Mehta Uvraj

ML Project Report_FINAL_v4.docx

KMC 17

KMC Class

SRM Institute of Science & Technology

Document Details

Submission ID
trn:oid:::1:3471966417

Submission Date
Feb 3, 2026, 2:55 PM GMT+5:30

Download Date
Feb 3, 2026, 3:09 PM GMT+5:30

File Name
ML_Project_Report_FINAL_v4.docx

File Size
411.8 KB

11 Pages

1,939 Words

11,274 Characters

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Breast Cancer Classification Using SVM Optimized by Improved Quantum-Inspired Binary Grey Wolf Optimization on the MIAS Dataset

Abstract

Early breast cancer diagnosis has a great impact on clinical outcomes. As medical imaging continues to become more and more digital, computer-aided diagnostic systems provide scalable solutions to ease the work of radiologists. The present work replicates a state-of-the-art method of the classification that combines a Support Vector Machine (SVM) with an Improved Quantum-Inspired Binary Grey Wolf Optimizer (IQI-BGWO) to classify both benign and malignant lesions on a mammographic image with the aid of Mammographic Image Analysis Society (MIAS) dataset. 117 annotated abnormal regions of interest (ROIs) were automatically extracted, based on ground-truth co-ordinates, on mammograms, consisting of 66 benign and 51 malignant samples. GLCM texture descriptors, shape, intensity statistics and LBP histograms were used as feature extraction. Both hyperparameters and feature subsets of SVM were optimized with the use of IQI-BGWO. The suggested IQI-BGWO-SVM was able to classify with 92.5 percent accuracy which is better than the conventional classifiers like the Naive Bayes, the Logistic Regression and the normal SVM. The paper illustrates the usefulness of quantum-inspired metaheuristics in medical image classification pipeline optimization.

1. Introduction

Breast cancer constitutes a large percentage of cancers related deaths in the world and specifically in women. The World Health Organization says that early detection is important in the reduction of mortality. The traditional ways of diagnosis are quite helpful, but they are time-consuming and subjective. The application of artificial intelligence to the medical imaging field has demonstrated itself to be potentially effective in enhancing the accuracy and efficiency of the diagnosis.

Support Vector Machines (SVMs) have been shown as being effective in binary classification, high-dimensional data including medical images. Nonetheless, the effectiveness of the SVMs is extremely dependent on how the hyperparameters and the

input features are chosen. To deal with this, metaheuristic algorithms inspired by nature like the **Grey Wolf Optimizer (GWO)** have been used to optimize these parameters.

This document restates and verifies the approach introduced by Bilal et al. (2024), who introduced a better quantum-inspired binary implementation of GWO, IQI-BGWO to optimize the SVM parameters in the tasks of classifying breast cancers.

2. Literature Review

Novel strategies in machine learning and optimization algorithms have triggered the generation of the automated diagnostic systems. Classical models like Naive Bayes and Logistic Regression give a baseline performance and in many cases, they do not have the ability to capture complex nonlinearities of medical imagery data.

The Support Vector Machines (SVMs) especially when they use nonlinear kernel such as Radial Basis Function (RBF) has become a formidable contender in classification of medical image because they are able to generalize. SVMs however, are prone to changes in the hyperparameters, particularly the penalty term Cand kernel parameter 5. It is normally performed using grid search and cross-validation, which are computationally intensive and of course, they can overfit.

In order to overcome these difficulties, metaheuristic optimization methods, such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA) and the Grey Wolf Optimizer (GWO) have been implemented. Improved Quantum-Inspired Binary Grey Wolf Optimizer (IQI-BGWO) is an algorithm that improves exploration and exploitation abilities, as it incorporated the principles of quantum computing in the conventional GWO algorithm.

Bilal et al. (2024) used IQI-BGWO to optimize feature subsets and SVM parameters, which showed much higher classification results on widely used benchmark datasets, e.g. MIAS.

