International Journal on
Advanced Science
Engineering
Information Technology

Vol.11 (2021) No. 2
ISSN: 2088-5334

Gradient Boosting Machine Based on PSO for prediction of Leukemia after a Breast Cancer Diagnosis

Mohanad A.Deif[a,*], Rania E. Hammam[a] , Ahmed A. A. Solyman[b]

[a] Department of Bioelectronics, Faculty of Engineering, Modern University of Technology and Information (MTI) University, Egypt
[b] Department of Electrical and Electronics Engineering, Istanbul Gelisim University, Turkey
Corresponding author: *Mohand.Deif@eng.mti.edu.eg

Abstract—**The purpose of this study is to develop an accurate risk predictive model for Chronic Myeloid Leukemia (CML) after an early diagnosis of Breast Cancer (BC). Gradient Boosting Machine (GBM) classification algorithm has been applied to the SEER breast cancer dataset for females diagnosed with BC from 2010 to 2016. A practical Swarm optimizer (PSO) was utilized to optimize the GBM algorithm's hyperparameters to find the SEER dataset's best attributes. Nine attributes were carefully selected to study the growth of CML after a lag time of 6 months following BC's diagnosis. The results revealed that the predictive model could classify patients with breast cancer only and patients with breast cancer with Leukemia by an achieved Accuracy, Sensitivity, and Specificity rates of 98.5 %, 99 %, 97.85 %, respectively. To verify the performance of the proposed algorithm, the accuracy of the suggested GBM classifier model was compared with another state-of-the-art model classifiers KNN (k-Nearest Neighbor), SVM (Support Vector Machine), and RF (Random Forest), which are commonly applied algorithms in most of the existing literature. The results also proved the superior ability of the implemented GBM model Classifier in the classification of breast cancer disease and prediction of patients having Leukemia developed after having breast cancer. These results are promising as they show the integral role of the GBM classifier to classify and predict the tumor with high accuracy and efficiency, which will further help in better cancer diagnosis and treatment of the disease.**

Keywords— **Risk predictive model; chronic myeloid leukemia; breast cancer; gradient boosting machine; classification algorithm.**

Manuscript received 17 Aug. 2020; revised 3 Dec. 2020; accepted 1 Jan. 2021. Date of publication 30 Apr. 2021.

I. Introduction

Breast Cancer is the most communal malignancy amongst females, leading to death in middle-aged females[1]. Survival after a breast cancer diagnosis has improved due to early diagnosis and effective treatments. However, the increased life probability of BC patients has led to the development of other malignancies. Patients treated with radiotherapy and chemotherapy, commonly including alkylating agents and anthracyclines, are susceptible to an increased risk of developing leukemia [2]. Many other studies also indicated an increased risk of acute Leukemia following breast cancer chemotherapy [3], [4]. In specific, the prediction of acute myeloid Leukemia (AML) after BC diagnosis was also suggested [5], [6]. Mehrdad et al. [7] reported a case study of a young female with breast cancer that developed acute myelogenous leukemia (AML) malignancy within a limited time of treatment. Balduzzi et al. [8] revealed that most of the patients who showed secondary Leukemia after chemotherapy for their primary breast cancer had been treated with the combined use of alkylating agents and radiation therapy.

Muneer et al.[9] applied a comprehensive statistical study on [SEER] data to inspect the survival and the risk of chronic myeloid Leukemia (CML) after breast cancer (BC) diagnosis. They utilized the Epidemiology, Surveillance, and End Results 'SEER' database. Female, BC diagnosed from 1992 to 2014, were chosen and monitored for the growth of CML after a delay time of 6 months following BC's diagnosis. The authors found that the observed/expected (O/E) ratios with 95% confidence intervals (CI) after BC diagnosis were 1.26 and that the probability of the CML during the first 5 years of diagnosis was significantly higher. They also concluded that hormonal receptors, radiotherapy, and chemotherapy were related to an effective elevated risk of CML in BC patients. The general permanence median of CML was 28 months after BC.

Machine learning techniques such as Gradient Boosting Machine (GBM) [10] have been employed to detect breast cancer. Machine learning is a subfield of artificial intelligence research that uses a range of statistical and optimization methods to "learn" from empirical data and then use the previous training to identify new data, recognize new trends or predict other outputs[11]. Machine learning is a more effective method than statistics because it allows decisions to be made that could not be made using traditional statistical methods [11], [12]. Austria et al. [13] conducted a comparison between different machine learning algorithms in breast cancer prediction. The authors concluded that Gradient Boosting (GBM) machine learning algorithm was the best classifier in predicting breast using the Coimbra Breast Cancer Dataset (CBCD) with an accuracy of 74.14%.

