Enhancing Cancer Prediction: Investigating ACO-Based Feature Selection with Logistic Regression, XGBoost, and LightGBM

ASHWINI CHIDURALA
CSE(AIML)
SR UNIVERSITY
Warangal ,India
ashwinichidurala2004@gmail.com

SHASHANKA BASANI
CSE(AIML)
SR UNIVERSITY
Warangal ,India
shashankabasani5@gmail.com

PRANAVI MULASTAM
CSE(AIML)
SR UNIVERSITY
Jagtial ,India
mulastampranavi@gmail.com

SANJAY NERELLA
CSE(AIML)
SR UNIVERSITY
Jagtial ,India
sanjunerella9010@gmai.com

Abstract—**This report mainly concentrates on cancer prediction, particularly focusing on breast cancer prediction, which remains a significant challenge in healthcare. Improving patient outcomes requires early identification, and machine learning approaches have enough opportunities in this area to improve the predicting capabilities. Feature selection plays a vital role in developing accurate predictive models, and this study investigates the effectiveness of Ant Colony Optimization (ACO) in enhancing feature selection for three commonly used machine learning algorithms: Logistic Regression, XGBoost, and LightGBM. The report begins with an introduction highlighting the importance of early cancer detection and the role of machine learning techniques in achieving the goal. Then presents a comprehensive literature review, summarizing previous research paper studies on cancer detection and classification using various machine learning approaches, with and with out implementing ACO. The proposed methodology outlines the steps followed in the research process, from importing necessary libraries to evaluating the performance of machine learning models.**

Keywords—Ant Colony Optimization, Feature Selection, Nature Inspired Computing, Breast Cancer Prediction, Machine Learning.

I. INTRODUCTION

Cancer is the main disease which does not detected in early stages in now a days we are developing a model using machine learning techniques which detects the breast cancer in early stages. cancer is the major disease which kills the patients if it is not detected in early stages. Ant Colony Optimization (ACO) is developed as a powerful feature selection, ACO has taken inspiration from the behaviour of ants to efficiently navigate search spaces. Inspired by this natural process, ACO optimizes feature subsets for machine learning models, mainly in medical applications. Our research concentrates on exploring ACO with three commonly used machine learning algorithms for breast cancer prediction. We objective to analyse how ACO-based feature selection enhances the performance of these models in predicting breast cancer. By reading the Research papers worked on Breast Cancer dataset, we implemented ACO to choose the most effective features, with the goal of improving the accuracy, precision, recall, and F1-score of the Logistic Regression, XGBoost, and LightGBM models. By comparing the results of machine learning models with and without ACO-based feature selection, we aim to explain the influence of ACO on breast cancer prediction. This study contributes to the growing of research on optimization techniques in healthcare, with specific focus on the efficiency of machine learning models for early cancer detection. Our work seeks to provide valuable insights for doctors and researchers in oncology,. Ultimately, our goal is to develop the cancer research by Utilizing the capabilities of ACO and machine learning to benefit the patient care and outcomes.

II. LITERATURE SURVEY

[1] In this paper Rajesh Saturi1 team presents a methodology employing Convolutional Neural Networks (CNN), Local Ternary Pattern (LTP), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO) for breast cancer detection. Achieving an accuracy rate of 93.68%, the approach combines various techniques, promising more reliable diagnoses. However, it suffers from time-consuming processes and requires further validation. Future prospects include integration with medical imaging techniques for real-time applications.

[2] In this paper R Rajakumari team explores breast cancer detection utilizing Google Net, Alex Net, and Adam with an accuracy rate of 90.16%. While it introduces novelty in deep CNN architecture and emphasizes interpretability and clinical relevance, concerns arise regarding overfitting and the black box nature of the model. Future directions aim to address these issues by enhancing explainability and interpretability through multi-task learning.

[3] paper present by Oladele and team have comparative studied utilizing Support Vector Machine (SVM), k-Nearest Neighbours (KNN), PSO, ACO, and Random Forest, achieving an accuracy rate of 96.99%. The study focuses on informative features and reducing computational costs but is limited by data dependence and scalability issues. Future prospects include implementing explainable AI and early-stage detection.

