MULTI-MODAL FEATURE SELECTION AND PARETO-BASED MODEL FUSION FOR BREAST CANCER COMPUTER-AIDED DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has been widely applied to breast cancer imaging analysis, yet most existing studies optimize individual metrics such as the area under the receiver operating characteristic curve (AUC) while overlooking feature acquisition cost, feature redundancy, and prediction reliability. This paper proposes a computer-aided diagnosis method based on feature-group multi-objective optimization and Pareto model fusion. The method jointly maximizes AUC and minimizes the number of feature groups, and dynamically fuses Pareto sub-models by integrating predictive performance, feature-group count, and predictive uncertainty, yielding patient-level benign/malignant probabilities with uncertainty estimates. Experiments on clinical data and digital breast tomosynthesis (DBT) images from The Cancer Imaging Archive (TCIA) demonstrate that the proposed method reduces the number of feature groups while maintaining diagnostic performance and provides a quantitative basis for evaluating prediction reliability and stability.

Keywords: computer-aided breast cancer diagnosis; multi-modality; multi-objective optimization; Monte Carlo dropout; uncertainty estimation; model fusion.

1 INTRODUCTION

Breast cancer is one of the most common malignancies among women worldwide, and its incidence continues to rise globally Xue et al. (2021). In 2020, approximately 2.3 million women were newly diagnosed with breast cancer, making it the most commonly diagnosed cancer globally, surpassing lung cancer Sung et al. (2021). Screening-facilitated early detection and advances in adjuvant therapy both contribute to reductions in breast cancer-related mortality Berry et al. (2005). For appropriately selected patients with early-stage breast cancer, breast-conserving treatment reduces surgical extent while maintaining long-term survival Narod et al. (2015). Improving early detection and diagnostic accuracy is therefore of considerable clinical importance.

AI is pivotal in the early detection and diagnosis of breast cancer Desai & Desai (2025). Its major strength lies in integrating multimodal information, including imaging features from mammography, ultrasonography, and MRI, as well as clinical and genomic data Arya & Saha (2021). McKinney et al. (2020) develop a deep learning-based AI system for breast cancer screening, while Rodriguez-Ruiz et al. (2019) compare a stand-alone AI system with 101 radiologists. Hamdy et al. (2022) develop a DenseNet201-based model with hierarchical fusion to integrate multimodal breast cancer images, and Akselrod-Ballin et al. (2019) jointly analyze mammograms and linked electronic health records, showing that multimodal models outperform single-source models. However, modality heterogeneity often requires complex architectures with substantial computational, annotation, and deployment costs, especially in resource-limited primary care Alzamil et al. (2025). Conversely, lightweighting alone may compromise sensitivity and specificity. AI-based breast cancer diagnosis has thus become a multi-objective optimization problem, requiring a balance between accuracy and cost. Multi-objective optimization offers a framework for this trade-off; by identifying Pareto-optimal solutions, it supports feature selection and model lightweighting while preserving clinically meaningful performance Chen et al. (2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Although AI models for breast cancer detection have progressed substantially, AI-assisted diagnosis still faces challenges in real-world clinical practice. First, existing models are typically optimized for classification metrics such as AUC, accuracy, and sensitivity McKinney et al. (2020). Yuan et al. (2021) note that AUC maximization has become a prominent objective in deep learning, especially in medical image classification. However, clinical utility is also influenced by the number of features required, examination cost, and completeness of patient information Kursa & Rudnicki (2010). If a model incorporates many clinical features for only marginal gains, it may increase data acquisition burden and reduce applicability in primary care or incomplete-data scenarios. AI-assisted breast cancer diagnosis models should therefore balance classification performance with model parsimony and clinical applicability.

Second, medical AI models should not only generate predictions but also indicate their reliability. Conventional classifiers output benign or malignant probabilities, but these do not necessarily quantify prediction reliability. With noisy or missing features, models may still produce incorrect predictions with high confidence Kim et al. (2023). Estimating predictive uncertainty is therefore essential for improving the trustworthiness of medical AI systems Nair et al. (2020); Leibig et al. (2017).

Existing feature-selection methods are predominantly single-objective and often select the feature subset with the highest AUC or accuracy Saeys et al. (2007). Although straightforward, this may overlook complementary information in multiple candidate models with comparable performance but different feature combinations. This limitation is particularly relevant to small-sample medical datasets, where a single model may be sensitive to data partitioning and feature redundancy, leading to unstable predictions. By contrast, ensemble learning that combines multiple model outputs may improve diagnostic performance and robustness Zhou et al. (2002). Table 1 compares the proposed method with representative works along four dimensions: multi-modal, feature selection, uncertainty, and fusion. Most existing studies address only one dimension, whereas the proposed method satisfies all four.

Table 1: Comparison of existing methods and the proposed method across four dimensions.

Article	Multi-modal	Feature Selection	Uncertainty	Fusion
McKinney et al.	×	×	×	×
Rodriguez-R et al.	×	×	×	×
Akselrod-B et al.	√	×	×	×
Yuan et al.	×	√	×	×
Kursa and Rudnicki	×	√	×	×
Leibig et al.	×	×	√	×
Nair et al.	×	×	√	×
Zhou et al.	×	×	×	√
Ours	√	√	√	√

To address these challenges, this study proposes an intelligent diagnostic method for AI-assisted breast cancer diagnosis based on multi-objective optimization of feature groups and Pareto-based model fusion, jointly considering diagnostic performance, feature-group count, and predictive uncertainty. First, deep imaging features from digital breast tomosynthesis (DBT) images and structured clinical features are grouped by medical semantics. A genetic algorithm then performs a multi-objective search at the feature-group level, yielding multiple Pareto-optimal models that balance diagnostic performance and model parsimony. Finally, Monte Carlo Dropout estimates model uncertainty, and the Pareto-optimal models are dynamically combined using preference-dependent weights reflecting the user's priorities regarding accuracy, parsimony, and reliability.

The main contributions of this study are as follows:

1)We develop a feature-group-level multi-objective optimization framework that maximizes AUC while minimizing the number of feature groups. A genetic algorithm identifies Pareto-nondominated solutions, balancing diagnostic performance and model parsimony.

2)We introduce Monte Carlo Dropout–based uncertainty estimation through multiple stochastic forward passes of the Pareto submodels. Entropy is used to quantify predictive uncertainty, providing both diagnostic probabilities and an indication of prediction reliability.

3)We design a decision-maker-preference-based dynamic Pareto model-fusion mechanism that jointly considers AUC, the number of feature groups, and predictive uncertainty. Model weights are dynamically assigned to different clinical requirements, improving stability and applicability.

2 Related Work

2.1 AI-Assisted Breast Cancer Diagnosis

Deep learning has become an important approach in medical image analysis. Litjens et al. (2017) systematically review deep learning in medical image analysis and report that deep neural networks are widely applied to image classification, object detection, lesion segmentation, image registration, and prognosis prediction. Shen et al. (2017) summarize major advances in deep learning for medical image analysis, noting that deep models can automatically learn high-level semantic representations from medical images, reducing reliance on manually engineered features.

In addition, the review by Lundervold & Lundervold (2019) highlights the broad potential of deep learning in magnetic resonance imaging (MRI) and other medical imaging applications. McKinney et al. (2020) develop a deep learning system using large-scale mammographic datasets and report strong performance in breast cancer screening. Rodriguez-Ruiz et al. (2019) compare a stand-alone artificial intelligence (AI) system with radiologists and report good performance in breast cancer detection. Wu et al. (2020) further show that deep neural networks can improve radiologists' performance in breast cancer screening.

Although medical images contain rich diagnostic information, imaging data alone may not fully capture a patient's clinical condition. Breast cancer diagnosis and risk assessment are also closely associated with clinical information such as age, family history, medical history, structured features from radiology reports, and Breast Imaging Reporting and Data System (BI-RADS) assessments Yala et al. (2021). Multimodal models integrating breast imaging with clinical risk factors may yield more robust breast cancer risk prediction.

2.2 Feature Selection and Multi-Objective Optimization

Feature selection reduces redundant variables while preserving relevant information, improving generalizability and interpretability. Guyon et al. (2002) propose support vector machine recursive feature elimination (SVM-RFE), which ranks feature importance and recursively eliminates the least important features. Kursa (2014) study random-forest-based gene selection, where a random forest classifier assesses feature importance and iteratively removes less informative features. Fortino et al. (2020) introduce a genetic algorithm for genome-scale biomarker discovery, demonstrating its potential for feature selection.

Conventional methods often optimize a single criterion such as accuracy or the area under the receiver operating characteristic curve (AUC), which may overemphasize one metric while neglecting model complexity, feature-acquisition costs, and clinical interpretability. Since fewer features reduce acquisition costs and improve usability, feature selection can be formulated as a multi-objective problem that jointly optimizes model performance and the number of selected features.

Multi-objective optimization has been applied to breast cancer diagnosis, recurrence prediction, and survival analysis, seeking trade-off solutions rather than a single optimum. Deb et al. (2002) propose the non-dominated sorting genetic algorithm II (NSGA-II), which uses fast non-dominated sorting and crowding distance to preserve diversity and is a classic method in multi-objective evolutionary optimization. The strength Pareto approach by Zitzler & Thiele (1999) provides an important foundation. Zhang & Li (2007) propose the multi-objective evolutionary algorithm based on decomposition (MOEA/D), which decomposes a multi-objective problem into scalar subproblems and performs strongly on complex tasks. The proposed framework defines AUC maximization and minimizing the number of feature groups as two objectives, using a genetic algorithm to search at the

feature-group level and yield non-dominated solutions that balance diagnostic performance and feature parsimony.

2.3 Uncertainty Estimation and Trustworthy Medical AI

The clinical utility of medical AI systems depends not only on predictive accuracy but also on the reliability and interpretability of their outputs. Uncertainty estimation enables clinicians to identify cases that are difficult for the model to classify, thereby improving the safety of human–AI collaborative diagnosis Gawlikowski et al. (2023).

Monte Carlo dropout is a widely adopted approach for uncertainty estimation. During inference, dropout layers remain active and multiple stochastic forward passes yield an empirical predictive distribution, from which model uncertainty is estimated Abdar et al. (2021). Leibig et al. (2017) apply uncertainty estimation to diabetic retinopathy detection and demonstrate that it can identify cases in which the model is likely to be unreliable, whereas Nair et al. (2020) investigate uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, showing that uncertainty estimates can provide useful information for quality control in medical image analysis.