AML is an aggressive hematologic cancer that causes the building up of immature cells in the blood and bone marrow [14] and could lead to other types of cancer in other organs of the body. In order to predict the hazard of CML after early BC diagnosis, a prediction model approach was presented using Gradient Boosting Machine (GBM) to examine the relationship between breast cancer (BC) and chronic myeloid Leukemia (CML). The tests were applied to the Surveillance, End Results 'SEER' database, and Epidemiology. Female diagnosed with BC from 2010 to 2016 were chosen and tracked for the growth of CML after a lag time of 6 months following BC's diagnosis. The test data set is loaded into the model classifier and then into a Particle Swarm Optimizer (PSO) to optimize the recognition system and find the best hyper-parameter values. To evaluate our prediction GBM model classifier, another state-of-the-art classifiers KNN (k-Nearest Neighbor), SV (Support Vector), and RF (Random Forest)) were implemented and compared with the proposed model. The concept of PSO optimization is to simulate the predation behavior of birds[15]. Each particle is a candidate solution and has a fitness value, position, and speed. Historical knowledge of the optimum solution instructs the particle to travel into a better location.

II. Materials and Method

A. Materials

Data were collected from the US National Cancer Institute's - SEER Database, employing the SEER Stat. Software (Version 8.3.4). The SEER 13 registries have been employed which cover about 13.4 percent of the US population between 1992 and 2016 (based on 2010 census) [16].

B. Methodology

The proposed methodology is presented in figure (1). The method consists of three stages. First Stage: Select attributes, second stage: SEER dataset preprocessing, and the third stage: Development of a classification model and classifier.

1) Selected attributes: The nine attributes were token-based on attributes that are selected in reference [9]. The following subsection provides information about the definitions for the selected attributes from the SEER Dictionary. Table I summarizes the description of Selected Attributes.

TABLE I
SELECTED ATTRIBUTES

Attributes Names	Attributes description
Age at diagnosis	This reflects the patient's age at diagnosis for this cancer.
Race [White, Black and Other**]**	Describes the patient ethnicity estimate.
Marital status [Married, Single, Widowed, Divorced and Separated**]**	Explains the marital status of patients at the time the reportable tumor was diagnosed.
Histology [Ductal and lobular neoplasms and Others**]**	Identifies the anatomical structure of given primary cells and/or tissue.
Grade [Well differentiated; Grade I, moderately differentiated; Grade II, Poorly differentiated; Grade III and Undifferentiated; anaplastic; Grade IV**]**	Describes a tumor in terms of how abnormal the tumor cells are when compared to normal cells.
Derived [Localized, Regional, direct extension, Regional, Distant]	Describes Behaviors associated with the histological description of the neoplasm
Derived HER2 Recode [Positive, Negative and Borderline**]**	The test uses to get information about the status of the HER2 proteins that can play a role in the development of breast cancer
Primary Site	Designates where the main tumor originated.
Laterality [Right and Left]	Defines the side of a paired organ or side of the body on which the reportable tumor developed.

2) SEER dataset preprocessing: A random sample training set of 1200 cases was used, and the classification rule system was then extended to the entire breast cancer dataset. The following steps were performed on the datasets to convert the raw data to appropriate study purposes.

- Extract all patients from the SEER dataset having BC (Breast Cancer only) and patients with CML (Breast Cancer patients who have developed to Chronic Myeloid Leukemia).
- Select 9 attributes that are related to cancer in the previous study.
- Filter cases for the period of interest. The period from 2010 to 2016
- Filter cases that have 6 months lag time between the diagnosis of BC and CML.
- Clean dataset by excluding cases that had an unknown status.
- Convert nominal attributes e.g., marital status, sex into numeric values.
- Derive binary attributes as targets to classify between cases BC and CML, where a value of 0 represents BC cases, while a 1 represents the CML cases.

After the SEER dataset preprocessing, it was found that 450 females with breast cancer have progressed to chronic myeloid Leukemia and 550 females with breast cancer without further progression.

3) Development of a classification model: The classification model for breast cancer classification is achieved by the Gradient Boosting Machine (GBM). Supervised classification methods are employed to construct the classification model. The classification schemes employed in our experiments include the following steps:

- Split the dataset into 70% as the training set, 15% as the validation set, and 15% as the testing set.
- Select the GBM model parameters [Tree depth - Minimum number of observations in terminal nodes] using the PSO optimization algorithm.
- Train the GBM model.
- Evaluation of the model wherein this stage the GBM classifier was compared with another model classifiers.

