while Ensemble 2 had AUC 0.74, accuracy 0.65, sensitivity 0.67, and specificity 0.63. This pattern of performance between the Ensemble models and the multi-stage model follows similarly compared to the ENET base models.

Discussion:

Overall, the multi-stage (ENET) model showed improved performance over the MLP version across all metrics; however, this difference in performance is stemming from differences in sample efficiency considering the low sample size of this analysis. Nevertheless, neural network models are more flexible with their input data and could be trained with rich non-structured data, such as SPECT MPI polar maps to enhance performance.

2. Additional validation on larger, multi-center datasets would strengthen its clinical applicability.

Unfortunately, we were not able to collect an external validation dataset due to the unavailability of compatible public datasets and data privacy restrictions.

3. Please use bullets to summarizes your contributions.

Thank you for this feedback. We have changed the author’s contributions to use bullet points in the title page. The changes are below:

Kristoffer Larsen

* Conception design
* Analysis and interpretation of data
* Drafting of the manuscript and revising it critically for important intellectual content
* Final approval of the submitted manuscript.

Chen Zhao

* Conception design
* Revising the manuscript for important intellectual content.

Zhuo He

* Conception design
* Revising the manuscript for important intellectual content.

Joyce Keyak

* Revising the manuscript for important intellectual content.

Qiuying Sha

* Analysis and interpretation of data
* Revising the manuscript for important intellectual content.

Diana Paez

* Active involvement in collecting data and participation in data analysis
* Revising the manuscript for important intellectual content.

Xinwei Zhang

* Active involvement in collecting data
* Revising the manuscript for important intellectual content.

Guang-Uei Hung

* Active involvement in collecting data
* Revising the manuscript for important intellectual content.

Jiangang Zou

* Active involvement in collecting data
* Revising the manuscript for important intellectual content.

Amalia Peix

* Active involvement in collecting data
* Revising the manuscript for important intellectual content.

Weihua Zhou

* Conception design
* Analysis and interpretation of data
* Revising the manuscript for important intellectual content
* Final approval of the submitted manuscript.

4. The machine learning Technologies need to be described in details in the paper.

A section has been added to the Methods section under multi-staged modeling which explains further details regarding the ML technology. See below for the addition:

The ML models used the Scikit-learn package version 1.2.2 [1]. Hyperparameter tuning was performed using Keras-Tuner; the elastic-net logistic regression base model used in the ensembles had its L1 and L2 ratios tuned over a large grid.

[1] Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python”. JMLR, 12: 2825-2830, 2011.

5. Please describe how do you train your model, what is the training dataset and test dataset.

Thank you for this comment, we have added additional information to the methods section under multi-stage modeling. See below for the changes:

The overall data modeling pipeline is shown in Fig. 2. A nested cross-validation structure was used where the outer training fold (train (9/10): n = 196, test (1/10): n = 22) was initially processed before two separate validation sets were sliced off (remaining train: n = 156, validation 1: n = 20, validation 2: n = 20). The outer training fold was then split again into an inner 5-fold stratified shuffle split cross validation (train (9/10): n = 176, test (1/10): n = 20), for hyperparameter tuning then training on the whole outer training fold with the optimal hyperparameter configuration. This is repeated for both ensembles’ models. Afterward, one of the validation sets is used for training of multiple multi-staged models each with different scaling hyperparameters which is then evaluated on the second validation set. The highest performing multi-stage model is then evaluated on the outer test fold for the final performance metrics.

**Reviewer 3:**  
Reviewer #3:

1. Include a concise literature-based classification of recent (2021-2025) CRT outcome prediction methods. Suggested thematic structure:  
\*       Traditional ML with full-data reliance (e.g., logistic regression, SVMs):  
\*       Deep learning-based prognostic models:  
\*       Feature selection and reduction for CRT decision support:  
\*       Multimodal imaging integration (e.g., echocardiography, SPECT, MRI):  
\*       Sequential decision models (cost-sensitive or cascading approaches):  
\*       Models with uncertainty quantification or probabilistic outputs:  
Please cite at least 10 relevant works post-2021 in a synthesized paragraph in the Introduction or Related Work section.

We have added a synthesized and thematically organized summary of recent CRT outcome prediction methods (2021–2025) to the Introduction section in addition to the already noted ML based methodologies for CRT prediction. The added paragraph briefly reviews relevant works across six suggested categories: (1) traditional machine learning approaches relying on full data input, (2) deep learning-based models, (3) methods focused on feature selection and reduction, (4) multimodal imaging integration, (5) sequential decision-making approaches including cost-sensitive and cascading models, and (6) models incorporating uncertainty quantification.

