Breast Cancer Detection in Histopathology images using Machine Learning Models with Optimized Features - A Comparative Study

MINI PROJECT REPORT
Submitted in partial fulfilment for the Degree of Integrated MSc in Mathematics with minor in Data Science

By
GOWRI SHAJU
AM.SC.I5MAT20010

DEPARTMENT OF MATHEMATICS
AMRITA SCHOOL OF PHYSICAL SCIENCES
AMRITA VISHWA VIDHYAPEETHAM
AMRITAPURI CAMPUS - 690525

January 2025

AMRITA SCHOOL OF PHYSICAL SCIENCES

AMRITA VISHWA VIDHYAPEETHAM

AMRITAPURI CAMPUS - 690525

BONAFIDE CERTIFICATE

This is to certify that the mini project entitled ***"Breast Cancer Detection in Histopathology images using Machine Learning Models with Optimized Features - A Comparative Study"*** submitted by **Gowri Shaju, Register Number - AM.SC.I5MAT20010**, for the award of the **Degree of Integrated MSc in Mathematics with minor in Data Science** under the **Faculty of Physical Sciences** is a bonafide record of the work carried out by her under my guidance and supervision at the Department of Mathematics, Amrita School of Physical Sciences, Amritapuri.

Dr. Rajan S
Chairperson
Department of Mathematics
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India

Ms. Lekhs S Nair
Project Supervisor
Assistant Professor
Department of Computer Science
Amrita Vishwa Vidyapeetham
Amritapuri Campus, India

AMRITA SCHOOL OF PHYSICAL SCIENCES
AMRITA VISHWA VIDHYAPEETHAM
AMRITAPURI CAMPUS - 690525

DECLARATION

I, **Gowri Shaju, Register Number - AM.SC.I5MAT20010**, hereby declare that this project work entitled **"Breast Cancer Detection in Histopathology images using Machine Learning Models with Optimized Features - A Comparative Study"**, is the record of the study done by me under the guidance of Ms. Lekha S Nair, Assistant Professor, Department of Computer Science, Amrita Vishwa Vidyapeetham, Amritapuri. To the best of my knowledge, this work has not formed the basis for awarding any degree/diploma/ associateship/fellowship/or a similar award to any candidate in any University.

Place : Amritapuri **Signature of the student**
Date : 06-01-2025

COUNTERSIGNED

Ms. Lekha S Nair
Assistant Professor
Department of Computer Science
Amrita Vishwa Vidyapeetham
Amritapuri Campus

Contents

Acknowledgement

I am deeply grateful to everyone who supported and guided me throughout the course of this project. I would like to express my sincere gratitude to my project guide Ms. Lekha S Nair for her expert advice, encouragement, and valuable feedback, which were crucial in completing this work. This mentorship has been an inspiring experience for me. I am also thankful to the Department of Mathematics at Amrita School of Physical Sciences, Amritapuri Campus for providing the necessary resources and facilities to carry out this project. The resources and facilities offered by the department were instrumental in the successful execution of this project. I am also indebted to the faculty members and my friends, whose constructive feedback and cooperation have greatly enriched this work. Lastly, I would like to thank my parents and God for their constant support and belief in me, which kept me going through every challenge.

List of Tables

List of Figures

Abstract

Determining the effectiveness of machine learning algorithms trained on specific optimal features derived from the examination of breast histopathology pictures is the aim of this project. Several techniques are used to optimize and choose the key features from the available feature set in order to increase the efficacy of the model. Support vector machines (SVM), the logistic regression model, and other classification models are tested using this feature set. To identify the most effective feature and set of methods, the metrics of accuracy, precision, recall, and f1 score are evaluated both with and without using all the features. The results offer a better model for detecting breast cancer since it has been shown that using feature sets that are tuned significantly increases the model's classification accuracy of benign and malignant tumours. This study highlights the use of contemporary machine learning algorithms in medical imaging and offers information on the value of feature optimization in breast cancer diagnosis models.

1 Introduction

One of the common reasons for the increase in death rate among women is breast cancer. Breast cancer occurs when the tissues in the breast region start growing uncontrollably and form tumours. In 2022, as per the estimation of World Health Organization(WHO), approximately 2.3 million women were diagnosed with breast cancer and about 670000 deaths were reported globally. Early detection of tumour is critical in reducing the death rate due to breast cancer. Histopathology, the study of tissues at a microscopic level is one among the various methods for the diagnosis of breast cancer. It begins with biopsy where the tissue of the breast region is collected and observed under the microscope for study. Histopathology images are the microscopic images of the cells.

In this work, we suggest a comparative methodology to assess how well machine learning models use features extracted from histopathology images to detect breast cancer. The Wincosin dataset, which comprises 30 features taken from histopathology pictures, is used for the analysis. To categorize cases of breast cancer, four popular machine learning models are used: Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. There are two stages to the investigation. To create baseline performance measurements, the models are trained and assessed using all 30 characteristics in the first phase. To find the most pertinent subset of characteristics, we combine feature optimization methods in the second phase, including Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Genetic Algorithm (GA), Differential Evolution (DE) and Whale Optimization Algorith (WOA). The chosen optimal feature set is used to train the machine learning models and the accuracies obtained in both stages are compared.

The significant improvement in the prediction of malignancy by machine learning models using optimized feature set is demonstrated by the findings of this project. The feature set derived from PSO, ACO, GA, DE and WOA yields higher accuracy compared to the feature set consisting of all the features of the dataset.

2 Proposed Methodology

2.1 Dataset Description

The dataset used for the implementation of this project is Wincosin dataset. It consists of 569 instances and 32 attributes of which two represent the diagnosis column (M - malignant and B - benign) and the ID number of the patients respectively. So, it consists of a total of 30 features.

Total	Benign	Malignant
569	357	212

Table 1: Wincosin Dataset

2.2 Phase 1 - Without Optimization

In the first phase, machine learning models, Logistic Regression, SVM, KNN and Random forest are trained using the dataset. The accuracy, precision, recall and F1 Score of each model are noted.

Machine learning have been contributing significantly to the medical field over the past years. among the various classification models, logistic regression, support vector machine (SVM), K-nearest neighbors (KNN), and random rorest have been widely applied to classify tumors as benign or malignant. Logistic regression is the statistical model that uses the sigmoid function to predict the output. SVM is a powerful classifier that identifies the optimal hyperplane to separate different classes with a maximum margin. KNN, a simple yet effective method, classifies data based on the majority vote of its nearest neighbors. Random Forest, an ensemble method, constructs multiple decision trees and aggregates their predictions to imporve the classification accuracy.

2.3 Phase 2 - With Optimization

Optimization algorithms are widely used in machine learning to find the best parameters that maximize or minimize the objective function, which here is the accuracy or the cost function. In phase 2 of this work, evolutionary optimization algorithms, PSO, ACO, GA, DE and WOA are used for feature selection. Evolutionary optimization algorithms are inspired by natural processes such as genetics, evolution and social behavior of organisms. These algorithms work by iteratively improving a set of candidate solutions by optimizing the given objective function.

2.3.1 Particle Swarm Optimization (PSO)

i. For each particle in the population,
randomly initialize the particle with some position and velocity.
END

ii. For each particle,
calculate the fitness value.
If the fitness value is better than the pBest in the history, set the current value as the pBest.
END
Choose the particle with the best value as the gBest.

iii. For each particle

- Update particle velocity

$$v_i[t+1] = v_i[t] + c_1 \cdot r_1 \cdot (pBest[t] - x_i) + c_2 \cdot r_2 \cdot (gBest[t] - x_i)$$

c_1 and c_2 are usually set to 2.
r_1 and r_2 are random numbers between 0 and 1.

- Update particle position by

$$x_i[t+1] = \begin{cases} 1, & \text{sigmoid}(\text{v}_\text{i}[\text{t}+1]) > \text{rand}(0,1) \\ 0, & \text{Otherwise} \end{cases}$$

END

iv. Repeat the process for a predefined number of iterations or until the stopping criteria is met.

2.3.2 Ant Colony Optimization (ACO)

i. For each particle in the population,
   initialize the pheromone value, τ_i
   Initialize parameters,
   No. of ants, n
   No. of iterations
   Pheromone evaporation rate, ρ
   Pheromone importance, α
   Pheromone deposit constant, Q
   Heuristic information, η_i
   Heuristic importance, β
   END

ii. For each ant,
   calculate the transition probability by

$$P_i = \frac{\tau_i^{\alpha} \cdot \eta_i^{\beta}}{\sum_{j \in Featuresset} \tau_j^{\alpha} \cdot \eta_j^{\beta}}$$

   Each ant contrsucts a feature subset by deciding whether to include-the particular feature or not depending upon the transition probability.
   END

iii. Evaluate the subsets
   Train a classifier using the subset of the feature set obtained in the above step and calculate the cost function (cost) and accuracy.

iv. Update the pheromone value

- Evaporate pheromones by

$$\tau_i \longrightarrow (1 - \rho) \cdot \tau_i$$

- Deposite new pheromones by

$$\tau_i \longrightarrow \tau_i + \sum_{ants} \frac{Q}{Cost}$$

   END

v. Repeat the process for a predefined number of iterations or until the stopping criteria is met.

2.3.3 Genetic Algorithm (GA)

i. Generate an initial population of N chromosomes.

Initialize the parameters,

Mutation factor (F)

Crossover probability (CR)

Maximum number of generations (Max_Generations)

Fitness function (Fitness_Function).

ii. For each chromosome in the population,

evaluate how well it performs by decoding the chromosomes to get the selected features, training a classifier and calculating the accuracy.
END

iii. Select parent chromosomes for crossover

Select the fittest individual to form next generation. Usually chromosomes with higher fitness values (accuracy) are selected.

iv. For each pair of selected parent,

combine parent chromosomes to produce offspring by selecting a cross-over point and swapping bits between two parents beyond this point.

