Beyond Predictive Accuracy: A Multi-Criteria Adaptive Framework for Reliability-Aware Machine Learning Model Selection in Cardiovascular Disease Prediction

Divya Priya R[1], Bagyamani J[2]

[1]Research Scholar, Department of Computer Science, Government Arts College, Dharmapuri

[2]Principal, Government Arts and Science College, Hosur

DOI: https://doi.org/10.51244/IJRSI.2026.1308000082

Received: 21 August 2026; Accepted: 26 August 2026; Published: 04 September 2026

ABSTRACT

Machine learning (ML) models are increasingly being used for cardiovascular disease prediction with model selection often guided by predictive performance measures such as accuracy. However, predictive accuracy per se may not adequately reflect the prediction reliability relevant to clinical decision support, especially when competing models differ in calibration, predictive uncertainty, confidence and decision reliability. The study proposes **MASE-RAMR**, a **Multi-Criteria Adaptive Selection Engine with Reliability-Aware Model Ranking**, to offer a wider basis for evaluation and ranking of cardiovascular disease prediction models. The proposed framework clusters heterogeneous evaluation metrics into four reliability-oriented meaningful dimensions: **Accuracy, Calibration, Uncertainty, and Decision Reliability**. Then, a hierarchical CRITIC-based weighting strategy is adopted, first the importance of the specific metrics is determined in each dimension and then the global importance of the resulting dimensions is estimated. The obtained adaptive criteria matrix is then evaluated by TOPSIS to compute **Reliability Aware Model Ranking (RAMR)** score for each candidate model. For the evaluation of the framework, six cardiovascular disease datasets were used and compared with the conventional CRITIC-TOPSIS model ranking approach with the same experimental setting. The comparison demonstrates two important behaviours of the proposed framework, **decision stability**, in which the same top-ranked model is retained for a number of datasets, and **adaptive ranking,** in which the proposed framework changes or refines the model rankings when the additional reliability dimensions provide discriminative information. Further analysis of traditional classifiers shows that the framework does not inherently favour complex ensemble or boosting models, as conventional classifiers achieve the highest reliability rankings on several datasets. These results show that MASE-RAMR provides a dataset-dependent and reliability-aware strategy for ML model selection beyond predictive accuracy alone, providing a more comprehensive basis for identifying candidate models for clinical decision support.

Keywords: Cardiovascular disease prediction; Machine learning; Reliability-aware model selection; Predictive uncertainty; Probability calibration; Decision reliability; Multi-criteria decision-making

INTRODUCTION

Cardiovascular disease (CVD) continues to be a major challenge for healthcare systems, and accurate risk prediction can enable earlier identification of individuals at high risk and more informed clinical decision-making. Machine learning (ML) has been increasingly used for CVD prediction due to its ability to capture complex relationships from heterogeneous clinical variables . Therefore, a variety of statistical, tree-based, boosting and ensemble algorithms have been explored for cardiovascular prediction with model selection often based on measures such as accuracy, sensitivity, specificity, F1-score and ROC-AUC.

Although predictive performance is an important aspect of clinical ML evaluation, it does not always reflect prediction reliability well. Models with similar discrimination may differ in terms of the calibration of the predicted probabilities, the uncertainty of the predictions, and the confidence with which the outputs can be interpreted. When predicted probabilities are used for risk assessment, calibration is important, and uncertainty

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

information can offer additional insight into situations where model predictions might be less reliable [1–6]. Thus, choosing a model only based on its predictive accuracy might miss features relevant to trustworthy clinical decision support.

However, a straightforward application of CRITIC–TOPSIS to a flat set of evaluation metrics treats all criteria as independent decision elements, even if many of the criteria are different aspects of model behaviour. For example, Test Accuracy, Cross-Validation Accuracy and Cross-Validation Standard Deviation together describe predictive performance, whereas Log Loss, Brier Score and Expected Calibration Error provide complementary information about calibration. In the same vein, predictive uncertainty and entropy are measures of uncertainty, while confidence and prediction margin are measures of decision reliability. In a flat representation, these higher level concepts are not explicitly identified prior to the final model ranking being produced.

To address this limitation, this study proposes **MASE-RAMR**, a Multi-Criteria Adaptive Selection Engine integrated with a Reliability-Aware Model Ranking framework. MASE organizes heterogeneous model-evaluation criteria into four reliability-oriented meaningful dimensions: **Accuracy, Calibration, Uncertainty, and Decision Reliability**. A hierarchical CRITIC procedure is then applied in two stages. First, local weights are determined for individual criteria within each dimension. Second, the resulting dimension-level information is weighted globally to form an adaptive criteria structure. TOPSIS is subsequently applied to this structure to obtain the **Reliability-Aware Model Ranking (RAMR)** score and final model ranking.

The objective of MASE-RAMR is therefore not to replace predictive accuracy with other criteria, but to place accuracy within a broader reliability-aware model-selection process. The framework allows the relative contribution of the evaluation criteria and dimensions to vary according to the information contained in each dataset. This makes the final model preference dataset-dependent rather than determined in advance by a fixed weighting scheme or by an assumed preference for a particular algorithmic family.

To evaluate the proposed framework, this study compares MASE-RAMR with conventional Flat CRITIC–TOPSIS across six cardiovascular disease datasets using a common pool of candidate machine-learning models. The comparison will be made based on two measures: the stability of the decision (two approaches give the same ranking of the top model), and the ordering of the competing models (the ordering of the models in the adaptive ranking formulation changes). Not only is the selection of the model to be analyzed, but also the difference in the full model ranking in the two decision structures are analyzed.

There are three main contributions of this study. First, it suggests a multi-level, multi-dimensional heterogeneity of model evaluation criteria with four criteria measured by reliability of the predictions: Accuracy, Calibration, Uncertainty and Decision Reliability. Second, it presents an adaptive CRITIC weighting process in two stages taking into account the importance level of each criterion and the importance level of each dimension. Third, it allows a multi-dataset comparison of MASE-RAMR and Flat CRITIC–TOPSIS, considering both the stability of the model selection and the variations in the complete ranking of the candidate models.

Therefore, the main research question is whether predictive performance, in addition to the dimension of calibration and uncertainty, can be integrated into a decision-reliability framework to be more informative when selecting cardiovascular disease prediction models. The purpose is not to claim that the model identified by MASE-RAM necessarily has higher predictive accuracy. In this study, however, the researchers ask whether a multi-tier decision model that evaluates various predictive characteristics as a set could provide a more robust basis for model selection.

Next, in section 2, background context and a synthesis of literature is provided. Section 3 presents the MASE-RAM framework, which outlines how to assess models, how to assign weights to them in multiple tiers and how to prioritize them based on reliability. Section 4 presents the empirical results and summarizes the learning from the six different cardiovascular disease datasets. Lastly, in Section 5, main results are summarized, main results are presented, framework limitations are discussed, and future research directions are discussed.

Related Work and Research Gap

Machine Learning for Cardiovascular Disease Prediction

The ability of machine learning (ML) to model complex relationships among clinical variables and detect trends not easily captured by traditional statistical approaches has led to increasing exploration of its use for cardiovascular disease (CVD) prediction. Recent systematic evidence demonstrates that ML models can achieve comparable discrimination to conventional CVD risk prediction approaches, and in some cases better discrimination. However, the clinical utility of these models is limited by significant heterogeneity between studies and methodological limitations such as inconsistent validation and reporting practices [7,8].

In most ML studies, the performance of candidate models is assessed by predictive-performance measures such as accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). These measures are important for discrimination assessment but do not fully characterise the reliability of a model's predictions. In particular, a model may be well discriminative, but badly calibrated in terms of probabilities. Thus, calibration is an important complementary property when the predictions are to support clinical decisions [18,19].

This distinction is particularly important in CVD prediction. Recent systematic evidence suggests that discrimination is reported more consistently than calibration in cardiovascular risk prediction studies. This may leave model evaluation focused primarily on how well a model distinguishes between outcomes, rather than on whether its predicted probabilities accurately reflect the observed risks [8,18].

Calibration, Uncertainty and Reliability in Clinical ML

Calibration assesses how closely a model's predicted probabilities correspond to the observed frequencies of the outcome. For example, a well-calibrated model should produce predicted risks that are close to the actual event frequency among patients assigned a particular risk level. Calibration is a valuable aspect of understanding model predictions, as it will give information about the accuracy of the predictions, unlike discrimination. In this sense, calibration is important when interpreting model predictions in clinical practice [2,6].