[4] paper presented by Yamuna Prasad team proposed a methodology employing ACO, PSO, SVM, and Genetic Algorithm with an accuracy rate of 86.76% for siRNA design. While it enhances siRNA design and offers biological validation, concerns regarding potential data dependence and computational expenses are noted. Future directions include in vitro and in vivo validation and high-throughput screening integration.

[5] paper by D. P. Aldryan,Adiwijaya,Aditsania Annisa team introduces a method utilizing microarray data analysis, MBP-CGP, and ACO, achieving an accuracy rate of 64.12%. While it combines multiple techniques and emphasizes early detection, challenges arise in result interpretation and comparison. Future directions include multi-cancer classification and further validation using new microarray data.

III. PROPOSED METHODOLOGY

To jump-start our machine learning journey for breast cancer classifications, we started by importing the required libraries. The libraries are the backbone of our research, in our paper we are using different python machine learning libraries such as NumPy & Pandas for data manipulation Matplotlib & Seaborn for data visualization Scikit-learn for access to machine learning algorithms Once our importing libraries was ready, then we loaded the dataset. The Python libraries were helpful for dealing large breast cancer dataset, which was pre-cleaned & structured, saved us a lot of time and effort for cleaning data. Breast cancer dataset has rich in features and observations, which will play major role in model building. The next step is to make sure the data is clean. We checked for null values and anomalies which will affect our model analysis. Data integrity is very important, as it directly affects the fulfilment of the machine learning models. Once we have a clean dataset, next step is to move on to feature selection. In this step we apply a new technique called Ant Colony Optimization (ACO), which is inspired by the behavior of ants looking for the shortest route to the food chain If we identify the necessary features for our model accuracy has improved once we are ready to teach our model. After selecting best features using ant colony optimization we will apply three different models. Gradient Boosting Classifier Support Vector Machine Logistic Regression Each model has its own strengths, but together they are a powerful tool. We trained one set with the features selected by our ACO algorithm and another set with features selected through traditional machine learning techniques without selecting any features using ACO. This process allowed us to draw a parallel and evaluate the efficacy of ACO in feature selection. We calculated various metrics such as accuracy, F1-score, and precision for both sets of models. These indicators gave us a useful way to assess the effectiveness of our models. Lastly, we contrasted the outcomes with and without ACO. This comparison went beyond a cursory glance at data to include a more thorough examination of the ways in which the performance of our three distinct models was influenced by ACO-selected features. It was a turning point that demonstrated how ACO may improve machine learning tasks. To sum up, our methodology combined cutting-edge techniques with conventional models. We were able to obtain important insights into the optimization of machine learning models for the classification of breast cancer by combining ACO with standard feature selection techniques.