For example:

$P_1 = 1001101011$

$P_2 = 0101011011$

Let the crossover point be 5, then,

Offspring 1 = 1001111011 and

Offspring 2 = 0101001011

END

v. For each offspring chromosome,

introduce diversity by flipping bits in the offspring with a small probability.
END

vi. Replacement of chromosomes

Replace the least fit chromosomes with the offspring.

vii. Stop the process when the fitness does not improve for several generations.

2.3.4 Differential Evolution (DE)

i. For each individual in the population,

initialize a population N_p of binary vectors, each of size D (here D = 30).

Initialize the parameters,

Mutation factor (F)

Crossover probability (CR)

Maximum number of generations (Max_Generations)

Penalty coefficient for feature subset size (λ, optional).

Fitness function (Fitness_Function).

END

ii. For each individual i in the population,

Create a mutant vector using three distinct random indivduals from the popolutation (r_1, r_2, r_3) by :

$$v_i = x_{r1} + \mathrm{F}\cdot(x_{r2} - x_{r3})$$

END

iii. For each individual i in the population

Generate a trial vector u_i by

$$u_{i,j} = \begin{cases} v_{i,j}, & \text{if } r \leq \mathrm{CR} \text{ or } \mathrm{j} = \mathrm{j_{rand}} \\ x_{i,j}, & \text{otherwise} \end{cases}$$

Convert this vector to a binary value by using a threshold.

END

iv. For each individual i,

Evaluate the fitness of the trial vector u_i by training a classifier and comparing the fitness value with the current x_i.

$$x_i = \begin{cases} u_i, & \mathrm{fitness(u_i)} > \mathrm{fitenss(x_i)} \\ x_i, & \text{otherwise} \end{cases}$$

END

v. Repeat the process for a fixed number of iterations or until convergence.

2.3.5 Whale Optimization Algorithm (WOA)

i. For each whale in the population,

Initialize a population X_i of binary vectors, each of size d = 30.

END

ii. For each whale,

calculate the fitness value using any classifier.

END

Choose the whale with highest fitness value as X_{best}.

iii. For each whale,

- Calculate

A = 2a·r_1 - a, where a decreases linearly from 2 to 0 and $r_1 \sim U(0,1)$.

$C = 2r_2$, where $r_2 \sim U(0,1)$.

Set $p \sim U(0,1)$ to decide between encircling prey, spiral update, or exploration.

- if ($p < 0.5$):

($|A| < 1$):

Update position based on the best whale by,

$$X_{i+1} = X_{best} \text{ - A } \cdot |C \cdot X_{best} \text{ - } X_i|$$

else:

select a random agent, X_{rand} and update current agent by,

$$X_{i+1} = X_{rand} \text{ - A } \cdot |C \cdot X_{rand} \text{ - } X_i|$$

else:

Compute

$$\text{D} = |X_{best} \text{ - X}|$$

Update the position by,

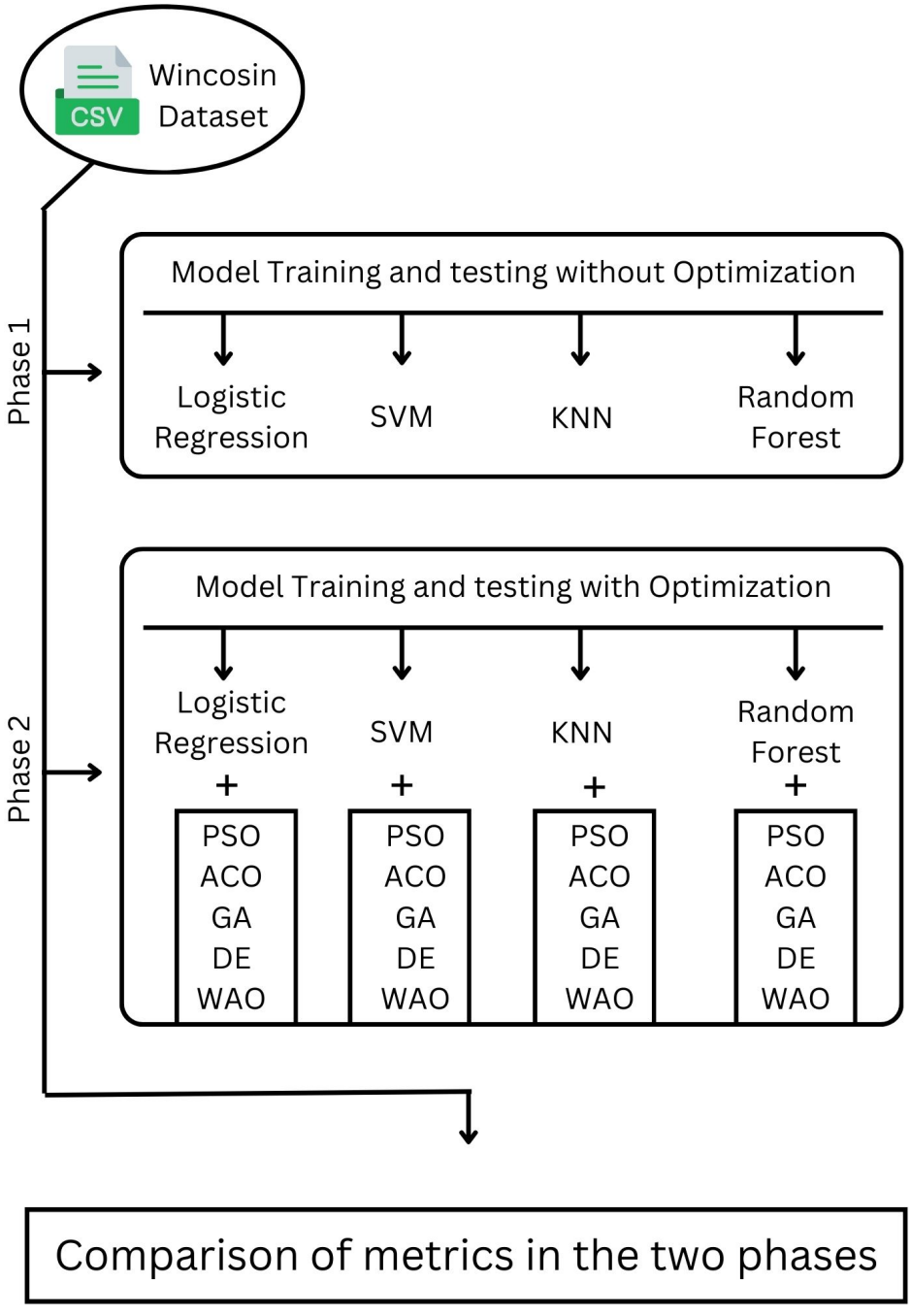
$$X_{i+1} = \text{D } \cdot e^{b \cdot t} \cdot \cos(2\pi l) + X_{best}$$

$l \sim$ U(-1,1).

END

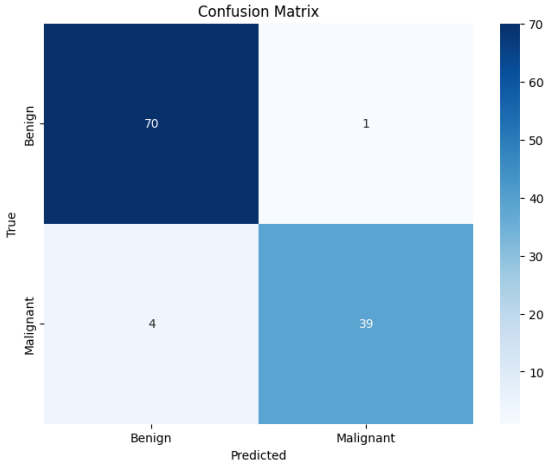
iv. Calculate the fitness of each whale and update X_{best} if there is any better solution.

v. Repeat the process for predefined number of iterations.


Wincosin Dataset
CSV
Phase 1
Model Training and testing without Optimization
Logistic Regression
SVM
KNN
Random Forest
Phase 2
Model Training and testing with Optimization
Logistic Regression
SVM
KNN
Random Forest
+
PSO
ACO
GA
DE
WAO
Comparison of metrics in the two phases

