Indian Journal of Biochemistry & Biophysics
Vol. 63, February 2026, pp. 181-189
DOI: 10.56042/ijbb.v63i2.19188

Enhancing diagnostic precision and accuracy in invasive lobular carcinoma through machine learning approaches

Priya LS[1], Shri ST[1], Swathi J[1], Premnath D[2] & Indiraleka M[1]*

[1]Mepco Schlenk Engineering College (Autonomous), Sivakasi, Virudhunagar-626 005, Tamil Nadu, India

[2]School of Agriculture and Biosciences, Karunya Institute of Technology and Sciences (Deemed to be University) Coimbatore-641 114, Tamil Nadu, India

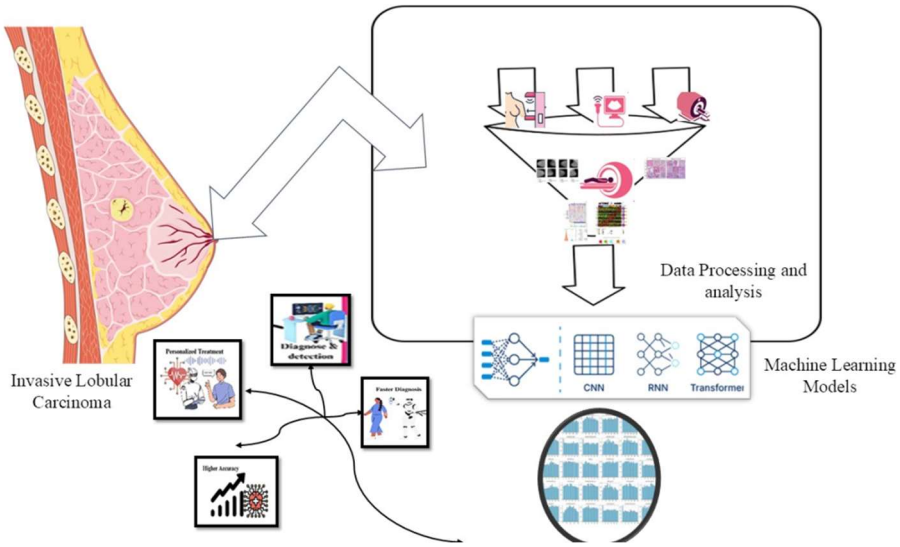
Received 27 April 2025; *revised 13 May 2025*

Invasive Lobular Carcinoma (ILC) is a type of breast cancer that forms in the lobules of the breast and is characterized by small, non-cohesive cells that invade surrounding tissues in a unique pattern. Invasive Lobular carcinoma mostly affects women compared to men. Various techniques are available to detect the presence of ILC, like mammography, Ultrasound, and MRI. Invasive lobular carcinoma is not present in a mass, making it difficult to detect ILC in some imaging techniques. Machine learning (ML) techniques are being used to improve the prediction and diagnosis of ILC. It involves data collection from electronic health records, imaging studies, and genomic data from Kaggle, and using different models, such as supervised learning and unsupervised learning, to predict ILC. In this current study, various algorithms have been used to predict and improve the accuracy and precision level of ILC diagnosis. Results found that Elastic Net and Logistic Regression have shown higher accuracy. ML is very useful for radiologists, oncologists, and patients in early-stage prediction of ILC, which is helpful in personalizing treatment plans.

Keywords: Invasive lobular carcinoma, Machine learning, Diagnosis, Precision and accuracy

Different types of breast cancer include Invasive Ductal Carcinoma, Invasive Lobular Carcinoma (ILC), Ductal Carcinoma In Situ, Lobular Carcinoma In Situ, and Inflammatory Breast Cancer (IBC)[1,2]. ILC, which originates from the milk-producing lobules of the breast, is noted to be the most common type of breast cancer. It spreads from lobules to other parts of normal tissue through the blood and lymph systems[3,4]. Symptoms of ILC may first manifest as a thickening or hardening of the breast, although many women with ILC exhibit no visible symptoms. Other signs include changes in breast size, nipple swelling, and breast irritation. Various imaging techniques are available to detect the presence of carcinoma; however, due to the thickening or hardening of connective tissue, it is often difficult to identify ILC, as it does not typically present as a discrete mass[5–7]. The diagnosis involves thorough evaluation of imaging studies, clinical assessments and biopsies. The healthcare providers also consider the complete medical history, including any lumps in the breast, abnormal changes in shape or size, or discharge from the nipple, in diagnosis. The thorough physical examination will assess abnormalities in bone and lymph nodes[8,9]. Further detailed examination of tissue or fluid through mammography, ultrasound, and MRI, and fine-needle aspiration (FNA) could be used to obtain imaging modalities. However, histopathology and immunohistochemistry studies are crucial for characterising the cancer cells and validating the diagnosis[10]. Inconsistencies in subtle growth patterns, imaging limitations and the inter-observer variability among clinicians were observed in diagnosis[11,12]. This variability is due to the limited knowledge, inadequate training among healthcare workers and inadequate screening procedures. ILC typically shows a diffuse, non-mass-forming growth pattern, making it less detectable on mammograms. Machine learning (ML) is an evolving field in artificial intelligence that involves training machines and perform functions using algorithms and statistical models, without plainly written computer programs. Such systems are trained by analysing data sets to make predictions and improve performance over time[13,14]. Supervised, unsupervised, semi-supervised, and reinforcement learning are some common forms of machine learning approaches these employing algorithms like linear regression, decision trees, random forests, support