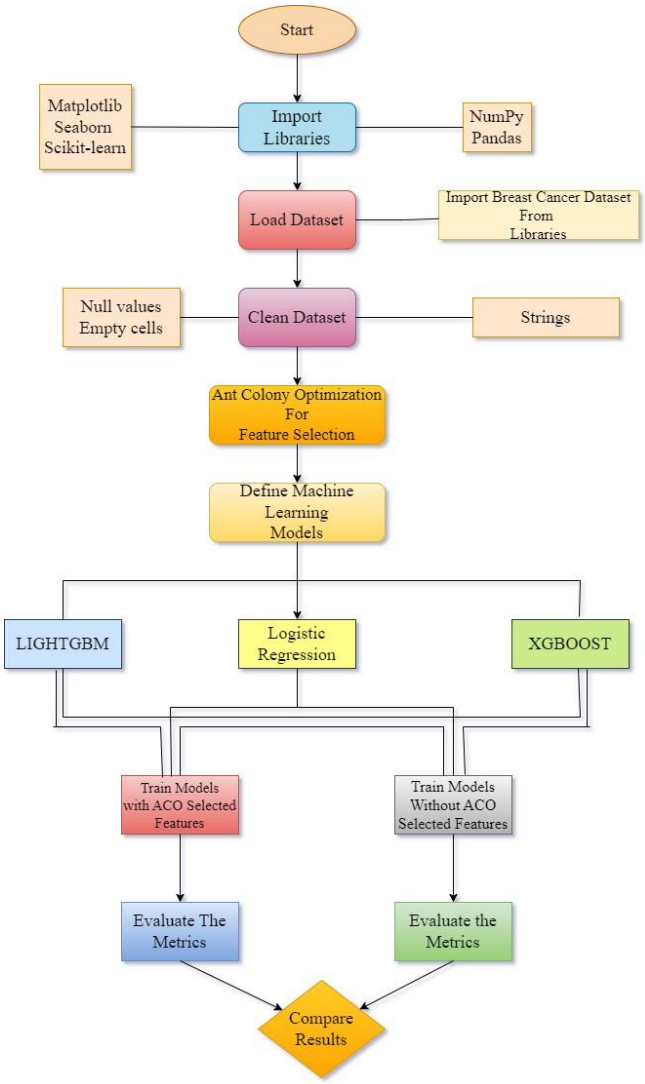

Fig-1- Flowchart of Proposed Methodology

A. *Ant Colony Optimization*

Ant Colony Optimization (ACO) is a interesting algorithmic exploration. Ant Colony Optimization technique is purely inspired from the foraging behavior of ant colonies. It's a mighty optimization technique used to solve combinatorial optimization problems, such as the traveling salesman problem and job scheduling. The algorithm imitates the way ants communicate through pheromone trails to find the shortest path from anthill to food sources.

Step 1: Initialization

Starting of Pheromone Trails:

Set pheromone trails to a constant value of $\tau 0$τ0 on each edge.

Formula: $\tau ij(0)$=τ0τij(0)=τ0

Initialization of Ant Positions:

Place mm ants randomly on nodes to start the search process.

Step 2: Ant Movement

Probability Calculation:

At each iteration, ants probabilistically choose the next node to move to base on pheromone levels and heuristic information.

Formula for calculating the probability of moving from node ii to node jj:

$$Pr(i,j) = \frac{\tau(i,j).[\eta(i,j)]\beta}{\sum all\ legal\ j\ \tau(i,j).[\eta(i,j)]\beta}$$

where $\alpha\alpha$ and $\beta\beta$ are parameters controlling the importance of pheromone and heuristic information, and NkNk is the set of feasible next nodes for the ant.

Step 3: Pheromone Update

Evaporation of Pheromone Trails:

Existing pheromone trails evaporate by a rate $\rho\rho$ after each iteration.

Formula: $\tau ij(t+1) = (1-\rho) \cdot \tau ij(t)$τij(t+1) =(1−ρ) ·τij(t)

Deposit of Pheromone:

Pheromone is deposited on edges traversed by ants.

Formula: $\tau ij(t+1) = \tau ij(t+1) + \Delta\tau ij$τij(t+1)=τij(t+1)+Δτij, where $\Delta\tau ij$Δτij is the amount of pheromone deposited.

Step 4: Termination

Termination Condition:

The process iterates until a termination condition is met, such as a maximum number of iterations or convergence criteria.

Note: The heuristic information ηijηij is defined as $1/dij$1/dij, where dijdij is the distance between nodes ii and jj, and $\beta=2\beta=2$. The parameter $\rho\rho$ controls the rate of pheromone evaporation. The amount of pheromone deposited on edges is determined by the pheromone update rule: $\tau ij = [\tau ij \cdot \rho] + \delta ij$τij =[τij·ρ]+δij, where δijδij is the amount of pheromone deposited.

