International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Breast Cancer Classification using LGBM and SVM

Denesh Das [1,*], Md Masum Billah [2], Amit Deb Nath [3], Numair Bin Sharif [4] and Kallol Kanti Mondal [5]

[1] *Department of Electrical and Electronic Engineering, Southern University Bangladesh, Chattogram, Bangladesh.*
[2] *Department of Electrical and Electronic Engineering, University of Rajshahi, Rajshahi, Bangladesh.*
[3] *Department of Electrical and Electronic Engineering, Leading University, Sylhet, Bangladesh.*
[4] *Department of CSE, United International University, Dhaka, Bangladesh.*
[5] *Institute of Biological Sciences, University of Rajshahi, Rajshahi, Bangladesh.*

International Journal of Science and Research Archive, 2022, 07(02), 876-881

Publication history: Received on 07 November 2022; revised on 19 November 2022; accepted on 28 November 2022

Article DOI: https://doi.org/10.30574/ijsra.2022.7.2.0313

Abstract

Breast cancer (BC) is among the most common cancers affecting women worldwide, highlighting the urgent need for early and accurate diagnosis. It develops in the breast tissue and is one of the most frequent causes of women's death. This cancer can be cured if it is diagnosed at preliminary stage. Malignant and benign are two types of tumor found in case of breast cancer. Malignant tumors are deadly as their rate of growth is much higher than benign tumors. So, early identification of tumor type is pivotal for the appropriate treatment of a patient having breast cancer. Machine learning (ML) has emerged as a powerful tool for BC classification, enhancing diagnostic precision and improving patient outcomes. In this work, Wisconsin Breast Cancer Dataset is used. Our goal is to analyze the dataset and evaluate the performance of LGBM and SVM for predicting breast cancer.

Keywords: Breast Cancer Diagnosis; WBCD Dataset; Malignant; Benign; Classification; Machine Learning; LGBM; SVM

1. Introduction

Cancer is currently one of the major causes of mortality globally. For women, breast cancer-related deaths are more common than deaths from other types of cancer due to the disease's annual death toll of thousands of people [1-2]. The rate of incidence for breast cancer differs by region, from 19.3 per 100,000 women in East Africa to 89.7 per 100,000 women in Western Europe, according to some statistics [3]. It is well known that the number of new cases has been rising these years and will likely exceed 27 million in 2030 [4]. On the other hand, breast lumps enable us to recognize breast tissue that differs from that found under normal circumstances [5].

To determine breast cancer, patients are frequently subjected to a barrage of examinations which include ultrasound, biopsy and mammography according to the varying nature of breast cancer symptoms. Of these methods, the most indicative is biopsy that involves the extraction of sample tissues or cells for investigation. The sample of cells is collected from a breast through Fine Needle Aspiration (FNA) procedure and then sent for analysis under a microscope to a pathology laboratory [6]. Numerical features such as radius, texture, perimeter and area can be calculated from microscopic images of cells and tissues. Data subsequently obtained from FNA are analyzed to predict the probability of the patient having malignant tumor in combination with various imaging data. The prophesy and probability of survival can be remarkably enhanced by early diagnosis of BC, as it allows patients to receive timely clinical treatment. The proper identification of BC and patient categorization into malignant or benign classes is an extremely significant avenue of research. Various methods for predicting breast cancer have been established in recent years. Classification techniques for instance, Random Forest (RF) [7], Support Vector Machine (SVM), Adaboost Classifier, K-Nearest Neighbors (KNN) and XGboost classifier have been used in the recent literature [8, 9].