Recent efforts to predict CRT outcomes using machine learning span a wide spectrum of modeling strategies. Traditional ML models such as logistic regression and SVMs have been employed with full-data reliance on clinical and imaging features, including work by Howell et al. and Manohar et al., who used combinations of clinical, ECG, echocardiographic, and CT-derived parameters for outcome prediction [13], [14]. Deep learning-based prognostic models have emerged using complex input representations, such as Chang et al.’s use of synthetic echocardiographic strain traces to train convolutional neural networks with strong predictive performance [15] and Larsen et al.’s fusion of SPECT MPI images with clinical and ECG features [16]. Other studies focus on feature selection and dimensionality reduction, such as Nejadeh et al., who proposed methods to identify the most informative feature subset for downstream prediction tasks [4]. The integration of multimodal imaging, including MRI, CT, and SPECT with clinical and ECG data, has also shown promise, as seen in works by Khamzin et al. and Haque et al. [5], [6]. To address clinical workflow realities, sequential decision models have gained attention, where information is acquired in stages, considering cost and diagnostic value; for instance, Khamzin et al. applied model-derived features from simulated electrophysiology to augment initial data [5]. Finally, a few recent approaches have explored uncertainty quantification, though rarely in CRT specifically, making our contribution in this area timely. Collectively, these studies highlight the growing interest in predictive modeling for CRT, but few have addressed the real-world clinical sequence and cost-efficiency trade-offs in diagnostic testing.

[13] Howell *et al.*, “Using Machine-Learning for Prediction of the Response to Cardiac Resynchronization Therapy: the SMART-AV Study”. JACC Clin Electrophysiol, 7: 1505-1515, 2021.

[14] Manohar *et al.*, “Prediction of Cardiac Resynchronization Therapy Response Using a Lead Placement Score Derived From 4-Dimensional Computed Tomography”. Circ Cardiovasc Imaging, DOI: 10.1161/circimaging.122.014165.

[15] Chang *et al.*, “Deep Learning Significantly Boosts CRT Response Prediction Using Synthetic Longitudinal Strain Data: Training on Synthetic Data and Testing on Real Patients”. Biomed. J., DOI: 10.1016/j.bj.2024.100803.

[16] Larsen *et al.*, “A New Method Using Deep Learning to Preidct the Response to Cardiac Resynchronization Therapy”. J Imaging Inform Med, DOI: 10.1007/s10278-024-01380-8.

[17] Nejadeh *et al.*, “Predicting the response to cardiac resynchronization therapy (CRT) using the deep learning approach”. Biocybern. Biomed Eng., 41: 758-778, 2021.

[18] Khamzin *et al.*, “Machine Learning Prediction of Cardiac Resynchronization Therapy Response From Combination of Clinical and Model-Driven Data”. Front Physio., DOI: 10.3389/fphys.2021.753282.

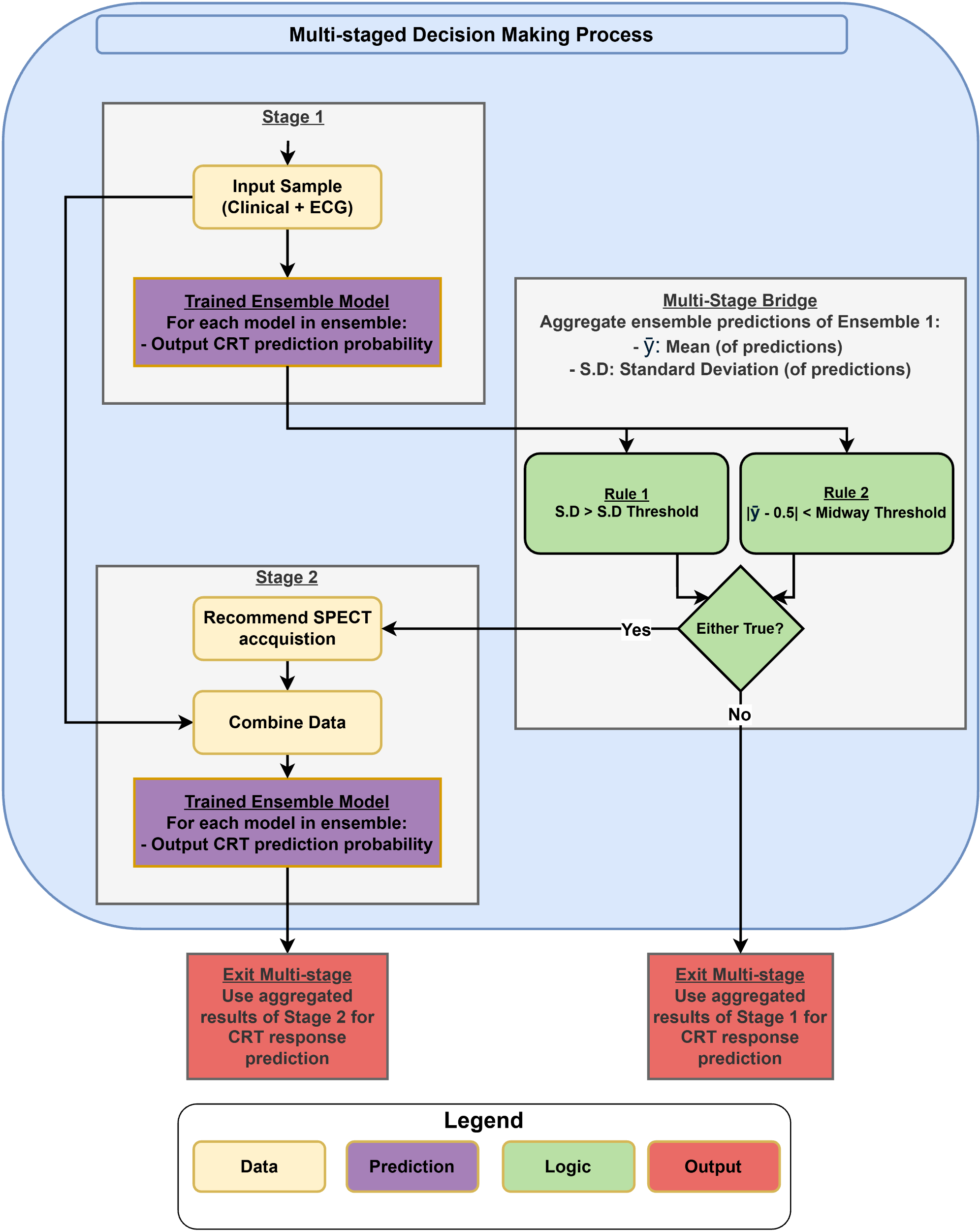
[19] Haque *et al.*, “Interpretable machine learning predicts cardiac resynchronization therapy responses from personalized biochemical and biomechanical features”. BMC Med. Inform. Decis. Mak., DOI: 10.1186/s12911-022-02015-0.

