Breast Cancer Detection Using Six Different Algorithms

Deshpande Arnav Sunil, Saloni Parekh, Anish Pattnaik and Ayushmaan Agarwal
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India
deshpandearnav.sunil2019@vitstudent.ac.in, saloni.parekh2019@vitstudent.ac.in,
anish.pattnaik2019@vitstudent.ac.in, ayushmaan.agarwal2019@vitstudent.ac.in

Dr. C. G. Mohan
School of Mechanical Engineering
Vellore Institute of Technology
Vellore, India
mohan.cg@vit.ac.in

Abstract

Cancer is among the most prevalent causes of mortality worldwide. Breast cancer is one of most often diagnosed cancers, with over 200 different varieties to choose from. Given the high prevalence of morbidity and death, early and accurate diagnosis are critical. For this purpose, symptoms must be carefully assessed and classified and this can be done by Machine Learning (ML) and Data Analytics techniques. The main goal of this research is to examine several ML and Deep Learning (DL) approaches for breast cancer diagnosis and accuracy prediction. The primary dataset being used for research purposes is the Wisconsin Breast Cancer Diagnosis dataset. The following are the findings of the algorithms used: 93.08 percent accuracy for Linear Regression, 93.61 percent accuracy for the K Nearest Neighbor, 96.50 percent accuracy for Support Vector Machine, and 95.10 percent accuracy for Multilayer Perceptron approaches.

Keywords

K Nearest Neighbor (KNN), Support Vector Machine (SVM), Machine Learning (ML), Random Forest (RF), and Multilayer Perceptron (MP).

1. Introduction

Sha et al. (2020) Breast cancer can be diagnosed using detection of the tumor first using ultrasound, mammogram, biopsy, MRI etc. But the cost of detection using these can be very expensive. Applying ML and DL algorithms to detect threats earlier can be much easier and less expensive. Even though the symptoms are moderate in the beginning, the odds of survival rise dramatically if somehow the diagnosis is obtained as a consequence of early identification. FNAC, ultrasound assisted surgical biopsy, and mammography are among the several diagnostic procedures used to identify breast cancer. But detection of breast cancer using mammography is very poor in case of dense breasts and about 10%-20% cases go undetected. Shen et al. (2019) ML can be used for breast cancer detection by analyzing the tumor size. Mechanical learning methods are the best ways to get good results among the problems of differentiation and prediction. Although a large number of AI algorithms for detection of breast cancer have been developed, the accuracy of the models is varied. So, the selection of the proper model for the given problem is very important. Mridha et al. (2021) have implemented various ML and DL techniques which could be utilized for detecting and classifying breast cancer. The algorithms are: KNN, NB, RF, MP, SVM, LR. D. S. Seah et al. (2022) Biopsy is the ultimate confirmation of breast cancer detection but these ML and DL techniques can act as the foundation of early detection using Magnetic Resonance Images (MRI), X-Rays and other factors. With the advancement in the field of automated detection, the accuracy and the prediction scores have been getting better and been very helpful.

2. Literature Review

Al-Azzam and Shatnawi (2021) Various ML and DL techniques have also been used for the prediction of breast cancer. The methods used are (SVM), KNNs, RFs, Artificial Neural Networks (ANNs), NB, Decision Tree (DT), LR deals with the comparison of four different types of ML algorithms that are used in breast cancer detection. Amethiya et al. (2021) Convolutional Neural Networks or CNNs, Recurrent Neural Networks or RNNs, Fuzzy logic, and Genetic algorithm are the four algorithms. The data comes from the WBCD at the University of California, Irvine. The accuracy, precision, and F scores are computed, and the optimal algorithm is identified. For one of the articles. M. B. Ammar et al. (2022) CNN had a 96.49 percent accuracy, RNN had a 63.15 percent accuracy, Genetic algorithm had an 80.39 percent accuracy, and Fuzzy Logic had an 88.81 percent accuracy. The scope of the paper is to use similar types of algorithms with a broad dataset and try to increase the accuracy. Bai et al. (2021) They got the maximum accuracy of 97.13%. Lotter et al. (2021) We would be using the normal machine for the accuracy and F1 calculations. H. Asri et al. (2016) we have seen that the authors have received an Accuracy of 97.60% for ANN. The other accuracies obtained are LDA (95.99%), SVM (96.70%), weighted KNN (96.70%) and cosine KNN (97.13). The author has studied 5 different classifiers in the paper. In A. Cruz-Roa et al. (2017) we can see that 6 different classifiers run on the Wisconsin Dataset, and the best performing model has been incorporated into the backend of a website. It uses the GLOBOCAN 2018 figures of Cancer. Further on exploring A. Das et al. (2021) The methods and techniques used to predict breast cancer using thermal pictures in this study are done on the basis of a deep CNN model. The study concludes with the major finding that based on the output spectrum and training data of 680 thermograms, 95.8% accuracy of prediction is achieved for breast cancer. In Dhahri et al. (2019) we can see that the paper has five techniques comparison done. The techniques are: SVM, KNN, LR, RFs, and ANNs. Another article based on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, Heena et al. (2019) proposes a comparison of six ML algorithms: NB, RF, Artificial Neural Networks, KNN, SVM, and DT. In another assessment, the study explores the benefits and limits of existing deep-learning-based architectures, investigates the datasets utilized, and discusses picture pre-processing approaches. In Islam et al. (2020) a full description of many imaging modalities is also presented, as well as performance metrics and discoveries, barriers, and research potential for future researchers.