3 Implementation

Logistic Regression



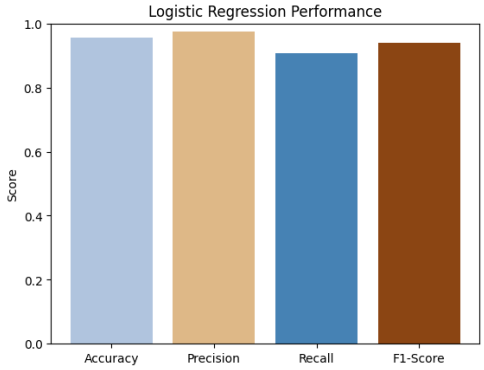


Figure 1: Logistic Regression

Accuracy: 0.956140350877193

Precision: 0.975

Recall: 0.9069767441860465

F1 Score: 0.9397590361445783

Logistic Regression + PSO

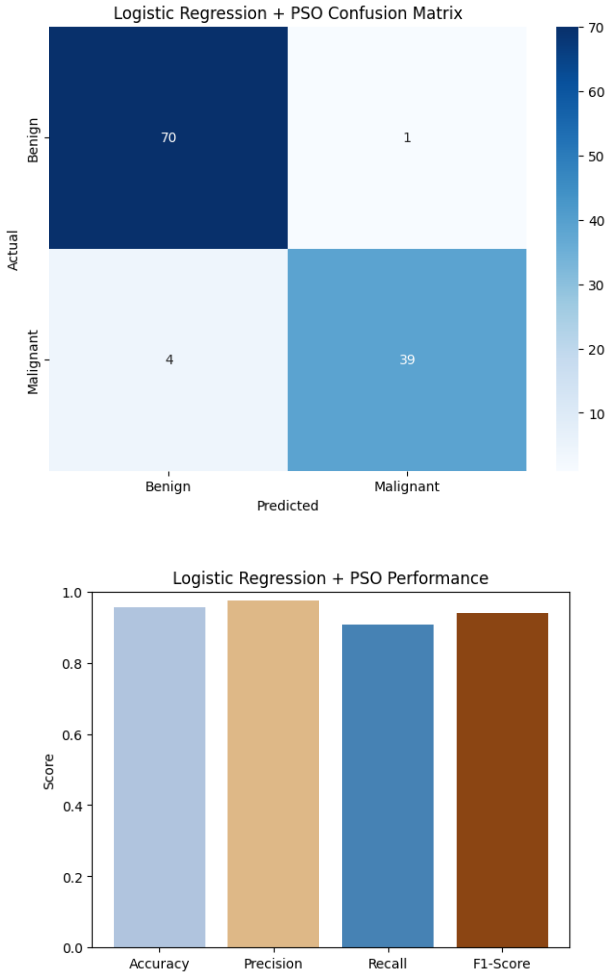


Figure 2: Logistic Regression + PSO

Accuracy: 0.956140350877193

Precision: 0.975

Recall: 0.9069767441860465

F1 Score: 0.9397590361445783

Logistic Regression + ACO

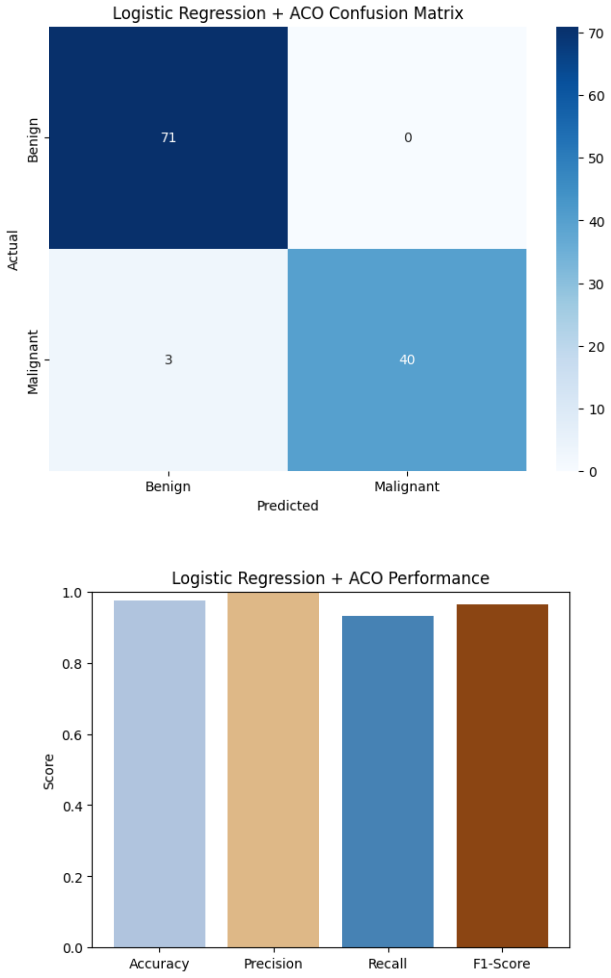


Figure 3: Logistic Regression + ACO

Accuracy: 0.9736842105263158

Precision: 1.0

Recall: 0.9302325581395349

F1 Score: 0.963855421686747

Logistic Regression + GA

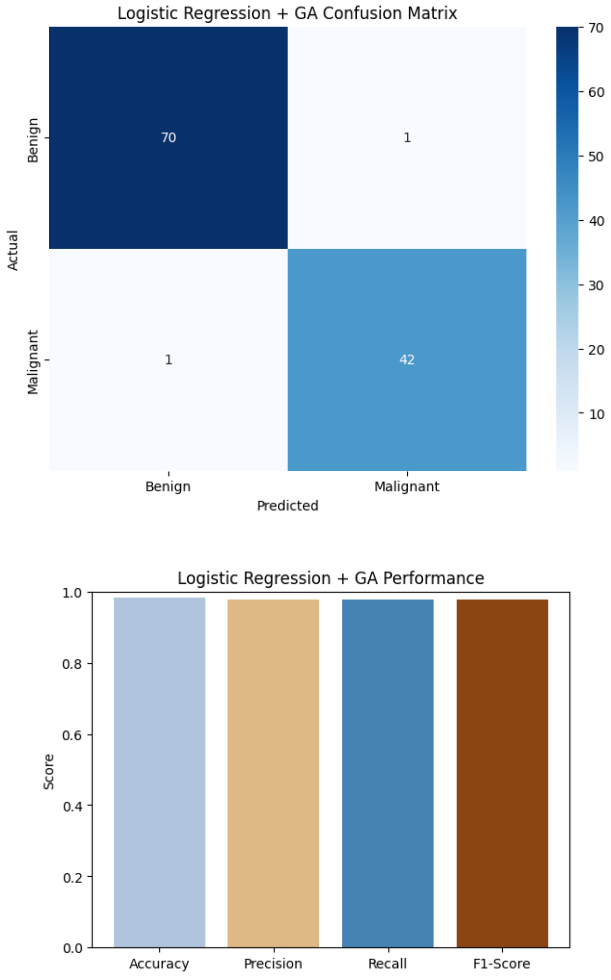


Figure 4: Logistic Regression + GA

Accuracy: 0.9824561403508771
Precision: 0.9767441860465116
Recall: 0.9767441860465116
F1 Score: 0.9767441860465116

Logistic Regression + DE

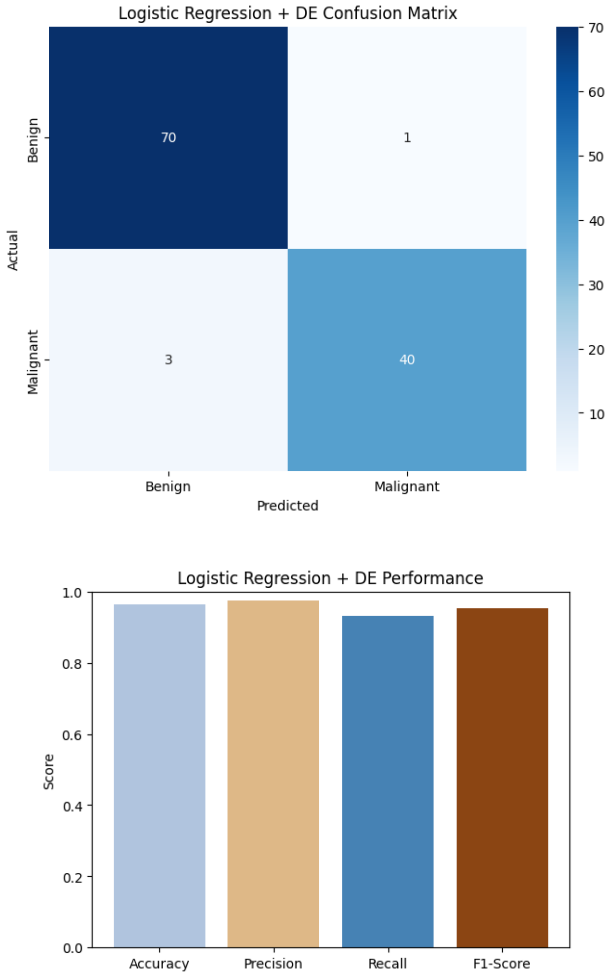


Figure 5: Logistic Regression + DE

Accuracy: 0.9649122807017544
Precision: 0.975609756097561
Recall: 0.9302325581395349
F1 Score: 0.9523809523809523

Logistic Regression + WOA

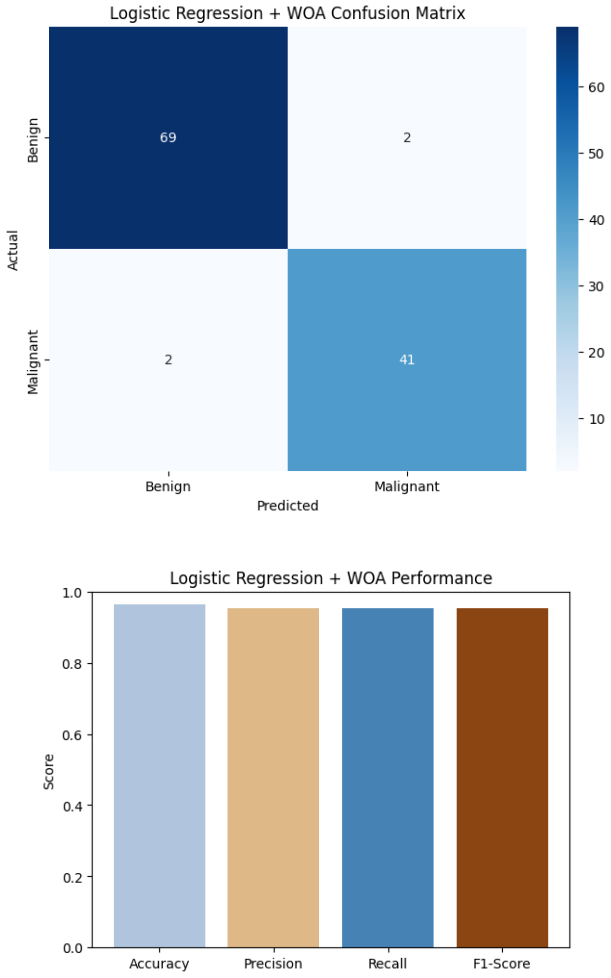


Figure 6: Logistic Regression + WOA

Accuracy: 0.9649122807017544
Precision: 0.9534883720930233
Recall: 0.9534883720930233
F1 Score: 0.9534883720930233