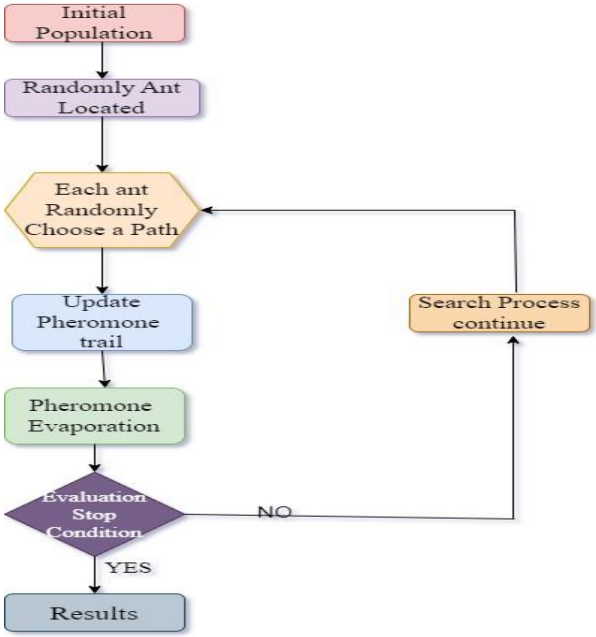

Fig-2- Flowchart of Ant Colony Optimization

B. Metrics for Calculations

Accuracy: One popular statistic for assessing a classification model's performance is accuracy. The ratio of accurately predicted instances to all instances in the dataset is measured.

Accuracy = $\frac{Correct\ predictions}{All\ predictions}$

Precision: Precision shows the proportion of correct positive predictions among all the positive predictions the model generates. The focus is on the degree of accuracy of optimistic projections.

Precision = $\frac{True\ Positive}{True\ Positive + False\ Positive}$

F1-score: The harmonic mean of recall and precision is the F1 score. It offers equal weight to both criteria, balancing recall and precision. It comes in very handy when working with unbalanced datasets.

F1 = 2. $\frac{precision\ .\ recall}{precison + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$

Covariance Matrix: A covariance matrix is a square matrix that summarizes the covariance relationships between variables in a dataset.

Compute "covariance matrix":

$\sum = \frac{1}{m}\sum_{i=1}^{n}\left(x^{(i)}\right)\left(x^{(i)}\right)^{T}$

XiXi and YiYi are individual data points of variables X and Y respectively.

$\bar{X}$X¯ and $\bar{Y}$Y¯ are the means of variables X and Y respectively.

nn is the total number of data points.

Correlation Matrix: A table that displays the correlation coefficients between variables in a dataset is called a correlation matrix. It shows how changes in one variable effect changes in another, offering insights into the linkages between many variables.

IV. RESULTS

A. Table

S.NO	MODEL	ACCURACY	PRECISION	F1-SCORE
1	LIGHT GBM (WITH ACO)	0.9736842105263158	0.9722222222222222	0.9790209790209791
2	LIGHT GBM (WITH OUT ACO)	0.9649122807017544	0.9589041095890411	0.9722222222222222
3	LOGISTIC REGRESSION (WITH ACO)	0.9649122807017544	0.9589041095890411	0.9722222222222222
4	LOGISTIC REGRESSION (WITH OUT ACO)	0.9561403508771931	0.9459459459459459	0.9655172413793103
5	XGBOOST (WITH ACO)	0.9649122807017544	0.9589041095890411	0.9722222222222222
6	XGBOOST (WITH OUT ACO)	0.9561403508771931	0.9583333333333334	0.9650349650349651

TABLE-1-Comparision of results of models with and without ACO

In order to compare the performance of the LightGBM, Logistic Regression, and XGBoost models in predicting breast cancer, Ant Colony Optimization (ACO) for feature selection was used in table-1. Applying ACO enhances predictive performance consistently across all models; greater outcomes for accuracy, precision, and F1-score value are observed when ACO is used. LightGBM always performs better than XGBoost and Logistic Regression models. Notably, the addition of ACO dramatically improves performance measures, even if Logistic Regression and XGBoost perform rather consistently in scenarios with and without ACO. These results reveal that ACO can enhance feature selection, boosting the overall precision of machine learning models in predicting cancer. Significantly, LightGBM surpasses its competitors in effectiveness.

B. Graphs and Matrices

1) Correlation matrix: When employing Ant Colony Optimization (ACO) for selecting features, the correlation matrix reveals how selected traits are linked to a target variable like cancer. A positive correlation indicates a direct relationship, while a negative one suggests an opposing connection..This potentially enhances model performance by assisting in the identification of the most pertinent elements for predictive modelling.