Gawlikowski et al. (2023) and Abdar et al. (2021) systematically review uncertainty quantification in deep learning and emphasize that uncertainty estimation has become an important component of trustworthy artificial intelligence. In AI-assisted breast cancer diagnosis, mammographic images are affected by factors such as breast density, indistinct lesion boundaries, and variations in image quality. For challenging cases, class probabilities may not adequately reflect the risk or reliability of a prediction. Providing uncertainty information alongside predicted benign or malignant probabilities may therefore help clinicians interpret and appropriately use the outputs of these systems.

2.4 Model Fusion and Medical Decision Support

Compared with a single model, an ensemble of multiple models can often reduce variance and produce more accurate and stable predictions. Zhou et al. (2002) demonstrate that, in neural network ensembles, selecting an appropriate subset of models may outperform the straightforward aggregation of all available models. Narod et al. (2015) reviews the applications of ensemble-based learning in decision-making systems and highlights its potential to improve the robustness of classification systems. Kittler et al. (1998) develop a theoretical framework for classifier combination, providing an important foundation for decision-level fusion.

Conventional fusion methods commonly use simple averaging, majority voting, or weighted averaging based on validation-set performance. However, these methods generally focus primarily on predictive performance and pay less attention to feature-acquisition requirements and predictive uncertainty. In AI-assisted breast cancer diagnosis, the practical value of a model may be limited if it relies on a large number of difficult-to-obtain features. Similarly, a model with high predictive uncertainty should not receive an excessively large weight in the final ensemble. Accordingly, the proposed Pareto-based model-fusion strategy jointly considers model AUC, the number of features, and predictive uncertainty, while incorporating preference parameters to regulate the trade-offs among these objectives.

3 Methodological Framework

3.1 Overall Framework

This study develops an auxiliary breast cancer diagnosis framework based on multi-objective optimization for multi-modal feature selection and Pareto-based model fusion in Figure 1. The multi-modal feature groups G1–G28 are used as inputs, and a genetic algorithm is subsequently employed to search for different combinations of feature groups. The search is formulated as a bi-objective optimization problem that maximizes diagnostic AUC while minimizing the number of required feature groups, thereby identifying approximate Pareto-nondominated solutions. For each Pareto-optimal combination, the corresponding model is retrained and its hyper-parameters are optimized to construct a pool of MLP submodels. Monte Carlo Dropout is used to perform multiple stochastic forward passes, and predictive entropy is calculated to quantify the uncertainty of each submodel.

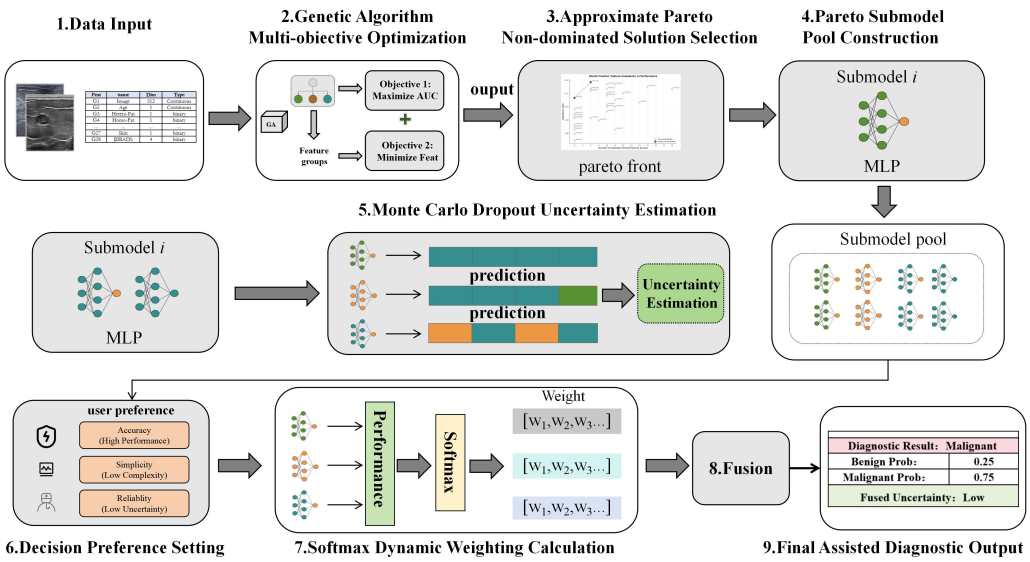


Figure 1: Overall Framework.It comprises a sequence of stages: data input, algorithmic processing, a Pareto pool, and final fusion output.

During the diagnostic stage, the preference settings are incorporated to reflect the preferences for accuracy, simplicity, and reliability. The score is calculated by jointly considering each model's AUC, the number of feature groups, and its uncertainty, after which the Softmax function is used to assign dynamic fusion weights. Finally, the benign and malignant probabilities generated by the individual models are fused at the prediction level to produce patient-level classification results and the corresponding fused uncertainty, thereby achieving a balance among diagnostic performance, model simplicity, and prediction reliability.

3.2 Notation Definitions

All notation definitions are provided in the AppendixA.1.

3.3 Multi-Objective Genetic Algorithm-Based Optimization

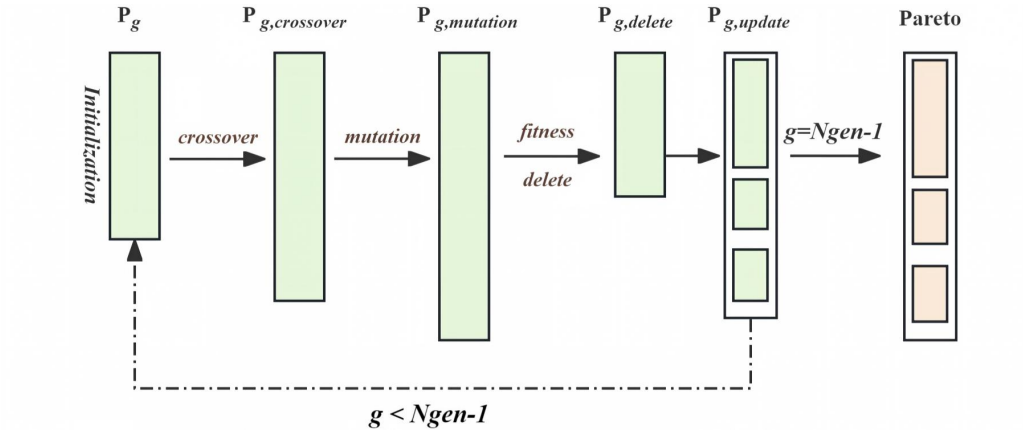


Figure 2: Flowchart of Genetic Algorithm-Based Pareto Pool Selection.

Each individual z is a binary vector of length G:

$$z = [z_1, z_2, \ldots, z_G], \tag{1}$$

where each bit indicates whether the corresponding feature group is selected. $z_1 = 1$ means the image feature group is always retained; $z_i = 1$ for $i > 1$ means the i-th group is selected. For example,

$$z = [1, 0, 1, 1, 0, \ldots, 1] \tag{2}$$

It indicates that the first, third, and fourth groups are selected, while the others are discarded. The selected groups are concatenated into the model input:

$$x = [x^{img}, x^{g_2}, x^{g_3}, \ldots]. \tag{3}$$

The genetic algorithm follows Figure 2.

1) Initialization The initial population contains N_{pop} random binary vectors.

2) Crossover Single-point crossover is applied to parent pairs. A crossover point is randomly chosen from $\{1, \ldots, G-1\}$, and segments after it are exchanged to produce two offspring. Crossover occurs with probability p_c.

3) Mutation Each non-mandatory bit is flipped independently with probability p_m: 1 becomes 0, and 0 becomes 1. The first bit is excluded because the image feature group must remain active.

4) Fitness Evaluation

$$F(x) = (1 - f_{AUC}(x),\ f_{feat}(x))^T \tag{4}$$

For each individual x, the following steps are performed: **feature extraction**, which extracts the subset defined by x; **feature concatenation**, which concatenates clinical and image features; **model training**, which trains a three-layer MLP with a fixed seed; and **performance evaluation**, which computes AUC on the validation set.

5) Algorithm Output Inferior solutions are not replaced by superior ones; all evaluated individuals from every generation are retained with their objective pairs $(1 - f_{AUC}(x), f_{feat}(x))$. After all generations, the historical candidate set is filtered by the approximate Pareto-dominance relation in the AppendixA.2, yielding the final set S_P.

4 Model Fusion and Prediction Output

4.1 Monte Carlo Uncertainty Estimation

Details are provided in the AppendixA.3.

4.2 Dynamic Preference-Based Model Fusion

A user-defined preference vector is introduced:

$$r = (r_{AUC}, r_f, r_u). \tag{5}$$

The parameter r_{AUC} specifies the preference for AUC and represents the minimum acceptable AUC, reflecting the emphasis on predictive accuracy. The parameter r_u specifies the preference for uncertainty and reflects the emphasis on diagnostic reliability. To prevent models with inadequate predictive performance or excessive uncertainty from adversely affecting fusion, two hard thresholds are established: a model is excluded if its validation AUC is lower than r_{AUC}, or if its mean validation uncertainty $\bar{u}_i$ exceeds $1.2r_u$.

The parameter r_f specifies the preferred feature dimensionality. Instead of simply minimizing the number of selected features, we introduce a target feature number and measure the absolute deviation between the candidate feature number and this target, followed by min–max normalization:

$$norm_aucs_i = \frac{AUC_i - \min_{S_j \in S_P} AUC_j}{\max_{S_j \in S_P} AUC_j - \min_{S_j \in S_P} AUC_j + \epsilon}, \tag{6}$$

$$norm_feats_i = \frac{|Feat_i - r_f| - \min_{S_j \in S_P} |Feat_j - r_f|}{\max_{S_j \in S_P} |Feat_j - r_f| - \min_{S_j \in S_P} |Feat_j - r_f| + \epsilon}, \tag{7}$$

$$norm_uncs_i = \frac{\bar{u}_i - \min_{S_j \in S_P} \bar{u}_j}{\max_{S_j \in S_P} \bar{u}_j - \min_{S_j \in S_P} \bar{u}_j + \epsilon}. \tag{8}$$

Here, $\epsilon > 0$ is a numerical stabilization term, and all three metrics use min–max normalization. The dynamic weights are then computed as

$$w_i = \alpha \cdot norm_aucs_i - \beta \cdot norm_feats_i - \gamma \cdot norm_uncs_i, \tag{9}$$

$$\bar{w}_i = \frac{\exp w_i}{\sum_{j=1}^{N_{fin}} \exp w_j}, \tag{10}$$

where α, β, and γ denote the relative importance of predictive accuracy, model simplicity, and diagnostic reliability, respectively; they are set to 1, 0.7, and 0.3 in this study.

