Dear Reviewers,

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

We have revised the manuscript to address each point raised. Below, we provide a detailed response to your comments.

In this document:

* Reviewers’ comments are in black.
* Our responses are in blue.
* The text from the revised manuscript is in red.
* The text from the unchanged original manuscript is in orange.

**Editor:**
1) Introduction needs a better review of the current relevant literature on this topic - core methods used, key results, gaps/limitations that led to the rationale for your approach.

Thank you for this suggestion. The introduction section has been expanded to focus on the how the methods used by the relevant literature led to interesting results which created the rationale for our approach.

Feeny et al.7 applied ML using 9 clinical features, such as QRS morphology, QRS duration, NYHA classification, LVEF to predict CRT response at a 6-month follow up which improves CRT response prediction beyond current guidelines. Apostolopoulos et al.8 has utilized a supervised DL method to predict coronary artery disease (CAD) using a VGG16 transfer learning (TL) model solely using SPECT MPI polar maps which is competitive with experts’ diagnostic ability. He et al.9 applied the autoencoder (AE) technique, an unsupervised deep learning method, to the polar maps of left-ventricular mechanical dyssynchrony to extract new predictors showcasing that currently extracted SPECT MPI parameters do not fully utilize all the semantic information from polar maps. To our knowledge, CRT response has only been modeled using simple clinical features while the task of predicting CAD has utilized images directly, such as by Apostolopoulos et al. 8. Therefore, using DL to process the high dimensional data, we hypothesize that combining clinical features and medical imagery can improve CRT response prediction.

2) Verify references conform to journal format requirements.

We have reviewed and corrected the references to ensure they adhere to the journal’s format requirements. Specifically:

* References [29]: The page numbers were incorrect and have been updated.
* Reference [1]: The journal name was previously listed incorrectly as *Am Heart Assoc* and has been corrected to *Circulation*.
* For the references “Not Checked”:
	+ [22] and [18]: These are conference papers, which are common in the computer science field
	+ [20]: This refers to a widely referenced preprint paper.
	+ [19]: This is a book chapter.

Below, we provide the corrected references with their updated details:

[[27] Gumuser G, Parlak Y, Topal G, Batok D, Ruksen E, Bilgin E. Comparison between ECTb and QGS for assessment of left ventricular function. *J Nucl Med* 2007; **48**; 408.](https://jnm.snmjournals.org/content/48/supplement_2/408P.4)

[[29] Okuda K, Nakajima K, Matsuo S, *et al.* Comparison of diagnostic performance of four software packages for phase dyssynchrony analysis in gated myocardial perfusion SPECT. *Eur J Nucl Med Mol Imaging* 2017; **7**: 27.](https://pmc.ncbi.nlm.nih.gov/articles/PMC5364119/)

[[1] Virani SS, Alonso A, Benjamin EJ, *et al.* Heart disease and stroke statistics—2020 update: A report from the American Heart Association. *Circulation* 2020; **141**: 139–596.](https://pubmed.ncbi.nlm.nih.gov/31992061/)

**Reviewer #1**

1) The authors applied VGG16 for image processing and classification. Following the literature review, there are DL methods more efficient in some cases from VGG16, such ar ResNet, AlexNet, MobileNet.
It would be important to the researchers to know the performance of the other well known DL methods for image classification of polar maps. A comparative analysis with state of the art DL methods could be useful to show the advantageous characteristics of DL methods.

We appreciate the reviewer’s insightful suggestion to compare the performance of our proposed VGG16-based model with other well-known deep learning (DL) architectures, such as ResNet, AlexNet, and MobileNet. In response, we conducted a comparative analysis with ResNet and MobileNet, but not AlexNet because it is not in common usage as it is one of the more historical DL image models. For ResNet and MobileNet, we only consider the cases of utilizing both the tabular and image modalities of data. We describe the models in the methods section.

To compare the TL image model, we also consider ResNet50 and MobileNet2, two widely used ImageNet models.

Below are summarizations from the results section.