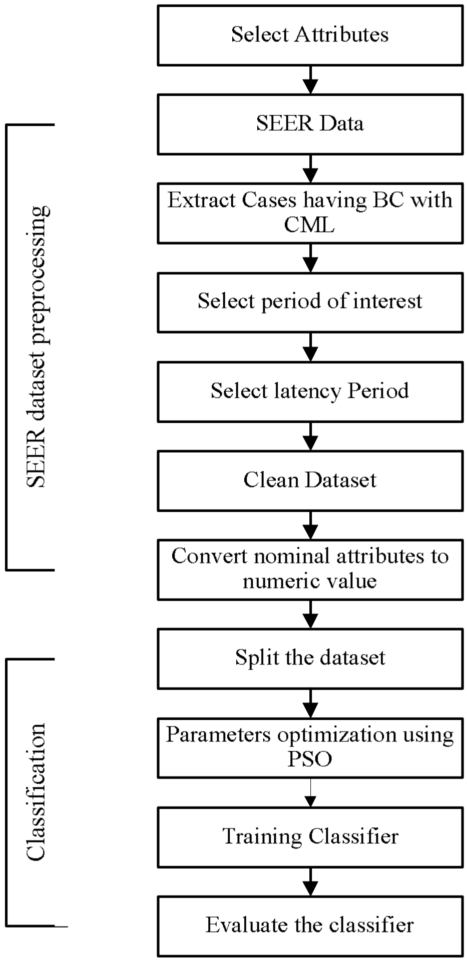

Select Attributes
SEER Data
Extract Cases having BC with CML
Select period of interest
Select latency Period
Clean Dataset
Convert nominal attributes to numeric value
Split the dataset
Parameters optimization using PSO
Training Classifier
Evaluate the classifier
SEER dataset preprocessing
Classification

Fig. 1 Overall methodology steps

C. *Gradient Boosting Machine*

The gradient boosting machine (GBM) is a method for the progressive improvement of error. GBM was drawn up by Freidman [17], who defined guesstimate of the functional dependency, $y = \eta(s(t))$. The loss function explains how stable the classification process is. The best approach for strengthening the categorization model is to put down the loss function in its gradient direction. The classifier in this study was trained using the GBM classifier. The training data set was expressed as M = $\{m_p\ , p = 1,2, \dots \dots , N\}$ where *N* is the total number of patients and expresses the feature vector of m_p selected attributes**.** The loss function $\psi(y, \eta)$ in this model is expressed as follows:

$$\hat{\eta}\big(s(t)\big) = \hat{y} = arg\ \ min\, \psi(y, \eta) \qquad (1)$$

The function estimation, $\hat{y} = \sum_{i=1}^{M} \widehat{y_\iota}$ is parametrized with $\widehat{y_\iota}$ which is a boost. The greedy method was created that estimates $\widehat{y_k} = \widehat{y_{k-1}} + \Delta_k . \xi(\overline{s(t)}, \theta_k)$ at each recursion, where $\xi(\overline{s(t_\iota)}, \theta)$ is called the base learner, which is a decision tree. The function is constructed as follows:

$$(\Delta_k, \theta_k) = arg\ \ min_{\Delta,\theta} \sum_{i=1}^{N} \psi(y^{(i)}, \widehat{\eta_{k-1}}) \ + \ \Delta . \xi\big(\overline{s(t_\iota)}, \theta\big) \quad (2)$$

Since optimization is a difficult problem for the base learner and the general loss function, Friedman recommended a novel function $\xi\big(\overline{s(t)}, \theta\big)$, that is the nearest to being parallel to the negative gradient along with the perceived data, whereby the optimization process is converted to the conventional least square minimization. The GBM algorithm is shown in Table II:

TABLE II
GRADIENT BOOSTING MACHINE ALGORITHM

```
Gradient Boosting Machine Algorithm
Data: n observed data features {T-F features, statistical features $\overline{s(t_\iota)}$}
Process: Compute the loss function $\psi(y, \eta)$ and the base learner classifier $\xi(\overline{s(t)}, \theta)$ to number of iterations M
    1. Build the predictive classifier $\hat{\eta}$ (s(t)) for $\overline{s(t)}$
    2. Initialize $\widehat{\eta_0} = arg\ \ min_{\Delta_k} \sum_{i=1}^{N} \psi(\overline{s(t_\iota)}, \Delta_k)$ for $m \in \{1,2, \dots . , M\}$
    3. Compute the negative gradient $\boldsymbol{\zeta_k(s(t))}$
    4. Fit a new base learner function $\xi(\overline{s(t)}, \theta_k)$
    5. Identify the best gradient descent step-size $\Delta_k$ to obtain a tree classifier.
       $\Delta_k = arg\ \ min_{\Delta\theta} \sum_{i=1}^{N} \psi\left(y^{(i)}, \eta_{k-1}\widehat{(s(t_\iota))}\right) + \Delta . \xi(\overline{s(t_\iota)}, \theta_k)$
    6. Update function $\eta_k = \Delta_k \zeta_k(s(t))$ and the GBM classifier $\eta(\overline{s(t_\iota)}) = \eta_k + \eta_{k-1}$
    end for.
    return $\eta(\overline{s(t_\iota)})$;
```

D. *Analysis and Visualizing the Distribution of a Dataset*

The most important step before developing the classifier is the analysis and understanding of the relations between each selected attribute and target (BC or CML). The following Figures show the distribution of selected attributes corresponding to the targets. From Fig. 2, it was found that patients having certain common variables as (Marital status: Married, Stage: Localized, and Progesterone/Estrogen receptors Status: Negative) are prone to a risk of CML which has increased significantly after the diagnosis of BC. From Fig. 3 it can be observed that the highest incidence of Leukemia after breast cancer was at age 50 and it was observed that it decreased after that age. Fig. 4 shows the incidence of leukemia patients was seen to be greater in cases where the primary site was recorded at the upper-outer quadrant of the breast. On the contrary, the histogram curves for patients prone to Leukemia and those who are not were identical at the region of overlapping lesion of the breast. Fig. 5 shows. Most of the Histology analysis has ductal and lobular neoplasms. There is no major discrepancy between the sides of the breast on which the reportable tumor originated.