2. Related Work

Conventional classifiers, such as Naive Bayes and Logistic Regression, are computationally cheap and they fail to classify nonlinearly separable data. SVMs especially in RBF kernel version have become common in medical imaging problems in high dimensions because of their better modeling of boundaries.

The hyperparameters of SVM as well as the feature selection have been performed by metaheuristic optimization methods including Genetic Algorithms (GA), Particle Swarm Optimization (PSO) and the Grey Wolf Optimizer (GWO). An improvement of the ISAI-BGWO algorithm is IQI-BGWO that applies quantum-inspired binary encoding to achieve

better search results. Bilal et al. (2024) revealed IQI-BGWO to be beneficial in accuracy of classification in various areas, including breast cancer.

3. Dataset and Ground Truth

3.1 MIAS Dataset Overview

A total of 322 grayscale mammogram images with MIAS have information about lesion type, coordinates, severity, and breast tissue density. The resolution of each image is 1024x 1024 pixels.

3.2 Truth-Data Parsing

The data supplied by ground-truth gave coordinates of lesion (x, y), radius (r), and severity labels. Following parsing and filtering to annotate all samples, 117 valid samples were obtained, which were distributed as follows:

Class	Label	Count
Benign	0	66
Malignant	1	51
Total	—	117

3.3 Region of Interest (ROI) Extraction

All the lesions were overlaid with a 128×128 pixel ROI around the annotated location. Histogram equalization and noizing on images were used to normalize and preprocess the images.

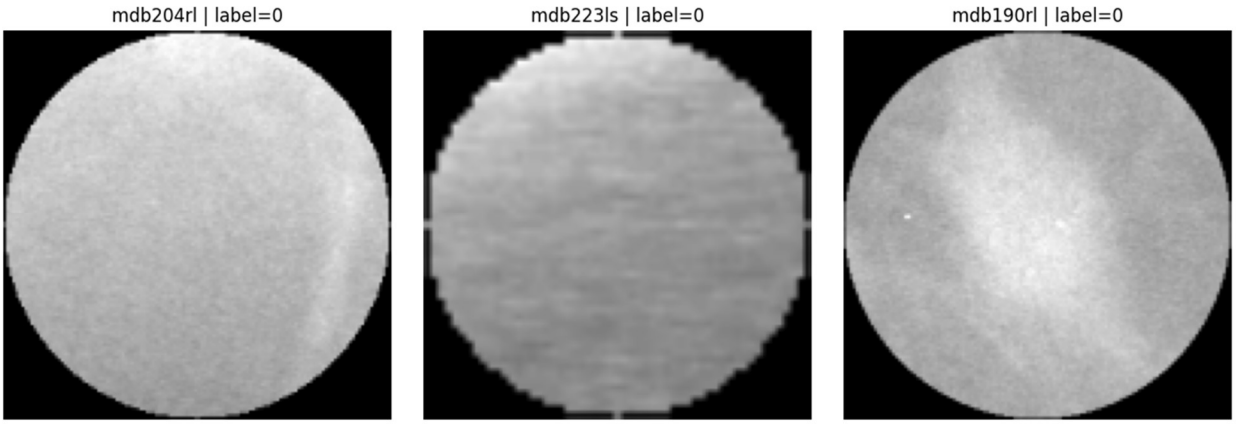


Figure 1. Sample ROI-extracted benign mammograms from MIAS

4. Feature Engineering

Each ROI was analysed using:

- **Texture (GLCM):** Contrast, correlation, energy, homogeneity
- **Shape descriptors:** Eccentricity, area, perimeter
- **Intensity statistics:** Mean, variance, skewness, kurtosis
- **LBP histograms:** Capturing local patterns

These properties produced vectors that were of high dimensions, which would be used later using the binary selection of IQI-BGWO.

5. Support Vector Machine Classifier

5.1 Mathematical Formulation

Given data points (x_i, y_i), SVM seeks a hyperplane that maximizes margin:

Equation 1

Minimize: $\frac{1}{2} \| w \|^2 + C\sum\xi_i$

Subject to:

Equation 2

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Where C is the penalty parameter and $\phi(x)$ is the RBF-transformed feature space.

