**­** **Text

Description automatically generated with low confidence**

**Optimization of Logistic Regression Algorithm with Genetic Algorithm for Breast Cancer Classification**

Marzieh Khalili Chachaki

Asia Pacific University of Technology & Innovation (APU), Technology Park Malaysia, Kuala Lumpur, Malaysia

[TP067265@mail.apu.edu.my](mailto:TP067265@mail.apu.edu.my)

[marzieh.khaliliy@gmail.com](mailto:marzieh.khaliliy@gmail.com)

**Abstract**

**Breast cancer is cancer that develops in the breast cells. The most prevalent cancer among women is breast cancer. Although both men and women may be diagnosed with breast cancer, women are substantially more often affected. Early detection and treatment of illnesses significantly improves patients' chances of survival. This article examines several research on the use of machine learning such as LR, SVM, ANN and genetic algorithms for breast cancer diagnosis. The research indicates that evolutionary algorithms may have a significant impact on the optimization of machine learning models. This work is intended to identify and justify the improvements proposed for the future study, in order to obtain better predictions and high accuracies using with GA and machine learning algorithms.**

***Key words\_\_ Breast Cancer, Machine learning, Genetic Algorithm, Logistic regression***

1.Introduction

Cancer is among the most serious illnesses. For women, breast cancer is the most dangerous form of the illness (Derisma, Silvana and Imelda, 2018). Pathological diagnosis is the most reliable approach for regulating and improving the disease's survival rate. When it comes to diagnosis, conventional approaches may rely on the specialists' own experiences, and the outcomes may be subjective. If cancer cells are found early enough, they may be attacked and prevented from spreading. A rapid diagnosis necessitates an accurate and steady diagnostic strategy that allows the physician to differentiate between benign and malignant breast tumours without the need for a surgical biopsy. Every minute, at least one new incidence of breast cancer is diagnosed in a woman somewhere in the globe, and the illness claims the lives of at least one individual (Sultana, Khader Jilani and ., 2018). In recent years, computer diagnostic tools and artificial intelligence methods have evolved (Huang *et al.*, 2019). This diagnosis has been categorised using several methods, including the Decision Tree technique, the KNN algorithm, the Logistic Regression algorithm, the ANN algorithm, etc., (Liu, 2018). Total detection accuracy depends on three factors: high-quality technologies, the performance of an experienced physician, and the interpretation and use of a multidisciplinary approach for detecting breast cancer by many disciplines. Consequently, data mining and artificial intelligence must continue to make contributions (Derisma, Silvana and Imelda, 2018). Evolutionary algorithms are a common name for nature-inspired algorithms. It accomplishes its mission by screening through a population and eliminating undesirables. Alternatively, healthy members may be able to bear the situation until a better solution is identified (Sethi, 2018).

The study of building computer algorithms that can recover data is known as machine learning. It is an application of artificial intelligence that can learn and develop independently via experience (Sethi, 2018).

2.Problem statement

Breast cancers are among the most prevalent. Assuming an early diagnosis, there is a fair likelihood of survival for at least five years. According to data, the death rate from this illness has increased considerably over the last few decades. In terms of treatment, early diagnosis is essential. In addition to medications, then, a Data Science solution is necessary to address the issue of death. This research aims to predict whether a tumour will be malignant or benign, as well as to uncover general patterns that may be utilised to drive model selection and hyper parameter selection. The breast cancer dataset is utilised to improve the accuracy of a logistic regression model, and then a genetic algorithm is employed to optimise it.

3.Aim

The objective of this study was to evaluate the performance of the Logistic Regression approach and enhance it using a Genetic Algorithm trained on Breast Cancer datasets.

4.Objectives

The objectives of the study are as follows:

1. Prediction accuracy of breast cancer based on Logistic Regression Algorithm
2. Optimize and improve accuracy of LR with Genetic Algorithm

5.Literature review

A genetic algorithm utilises n chromosomes to generate a new population. After that, the fitness of each chromosome is examined. By choosing new chromosomes, crossing two chromosomes, and modifying the chromosomal organisation, a new population may be generated (with a chance of doing so) (Assaf and Saleh, 2017). GA is an important tool in the pursuit of the optimal solution set for a given design. The genetic algorithm flowchart of the procedure is shown below.

Diagram

Description automatically generated

This section discusses machine learning and evolutionary methodologies in relation to breast cancer classification.

According to recent study, machine learning algorithms may be utilised to solve the problem of breast cancer detection. In the first test, it was shown that, with correct configuration, the three most popular evolutionary algorithms may achieve identical performance. In the second experiment, we examined how combining multiple techniques to feature selection may improve accuracy. Their last experiment revealed a way for automatically building the machine learning classifier. Because hyperparameters constitute a problem for machine learning algorithms, the GP technique was developed to address this issue. From a number of conceivable setups, appropriate algorithms were chosen based on the algorithm they gave. Using the proposed strategy, which examined an ensemble of strategies using an extensive machine learning methodology, it took much longer to get relevant results. Consequently, the proposed model seems to be a perfect match for both the control parameter setup and the automated breast cancer detection of machine learning algorithms (Dhahri *et al.*, 2019).