On the test set, each submodel S_i produces diagnostic predictions for sample j:

$$u_{i,j} = -\sum_{k}^{K} p_{k,i,j} \log_2 p_{k,i,j}, \tag{11}$$

$$p'_{k,i,j} = \frac{p_{k,i,j}}{\sum_k^K p_{k,i,j} + u_{i,j}}, \tag{12}$$

$$u'_{i,j} = \frac{u_{i,j}}{\sum_k^K p_{k,i,j} + u_{i,j}}. \tag{13}$$

The weighted product of evidence for class k, the evidence conflict term, and the global uncertainty baseline are defined as

$$A_k = \prod_{i=1}^{N} \left(\bar{w}_i p'_{k,i,j} + 1 - \bar{w}_i \sum_{l=1}^{K} p'_{l,i,j} \right), \tag{14}$$

$$B = \prod_{i=1}^{N} \left(1 - \bar{w}_i \sum_{k=1}^{K} p'_{k,i,j} \right), \tag{15}$$

$$C = \prod_{i=1}^{N} \left(1 - \bar{w}_i \right). \tag{16}$$

A_k aggregates the support from all submodels for class k, with larger weights indicating more reliable models; B measures disagreement among submodels and is subtracted to reduce conflicting evidence; C is the baseline when all models are completely uncertain. The normalization factor and the fused outputs are

$$\mu = \sum_{k=1}^{K} A_k - (K-1)B - C, \tag{17}$$

$$p_{k,j} = \frac{A_k - B}{\mu}, \tag{18}$$

$$u_j = \frac{B - C}{\mu}, \tag{19}$$

where $p_{k,j}$ is the fused diagnostic probability for class k, and u_j is the fused global uncertainty; $B - C$ represents conflict-induced uncertainty, so a larger u_j indicates lower reliability.

4.3 RSC and Mean Fusion

For RSC, the submodel with the highest score is selected:

$$i^* = \arg\max s_i, \quad i \in \{1, \ldots, N_{fin}\}, \qquad p_{k,j} = p'_{k,i^*,j}, \qquad u = u'_{i^*,j}. \tag{20}$$

For mean fusion, the outputs are averaged over all selected submodels:

$$p_{k,j} = \frac{1}{N_{fin}} \sum_{i=1}^{N_{fin}} p'_{k,i,j}, \qquad u_j = \frac{1}{N_{fin}} \sum_{i=1}^{N_{fin}} u'_{i,j}. \tag{21}$$

Table 2: AUC and AUPRC comparison under different model-pool sizes.

	AUC				AUPRC			
Model	$m = 7$	$m = 15$	$m = 22$	$m = 30$	$m = 7$	$m = 15$	$m = 22$	$m = 30$
Dynamic Preference (DP)	**0.8613**	**0.8597**	**0.8532**	**0.8548**	**0.8484**	**0.8547**	**0.8461**	**0.8500**
Mean Fusion (MF)	0.8661	0.8565	0.8500	0.8516	0.8592	0.8509	0.8432	0.8484
BestSingle (BS)	0.8484	0.8484	0.8484	0.8484	0.8460	0.8460	0.8460	0.8460
Logistic Regression (LR)	0.8629	0.8629	0.8629	0.8629	0.8190	0.8190	0.8190	0.8190
Decision Tree (DT)	0.8435	0.8435	0.8435	0.8435	0.6985	0.6985	0.6985	0.6985
SVM-RBF	0.8532	0.8532	0.8532	0.8532	0.8214	0.8214	0.8214	0.8214
KNN	0.7403	0.7403	0.7403	0.7403	0.6595	0.6595	0.6595	0.6595
XGBoost (XGB)	0.8274	0.8274	0.8274	0.8274	0.7337	0.7337	0.7337	0.7337

5 EXPERIMENTS

5.1 DATA ACQUISITION AND PREPROCESSING

We obtain clinical data and DBT images from 256 breast cancer patients in TCIA. The dataset is split into training (60%), validation (20%), and test (20%) sets. The validation set is used for model selection and fusion-weight determination, and the test set is used only for final evaluation. Data preprocessing and augmentation details are provided in the AppendixA.4.

5.2 PARETO MODELS

In this study, the genetic algorithm runs for 200 iterations, and all trained models are evaluated to obtain an approximate Pareto-optimal solution set. The final Pareto front and feature-selection heatmaps are shown in Figure 7. The Pareto front reveals a trade-off between diagnostic performance and feature-group count, and all candidate models in the Pareto pool are retained for subsequent analyses. Detailed results, including the seven-stage feature-addition experiment and the genetic algorithm search trajectory, are provided in Appendix A.5 and Appendix A.6.

5.3 COMPARISON WITH BASELINES AND FUSION STRATEGIES

To evaluate the proposed dynamic preference fusion method, we compare three fusion settings: BestSingle, Mean Fusion, and Dynamic Preference. BestSingle is the best single Pareto-optimal model selected by validation performance, corresponding to model_018. We also include classical baselines: LR, DT, SVM-RBF, KNN, and XGBoost. All models are trained on the same Pareto-optimal feature subsets under a strict train–validation–test framework.

Classification performance is evaluated using AUC and AUPRC; full results with additional metrics are provided in Appendix A.8. To study the effect of model-pool size, we rank candidate models by validation dynamic scores and construct pools of size $m = 7, 15, 22, 30$, corresponding to the top 25%, 50%, 75%, and 100% of the ranked models.

Table 2 reports AUC and AUPRC under the four pool sizes. DP achieves the highest AUPRC in all four settings and consistently outperforms BS in AUC. For example, at $m = 15$, DP reaches an AUC of 0.8597 and an AUPRC of 0.8547, surpassing MF (0.8565, 0.8509) and BS (0.8484, 0.8460). LR attains a slightly higher AUC (0.8629) in some settings, but its AUPRC (0.8190) is substantially lower, indicating weaker ranking for positive cases. Overall, DP provides the best AUPRC and a more balanced sensitivity–specificity trade-off, demonstrating a clear advantage in identifying positive breast cancer cases. These results show that DP effectively exploits complementary information from multiple Pareto models and dynamically adjusts weights according to predictive performance, feature complexity, and predictive uncertainty.

5.4 EVALUATION OF UNCERTAINTY ESTIMATION

To evaluate the practical value of uncertainty estimation, we progressively remove the most uncertain test samples at 10%, 30%, and 50%, and examine the AUC of each model on the remaining

samples. If uncertainty estimation is effective, performance should improve as uncertain samples are removed. Table 3 reports the $m = 30$ setting. Removing high-uncertainty samples consistently

Table 3: Uncertainty-based sample filtering for m=30.

model_name	AUC	AUPRC	Accuracy	Sensitivity	Specificity
model_018	0.8710	0.8700	0.7826	0.7222	0.8214
dynamic_preference	**0.8631**	**0.8691**	**0.7826**	**0.7222**	**0.8214**
mean	0.8631	0.8691	0.7826	0.6667	0.8571
percentile_removed:10%,m=30					
model_018	0.9063	0.9142	0.8611	0.8125	0.9000
dynamic_preference	**0.9094**	**0.9266**	**0.8611**	**0.8125**	**0.9000**
mean	0.9094	0.9266	0.8333	0.7500	0.9000
percentile_removed:30%,m=30					
model_018	0.9273	0.9343	0.9231	0.8182	1.0000
dynamic_preference	**0.9333**	**0.9545**	**0.9231**	**0.8182**	**1.0000**
mean	0.9333	0.9545	0.9231	0.8182	1.0000

improves AUC, AUPRC, accuracy, and specificity across all models. At 30% removal, Dynamic Preference and Mean Fusion reach an AUC of 0.9333 and an AUPRC of 0.9545, outperforming model_018 (AUC 0.9273, AUPRC 0.9343). These gains suggest that high-uncertainty samples are more difficult to classify. A detailed robustness comparison across all model-pool sizes and baseline models is provided in Appendix A.10. In addition, the uncertainty-based filtering results suggest that uncertainty estimation effectively identifies difficult samples and improves performance after their removal, supporting its practical value in reliable diagnosis.The uncertainty score therefore serves as a reliability indicator that can complement conventional probability outputs in clinical decision-making.

5.5 Robustness Evaluation

To assess tolerance to data-quality degradation, we randomly mask test-set features at the feature-group level, with missing ratios of 5%, 10%, 20%, and 30%, and repeat each condition five times. We further introduce mild, moderate, and severe Gaussian noise to examine performance under combined missing-feature and noise perturbations. We take $m = 30$ as a representative setting; results for other pool sizes and baselines are provided in Appendix A.10.

Across all settings, AUC generally decreases as feature missingness and noise increase. SVM-RBF is the most stable model, whereas Dynamic Preference ranks second overall and generally outperforms or closely approaches BestSingle and Mean Fusion. Notably, under moderate and severe noise and relatively high missing ratios, Dynamic Preference maintains a stable AUC without substantial collapse, suggesting that dynamic scoring and adaptive weight allocation can mitigate the effects of low-quality models and perturbation-sensitive predictions. Mean Fusion shows similar trends but is less stable across pool sizes. BestSingle, KNN, and XGBoost are more sensitive to feature missingness and noise, while Decision Tree exhibits transient fluctuations under mild or moderate noise. Although Dynamic Preference does not outperform SVM-RBF, it consistently ranks second and demonstrates good tolerance to missing clinical features and combined noise perturbations.