* Corresponding author: Denesh Das

2. Literature Review

In [7], the capability of the classification of Naïve Bayes, Random Forest, Logistic Regression, Multilayer Perceptron, K-nearest neighbors in evaluating the Breast Cancer Disease dataset culled from UCI machine learning repository, was observed to predict the existence of Breast cancer. In [10], authors have used Random Forest classifier for the identification and prediction of breast cancer to determine whether the person has breast cancer. This offers the highest identification accuracy since both classification and regression approaches are used for Random Forest algorithm. In [11], a study on breast cancer was provided by the authors to develop predictive models for breast cancer survival. In this paper, three breast cancer survivability prediction models were applied to two classes: benign and malignant cancer. In [12], authors highlighted all previous research on ML algorithms used for breast cancer determination. They suggested that the problem of limited available data sets can be solved by data augmentation techniques. In [13], authors presented a technique that can be used to detect and identify cell morphology in automated systems that carry out the classification using computer-aided mammogram image features. In [14], authors have compared various classification and clustering algorithms in the survey. The result shows that the algorithms for classification are better predictors than the clustering algorithms. In [15], the method of automatic detection of anomalies in mammograms is discussed. Applying the fuzzy-C-means and thresholding strategy, suspect region-of-interest (ROI) was segmented. The proposed algorithm for the Mini-MIAS dataset was validated. They concluded that the performance of suspicious region detection in mammograms can be improved by subtracting preprocessed enhanced and preprocessed enhanced inverted images. In [16], for the identification of the malignant and benign state, an algorithm was proposed by the authors depending on a fuzzy inference system. Comparison of conventional performance criteria such as sensitivity, accuracy and specificity suggest that their introduced solution outperforms Artificial Neural Network (ANN) and SVM classification.

3. Dataset

In this work, Wisconsin Breast Cancer Dataset (WBCD) of the FNA biopsy system has been used and different machine learning (ML) classifiers have been implemented to determine the form of breast cancer in a suspected patient.

Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, Wisconsin, USA has developed the WBCD dataset used for this paper which is publicly accessible [17]. This dataset includes 357 and 212 cases of benign and malignant breast cancer respectively as shown in Fig. 1.

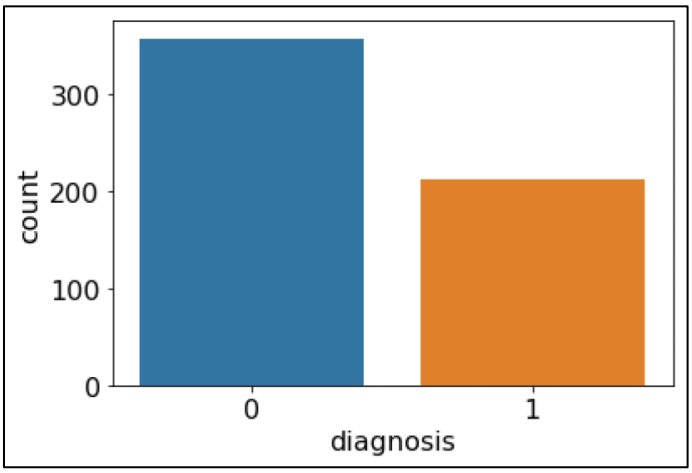

Figure 1 Class Distribution

The dataset comprises 32 columns, with the ID number being the first column and the diagnosis outcome (0-benign and 1-malignant) being the second column. The rest of the columns (3-32) contain three measurements (mean, standard deviation, and mean of worst) of ten features. These features represent the shape and size of the target cancer cell nucleus. The sample of cells is collected from a breast through Fine Needle Aspiration (FNA) procedure in biopsy test. For each cell nucleus, these features are determined by analyzing under a microscope in a pathology laboratory. The 10 real-value features are described in Figure 2. In the pre-processing phase, an exploratory analysis is conducted to better understand the dataset. Data cleansing is applied to improve the overall accuracy, followed by normalization of feature

values using a standardized scale. Additionally, categorical variables are converted into numerical format to ensure compatibility with machine learning models [18].

Feature Name	Feature Description
Radius	Average of distances from center to circumference points.
Texture	Standard deviation (SD) of gray-scale value.
Perimeter	Gross distance between the snake points.
Area	Total number of pixels on the inside of the snake along with one half of the pixels in the circumference.
Smoothness	Local variance in length of radius, quantified by calculating the length difference.
Compactness	$\text{Perimeter}^2/Area$.
Concavity	Intensity of the contour's concave parts.
Concave points	The number of contour concavities.
Symmetry	The difference in length between lines perpendicular to the major axis in both directions to the cell boundary.
Fractal Dimension	Coastline estimation. A higher value leads to a less normal contour representing a higher risk of malignancy.

Figure 2 Feature Description

4. LGBM

LGBM is a gradient boosting framework optimized for speed and memory efficiency. Unlike traditional boosting methods, it grows trees in a leaf-wise manner, selecting splits that minimize error at each step rather than growing trees level-wise. This approach allows LGBM to focus on the most critical regions of the data, leading to faster convergence and improved accuracy. The model is particularly effective in handling large datasets, efficiently managing high-dimensional feature spaces, and reducing computational complexity by leveraging histogram-based learning for optimal split selection. In BC classification, LGBM has been successfully integrated into ensemble learning frameworks, significant.