In another work, (Murtirawat *et al.*, 2020), the repository for machine learning at UCI was used to train and validate models utilising Wisconsin breast cancer diagnostic data. Before extracting features using Principal Component Analysis (PCA), data were pre-processed (PCA). K-Nearest Neighbours (98.60 percent), Logistic Regression (97.90 percent), and Ensemble Learning (99.30 percent) were some of the Machine Learning algorithms recommended in the research, with an accuracy of 98.60 percent, 97.90 percent, and 99.30 percent, respectively.

They developed an ensemble model with a 97.07% accuracy rate by using a variety of well-established techniques. It was a comparison of numerous machine learning algorithms and an assessment of the time complexity of all the strategies used in this paper; their successful work led to a 96 percent accuracy in breast cancer cell identification. (Hazra, Kumar and Gupta, 2016).

The cancer features found by Subrata kumar Mandal (Kumar Mandal, 2017), accurately differentiate between malignant and benign breast cancer. The temporal complexity and accuracy of several classifiers, such as Naive Bayes, Logistic Regression, and Decision Trees, are compared. The author discovered via experimentation that Logistic Regression (LR) gave the maximum accuracy and execution time.

Wang and colleagues (Wang *et al.*, 2018) proposed ensemble learning using the weighted area under the receiver operating characteristic curve (ROC) to reduce diagnostic variability and increase diagnostic accuracy. In comparison to the present SVM model, their model achieved 76.4% accuracy on the balanced SEER BC dataset, with a variance reduction of 97.9% and an increase in training accuracy of 33.3%, which sufficiently 85 supports the efficacy of ensemble modelling approaches.

Kawashima (Kawashima, Bai and Quan, 2017) use text mining and pattern clustering to analyse their study on breast cancer and its associated genes. Due to its reliance on unsupervised learning, data labelling is not required. Breast cancer is associated with a limited number of genes, which may be narrowed down depending on their physical appearance. Using the clustering method, the link between the two is determined. In this research, clustering is proven to be superior than k-means.

They demonstrate that breast cancer is a prominent issue for women using data from Rajaguru and Prabhakar (Rajaguru and Kumar Prabhakar, 2018). As a result, tumours may be benign or malignant. There is a distinction between cancerous and noncancerous tumours. Breast cancer is a cancerous tumour that is malignant. Utilizing artificial approaches, breast cancer may be efficiently treated. In the realm of breast cancer prognosis, machine learning and computer technology play a crucial role. The accuracy, performance index, sensitivity, and specificity of the Bayesian classifier are used to describe the findings of this investigation. The actual population percentage is 83.45 percent.

The study of Zarbakhsh (Koyuncu *et al.*, 2017) focuses on breast cancer in women aged 35 and older. In this study, neuro-fuzzy inference is utilised to categorise and choose features for pre-processing. The Cuckoo technique is used to compute radius. To show the algorithm's precision, it was applied to the Wisconsin Breast Cancer Dataset.

Heenaye and Khan (Khan, 2018) advocated for the early identification of breast cancer. To account for the possibility of breast cancer, a successful algorithm was created. This investigation included classification, feature selection, and segmentation on breast cancer ultrasound and mammogram pictures. This study adopts a Bayesian Neural Network methodology. Consequently, the model's sensitivity is increased to 100 percent, making it more accurate. Using a technology called support vector machine, cancer has been discovered. Sensitivity, false negatives, false positives, and specificity will be shown for the findings.

Hawaz (Nawaz, Sewissy and Soliman, 2018) created a multiclass categorization of breast cancer using CNN. With their proposed technique and the DenseNet and BreakHis datasets, they achieved 95.4% accuracy. Kasani (Khan, 2018) categorised histopathological biopsies into malignant and benign patients using an approach based on ensemble deep learning. When evaluated on a range of datasets, including BreakHis, ICIAR, PatchCamelyon, and Bioimaging, the ensemble deep learning-based method acquired a classification accuracy of 83.10–98.13 percent.

Mohanty(Mohanty *et al.*, 2019) introduced FS as an optimization strategy for forests. Mammographic benign and pathological lesions are classified using SVM, KNN, NB, and C4.5. 99.08 percent was the greatest accuracy of the C4.5 classification for the DDSM data collection. Using the MIAS dataset, the authors were able to attain a Nave Bayes classification accuracy of 97.86 percent. According to Badr (Badr, Salam and Ahmed, 2019), employing novel hybrid approaches like SNM and Gray Wolf Optimizer to determine if a breast tumour is malignant or benign (GWO). With a minimum of 569 rows and 32 columns, the maximum accuracy of the GWO-SVM model was 97.0 percent.

In another experiment, the findings of which were only just released, GAs achieved a higher rate of accuracy. The low proportion of false-positive predictions produced by our approach reveals something crucial: (less than 0.1). On the other hand, neither optimizer's expectations for recall are met. Acceptable: F1-scoring. All three optimizers need more time than GA to finish. The stability and accuracy of the GA are established through a lengthy process of selection, crossover, and mutation. A uniform distribution gave classification accuracy that was slightly better to that supplied by a normal distribution when creating the initial weights(Davoudi and Thulasiraman, 2021).