6 Conclusion

The experimental results demonstrate the effectiveness of the proposed dynamic preference fusion method from three perspectives: classification performance, uncertainty estimation, and robustness. Dynamic Preference integrates complementary information from multiple Pareto-optimal models, achieves higher AUC and AUPRC than BestSingle, and maintains a more balanced sensitivity–specificity trade-off than individual baselines. Removing high-uncertainty samples consistently improves performance, confirming that uncertainty scores identify difficult cases. Robustness evaluation shows that Dynamic Preference tolerates feature missingness and input noise, ranks second overall, and generally outperforms BestSingle and Mean Fusion.

AI USE STATEMENT

In this work, we used generative AI tools for three purposes: (1) generating initial code for data preprocessing, model training, and evaluation, which was subsequently modified, tested, and verified by the authors; (2) polishing the language and improving the readability of the manuscript; and (3) assisting in searching for and retrieving part of the related literature.

For AI-generated code, the authors reviewed, modified, and tested all generated code, and verified its correctness by reproducing the experimental results reported in this paper. For language polishing, the authors reviewed and edited all AI-assisted text to ensure that the scientific content, claims, and interpretations remained accurate. For literature search, the authors manually verified the existence and relevance of all cited references and read the corresponding papers before citing them. We did not use generative AI tools to generate research ideas, design the methodology, fabricate data, or produce experimental results. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses the publicly available Breast Cancer dataset from The Cancer Imaging Archive (TCIA). All patient data were de-identified by the dataset providers before release, and no personally identifiable information was accessed or used in this work. The study does not involve new human-subject data collection, and no additional ethical approval was required. The proposed method is intended to assist, rather than replace, clinical decision-making, and all diagnostic outputs should be interpreted by qualified clinicians. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the dataset, data preprocessing, model architecture, and training procedure in the main text and the appendix. The data source, dataset split, missing-data handling, feature encoding, image feature extraction, and data augmentation strategy are detailed in Appendix A.4. The feature groups used in this study are listed in Table 4. The genetic algorithm settings, including population size, crossover probability, mutation probability, and the number of generations, are described in Section 3.3 and Appendix A.6. The Monte Carlo dropout procedure, the entropy-based uncertainty definition, and the hyper-parameter settings of the MLP models are provided in Appendix A.3 and Table 6. The complete experimental results, including the performance comparison across model-pool sizes and the robustness evaluation, are reported in Appendix A.8, Appendix A.9, and Appendix A.10. The source code will be released in an anonymized repository upon acceptance.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

A. Akselrod-Ballin, M. Chorev, Y. Shoshan, A. Spiro, A. Hazan, R. Melamed, E. Barkan, E. Herzel, S. Naor, E. Karavani, G. Koren, Y. Goldschmidt, V. Shalev, M. Rosen-Zvi, and M. Guindy. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, 292(2):331–342, 2019.

Deema Alzamil, Bader Alkhamees, and Mohammad Mehedi Hassan. A systematic review of multimodal fusion and explainable AI applications in breast cancer diagnosis. *CMES - Computer Modeling in Engineering and Sciences*, 145(3):2971–3027, 2025.

Nikhilanand Arya and Sriparna Saha. Multi-modal advanced deep learning architectures for breast cancer survival prediction. *Knowledge-Based Systems*, 221:106965, 2021.

D. A. Berry, K. A. Cronin, S. K. Plevritis, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. *The New England Journal of Medicine*, 353(17):1784–1792, 2005.

Xi Chen, Jiahuan Lv, Zeyu Wang, Genggeng Qin, and Zhiguo Zhou. Deep-automo: Deep automated multiobjective neural network for trustworthy lesion malignancy diagnosis in the early stage via digital breast tomosynthesis. *Computers in Biology and Medicine*, 183:109299, 2024.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

M. Desai and B. Desai. Evolution and integration of artificial intelligence across the cancer continuum in women: advances in risk assessment, prevention, and early detection. *Cancer Causes and Control*, 36:1343–1360, 2025.

V. Fortino, G. Scala, and D. Greco. Feature set optimization in biomarker discovery from genome-scale data. *Bioinformatics*, 36(11):3393–3400, 2020.

J. Gawlikowski, C. R. N. Tassi, M. Ali, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56:1513–1589, 2023.

I. Guyon, J. Weston, S. Barnhill, et al. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

E. Hamdy, O. Badawy, and M. Zaghloul. Densely convolutional networks for breast cancer classification with multi-modal image fusion. *The International Arab Journal of Information Technology*, 19(3A):463–469, 2022.

Jihyo Kim, Jiin Koo, and Sangheum Hwang. A unified benchmark for the unknown detection capability of deep neural networks. *Expert Systems with Applications*, 229:120461, 2023.

J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

M. B. Kursa. Robustness of random forest-based gene selection methods. *BMC Bioinformatics*, 15: 8, 2014.

M. B. Kursa and W. R. Rudnicki. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

C. Leibig, V. Allken, M. S. Ayhan, et al. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7:17816, 2017.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

S. M. McKinney, M. Sieniek, V. Godbole, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577:89–94, 2020.

Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, 2020.

S. A. Narod, J. Iqbal, V. Giannakeas, V. Sopik, and P. Sun. Breast cancer mortality after a diagnosis of ductal carcinoma in situ. *JAMA Oncology*, 1(7):888–896, 2015.

Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H. Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, Matthew G. Wallis, Ingvar Andersson, Sophia Zackrisson, Ritse M. Mann, and Ioannis Sechopoulos. Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute*, 111(9):916–922, 2019.

Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

N. Wu et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194, 2020.

Z. Xue et al. Research progress on epidemiological trend and risk factors of female breast cancer. *Tumor Preventive Treatment Research*, 48(1):87–92, 2021.

Adam Yala et al. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13:eaba4373, 2021.

Z. Yuan, Y. Yan, M. Sonka, and T. Yang. Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 512–522, 2021.

Q. Zhang and H. Li. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on Evolutionary Computation*, 11(6):712–731, 2007.

Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1–2):239–263, 2002.

E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

A Appendix

A.1 Notation definitions

N_{gen} denotes the total number of generations of the genetic algorithm, and N_{pop} denotes the population size. The parameters p_c and p_m represent the crossover probability and bitwise mutation probability, respectively. $F(x)$ denotes the overall fitness value of individual x, which simultaneously evaluates the diagnostic performance and feature complexity of the corresponding model. $f_{AUC}(x)$ denotes the AUC of the model trained using the feature-group combination represented by individual x on the validation set, whereas $f_{feat}(x)$ denotes the number of features selected or included by individual x. $\Omega = \{0, 1\}$ denotes the set of classes in the binary classification task, where 0 and 1 represent the two diagnostic categories, respectively. D_{tr} denotes the training dataset used to train the models, and D_{val} denotes the validation dataset used for model validation, performance evaluation, and hyperparameter selection. (x_n, y_n) denotes the n-th sample and its corresponding ground-truth label, where $y_n \in \{0, 1\}$. $p_{k,i}$ denotes the original predicted probability assigned by submodel S_i to the input sample belonging to class k. $S_{\boldsymbol{P}}$ denotes the set of approximate Pareto-optimal solutions obtained through multi-objective genetic algorithm-based selection. $p_{k,i}^t$ denotes the predicted probability assigned to class k by submodel S_i during the t-th stochastic forward pass using Monte Carlo Dropout. $\bar{p}_{k,i}$ denotes the mean predicted probability assigned to class k by submodel S_i across multiple Monte Carlo Dropout inferences. AUC_i denotes the performance of submodel S_i on the validation set, and $Feat_i$ denotes the number of features used by submodel S_i. u_i denotes the sample-level uncertainty estimated from the prediction of submodel S_i for the sample, whereas $\bar{u}_i$ denotes the mean uncertainty of submodel S_i across the sample set. The quantities $(norm_aucs_i, norm_feats_i, norm_uncs_i)$ denote the normalized AUC, feature-count, and uncertainty metrics of submodel S_i, respectively. α, β, and γ denote the weighting coefficients assigned to diagnostic performance, model complexity, and predictive uncertainty, respectively. w_i denotes the overall score obtained using the Softmax function, and $\bar{w}_i$ denotes the resulting dynamic fusion weight assigned to submodel S_i. $p_{k,j}$ denotes the final predicted probability that the input sample belongs to class k after the weighted fusion of the predictions from all submodels. $u_{i,j}$ denotes the final uncertainty output obtained by fusing the predictive results of multiple submodels. μ denotes the adjustment factor used for metric normalization, and ϵ denotes a small numerical stability constant introduced to prevent division by zero; in this study, ϵ was set to 10^{-12}.

A.2 Approximate Pareto-Dominance Relation

For two candidate solutions x_i and x_j, conventional Pareto dominance is defined as

$$\begin{aligned} x_i \prec x_j \iff & (f_{AUC}(x_i) \geq f_{AUC}(x_j)) \\ & \wedge (f_{feat}(x_i) \leq f_{feat}(x_j)) \\ & \wedge [(f_{AUC}(x_i) > f_{AUC}(x_j)) \vee (f_{feat}(x_i) < f_{feat}(x_j))] . \end{aligned} \tag{22}$$

In practice, very small differences in AUC may arise from computational variation, random initialization, or sampling variability and may not represent meaningful improvements in discriminative performance. Therefore, an AUC tolerance threshold ε_{AUC} ($\varepsilon_{AUC} > 0$) is introduced to define an approximate Pareto-dominance relation:

$$\begin{aligned} x_i \prec_\varepsilon x_j \iff & (f_{AUC}(x_i) \geq f_{AUC}(x_j)) \\ & \wedge (f_{feat}(x_i) \leq f_{feat}(x_j)) \\ & \wedge [(f_{AUC}(x_i) - f_{AUC}(x_j) > \varepsilon_{AUC}) \vee (f_{feat}(x_i) < f_{feat}(x_j))] . \end{aligned} \tag{23}$$

According to this definition, x_i approximately dominates x_j only if:

1. the AUC of x_i is not lower than that of x_j, and the number of features selected by x_i is not greater than that selected by x_j; and
2. at least one of the following conditions is satisfied:
   - the AUC improvement exceeds ε_{AUC}; or
   - x_i uses strictly fewer features.

