An Ensemble Learning Approach for Cancer Detection

Ratnesh Kumar Sharma[1], Prof. (Dr.) Satya Singh[2]

Department of Computer Science and Engineering, Mahatma Gandhi Kashi Vidyapith, Varanasi, Uttar Pradesh, India

DOI: https://doi.org/10.51244/IJRSI.2026.1305000234

Received: 01 June 2026; Accepted: 07 June 2026; Published: 11 June 2026

ABSTRACT

Cancer disease classification using high dimensional microarray datasets has become an important research area in healthcare analytics, bioinformatics, and intelligent clinical decision support systems because conventional machine learning approaches frequently experience challenges related to feature redundancy, noisy attributes, overfitting, computational complexity, and reduced predictive stability. This research paper presents an efficient hybrid and ensemble machine learning framework for accurate cancer disease classification using binary and multiclass cancer microarray datasets. The proposed framework integrates advanced feature selection techniques including Recursive Feature Elimination, Maximum Relevance Minimum Redundancy, Boruta, Correlation Feature Selection, and Principal Component Analysis with metaheuristic optimization algorithms such as Ant Colony Optimization, Particle Swarm Optimization, Improved Grey Wolf Optimization, Ant Lion Optimization, and Salp Swarm Optimization for identifying the most informative gene expression features and reducing dimensionality. Furthermore, multiple machine learning classifiers including Support Vector Machine, Random Forest, AdaBoost, XG Boost, Extreme Learning Machine, and ensemble voting approaches are incorporated to improve predictive reliability, robustness, and generalization capability. Experimental analysis performed on lung cancer, colon cancer, prostate cancer, leukemia, breast cancer, ALL-AML, lymphoma, and SRBCT microarray datasets demonstrated significant improvements in classification accuracy, sensitivity, specificity, precision, recall, Matthews Correlation Coefficient, and F1 score compared with conventional machine learning classifiers. The proposed hybrid ensemble framework effectively minimizes misclassification, enhances feature optimization, improves classification stability, and provides a reliable computational approach for intelligent cancer diagnosis, healthcare analytics, and precision clinical decision support systems [1], [2].

Index Terms: Cancer Disease Classification, Microarray Dataset, Machine Learning, Ensemble Learning, Feature Selection, Optimization Algorithms, Healthcare Analytics

INTRODUCTION

Cancer remains one of the leading causes of mortality worldwide, and early as well as accurate diagnosis plays a crucial role in improving patient survival rate, treatment planning, and healthcare outcomes. Recent advancements in microarray technology have enabled researchers to analyze thousands of gene expression features simultaneously for detecting complex cancer related abnormalities and molecular patterns. However, microarray datasets generally contain high dimensional features, noisy attributes, redundant information, irrelevant genes, and limited sample sizes, which reduce the effectiveness, stability, and generalization capability of conventional machine learning classifiers [3], [4]. These challenges increase computational complexity and may lead to overfitting and reduced prediction accuracy in cancer diagnosis systems. Therefore, intelligent hybrid and ensemble machine learning frameworks integrated with feature selection and optimization techniques are required to improve classification performance, predictive reliability, dimensionality reduction, and clinical decision support capability for efficient cancer disease diagnosis and healthcare analytics [5].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

LITERATURE REVIEW

Several researchers have proposed machine learning and ensemble learning approaches for cancer disease classification using high dimensional microarray datasets. Feature selection techniques such as Recursive Feature Elimination, Principal Component Analysis, Maximum Relevance Minimum Redundancy, Boruta, Information Gain, and Correlation Feature Selection have been widely utilized for identifying significant gene expression features and reducing dataset dimensionality [6], [7]. In addition, metaheuristic optimization algorithms including Ant Colony Optimization, Particle Swarm Optimization, Grey Wolf Optimization, Ant Lion Optimization, and Salp Swarm Optimization demonstrated strong capability for selecting optimal feature subsets and improving classification efficiency [8], [9]. Researchers also integrated machine learning classifiers such as Support Vector Machine, Random Forest, AdaBoost, XGBoost, K Nearest Neighbor, and Extreme Learning Machine for enhancing predictive performance and classification stability. Furthermore, ensemble learning approaches based on majority voting and weighted averaging significantly improved robustness, generalization capability, and diagnostic reliability for intelligent healthcare prediction and cancer diagnosis systems [10], [11].

Problem Statement

Existing machine learning approaches experience difficulties handling high dimensional microarray datasets and noisy gene expression features. Several approaches focus only on binary cancer classification while multiclass classification remains challenging. Existing frameworks do not sufficiently integrate feature selection, optimization algorithms, ensemble learning, and Explainable AI. Therefore, efficient and scalable machine learning frameworks are required for reliable cancer disease classification.

Research Objectives

1. To analyze cancer microarray datasets using machine learning techniques.
2. To develop hybrid machine learning models for cancer disease classification.
3. To integrate optimization algorithms with feature selection techniques.
4. To improve classification accuracy, specificity, sensitivity, precision, and F1 score.
5. To develop ensemble learning approaches for binary and multiclass cancer diagnosis.