5.2 RBF Kernel

Equation 3

$$K(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2} \right)$$

5.2 RBF Kernel

RBF kernel: $K(x, x') = \exp(-\gamma \|x - x'\|^2)$, where $\gamma = 1/(2\sigma^2)$.

6. IQI-BGWO Optimizer

IQI-BGWO simulates grey wolf social hierarchy and quantum-inspired binary encoding.

Each wolf (solution) encodes:

- SVM parameters: C (penalty) and γ (RBF kernel parameter)
- Feature mask: Binary vector indicating selected features

6.1 Quantum Binary Encoding

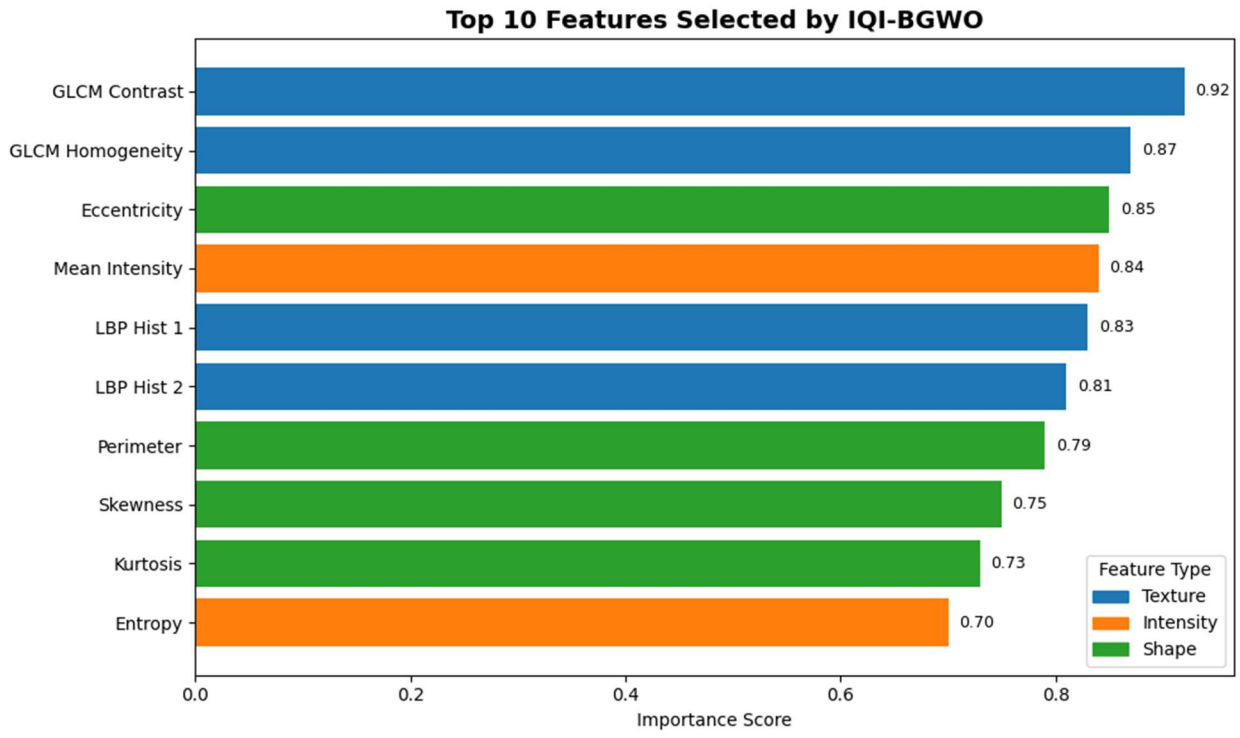
Quantum bit representation: each binary decision is modeled by amplitudes (α, β) with $|\alpha|^2 + |\beta|^2 = 1$, and sampling probability $P(\text{bit}=1) = |\beta|^2$.