Support Vector Machine

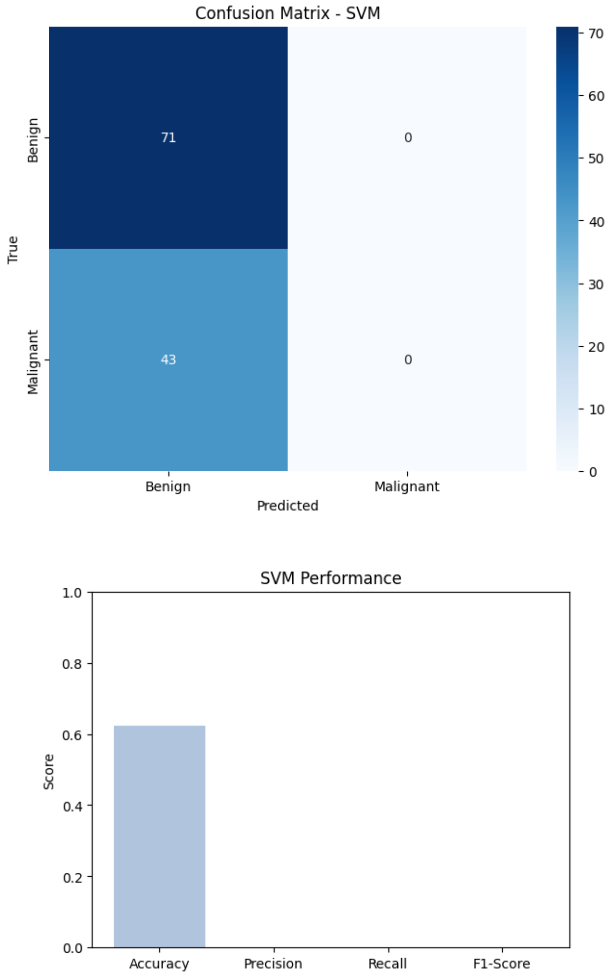


Figure 7: Support Vector Machine

Accuracy: 0.6228070175438597
Precision: 0.0
Recall: 0.0
F1 Score: 0.0

Support Vector Machine + PSO

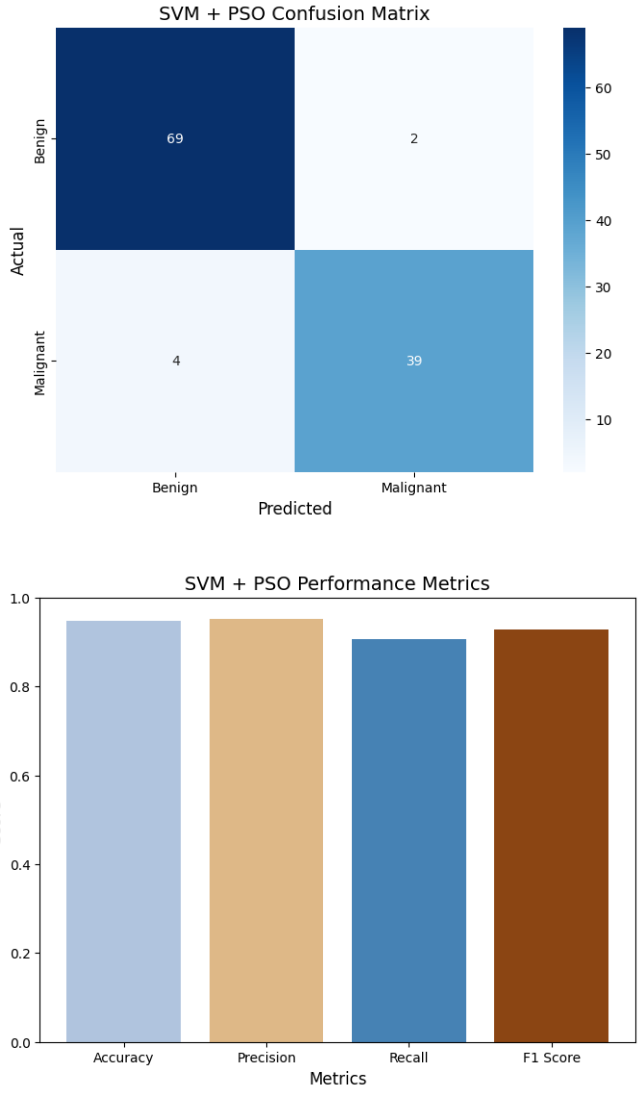


Figure 8: Support Vector Machine + PSO

Accuracy: 0.9473684210526315
Precision: 0.9512195121951219
Recall: 0.9069767441860465
F1 Score: 0.9285714285714286

Support Vector Machine + ACO

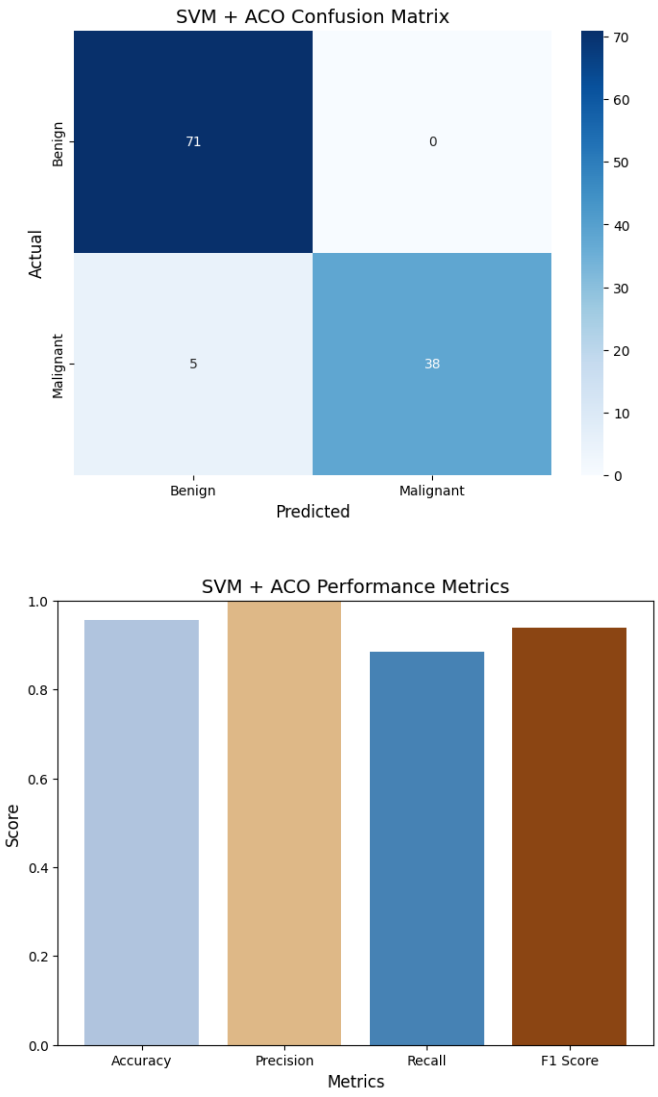


Figure 9: Support Vector Machine + ACO

Accuracy: 0.956140350877193

Precision: 1.0

Recall: 0.8837209302325582

F1 Score: 0.9382716049382716

Support Vector Machine + GA

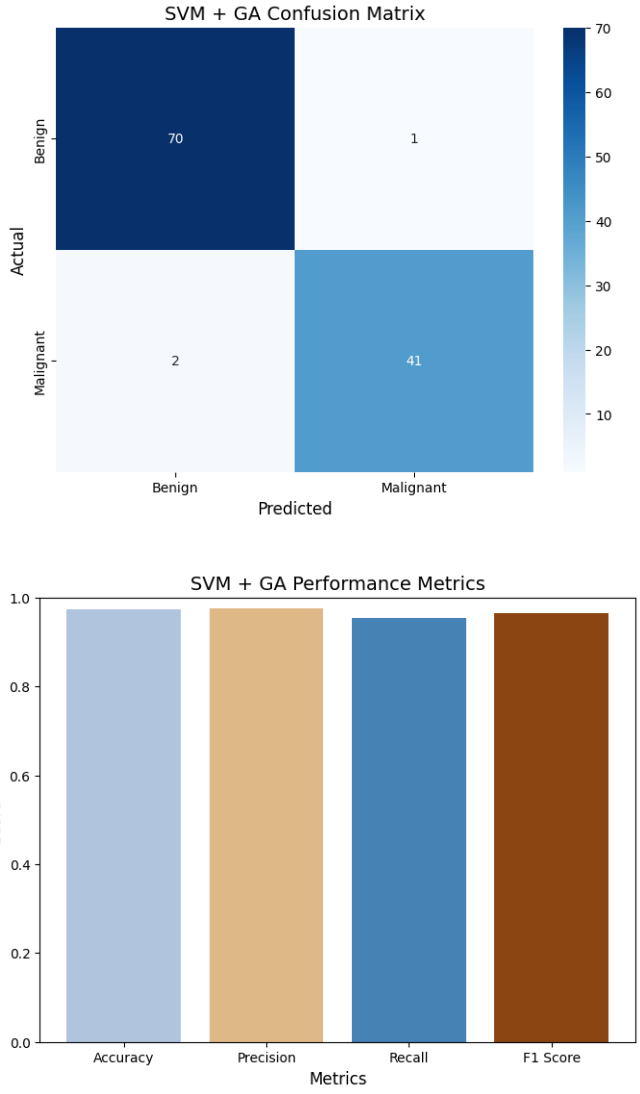


Figure 10: Support Vector Machine + GA

Accuracy: 0.9736842105263158
Precision: 0.9761904761904762
Recall: 0.9534883720930233
F1 Score: 0.9647058823529412