The multi-input DL model utilizing the ResNet50 image model architecture had AUC 0.84, accuracy 0.71, sensitivity 0.71, and specificity 0.71, while the MobilNeet2 mage model architecture version had AUC 0.75, accuracy 0.68, sensitivity 0.80, and specificity 0.53.

And the discussion section.

The multi-input DL (MobileNet2) model showed the lowest performances when compared to the multi-input VGG16 and ResNet50 architectures which follows since MobileNet2 was designed to be light

weight for mobile devices. In comparison of the multi-input VGG16 to ResNet50, slightly lower AUC was noted (0.83 vs 0.84) similarly for specificity (0.69 vs. 0.71), however, accuracy was higher (0.73 vs. 0.71) and similarly for sensitivity (0.76 vs. 0.71). Notably, the ResNet50 architecture had substantially high standard deviation in specificity. Overall, we reaffirm that the VGG16 architecture performs better in an aggregate view across the metrics, though minorly. We expect that better performance can be achieved via fine-tuning of the image models to the polar map data given enough images.

2) Also, there is no performance analysis on how DL methods work solely on images and in which degree the results are improved including clinical, or vice versa. Further experiments are needed and to be included in the discussion of results.

We appreciate the reviewer’s comment on the need for a performance analysis of deep learning (DL) methods solely on images and the inclusion of such results in the discussion. A DL model which works solely only the images (polar maps) was already considered in the paper (“DL only Polar map”), its performance was noted in the results section, but further discussion was missing from the discussion section. It has now been added:

When compared with the single-input DL models the multi-input DL model trended towards improved performance across all metrics. Additionally, the multi-input DL model had a significant difference in AUC above the polar map only DL model. When comparing the single-input DL models, the tabular DL model show higher AUC and specificity, but lower accuracy and sensitivity. It is interesting that the tabular features were able to detect CRT responders more often, while the images themselves lead to detecting CRT non-responders more often. However, it should be noted that the tabular DL model had roughly balanced sensitivity and specificity in comparison to the polar map DL model.

3) Concerning the implementation of the GradCam, it is observed that it is applied in gray scale images of polar maps. As the polar maps produced by SPECT-MPI are colourful, which are the results of the GradCam application?

We appreciate the reviewer’s comment regarding the use of grayscale polar maps for Grad-CAM analysis, given that SPECT-MPI polar maps are commonly visualized in color. It is important to note that the polar maps generated by SPECT-MPI are naturally in grayscale, as they represent numerical data values corresponding to myocardial perfusion. The colorization typically seen in clinical settings is applied by specific software tools to enhance visual interpretation for clinicians.

To clarify, in our study, we used the grayscale polar maps directly as input to the Grad-CAM analysis. The case demos Figure 5 and 6, show both the colored slices (Top) and polar maps (Bottom Left) as output from the Emory Reconstruction Toolbox, the grayscale polar maps (Middle Right), and Grad-CAM overlayed heatmap of the polar maps (Bottom Right). To minimize confusion, we have added labels to Figures 5 and 6, explicitly indicating the grayscale polar maps and the Grad-CAM overlay outputs. This ensures clearer differentiation between the different representations.

We hope this additional clarification addresses the reviewer’s concern.

4) Open issues, limitations of the existing approaches or prevalent trend /issues of utilizing SPECT MPI polar maps, incorporating additional patient data directly in the form of medical imagery should be elucidated.

We appreciate the reviewer’s suggestion to elaborate on these topics. A section has been added to the discussion section which dives into issues of incorporating additional patient data in a multi-modal way.