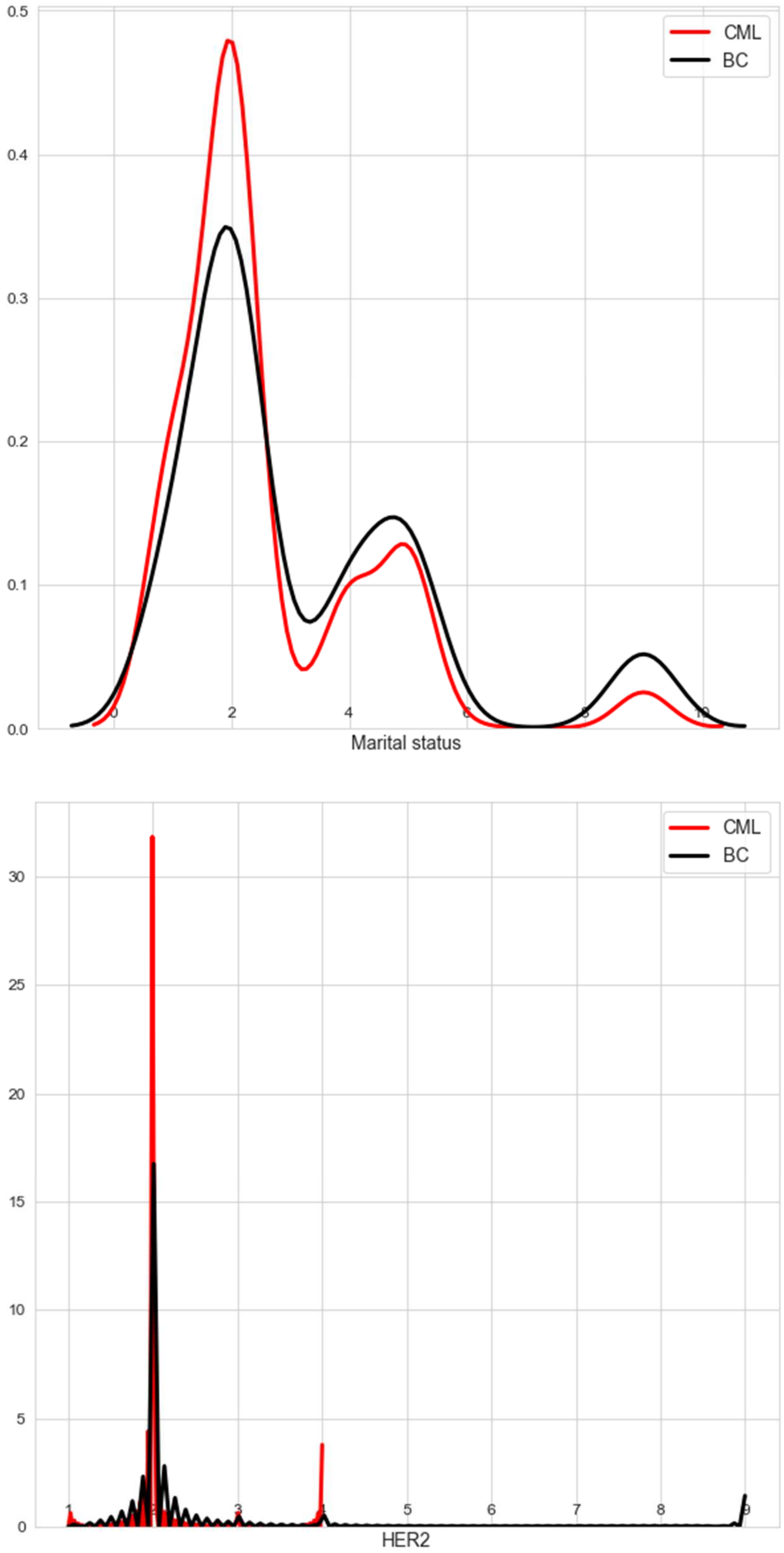


Fig. 2 Distribution curve for Marital status and Derived HER2 Recode features with target values