Another important way of looking at the reliability of prediction is uncertainty. Medical ML systems can receive noise measurements, missing data, unusual measurements, or measurements that are outside of the range of measurements used for training. When this is the case, a prediction along with information about the level of confidence that the model has in its prediction can be helpful. Kompa et al. highlighted the need for quantifying and communicating uncertainty in medical ML, which can help in using model predictions in a safer and informed manner [4]. Similarly, Seoni et al. highlighted Page 5 of 31 - AI Writing Submission Submission ID trn:oid:::3117:614343109 uncertainty estimation as an important consideration for the safe application of AI in healthcare [5].

Calibration and uncertainty go together and are not interchangeable. Calibration is an indicator of the reliability of predicted probabilities, while uncertainty gives insights into the level of confidence or uncertainty in specific predictions. More recent efforts around the calibration of probabilities for diagnostic uncertainty have also highlighted the need to consider probability assessments in diagnostic contexts in which there is disagreement and uncertainty in the diagnostic label itself [6].

Based on these observations, it does not seem that predictive accuracy is enough to use in the selection of a model to be used for clinical decision support. Using a model-selection procedure based on predictive performance and other characteristics of reliability such as calibration, uncertainty, and confidence, a more comprehensive framework can be established to compare competing models.

Multi-Criteria Approaches for Machine Learning Model Selection

A multi-criteria decision making (MCDM) problem can be used for the evaluation of various parameters of alternative models. When several models exhibit conflicting behaviors in various aspects of the evaluation, MCDM methods can be used to take several criteria into account. This is important for selecting the most

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

appropriate model for a ML model, for example, ML model A may be more accurate in predicting whereas ML model B may be better at being calibrated or have lower uncertainty.

The Criteria Importance Through Intercriteria Correlation (CRITIC) method is based on an objective way to assign the weights for the criteria, which takes into account the intensity of each criterion's contrast and the correlation between the other criteria [9]. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is used to rank the alternatives based on the closeness of the alternatives to the positive ideal solution and the distance of the alternatives from the negative ideal solution [10]. CRITIC and TOPSIS are used together to offer a structured method of weighing several criteria and rank the various alternatives.

The application of MCDM has also been directly made on the selection of ML models. Ali et al. proposed a multi-criteria approach to suggest the best ML algorithms from a set of performance metrics and illustrated the approach using different datasets [11]. In the same way, Leyva-Vázquez et al. proposed a TOPSIS based method to obtain the best ML model from several evaluation criteria [12]. Combined, these studies demonstrate that model selection may be more than just a single performance metric and that MCDM provides a systematic approach to comparing different algorithms.

But these methods focus mainly on the evaluation measures as individual criteria. If a lot of dissimilar metrics are added in one decision matrix, then all the metrics that represent a similar concept are placed at the same level in the decision matrix. For clinical ML, this may be problematic because measures such as accuracy and cross-validation performance represent predictive performance, whereas Log Loss, Brier Score, and ECE describe calibration, and uncertainty and confidence measures describe other aspects of prediction reliability.

Existing Clinical ML Frameworks and Research Gap

Recent clinical ML studies have also explored ensemble learning and explainability to improve predictive performance and transparency. In [13] Ganie et al. developed a stacking ensemble framework for predicting multiple cancer types and incorporated SHAP analysis to improve the interpretability of the resulting models [13].

The present study is related to this research direction but addresses a **different problem**. The focus is not on developing another stacking architecture or introducing an explainability technique. Instead, the focus is on **how competing ML models should be evaluated and selected when predictive performance and reliability are considered jointly**.

Based on the literature, three specific gaps motivate the present work:

1. **Accuracy-centered selection:** predictive performance remains a dominant basis for comparing ML models, while calibration and uncertainty are not consistently integrated into the final selection decision[1,6–8,18–20] .

2. **Flat multi-criteria evaluation:** existing MCDM-based model-selection approaches demonstrate the value of combining multiple criteria, but related criteria are generally treated as individual measures rather than being organized into higher-level clinical reliability dimensions [11,12].

3. **Limited reliability-aware model selection:** existing clinical ML work has addressed prediction performance, ensemble learning, explainability, calibration, and uncertainty as separate concerns, but there remains an opportunity to integrate these characteristics into a structured model-selection framework.

To address these gaps, this study proposes the **Multi-Criteria Adaptive Selection Engine with Reliability-Aware Model Ranking (MASE-RAMR)**. The framework categorizes model-evaluation measures into the four dimensions of Accuracy, Calibration, Uncertainty and Decision Reliability, and uses hierarchical weighting to create the final model ranking. The proposed approach addresses the question of: “Which model has the best overall prediction profile of all models evaluated?” rather than the more common question: “Which model performs best?”

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

It is not MASE-RAMR's strategy to guarantee the selected model will consistently achieve the best Test Accuracy. Rather, it explores whether there are other dimensions of reliability that offer knowledge beyond predictive accuracy to choose a model for clinical decision support. The next sections explain the new framework proposed and its experimental comparison with the traditional CRITIC–TOPSIS approach.

Proposed MASE-RAMR Framework

Framework Overview

A multi-criteria approach for evaluating and ranking competing machine learning (ML) models is proposed, the Multi-Criteria Adaptive Selection Engine with Reliability-Aware Model Ranking (MASE-RAMR). The framework is not just based on predictive accuracy; it takes into account four evaluation dimensions: Accuracy, Calibration, Uncertainty, and Decision Reliability. The evaluation is based on the following 10 criteria: Test Accuracy, Cross-Validation (CV) Accuracy, CV Standard Deviation, Log Loss, Brier Score, Expected Calibration Error (ECE), Predictive Uncertainty, Entropy, Confidence, and Prediction Margin. These criteria are classified into four dimensions before weightage and ranking are done.

The structure is hierarchical and takes a weighting approach. The CRITIC method is first applied independently in each dimension to identify the relative importance of the individual criteria. CRITIC is an objective weighting method which takes into account criterion variability and inter-criterion correlation [9]. The local weights are then applied to sum the criteria to create the Adaptive Criteria Matrix (ACM). The ACM is then subjected to a second CRITIC stage, to assess the global importance of the four evaluation dimensions. This weighted matrix is then used to analyse using TOPSIS method that ranks the alternatives according to their closeness to the positive ideal solution and distance from the negative ideal solution [10]. The closeness coefficient "TOPSIS" is thus calculated and used as a RAMR score.

The entire MASE-RAMR process is broken down into four steps:

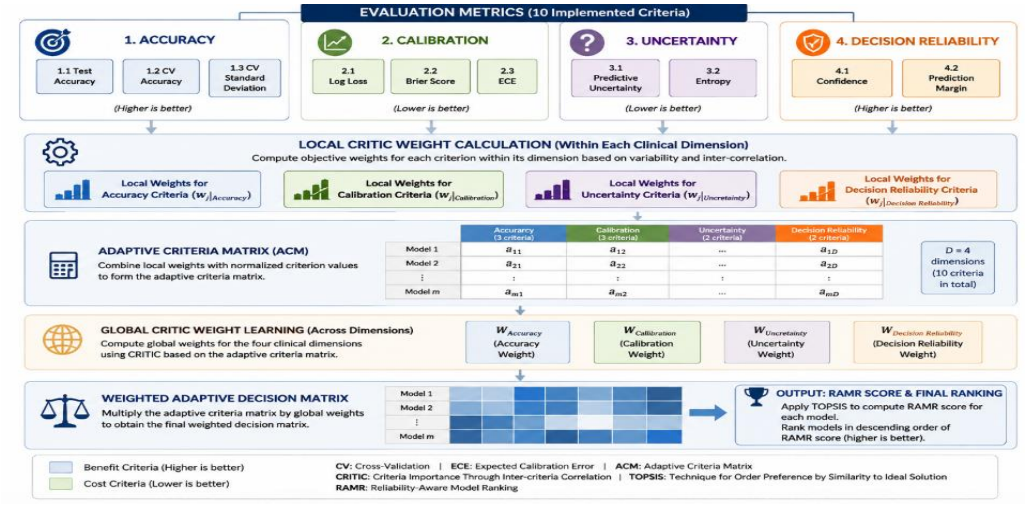
Stage 2 — Metric evaluation: Candidate ML models → Metric evaluation

Stage 2 Adaptive Criteria Formation: MASE dimensions to local CRITIC weighting to Adaptive Criteria Matrix

Stage 3 – Global Weighting: Global CRITIC weighting → Weighted decision matrix

The third stage is Reliability-Aware Ranking (RAMR score) to final model ranking, as indicated by the arrow.