If neither solution approximately dominates the other, they are regarded as mutually non-dominated:

$$x_i \parallel_\varepsilon x_j \iff (x_j \not\prec_\varepsilon x_i) \wedge (x_i \not\prec_\varepsilon x_j) . \tag{24}$$

In particular, if

$$\begin{aligned} & f_{AUC}(x_i) \geq f_{AUC}(x_j), \quad f_{AUC}(x_i) - f_{AUC}(x_j) \leq \varepsilon_{AUC}, \\ & f_{feat}(x_i) \geq f_{feat}(x_j), \end{aligned} \tag{25}$$

the two solutions may be retained simultaneously because the marginal AUC improvement of x_i does not exceed the predefined tolerance and is achieved without a reduction in feature complexity.

Given the complete candidate set PS_ε, the retained approximate non-dominated set is defined as

$$PS_\varepsilon = \{x \in S \mid \nexists y \in S,\ y \prec_\varepsilon x\} . \tag{26}$$

After all generations have been completed, the global approximate Pareto-optimal solution set S_P is extracted from all evaluated candidate solutions.

A.3 Monte Carlo uncertainty estimation

Each submodel S_i in the Pareto pool, where $i \in \{1, 2, \ldots, N_{fin}\}$, corresponds to the i-th Pareto-optimal feature subset. Before model fusion, an uncertainty estimation procedure is performed for each submodel to assess the uncertainty of its predictions.

1. **Submodel loading:** the multilayer perceptron (MLP) architecture and model weights corresponding to submodel S_i are loaded to ensure that the model parameters remain consistent with their state after training.
2. **Activating the Dropout layers:** the Dropout layers in the model are activated to introduce stochasticity during inference.
3. **Input processing:** the sample to be predicted is selected and input into submodel S_i.
4. **Multiple independent forward passes:** for the input sample j, T stochastic forward passes are performed with Dropout enabled. Each pass uses a newly sampled Dropout mask, resulting in T predictive probability vectors.

5. **Statistical feature calculation:** based on the sequence of T prediction results, the mean predictive probability $\bar{p}_{k,i}$ for sample j is calculated to represent the model's primary prediction. Predictive uncertainty is then estimated based on the resulting diagnostic probabilities.

For patient j, the probability assigned by submodel S_i to class k during the t-th stochastic forward pass is denoted by

$$\bar{p}_{k,i} = \frac{1}{T} \sum_{t=1}^{T} p_{k,i}^{t}. \tag{27}$$

For a binary classification task, $\Omega = \{0, 1\}$ and $K = 2$.

The sample-level predictive entropy of submodel S_i for sample j is defined as

$$u_i = -\sum_{k=1}^{K} \bar{p}_{k,i} \log_2 \bar{p}_{k,i}. \tag{28}$$

The training and validation datasets are defined as

$$D_{tr} = \{(x_n, y_n)\}_{n=1}^{N_{tr}}, \quad D_{val} = \{(x_n, y_n)\}_{n=1}^{N_{val}}, \tag{29}$$

where x_n denotes the input feature vector and $y_n \in \{0, 1\}$ denotes the ground-truth label.

The model-level mean uncertainty of submodel S_i on the validation set is calculated as

$$\bar{u}_i = \frac{1}{|D_{val}|} \sum_{(x_n, y_n) \in D_{val}} u_i. \tag{30}$$

The validation-set mean uncertainty $\bar{u}_i$ is used for candidate filtering and weight calculation, whereas $u_{i,j}$ is used to quantify the uncertainty associated with the prediction of an individual sample.

A.4 Data Preprocessing

Four patients had extensive missing data. Therefore, no imputation was performed for these cases. The clinical data were encoded, while the image data were transformed into 512-dimensional vector representations using a Transformer. For the training set, data augmentation is performed, with continuous feature interpolation and binary feature inheritance applied to continuous and binary features, respectively. Figure3 shows that the input data is divided into two parts: clinical data and DBT

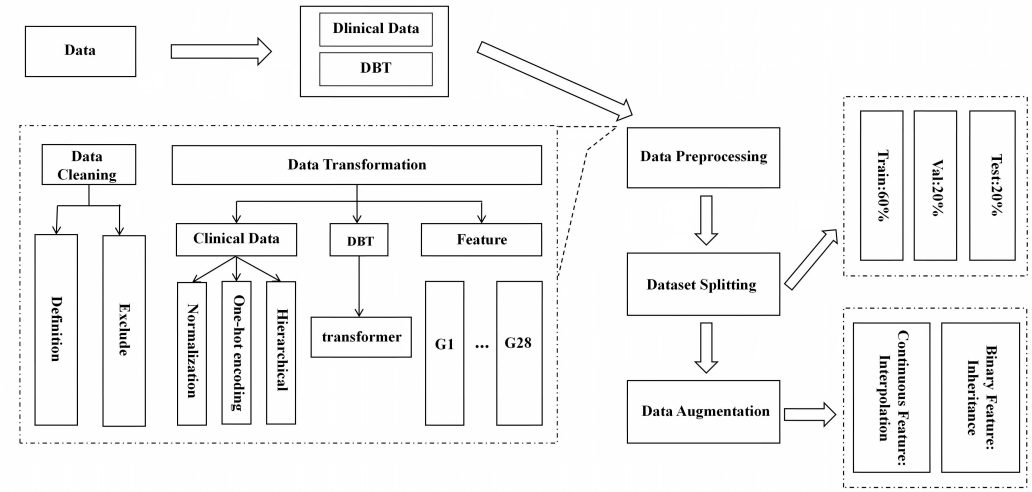


Figure 3: Data processing pipeline.

images. The first stage is data processing. First, the data is defined and excluded;. Then, the clinical

Table 4: Feature groups used in this study.

Feat	Name	Dimension	Type
G1	Image	512	Continuous
G2	Age	1	Continuous
G3	Hetero-Fat	1	binary
G4	Homo-Fat	1	binary
G5	Hetero-FG	1	binary
G6	Homo-FG	1	binary
G7	Lactating	1	binary
G8	Signs	8	binary
G9	Injury	2	binary
G10	Family history	2	binary
G11	Personal history	2	binary
G12	HRT	2	binary
G13	Nipple discharge	2	binary
G14	Shape	3	binary
G15	Margin	4	binary
G16	Anechoic	1	binary
G17	Solid	1	binary
G18	Heterogeneous	1	binary
G19	Hyperechoic	1	binary
G20	Hypoechoic	1	binary
G21	Isoechoic	1	binary
G22	Combined	1	binary
G23	Enhancement	1	binary
G24	Shadowing	1	binary
G25	Halo	1	binary
G26	Calcifications	3	binary
G27	Skin_thickening	1	binary
G28	BIRADS	4	binary
/	Classification	1	binary

data is processed sequentially through normalization, one-hot encoding, and hierarchical encoding, while the DBT images are processed by a transformer, ultimately forming feature groups G1 to G28 in Table4. In the second stage, the transformed data is divided. Finally, data augmentation is performed, which includes continuous feature interpolation and binary feature inheritance. Classification is a binary classification label representing the final diagnostic outcome and is therefore excluded from feature-group selection. A two-dimensional hierarchical encoding was additionally introduced for the Injury, Family history, Personal history, HRT, Nipple discharge, and Signs feature groups:

$$k \in \{0, 1\}. \tag{31}$$

In this encoding scheme, $k = 1$ indicates that the corresponding field has an explicit record, whereas $k = 0$ indicates that the field is unknown or missing.

$$x' = [k, x]. \tag{32}$$

Data Augmentation Scheme Continuous features were augmented using an interpolation-based strategy:

$$x_{new} = \lambda x_i + (1 - \lambda)x_j, \quad \lambda \in (0, 1). \tag{33}$$

For a binary feature denoted by b, the corresponding generation rule is defined as:

$$b_{new} \in \{b_i, b_j\}. \tag{34}$$

Let N_c denote the number of samples belonging to class c in the training set, and let δ denote the augmentation factor. The number of newly generated samples for class c is calculated as follows:

$$N_c^{new} = \delta N_c. \tag{35}$$

In the experiments conducted in this study, δ is set to 1. Accordingly, the number of newly generated samples for each class was equal to the number of samples from that class in the original training set, resulting in an augmented training set twice the size of the original training set.

A.5 Feature Selection Experiment

To evaluate the contributions of features from different modalities and feature dimensions to the predictive performance of the model, this study designs a seven-stage Sequential Feature Addition Experiment based on a multi-layer perceptron (MLP). First, a 512-dimensional image-feature representation is used to construct an image-only baseline model. Subsequently, clinical and ultrasound features are progressively incorporated according to the clinical diagnostic workflow, resulting in seven cumulative input settings. Table 5 summarizes the experimental results.Step 1:Including only image features.

Step 2:Adding Age,heterogeneous: predominantly fat,homogeneous: fat,heterogeneous: predominantly fibro glandular,homogeneous: fibro glandular, and lactating to the features used in Step 1.

Step 3:Further incorporating Signs,breast injury,family history,personal history,HRT, and nipple discharge.

Step 4:Adding morphology-related features, including Shape and Margin.

Step 5:Adding echo-pattern features, including Hyperechoic, Hypoechoic, Isoechoic, anechoic,complex cystic/solid, and heterogeneous.

Step 6:Adding associated ultrasound findings, including Enhancement, Shadowing, Halo, Combined, Calcifications, and Skin_thickening.

Step 7:Finally,BIRADS was incorporated.

Through this progressive feature-accumulation strategy, the study systematically evaluated the incremental contributions of different feature groups beyond the image features.

Table 5: MLP performance with sequential feature addition.

Step	AUC	ACC	Sen.	Spe.
1	0.7377	0.7020	0.6500	0.7355
2	0.7410	0.7333	0.7000	0.7548
3	0.7910	0.7294	0.7100	0.7419
4	0.8394	0.7843	0.7800	0.7871
5	0.8297	0.7882	0.7500	0.8120
6	0.8290	0.7843	0.7900	0.7800
7	0.8458	0.7686	0.7600	0.7742

With the progressive incorporation of structured clinical and ultrasound features into the image-only baseline, the model demonstrates an overall improvement in classification performance, suggesting that multi-modal information fusion can enhance diagnostic discrimination. Nevertheless, model performance does not improve monotonically as the number of input features increases. Specifically, although additional features are introduced from Step 4 to Steps 5 and 6, the AUC decreases slightly. The inclusion of BIRADS in Step 7 partially restores the AUC; however, neither sensitivity nor specificity improves further. These findings indicate that increasing the dimensionality of the input feature set does not necessarily lead to superior predictive performance. Certain newly incorporated features may provide limited additional discriminative information or introduce redundancy and noise, thereby compromising the model's generalization capability. Consequently, optimizing the feature subset—rather than indiscriminately expanding the feature space—is essential for improving model performance.