Support Vector Machine + DE

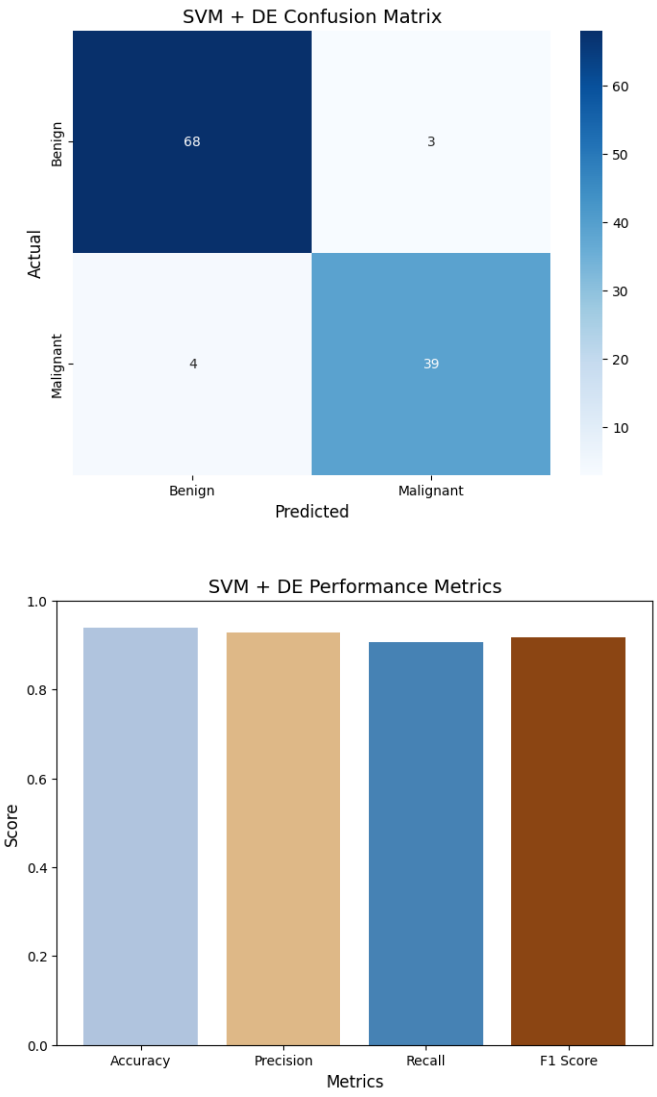


Figure 11: Support Vector Machine + DE

Accuracy: 0.9385964912280702
Precision: 0.9285714285714286
Recall: 0.9069767441860465
F1 Score: 0.9176470588235294

Support Vector Machine + WOA

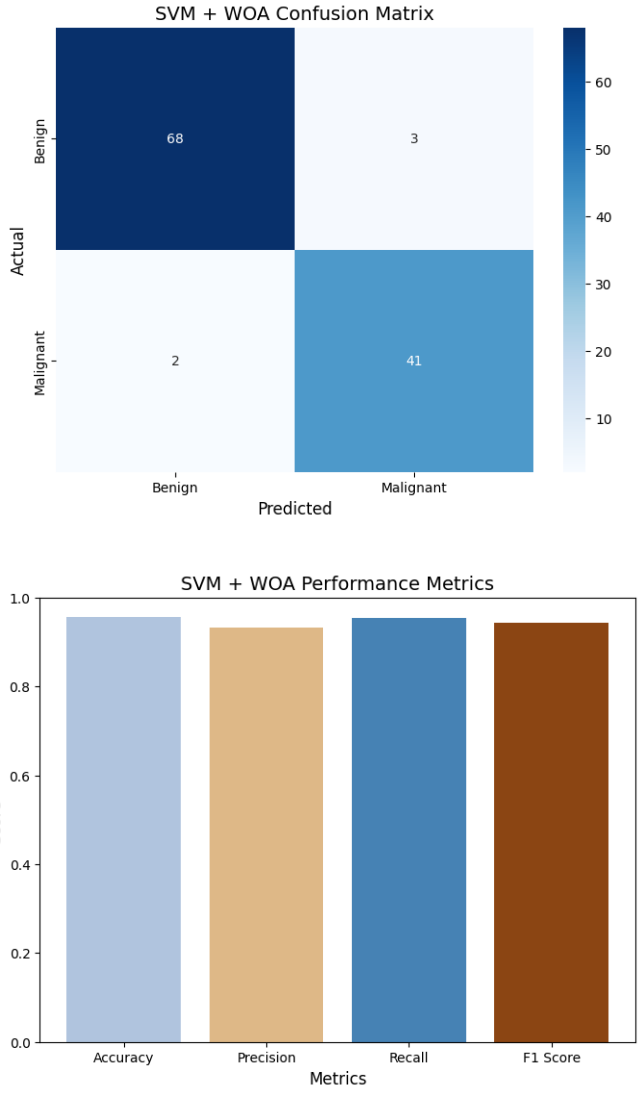


Figure 12: Support Vector Machine + WOA

Accuracy: 0.956140350877193
Precision: 0.9318181818181818
Recall: 0.9534883720930233
F1 Score: 0.9425287356321839

K-Nearest Neighbors

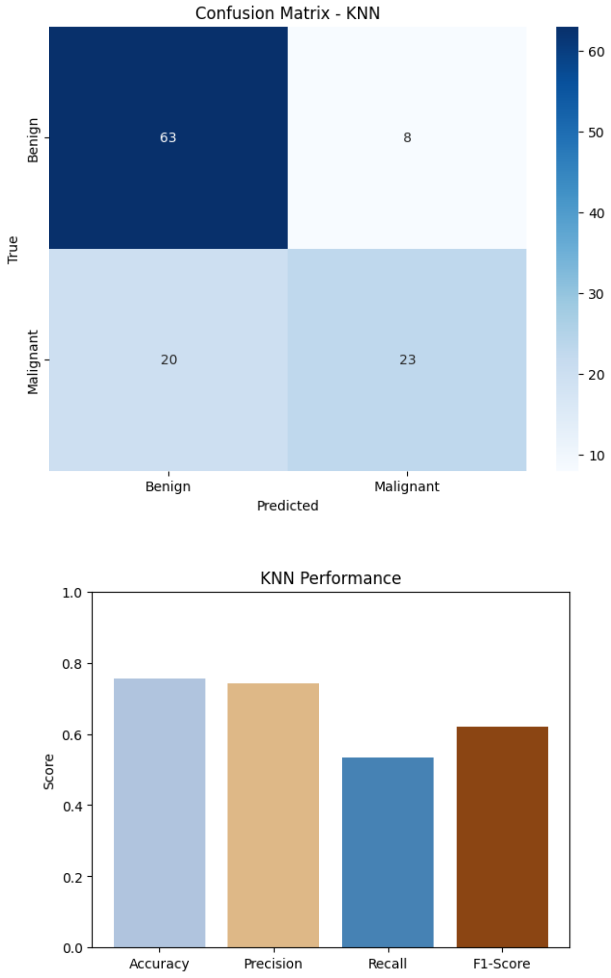


Figure 13: K-Nearest Neighbors

Accuracy: 0.7543859649122807
Precision: 0.7419354838709677
Recall: 0.5348837209302325
F1 Score: 0.6216216216216216