Figure 1. Architecture of the Multi-Criteria Adaptive Selection Engine (MASE).



Evaluation Criteria and Dimension Formation

The candidate models are assessed by metrics that are based on four complementary dimensions. The Accuracy

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

dimension measures both Test Accuracy and CV Accuracy as well as CV Standard Deviation. Log Loss, Brier Score and ECE are part of the Calibration dimension, measuring the alignment between the probabilities forecasted and the outcomes that are observed [14]. The Uncertainty dimension comprises of Predictive Uncertainty and Entropy, and Decision Reliability of Confidence and Prediction Margin.

Test Accuracy, CV Accuracy, Confidence, and Margin are considered benefit criteria, where higher numbers are desirable. The rest of the metrics—CV Standard Deviation, Predictive Uncertainty, Entropy, Log Loss, Brier Score, and ECE—are used as cost metrics, in which smaller values are better.

This group is used as the intermediate structure for the hierarchical weighting procedure and to isolate the individual evaluation criteria to the larger dimensions in the final model ranking.

Mathematical Formulation

To ensure comparability among criteria with different scales and optimization directions, the evaluation values are first transformed into a common benefit-oriented representation. For cost criteria, the implementation applies

$$x'_{ij} = max(x_j) - x_{ij}$$

where x_{ij} represents the value of criterion j for model i, and x'_{ij} is the corresponding benefit-oriented value. Min-max normalisation is then applied:

$$z_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)}$$

Following normalization, larger values consistently indicate more desirable performance.

The complete mathematical formulation implemented in MASE-RAMR is summarized in **Table 1**.

Table 1. Mathematical formulation of the proposed MASE-RAMR framework

Stage	Operation	Formulation
1	Cost-to-benefit transformation	$x'_{ij} = max(x_j) - x_{ij}$
2	Min-max normalization	$z_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)}$
3	Local CRITIC information content	$C_j = \sigma_j \sum_{k=1}^{m} (1 - r_{jk})$
4	Local CRITIC weight	$w_j = \frac{C_j}{\sum_{q=1}^{m} C_q}$
5	Adaptive Criteria Matrix	$A_{id} = \sum_{j \in d} z_{ij} w_j$
6	ACM normalization	$R_{id} = \frac{A_{id} - min(A_d)}{max(A_d) - min(A_d)}$
7	Global CRITIC information content	$C_d = \sigma_d \sum_{k=1}^{D} (1 - r_{dk})$

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

8	Global dimension weight	$W_d = \frac{C_d}{\sum_{q=1}^{D} C_q}$
9	Weighted decision matrix	$V_{id} = R_{id} W_d$
10	Positive ideal solution	$V_d{}^{+} = max_i(V_{id})$
11	Negative ideal solution	$V_d{}^{-} = min_i(V_{id})$
12	Distance from positive ideal	$D_i{}^{+} = \sqrt{\sum_{d=1}^{D} (V_{id} - V_d{}^{+})^2}$
13	Distance from negative ideal	$D_i{}^{-} = \sqrt{\sum_{d=1}^{D} (V_{id} - V_d{}^{-})^2}$
14	**RAMR score**	$RAMR_i = \frac{D_i{}^{-}}{D_i{}^{+} + D_i{}^{-}}$

Here, σ denotes standard deviation and r denotes the correlation coefficient. The indices i, j, d, k, m, and D represent candidate model, criterion, dimension, comparison criterion/dimension, number of criteria within a dimension, and total number of dimensions, respectively.

Hierarchical CRITIC Weighting

The first CRITIC stage is performed independently within each evaluation dimension. For criterion j, the information content is determined from its standard deviation and its correlation with the other criteria in the same dimension [9]. The resulting local weights are used to aggregate the normalized criteria belonging to each dimension according to

$$A_{id} = \sum_{j \in d} z_{ij} w_j$$

The resulting Adaptive Criteria Matrix represents each candidate model using four dimension-level scores.

A second CRITIC analysis is then applied to the normalized Adaptive Criteria Matrix. This stage determines the global importance of the four dimensions:

$$W_d = \frac{C_d}{\sum_{q=1}^{d} C_q}$$

Thus, the two weighting stages serve different purposes: **local CRITIC determines the relative importance of individual criteria within each dimension, whereas global CRITIC determines the relative importance of the dimensions themselves**. This hierarchical weighting structure forms the adaptive component of MASE.

Reliability-Aware Model Ranking

The normalized Adaptive Criteria Matrix is then combined with the global dimension weights to form the weighted decision matrix:

$$V_{id} = R_{id} W_d$$

The resulting matrix is then evaluated using TOPSIS procedure [10]. Since the cost criteria have already been

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

transformed into benefit-oriented values, the positive and negative ideal solutions are determined using the maximum and minimum values, respectively:

$$V_d^+ = max_I(V_{id}) \qquad V_d^- = min_i(V_{id})$$

The Euclidean distances of each candidate model from the positive and negative ideal solutions are then calculated as

$$D_i^+ = \sqrt{\sum_{d=1}^{D} (V_{id} - V_d^+)^2}$$

And

$$D_i^- = \sqrt{\sum_{d=1}^{D} (V_{id} - V_d^-)^2}$$

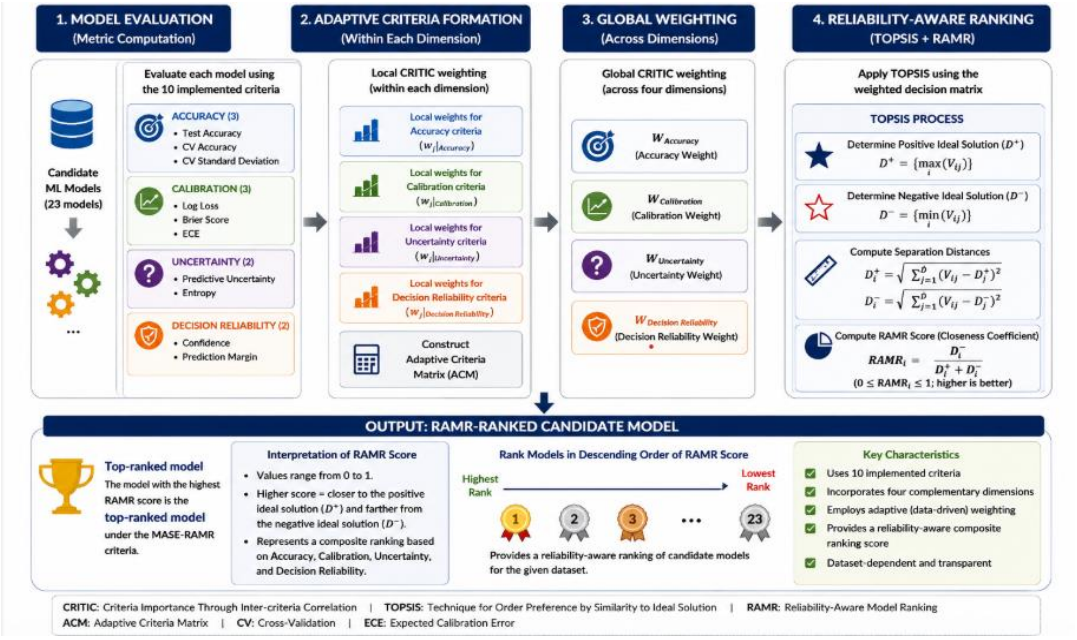
The final RAMR score is the TOPSIS closeness coefficient:

$$RAMR_i = \frac{D_i^-}{D_i^+ + D_i^-}$$

A higher RAMR score indicates greater relative closeness to the positive ideal solution across the dimensions incorporated into the framework. Candidate models are therefore ranked in descending order of their RAMR scores.

Figure 2. Reliability-Aware Model Ranking (RAMR) framework

(From Metric Evaluation to RAMR-Ranked Candidate Model)



Interpretation of RAMR

RAMR is a **composite model-ranking measure**, rather than a replacement for predictive accuracy. Test

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

Accuracy describes a specific aspect of predictive performance, whereas RAMR evaluates the relative position of each candidate model across the multiple dimensions incorporated into the proposed framework.