3. Methods

We use a variety of ML and DL techniques to diagnose breast cancer. The data must first be preprocessed before the model can be classified. The WBCD dataset was utilized. There are 569 samples in the dataset, with 33 features in the column. texture, smoothness, radius, area, compactness, curvature, concave points, symmetry, and fractal dimensions are some of the properties (Figure 1).

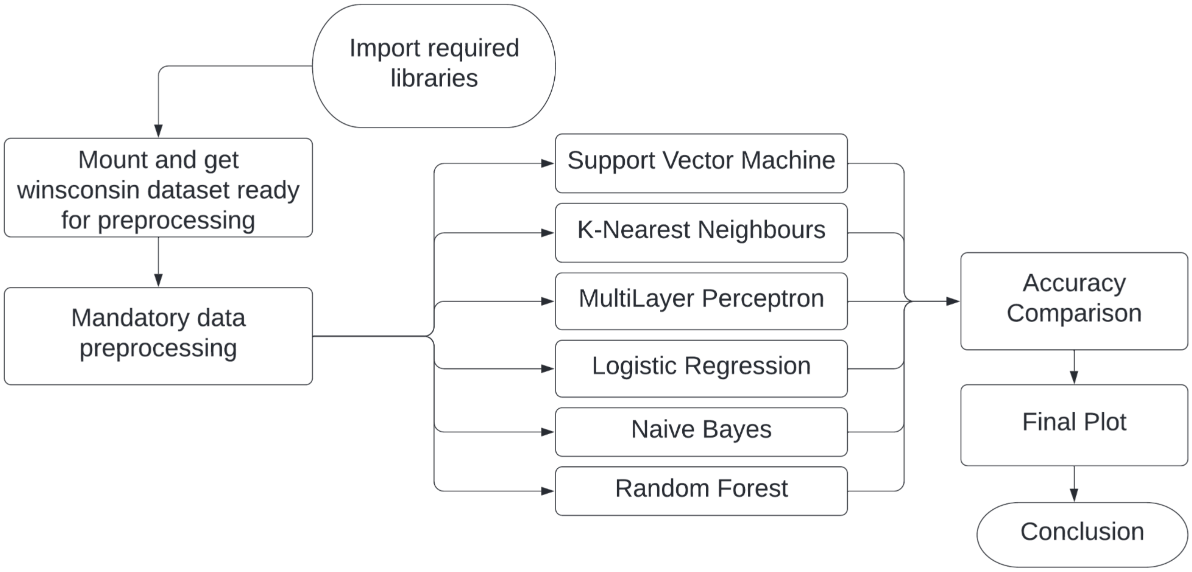

Figure 1. Flowchart of approach

4. Data Collection

The data collected is analyzed and then unnecessary data is removed as they can be a hindrance in training and testing. The type of breast cancer mentioned in the dataset are basically of two types: malignant and benign. Malignant is when the cells migrate and start to spread to other organs which can be very dangerous, while benign means the initial pre-cancer phase.