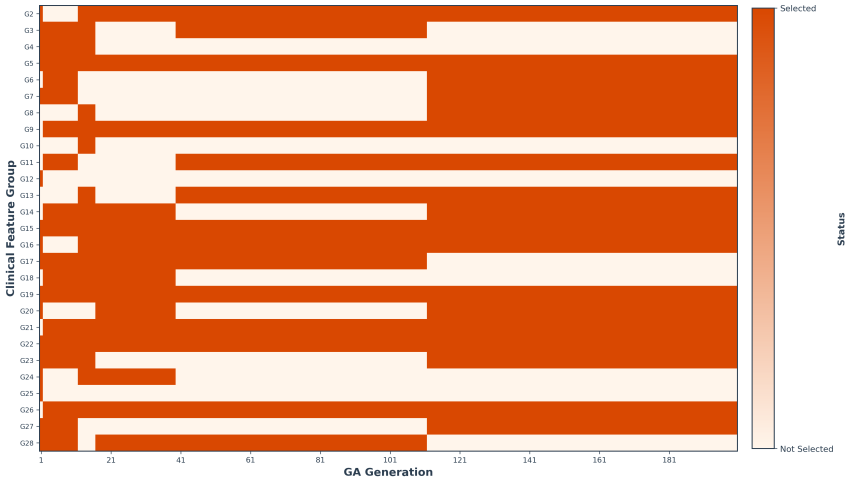


Figure 4: Heatmap of feature-selection patterns across genetic algorithm iterations.

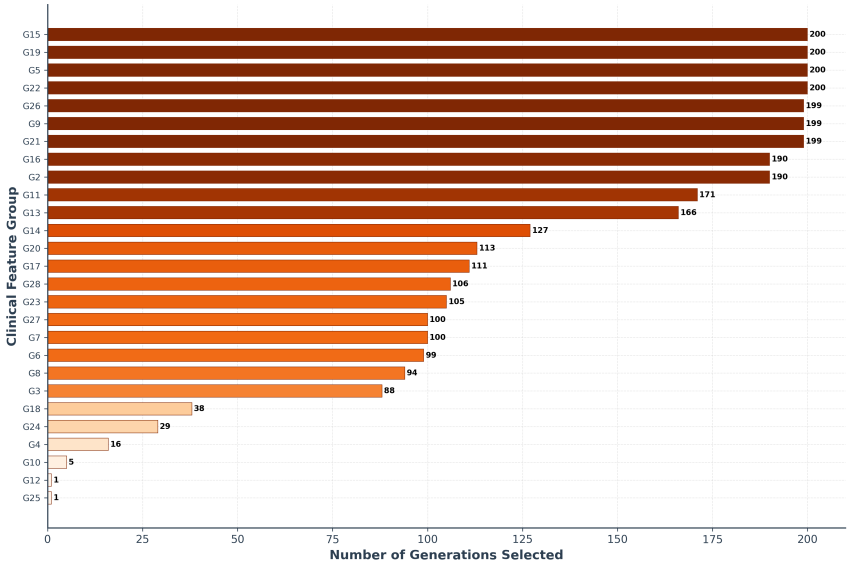


Figure 5: Feature-selection frequency statistics.

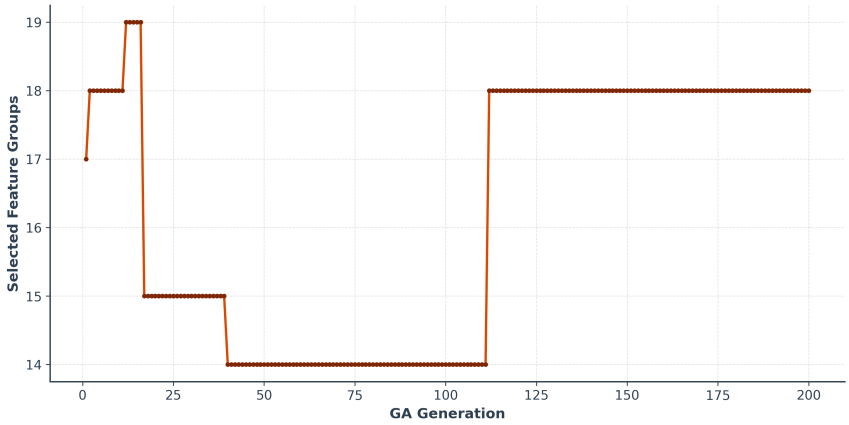
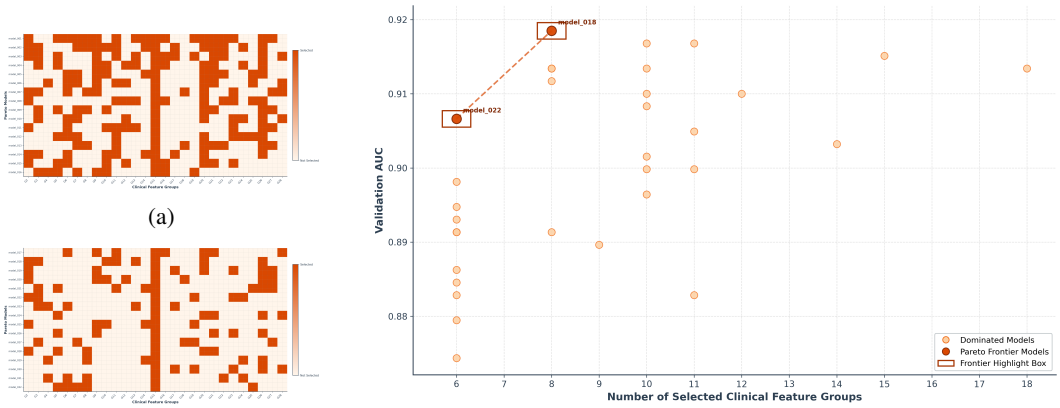


Figure 6: Convergence of the Number of Selected Features During Genetic Algorithm Optimization.

A.6 Trajectory of the Genetic Algorithm

This set of figures illustrates the search trajectory of the genetic algorithm during feature-group selection.As the evolutionary process progressed, distinct differences gradually emerged in the selection frequencies of the feature groups.In particular, feature groups such as Margin were consistently selected across nearly all generations, forming a relatively stable set of core candidate features. This finding suggests that these features may have substantial discriminative value and therefore warrant further investigation and validation in subsequent clinical studies.By contrast, a small number of feature groups were rarely selected across generations, suggesting that they may provide limited incremental predictive value in the current dataset. Subject to further external validation, this finding may help inform the reduction of low-yield feature acquisition and the associated examination burden.The number of feature groups included in the best-performing individual of each generation initially fluctuated, subsequently decreased, and eventually stabilized. This pattern indicates that the genetic algorithm did not merely minimize the number of selected features but instead sought a balance between predictive performance and model complexity.Overall, the feature combinations evolved from an unstable state during the early generations toward convergence, ultimately exhibiting a relatively stable selection pattern.These results demonstrate that the genetic algorithm can retain consistently selected candidate feature groups while reducing potentially redundant information, thereby providing a basis for subsequent model development and the identification of Pareto-optimal candidate solutions.

A.7 Pareto Models



(a) (b) (c)

Figure 7: (a, b) Feature-selection heatmaps of the Pareto solutions. (c) Pareto front.

The Pareto front shows the distribution of candidate models according to the number of clinical feature groups and validation AUC. Dark nodes and the dashed line denote Pareto-front models. Performance does not improve continuously with more feature groups, indicating that additional clinical features do not necessarily yield better diagnostic performance. Model_022 and Model_018 form the Pareto front, representing the trade-off between lower feature-acquisition cost and higher diagnostic performance, respectively. This study does not use only the Pareto-front models; all models in the Pareto candidate pool are included in subsequent analyses to compare different feature combinations, model complexities, and predictive performances. The genetic algorithm identified 32 Pareto candidate submodels that satisfied the predefined selection criteria. To further improve predictive performance on each fixed Pareto-optimal feature subset, the multi-layer perceptron (MLP) models were subjected to hyper-parameter optimization and retraining based on the Pareto candidate pool generated by the genetic algorithm. Specifically, hyper-parameter optimization focused primarily on parameters related to network architecture and regularization, including the hidden-layer dimensions, dropout rate, learning rate, and weight-decay coefficient. For each hyper-parameter configuration, the corresponding MLP models were independently retrained from random initialization using the same collection of Pareto-optimal feature subsets, and their predictive performance was evaluated on the validation set in terms of the area under the receiver operating characteris-

Table 6: Parameter Settings for the MLP Model.

parameter	value
learning_rate	[5e-4, 1e-3, 2e-3]
batch_size	[16, 32]
dropout	[0.1, 0.2]
weight_decay	[1e-5, 1e-4]
hidden_dims	{[[32, 16], [64, 16], [64, 32], [128, 64]}
epochs	[30]
patience	[8]

tic curve (AUC). The aggregate performance of the Pareto model pool was subsequently compared across hyper-parameter configurations to determine the optimal configuration.

A.8 Complete Experimental Results of Model Performance Comparison

To comprehensively evaluate the proposed dynamic preference fusion method, the comparison methods are categorized into three groups: the best single Pareto-optimal model (BestSingle), ensemble fusion methods, and classical machine-learning baseline models. BestSingle denotes the best individual model selected from the Pareto-optimal model set according to its performance on the validation set; in this study, BestSingle corresponds to model_018. The ensemble methods include two fusion strategies: Dynamic Preference and Mean Fusion. The classical machine-learning baselines include Logistic Regression (LR), Decision Tree (DT), support vector machine with a radial basis function kernel (SVM-RBF), K-Nearest Neighbors (KNN), and XGBoost.

To ensure a fair comparison, all baseline models are trained using the Pareto-optimal feature subsets identified by the genetic algorithm. First, candidate models are trained on the training set using different feature subsets and hyperparameter configurations. Subsequently, the optimal feature subset and hyperparameter configuration for each type of baseline model are selected according to the AUC achieved on the validation set. Finally, the optimal configuration identified on the validation set is applied to the independent test set to obtain the final performance. The entire modeling procedure follows a strict train–validation–test framework.