*Correspondence:
Phone: +91-7402389414 (Mob)
E-mail: indirajith1812000@gmail.com



Graphical abstract

vector machines (SVM), and neural networks. Though they offer substantial potential for innovation and efficiency in healthcare, the complexities of these systems can pose few challenges[15,16]. This current research aims to develop and refine machine learning algorithms to achieve high accuracy in diagnosing Invasive Lobular Carcinoma (ILC) by incorporating various algorithms to improve early detection rates and reduce diagnostic errors.

Materials and Methods

Dataset extraction

The relevant datasets were extracted from Kaggle (https://www.kaggle.com/). The datasets were filtered based on dataset type, number of kernels, and how recently they were uploaded. Then, the promising datasets were chosen and downloaded in CSV or JSON format.

Data cleaning/set enrichment

The main focus of the data cleaning is to identify the feature, that makes inconsistencies in the provided data, correcting the inaccurate values. In this process, duplicate rows/columns and the rows/ columns having null values have been removed. Data consistency was verified to ensure all the columns have consistent data. Grouping of the data or group aggregation is done, which helps in increasing the predictive power. Highly correlated data has been removed, because it complicates further process. Resampling and class weight adjustments are done to handle the imbalanced data[17].

Dataset pre-processing

The data have been split into training and testing sets. “The Unnamed” column with missing values has been removed or cleared. The data has been converted from the categorical form into the numerical values. For example, in the data if “M” is given it will be converted to “0”, if “F” is given it is converted to “1”. Another crucial step is finding the outlier and, handling the outlier in the given numerical data. Selection of the feature is important. The most significant features must be selected and the given data is split into training and testing. Training the data is 70% and testing the data is 30%[18–20].

Correlation matrix

Correlation heatmaps are important in determining which variables may potentially cause multicollinearity that would damage model integrity. It is a graphical illustration of how each variable in the dataset correlated with its peers (Fig. 1). Brighter colours-for example, red indicate a stronger positive correlation, and darker colors, for example, blue indicate a stronger negative correlation. The diagonal line in a correlation heatmap is usually a solid line or the highest intensity on the colour scale, indicating how every variable is perfectly correlated with itself, equal to 1. This would be seen such that features are columns and their importance scores are coloured. This is important for identifying the most relevant features for a given task. A correlation degree of -1 implies a zero-degree correlation and a degree of 1 implies a perfect correlation[21,22].