5. Support Vector Machine (SVM)

SVM is another supervised learning model that analyzes data for regression and classification in machine learning [19]. Given a set of training examples, each labelled as belonging to one of two categories, an SVM training algorithm develops a model that categorizes new examples into one of two groups, resulting in a non-probabilistic binary linear classifier. SVM is preferred for its computational power and its ability to detect outliers. Apart from that the prediction's precision and its effectiveness in small datasets are among SVM advantages.

6. Results and discussion

The effectiveness of the ML model is evaluated using several essential performance metrics. Accuracy measures the overall proportion of correctly classified instances. Precision evaluates how many of the identified cases were malignant, focusing on the model's correctness. Recall measures how well the model can detect all malignant cases, which is particularly important for medical diagnoses. The F1-score balances precision and recall, offering a harmonic means to ensure both metrics are considered. 20% of the entire dataset was included in the test dataset. A confusion matrix is generated for the actual and predicted result consisting of TP, FP, TN, and FN for calculating accuracy for each algorithm used. Below, the meaning of the terms is mentioned.

- TP = True Positive (Accurately Identified)
- TN = True Negative (Inaccurately Identified)
- FP = False Positive (Accurately Rejected)
- FN = False Negative (Inaccurately Rejected)

Confusion matrix for LGBM and SVM is shown in Fig 3 and 4. CR of LGBM and SVM is shown in Fig 5 and 6. Table 1 show comparison with state-of-the-art works.

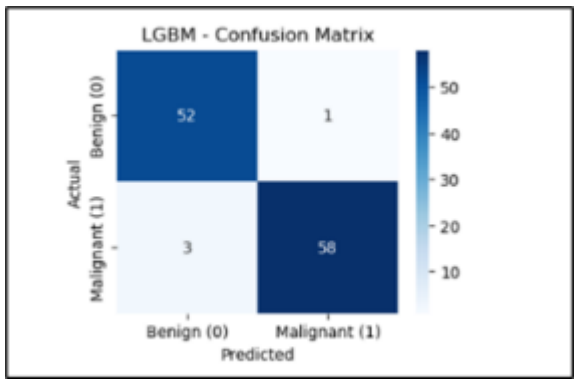

Figure 3 Confusion Matrix

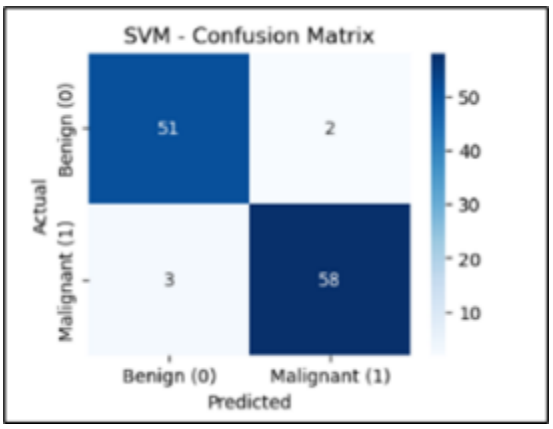

Figure 4 Confusion Matrix

Classification Report:	precision	recall	f1-score	support
Benign (0)	0.9455	0.9811	0.9630	53
Malignant (1)	0.9831	0.9508	0.9667	61
accuracy			0.9649	114
macro avg	0.9643	0.9660	0.9648	114
weighted avg	0.9656	0.9649	0.9649	114

Figure 5 CR of LGBM

Classification Report:	precision	recall	f1-score	support
Benign (0)	0.9444	0.9623	0.9533	53
Malignant (1)	0.9667	0.9508	0.9587	61
accuracy			0.9561	114
macro avg	0.9556	0.9565	0.9560	114
weighted avg	0.9563	0.9561	0.9562	114

Figure 6 CR of SVM

Table 1 Comparison

Ref.	Model	Accuracy
[16]	RF	91%
	SVM	90%
	KNN	84%
[19]	DT	94.73%
	KNN	93.85%

	AdaBoost	94.73%
[20]	MLP	95.28%
	J48	95.14%
	IBK	94.56%
[21]	RNN	63.15%
	Genetic Algorithm	80.39%
	Fuzzy Logic	88.81%
Ours	LGBM	96.49%
	SVM	95.61%