K-Nearest Neighbors + PSO

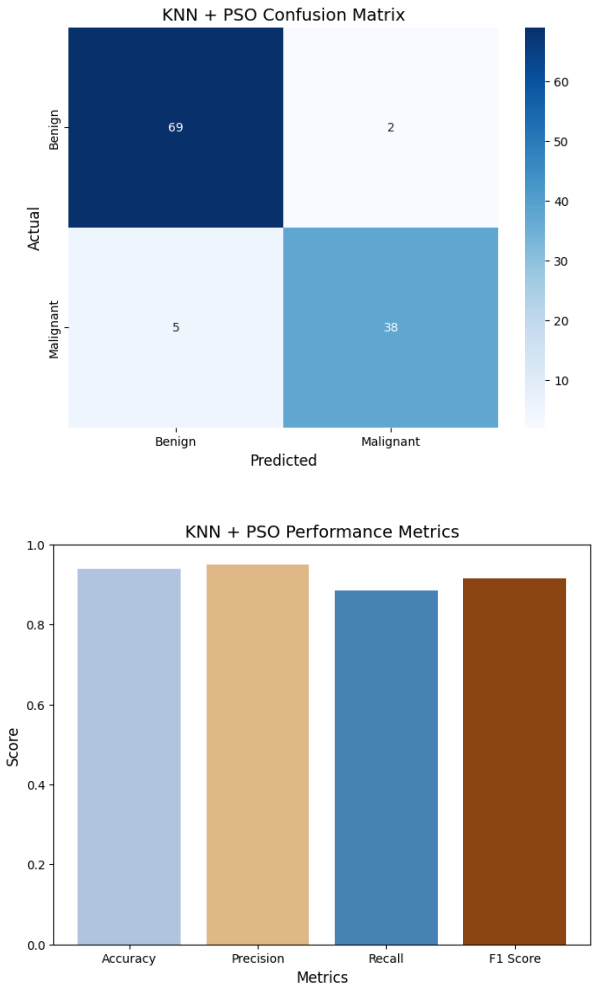


Figure 14: K-Nearest Neighbors + PSO

Accuracy: 0.9385964912280702

Precision: 0.95

Recall: 0.8837209302325582

F1 Score: 0.9156626506024096

K-Nearest Neighbors + ACO

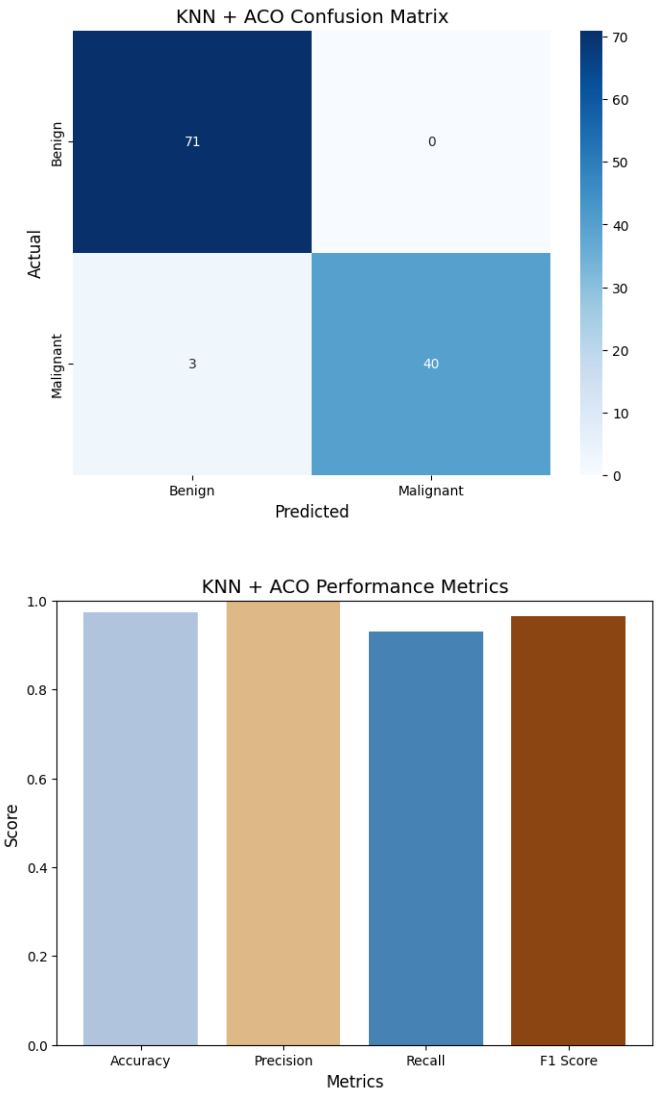


Figure 15: K-Nearest Neighbors + ACO

Accuracy: 0.9736842105263158

Precision: 1.0

Recall: 0.9302325581395349

F1 Score: 0.963855421686747

K-Nearest Neighbors + GA

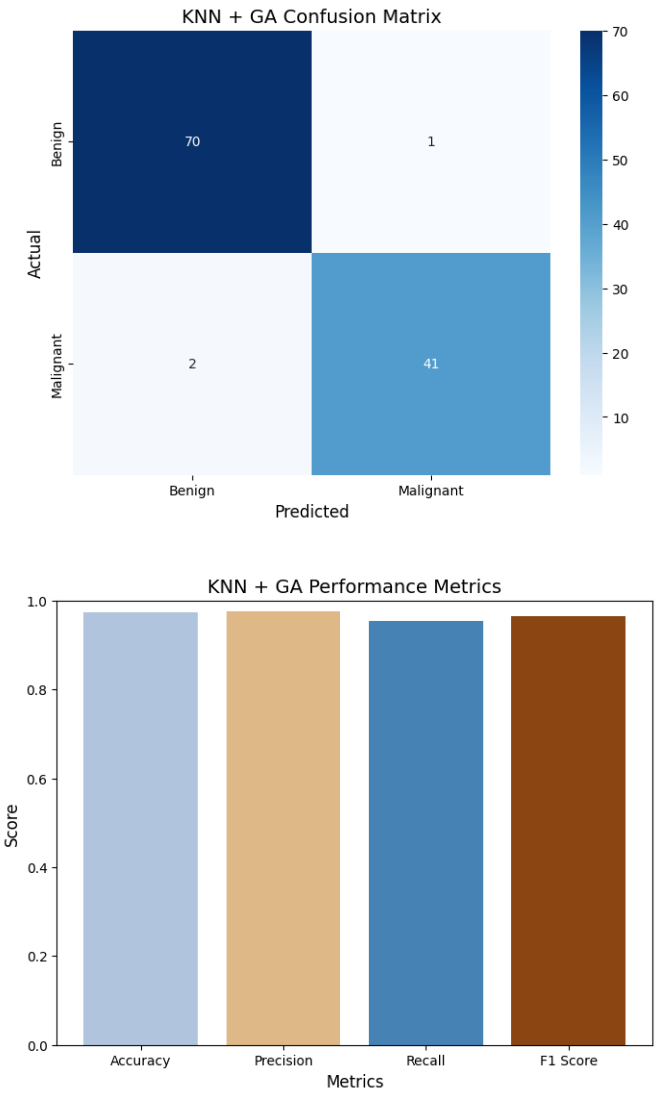


Figure 16: K-Nearest Neighbors + GA

Accuracy: 0.9736842105263158
Precision: 0.9761904761904762
Recall: 0.9534883720930233
F1 Score: 0.9647058823529412

K-Nearest Neighbors + DE

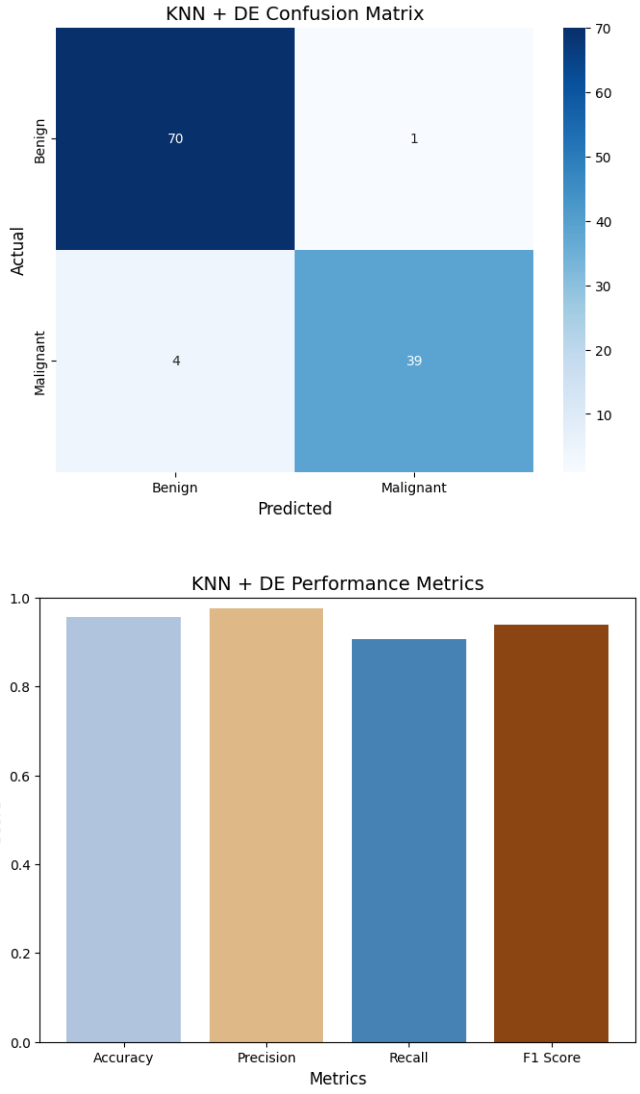


Figure 17: K-Nearest Neighbors + DE

Accuracy: 0.956140350877193

Precision: 0.975

Recall: 0.9069767441860465

F1 Score: 0.9397590361445783

K-Nearest Neighbors + WOA

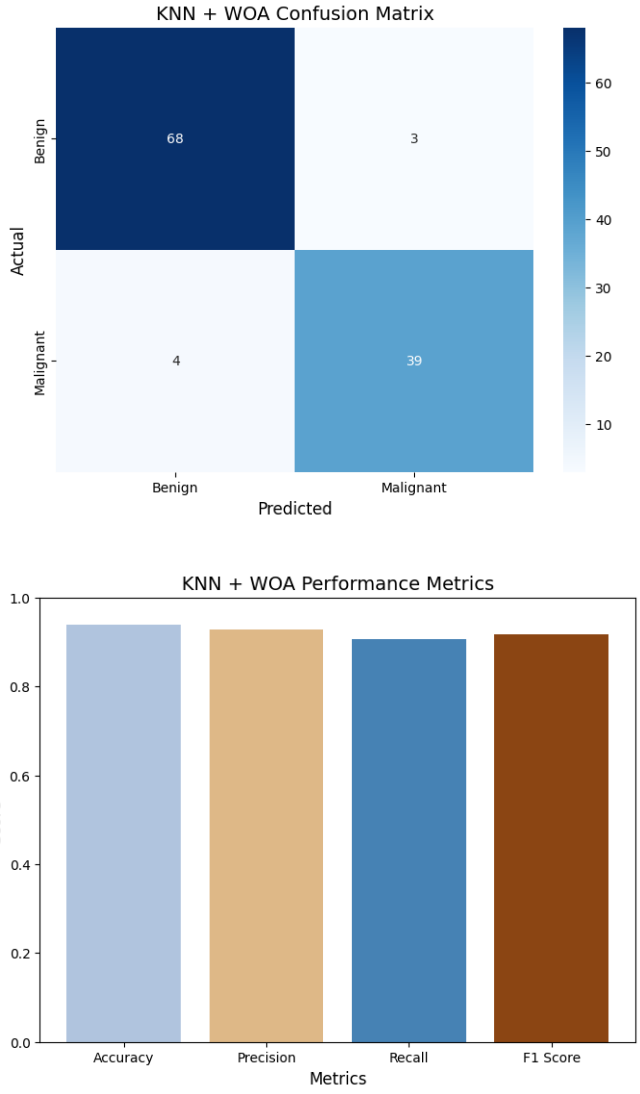


Figure 18: K-Nearest Neighbors + WOA

Accuracy: 0.9385964912280702
Precision: 0.9285714285714286
Recall: 0.9069767441860465
F1 Score: 0.9176470588235294