5. Results and Discussion

5.1 Numerical Results

A highly used and widely accepted metric to test a differentiation model is that via a confusion matrix which makes use of existing and true figures present in a test dataset. When making use of this, the terms ‘true negative’ and ‘true positive’ directly reflect that the obtained results by the model are correct. On the contrary, a result indicating a ‘false negative’ or ‘false positive’ signify the opposite. This same fact can be summarized as follows for this particular case:

1. *True Positive or TP*: When the output of the model is yes and the patient does have cancer.
2. *True Negative or TN:* In the case where the output of the model is no, and the patient doesn’t have cancer.
3. *False Positive or FP:* Wherein the output of the model is yes yet the patient doesn’t actually have cancer.
4. *False Negative or FN:* When the prediction is no but the patient actually does have cancer.
5. *Accuracy:* This performance measure is the fractional ratio of the correctly identified results to the overall number of results

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

.

6. *Precision:* It is the ratio of correctly identified labels and the overall positive results

$$Precision = \frac{TP}{TP + FP}$$

7. *Recall*: Recall is the ratio of the results that are positive and have been identified correctly to the total number of values in the class. It is also known as sensitivity.

$$Recall = \frac{TP}{TP + FN}$$

8. *F1 Score:* The F1 school combines the accuracy and memory of the editor into a single metric by taking its own harmonic mean. It is used to compare the performance of two classifiers.

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

9. *Support:* It is wont to show the number of events of a particular class in real responses and is done by calculating the lines of the confusion matrix (Table 1).

Table 1. Accuracy of different models

Algorithm		Precision	Recall	F1-Score	Accuracy
K-Nearest Neighbour		0.97	0.95	0.96	97.340
	Benign %	93	97	95	

	Malignant%	93	83	88	
Logistic Regression		0.97	0.97	0.97	97.872
	Benign %	93	97	95	
	Malignant%	94	87	90	
Support Vector Machine		0.96	0.94	0.95	96.503
	Benign %	97	98	97	
	Malignant%	96	94	95	
MultiLayer Perceptron		0.92	0.94	0.93	95.104
	Benign %	97	96	96	
	Malignant%	92	94	93	
Random Forest		0.96	0.98	0.97	97.902
	Benign %	98	99	0.98	
	Malignant%	98	96	97	
Naive Bayes		0.98	0.88	0.93	95.104
	Benign %	94	99	0.96	
	Malignant%	98	88	0.93	

5.2 Graphical Results

Data Analysis

In the Figure 2, the correlation is signified via the colors. The intensity of the color is directly proportional to the amount of correlation.

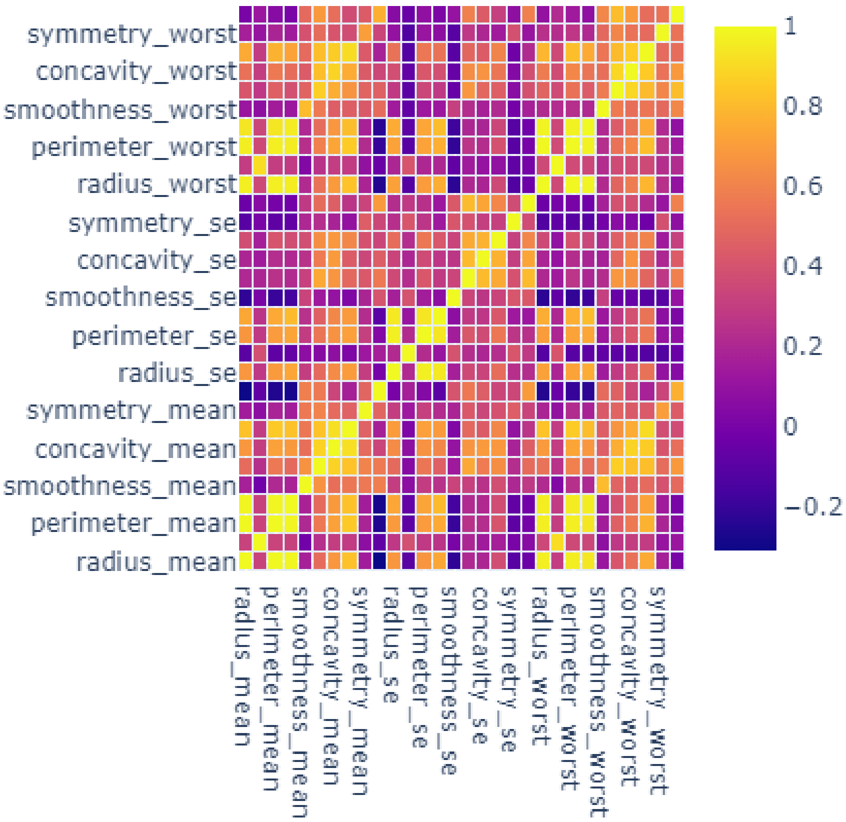

Figure 2. Heat map for the correlation matrix

Distribution

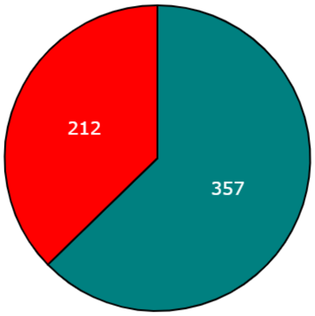

Figure 3. Data Distribution (Red-Malignant, Teal-Benign)