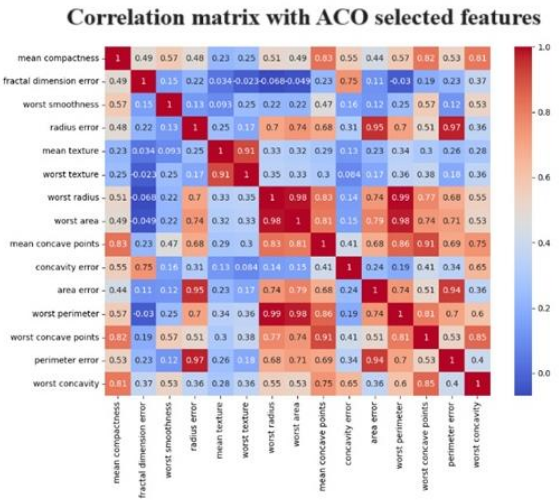

Fig-3- Correlation Matrix with ACO selected features

Without Applying ACO: All attributes, regardless of their importance, may be included in the correlation matrix if ACO is not applied. Consequently, the model might be muddled with irrelevant information or noise, potentially diminishing its predictive precision. Without ACO, while the correlation matrix offers a broader analysis of feature relationships, it struggles to effectively prioritize the most crucial features for accurate forecasting.

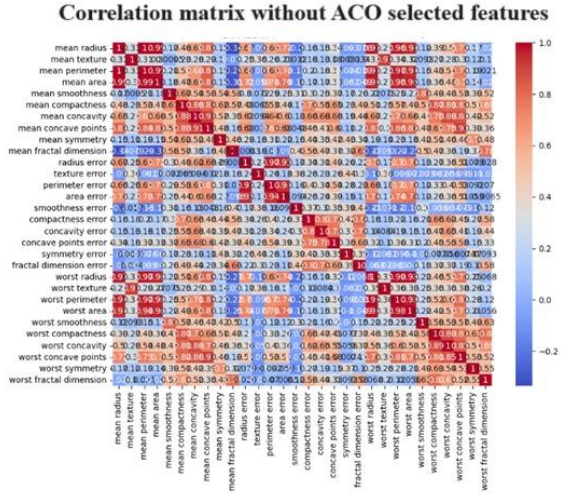

Fig-4-Correlation Matrix without ACO selected features

2) Confusion Matrix: The confusion matrix acts like a detailed map for navigating the landscape of model predictions, highlighting paths taken by true positives and false negatives alike. It sheds light on the accuracy, precision, recall, and F1-score when ACO is in play. When features selected by ACO are used, it's as if one has switched to a sharper lens on a camera; suddenly everything comes into clearer focus suggesting an enhanced ability to classify correctly.

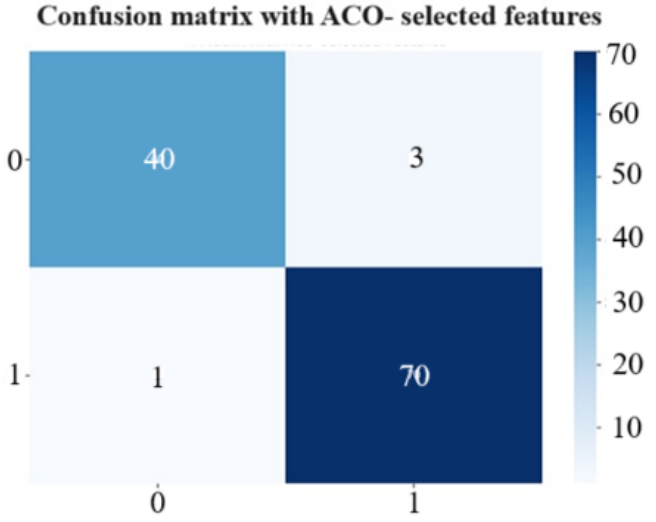

Fig-5- confusion matrix with ACO selected features

When ACO is not used, the confusion matrix still assesses a model's performance but might include extraneous or disruptive elements. This can impair the model's precision in classifying events, altering key metrics such as accuracy, precision, recall, and F1-score.