2.      Lack of code availability or pseudo-code blocks reduces transparency for reproducibility.

The code is available from within the data availability statement. For your convenience it can be found here: https://github.com/MIILab-MTU/CRT\_MultiStageML\_Uncertainty.

3.      Key thresholds (standard deviation, midway) and their tuning logic, although mentioned, are complex and should be graphically explained (e.g., thresholding flow diagram).

We have updated Figure 1 to include the thresholding flow diagram in the specific area referencing the multi-stage logic. See here for the updated figure:



4.      There is no mention of random seed usage for reproducibility.

A brief description of random seed usage has been added to the methods section. The code is publicly available for a more detailed view of the random seed usage.

Random seeds were fixed during the start of the experiment, moreover, random seeds were set during data splitting and preprocessing along with each model’s initial configuration for reproducibility.

5.      External validation using an independent dataset (e.g., from another institution or imaging vendor).

Unfortunately, we were not able to collect an external validation dataset due to the unavailability of compatible public datasets and data privacy restrictions.

6.      Ablation study isolating uncertainty quantification to quantify its real contribution.

The comparison of the multi-stage model against its component ensemble models highlights the contribution of the uncertainty quantification. The uncertainty quantification is not used as a feature of the ensemble models, but a value to gate which allows the traversal between ensemble models. A section has been added to the discussion section to address this. Moreover, we considered the performance of the ensemble models from the multi-stage model only in the context of patients that are available to the model.

The Ensemble 2 model achieved the highest AUC and sensitivity, while the multi-stage model achieved the highest accuracy and specificity. Considering that the Ensemble 2 model always had the additional gated SPECT MPI modality while the multi-stage model only used this modality on average 52.7% of the time, shows modest potential. The value of uncertainty quantification is highlighted in comparison between the Ensemble 2 and multi-stage model, patients with lower uncertainty during stage 1 can be considered more trustworthy. For the ensemble 1 model when restricting sample with high confidence, i.e. samples not requiring additional data acquisition, we observe AUC 0.76, accuracy 0.77. Sensitivity 0.58, and specificity 0.82. In comparison with the ensemble 1 model applied to all samples, the performance on the subset of data is improved leading to decreased selective risk; however, due to the small sample size of this analysis, it is possible out-of-distribution samples which would be harder to predict may no longer be evaluated.

7.      Comparative modeling including a deep learning architecture for performance benchmarking.

To further benchmark our model, another comparison multi-stage model using a multi-layered perceptron model neural network as its base model is introduced in the methods section and its results presented in the results section. A section has also been added to the discussion for comparison against the elastic-net logistic regression model. Table 2 and Table 3 have been updated to include the performance metrics.

Methods:

Considering the growing attention and demonstrated capabilities of neural networks in clinical prediction tasks, a multilayer perceptron (MLP) model was also incorporated as a base learner model with corresponding Ensemble models and a multi-stage model.

Results:

The multi-stage (MLP) model had AUC 0.70, accuracy 0.64, sensitivity 0.62, and specificity 0.67. Its Ensemble 1 had AUC 0.68, accuracy 0.64, sensitivity 0.68, and specificity 0.60 while Ensemble 2 had AUC 0.74, accuracy 0.65, sensitivity 0.67, and specificity 0.63. This pattern of performance between the Ensemble models and the multi-stage model follows similarly compared to the ENET base models.

Discussion:

Overall, the multi-stage (ENET) model showed improved performance over the MLP version across all metrics; however, this difference in performance is stemming from differences in sample efficiency considering the low sample size of this analysis. Nevertheless, neural network models are more flexible with their input data and could be trained with rich non-structured data, such as SPECT MPI polar maps to enhance performance.