As we can see from Figure 3. the data is divided into malignant and benign with more samples which are benign.

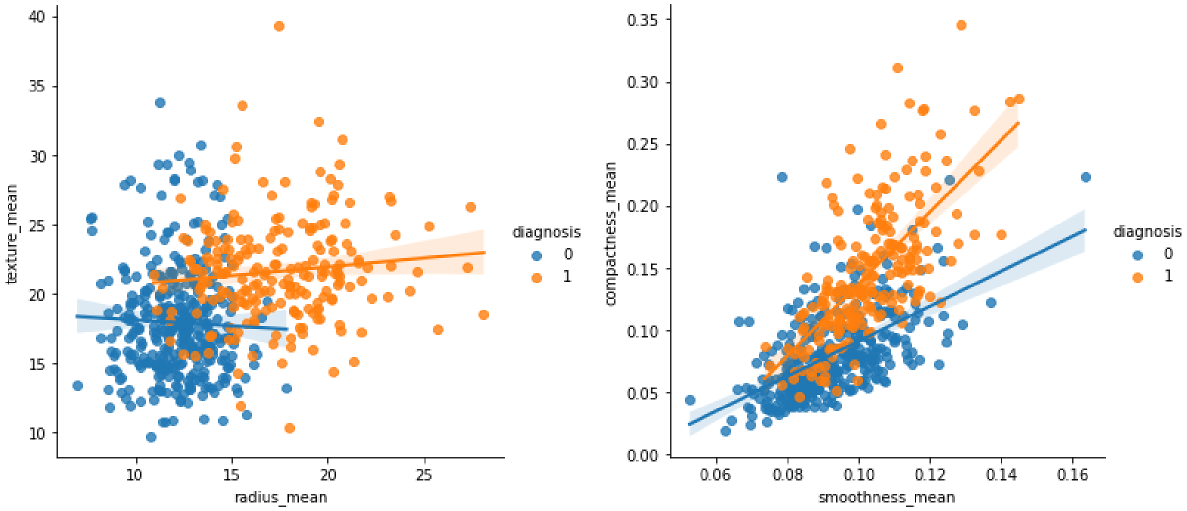

Figure 4. Relational Overview

Figure 4 shows us that the relational data analysis can be done using different metrics and we choose those data which have higher correlation and show better results when used after feature selection. Diverse metrics such as the Compactness mean, and texture mean are considered for a relational overview.

5.3 Validation

The performance of each algorithm is evaluated. The visualization of the data takes place after the data has been pre-processed and prepared via various methods. Herein, as per their efficiency and effectiveness, an evaluation is conducted. We conducted experiments using different algorithms and the best accuracy is tabulated. As we can see with an accuracy of 97.902% RF outperforms the other algorithm in terms of accuracy and hence proves to be the best possible algorithm for cancer prediction. Followed by RF we have LR and KNN which are ML algorithms.

6. Conclusion

To restate, this paper is primarily about comparing the different ML and DL methods that can be used to detect breast cancer and its predictive accuracy. Our model followed a method wherein the first process is that of extracting features and using those for testing and training a transfer learning model. As indicated by the observations in this study, it was seen that the techniques could be used on the WBCD dataset for boosting the accuracy of classifying lesions in breast cancer. After training the model, it was found that all algorithms achieved an accuracy of 90% or higher, and the accuracy of the Random Forest model was the highest at 97.92%.

References

Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. Shih, J. Tomaszewski, F. A. González, and A. Madabhushi, “Accurate and reproducible invasive breast cancer detection in whole-slide images: A DL approach for quantifying tumor extent,” *Scientific Reports*, vol. 7, no. 1, 2017.

Das, M. N. Mohanty, P. K. Mallick, P. Tiwari, K. Muhammad, and H. Zhu, “Breast cancer detection using an ensemble DL method,” *Biomedical Signal Processing and Control*, vol. 70, p. 103009, 2021.