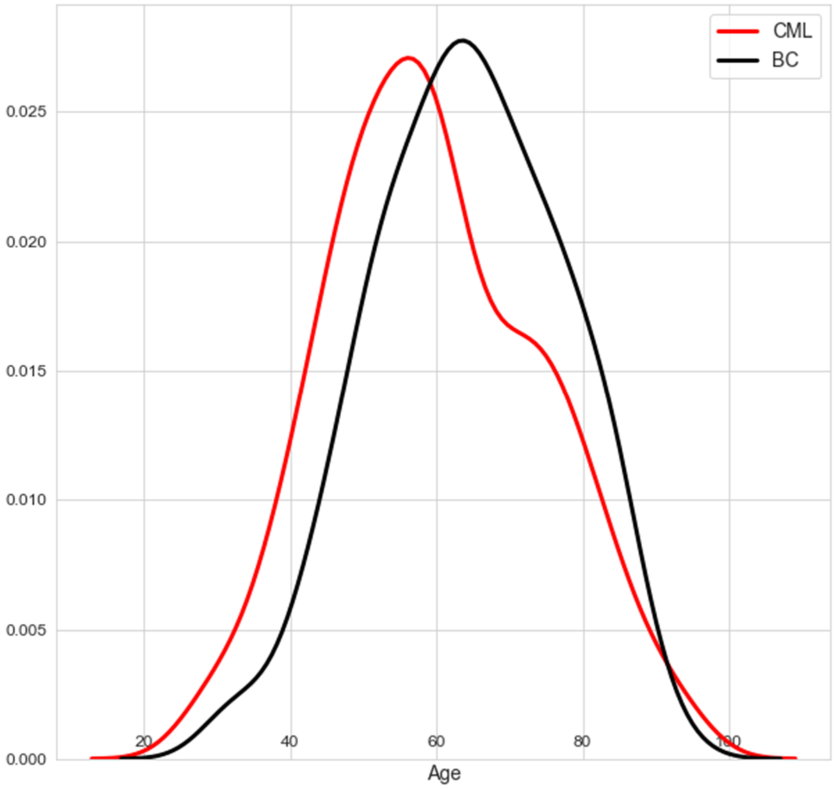


Fig. 3 Distribution curve for Age feature with target values

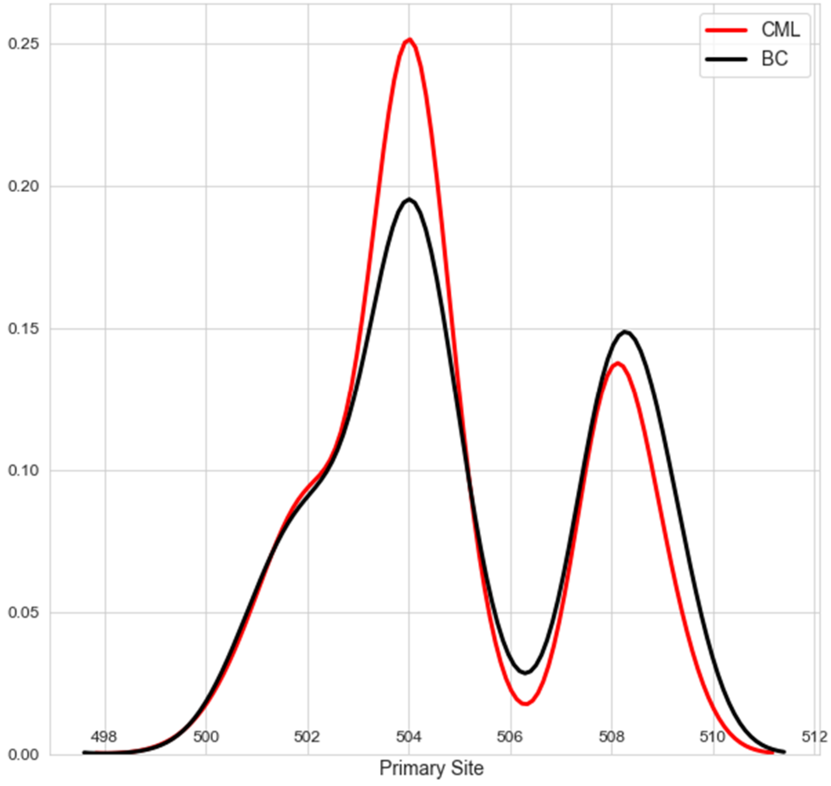


Fig. 4 Distribution curve for Primary Site feature with target values

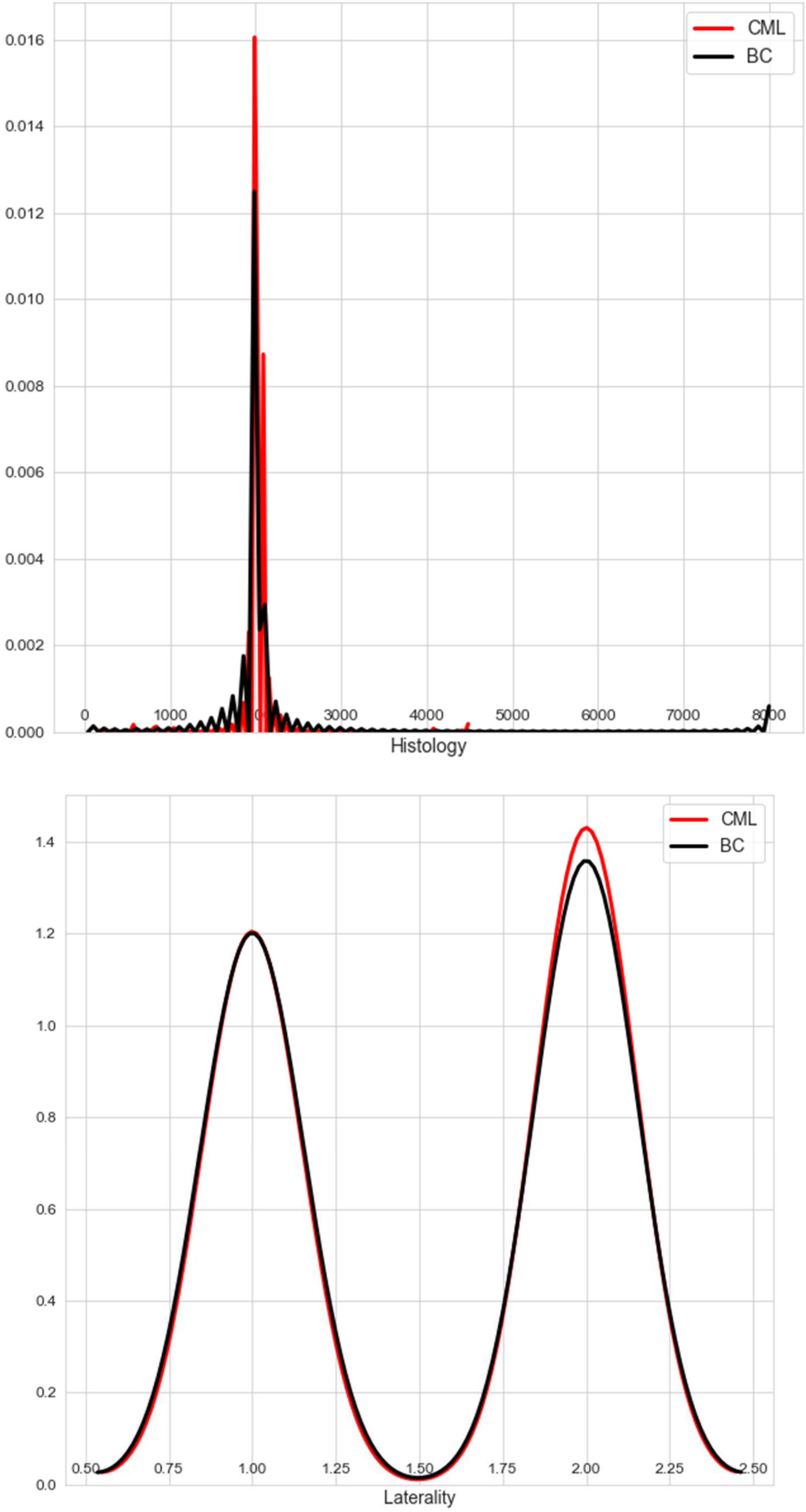


Fig. 5 Distribution curves for Histology and Laterality features with target values

E. *Performance Analysis*

The selected features were employed to train the Gradient Boosting Machine (GBM) classifier