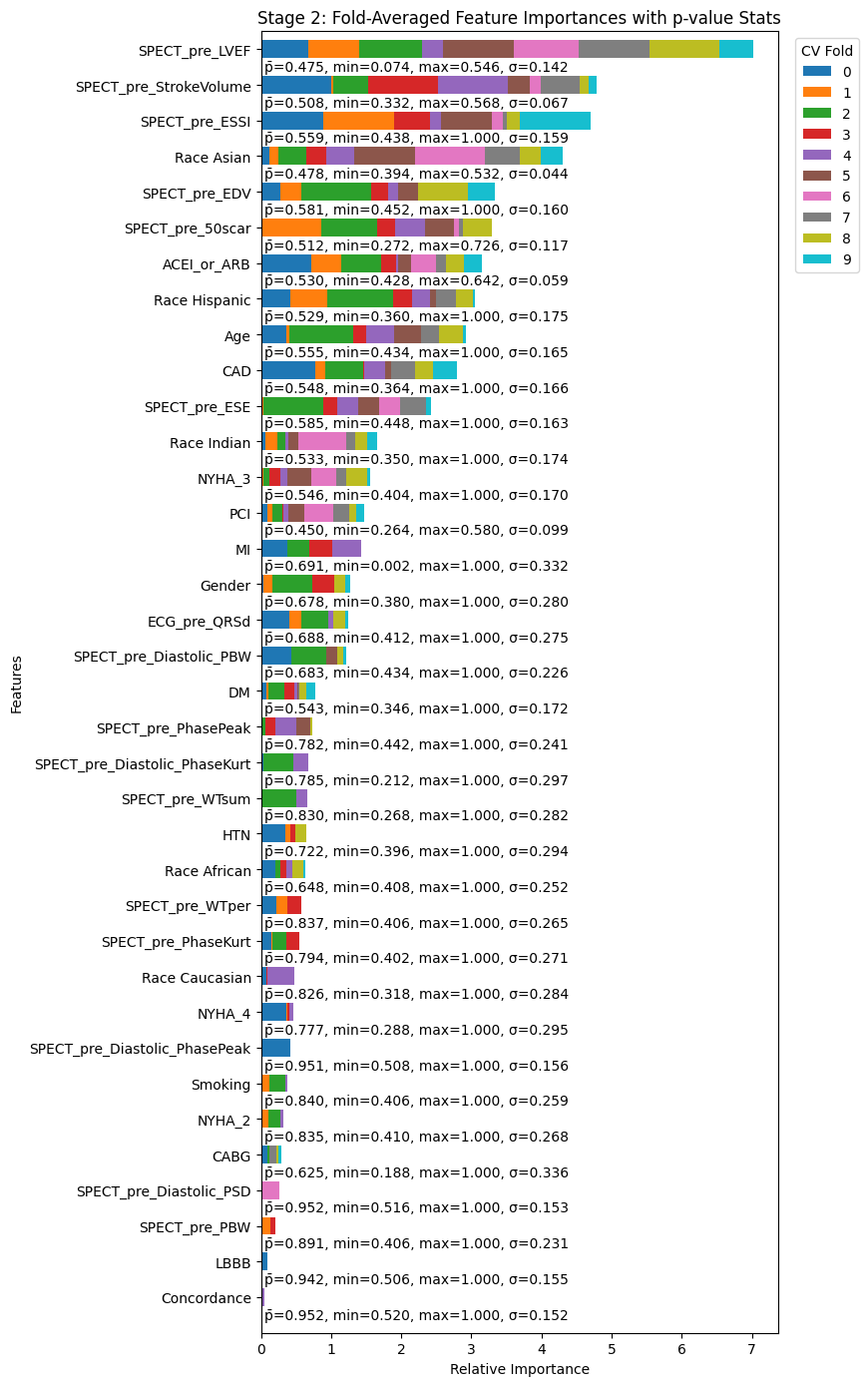
8.      Figures 4-6 are critical and should be more detailed (e.g., indicate statistical significance).

For figure 4, p-values from the DeLong test against the null hypothesis that the distribution of the ranks between CRT responders and non-responders are equal (AUC: 0.50) have been added.

For figure 5, the average p-value from the permutation test across all folds is provided for each feature below its relative importance bar.

A screenshot of a computer

AI-generated content may be incorrect.