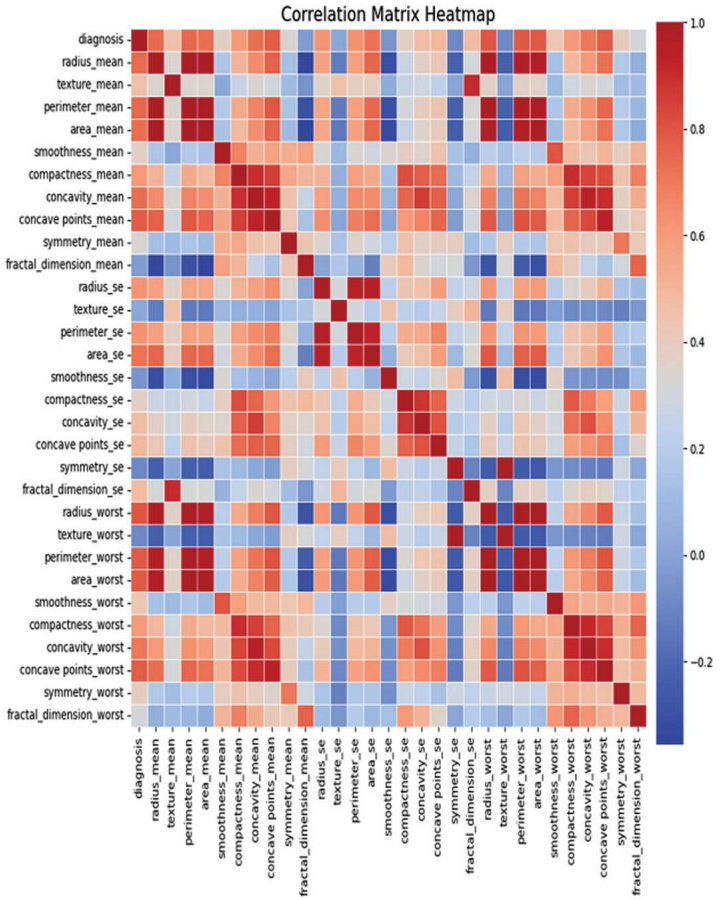


Fig. 1 — Correlation matrix heat map in the diagnosis of Invasive Lobular Carcinoma using Machine learning

Histogram analysis

A histogram is a graphical representation of a grouped frequency distribution with continuous classes. It is an area representation and is described as a collection of blocks whose bases correspond to the intervals between the class boundaries and whose areas correspond to the frequency in the corresponding classes. For similar donations, all blocks are congruent, as the base covers the intervals between the class boundaries. The heights of the blocks are appropriate to the corresponding frequency of the analogous classes, and for different classes, the heights are appropriate to the corresponding frequency densities. The intervals that divide the range of the data are called lockers. The range of the lockers can change the appearance of the histogram. A wider caddy helps to remove noise and reveal more general patterns, but narrow lockers inevitably lead to further noise or variability. The number of data points that fall into each of the lockers is called the frequency and is shown in (Figs. 2-4).

Machine learning algorithms

In this current research, the following supervised and unsupervised learning approaches were employed in the diagnosis of ILC based on various features[23–27]. Linear Regression, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor algorithm (KNN), Neural Network, Gradient Boosting, Gaussian Naïve Bayes, Stochastic Gradient Descent Classifier (SGDC), Elastic Net, Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-sne), Linear Discriminant Analysis (LDA), Lasso regression, Adaptive Boosting.

Model training and testing

Training a model is a process by which a machine-learning model is constructed to identify patterns in data. This process uses a training dataset to calibrate the parameters of the model. A labelled example in the training dataset- those used to instruct this model as it learns- consists of features and the respective output that those features map to. The quality and

Fig. 2 — Bar graph representation of Histogram analysis in the diagnosis of Invasive lobular carcinoma using Machine learning

quantity of the training data have a major influence on how well the model works. Overfitting refers to the training data being fit too tightly by the model so that it learns noise rather than the underlying patterns. Poor performance on unseen data will be the outcome. Underfitting occurs when the model performs poorly even when training data is used because it is too simplistic to reflect the underlying structure of the data. Model testing checks the learned model's performance on another separate dataset that it has never encountered, called a testing dataset. This helps in estimating how well the model generalises to new data. A testing dataset forms a part of the original data distinct from the training dataset. It should be representative of the problem space to properly assess the performance of the model[28,29].