Consequently, the model with the highest Test Accuracy does not necessarily have to obtain the highest RAMR score. A candidate model could be less accurate but get a higher RAMR score due to its performance against the calibration, uncertainty and decision-reliability criteria.

The purpose of RAMR is thus to investigate whether selection on accuracy alone yields the "best" model or whether adding other complementary dimensions of model evaluation results in different or more informative rankings of the candidate models.

Comparison with Conventional CRITIC-TOPSIS

The proposed framework is compared with conventional flat CRITIC–TOPSIS approach. Both methods follow the already developed CRITIC method and TOPSIS method [9,10] and the only difference between these two methods is in the arrangement and weighting of the evaluation criteria.

The conventional approach can be shown as:

Evaluation Metrics → CRITIC → TOPSIS → Model Ranking.

In contrast, MASE-RAMR follows:

Evaluation Metrics→Clinical Dimensions→Local CRITIC→Adaptive Criteria Matrix→Global CRITIC→TOPSIS→RAMR.

The main contribution of the proposed framework is the hierarchical integration of heterogeneous evaluation criteria rather than the introduction of a new CRITIC or TOPSIS algorithm. .

Model Selection Procedure

For each dataset, the same candidate ML models and evaluation criteria are used in both the proposed and conventional frameworks. The MASE-RAMR procedure follows these steps:

1. Assess the candidate ML models on the pre-defined metrics.
2. Convert cost criteria to values based on benefits, and normalize metrics.
3. Classify metrics in the four evaluation dimensions.
4. Use local CRITIC weighting in each dimension.
5. Create the Adaptive Criteria Matrix.
6. Normalize the Adaptive Criteria Matrix and then use the global CRITIC weighting.
7. Create the weighted decision matrix.
8. Use TOPSIS and determine RAMR score.
9. List the candidate models in order of the RAMR score.
10. Compare the ranking obtained with the traditional CRITIC-TOPSIS ranking.

The model having the highest RAMR score is termed as RAMR-selected model for each dataset. The following analysis investigates if this selection conforms or diverges from the traditional ranking and how decisions based on the precision and the extra reliability dimensions affect it.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

Experimental Setup and Results

Experimental Setup

Benchmark Datasets

The proposed framework MASE-RAM was tested on six cardiac-related databases—namely, Stroke, Statlog, Hungarian, Cleveland, Processed VA and Processed Switzerland. Each Data set was evaluated independently using the same candidate model evaluation procedure and the same evaluation criteria. The design of this dataset-wise structure enables the weighting matrix structure and thus the model ranking to be set depending on the properties of the individual evaluation matrix and not fixed for all datasets in a common matrix structure.

Three of the datasets—Cleveland, Hungarian, Switzerland, and Long Beach VA—are from the UCI Machine Learning Repository, and Stroke was downloaded from Kaggle. During the experiments the VA and Switzerland datasets were used in this processed state and are called Processed VA and Processed Switzerland, respectively.

Candidate Machine-Learning Models

Each of the 6 datasets was tested independently on a common set of 23 candidate machine-learning models. The classifiers in the pool encompassed conventional ones (such as statistical classifiers), instance-based ones, tree-based classifiers, boosting methods, neural-network classifiers, and ensemble classifiers.

The same set of candidates across the datasets offers a common ground to study the evolution of model preferences across evaluation features and weights. The goal is therefore not to find the best all-around algorithm, but to find the algorithm that has the best overall evaluation profile in each set of data.

Evaluation Criteria and MASE Dimensions

Ten criteria as specified in Section 3 were used to assess each candidate model. The criteria were grouped under four MASE dimensions: Accuracy, Calibration, Uncertainty and Decision Reliability. Accuracy refers to Test Accuracy, Cross-Validation Accuracy, and Cross-Validation Standard Deviation; Uncertainty refers to Predictive Uncertainty and Entropy; Calibration refers to Log Loss, Brier Score, and Expected Calibration Error (ECE); and Decision Reliability refers to Confidence and Prediction Margin.

The grouping provides the intermediate structure required by the hierarchical weighting procedure. Rather than allowing all ten criteria to enter the final decision as a single undifferentiated set, related criteria are first evaluated within their respective dimensions. Local CRITIC weighting is then applied within each dimension, followed by global CRITIC weighting of the resulting dimensions. This follows the hierarchical procedure defined in Section 3.

Comparative Methods

Two model-selection approaches were evaluated using the same underlying model-evaluation results.

The first was **Flat CRITIC–TOPSIS**, in which the ten evaluation criteria were treated as a single decision matrix. CRITIC provides objective criterion weighting by considering the contrast intensity and correlation structure of the criteria, while TOPSIS ranks alternatives according to their relative closeness to positive and negative ideal solutions [9,10].

The second was the proposed **MASE-RAMR** framework. The ten criteria were first organized into the four MASE dimensions, followed by local CRITIC weighting, construction of the Adaptive Criteria Matrix, global CRITIC weighting, and final TOPSIS ranking. The distinction between the two approaches is therefore the organization and weighting of the evaluation evidence rather than the introduction of a new CRITIC or TOPSIS algorithm.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

Experimental Workflow

For each dataset, the 23 candidate models were first evaluated using the predefined criteria. The resulting evaluation matrix was then processed independently using Flat CRITIC–TOPSIS and MASE-RAMR.

The conventional pathway was:

Evaluation criteria → CRITIC weighting → TOPSIS → model score and ranking

The proposed pathway was:

Evaluation criteria → MASE dimensions → local CRITIC weighting → Adaptive Criteria Matrix → global CRITIC weighting → TOPSIS → RAMR score and ranking.

The resulting rankings were compared at both the top-model and complete-ranking levels. Global and local CRITIC weights were also examined to determine how the relative importance of the dimensions and individual criteria varied across datasets.

Figure 3. Experimental workflow for the comparative evaluation of Flat CRITIC–TOPSIS and MASE-RAMR.

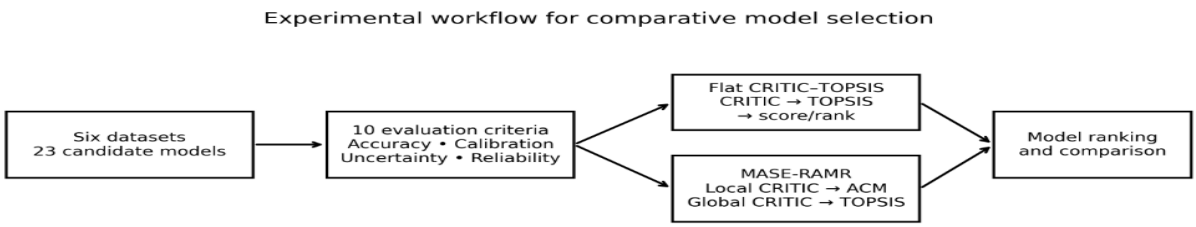
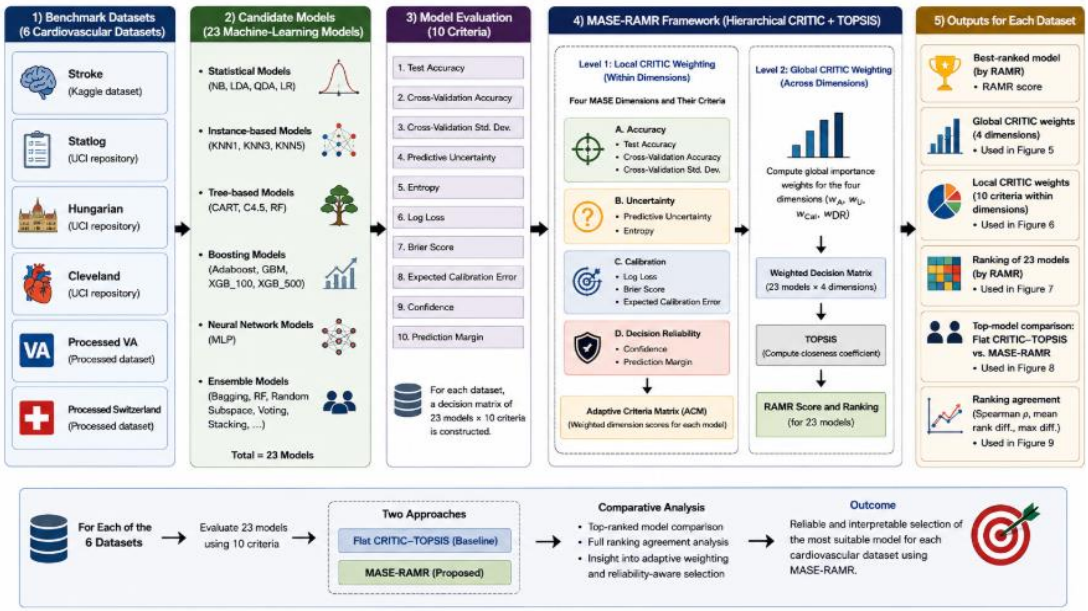


Figure 4. Dataset-wise MASE-RAMR model-selection overview.



