



**ISSN: 2454-9940**



**INTERNATIONAL JOURNAL OF APPLIED  
SCIENCE ENGINEERING AND MANAGEMENT**

**E-Mail :**  
**editor.ijasem@gmail.com**  
**editor@ijasem.org**

**[www.ijasem.org](http://www.ijasem.org)**

# A COMPARATIVE ANALYSIS OF HYBRIDIZED GENETIC ALGORITHM IN PREDICTIVE MODELS OF BREAST CANCER TUMORS

Gajjela Soumya<sup>1</sup>, Darma Sathwika<sup>2</sup>,  
Peechu Nanditha<sup>3</sup>, Dr. K. Niranjana Reddy<sup>4</sup>  
<sup>1,2,3</sup> UG Student, Dept. of ECE, CMR Institute of Technology, Hyderabad  
<sup>4</sup> Associate Professor, Dept. of ECE,  
CMR Institute of Technology, Hyderabad

## ABSTRACT

Advancement in computer-aided tools towards accurate breast cancer early prediction models has proven to be advantageous, which in turn helps to reduce the mortality rate associated with this cancer. From the literature, random forest predictor has been observed to have high accuracy in comparison to other machine learning regressors, also genetic algorithm has been observed to be a good feature selection method in data pre-processing. In a bid to improve the accuracy of breast cancer predictive models, several studies have developed hybridized genetic algorithm models for feature selection, however, the order of hybridization may not have been taken into consideration, as this can have an impact on the hybridized model's performance. Therefore, this paper proposes several high-performing predictive models using hybridized genetic algorithm, based on other learning models, while taking into consideration the placement order of the feature selection

algorithms in the hybridized models. The Wisconsin Breast Cancer dataset was used as the test bench, while filter, wrapper and embedded feature selection algorithms were used in the proposed hybridized models. The performances of proposed hybridized models were compared with those of the individual learning models, considered in this work. These models include Fisher\_Score, Mutual Information Gain, Correlation Chi-square test, Coefficient, Variance, Genetic Algorithm, Lasso and Linear Regressors with L1 regularization, Ridge Regressor with L2 regularization, Tree-based methods. From the performance evaluation results, the proposed hybridized Genetic Algorithm with Fisher\_Score (GA + Fisher\_Score) model showed promising results, as it had an accuracy score of 99.12%, thereby outperforming other proposed hybridized genetic algorithm models considered.

## INTRODUCTION

Over the last few decades, Cancer has been known, universally, to be a deadly disease.

In a 2022 USA cancer report by the American Cancer Society, heart disease was ranked as the most common disease closely followed by cancerous diseases. From the report, the total number of expected new cancer cases recorded was almost 2 million with over 609 thousand expected deaths [1]. The good news is that within the last 30 years, scientists have done tremendous work to help reduce the mortality rate associated with cancer disease. According to the American Cancer Society report [1], a significant decline in cancer related deaths was observed in the last 3 decades, to around 3.5 million fewer cases than expected. One of the most common cancers, in the year 2022, is breast cancer which accounts for 12.5% of all new annual cancer cases, globally. From the American Cancer Society 2022 report, the most often diagnosed cancer cases recorded, among American women, was Breast Cancer. At least, 1 in every 8 American women will experience an invasive breast cancer during their lifetime. The estimated number of new cases associated with breast cancer in 2022 was over 300 thousand women, with almost 289 thousand invasive cases and a little over 51 thousand non-invasive cases [2].