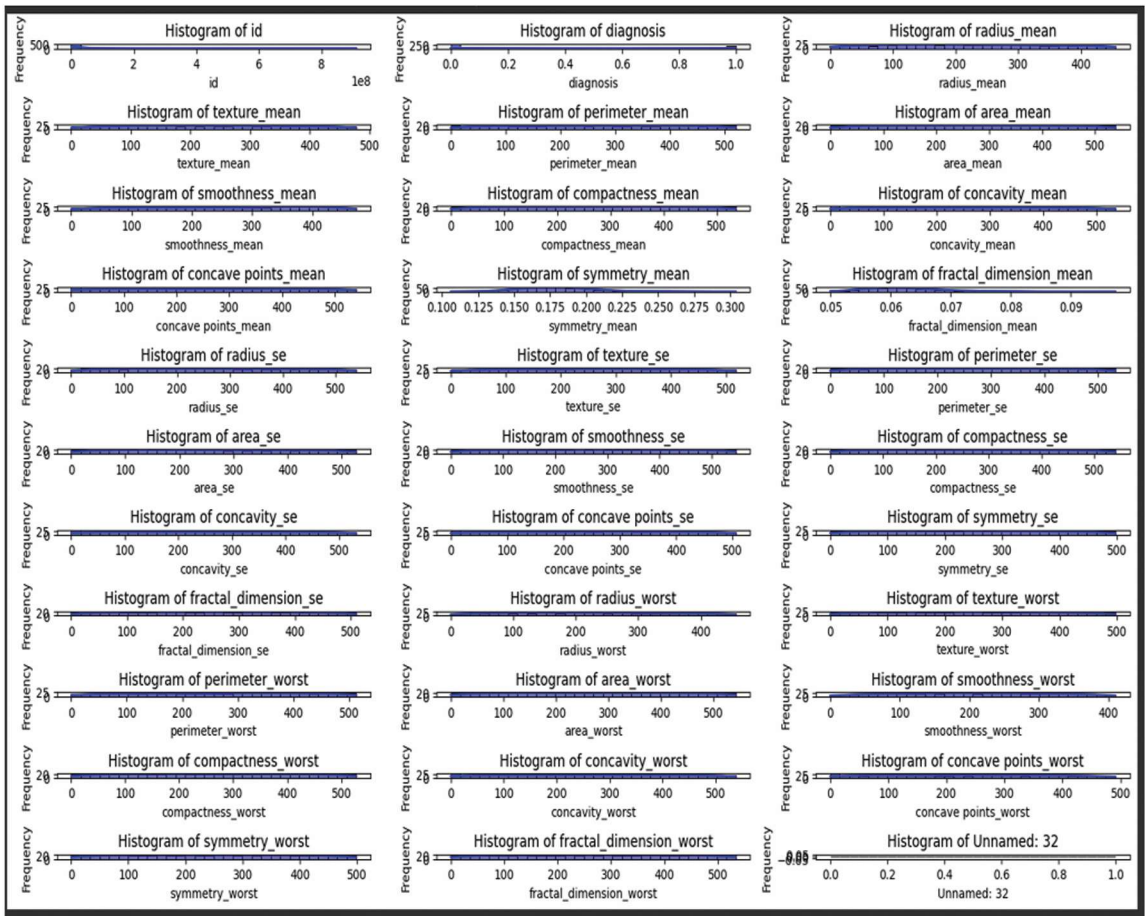


Fig. 3 — Frequency range of histogram analysis in the diagnosis of Invasive lobular carcinoma using Machine Learning

Model evaluation

Confusion Matrix

The confusion matrix is a tabular representation of the performance of a classification algorithm, summarizing true positives, false positives, true negatives, and false negatives. The confusion matrix is hugely helpful for binary and multi-class classification problems.

True Positives (TP): The number of rightly predicted positive instances.

True Negatives (TN): The number of instances that were correctly classified as negative.

False Positives (FP): Mis-classified number of negative cases as positive (Type I Error).

False Negatives (FN) is described by the number of true positives is misclassified as false negatives and this represents type II error.

The metrics including accuracy, precision, recall, and F1-score were used to assess the model. This is important to evaluate how effectively the model differentiates ILC from other conditions.

Accuracy is determined by calculating the ratio of correct predictions to the overall number of predictions made and this describes how frequently a model produces correct predictions.

$$ACCURACY=(TP+TN)/(TP+FP+FN+TN)$$

Precision shows how many of the instances that the model identified as positive are indeed positive and it is determined by calculating the ratio of true positives to the total of true positives and false positives.

$$Precision=True\ \ Positives\ \ (TP)/False\ \ Positives\ (FP)+True\ Positives\ (TP)$$

The recall refers to the proportion of actual positive instances that the model correctly identifies. It is determined by calculating the ratio between true positives to the sum of true positives and false negatives.

$$Recall\ =\ True\ Positive\ (TP)\ /\ True\ Positive\ (TP)\ +\ False\ Negative\ (FN)$$