then the classifier performance was evaluated in terms of Sensitivity, Specificity, and Accuracy. These terms are calculated as shown in equations (3, 4, and 5). The explanation of the terms of TP, FP, TN, FN are shown in Table III.

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \qquad (3)$$

$$Specifity = \frac{TN}{FP+TN} \times 100\% \qquad (4)$$

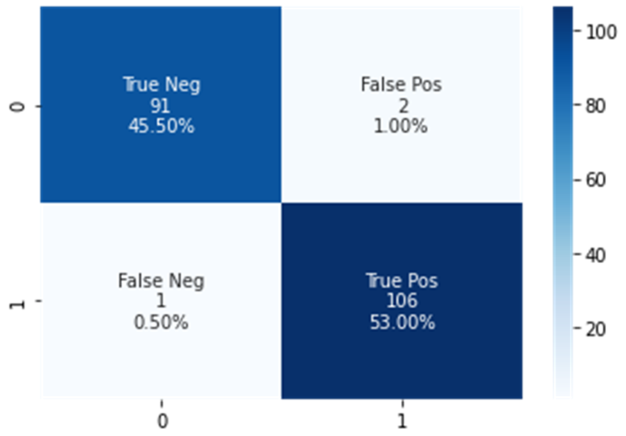
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \qquad (5)$$