## LITERATURE REVIEW

### **Myths and Misconceptions of Breast Cancer in the Pakistani Population**

Breast cancer is one of the most prevalent types of cancer in the female population. The cultural diversity, religious beliefs, myths, and misinformation regarding the disease contribute to diagnostic delays and enhanced burden on the healthcare system. This study aimed to ascertain the extent of knowledge and prevalence of erroneous beliefs and misconceptions regarding breast cancer among Pakistani women belonging to diverse socioeconomic and educational backgrounds. This cross-sectional study was performed in a tertiary care hospital in Karachi, Pakistan. A total of 350 women were enrolled in the study as a representative female population, and 300 participants were included who met the inclusion criteria. Participants were conveniently interviewed using a pre-piloted questionnaire designed to assess the prevalent myths and misconceptions about breast cancer. The data were analyzed by SPSS version 23 (IBM Corp., Armonk, NY, USA) using descriptive statistics. The study findings point to a significant prevalence of erroneous beliefs and a lack of accurate information on breast cancer. The mean age of the

participants was  $20.8 \pm 10.4$  years. The majority of the participants belonged to a middle socioeconomic status (70%) and were undergraduates (61.4%). The participants' friends and family members were the most frequent sources of information regarding breast cancer. The most common myth was "breast-feeding offers immunity to breast cancer completely" (76.6%), followed by "breast cancer spreads after biopsy" (63.8%). Participants also believed that breast tissue biopsy can lead to the spread of cancer (63.4%) and that faith healers and alternative medicine can cure breast cancer (47.5%). One-third (33.3%) of the participants considered all lumps to be breast cancer; however, approximately half (41.6%) of the participants thought that only painful lumps were associated with breast cancer. A significant number of participants believed breast cancer to be a result of God's curse (31.4%) or evil eye (38.7%). The findings suggest a critical need for community-based breast health education initiatives that take into account Pakistani women's distinctive cultural and societal attitudes and work to dispel common misconceptions about the condition.

## Breast cancer detection by leveraging Machine Learning

India has witnessed 30% of the cases of breast cancer during the last few years and it is likely to increase. Breast cancer in India accounts that one woman is diagnosed every two minutes and every nine minutes, one woman dies. Early detection and diagnosis can save the lives of cancer patients. This paper presents a novel method to detect breast cancer by employing techniques of Machine Learning. The authors carried out an experimental analysis on a dataset to evaluate the performance. The proposed method has produced highly accurate and efficient results when compared to the existing methods.

Breast cancer (BC) is the malignant tumor that activates in the cells of the breast. A tumor has the potential to spread to other parts of the body [1], [2]. BC is a universal disease that hammers the lives of women typically in the age group of 25–50. With the potential rise in the number of BC cases in India, the distress reaching is alarming. During the past five years, the survival rates of BC patients are about 90% in the USA and whereas in India the figure reports approximately 60% [3]. BC projection for India during 2020

suggests the number to go as high as two millions [4].

## **Fine-Grained Object Detection in Remote Sensing Images via Adaptive Label Assignment and Refined-Balanced Feature Pyramid Network**

—Object detection in high-resolution remote sensing images remains a challenging task due to the uniqueness of its viewing perspective, complex background, arbitrary orientation, etc. For fine-grained object detection in high-resolution remote sensing images, the high intra-class similarity is even more severe, which makes it difficult for the object detector to recognize the correct classes. In this article, we propose the refined and balanced feature pyramid network (RB-FPN) and center-scale aware (CSA) label assignment strategy to address the problems of fine-grained object detection in remote sensing images. RB-FPN fuses features from different layers and suppresses background information when focusing on regions that may contain objects, providing high-quality semantic information for fine-grained object detection. Intersection over Union (IoU) is usually applied to select the positive candidate samples for training. However,

IoU is sensitive to the angle variation of oriented objects with large aspect ratios, and a fixed IoU threshold will cause the narrow oriented objects without enough positive samples to participate in the training. In order to solve the problem, we propose the CSA label assignment strategy that adaptively adjusts the IoU threshold according to statistical characteristics of oriented objects.

Experiments on FAIR1M dataset demonstrate that the proposed approach is superior. Moreover, the proposed method was applied to the fine-grained object detection in high-resolution optical images of 2021 Gaofen challenge. Our team ranked sixth and was awarded as the winning team in the final.