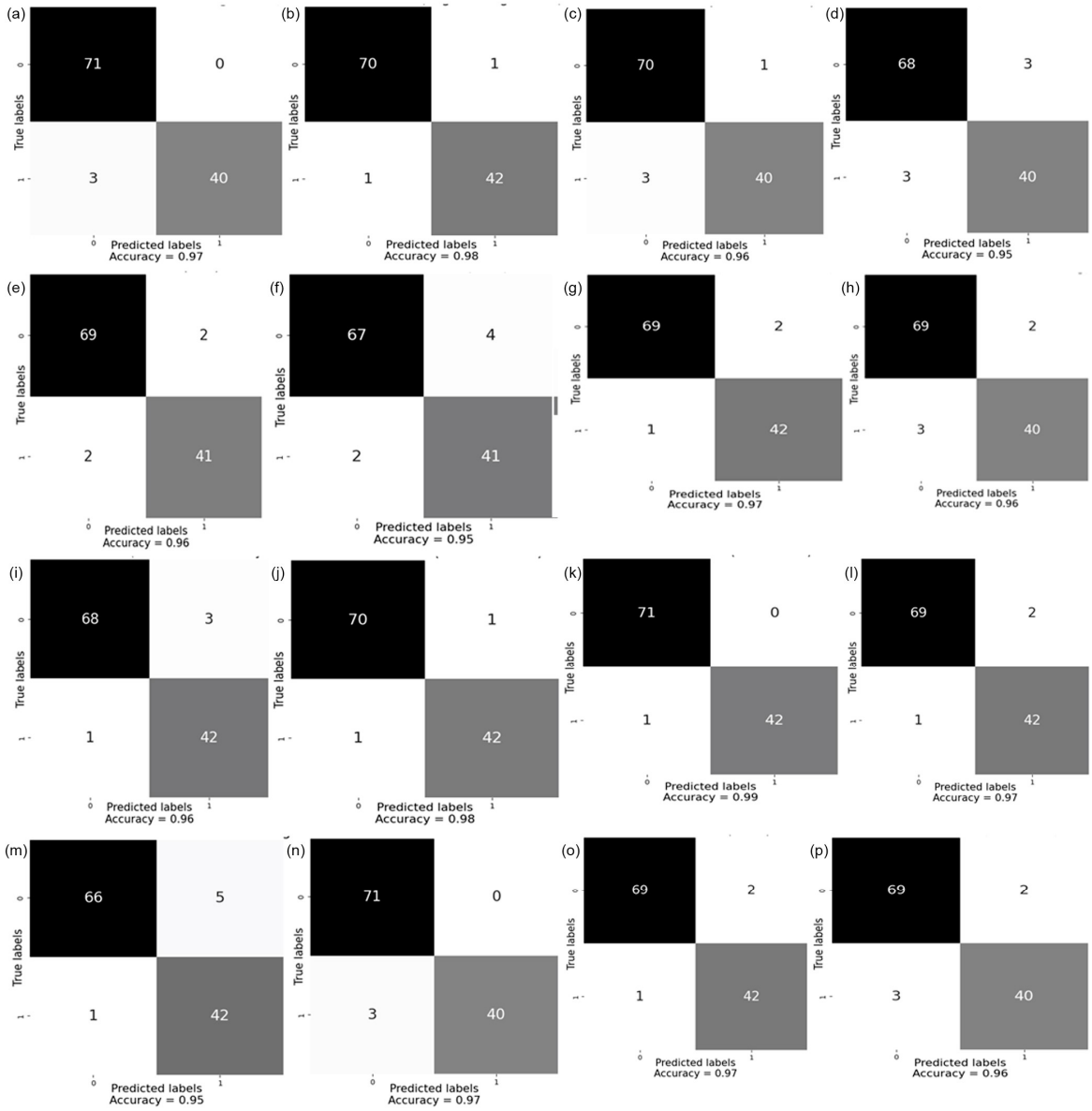


Fig. 4 — Confusion matrix analysis: (a) Linear regression; (b) Logistic regression; (c) Random forest; (d) Decision tree; (e) SVM; (f) KNN; (g) Neural network; (h) Gradient boosting; (i) Gaussian naïve bayes; (j) SGDC; (k) Elastic net; (l) PCA; (m) t-sne; (n) LDA; (o) Lasso regression; and (p) Adaptive boosting

F1-Score is a metric that balances precision and recall, offering a single measure of a model's performance. It is determined by calculating the harmonic mean of precision and recall, treating both metrics with equal importance.

$F1 = 2 * (precision * recall) / (precision + recall)$

Results and Discussion

The performance of various machine learning models was tested by reporting their accuracy rates in making outcome predictions. Summarizes the accuracy rates of models analyzed in this study. By performing a Confusion Matrix for each Machine Learning Algorithm (ML) for diagnosing Invasive Lobular Carcinoma (ILC), different metrics such as accuracy, precision, recall, and F1-score were obtained. The accuracy level we obtain for ILC using Machine Learning (ML) is increased from the previous work done[30]. Confusion matrix for ILC using Linear Regression showed TP=71, FP=0, FN=3, TN=40. Linear regression obtains an accuracy level of 97% (Fig. 4a). Confusion matrix for ILC using