Incorporating the additional SPECT MPI polar images has been seen to improve predictive capabilities over usage of tabular data only. However, this approach introduces challenges. A key limitation is the inability to determine the relative contributions of the two data sources to the final ultimate prediction, e.g., the polar map images played a larger role in determining CRT response for this patient. Additionally, from Kline et al., we see that in general across health fields, utilizing multiple modalities of data sources on average increases predictive performance. However, this comes with increased complexity in model design and implementation. Furthermore, models trained on SPECT MPI polar maps require a substantial number of samples due to the high dimensionality and inherent noise of the images, as noted by Apostolopoulos et al.8 While our study incorporates a reasonably large dataset, the sample size may still not be fully adequate to develop a classifier with optimal performance. These limitations underscore the importance of continued work to refine and validate multi-modal approaches and to develop methods that can disentangle the contributions of different data sources in clinical predictions.

5) Conclusion section need to be rewritten as there is no explanation on the main advantages of the proposed methods in this domain, and the main limitations, or other merits of them. There is no sufficient explanation of the comparisons tasks and future directions are completely not clear and useful for the readers and future researchers in this domain.

Thank you for pointing this out. The conclusion section has been rewritten with a focus on specifying the main advantages, limitations, comparisons to the other methods in the study, and the future direction.

The proposed multi-input DL model shows values over the use of ML models utilizing tabular medical data and current guidelines by improving the performance towards predicting CRT by including polar map images. Polar maps are easily generated from SPECT MPI scans at no extra cost but are not directly used towards CRT response prediction currently. The main limitation, despite the explainability of Grad-CAM, is the lack of interpretability from the model. Incorporating additional volumetric imaging data, such as sequential frames from gated SPECT MPI could lead towards improved CRT response prediction through DL methods.

**Reviewer #2**

1) Page 4 Clinical Data - I would suggest including a table in the supplementary section that lists the names of the 44 features used in detail.

Thank you for the suggestion. While Table 1 provides and overview of the used in the model after feature selection, we have added a supplementary table showing all the features that were available before feature selection.

2) The ENET model was not described in the Methods section, but it is mentioned in the Results. Please include the description of the ENET model in the Methods section.
We appreciate the comment. The ENET model is a specific augmentation of the logistic regression model that constrains the model through addition regularization. In the methods section, we have changed the mention of logistic regression to the elastic net logistic regression.

The four models include Elastic Net Logistic Regression (ENET), Random Forest (RF), Adaboost Decision Tree (ADA), and Support Vector Machine (SVM).

3) The title of Figure 4 does not indicate that it refers to the DL model.

Thank you for bringing this to our attention. The title has been renamed.

Figure 4. Sample size simulation on test set using randomized subsampling on training set for the multi-input DL (VGG16) model.

4)  Page 4, Lines 21-23: "Further exclusion of patients with significant missing data resulted in a final analytical cohort of 218 patients." What exactly is meant by "significant missing data"? Which parameters were missing, and how many were lacking?

Thank you for pointing this out. A further breakdown of missing data and relevant exclusions have been added.

Further exclusion of patients with significant missing data (ECG QRSd, percutaneous coronary intervention (PCI), myocardial infarction (MI), hypertension (HTN), ACEI/ARB medications, or SPECT Concordance) resulted in a final analytical cohort of 218 patients.

5) Did the differing SPECT protocols employed in the various cohorts (GUIDE-CRT and VISION-CRT) influence the observed results? If there is potential bias, I suggest including this in the limitations section.

Thank you for raising this important point. We acknowledge that the differing SPECT acquisition protocols between GUIDE-CRT and VISION-CRT could introduce bias into the results. While such systematic differences in SPECT MPI acquisition protocols are well-documented and not unique to this study, they may have influenced the derived SPECT parameters. In addition, variations in software used to generate SPECT MPI parameters could further contribute to potential biases.

To address this, we have added the following to the limitations section:

Moreover, differences in SPECT MPI acquisition between VISION-CRT and GUIDE-CRT have the potential to bias generated parameters in a systematic way between the different trials.

6) What is the rationale for comparing the obtained model solely with ML models that use only tabular data and do not incorporate SPECT MPI data? Although the multi-input DL model was not significantly better than the ML models, why did you not combine the results obtained from the VGG16 module with those from the LR, RF, ADA, and SVM classifiers, and compare these combinations with the VGG16+MLP model?