PROPOSED METHODOLOGY

The proposed framework integrates advanced feature selection techniques, metaheuristic optimization algorithms, and ensemble learning methods for efficient cancer disease classification using microarray datasets. Initially, preprocessing and normalization techniques are applied to remove noisy records, redundant attributes, irrelevant gene expression features, and inconsistencies present in the datasets. After preprocessing, feature selection is performed using Recursive Feature Elimination, Maximum Relevance Minimum Redundancy, Boruta, and Correlation Feature Selection methods to identify the most informative and discriminative gene expression features. Subsequently, optimization algorithms including Ant Colony Optimization, Particle Swarm Optimization, Improved Grey Wolf Optimization, Ant Lion Optimization, and Salp Swarm Optimization are applied to obtain optimal feature subsets and reduce dimensionality while preserving classification capability [12], [13]. Finally, multiple machine learning classifiers such as Support Vector Machine, Random Forest, AdaBoost, XGBoost, K Nearest Neighbor, and Extreme Learning Machine are integrated through ensemble voting and weighted averaging techniques to improve classification accuracy, predictive stability, robustness, and generalization capability for intelligent cancer diagnosis and healthcare decision support systems [14], [15].

RESULTS AND DISCUSSION

Experimental analysis demonstrated that the proposed hybrid and ensemble models achieved superior classification performance compared with conventional machine learning classifiers. The RFE-ACO-RF model achieved 97.8 percent classification accuracy, while the PMP-SVM model achieved 98.2 percent F1 score with improved specificity and sensitivity. Ensemble voting methods demonstrated improved classification

robustness across multiple cancer datasets [16], [17]. The proposed feature optimization approaches significantly reduced dimensionality and improved classifier stability for healthcare prediction systems [18].

CONCLUSION AND FUTURE SCOPE

This research paper presented hybrid and ensemble machine learning approaches for efficient cancer disease classification using microarray datasets. Experimental analysis confirmed that the proposed models significantly outperformed conventional classifiers in terms of accuracy, specificity, sensitivity, precision, and F1 score. Future research may focus on integrating deep learning architectures, Explainable AI frameworks, and intelligent clinical decision support systems for precision healthcare applications [19], [20].

Table I. Cancer Microarray Dataset Description

Dataset	Samples	Features	Classes
Lung Cancer	62	2178	2
Colon Cancer	63	2000	2
Prostate Cancer	103	341	2
ALL AML	72	7129	3
SRBCT	83	2308	4

Table II. Comparative Performance Analysis

Model	Accuracy	Sensitivity	Specificity	F1 Score
RFE-ACO-RF	97.8%	96.9%	95.4%	96.9%
PMP-SVM	98.2%	98.3%	98.1%	98.2%
CFS-IGWO	96.8%	95.8%	95.2%	95.7%
BIMSSA	97.5%	97.1%	96.8%	97.0%

Table III. Feature Reduction Analysis

Dataset	Original Features	After Feature Selection	Optimized Features
Lung Cancer	12533	1250	320
Colon Tumor	2000	850	200
Breast Cancer	570	217	95
Leukemia	7129	1500	410

REFERENCES

1. R. H. Abiyev and S. Abizade, Diagnosing Parkinson diseases using fuzzy neural system, 2016.
2. T. H. H. Aldhyani et al., Soft clustering for chronic disease diagnosis, 2020.

3. Ding C. et al., Maximum Relevance Minimum Redundancy feature selection, 2005.
4. Rudnicki and Kursa, Boruta feature selection algorithm, 2010.
5. Sun L. et al., Machine learning in healthcare analytics, 2021.
6. Blessie et al., Correlation Feature Selection techniques, 2012.
7. Khaire U. M. et al., ReliefF feature selection algorithm, 2022.
8. Dorigo M. et al., Ant Colony Optimization, 2004.
9. Kennedy J. and Eberhart R., Particle Swarm Optimization, 1995.
10. Breiman L., Random Forests, 2001.
11. Dietterich T., Ensemble learning methods, 2000.
12. Mirjalili S., Grey Wolf Optimizer, 2014.
13. Mirjalili S. et al., Salp Swarm Algorithm, 2017.
14. Cortes C. and Vapnik V., Support Vector Networks, 1995.
15. Chen T. and Guestrin C., XGBoost scalable tree boosting system, 2016.
16. Sharma R. K. et al., Hybrid cancer classification model using RFE-ACO-RF, 2025.
17. Sharma R. K. et al., PMP-SVM model for cancer diagnosis, 2025.
18. Liu H. and Motoda H., Feature selection for knowledge discovery, 1998.
19. Lundberg S. and Lee S., SHAP explainable AI framework, 2017.
20. Ribeiro M. et al., Explaining predictions of any classifier, 2016.