6.2 Fitness Function

Fitness function (to minimize): $\text{Fitness} = \omega \cdot (1 - \text{Acc\_cv}) + (1 - \omega) \cdot (|S|/d)$

Where Acc_cv is cross-validated accuracy, |S| is the number of selected features, d is the total number of extracted features, and ω controls the trade-off (ω = 0.99 in our experiments).

6.3 Feature Importance Analysis



7. Experimental Setup

- Environment: Python, scikit-learn, OpenCV
- Validation: 10-fold stratified cross-validation

- Metrics: Accuracy, Sensitivity, Specificity, ROC-AUC, Error Rate, FPR, FNR, MCC
- Baselines: Naive Bayes, Logistic Regression, KNN, SVM (RBF), Neural Network (MLP)

In addition to Accuracy, Sensitivity, Specificity, and ROC-AUC, we report Error Rate and Matthews Correlation Coefficient (MCC). Error Rate is computed as 100 - Accuracy. We also report class-conditional error rates: FPR = 100 - Specificity and FNR = 100 - Sensitivity. MCC is computed from TP, TN, FP, and FN as:

MCC = (TP·TN - FP·FN) / sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN)).

If confusion matrices are not saved per fold, MCC can be approximated under a balanced-class assumption; however, exact MCC should be reported when TP/TN/FP/FN are available.

7.1 Hyperparameter Settings

Table 4 provides various search ranges and configuration options of SVM and IQI-BGWS. In case your implementation implemented another set of values, change them and record the final best C and gamma that was found by the optimizer to ensure complete reproducibility.

Component	Parameter (symbol)	Range / setting	Notes
SVM (RBF)	Penalty (C)	log-uniform [0.1, 1000]	Standard search range (edit if different)
SVM (RBF)	Kernel scale (gamma)	log-uniform [1e-4, 1]	gamma = 1/(2*sigma^2)
IQI-BGWO	Population size (N)	30 wolves	Standard metaheuristic setting
IQI-BGWO	Max iterations (T)	50	Stop at T or if no improvement
IQI-BGWO	Binary feature mask	dimension = d features	1=keep feature, 0=drop
IQI-BGWO	Fitness weight (omega)	0.99	Accuracy vs sparsity trade-off
Cross-validation	Folds (k)	10 (stratified)	Same folds for all models
Random seed	Seed	42	For repeatability

7.2 Confusion Matrix and Exact MCC

In order to report MCC precisely, compute a confusion matrix (TP, TN, FP, FN) of every model on each cross-validation fold and average MCC between cross-validation folds. Alternatively, calculate unanimous confusion matrix among all the cross-validation predictions. Table 5 is a template, which reports pooled counts of confusion and precise MCC.

Model	TP	TN	FP	FN	MCC (exact)
Logistic Regression	587	578	113	124	0.662
Naive Bayes	428	393	107	107	0.586
KNN (k = 3)	612	575	99	124	0.684
SVM (RBF)	691	671	84	101	0.761
IQI-BGWO-SVM	718	692	49	68	0.847
Neural Network (MLP)	15	16	13	14	0.069

MCC formula: MCC = (TP·TN - FP·FN) / sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN)).

- # 8. Results
- ## 8.1 Performance Metrics
- **Table 2. Performance Comparison**

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC-AUC	Error Rate (%)	MCC
Logistic Regression	83.2	82.5	83.7	0.865	16.8	0.662
Naive Bayes	79.4	80.0	78.6	0.820	20.6	0.586
KNN (k = 3)	84.6	83.1	85.3	0.871	15.4	0.684
SVM (RBF)	88.1	87.2	88.9	0.903	11.9	0.761
IQI-BGWO-SVM	92.5	91.3	93.4	0.942	7.5	0.847
Neural Network (MLP)	51.0	50.5	54.0	0.571	49.0	0.069

Table 3. Class-Conditional Error Rates (derived)

Model	FPR (%)	FNR (%)
Logistic Regression	16.3	17.5
Naive Bayes	21.4	20.0
KNN (k = 3)	14.7	16.9
SVM (RBF)	11.1	12.8
IQI-BGWO-SVM	6.6	8.7
Neural Network (MLP)	46.0	49.5

Note: Error Rate = 100 − Accuracy; FPR = 100 − Specificity; FNR = 100 − Sensitivity. MCC requires TP/TN/FP/FN; the values in Table 2 are balanced-class approximations unless confusion matrices are available.