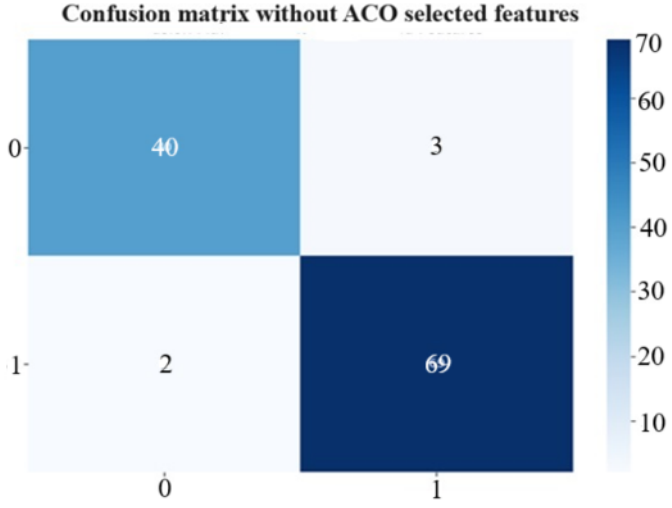

Fig-6- Confusion Matrix without ACO selected features

3) Comparision of Accuracy: Using Ant Colony Optimization (ACO) and without it, this graph contrasts the accuracy of several machine learning models. A bar is used to represent each model, and two bars are used to indicate scenarios with and without ACO for each model. Among the models, LightGBM consistently obtains the highest accuracy, followed by XGBoost and Logistic Regression. Moreover, in the case of each model, the scenario that applies ACO typically yields a higher accuracy than the scenario that does not. The vertical distance between the bars for each model shows this variation in accuracy.

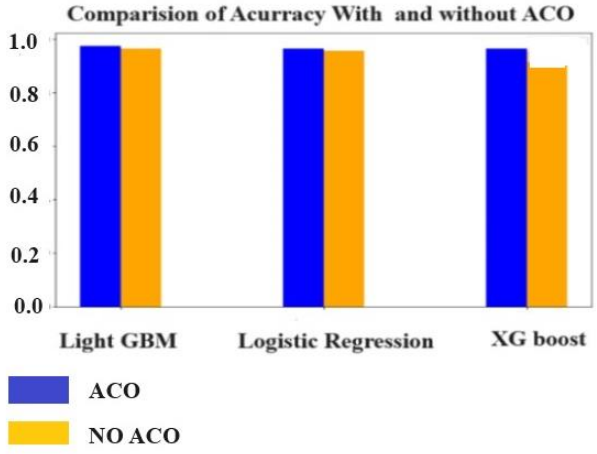

Fig-7- Comparison of Accuracy with and without ACO

Comparision of Precision: The precision of a model indicates its ability to calculate the proportion of true positive predictions among all positive predictions in order to prevent false positives. When ACO is implemented, we find that, generally speaking, the precision is marginally higher than in situations when ACO is not used. The maximum precision is achieved by LightGBM with ACO, closely followed by XGBoost and Logistic Regression.

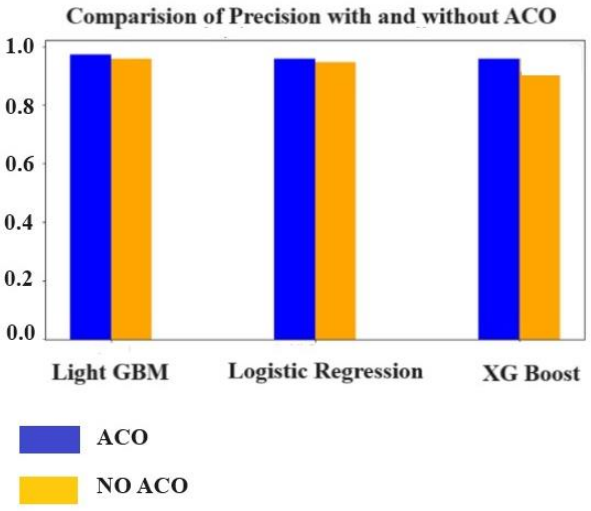

Fig-8-Comparision of Precision with and without ACO

4) Comparision of F1-Score: This graph contrasts the F1-scores of many machine learning models with and without the use of an ACO. Every model is represented by a bar, same like in the previous graphs; two bars for each model indicate possibilities with and without ACO. The F1-score provides a balance between precision and recall by being the harmonic mean of the two metrics. When ACO is applied, we typically find that the F1-score is marginally higher than in scenarios without ACO. Among the models, LightGBM with ACO regularly has the greatest F1-score, demonstrating its balanced precision and recall ability.