## **Clinical Information Extraction**

### **Applications: A Literature Review**

With the rapid adoption of electronic health records (EHRs), it is desirable to harvest information and knowledge from EHRs to support automated systems at the point of care and to enable secondary use of EHRs for clinical and translational research. One critical component used to facilitate the secondary use of EHR data is the information extraction (IE) task, which automatically extracts and encodes clinical



information from text. In this literature review, we present a review of recent published research on clinical information extraction (IE) applications. A literature search was conducted for articles published from January 2009 to September 2016 based on Ovid MEDLINE In-Process & Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Scopus, Web of Science, and ACM Digital Library. A total of 1,917 publications were identified for title and abstract screening. Of these publications, 263 articles were selected and discussed in this review in terms of publication venues and data sources, clinical IE tools, methods, and applications (including disease areas, drug-related studies, and clinical workflow optimizations). Clinical IE has been used for a wide range of applications, however, there is a considerable gap between clinical studies using EHR data and studies using clinical IE. This study enabled us to gain a more concrete understanding of the gap and to provide potential solutions to bridge this gap.

### **Construction and Validation of a Lung Cancer Diagnostic Model Based on 6-Gene Methylation Frequency in Blood, Clinical**

### **Features, and Serum Tumor Markers**

Lung cancer has a high mortality rate. Promoting early diagnosis and screening of lung cancer is the most effective way to enhance the survival rate of lung cancer patients. Through computer technology, a comprehensive evaluation of genetic testing results and basic clinical information of lung cancer patients could effectively diagnose early lung cancer and indicate cancer risks. This study retrospectively collected 70 pairs of lung cancer tissue samples and normal human tissue samples. The methylation frequencies of 6 genes (FHIT, p16, MGMT, RASSF1A, APC, DAPK) in lung cancer patients, the basic clinical information, and tumor marker levels of these patients were analyzed. Then, the python package “sklearn” was employed to build a support vector machine (SVM) classifier which performed 10-fold cross-validation to construct diagnostic models that could identify lung cancer risk of suspected cases. Receiver operation characteristic (ROC) curves were drawn, and the performance of the combined diagnostic model based on several factors (clinical information, tumor marker level, and methylation frequency of 6 genes in

blood) was shown to be better than that of models with only one pathological feature. The AUC value of the combined model was 0.963, and the sensitivity, specificity, and accuracy were 0.900, 0.971, and 0.936, respectively. The above results revealed that the diagnostic model based on these features was highly reliable, which could screen and diagnose suspected early lung cancer patients, contributing to increasing diagnosis rate and survival rate of lung cancer patients.

## EXISTING SYSTEM

In a recent study conducted by the authors [7], Random Forest was observed to be a highly effective algorithm for breast cancer prediction and its performance was improved further when a feature selection technique, Genetic Algorithm, was included in its pre-processing phase. Feature selection is said to have relevance in machine learning as it is helpful in obtaining pertinent and germane data from a large multidimensional dataset, which may also help to reduce computational cost and improve the classification performance by taking out irrelevancies and noise from the data [8], [9], [10], [11]. In comparison to other dimensionality reduction techniques which make alterations to the initial data set, feature selection methods do

not make any alterations [12]. Although much research has been done around feature selection, improvements are still needed, such as in computational complexity and noise reduction.

### Disadvantages

- The complexity of data: Most of the existing machine learning models must be able to accurately interpret large and complex datasets to detect Breast Cancer Tumors.
- Data availability: Most machine learning models require large amounts of data to create accurate predictions. If data is unavailable in sufficient quantities, then model accuracy may suffer.
- Incorrect labeling: The existing machine learning models are only as accurate as the data trained using the input dataset. If the data has been incorrectly labeled, the model cannot make accurate predictions.

### Proposed System

The motivation behind the proposed hybridization of genetic algorithm based on other learning models, stems from the need to improve model performance, interpretability, and generalization ability by selecting relevant and non-redundant features. The main challenges faced include handling high-dimensional datasets, preventing over fitting, exploring the large search space, striking a balance between

feature relevance and redundancy, and ensuring scalability for large-scale datasets.