Random Forest

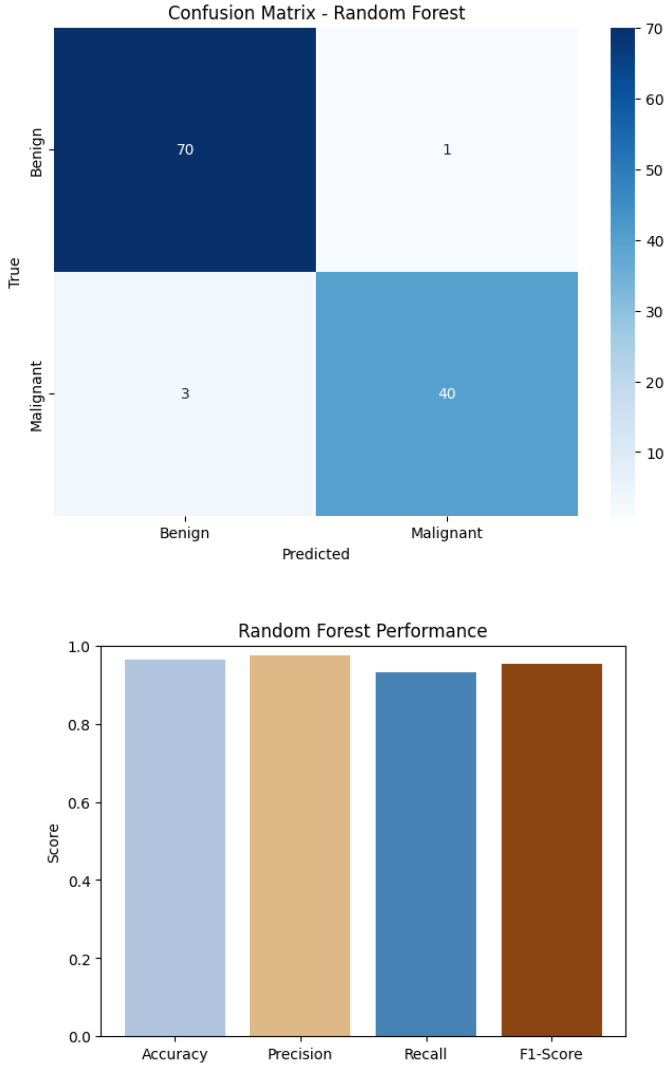


Figure 19: Random Forest

Accuracy: 0.9649122807017544
Precision: 0.975609756097561
Recall: 0.9302325581395349
F1 Score: 0.9523809523809523

Random Forest + PSO

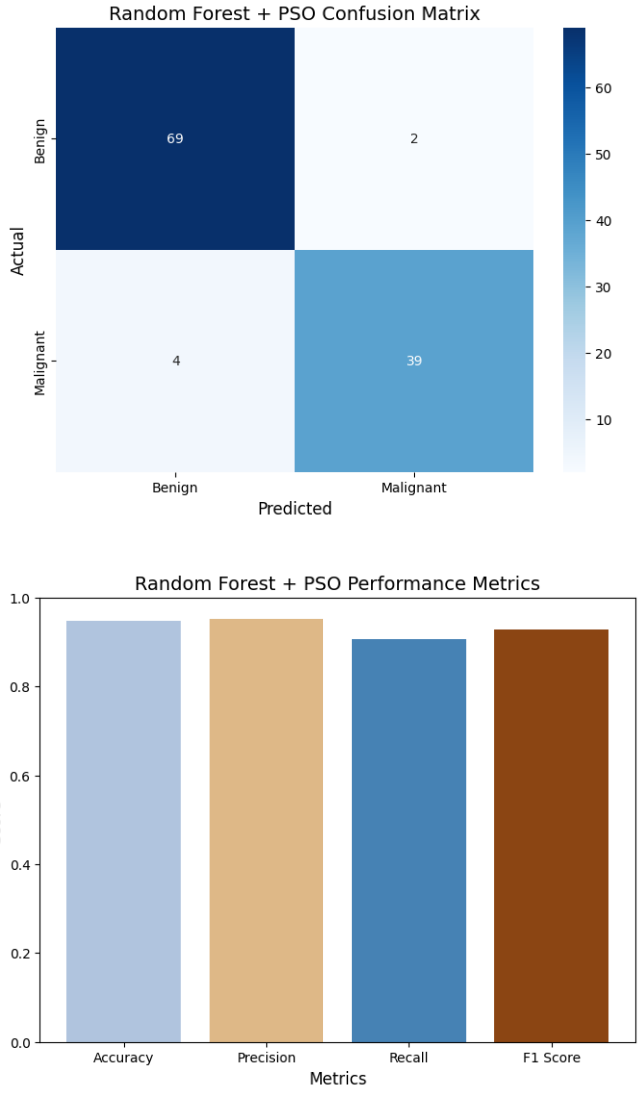
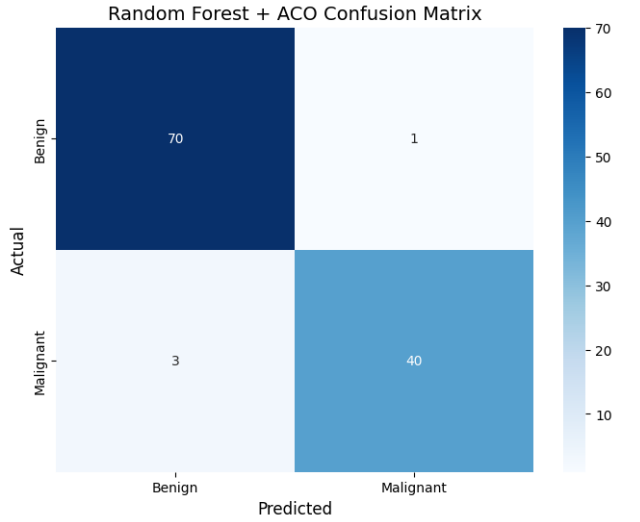


Figure 20: Random Forest + PSO

Accuracy: 0.9473684210526315
Precision: 0.9512195121951219
Recall: 0.9069767441860465
F1 Score: 0.9285714285714286

Random Forest + ACO



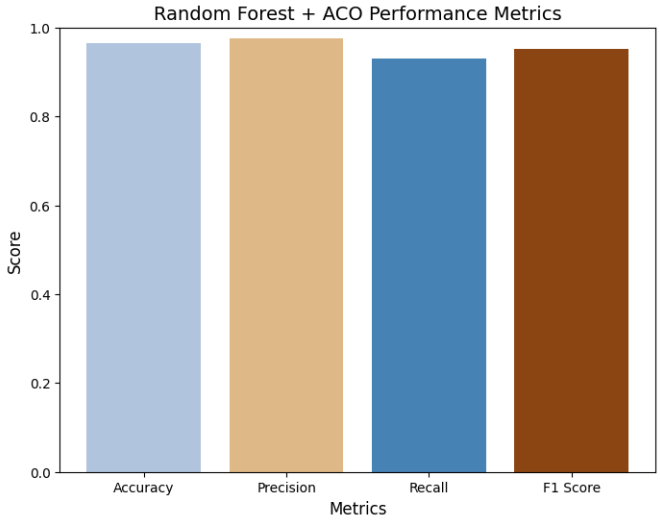


Figure 21: Random Forest + ACO

Accuracy: 0.9649122807017544
Precision: 0.975609756097561
Recall: 0.9302325581395349
F1 Score: 0.9523809523809523

Random Forest + GA

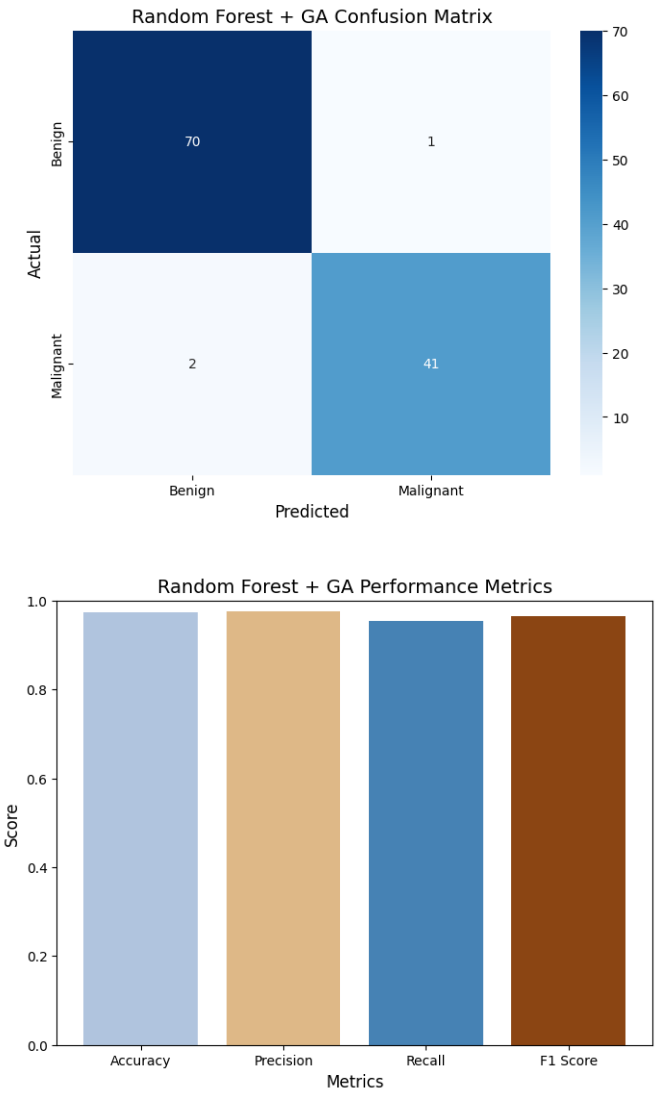


Figure 22: Random Forest + GA

Accuracy: 0.9736842105263158
Precision: 0.9761904761904762
Recall: 0.9534883720930233
F1 Score: 0.9647058823529412