6.Conclusion

Literature papers have provided insight into the breast cancer prediction applications of some Machine Learning algorithms and Genetic Algorithms. As stated, we want to apply Logistic Regression and genetic algorithm techniques in the future to improve the accuracy of our method after data scaling. In addition, further data samples should be gathered in the future to improve the predictive capabilities of the proposed system.

7.References

Assaf, R. and Saleh, Y. (2017) ‘Vehicle-Routing Optimization for Municipal Solid Waste Collection Using Genetic Algorithm: The Case of Southern Nablus City’, *Civil and Environmental Engineering Reports*, 26(3), pp. 43–57. doi: 10.1515/ceer-2017-0034.

Badr, E. M., Salam, M. A. and Ahmed, H. (2019) ‘Optimizing Support Vector Machine using Gray Wolf Optimizer Algorithm for Breast Cancer Detection’, (November).

Davoudi, K. and Thulasiraman, P. (2021) ‘Evolving convolutional neural network parameters through the genetic algorithm for the breast cancer classification problem’, *Simulation*, 97(8), pp. 511–527. doi: 10.1177/0037549721996031.

Derisma, Silvana, M. and Imelda (2018) ‘Optimization of Neural Network with Genetic Algorithm for Breast Cancer Classification’, *2018 International Conference on Information Technology Systems and Innovation, ICITSI 2018 - Proceedings*, pp. 398–403. doi: 10.1109/ICITSI.2018.8696014.

Dhahri, H. *et al.* (2019) ‘Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms’, *Journal of Healthcare Engineering*, 2019. doi: 10.1155/2019/4253641.

Hazra, A., Kumar, S. and Gupta, A. (2016) ‘Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms’, *International Journal of Computer Applications*, 145(2), pp. 39–45. doi: 10.5120/ijca2016910595.

Huang, H. *et al.* (2019) ‘A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features’, *BMC Bioinformatics*, 20(Suppl 8), pp. 1–14. doi: 10.1186/s12859-019-2771-z.

Kawashima, K., Bai, W. and Quan, C. (2017) ‘Text mining and pattern clustering for relation extraction of breast cancer and related genes’, *Proceedings - 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2017*, pp. 59–63. doi: 10.1109/SNPD.2017.8022701.

Khan, M. H. M. (2018) ‘Automated breast cancer diagnosis using artificial neural network (ANN)’, *Proceedings - 3rd Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2017*, 2017-December, pp. 54–58. doi: 10.1109/ICSPIS.2017.8311589.

Koyuncu, B. *et al.* (2017) ‘9th International Conference on Computational Intelligence and Comunication Networks : proceedings : venue: Final International University, GIRNE, Cyprus’, pp. 39–42. doi: 10.1109/CICN.2017.11.

Kumar Mandal, S. (2017) ‘Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree’, *International Journal Of Engineering And Computer Science*, 6(2), pp. 2319–7242. doi: 10.18535/ijecs/v6i2.40.

Liu, L. (2018) ‘Research on logistic regression algorithm of breast cancer diagnose data by machine learning’, *Proceedings - 2018 International Conference on Robots and Intelligent System, ICRIS 2018*, pp. 157–160. doi: 10.1109/ICRIS.2018.00049.

Mohanty, F. *et al.* (2019) ‘Mammogram classification using contourlet features with forest optimization-based feature selection approach’, *Multimedia Tools and Applications*, 78(10), pp. 12805–12834. doi: 10.1007/s11042-018-5804-0.

Murtirawat, R. *et al.* (2020) ‘Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning’, *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, (Icesc), pp. 534–540. doi: 10.1109/ICESC48915.2020.9155783.

Nawaz, M., Sewissy, A. A. and Soliman, T. H. A. (2018) ‘Multi-class breast cancer classification using deep learning convolutional neural network’, *International Journal of Advanced Computer Science and Applications*, 9(6), pp. 316–322. doi: 10.14569/IJACSA.2018.090645.

Rajaguru, H. and Kumar Prabhakar, S. (2018) ‘Bayesian linear discriminant analysis for breast cancer classification’, *Proceedings of the 2nd International Conference on Communication and Electronics Systems, ICCES 2017*, 2018-January(Icces), pp. 266–269. doi: 10.1109/CESYS.2017.8321279.

Sethi, A. (2018) ‘Analogizing of Evolutionary and Machine Learning Algorithms for Prognosis of Breast Cancer’, *2018 7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2018*, pp. 252–255. doi: 10.1109/ICRITO.2018.8748502.

Sultana, J., Khader Jilani, A. and .. (2018) ‘Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers’, *International Journal of Engineering & Technology*, 7(4.20), p. 22. doi: 10.14419/ijet.v7i4.20.22115.

Wang, H. *et al.* (2018) ‘A support vector machine-based ensemble algorithm for breast cancer diagnosis’, *European Journal of Operational Research*, 267(2), pp. 687–699. doi: 10.1016/j.ejor.2017.12.001.