TABLE III
TERMS EMPLOYED FOR MEASURING THE PERFORMANCE

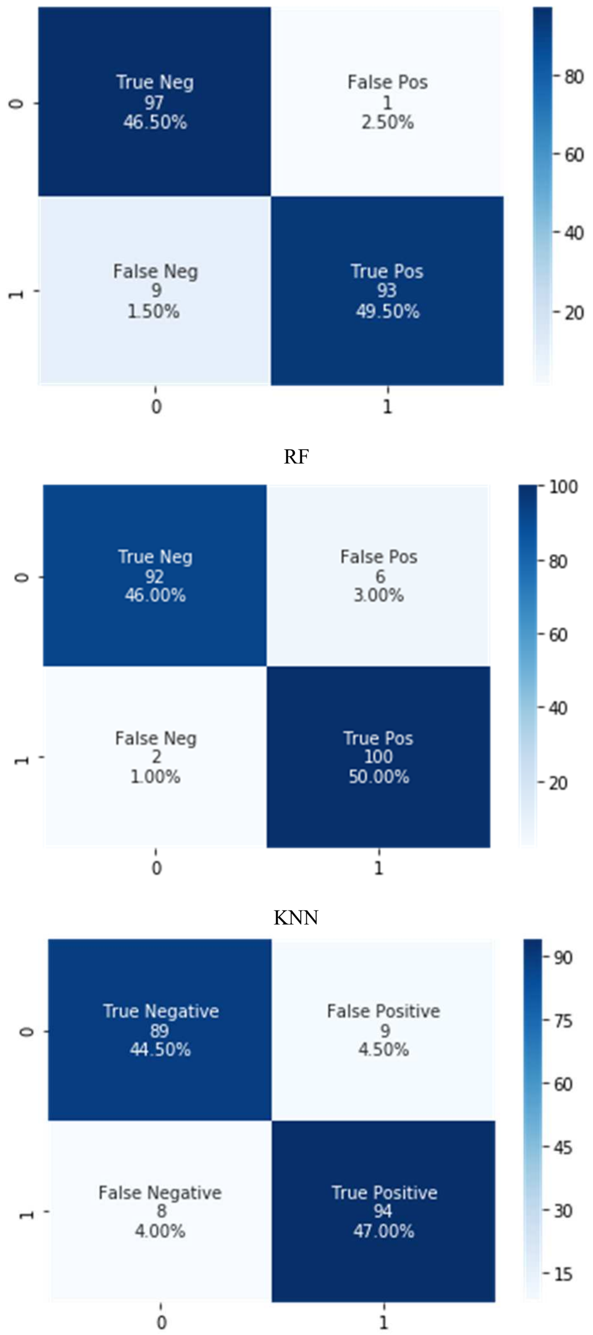
Term	Meaning
True Positives (TP)	The number of positive cases that are correctly categorized as positive cases (CML Patients categorized as CML).
False Positives (FP)	The number of positive cases that are wrongly categorized as negative cases (CML Patients categorized as BC).
True Negatives (TN)	The number of negative cases that are properly categorized as negative cases. (BC patients categorized as BC)
False negatives (FN)	The number of negative cases that are wrongly categorized as positive cases (BC patients categorized as CML)

III. RESULTS AND DISCUSSION

The experiments were performed on a Google Colab kernel. Google Colab is a cloud computational environment that enables to development of deep learning applications using popular libraries such as Keras, TensorFlow, PyTorch, and OpenCV. Besides, Google Colab provides 2 GPU cores, 13 GB of RAM [18]. The predicted values that are produced from the GBM classifier and the counterpart classifiers [KNN (k-Nearest Neighbor), SV (Support Vector), and RF (Random Forest)] are shown in the confusion matrix in Fig. 6 in the purpose of evaluation of our GBM model classifier.



GBM



RF

KNN

SVM

Fig.6 Confusion matrices of GBM classifier and other counterpart classifiers.

The testing data set had a total of 200 cases; it is clear from the confusion matrix that the trained GBM classifier could correctly identify 106 cases as CML patients "patients having breast cancer and leukemia" (TP) and 91 cases are correctly identified as BC patients "patients having breast cancer only" (TN). 2 CML cases were wrongly identified as BC (FP) and 1 BC patient only was wrongly identified as CML patient (FN). Table IV shows the computed sensitivity, specificity, and accuracy for testing data of the Gradient Boosting Machine (GBM) classifier and the other state of art classifiers.

TABLE IV
COMPARISON BETWEEN THE PERFORMANCE OF THE GBM CLASSIFIER AND OTHER STATE OF ART CLASSIFIERS

	GBM	KNN	RF	SVM
Accuracy	0.9850	0.9600	0.9500	0.9150
Sensitivity	0.9907	0.9804	0.9118	0.9216
Precision	0.9815	0.9434	0.9894	0.9126
Specificity	0.9785	0.9388	0.9898	0.9082
F1 Score	0.9860	0.9615	0.9490	0.9171

Table IV's results indicate that the GBM classifier could classify BC and CML patients by an achieved accuracy, sensitivity, and specificity rates of 98.5 %, 99 %, and 97.85 %, respectively. It is also noticeable that the GBM classifier attained a significantly higher accuracy than the corresponding counterpart classifiers. The better performance of the GBM classifier as compared to other state-of-the-art classifiers reveals the ability of GBM to classify the applied input into either of the two classes BC patients or CML patients and therefore predict the CML patients after early BC diagnosis.

To evaluate the proposed model, the F1 score was calculated (weighted harmonic mean of the test's precision and recall), shown in Table IV. The GBM classifier had the highest F1 score due to high values of precision and recall. So, by evaluating and comparing the overall performance of the presented classifiers, it can be observed that the GBM predictive model had the superior classification and prediction ability of breast cancer than the other counterparts.