The classification performance of each method is evaluated using AUC, AUPRC, accuracy, sensitivity, and specificity. To further investigate the effect of model-pool size on ensemble performance, the candidate models are ranked according to their dynamic scores on the validation set. Model pools of different sizes are then constructed by retaining the top 25%, 50%, 75%, and 100% of the ranked candidate models. Following the predefined hard-selection procedure, 30 candidate models are included in the fusion stage. Accordingly, the number of models participating in the ensemble is set to $m = 7$, $m = 15$, $m = 22$, and $m = 30$, corresponding approximately to the top 25%, 50%, 75%, and 100% of the candidate models, respectively.

Table 7 reports the performance comparison when $m = 7$. When $m = 7$, Mean Fusion achieves an AUC of 0.8661 and an AUPRC of 0.8592, exceeding the corresponding values of 0.8484 and 0.8460 obtained by BestSingle. This finding indicates that integrating even a small number of high-quality candidate models can exploit their complementary information and further improve overall predictive performance. Dynamic Preference achieves an AUC of 0.8613 and an AUPRC of 0.8484. Although these values are slightly lower than those of Mean Fusion, the model still demonstrates strong discriminative ability. Compared with BestSingle, Dynamic Preference improves the AUC from 0.8484 to 0.8613, the accuracy from 0.7647 to 0.7843, and the sensitivity from 0.7000 to 0.8000. Although the specificity decreases from 0.8065 to 0.7742, Dynamic Preference substantially improves the detection of positive cases and achieves a more balanced trade-off between sensitivity and specificity. Compared with the machine-learning baseline models, Dynamic Preference achieves an AUC of 0.8613, which is close to that of Logistic Regression (0.8629) and higher than those of Decision Tree, SVM-RBF, KNN, and XGBoost. Its AUPRC of 0.8484 is higher than that of all baseline models, indicating a clear advantage in ranking positive cases.

Table 7: Performance comparison of various models when $m = 7$.

Model	AUC	AUPRC	Accuracy	Sensitivity	Specificity
Dynamic Preference	0.8613	0.8484	0.7843	0.8000	0.7742
Mean Fusion	0.8661	0.8592	0.7843	0.8000	0.7742
BestSingle	0.8484	0.8460	0.7647	0.7000	0.8065
Logistic Regression	0.8629	0.8190	0.7647	0.8500	0.7097
Decision Tree	0.8435	0.6985	0.7255	0.9000	0.6129
SVM-RBF	0.8532	0.8214	0.8431	0.7500	0.9032
KNN	0.7403	0.6595	0.7059	0.7500	0.6774
XGBoost	0.8274	0.7337	0.8039	0.9000	0.7419

Table 8 reports the performance comparison when $m = 15$. When $m = 15$, the top 50% of candidate models ranked by dynamic score are selected for ensemble fusion. Dynamic Preference achieves an AUC of 0.8597, outperforming both Mean Fusion (0.8565) and BestSingle (0.8484). This result suggests that, as the model pool expands, the dynamic weighting mechanism can effectively distinguish the relative contributions of different candidate models and mitigate the adverse influence of weaker models on the ensemble outcome. Dynamic Preference achieves an AUPRC of 0.8547, the highest among all evaluated models. This finding demonstrates its clear advantage in identifying positive samples and ranking positive cases, making it particularly suitable for breast cancer computer-aided diagnosis, where the detection of malignant cases is of primary concern. Dynamic Preference achieves an accuracy of 0.7647, a sensitivity of 0.7500, and a specificity of 0.7742. Compared with BestSingle, Dynamic Preference maintains the same accuracy while increasing the sensitivity from 0.7000 to 0.7500; however, its specificity decreases from 0.8065 to 0.7742. These results indicate that Dynamic Preference enhances the detection of positive cases while achieving a relatively balanced trade-off between the identification of positive and negative cases. Among the baseline models, Logistic Regression achieves the highest AUC for an individual model (0.8629), slightly exceeding the AUC of Dynamic Preference (0.8597). However, its AUPRC is only 0.8190, which is substantially lower than that of Dynamic Preference. SVM-RBF achieves the highest accuracy (0.8431) and specificity (0.9032); nevertheless, its AUPRC is 0.8214, also lower than that of Dynamic Preference. Decision Tree and XGBoost both achieve a sensitivity of 0.9000, but their specificities are 0.6129 and 0.7419, respectively. This indicates that their higher positive-case detection rates are accompanied by a greater number of false-positive predictions. In contrast, Dynamic Preference demonstrates greater stability in terms of AUC, AUPRC, and the overall balance between classification performance for the two classes.

Table 8: Performance comparison of various models when $m = 15$.

Model	AUC	AUPRC	Accuracy	Sensitivity	Specificity
Dynamic Preference	0.8597	0.8547	0.7647	0.7500	0.7742
Mean Fusion	0.8565	0.8509	0.7647	0.7500	0.7742
BestSingle	0.8484	0.8460	0.7647	0.7000	0.8065
Logistic Regression	0.8629	0.8190	0.7647	0.8500	0.7097
Decision Tree	0.8435	0.6985	0.7255	0.9000	0.6129
SVM-RBF	0.8532	0.8214	0.8431	0.7500	0.9032
KNN	0.7403	0.6595	0.7059	0.7500	0.6774
XGBoost	0.8274	0.7337	0.8039	0.9000	0.7419

Table 9 reports the performance comparison when $m = 22$. When $m = 22$, the overall performance is broadly consistent with that observed when $m = 15$. The AUC of Dynamic Preference remains higher than that of Mean Fusion. In addition, Dynamic Preference continues to achieve the highest AUPRC among all evaluated models.

Table 9: Performance comparison of various models when $m = 22$.

Model	AUC	AUPRC	Accuracy	Sensitivity	Specificity
Dynamic Preference	0.8532	0.8461	0.7647	0.7500	0.7742
Mean Fusion	0.8500	0.8432	0.7647	0.7500	0.7742
BestSingle	0.8484	0.8460	0.7647	0.7000	0.8065
Logistic Regression	0.8629	0.8190	0.7647	0.8500	0.7097
Decision Tree	0.8435	0.6985	0.7255	0.9000	0.6129
SVM-RBF	0.8532	0.8214	0.8431	0.7500	0.9032
KNN	0.7403	0.6595	0.7059	0.7500	0.6774
XGBoost	0.8274	0.7337	0.8039	0.9000	0.7419

Table 10 reports the performance comparison when $m = 30$. When $m = 30$, meaning that all candidate models are included in the ensemble, Dynamic Preference again achieves the highest AUPRC among all evaluated models. Under the full model-pool setting, Dynamic Preference exhibits a more pronounced advantage over Mean Fusion. This finding suggests that, as the number of models increases and the performance heterogeneity among models becomes more pronounced, Dynamic Preference can maintain favorable predictive performance and stability through dynamic scoring and adaptive weight allocation.

Table 10: Performance comparison of various models when $m = 30$.

Model	AUC	AUPRC	Accuracy	Sensitivity	Specificity
Dynamic Preference	0.8548	0.8500	0.7647	0.7500	0.7742
Mean Fusion	0.8516	0.8484	0.7451	0.7000	0.7742
BestSingle	0.8484	0.8460	0.7647	0.7000	0.8065
Logistic Regression	0.8629	0.8190	0.7647	0.8500	0.7097
Decision Tree	0.8435	0.6985	0.7255	0.9000	0.6129
SVM-RBF	0.8532	0.8214	0.8431	0.7500	0.9032
KNN	0.7403	0.6595	0.7059	0.7500	0.6774
XGBoost	0.8274	0.7337	0.8039	0.9000	0.7419

The results of Dynamic Preference under different model-pool sizes can be summarized as follows. In terms of AUC, Dynamic Preference achieves its highest value of 0.8613 when $m = 7$. This result indicates that a compact model pool composed of a small number of highly rated models can provide strong overall discriminative ability. As the number of models increases from 7 to 22, the AUC of Dynamic Preference generally decreases. This decline suggests that continuously expanding the model pool may introduce weaker models, redundant information, or highly correlated predictions, thereby diluting the contributions of high-quality models. When the number of models is further increased to 30, the AUC increases slightly from 0.8532 to 0.8548, indicating that the dynamic weighting mechanism can partially mitigate the adverse effects associated with full-model fusion. In terms of AUPRC, Dynamic Preference achieves its highest value of 0.8547 when $m = 15$, followed by 0.8500 when $m = 30$, 0.8484 when $m = 7$, and 0.8461 when $m = 22$. Therefore, when the primary focus is the ranking performance of positive cases, the medium-sized model pool containing 15 models provides the best performance.

Taken together, the experimental results obtained with the four model-pool sizes show that model fusion generally outperforms the single best Pareto model. Combining multiple candidate models with diverse and complementary predictive characteristics can improve overall discriminative ability and positive-case ranking performance under most settings. However, comparisons across different model-pool sizes also demonstrate that increasing the number of models included in the ensemble does not necessarily lead to continuous performance gains. The dynamic scoring and weight allocation mechanisms can better address inconsistencies in model quality and the increased uncertainty associated with model predictions. In particular, under the full-model fusion setting, Dynamic Preference improves the AUC, AUPRC, accuracy, and sensitivity while maintaining the same specificity, demonstrating good robustness against interference from less informative models.

Compared with the classical machine-learning baseline models, Logistic Regression achieves the highest baseline AUC, SVM-RBF achieves the highest accuracy and specificity, and Decision Tree and XGBoost achieve the highest sensitivity. These results indicate that the individual models have distinct strengths across different evaluation metrics, but also exhibit relatively pronounced performance biases. Although Dynamic Preference does not achieve the highest value for every individual metric, it consistently maintains relatively high AUC and AUPRC values across different model-pool sizes and achieves a comparatively balanced trade-off between sensitivity and specificity. Notably, the AUPRC of Dynamic Preference is higher than that of all machine-learning baseline models in all four experiments, indicating a stable advantage in identifying positive breast cancer cases and ranking positive samples.

Overall, the Dynamic Preference fusion method proposed in this study can comprehensively exploit the predictive information provided by different Pareto models and dynamically adjust their weights according to model performance, feature complexity, and predictive uncertainty. These characteristics suggest that the proposed method has potential practical value for breast cancer computer-aided diagnosis.