Random Forest + DE

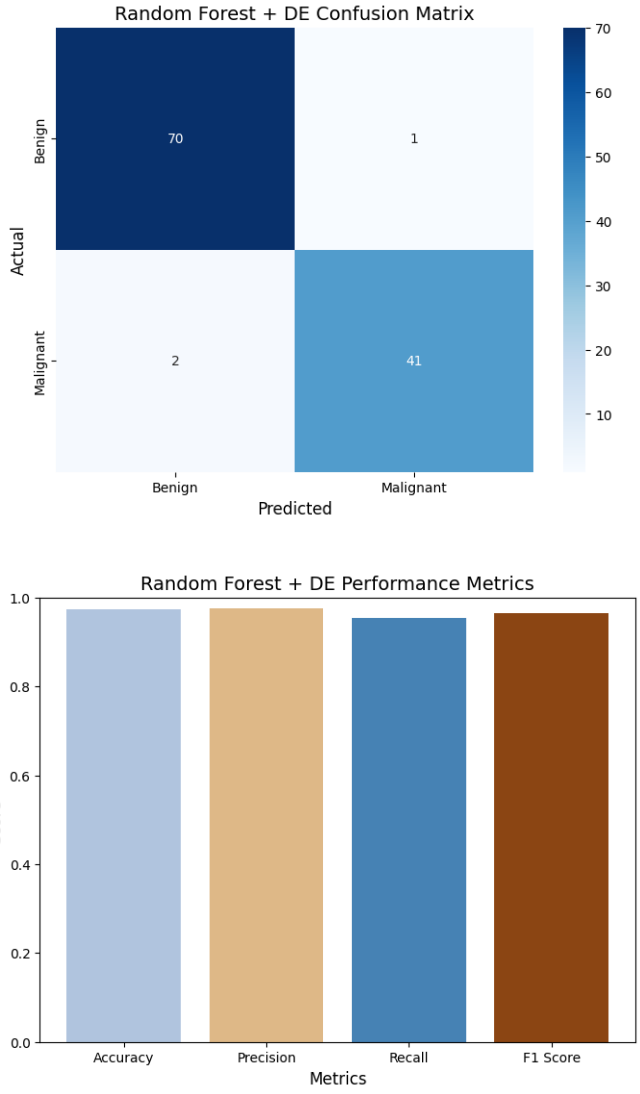


Figure 23: Random Forest + DE

Accuracy: 0.9736842105263158
Precision: 0.9761904761904762
Recall: 0.9534883720930233
F1 Score: 0.9647058823529412

Random Forest + WOA

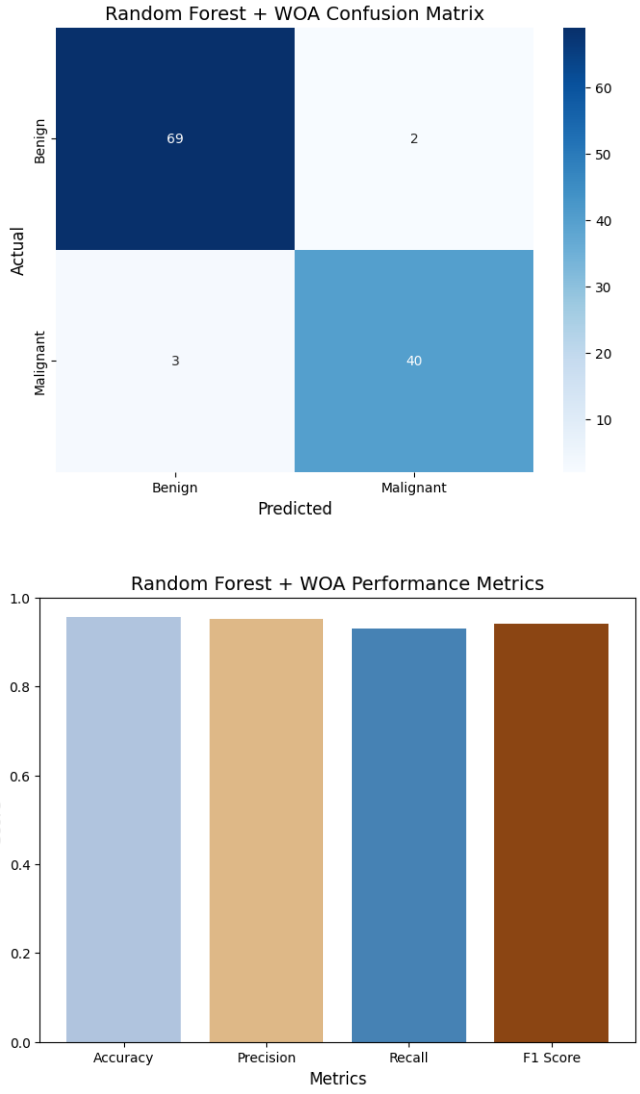


Figure 24: Random Forest + WOA

Accuracy: 0.956140350877193
Precision: 0.9523809523809523
Recall: 0.9302325581395349
F1 Score: 0.9411764705882353

4 Result Analysis

Accuracy

Row/Col	Logistic Regression	SVM	KNN	Random Forest
Normal	95.61%	62.28%	75.43%	96.49%
PSO	95.61%	94.73%	93.85%	94.73%
ACO	97.36%	95.61%	97.36%	96.49%
GA	98.24%	97.36%	97.36%	97.36%
DE	96.49%	93.85%	95.61%	97.36%
WOA	96.49%	95.61%	93.85%	95.61%

Table 2: Comparison of accuracies table

Precision

Row/Col	Logistic Regression	SVM	KNN	Random Forest
Normal	97.5%	0.0%	74.19%	97.56%
PSO	97.5%	95.12%	95%	95.12%
ACO	1.0%	1.0%	1.0%	97.56%
GA	97.67%	97.61%	97..61%	97.61%
DE	97.56%	92.85%	97.5%	97.61%
WOA	95.34%	93.18%	92.85%	95.23%

Table 3: Comparison of precision table

Recall

Row/Col	Logistic Regression	SVM	KNN	Random Forest
Normal	90.69%	0.0%	53.48%	93.02%
PSO	90.69%	90.69%	88.37%	90.69%
ACO	93.02%	88.37%	93.02%	93.02%
GA	97.67%	95.34%	95.34%	95.34%
DE	93.02%	90.69%	90.69%	95.34%
WOA	95.34%	95.34%	90.69%	93.02%

Table 4: Comparison of recall table

F1 Score

Row/Col	Logistic Regression	SVM	KNN	Random Forest
Normal	93.97%	0.0%	62.16%	95.23%
PSO	93.97%	92.85%	91.56%	92.85%
ACO	96.38%	93.82%	96.38%	95.23%
GA	97.67%	96.47%	96.47%	96.47%
DE	95.23%	91.76%	93.97%	96.47%
WOA	95.34%	94.25%	91.76%	94.11%

Table 5: Comparison of F1 Score table

5 Conclusion

The implementation and comparison of the performance of commonly used machine learning models, Logistic regression model, SVM model, KNN model and Random forest model along with evolutionary optimization algorithms, PSO, ACO, GA, DE AND WOA are done. Among these models, the combination of the Logistic regression model trained using the features obtained from Genetic Algorithm achieved the best results. The Logistic regression combined with GA model outperformed other models with 98.24% accuracy, 97.67% precision, 97.67% recall and 97.67% F1 score.

The exceptional performance of this approach can be attributed to the complementary strengths of the two techniques. Logistic Regression, known for its simplicity and interpretability, serves as a reliable classifier for binary classification tasks. Meanwhile, the Genetic Algorithm effectively reduces the feature space by selecting the most relevant subset of features, minimizing noise, and improving computational efficiency. Together, these methods provide a powerful framework for achieving optimal classification results.

This comparative study shows how different widely used classification models perform with the feature sets obtained using different evolutionary optimization algorithms. Even though some models gave 95% and more accuracies while trained using the full dataset, an improvement in those accuracies is also shown clearly.

Bibliography

[1] Leena Vig, *Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset*, 2014.

[2] Archit Aggarwal, Shubham Sharma, Tanupriya Choudhury, *Breast Cancer Detection Using Machine Learning Algorithms*, 2018.

[3] Anuj Vaghani, Manav Mangukiya, Meet Savani, *Breast Cancer Detection with Machine Learning*, 2022.

[4] Ahmed S. Elkorany, Khaled M. Almustafa, Mohamed Marey, Zeinab F. Elsharkawy, *Breast Cancer Diagnosis Using Support Vector Machines Optimized by Whale Optimization and Dragonfly Algorithms*, 2022.

[5] Pradeep Kumar Gupta, Poonam Rana, Vineet Sharma, *A Novel Deep Learning-based Whale Optimization Algorithm for Prediction of Breast Cancer*, 2020.

[6] Reza Rabiei, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, Alireza Atashi, *Prediction of Breast Cancer using Machine Learning Approaches*, 2022.

[7] Abebe Alemu Balcha, Samuel Alemu Woldie, *Impact of Genetic Algorithm for the Diagnosis of Breast Cancer: Literature Review*, 2023.

[8] Arslan Khalid, Arif Mehmood, Amerah Alabrah, Bader Fahad Alkhamees, Farhan Amin, Hussain AlSalman, Gyu Sang Choi, *Breast Cancer Detection and Prevention Using Machine Learning*, 2023..

[9] A.I. Diveev, S.V. Konstantinov, E.A. Sofronova, *A Comparison of Evolutionary Algorithms and Gradient-based Methods for the Optimal Control Problem*, 2018.