Experimental Results

The finalized MASE-RAMR results were obtained independently for the six benchmark datasets using the same pool of 23 candidate machine-learning models. Unlike a single aggregate score applied across all datasets, the RAMR score and the resulting ranking are calculated separately for each dataset. This allows the framework to

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

reflect differences in the information content and relative importance of the Accuracy, Uncertainty, Calibration, and Decision Reliability dimensions across datasets.

Table 4. Best-ranked model selected by the finalized MASE-RAMR framework

Dataset	**Best-ranked model**	**RAMR score**
Stroke	XGB_500	0.763885
Statlog	NB	0.762437
Hungarian	NB	0.798394
Cleveland	NB	0.790458
Processed VA	GBM	0.767795
Processed Switzerland	GBM	0.941427

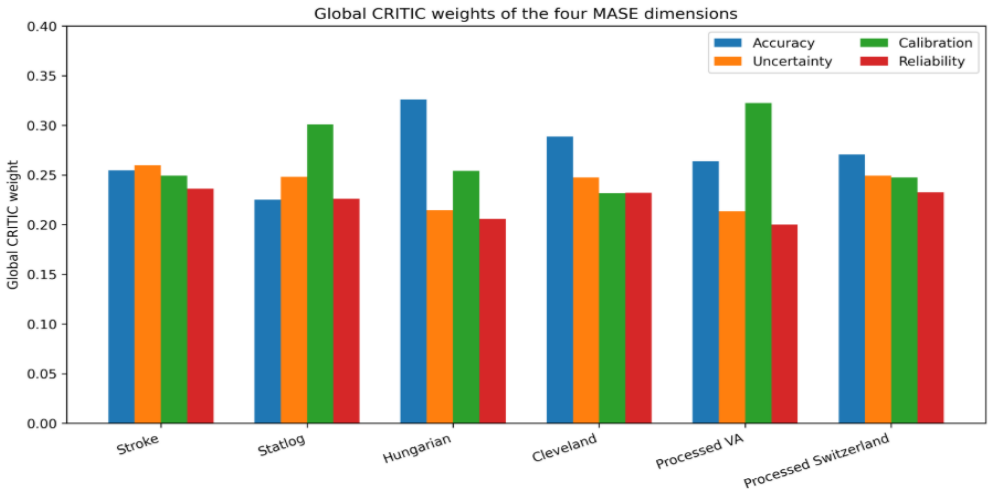
The finalized rankings identify **XGB_500** as the highest-ranked model for Stroke, **NB** for Statlog, Hungarian, and Cleveland, and **GBM** for Processed VA and Processed Switzerland. Thus, the preferred candidate varies across datasets rather than following a single model-family preference.

The RAMR score should consequently be interpreted as a relative decision measure within each evaluation setting. A higher score represents a higher distance to the positive ideal solution in the respective dataset, underpinning the TOPSIS formulation given in Section 3.

Global CRITIC Weight Analysis

The global CRITIC stage uses the Adaptive Criteria Matrix to determine the contribution of each MASE dimension (CI, EI, DI, and WDI) to the global evaluation after the local weights have been calculated. Because CRITIC weighting is performed separately for each dataset, the resulting dimension weights reflect the information structure of the corresponding candidate-model evaluation matrix.

Figure 5. Global CRITIC weights of the four MASE evaluation dimensions across the six datasets.



The four MASE dimensions do not receive weights that are not identical across the six datasets. Accuracy receives the largest weight for Hungarian and Cleveland, whereas Calibration is the largest contributor for Statlog and Processed VA. Stroke and Processed Switzerland show comparatively balanced contributions from the four dimensions.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

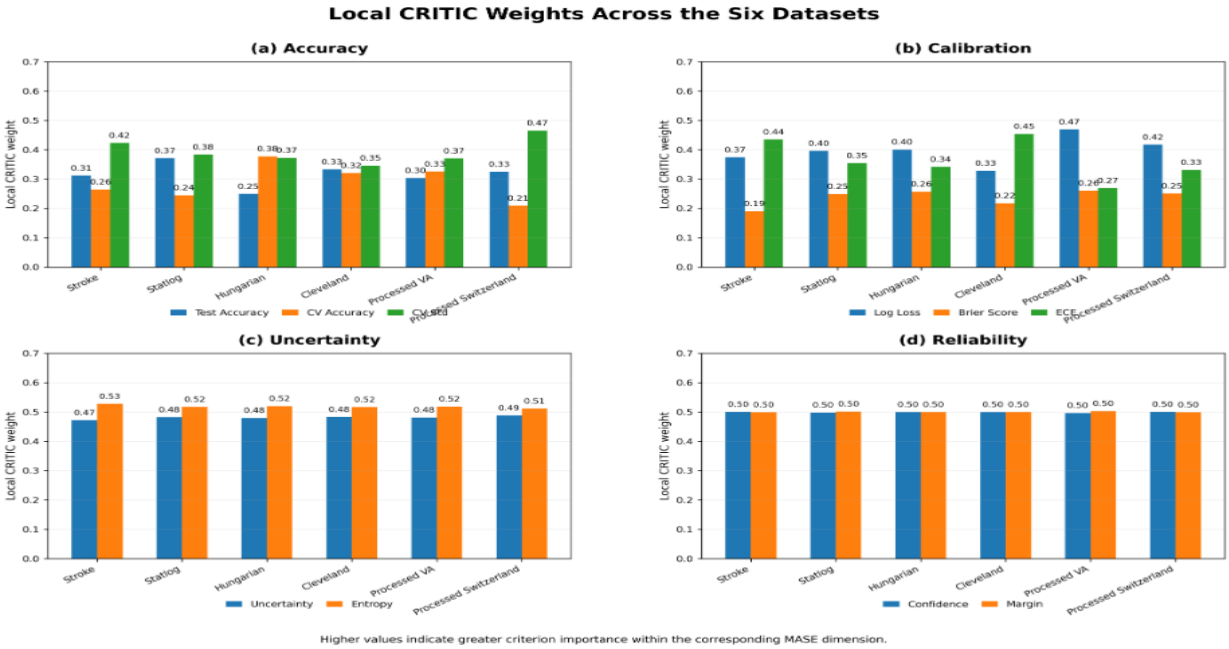
For Statlog, Calibration receives the highest weight at 0.3009, compared with 0.2250 for Accuracy, 0.2481 for Uncertainty, and 0.2261 for Reliability. In the Hungarian dataset, Accuracy has the largest global weight at 0.3260. These differences reveals that the relative contribution of the dimensions is not fixed in advance but varies across datasets.

The global weighting results show that therefore the weighting structure is dataset dependent. It is not necessary for Accuracy to have a strong influence on every component of the framework; each dimension's contribution depends on the characteristics of the information of the candidate – model evaluation results.

Local CRITIC Weight Analysis

Local CRITIC weighting determines the relative contribution of the individual criteria within each MASE dimension. Whereas Figure 5 describes the contribution of the four dimensions, Figure 6 shows how individual criteria contribute within those dimensions.

Figure 6. Local CRITIC weights of evaluation criteria within the four MASE dimensions across the six datasets.



The results show that criteria within the same dimension do not necessarily contribute equally. Their weights depend on the variability and inter-criterion relationships observed in the corresponding dataset. This follows the objective-weighting principle underlying CRITIC [20].

Within Accuracy, the local weighting distinguishes Test Accuracy, Cross-Validation Accuracy, and Cross-Validation Standard Deviation. Within Calibration, Log Loss, Brier Score, and ECE provide complementary information regarding probabilistic prediction quality. Calibration is particularly relevant because discrimination and calibration describe different aspects of clinical prediction-model performance [6–10].