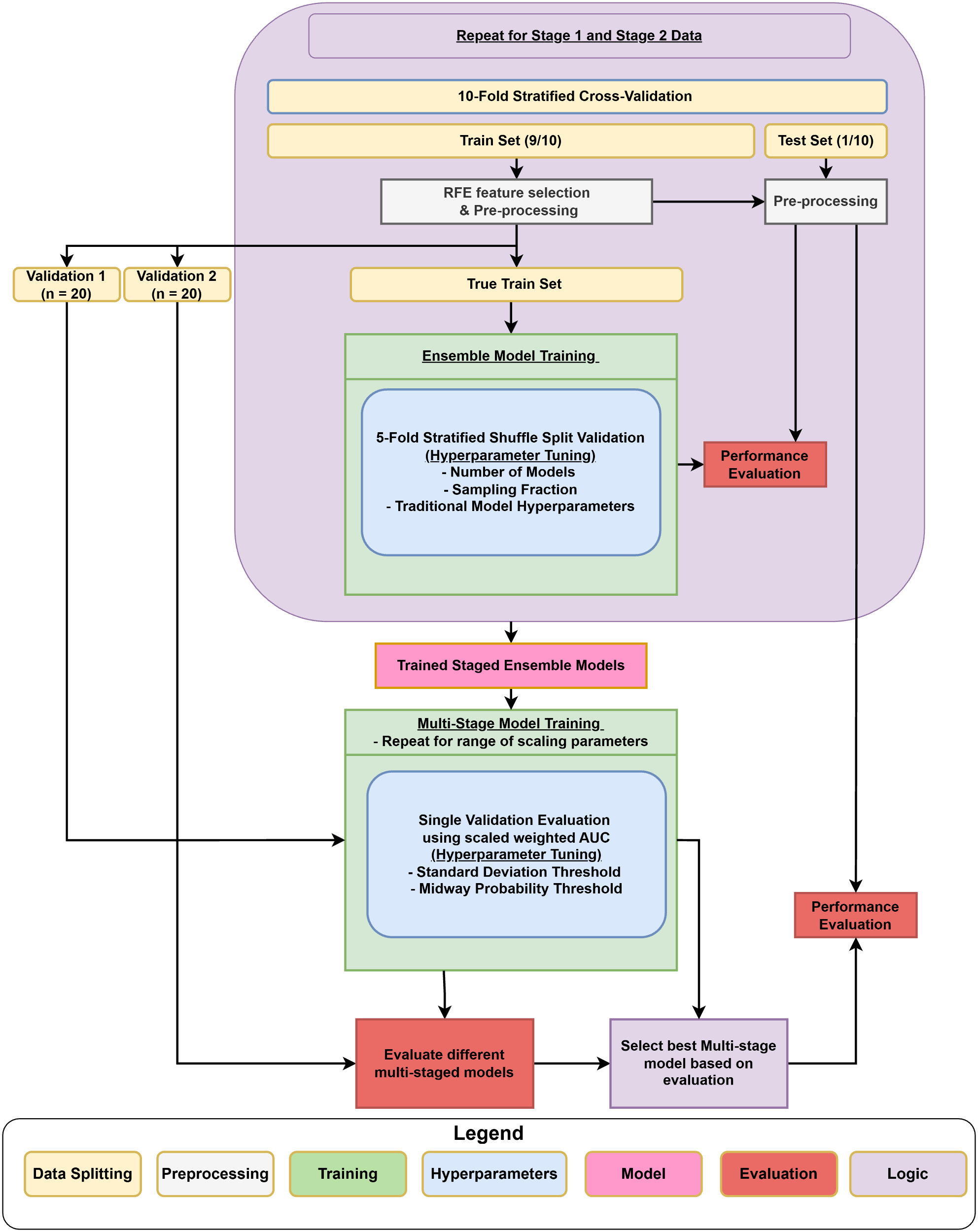
For figure 6, statistical significance was not added to the plot because the purpose of the simulation was to visualize how the training sample size influenced the performance of the model. We do not believe the addition of statistical significance values would meaningfully enhance the figure. We have already provided statistical test results for the case where all available training data is utilized.

9.      Table I includes dense data — consider summarizing or splitting based on clinical vs. imaging variables.

To reduce complexity and ease reading comprehension, Table 1 has been split into clinical and image variable groups. With imaging having two sections ECG and SPECT.

10.     Figure 1 and 2 (workflow diagrams) should include clear legends and improved resolution for clarity.

Legends have been added to the figures. See above for the updated Figure 1 which has a legend. Both figures have been exported as high resolution .SVG instead of .PNG to improve clarity. See below for the update Figure 2 with legend:



11.     There is no heatmap or decision pathway visualization to show model prediction paths.

This is a multivariate model; heatmaps would only be viable for showing bivariate relationships against the outcome. Decision pathway visualizations are for decision tree-based models. And showing a heatmap would not be proper for many samples (e.g., compare ECQ vs. LVEF against response, but not all samples used LVEF) and would have to balance this prediction probably across both models

Thank you for this comment; we have added some clarifying comments in the methods section under the uncertainty quantification section to help elucidate the model pathways. Moreover, Figure 1 shows cases the multi-stage decision making process for a patient prediction. See here for the addition:

The prediction pathway for a sample starts with generating the array of predictions from Ensemble 1 using clinical and ECG data, from here it is possible that a patient’s CRT response predictions have low uncertainty; if so, then the predictions from Ensemble 1 are taken as final; if not, then additional SPECT MPI data is acquired and joined onto the Stage 1 data generating the Ensemble 2 prediction array which is aggregated to create the final CRT response prediction.