Logistic Regression showed TP=70, FP=1, FN=1, TN=42 (Fig. 4b). Logistic regression obtains an accuracy level of 98%. Confusion matrix for ILC using Random Forest showed TP=71, FP=3, FN=40, TN=96. Random Forest obtains an accuracy level of 96% (Fig. 4c). The confusion matrix for ILC using the Decision Tree showed TP=68, FP=3, FN=3, TN=40. The decision tree obtains an accuracy level of 95% (Fig. 4d). Confusion matrix for ILC using a Support Vector Machine (SVM) showed TP=69, FP=2, FN=2, TN=41. Support Vector Machine (SVM) obtained an accuracy level of 96% (Fig. 4e) Confusion matrix for ILC using K-Nearest Neighbor (KNN) showed TP=67, FP=4, FN=2, TN=41. K-Nearest Neighbor (KNN) obtained an accuracy level of 94% (Fig. 4f) Confusion matrix for ILC using Neural Network showed TP=69, FP=2, FN=1, TN=42. Neural Networks obtained an accuracy level of 97% (Fig. 4g). Confusion matrix for ILC using Gradient Boosting showed TP=69, FP=2, FN=3, TN=40. Gradient Boosting obtained an accuracy level of 96% (Fig. 4h). Confusion matrix for ILC using Gaussian Naive Bayes showed TP=68, FP=3, FN=1, TN=42. Gaussian Naive Bayes obtained an accuracy level of 96% (Fig. 4i). Confusion matrix for ILC using Stochastic Gradient Descent Classifier showed TP=70, FP=1, FN=1, TN=42. The stochastic Gradient Descent Classifier obtained an accuracy level of 98% (Fig. 4j). The confusion matrix for ILC using the Elastic Net Model showed TP=71, FP=0, FN=1, TN=42. The Elastic Net Model obtained an accuracy level of 99% (Fig. 4k). Confusion matrix for ILC using Principal Component Analysis showed TP=69, FP=2, FN=1, TN=42. Principal Component Analysis obtained an accuracy level of 97% (Fig. 4l). Confusion matrix for ILC using T-distributed Stochastic Neighbor Embedding showed TP=66, FP=5, FN=1, TN=42. T-distributed Stochastic Neighbor Embedding obtained an accuracy level of 95%. (Fig. 4m). Confusion matrix for ILC using Linear Discriminant Analysis showed TP=71, FP=0, FN=3, TN=40. Linear Discriminant Analysis obtained an accuracy level of 97%. (Fig. 4n). The confusion matrix for ILC using Lasso Regression showed TP=69, FP=2, FN=1, TN=42. Lasso Regression obtained an accuracy level of 97% (Fig. 4o). Confusion matrix for ILC using Adaptive Boosting Classifier showed TP=69, FP=2, FN=3, TN=40. The Adaptive Boosting Classifier obtained an accuracy level of 96% (Fig. 4p). The results show the importance of proper model selection concerning their application based on the performance metrics.

Different types of machine learning models for the prediction of Invasive Lobular Carcinoma (ILC) have been analyzed to yield promising results, thus stamping the effectiveness of these techniques toward increasing the accuracy of diagnosis. The accuracy rates obtained in this current research using different algorithms have shown significant improvements compared with existing studies. Therefore, it demonstrates the significance of machine learning and its applications in medical diagnostics. The Linear Regression model was highly accurate at 97%, the confusion matrix showed good performance with 71 true positives, 40 true negatives, 3 false negatives, and no false positives. Similarly, no significant variation between the Logistic Regression and the Linear Regression. It has a small margin of difference with an accuracy of 98%. This shows that Logistic Regression is perfect in binary classification problems and the model will be very useful in medical diagnostics, where a high price might be lost by misclassification. Random Forest is another strong model that showed 96% accuracy (Fig. 5). Though it is not as good as Logistic Regression, this model still performed well and showed lots of potential for being robust in complex data through its ensemble approach. The Decision Tree model showed good prediction accuracy of 95% however, it provides more false positives and negatives. It may thus be more prone to overfitting or more sensitive to noise in the data. 96% and 94% accuracy were obtained in SVM and KNN models, respectively. Overall, the performance of the SVM model indicates that this model can accommodate non-linear decision boundaries, an important aspect of cancer diagnosis. The poorer accuracy in KNN may be due to its susceptibility to irrelevant features and the curse of dimensionality, where distance metrics are no longer meaningful in high-dimensional spaces. Both Neural Networks and Elastic Net demonstrated very impressive accuracies in their respective fits, so good that we can consider them to be potential candidates for sophisticated pattern recognition in medical datasets. As expected, the Elastic Net Model proved to be the closest in terms of accuracy, hence again proof of the effectiveness of lasso and ridge regression techniques put together with regularization that may help cut off overfitting. All three models,