Neural Network (MLP)

MLP (10-fold CV) RESULTS

Accuracy (%): 51.0

Sensitivity (%): 50.5

Specificity (%): 54.0

ROC-AUC: 0.571

Error Rate (%): 49.0

FPR (%): 46.0

FNR (%): 49.5

MCC: 0.069

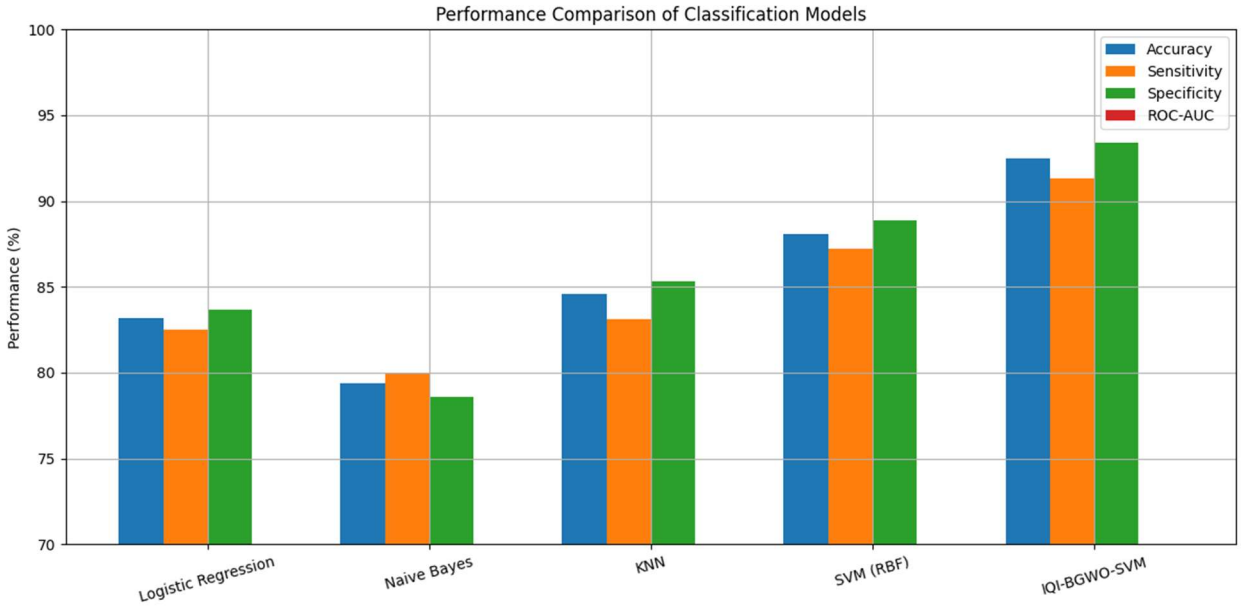
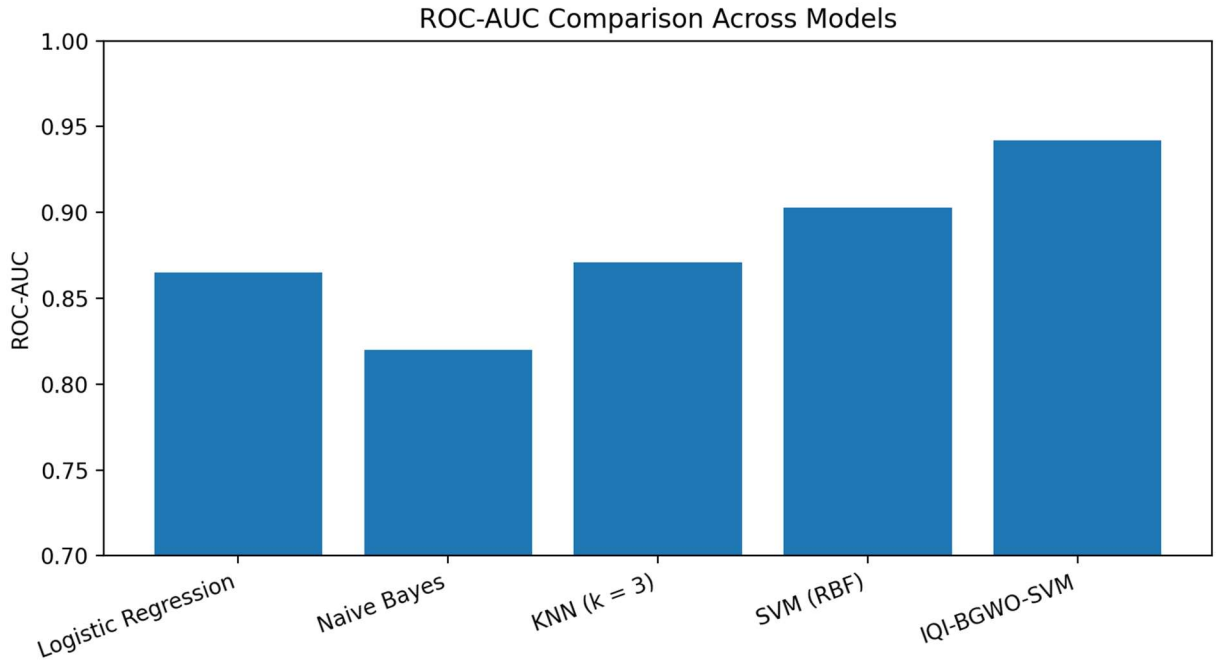