Thank you for the insightful comment. Simple ML models have been extensively used not only in the context of CRT, but in many other situations1. The ML models do utilize data from SPECT MPI, but only the tabular version of the SPECT MPI parameters, not the imaging component of SPECT MPI from where these tabular parameters are extracted from. Utilizing this method, we can compare simple classifiers which use the tabular SPECCT MPI data among other clinical variables (ML models) against more powerful classifiers which either also use the same tabular data (DL only tabular), do not use the tabular data but only the imaging data (DL only polar map), and finally the most comprehensive model which combines both the tabular and imaging components.

To reduce confusion, some more description is added to the methods section to indicate that SPECT MPI parameters are included in the tabular data for the ML models.

For performance comparison, four ML models were considered solely using the tabular data which includes clinical variables and derived parameters from SPECT MPI.

[1] K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918

7) External validation is particularly important. I would recommend externally validating your model as well. Do the authors have the capability to do this?

Thank you for your valuable suggestion. Both VISION-CRT and GUIDE-CRT are prospective clinical trials and they have been completed. Unfortunately, we do not have access to an external dataset at this time. Therefore, we were unable to conduct an externally validation to our models. However, we have performed a validation using holdout test set from our existing dataset to ensure robust evaluation of the model's performance. Future work would involve seeking opportunities for external validation to further assess the generalizability of our findings.

8) In Tables 1 and 2, it is not reflected whether the patients were on Optimal Medical Treatment (OMT); I only see ACE-I/ARB. Please include additional medications for the patients in Table 2. What percentage were on OMT? Are LBBB/non-LBBB data available?

Thank you for pointing this out. We have now included additional medication details in Table 2, specifically Beta Blockers (BB) and Mineralocorticoids (MRA), as well as information regarding whether patients were on Optimal Medical Treatment (OMT). Out of the total cohort, only 14 patients (6.4%) were on OMT. Regarding the LBBB data, although we had information on the patients' LBBB status, the majority of the patients had LBBB, which did not serve as a significant predictor in our dataset. Nevertheless, we have added the LBBB data to Table 2 for transparency.

9)  "A simulation validating the sample size is shown in Figure 4. The test set performance against the sample size shows that around 150 samples, the DL model has reached a plateau of performance indicating the robustness of the model at the current training sample size." Yes, but the confidence interval (CI) still visibly narrows, particularly in specificity and sensitivity, so it would be advisable to further expand the training data if possible.

Thank you for your suggestion. Unfortunately, expanding the training data is not feasible, as increasing the size of training set would limit the amount of data available for the holdout test set, which is critical for evaluating the model’s performance.

The narrowing of the CI reinforces the conclusion in fact since low variance in the performance (specificity and sensitivity) in the model is a desirable trait. We desire lower variance in the model performance because it indicates the predictions of the model tend to be more accurate more often. Ideally, we want the point estimate of the confidence interval (solid line) to increase until a plateau while at the same time the CI narrows indicating a higher certainty in the performance of our model which is what we are seeing. Comments regarding the narrowing of the CI have been added to the discussion section.

Moreover, the confidence interval (CI) for sensitivity and specificity are narrowing further indicating a higher certainty in the performance of the multi-input DL model as the sample size increases, especially past 175 samples.

**Grammatical and Formatting Errors:**

Page 4 Line 56: "complementary8 or 16 frames" is missing a space.

Thank you. This has been corrected.

Page 6 Line 51: Typo "variables fos the models" (should be "for the models").

Thank you. This has been corrected.

"VGG16 model": The model name is not consistently formatted.

VGG16 model is now the common name.

"Grad-CAM": In some places, the formatting is not consistent, for example, the Grad-CAM abbreviation is not used uniformly in terms of writing style.
Grad-CAM is now the common name.

**Additional Edits:**

1. Instances of the incorrect spelling “polarmap” were replaced by “polar map” in the manuscript.
2. In table 2 of the baseline metric, the percentage specify response and non-response were incorrect and have been corrected.