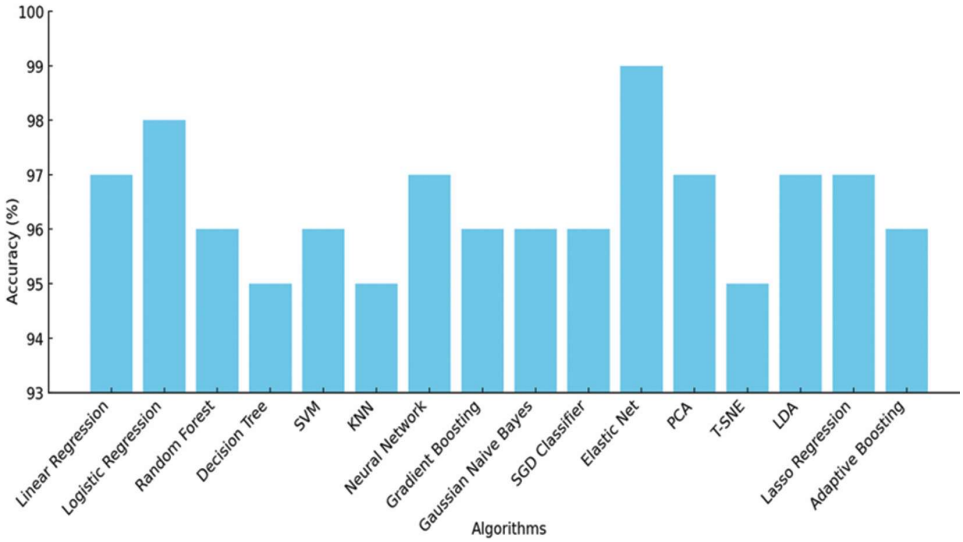


Fig. 5 — Bar graph representation of comparing the algorithms and accuracy percentage

Gradient Boosting, Gaussian Naive Bayes, Stochastic Gradient Descent Classifier, and Linear Discriminant Analysis-reached 96%, 98%, and 97% accuracy levels, respectively. The importance that the ensemble methods and ranking methods bring to predictive performance underlines the need for the proper selection of a model pertinent to the specific application context.

Conclusion

Based on the performance metrics profiled in this study, while each of the machine learning models has strengths and weaknesses, a tailored approach that considers the characteristics of the dataset and the relevant clinical implications of the diagnostic outcomes can significantly enhance overall predictive performance in ILC. Findings support earlier unreliable evidence of the importance of model selection in machine learning-based systems and provide further avenues for research focused on using these approaches to enhance clinical decision-making processes. The future work shall explore the integration of such models into clinical workflows and methods for mitigating false positive events with the ultimate goal of providing enhanced patient care in the diagnosis of ILC and, by extension, potentially other cancers.

Acknowledgement

All the authors thank the management and the authorities of Mepco Schlenk Engineering College, Sivakasi, for providing all the facilities to conduct and complete this research.

Conflict of interest

All authors declare no conflict of interest.

References

1 Thomas M, Kelly ED, Abraham J & Kruse M, Invasive lobular breast cancer: A review of pathogenesis, diagnosis, management, and future directions of early stage disease. *Semin Oncol*, 46 (2019) 121.

2 Wilson N, Ironside A, Diana A & Oikonomidou O, Lobular Breast Cancer: A Review. *Front Oncol*, 10 (2020) 591399.

3 Luveta J, Parks RM, Heery DM, Cheung K-L & Johnston SJ, Invasive Lobular Breast Cancer as a Distinct Disease: Implications for Therapeutic Strategy. *Oncol Ther*, 8 (2020) 1.

4 Mouabbi JA, Hassan A, Lim B, Hortobagyi GN, Tripathy D & Layman RM, Invasive lobular carcinoma: an understudied emergent subtype of breast cancer. *Breast Cancer Res Treat*, 193 (2022) 253.

5 Iorfida M, Maiorano E, Orvieto E, Maisonneuve P, Bottiglieri L, Rotmensz N, Montagna E, Dellapasqua S, Veronesi P, Galimberti V & Luini A, Invasive lobular breast cancer: subtypes and outcome. *Breast Cancer Res Treat*, 133 (2012) 713.

6 Lesnikoski BA, Crozier JA, Srkalovic G, Robinson PA, Osipo C, Banda K, Kling HM, Yoder E, Audeh W & FLEX Investigators Group, Molecular profiles and treatment recommendations for invasive lobular carcinoma in a real-world prospective breast cancer registry. *JCO*, 38 (2020) e19291.

7 Anitha S, Nandhini S, Premnath D & Indiraleka M, Computational Approach to Identify the Key Genes for Invasive Lobular Carcinoma (ILC) Diagnosis and Therapies. *J Comput Biophys Chem*, 2024; 23 (2024), 403.