Predictive Uncertainty and Entropy form the Uncertainty dimension, while Confidence and Prediction Margin form Decision Reliability. The inclusion of uncertainty is consistent with work emphasizing the importance of communicating uncertainty in medical machine-learning systems [12–16].

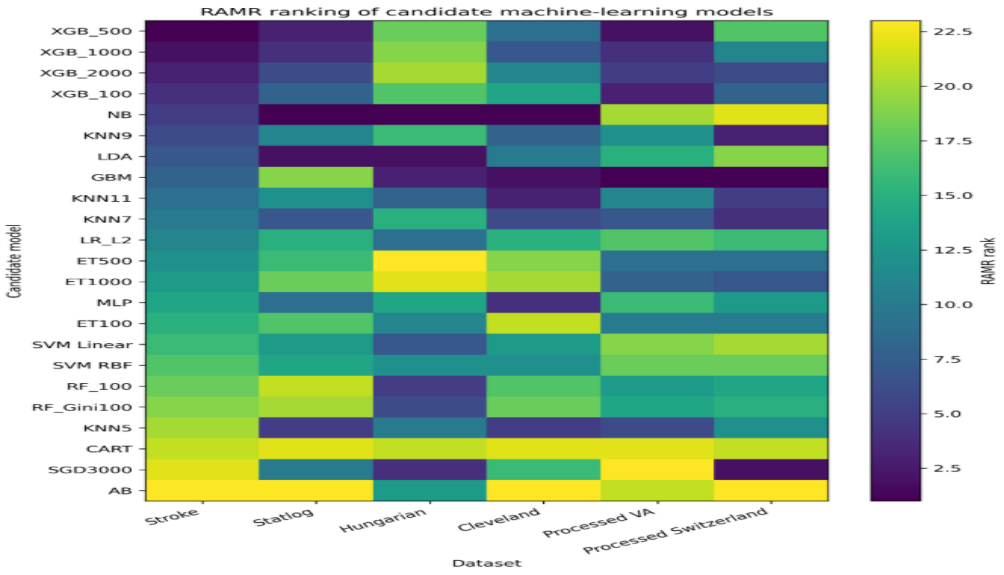
The local weights also vary across datasets. Thus, adaptation occurs at both levels of the hierarchy: the importance of individual criteria changes within dimensions, while the importance of the four dimensions changes at the global level.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

RAMR Ranking of Candidate Models

The final RAMR scores were used to rank all 23 candidate models independently for each dataset.

Figure 7. RAMR ranking of the 23 candidate machine-learning models across the six datasets.



The ranking heatmap shows considerable variation in model positions across the six datasets. XGB_500 ranks first for Stroke, NB ranks first for Statlog, Hungarian, and Cleveland, while GBM ranks first for Processed VA and Processed Switzerland.

The changes extend beyond the top-ranked model. Several candidates shift positions across datasets, reflecting differences in their overall multidimensional evaluation profiles. A model that performs well under one dataset-specific decision structure may therefore occupy a different position under another.

The ranking is therefore not simply an ordering based on accuracy. Instead, it reflects the relative position of each candidate after the Accuracy, Uncertainty, Calibration, and Decision Reliability dimensions have been incorporated into the RAMR decision process.

Comparative Analysis with Flat CRITIC–TOPSIS

The comparison with Flat CRITIC–TOPSIS was conducted to examine how the hierarchical organization of the evaluation criteria influences model selection. Both methods use the same ten criteria and 23 candidate models, but the criteria are aggregated differently before the final TOPSIS ranking.

Comparison of Top-Ranked Models

The first level of comparison considers the model selected as the highest-ranked candidate by each method for each dataset.

Table 5. Comparison of the top-ranked models obtained using Flat CRITIC–TOPSIS and MASE-RAMR

Dataset	**Flat CRITIC–TOPSIS**	**Flat Score**	**MASE-RAMR**	**RAMR Score**
Stroke	XGB_500	0.736922	**XGB_500**	**0.763885**
Statlog	LDA	0.747376	**NB**	**0.762437**
Hungarian	NB	0.727227	**NB**	**0.798394**

Cleveland	NB	0.768015	**NB**	**0.790458**
Processed VA	KNN5	0.692107	**GBM**	**0.767795**
Processed Switzerland	GBM	0.929751	**GBM**	**0.941427**

The two approaches identify the same top-ranked model for **Stroke, Hungarian, Cleveland, and Processed Switzerland**. The selections differ for **Statlog**, where Flat CRITIC–TOPSIS selects LDA and MASE-RAMR selects NB, and for **Processed VA**, where the respective selections are KNN5 and GBM.

Figure 8. Comparison of the top-ranked model selected by Flat CRITIC–TOPSIS and MASE-RAMR across the six datasets.

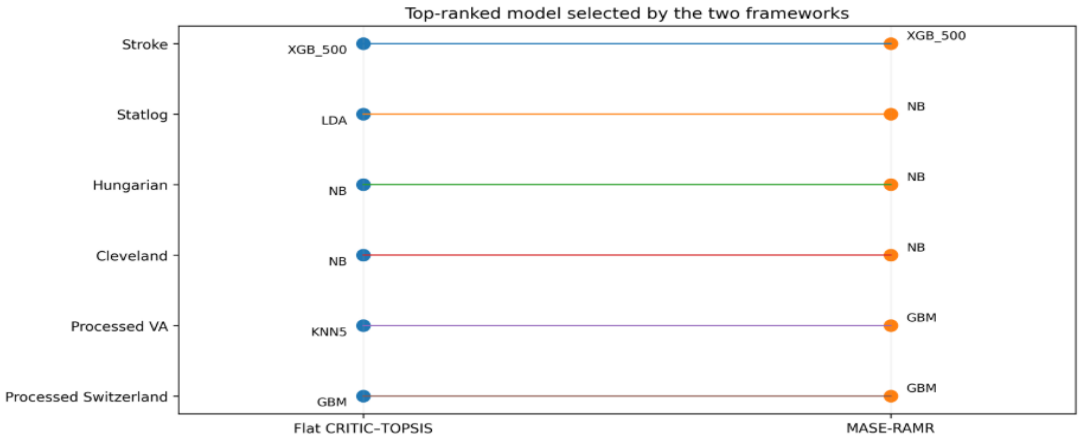


Figure 8 provides a visual representation of these stable and changed selections. The four datasets in which the two approaches agree demonstrate **decision stability**, whereas the Statlog and Processed VA datasets demonstrate **adaptive model selection** under the proposed framework.

The results show both **decision stability** and **adaptive selection**. In four datasets, the hierarchical procedure preserves the conventional top-ranked model. In two datasets, the additional organization of the evaluation criteria changes the preferred candidate. These changes should not be interpreted as direct evidence of improved predictive accuracy; they indicate a difference in the overall multi-criteria decision produced by the two weighting structures.

Complete Ranking Agreement Across the 23 Candidate Models

Top-model agreement alone does not describe how the remaining candidates are affected. Therefore, the complete rankings of all 23 models were compared using Spearman's rank correlation, mean absolute rank difference, and maximum rank difference.

Table 6. Agreement between Flat CRITIC–TOPSIS and MASE-RAMR rankings

Dataset	**Spearman ρ**	**Mean absolute rank difference**	**Maximum rank difference**
Stroke	0.966	1.39	5
Statlog	0.729	4.09	10
Hungarian	0.661	3.26	19
Cleveland	0.920	2.26	7

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

Processed VA	0.945	1.74	5
Processed Switzerland	0.967	1.48	4

Figure 9 shows the agreement between Flat CRITIC–TOPSIS and MASE-RAMR rankings across the 23 candidate machine-learning models for the six datasets.