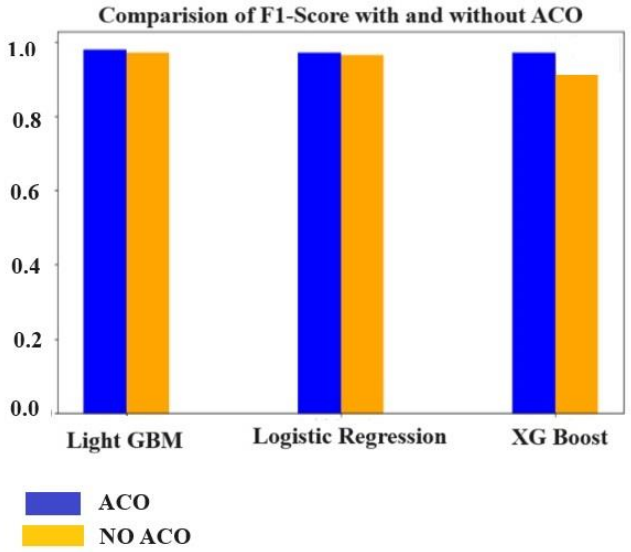

Fig-9- Comparison of F1-Score with and without ACO

V. CONCLUSION AND FUTURE SCOPE

Our research delves into the realm of cancer prognosis, a field where the early detection of disease is akin to spotting a storm on the horizon—vital for preparation and survival. We focused on enhancing diagnostic accuracy for breast cancer by marrying Ant Colony Optimization (ACO) with three established machine learning techniques: XGBoost, LightGBM, and Logistic Regression. Utilizing data from the Breast Cancer dataset allowed us to assess how ACO-driven feature selection could sharpen model performance in terms of accuracy, precision, and F1-score. The results were illuminating; ACO acted like a skilled sculptor chiseling away at marble to reveal more defined forms beneath—the models' capabilities were significantly enhanced across all metrics. Notably, LightGBM emerged as a champion gladiator besting its competitors in this arena of predictive prowess. Imagine walking through an orchard where each tree represents different data points; ACO serves as an adept gardener deciding which trees bear fruit worth harvesting—a process that not only improves yield but ensures quality. This strategic pruning helped demonstrate that when it comes to predicting life-altering diseases like cancer, refining our tools can lead directly to saving lives. Looking ahead offers exciting possibilities: tweaking ACO's settings might further refine these diagnostic instruments or exploring its utility across various types of cancers could broaden its applicability in oncology diagnostics. Moreover, integrating other optimization algorithms with ACO may forge even stronger shields against misdiagnosis—an ensemble approach creating robust defenses against this relentless foe. In essence our study contributes crucial brushstrokes to the evolving masterpiece that is medical technology optimization—each stroke making future predictions more precise and reliable thereby transforming patient care landscapes dramatically.

REFRENCES-

[1] Saturi, R., & Premchand, P. (2021). Multi-Objective Feature Selection Method by Using ACO with PSO Algorithm for Breast Cancer Detection. *International Journal of Intelligent Engineering & Systems*, *14*(5).

[2] Rajakumari, R., & Kalaivani, L. (2022). Breast Cancer Detection and Classification Using Deep CNN Techniques. *Intelligent Automation & Soft Computing*, *32*(2).

[3] Oladele, T. O., Olorunsola, B. J., Aro, T. O., Akande, H. B., & Olukiran, O. A. (2021). Nature-inspired meta-heuristic optimization algorithms for breast cancer diagnostic model: A comparative study.

[4] Prasad, Y., Biswas, K. K., & Jain, C. K. (2010). SVM classifier-based feature selection using GA, ACO and PSO for siRNA design. In *Advances in Swarm Intelligence: First International Conference, ICSI 2010, Beijing, China, June 12-15, 2010, Proceedings, Part II 1* (pp. 307-314). Springer Berlin Heidelberg.

[5] Aldryan, D. P., & Annisa, A. (2018, November). Cancer detection based on microarray data classification with ant colony optimization and modified backpropagation conjugate gradient Polak-Ribiére. In *2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA)* (pp. 13-16). IEEE.