### Advantages

- 1) We propose a scalable and efficient hybridization Machine Learning models for exact predictions.
- 2) We conducted several hybridizations of genetic algorithm with filter and embedded feature selection methods, in the data pre-processing phase of Random Forest predictive model, with the aim of improving its performance.

### CONCLUSION

In conclusion, this paper explored three classes of feature selection techniques: Filter methods, Wrapper methods, and Embedded methods. Filter methods evaluate feature relevance based on correlation with the dependent variable, while Wrapper methods use model training to measure the usefulness of feature subsets. Embedded methods enhance the objective function during the learning process. The study used the Wisconsin Hospitals Madison Breast Cancer Database, which consisted of 569 samples and 32 features. The Random Forest classifier was employed to predict benign and malignant tumors. The research compared different hybridization methodologies and evaluated their

performance using accuracy as the metric.

In Case 1, the hybrid model GA + Fisher\_Score achieved the highest accuracy of 99.12%.

### REFERENCES

- [1] Cancer Facts and Figures 2022, Atlanta: American Cancer Society, American Cancer Society, Atlanta, GA, USA, 2022.
- [2] (2022). Breast Cancer Facts and Statistics. [Online]. Available: <https://www.breastcancer.org/facts-statistics>
- [3] A. R. Vaka, B. Soni, and S. Reddy, "Breast cancer detection by leveraging machine learning," *ICT Exp.*, vol. 6, no. 4, pp. 320–324, Dec. 2020.
- [4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
- [5] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018.



- [6] K.-H. Chen, K.-J. Wang, A. M. Adrian, K.-M. Wang, and N.-C. Teng, "Diagnosis of brain metastases from lung cancer using a modified electromagnetism like mechanism algorithm," *J. Med. Syst.*, vol. 40, no. 1, pp. 1–14, Jan. 2016.
- [7] J. Ayoola and T. Ogunfunmi, "A comparative analysis of regression algorithms with genetic algorithm in the prediction of breast cancer tumors," in *Proc. IEEE Global Humanitarian Technol. Conf. (GHTC)*, Sep. 2022, pp. 143–149.
- [8] Y. Sun, C. F. Babbs, and E. J. Delp, "A comparison of feature selection methods for the detection of breast cancers in mammograms: Adaptive sequential floating search vs. genetic algorithm," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Jan. 2005, pp. 6532–6535.
- [9] A. Alzubaidi, G. Cosma, D. Brown, and A. G. Pockley, "Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information," in *Proc. Int. Conf. Interact. Technol. Games (ITAG)*, Oct. 2016, pp. 70–76.
- [10] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, "A comparative study for breast cancer prediction using machine learning and feature selection," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 1049–1055.
- [9] Karne, R. K. ., & Sreeja, T. K. . (2023). PMLC- Predictions of Mobility and Transmission in a Lane-Based Cluster VANET Validated on Machine Learning. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(5s), 477–483. <https://doi.org/10.17762/ijritcc.v11i5s.7109>
- [10] Radha Krishna Karne and Dr. T. K. Sreeja (2022), A Novel Approach for Dynamic Stable Clustering in VANET Using Deep Learning (LSTM) Model. *IJEER* 10(4), 1092-1098. DOI: 10.37391/IJEER.100454.
- [11] Reddy, Kallem Niranjan, and Pappu Venkata Yasoda Jayasree. "Low Power Strain and Dimension Aware SRAM Cell Design Using a New Tunnel FET and Domino Independent Logic." *International Journal of Intelligent Engineering & Systems* 11, no. 4 (2018).
- [12] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Design of a Dual Doping Less Double Gate Tfet and Its Material Optimization Analysis on a 6t Sram Cells."
- [13] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Low power process, voltage, and temperature (PVT) variations aware improved tunnel FET on 6T SRAM cells." *Sustainable Computing: Informatics and Systems* 21 (2019): 143-153.
- [14] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Survey on improvement of PVT aware variations in tunnel FET on SRAM cells." In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, pp. 703-705. IEEE, 2017.