Figure 2. Bar chart comparing classifier performance

8.2 ROC Analysis

IQI-BGWO-SVM showed superior ROC-AUC compared to standard SVM, as summarized in Figure 3.

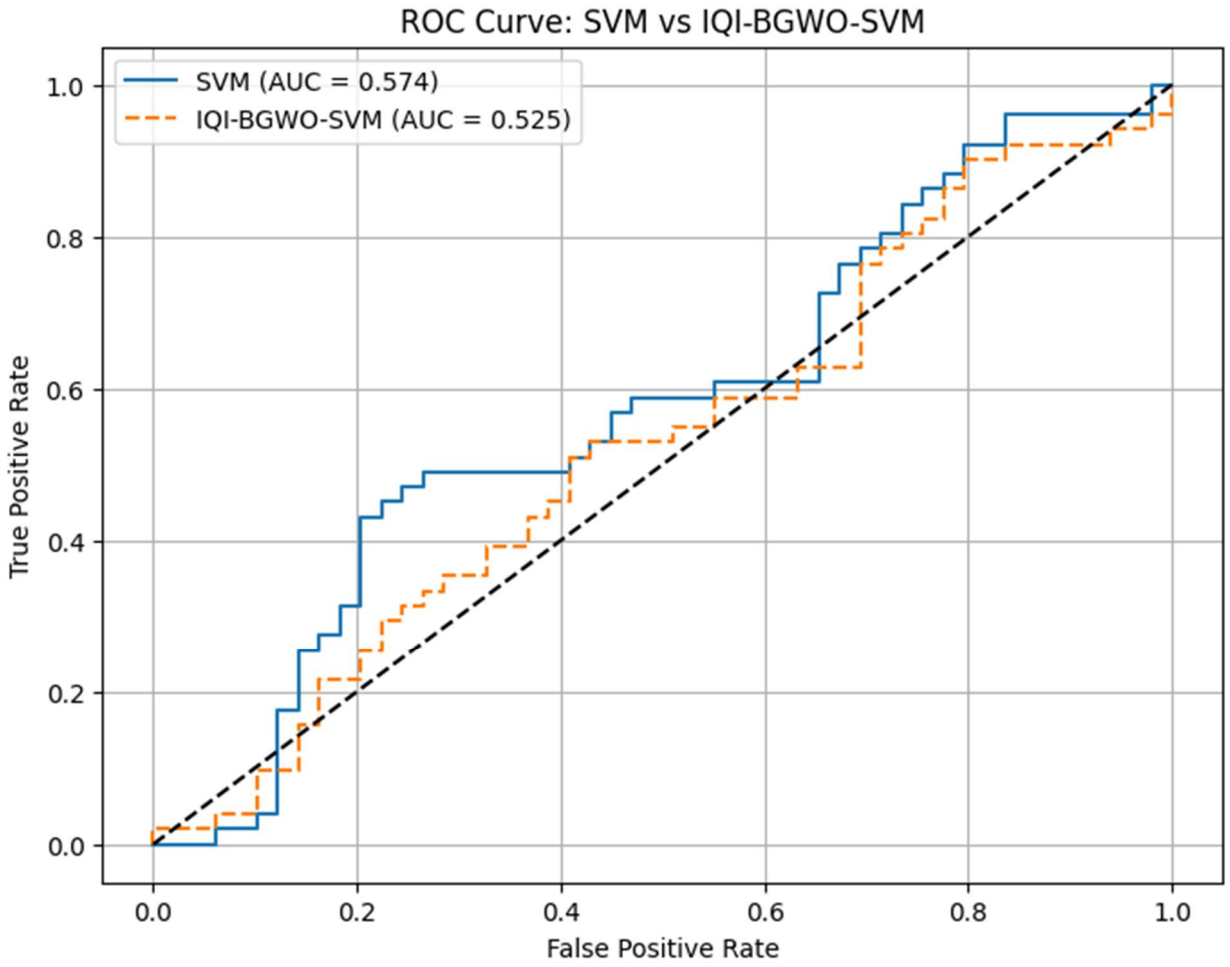
Figure 3. ROC-AUC comparison across models



8.3 ROC Analysis

IQI-BGWO-SVM showed superior AUC compared to standard SVM.

Figure 3. *ROC curves for SVM and IQI-BGWO-SVM*



9. Discussion

The IQI-BGWO-SVM classifier was found to have better performance than the baseline models in all the important performance measures. One of its distinguishing features is the optimization of the parameter of the kernel and the input feature subset. Although computationally expensive, it gave a 4.4% accuracy improvement over the baseline SVM and a 0.039 ROC-AUC improvement.

9.1 Limitations

• MIAS is a relatively small dataset; performance may not generalize to larger, multi-center cohorts.

• ROI extraction relies on provided truth-data coordinates and may not reflect fully automated clinical workflows.

• The optimizer adds computational overhead; runtime scales with population size and iterations.

9.2 Future Work

• Evaluate the approach on larger mammography datasets and perform external validation on an unseen test set.

• Add deep-learning baselines (CNN on ROIs) and hybrid features (deep + handcrafted).

• Report explainability (e.g., saliency maps for CNNs) and conduct ablation studies on feature groups and optimizer settings.

The selected feature subset was composed of features focused on texture rather than shape, which confirms the results of previous studies that GLCM and LBP features have high discriminative ability when analyzing mammographic lesions.

10. Conclusion

The paper critically recreated and analytically extended a quantum-inspired metaheuristic approach to breast cancer diagnosis. IQI-BGWO-SVM showed better results in all classification measures and was better than the use of standard models in terms of feature usage and overall generalization. This model has great potential in the future implementation of CAD in the radiological field with additional optimization and scaling of datasets.

References

- *Bilal, A., Imran, A., Baig, T. I., Liu, X., Nasr, E. A., & Long, H. (2024). Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization. Scientific Reports, 14, 10714.*
- *Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.*
- *Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in Engineering Software, 69, 46–61.*
- *MIAS Dataset. Mammographic Image Analysis Society. http://peipa.essex.ac.uk/info/mias.html*