However, we do not believe that it is possible to add either heatmaps or decision pathways to show the model prediction paths in a way that would be helpful to the reader.

Heatmaps would only be able to show bivariate (2D) or tri-variate (3D) relationships between explanatory variables against the ultimate CRT response prediction. Since the model is multivariate in nature, it would require the generation of hundreds of these plots. Moreover, the ultimate CRT response prediction depends on whether the patient needed the additional SPECT MPI data which would change which model the ultimate prediction came from. Meaning that the CRT response value is dependent on the specific model which is dependent on that stage’s available data.

For decision pathway visualizations which stem from decision tree models, since the base model is not built from decision trees, moreover, since the ensemble models are a collection of base models, it would not be possible to generate an easily interpretability representation of the model prediction paths.

12.     Delong and McNemar's tests should include confidence intervals.

In table 2, we have included confidence intervals for both the AUC from DeLong’s test and Sensitivity/Specificity from McNemar/s test.

13.     No multivariate analysis to determine independent predictors among the top-ranked features.

Table V has been added to account for multivariate analysis using linear regression of the top-ranked selected features used in both Ensemble 1 and 2 models. See below for the addition to the methods section addition:

Additionally, the features used for the Ensemble 1 model are considered for a multivariate analysis using logistic regression to determine statistical significance of features; the analysis is similarly repeated for Ensemble 2.

Result section addition:

The results of the multivariate analysis on the Ensemble 1 and Ensemble 2 features are shown in Table V. In comparison with the univariate significance presented in Table 1, Age, Race, and CAD remain significant among both Stage 1 and Stage 2 models. All features significant in Stage 1 remained significant in Stage 2 with the addition of SPECT features; however, DM, CABG, PCI, and ECG QRSd become significant only in Stage 2. SPECT LVEF and scar percentage are significant in Stage 2 which follows given their clinical importance. It is noteworthy that from a univariate perspective ECG QRSd is not significant (p = 0.718), when considered in a multivariate fashion amongst Stage 1 features it remains not significant (p = 0.156), until Stage 2 with the addition of SPECT features (p = 0.045).

14.     Deeper insight into why QRS duration and LBBB underperform.

We have expanded the discussion section to address this concern.

While LBBB morphology has been traditionally associated with improved CRT outcomes, recent studies indicate that not all patients with LBBB respond favorably to CRT [1]-[3]. Sassone et al. noted in their study among 243 patients undergoing CRT found that although LBBB was present in 70%, only 60% of LBBB patients experienced CRT response [4]. A “U-shaped” distribution of QRS duration was observed, with non-responders clustered between 120 and 130 ms and above 180 ms. This suggests that both excessively short and long QRS durations may be less predictive of CRT response in LBBB patients.

Furthermore, the VISION-CRT trial, which exclusively included patients with LBBB, revealed that despite the presence of LBBB, a significant proportion of patients did not exhibit the expected clinical benefits from CRT. This underscores the complexity of CRT response and the limitations of relying solely on QRS morphology as a selection criterion in favor of multivariate considerations for both mechanical and electrical dyssynchrony measures.

[1] Linde *et al.*, “Long-term benefits of biventricular pacing in congestive heart failure: results from the Multisite Stimulation in cardiomyopathy (MUSTIC) study”. JACC, 40: 111-118, 2002.

[2] Moss *et al.*, “Cardiac-Resynchronization Therapy for the Prevention of Heart-Failure Events”. NEJM, 361: 1329-1338, 2009.

[3] Gold *et al.*, “Effect of QRS Duration and Morphology on Cardiac Resynchronization Therapy Outcomes in Mild Heart Failure: Results from the Resynchronization Reverses Remodeling in Systolic Left Ventricular Dysfunction (REVERSE) study”. Circ, 126: 822-829, 2012.

[4] Sassone *et al.*, “Relation of QRS duration to response to cardiac resynchronization therapy”. Am J Cardiol, 115: 214-219, 2014.

15.     No exploration of alternative biomarkers or imaging features not captured.

Previously we mentioned in the limitations and future work section:  
In addition to clinical and ECG parameters, it is common to use features derived from echocardiography, specifically, speckle-tracking variables (STE). For future work, incorporating STE features as a new stage in the CRT decision-making process, as a replacement for gated SPECT MPI, or as an insertion between the two existing stages may yield interesting results. This comparative analysis could not be conducted in this study as STE data was not collected in VISION-CRT.

The section has been changed to:

This analysis considers ECG and SPECT-MPI due to their wide spread usage; however, other alternative imaging features could be considered either in replacement of SPECT-MPI or inserted as intermediate stages, such as from Speckle tracking echocardiography for strain features, PET-MPI for quantitative perfusion features in addition to those provided by SPECT-MPI, or Cardiac Magnetic Resonance Imaging for tissue characterization. Ultimately, the multi-staged modeling framework can be adapted to any sequential set of available imaging modalities.

16.     Inter-software variability (e.g., ECTb vs. QGS) is only briefly noted.

In this study, all SPECT image processing was performed by a single, experienced operator utilizing a consistent software package, Emory Cardiac Toolbox (ECTb4, Atlanta, GA). We added some content about this in the method.

