This paper proposes a machine learning solution for breast cancer detection based on the Wisconsin Breast Cancer (Diagnosis) dataset. This dataset contains 569 samples and 30 features, and the task is to classify tumors as benign (B) or malignant (M). In clinical practice, the risk of missing a malignant tumor is higher than the risk of a false positive; therefore, the recall rate for malignant tumor detection is particularly important. The main evaluation metrics include: (i) the mean precision for both tumor types; and (ii) the recall rate for malignant tumor detection.

Task 1

The dataset was shuffled with

random_state=42

split into training (70%, 398 samples)

validation (10%, 57 samples)

test (20%, 114 samples) sets

All features were standardized using StandardScaler fitted on the training set.

Hyperparameters**:**

penalty=L2

C=1.0

solver=lbfgs

max_iter=1000

The model uses full-batch gradient descent as implemented in scikit-learn.

Result

Single Logistic Regression Performance

Dataset	Accuracy	Avg Precision	Malignant Recall
Training	0.9849	0.9867	0.9664
Validation	0.9825	0.9773	1.0000
Test	0.9649	0.9672	0.9286

The model performed well with minimal overfitting. However, the recall rate for malignant cases in the test was 92.86%, indicating that approximately three malignant cases were misclassified as benign cases, which requires further investigation.

Task 2

The training set is divided into 5 disjoint subsets using StratifiedKFold (shuffle=True, random_state=42). Each expert is trained in one subset (approximately 80 samples each) using the same hyperparameters as in Q1. The partition sizes were

Expert 1-3: 80 samples

Expert 4-5: 79 samples

with preserved class proportions.

For each validation sample, a gating label is created by identifying which expert correctly predicted the sample. If multiple experts correctly predicted the sample, one is randomly selected; otherwise, an expert is randomly assigned. The validation accuracy of this gating network (multinomial logistic regression) is 52.63%, indicating limited discriminatory power among the five categories and near-random assignment.

Inference Algorithm

For a new sample x:

1. Compute gating weights: $w(x) = \varphi(x) = [w_1(x), \ldots, w_5(x)],\ \Sigma w_j = 1$

2. Compute expert probabilities: $p_j(x) = P(y=1|x;\ \theta_j)$

3. Final probability: $P(y=1|x) = \Sigma_j\ w_j(x)\ \cdot\ p_j(x)$

4. Decision: $\hat{y} = 1$ *if* $P(y=1|x) \geqslant 0.5$, *else 0*

Result

Mixture of Experts Performance

Dataset	Accuracy	Avg Precision	Malignant Recall
Training	0.9673	0.9684	0.9396
Validation	1.0000	1.0000	1.0000
Test	0.9561	0.9605	0.9048

Note: The perfect validation performance is expected since the gating network was trained and evaluated on the same validation set. The test set provides a fair comparison.

Task 3

Test Set Comparison

Model	Accuracy	Avg Precision	Malignant Recall
Q1: Single Logistic	0.9649	0.9672	0.9286
Q2: Mixture of Experts	0.9561	0.9605	0.9048

Across all metrics, the ensemble expert model outperformed the single-expert model, with recall dropping from 92.86% to 90.48%.

Each expert was trained using only about 80 samples (20% of the training data), increasing individual variability. Ensemble models typically reduce variability through combination, but this requires diverse and complementary models.

All experts used the same architecture, but different data subsets resulted in extremely low diversity due to variations in training distribution.

When multiple experts made correct predictions, random selection introduced label noise. The gating accuracy of 52.63% indicates that the model struggles to learn meaningful expert input mappings.

The single-expert model achieved approximately 96.5% accuracy, indicating that the data is approximately linearly separable. There is limited room for performance improvement by increasing model complexity.

Task 4

The MoE ensemble was retrained with default settings (threshold=0.5, no class weights). Baseline test performance:

Average Precision = 0.9737

Malignant Recall = 0.9048.