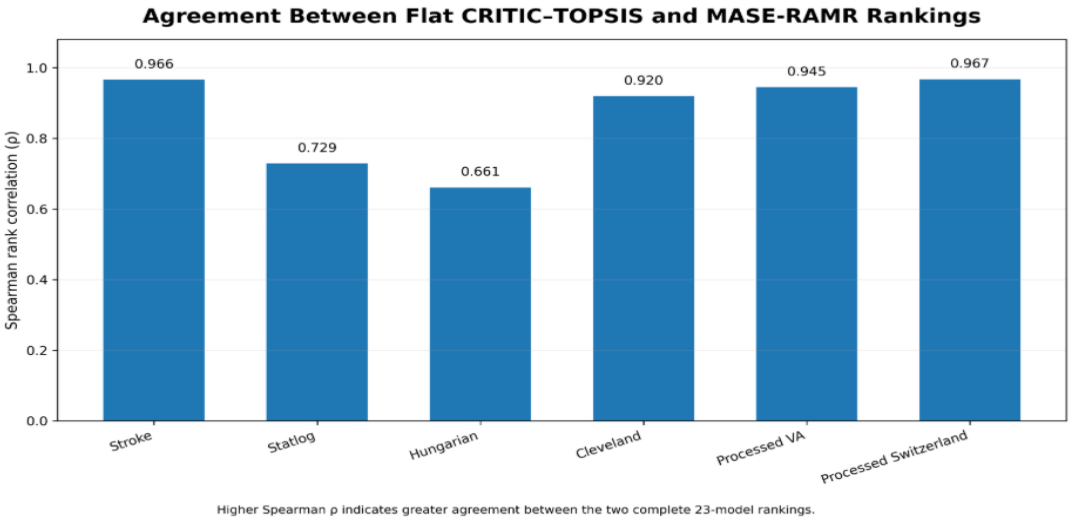


Figure 9. Overall consistency of the rankings of the 23 candidate machine-learning models for the six datasets obtained by Flat CRITIC–TOPSIS and MASE-RAMR methods.

The strongest ranking agreement occurs for Processed Switzerland (ρ = 0.967), Stroke (ρ = 0.966), Processed VA (ρ = 0.945), and Cleveland (ρ = 0.920). Lower agreement is observed for Statlog (ρ = 0.729) and Hungarian (ρ = 0.661).

The mean absolute rank differences give an alternative perspective on the ranking changes. For Stroke, the average error is 1.39 positions, while the average error for Processed Switzerland is 1.48 positions, with the Hungarian showing 4.09 positions and the Statlog showing 3.26 positions. The most significant rank difference is seen in Hungarian, where the maximum rank difference is 19.

The results make quite a difference between top-model stability and complete-ranking stability. When there is significant reordering of the lower candidates, the same winning model might still emerge from a dataset.

Interpretation of Ranking Changes

The main difference between the two methods lies in the way the evaluation evidence is arranged prior to being used in TOPSIS. Flat CRITIC–TOPSIS assumes that all ten criteria are at the same decision level while MASE-RAMR assigns local weights to criteria and then calculates the global dimension weights in four groups of related criteria.

There are two distinct patterns in the comparison. The top ranked model is the same for both in the cases of Stroke, Hungarian, Cleveland and Processed Switzerland. The preferred model, on the other hand, alters for Statlog and Processed VA.

The complete rankings show why top-model stability and complete-ranking stability should not be treated as the same thing. Hungarian retains NB as the top-ranked model despite having the lowest rank correlation, whereas Processed VA changes its top-ranked model despite maintaining a relatively high overall correlation.

The comparison therefore indicates that MASE-RAMR provides adaptive discrimination without substantially altering the conventional ranking. Its influence depends on the structure of the candidate-model evaluation

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

profiles within each dataset.

Cross-Dataset Analysis

Across the six datasets, MASE-RAMR shows three related characteristics: adaptive weighting, dataset-dependent model preference, and varying levels of ranking stability.

Adaptive Weighting Across Datasets

The global and local CRITIC results show that the relative contribution of the dimensions and individual criteria varies across datasets. This suggests that the weighting structure is driven by the information contained in the evaluation data rather than imposed as a fixed hierarchy.

There are two levels of adaptation: Global CRITIC is used to estimate the relative importance of the four dimensions of MASE, and local CRITIC is used to estimate the contribution of individual criteria within the four dimensions of MASE. The proposed framework is designed to be adaptive, with a two-stage weighting process.

Dataset-Dependent Model Selection

In the 6 datasets, the 3 models: XGB_500, NB and GBM are selected using MASE-RAMR. The results of their distribution reveals that this model does not yield one common winner.

Three sets are selected for NB, two sets for GBM and one set for XGB_500. These models were repeatedly chosen, indicating their positive multi-dimensionality for the corresponding evaluation context, but not in any sense superior per se.

This is confirmed by the baseline comparison. In four datasets, the conventional model selection remains the same, while in two datasets it changes. So, the impact of hierarchical weighting is not systematic but varies.

Ranking Stability and Adaptive Selection

A more detailed picture of this adaptation is given in the complete-ranking analysis. The highest level of agreement between the two methods is found in the two cases of Stroke and Processed Switzerland, whereas there is more disagreement in the reordering of candidate rankings in the two other cases, Statlog and Hungarian.

This is especially crucial to distinguish between the top-model stability and the complete-ranking stability. Processed VA shows minor change of selected model with relatively high rank correlation value, while the Hungarian model maintains the selected model with a large overall reordering of the candidate models.

This indicates that the stability of the models should not be evaluated just on the basis of the same model being the winner. When the decision structure has a significant impact on the relative position of competing candidates, complete-ranking analysis can offer further insight into this structure.

Cross-Dataset Interpretation

Combined, the results indicate that MASE-RAMR is not biased in favor of or against any specific class of algorithms. Rather, the direction and intensity of its impact is determined by the information collected in the Accuracy, Uncertainty, Calibration, and Decision Reliability criteria.

The latter is a natural extension of the purpose of the framework, that is, to combine multiple complementary aspects of prediction, not to maximize any given measure of predictive performance. This is also in line with the clinical prediction literature where discrimination and calibration and other aspects of prediction performance carry different information about the model's behaviour [1,2,6,17,20].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

OVERALL FINDINGS AND DISCUSSION

Main Empirical Findings

Based on the experimental analysis, there are three major findings.

The first one is that preference is data dependent. The candidate who is chosen from each of the six datasets does not necessarily win all of them, thus the final RAMR decision is determined by the multidimensional evaluation profile of the candidate models of each dataset.

Second, the weighting system is adaptive. Each dataset is unique and the value of each criterion and the four MASE dimensions will differ from dataset to dataset. The framework does not have criterion or dimension weights or any other fixed criterion.

Third, MASE-RAM has a ranking stability and adaptive differentiation. Four datasets have the same best model determined by Flat CRITIC–TOPSIS while the two datasets yield different best model. The full ranking analysis also reveals a wide variation in the amount of reordering among the datasets.

The results in total answer the main research question: Can the selection of models that are primarily based on predictive accuracy be extended using complementary information related to reliability?

Contribution of Hierarchical Adaptive Weighting

The principal methodological contribution of MASE-RAMR is the separation of criterion-level and dimension-level weighting.

At the first level, local CRITIC determines the relative contribution of individual criteria within each MASE dimension. At the second level, global CRITIC determines the contribution of the four dimensions after construction of the Adaptive Criteria Matrix. CRITIC's objective weighting principle is based on criterion variability and inter-criterion correlation [9].

This structure differs from a flat MCDM formulation in which all criteria enter the decision at the same level. Previous work has demonstrated the usefulness of MCDM for machine-learning model selection [11,12], while MASE-RAMR extends this idea by introducing an intermediate organization of related evaluation measures.

The experimental results show that the hierarchy does not systematically reverse conventional rankings. Instead, it preserves the top decision in some datasets and changes it in others. The contribution of the hierarchy is therefore best understood as **structured, data-dependent aggregation of evaluation evidence**.

Reliability Beyond Predictive Accuracy

The results support the central premise of the framework that predictive accuracy does not necessarily provide a complete representation of prediction-model performance. The assessment of clinical prediction-models often focuses on discrimination aspects of predictive performance in addition to calibration aspects [1,6,9,17,20]. The role of uncertainty in interpreting and communicating medical ML predictions is also growing in importance, and is increasingly recognised [3–5].

All these complementary features are combined in a unified model-selection framework by MASE-RAMR. The results in comparison to Flat CRITIC–TOPSIS and MASE-RAMR reveal that the end model preference could be influenced by other factors other than the predictive performance. Under this model, does not mean that accuracy is not important.

It is one of the 4 MASE dimensions that is still present. Instead, the framework looks at the candidate's predictive performance in conjunction with a positive profile of calibration, uncertainty, and decision reliability.

Therefore, the results support a multidimensional perspective of model quality but not that a higher RAMR score

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VIII August 2026

equates to increased clinical benefit.

Practical Implications for Model Selection

The results indicate that cardiovascular prediction modelling can be considered as a multi-criteria decision problem, which is dataset specific. Instead of taking for granted that a single algorithm would be the best in all scenarios, researchers can assess competing models using multiple complementary properties.