Despite these variations in initial acquisition parameters across the different trials, a crucial step was taken to ensure post-acquisition data consistency: all SPECT image processing for the entire combined dataset was performed centrally and consistently by the same experienced operator. This processing was executed exclusively using Emory Cardiac Toolbox (ECTb4, Atlanta, GA), which was implemented to quantify short-axis and planar projection images for automated measurement of left ventricular (LV) function and LV mechanical dyssynchrony (LVMD). This standardized approach in image reconstruction and quantitative analysis aimed to mitigate potential biases that could arise from using disparate software platforms or varied operator techniques.

17.     Impact of acquisition heterogeneity is underestimated.

Thank you for the insightful comment. We fully acknowledge that variability in image acquisition protocols is a recognized and persistent challenge in multi-center imaging studies, and despite our efforts, our dataset likely contains some degree of this inherent variability.

While the retrospective nature and modest sample size of this analysis precluded formal adjustment for all potential acquisition effects, we implemented crucial steps to minimize post-acquisition variability and enhance consistency. Specifically, all SPECT image processing for this study was performed by a single, experienced operator utilizing a consistent software package, Emory Cardiac Toolbox (ECTb4, Atlanta, GA). We emphasize that ECTb4's phase analysis, a key component of our data, is well-established for its robust performance, demonstrating strong temporal resolution[44], high reproducibility and repeatability[45,46], and reliability even with lower count data [47]. This standardization in processing helps mitigate potential influences stemming from different software or varied operator approaches.

Methods:

Despite these variations in initial acquisition parameters across the different trials, a crucial step was taken to ensure post-acquisition data consistency: all SPECT image processing for the entire combined dataset was performed centrally and consistently by the same experienced operator. This processing was executed exclusively using Emory Cardiac Toolbox (ECTb4, Atlanta, GA), which was implemented to quantify short-axis and planar projection images for automated measurement of left ventricular (LV) function and LV mechanical dyssynchrony (LVMD). This standardized approach in image reconstruction and quantitative analysis aimed to mitigate potential biases that could arise from using disparate software platforms or varied operator techniques.

Discussion:

Due to the retrospective nature of the dataset, we were unable to formally adjust for these differences. While we attempted to minimize variation through standardized preprocessing and internal validation, some variability in our results may reflect technical rather than physiological differences. Our study acknowledges limitations concerning data heterogeneity. As detailed in the Methods, the contributing clinical trials (VISION-CRT, Taiwan-CRT, and GUIDE-CRT) employed different SPECT acquisition protocols, which could influence cardiac function metrics like phase SD and bandwidth. While formal adjustment for these retrospective differences was not feasible, all SPECT image processing was consistently performed by a single operator using Emory Cardiac Toolbox (ECTb4). ECTb4's phase analysis, a key data component, is well-established for its robust performance, demonstrating strong temporal resolution [44], high reproducibility and repeatability [45,46], and reliability even with lower count data [47]. This processing standardization helps mitigate potential influences from varying software or operator approaches. For future work, encoding the different acquisition protocol as a predictor itself may prove useful in modeling, specifically with differences in SPECT MPI derived parameters.

[44]. Chen J, Faber TL, Cooke CD, Garcia EV. Temporal resolution of multi-harmonic phase analysis of ECGgated myocardial perfusion SPECT studies. J Nucl Cardiol 2008;15:383-91. [PMID: 18513645] [PMCID: PMC2992837]

[45]. Trimble MA, Velazquez EJ, Adams GL, Honeycutt EF, Pagnanelli RA, Barnhart HX, Chen J, Iskandrian AE, Garcia EV, Borges-Neto S. Repeatability and reproducibility of phase analysis of gated SPECT myocardial perfusion imaging used to quantify cardiac dyssynchrony. Nucl Med Commun 2008;29:374-81. [PMID: 18317303] [PMCID: PMC3048057]

[46]. Lin X, Xu H, Zhao X, Folks RD, Faber TL, Garcia EV, Chen J. Repeatability of left ventricular dyssynchrony and function parameters in serial gated myocardial perfusion SPECT studies. J Nucl Cardiol 2010;17:811-6. [PMID: 20440590] [PMCID: PMC2992839]

[47]. Cheung A, Zhou Y, Faber TL, Garcia EV, Zhu L, Chen J. The performance of phase analysis of gated SPECT myocardial perfusion imaging in the presence of perfusion defects: A simulation study. J Nucl Cardiol 2012;19:500-6 [PMID: 22203443] [PMCID: PMC3731539]

18.     Sample size and ethnic distribution may limit generalizability.

We agree that the small sample size (n=218), along with limited diversity in the ethnic composition of the cohort, may constrain the generalizability of our findings to broader populations. While our primary goal was to explore associations within the available dataset, we recognize that these demographic limitations may impact the external validity of the results. We have revised the manuscript to explicitly acknowledge this and have expanded the limitations section accordingly.