IV. CONCLUSION

This study showed the significant ability of the Gradient Boosting Machine (GBM) classifier to classify SEER breast cancer data of BCs “breast cancer only” and CML “breast cancer with leukemia” patients. A training set of a random sample of 1200 cases was employed, and then applied the categorization rule set obtained to the full breast cancer dataset. The results showed that the predictive model had the ability to classify between BC and CML by achieved accuracy, sensitivity, and specificity rates of 98.5 %, 99 %, and 97.85 %, respectively. The GBM algorithm's performance was further compared with other models that revealed the superior ability of the GBM classifier in the classification of breast cancer disease and the prediction of patients having Leukemia after having early breast cancer. Future enhancement of this work includes an increase in the number of features to find the best features in the disease identification dataset.

REFERENCES

[1] R. Shenolikar, E. Durden, N. Meyer, G. Lenhart, and K. Moore, “Incidence of secondary myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) in patients with ovarian or breast cancer in a real-world setting in the United States,” *Gynecol. Oncol.*, vol. 151, no. 2, pp. 190–195, 2018.

[2] I. Vanidassane, A. Gogia, V. Raina, and R. Gupta, “Treatment Related Acute Myeloid Leukemia in Breast Cancer Survivors: A Single Institutional Experience,” *Indian J. Hematol. Blood Transfus.*, vol. 35, no. 3, pp. 561–562, 2019.

[3] A. Matikas *et al.*, “Long-term safety and survival outcomes from the Scandinavian Breast Group 2004-1 randomized phase II trial of tailored dose-dense adjuvant chemotherapy for early breast cancer,” *Breast Cancer Res. Treat.*, vol. 168, no. 2, pp. 349–355, 2018.

[4] H. Choi *et al.*, “A Case of Preleukemic Chronic Myeloid Leukemia Following Chemotherapy and Autologous Transplantation for T-lymphoblastic Lymphoma.,” *Ann. Lab. Med.*, vol. 40, no. 5, pp. 417–420, 2020.

[5] I. Lalya, I. Essadi, R. Belbaraka, A. El Omrani, and M. Khouchani, “Acute Myeloid Leukemia After Treatment of Early Breast Cancer: Case Report and Literature Review,” *Indian J. Gynecol. Oncol.*, vol. 17, no. 3, p. 60, 2019.

[6] T. Radivoyevitch *et al.*, “Risk of acute myeloid leukemia and myelodysplastic syndrome after autotransplants for lymphomas and plasma cell myeloma,” *Leuk. Res.*, vol. 74, pp. 130–136, 2018.

[7] M. Payandeh, R. Khodarahmi, M. Sadeghi, and E. Sadeghi, “Appearance of acute myelogenous leukemia (AML) in a patient with breast cancer after adjuvant chemotherapy: case report and review of the literature,” *Iran. J. cancer Prev.*, vol. 8, no. 2, p. 125, 2015.

[8] A. Balduzzi and M. Castiglione-Gertsch, “Leukemia risk after adjuvant treatment of early breast cancer,” *Women’s Heal.*, vol. 1, no. 1, pp. 73–85, 2005.

[9] M. J. Al-Husseini *et al.*, “Risk and survival of chronic myeloid leukemia after breast cancer: A population-based study,” *Curr. Probl. Cancer*, vol. 43, no. 3, pp. 213–221, 2019.

[10] S. V. Ezhilraman, S. Srinivasan, and G. Suseendran, “Breast Cancer Detection using Gradient Boost Ensemble Decision Tree Classifier.”

[11] T. Mitchell and M. L. McGraw-Hill, “Edition.” New York: McGraw-Hill, Inc, 1997.

[12] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” *PLoS One*, vol. 13, no. 3, p. e0194889, 2018.

[13] Y. D. Austria, P. L. Jay-ar, L. B. S. Maria Jr, J. E. E. Goh, M. L. I. Goh, and H. N. Vicente, “Comparison of Machine Learning Algorithms in Breast Cancer Prediction using the Coimbra Dataset,” *Cancer*, vol. 7, p. 10, 2019.

[14] H. G. Kaplan, G. S. Calip, and J. A. Malmgren, “Maximizing Breast Cancer Therapy with Awareness of Potential Treatment-Related Blood Disorders,” *Oncologist*, vol. 25, no. 5, p. 391, 2020.

[15] H. Chen, L. Cui, Q. Guo, and J. Zhang, “Improved Particle Swarm Optimization Using Wolf Pack Search,” in *Journal of Physics: Conference Series*, 2019, vol. 1176, no. 5, p. 5, 2009.

[16] A.-A. RATES and A.-S. RATES, “SEER cancer statistics review 1975-2005,” 2008.

[17] S. I. Abed, “Predicting Breast Cancer Using Gradient Boosting Machine,” vol. 8, no. 6, pp. 885–891, 2019.

[18] A. Talele, A. Patil, and B. Barse, “Detection of Real Time Objects Using TensorFlow and OpenCV,” *Asian J. Converg. Technol.*, 2019.