The adaptive weighting mechanism also minimises the need to set criterion weights by hand. The relative importance of the criteria/dimensions, however, is not set a priori by subjective preference but rather observed from the decision matrix with the use of CRITIC.

The framework can thus provide a decision support layer following conventional model evaluation. It is not a substitute for the standard measures of predictive performance, but rather these are supplemented by calibration, uncertainty, and decision reliability information.

But RAMR is not to be taken as a replacement for clinical validation. Not choosing a model on a benchmark dataset does not make it a clinically superior model. Modern counselling on clinical prediction models focuses more on the validation, reporting, applicability, and performance of these models in the development data set than in the clinical data set of interest [15,16].

Technical Implications

The findings have several technical implications for reliability-aware machine-learning model selection.

First, MASE-RAMR demonstrates that model selection can consider accuracy, uncertainty, calibration, and decision reliability rather than relying on predictive accuracy alone.

Second, the adaptive weighting mechanism allows the relative importance of these criteria to vary across **datasets**, reducing dependence on manually fixed weights. The hierarchical structure also provides a transparent pathway from individual evaluation criteria to the final model ranking.

Third, comparison with Flat CRITIC–TOPSIS shows that hierarchical weighting can preserve overall ranking **patterns** while producing meaningful differences in individual model positions. Therefore, both top-model agreement and complete-ranking stability are important when evaluating multi-criteria ranking methods.

Finally, MASE-RAMR is a model-ranking framework rather than a clinical outcome optimization method. A higher RAMR score indicates a stronger overall profile according to the selected criteria, but does not establish clinical superiority. External datasets, clinical utility analysis, and prospective validation are therefore required before clinical deployment.

Limitations and Future Directions

There are some restrictions to be noted regarding the present results. First, it was experimented on a set of six cardiovascular datasets with a constant set of 23 candidate models. Conclusions may then not be extended beyond the results obtained with the other clinical populations, disease domains, or distinctly different pools of candidate models.

Second, the positions will be based on the factors that are incorporated into the decision matrix. Under other possible levels of uncertainty, calibration or clinically oriented utility, the weights and the model ranking might be different.

Third, weights obtained with CRITIC are objective and depend on data, but might vary as a function of the pool of candidates and the data characteristics. Thus, it is important to note that the RAMR should not be used as an absolute value, but rather compared to it within a specific evaluation context.

Fourth, the model selection and the ranking procedure is different between the two approaches (Flat CRITIC–TOPSIS) and their superiority in this sense cannot be claimed. A move in rank does not necessarily mean that there is a greater clinical benefit.

A larger and more diverse clinical cohort should be tested, in addition to a larger candidate model set, with a focus on the sensitivity and stability of the weighting and ranking methods, for future studies. Other clinically relevant utility criteria could broaden the scope toward model selection for decisions. Finally, external and prospective validation will be required to assess the relationship between the multidimensional reliability assessment and better clinical decision support.

CONCLUSION

In this study, the authors introduced a Multi-Criteria Adaptive Selection Engine with Reliability-Aware Model Ranking for cardiovascular disease prediction called MASE-RAMR. The framework goes beyond traditional model-selection methods by combining Accuracy, Calibration, Uncertainty and Decision Reliability in a hierarchy and within an adaptive weighting scheme.

Across the six datasets for cardiovascular diseases, and 23 candidate machine learning models, the top model was found to be specific to the dataset. For Stroke, NB was chosen for Statlog, Hungarian and Cleveland, and GBM was chosen for Processed VA and Processed Switzerland. The results also indicated that the relative value of the individual criteria and the dimensions of the MASE is different across different datasets which is further in favor of adaptive weighting.

Comparison with Flat CRITIC–TOPSIS showed substantial agreement across several datasets, while also revealing differences in model selection. The two approaches selected the same top model for four datasets, whereas MASE-RAMR selected a different model for Statlog and Processed VA. Complete-ranking correlations further showed that the framework can preserve the overall ranking structure while providing additional differentiation among competing models.

In summary, the results align with the idea of adopting multidimensional reliability-aware model selection over the predictive accuracy only. MASE-RAMR offers a framework with a structured and dataset-adaptive methodology to incorporate both predictive performance and calibration, uncertainty, and decision reliability. However, the data used as benchmarks here and thus the present results are not to be considered as proof of clinical superiority. The framework needs to be tested on more and independent clinical data sets and include clinically relevant utility measures and prospective testing.

REFERENCES

1. **Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P. J., McGinn, T., & Guyatt, G. (2017).** Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. JAMA, 318(14), 1377–1384. **https://doi.org/10.1001/jama.2017.12126**
2. **Huang, Y., Li, W., Macheret, F., Gabriel, R. A., & Ohno-Machado, L. (2020).** A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association, 27(4), 621–633. https://doi.org/10.1093/jamia/ocz228
3. **Huang, L., Ruan, S., Xing, Y., & Feng, M. (2024).** A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. Medical Image Analysis, 97, 103223. https://doi.org/10.1016/j.media.2024.103223
4. **Kompa, B., Snoek, J., & Beam, A. L. (2021).** Second opinion needed: Communicating uncertainty in medical machine learning. npj Digital Medicine, 4, 4. https://doi.org/10.1038/s41746-020-00367-3
5. **Seoni, S., Jahmunah, V., Salvi, M., Barua, P. D., Molinari, F., & Acharya, U. R. (2023).** Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). Computers in Biology and Medicine, 165, 107441. https://doi.org/10.1016/j.compbiomed.2023.107441
6. **Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019).** Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7

7. **Liu, T., Krentz, A., Lu, L., & Curcin, V. (2024).** Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. European Heart Journal – Digital Health, 6(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080
8. **Alemu, Y. M., Alemu, S. M., Bagheri, N., Wangdi, K., & Chateau, D. (2025).** Discrimination and calibration performances of non-laboratory-based and laboratory-based cardiovascular risk predictions: A systematic review. Open Heart, 12(1), e003147.
9. **Diakoulaki, D., Mavrotas, G., & Papayannakis, L. (1995).** Determining objective weights in multiple criteria problems: The CRITIC method. Computers & Operations Research, 22(7), 763–770. https://doi.org/10.1016/0305-0548(94)00059-H
10. **Hwang, C.-L., & Yoon, K. (1981).** Multiple attribute decision making: Methods and applications: A state-of-the-art survey. Springer. https://doi.org/10.1007/978-3-642-48318-9
11. **11.Ali, R., Lee, S., & Chung, T. C. (2017).** Accurate multi-criteria decision making methodology for recommending machine learning algorithm. Expert Systems with Applications, 71, 257–278. https://doi.org/10.1016/j.eswa.2016.11.034
12. **Leyva-Vázquez, M. Y., Briones Peñafiel, L. A., Sanchez Muñoz, S. X., & Quiroz Martinez, M. A. (2021).** A framework for selecting machine learning models using TOPSIS. In T. Ahram (Ed.), Advances in artificial intelligence, software and systems engineering (pp. 119–126). Springer. https://doi.org/10.1007/978-3-030-51328-3_18
13. **Ganie, S. M., Dutta Pramanik, P. K., & Zhao, Z. (2025).** Enhanced and interpretable prediction of multiple cancer types using a stacking ensemble approach with SHAP analysis. Bioengineering, 12(5), 472. https://doi.org/10.3390/bioengineering12050472
14. **Brier, G. W. (1950).** Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
15. **Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024).** TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378
16. **Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Oakden-Rayner, L., ... van Smeden, M. (2025).** PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, e082505. https://doi.org/10.1136/bmj-2024-082505
17. **Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., & Ziegler, A. (2023).** Calibrating machine learning approaches for probability estimation: A comprehensive comparison. Statistics in Medicine, 42(29), 5451–5478. https://doi.org/10.1002/sim.9921
18. **Zargarzadeh, A., Javanshir, E., Ghaffari, A., Mosharkesh, E., & Anari, B. (2023).** Artificial intelligence in cardiovascular medicine: An updated review of the literature. Journal of Cardiovascular and Thoracic Research, 15(4), 204–209. https://doi.org/10.34172/jcvtr.2023.33031
19. X. Sun, Y. Yin, Q. Yang, and T. Huo, "Artificial intelligence in cardiovascular diseases: diagnostic and therapeutic perspectives," European Journal of Medical Research, vol. 28, no. 1, p. 242, 2023, doi: 10.1186/s40001-023-01065-y.
20. **Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010).** Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology, 21(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2