7. Conclusion

In our work, the Wisconsin Breast Cancer Dataset was utilized and LGBM and SVM were applied to assimilate the efficacy and usefulness of these algorithms to find the highest accuracy of classifying malignant and benign breast cancer. The correlation between different features of the dataset has been analyzed for feature selection. The results will assist in picking the best ML algorithm for the construction of an automatic breast cancer diagnostic system. From our study, we can conclude that LGBM gives the maximum accuracy with an accuracy of 92.98%. We will try to strengthen our work in future by handling a comparatively large dataset and incorporating some more functions such as breast cancer phase detection and so on. We hope that this study will contribute to the clinical application of breast cancer treatment.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2017,” CA Cancer J Clin, vol. 67, no. 1, pp. 7–30, Jan. 2017.

[2] Cancer Genome Atlas Network, “Comprehensive molecular portraits of human breast tumours,” Nature, vol. 490, no. 7418, pp. 61–70, Sep. 2012.

[3] M. M. A. Rahhal, “Breast cancer classification in histopathological images using convolutional neural network,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 3, 2018. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2018.090310

[4] T. Araujo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Pol ´ onia, ´ and A. Campilho, “Classification of breast cancer histology images using convolutional neural networks,” PLoS One, vol. 12, no. 6, p. e0177544, Jun. 2017

[5] K. R and N. K, “Automated diagnosis of breast cancer using wavelet based entropy features,” 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 274–279, 2018.

[6] S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.

[7] Bharati, Subrato, Mohammad Atikur Rahman, and Prajoy Podder. "Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA." 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT). IEEE, 2018.

[8] Y. Khourdifi and M. Bahaj, “Applying best machine learning algorithms for breast cancer prediction and classification,” in 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS). IEEE, 2018, pp. 1-5.

[9] N. K. Sinha, M. Khulal, M. Gurung, and A. Lal, “Developing a web based system for breast cancer prediction using xgboost classifier,” International Journal of Engineering Research Technology (IJERT), vol. 9, 2020

[10] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, “Breast cancer prediction via machine learning,” in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2019, pp. 121-124.

[11] V. Chaurasia, S. Pal, and B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” Journal of Algorithms & Computational Technology, vol. 12, no. 2, pp. 119-126, 2018.

[12] N. Fatima, L. Liu, S. Hong, and H. Ahmed, “Prediction of breast cancer, comparative review of machine learning techniques, and their analysis,” IEEE Access, vol. 8, pp. 150360-150376, 2020.

[13] A. Toprak, “Extreme learning machine (elm)-based classification of benign and malignant cells in breast cancer,” Medical science monitor: international medical journal of experimental and clinical research, vol. 24, p. 6537, 2018.

[14] D. S. Jacob, R. Viswan, V. Manju, L. Padma Suresh, and S. Raj, “A survey on breast cancer prediction using data mining techniques,” in 2018 Conference on Emerging Devices and Smart Systems (ICEDSS). IEEE, 2018, pp. 256-258.

[15] K. L. Kashyap, M. K. Bajpai, and P. Khanna, “Breast cancer detection in digital mammograms,” in 2015 IEEE international conference on imaging systems and techniques (IST). IEEE, 2015, pp. 1-6.

[16] A. Bah and M. Davud, "Analysis of Breast Cancer Classification with Machine Learning based Algorithms," 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), Istanbul, Turkey, 2022, pp. 1-4, doi: 10.1109/ICMI55296.2022.9873696

[17] Wolberg, W. (1990). Breast Cancer Wisconsin (Original) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z.

[18] Md Boktiar Hossain, Rashedur Rahman, Khandoker Hoque, “Feature-Driven Supervised Learning for Detecting DDoS Attack”, International Journal of Science and Research Archive, 2021, 04(01), 393-402.

[19] S. Pawar, P. Bagal, P. Shukla and A. Dawkhar, "Detection of Breast Cancer using Machine Learning Classifier," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-5, doi: 10.1109/ASIANCON51346.2021.9544767.

[20] Ö. S. Keskin, A. Durdu, M. F. Aslan and A. Yusefi, "Performance comparison of Extreme Learning Machines and other machine learning methods on WBCD data set," 2021 29th Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, 2021, pp. 1-4, doi: 10.1109/SIU53274.2021.9477984.

[21] D. Sandeep and G. N. B. Bethel, "Accurate Breast Cancer Detection and Classification by Machine Learning Approach," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 366-371, doi: 10.1109/I-SMAC52330.2021.9640710.