Seah, N. Tayob, J. P. Leone, J. Hu, J. Yin, M. Hughes, S. M. Scott, R. I. Lederman, E. Frank, J. J. Sohl, Z. K. Stadler, T. K. Erick, J. Peppercorn, E. P. Winer, S. G. Silverman, S. E. Come, and N. U. Lin, “Perceptions of patients and medical oncologists toward biospecimen donation in the setting of abnormal breast imaging findings,” *Breast Cancer Research and Treatment*, vol. 192, no. 1, pp. 201–210, 2022.

Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using ML Algorithms for Breast Cancer Risk Prediction and Diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

Dhahri, E. A. Maghayreh, A. Mahmood, W. Elkilani, and M. F. Nagi, “Automated Breast Cancer Diagnosis Based on ML Algorithms,” *Journal of Healthcare Engineering*, vol. 2019, pp. 1–11, 2019.

Heena, S. Durrani, M. Riaz, I. Alfayyad, R. Tabasim, G. Parvez, and A. Abu-Shaheen, “Knowledge, attitudes, and practices related to breast cancer screening among female health care professionals: a cross sectional study," *BMC Women's Health*, vol. 19, no. 1, 2019.

Bai, R. Posner, T. Wang, C. Yang, and S. Nabavi, "Applying DL in digital breast tomosynthesis for automatic breast cancer detection: A review," *Medical Image Analysis*, vol. 71, p. 102049, 2021.

Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. Mcbride, and W. Sieh, "DL to Improve Breast Cancer Detection on Screening Mammography," *Scientific Reports*, vol. 9, no. 1, 2019.

Ammar, F. L. Ayachi, R. Ksantini, and H. Mahjoubi, "Data warehouse for ML: application to breast cancer diagnosis," *Procedia Computer Science*, vol. 196, pp. 692–698, 2022.

Mridha, M. A. Hamid, M. M. Monowar, A. J. Keya, A. Q. Ohi, M. R. Islam, and J.-M. Kim, "A Comprehensive Survey on Deep-Learning-Based Breast Cancer Diagnosis," *Cancers*, vol. 13, no. 23, p. 6116, 2021.

Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using ML Techniques," *SN Computer Science*, vol. 1, no. 5, 2020.

l-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised ML Models on Diagnosing Breast Cancer," *Annals of Medicine and Surgery*, vol. 62, pp. 53–64, 2021.

Lotter, A. R. Diab, B. Haslam, J. G. Kim, G. Grisot, E. Wu, K. Wu, J. O. Onieva, Y. Boyer, J. L. Boxerman, M. Wang, M. Bandler, G. R. Vijayaraghavan, and A. G. Sorensen, "Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient DL approach," *Nature Medicine*, vol. 27, no. 2, pp. 244–249, 2021.

Amethiya, P. Pipariya, S. Patel, and M. Shah, "Comparative analysis of breast cancer detection using ML and biosensors," Intelligent Medicine, 2021.

Sha, L. Hu, and B. D. Rouyendegh, "DL and optimization algorithms for automatic breast cancer detection," *International Journal of Imaging Systems and Technology*, vol. 30, no. 2, pp. 495–506, 2020.

Biographies

Deshpande Arnav Sunil is a Student at School of Computer Science and Engineering, Vellore Institute of Technology, Vellore. Interested in applying Computer Science solutions to Biological Problems.

Saloni Parekh is a Student at School of Computer Science and Engineering, Vellore Institute of Technology, Vellore. Intrigued by the vast applications Bioinformatics has in Computer Science. Willing to Explore more.

Anish Pattnaik is a third-year B. Tech student pursuing Computer Science and Engineering at Vellore Institute of Technology, Vellore. I have an inherent and deep interest in high-level programming languages such as C++, Java, Python, JavaScript, etc. I have a cognitive and analytical mindset. I possess knowledge and experience in the fields of Artificial Intelligence and Machine Learning and web development.

Ayushmaan Agarwal is a third-year computer science student seeking and learning various fields of tech, possessing skills in front-end and backend development. An avid reader and chess player with deep interest in learning new frameworks and fields like blockchain, machine learning and artificial intelligence.

C.G. Mohan is an Associate Professor at the department of Thermal and Energy Engineering. His area of interest is in applying computational solutions to Biological Problems using modern day technologies such as Deep Learning and Machine Learning. His previous research lies in characterization of renewable energy.