A.9 Complete Experimental Results of Evaluation of Uncertainty Estimation

Table 11: Uncertainty-based sample filtering for baseline models.

model_name	**AUC**	**AUPRC**	**Accuracy**	**Sensitivity**	**Specificity**
LogisticRegression	0.8571	0.8151	0.7609	0.8333	0.7143
DecisionTree	0.8512	0.8178	0.8261	0.7222	0.8929
SVM_RBF	0.8433	0.7024	0.7174	0.8889	0.6071
KNN	0.8294	0.7378	0.8261	0.8889	0.7857
XGBoost	0.7302	0.6527	0.7174	0.7222	0.7143
percentile_removed:10%					
LogisticRegression	0.8799	0.859	0.8611	0.7857	0.9091
DecisionTree	0.8766	0.8265	0.8889	0.9286	0.8636
SVM_RBF	0.875	0.7732	0.75	0.8571	0.6818
KNN	0.8734	0.8312	0.8056	0.8571	0.7727
XGBoost	0.7532	0.6409	0.75	0.7857	0.7273
percentile_removed:30%					
LogisticRegression	0.9969	0.9909	0.9231	1	0.875
DecisionTree	0.9563	0.9588	0.9231	0.8	1
SVM_RBF	0.9375	0.9252	0.8846	0.9	0.875
KNN	0.9125	0.9019	0.8462	0.8	0.875
XGBoost	0.9063	0.94	0.9615	0.9	1
percentile_removed:50%					

Table 12: Uncertainty-based sample filtering for m=7.

model_name	AUC	AUPRC	Accuracy	Sensitivity	Specificity
model_018	0.8512	0.8430	0.7826	0.7222	0.8214
dynamic_preference	0.8611	0.8550	0.7826	0.7222	0.8214
mean	0.8710	0.8750	0.7826	0.7222	0.8214
percentile_removed:10%,m=7					
model_018	0.8831	0.8912	0.8333	0.7857	0.8636
dynamic_preference	0.8896	0.8910	0.8333	0.7857	0.8636
mean	0.9026	0.9138	0.8333	0.7857	0.8636
percentile_removed:30%,m=7					
model_018	0.8938	0.9076	0.9231	0.8000	1.0000
dynamic_preference	0.9313	0.9476	0.9231	0.8000	1.0000
mean	0.9313	0.9476	0.9231	0.8000	1.0000
percentile_removed:50%,m=7					

Table 13: Uncertainty-based sample filtering for m=15.

model_name	AUC	AUPRC	Accuracy	Sensitivity	Specificity
model_018	0.8631	0.8561	0.7826	0.7222	0.8214
dynamic_preference	0.8750	0.8781	0.8043	0.7222	0.8571
mean	0.8671	0.8704	0.7826	0.6667	0.8571
percentile_removed:10%,m=15					
model_018	0.8864	0.8934	0.8611	0.7857	0.9091
dynamic_preference	0.9058	0.9170	0.8611	0.7857	0.9091
mean	0.9058	0.9170	0.8611	0.7857	0.9091
percentile_removed:30%,m=15					
model_018	0.9091	0.9264	0.9231	0.8182	1.0000
dynamic_preference	0.9333	0.9545	0.9231	0.8182	1.0000
mean	0.9333	0.9545	0.9231	0.8182	1.0000
percentile_removed:50%,m=15					

Table 14: Uncertainty-based sample filtering for m=22.

model_name	AUC	AUPRC	Accuracy	Sensitivity	Specificity
model_018	0.8591	0.8573	0.7826	0.7222	0.8214
dynamic_preference	0.8671	0.8687	0.7826	0.7222	0.8214
mean	0.8651	0.8689	0.7826	0.6667	0.8571
percentile_removed:10%,m=22					
model_018	0.8857	0.8966	0.8611	0.8000	0.9048
dynamic_preference	0.9048	0.9161	0.8611	0.8000	0.9048
mean	0.9048	0.9172	0.8333	0.7333	0.9048
percentile_removed:30%,m=22					
model_018	0.8938	0.9076	0.9231	0.8000	1.0000
dynamic_preference	0.9188	0.9435	0.9231	0.8000	1.0000
mean	0.9250	0.9455	0.9231	0.8000	1.0000
percentile_removed:50%,m=22					

A.10 Complete Experimental Results of Evaluation of Robustness Evaluation

This appendix provides the robustness evaluation results for the remaining model-pool sizes ($m = 7, 15, 22$) and the baseline models. Figures 8–12 show the AUC of each model as the feature-missing ratio and noise intensity increase.

Figure 8 reports the robustness results when $m = 7$. Dynamic preference fusion and Mean Fusion generally outperform BestSingle as feature missingness and noise increase, and dynamic preference fusion maintains a relatively stable AUC under moderate and severe noise conditions. Figure 9 presents the results when $m = 15$. The overall trend is consistent with that observed for $m = 7$: all models degrade as data quality decreases, and dynamic preference fusion remains competitive with or superior to Mean Fusion and BestSingle in most scenarios. Figure 10 shows the results when $m = 22$. Similar to smaller model pools, dynamic preference fusion exhibits good tolerance to feature missingness and noise perturbations and continues to rank among the top-performing methods. Figure 12 reports the robustness results for the baseline models. KNN and XGBoost exhibit more pronounced performance degradation as data quality deteriorates, whereas Decision Tree shows transient fluctuations under certain mild or moderate noise conditions. These results further support the robustness of dynamic preference fusion under complex data-quality conditions.

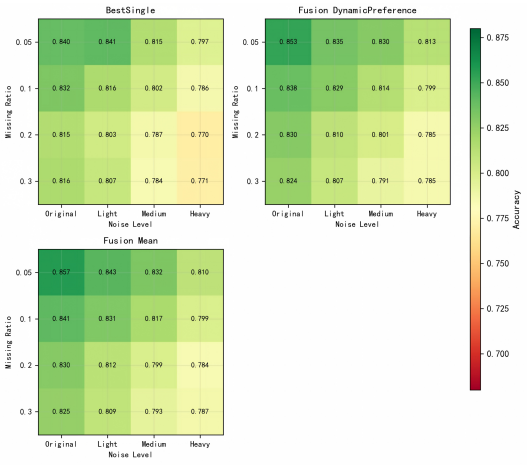


Figure 8: Robustness evaluation when $m = 7$.

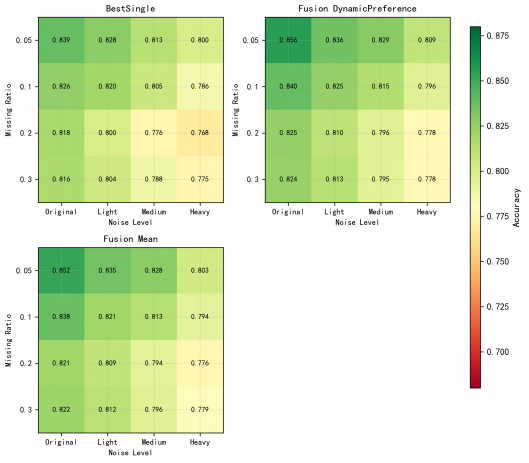


Figure 9: Robustness evaluation when $m = 15$.

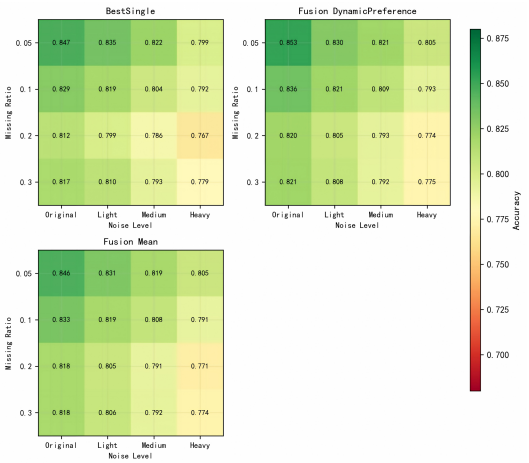


Figure 10: Robustness evaluation when $m = 22$.

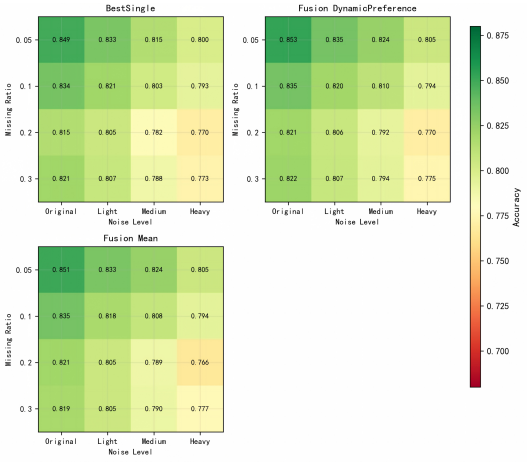


Figure 11: Robustness evaluation when $m = 30$.

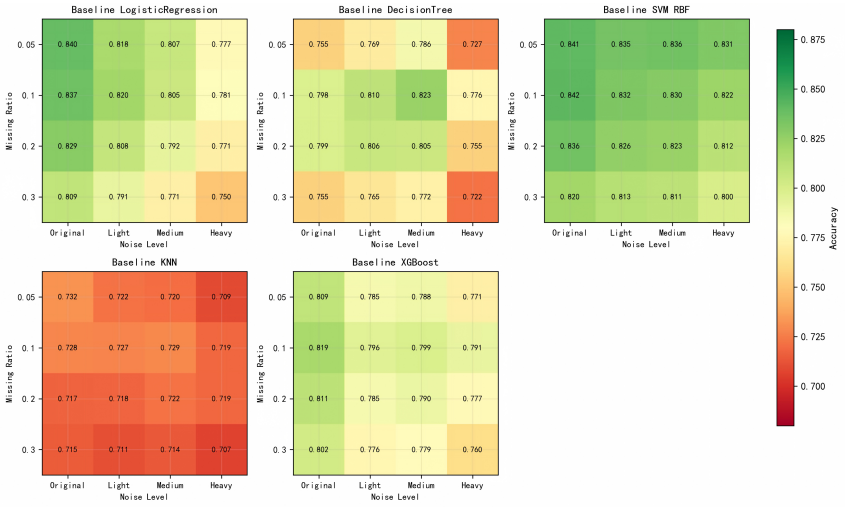


Figure 12: Robustness evaluation for baseline models.