The small number of patients (n=218) enrolled in this study increases the potential for variability in the analysis and may limit the statistical power to detect certain associations. Additionally, the limited representation of ethnic groups in the dataset may affect the generalizability of our findings to more diverse populations. Future studies with larger and more ethnically diverse samples are needed to validate and extend these observations.

19.     Clearly restate the novelty: sequential uncertainty-driven modeling

The discussion has been updated to restate the novelty. See below:

The novelty of this method lies in its multi-stage clinical decision support framework which adaptively incorporates additional medical test data as necessary based on predictive uncertainty, thus enabling refinement of personalized decisions; augmenting convention ML models to consider sequential aspects of increasing data dimensionality.

20.     Summarize three core contributions (performance, efficiency, resource optimization)

We have updated the discussion section to include three core contributions. See below:

Building off previous applications of ML to aid in CRT decision making, we develop a methodology to predict CRT response by sequential escalating data acquisition depending on uncertainty: starting with basic clinical and ECG data, then into more costly SPECT MPI. Moreover, the multi-stage model achieves comparable predictive performance against the full data model demonstrating a reduction in data acquisition burden. The methodology is flexible for a variety of multi-stage clinical workflows involving the successive collection of medical tests.

21.     Suggest next steps like multimodal fusion or transfer learning to broaden applicability

We have added a section to the discussion section for broaden applicability. See below for the addition:

Algorithmic improvements to the methodology include utilizing reinforcement [1], a paradigm of machine learning which works with sequential staged data that does not need explicit labels, in doing so the loosely defined definition of CRT response can be subverted. From a clinical perspective, utilizing high dimensional data such as the gated SPECT MPI 4D scans themselves in stage 2, along with the inclusion of further stages, such as genetic testing could improve CRT response [2].

[1] Liu *et al.*, “Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review”. JMIR, 22. 2020.

[2] Ferro *et al.*, “Impact of DCM-Causing Genetic Background on Long-Term Response to Cardiac Resynchronization Therapy”. JACC: Clinical Electrophysiology, 10: 1455-1464, 2024.

22.     a stronger contrast with previous single-stage or deep models is needed.

We have added a section to the discussion section for comparison of the multi-staged model to previous single-staged models including deep learning models.

In this study, we developed a multi-staged based ML ensemble model which utilized uncertainty quantification and internal logic to predict CRT response using different stages of data. This work demonstrates the potential for a more patient-centric and cost-effective approach to CRT evaluation. By prioritizing the use of less invasive data initially and only proceeding to more expensive modalities when necessary, this approach can improve patient care by reducing the burden of testing and potentially optimizing the selection of patients who are most likely to benefit from CRT, ultimately leading to more efficient resource allocation within the healthcare system.

The data for this study was constructed from 218 patients from the VISION-CRT, Taiwan-CRT, and GUIDE-CRT clinical trials. The multi-stage model’s performance, which includes the internal logic rules for cascading patients to the next stage was compared with the performance of two composite ensemble models. Ensemble 1 contained only baseline clinical and ECG multi-stage model always utilized Ensemble 1 and only used variables, and Ensemble 2 contained the first stage’s data in Ensemble 2 if a given patient’s CRT response prediction’s addition to the features derived from gated SPECT MPI. The uncertainty was below the uncertainty threshold or midway threshold. The Ensemble 2 model achieved the highest AUC and sensitivity, while the multi-stage model achieved the highest accuracy and specificity. Considering that the Ensemble 2 model always had the additional gated SPECT MPI modality while the multi-stage model only used this modality on average 52.7% of the time, shows modest potential. Although, these models still show moderate accuracy, utilizing higher optimized and non-linear models has the potential for improved performance.

In comparison, Tokodi et al. developed the SEMMELWEIS-CRT score, a single staged ML risk stratification system for predicting 1- to 5-year all-cause mortality in CRT patients, which achieved AUCs ranged from 0.768 to 0.803 across time points [9]. While useful for long-term mortality predictions, it does not address short-term CRT response decision making, which is the focus of our study. Feeny et al. employed a single staged naïve Bayes classifier ML model to predict CRT response [10]; like the Ensemble 1 model, clinical variables and ECG parameters were considered for modeling, but this does not effectively leverage potentially acquired myocardial perfusion data, such as from SPECT MPI. Though, similar AUC of 0.70 is observed in both the Ensemble 1 model and the naïve Bayes model. DL models have the capability to utilize higher dimensional data natively, such as directly from imaging modalities. Larsen et al. constructs a DL model drawing from clinical, ECG, SPECT MPI derived parameters, and SPECT MPI polar map images to predict CRT response, which result in AUC of 0.83 higher than the performance of the Ensemble 1 or multi-staged model [16]. However, it should be considered that the multi-staged modeling framework is agnostic to the choice of base model—therefore, it is possible to construct the Ensemble 2 model to include SPECT MPI polar map images to improve performance.

**Other:**

We have detected an issue in how the p-values from Delong AUC test and McNemars Sensitivity/Specificity tests were performed resulting in slightly inaccurate p-values. The issue has been corrected in Table Ⅲ. Apologies for any inconvenience.