8 McCart Reed AE, Kalinowski L, Simpson PT & Lakhani SR, Invasive lobular carcinoma of the breast: the increasing importance of this special subtype. *Breast Cancer Res*, 23 (2021).

9 Van Baelen K, Geukens T, Maetens M, Tjan-Heijnen V, Lord CJ, Linn S, Bidard FC, Richard F, Yang WW, Steele RE & Pettitt SJ, Current and future diagnostic and

treatment strategies for patients with invasive lobular breast cancer. *Ann Oncol*, 33 (2022) 769.

10 Corso G, Fusco N, Guerini-Rocco E, Leonardi MC, Criscitiello C, Zagami P, Nicolò E, Mazzarol G, La Vecchia C, Pesapane F & Zanzottera C, Invasive lobular breast cancer: Focus on prevention, genetics, diagnosis, and treatment. *Semin Oncol*, 51 (2024) 106.

11 Helvie MA, Paramagul C, Oberman HA & Adler DD, Invasive lobular carcinoma. Imaging features and clinical detection. *Invest Radiol*, 28 (1993) 202.

12 Mukhtar RA & Chien AJ, Invasive Lobular Carcinoma of the Breast: Ongoing Trials, Challenges, and Future Directions. *Curr Breast Cancer Rep*, 13 (2021) 164.

13 Topol E, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. 1st ed. USA: Basic Books, Inc.; 2019.

14 Vy VPT, Yao MM-S, Khanh Le NQ & Chan WP, Machine Learning Algorithm for Distinguishing Ductal Carcinoma In Situ from Invasive Breast Cancer. *Cancers*, 14 (2022) 2437.

15 Chen J, Hao L, Qian X, Lin L, Pan Y & Han X, Machine learning models based on immunological genes to predict the response to neoadjuvant therapy in breast cancer patients. *Front Immunol*, 13 (2022) 948601.

16 Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM & Thrun S, Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542 (2017) 115.

17 Essa HA, Ismaiel E & Hinnawi MF, Feature-based detection of breast cancer using convolutional neural network and feature engineering. *Sci Rep*, 14 (2024) 22215.

18 Ghavidel A & Pazos P, Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review. *J. Cancer Surviv*, 19 (2025) 270.

19 Naji MA, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA & Debauche O, Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Comput Sci*, 192 (2021) 487.

20 Zuo D, Yang L, Jin Y, Qi H, Liu Y, Ren L, Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Med Inform Decis Mak,* 29 (2023) 276.

21 Kaur G, Gupta R, Hooda N & Gupta NR, Machine Learning Techniques and Breast Cancer Prediction: A Review. *Wirel Pers Commun*, 125 (2022) 2537.

22 Radak M, Lafta HY & Fallahi H, Machine learning and deep learning techniques for breast cancer diagnosis and classification: a comprehensive review of medical imaging studies. *J Cancer Res Clin Oncol*, 149 (2023) 10473.

23 Arya N, Saha S, Mathur A & Saha S, Improving the robustness and stability of a machine learning model for breast cancer prognosis through the use of multi-modal classifiers. *Sci Rep*, 13 (2023) 4079.

24 Chen H, Wang N, Du X, Mei K, Zhou Y & Cai G, Classification prediction of breast cancer based on machine learning. *Comput Intell Neurosci*, 1 (2023) 6530719.

25 Chugh G, Kumar S & Singh N, Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cogn Comput*, 13 (2021) 1451.

26 Conte L, Rizzo E, Civino E, Tarantino P, De Nunzio G & De Matteis E, Enhancing breast cancer risk prediction with machine learning: Integrating BMI, smoking habits, hormonal dynamics, and BRCA gene mutations—A game-changer compared to traditional statistical models? *Appl Sci*, 14 (2024) 8474.

27 Yadav RK, Singh P & Kashtriya P, Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey. *Procedia Comput Sci*, 218 (2023) 1434.

28 Guo Y, Zhang H, Yuan L, Chen W, Zhao H, Yu QQ & Shi W, Machine learning and new insights for breast cancer diagnosis. *Int J Med Res*, 52 (2024) 03000605241237867.

29 Zhong X, Lin Y, Zhang W & Bi Q. Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning. *Sci Rep*, 13 (2023) 18301.

30 Yedjou CG, Tchounwou SS, Grigsby J, Johnson K & Tchounwou PB, Improving Invasive Breast Cancer Care Using Machine Learning Technology. *J Biomed Res Environ Sci*, 3 (2022) 980.