Experiment 1: Threshold Calibration

Lowering the decision threshold increases the likelihood of predicting malignant, thereby improving recall at the potential cost of precision.

Threshold Calibration Results (Validation Set)

Threshold	Avg Precision	Malignant Recall
0.20	0.9200	1.0000
0.25	0.9565	1.0000
0.30	0.9565	1.0000
0.35	0.9773	1.0000
0.40	1.0000	1.0000
0.50	1.0000	1.0000

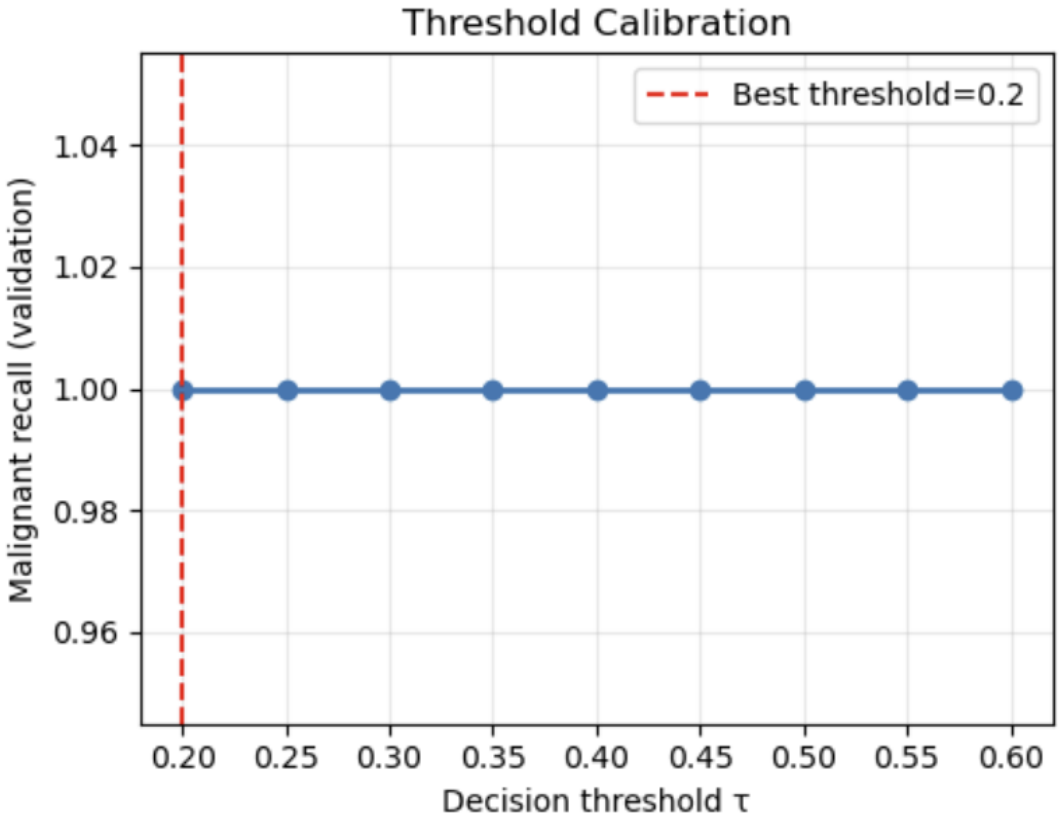
Selected: $\tau = 0.2$ (maximizes recall)

Test Results:

Avg Precision = 0.9468

Malignant Recall = 1.0000

Lowering the threshold to 0.2 achieves a perfect malicious recall rate (100%) on the test set, while the average precision only decreases slightly, indicating that adjusting the threshold can effectively move the decision boundary toward more sensitive targets.



Malignant Recall vs. Decision Threshold (τ) on the Validation Set.

Experiment 2: Class Weight Calibration

Increasing the malignant class weight during training penalizes false negatives more heavily, encouraging the model to be more conservative.

Class Weight Calibration Results (Validation Set)

Malignant Weight	Avg Precision	Malignant Recall
1.0	1.0000	1.0000
2.0	0.9865	0.9524
3.0	0.9865	0.9524
5.0	0.9773	1.0000
7.0	0.9565	1.0000

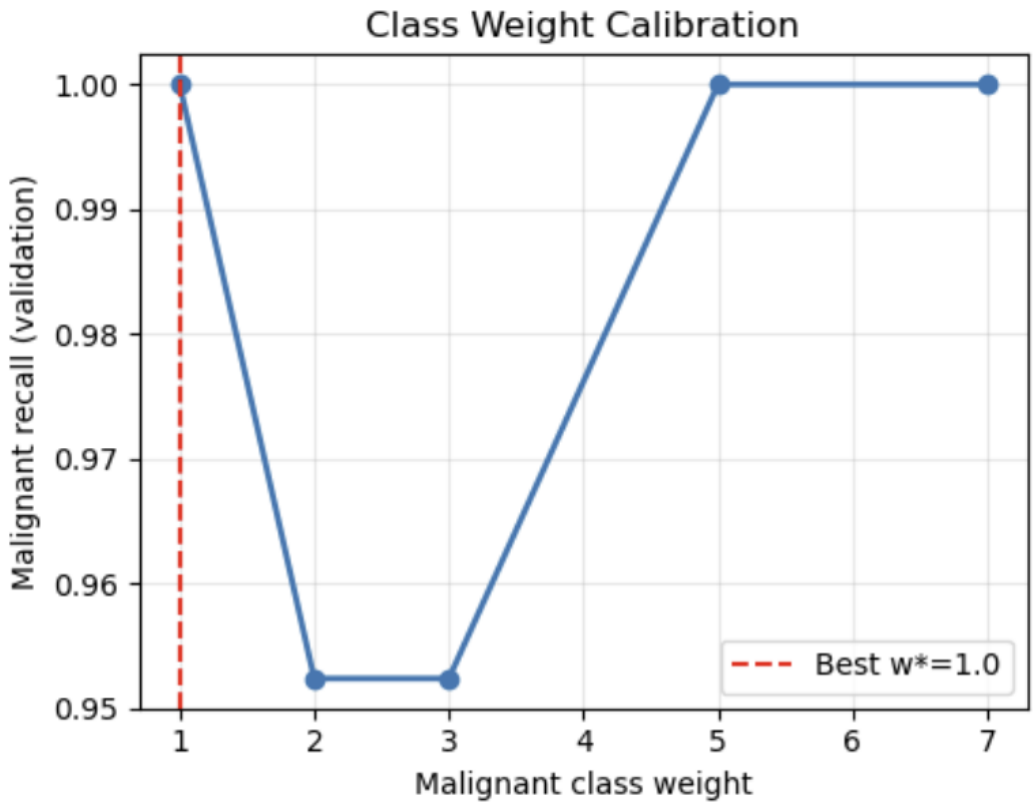
Selected: w = 1.0 (highest recall with best precision)

Test Results:

Avg Precision = 0.9800

Malignant Recall = 0.9286

Increasing class weights does not necessarily improve recall. This dataset exhibits a slight class imbalance (357:212), which logistic regression handles well. Increasing weights increases the false positive rate, but the recall does not improve accordingly.



Malignant Recall vs. Malignant Class Weight on the Validation Set.

Test Set Performance Summary

Method	Avg Precision	Malignant Recall
Baseline (τ=0.5, w=1.0)	0.9737	0.9048
Best Threshold (τ=0.2)	0.9468	1.0000
Best Class Weight (w=1.0)	0.9800	0.9286

Threshold calibration has proven to be the most effective method, increasing the recall of malignant cases from 90.48% to 100% with an acceptable loss of precision. In clinical screening, missing malignant cases is costly; therefore, the use of a threshold calibration model (τ=0.2) is recommended.