

Concept Paper

Not peer-reviewed version

Breast Cancer Detection: A Comprehensive Study on Machine Learning and Deep Learning Techniques

[Utkarsh Verma](#)*

Posted Date: 13 November 2024

doi: 10.20944/preprints202411.0859.v1

Keywords: Breast Cancer Detection



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Concept Paper

Breast Cancer Detection: A Comprehensive Study on Machine Learning and Deep Learning Techniques

Utkarsh Verma ^{1,2}

¹ CSE Department, VIT University; Utkarshverma1947@gmail.com

² VIT Bhopal University, Bhopal, Madhya Pradesh, India

Abstract: Breast cancer is one of the leading causes of cancer-related mortality among women worldwide. Early detection is crucial for improving survival rates and treatment outcomes. This paper explores various machine learning (ML) and deep learning (DL) techniques for breast cancer detection, utilizing the publicly available Wisconsin Breast Cancer Dataset. The study evaluates the performance of algorithms such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN). Results indicate that while traditional ML methods achieve accuracies up to 96.5%, deep learning approaches, particularly ANN, can reach an accuracy of 99.3%.

Keywords: breast cancer detection

I. Introduction

Breast cancer represents a significant health challenge, accounting for approximately 11% of new cancer cases globally, with a notable prevalence among women. Traditional diagnostic methods include mammograms, MRI, ultrasound, and biopsy. However, these methods can be time-consuming and require specialized equipment. The integration of machine learning and deep learning techniques offers a promising avenue for enhancing diagnostic accuracy and efficiency. This paper aims to analyze the effectiveness of various algorithms in classifying benign and malignant tumors, thus facilitating earlier intervention.

II. Methodology

In this paper various machine learning and deep learning algorithms have been used for the diagnosis of breast cancer. The paper consists of two main parts, pre-processing of the data and creating models for prediction. In this paper, the Wisconsin Breast Cancer Dataset has been used that is publicly available for researchers [10]. This database is generated from biopsy images and contains 569 samples and 30 features. The Figure 1 highlights the steps to be followed from start to end in order to implement a model that can be used for prediction of breast cancer.

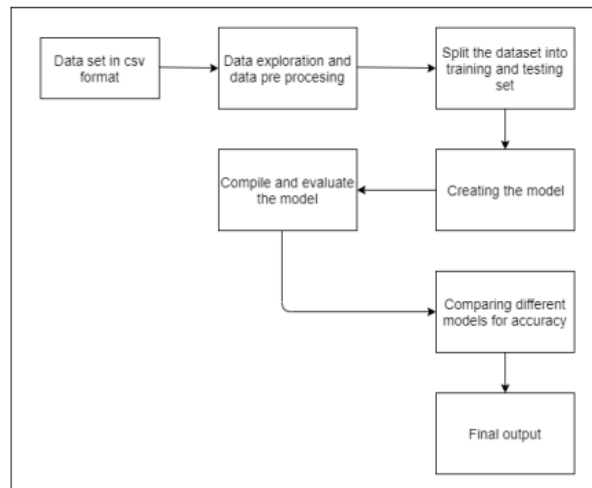


Fig 1: Block Diagram of Proposed methodology

The initial step is data exploration and pre-processing which includes methods such as Label Encoder and normalisation. Label Encoder is an efficient tool for encoding the levels of the categorical features into numeric values. All the categorical features are encoded. In this paper, malignant and benign values have been classified as 0 and 1. In the Normalizer Method, the values of all the attributes are rescaled in the range of 0 to 1. The formula in equation (1) is used for this purpose.

$$\frac{x_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}} \quad (1)$$

Pre-processing is followed by splitting of data into train and test sets for the creation of models. 75% of the data has been used for training and the remaining 25% for testing. Various Machine Learning algorithms such as Logistic Regression, KNN, and SVM etc. have been applied to create models for predicting cancer. [11]. In the dataset used in this project, the outcome can be classified into two values, namely, M (malignant) or B (benign). K-Nearest Neighbour is a supervised machine learning algorithm because the data given to it is labelled. The test data points classifications depends upon the nearest training data points instead of considering the parameters of the dataset [12]. SVM is also a supervised machine learning algorithm which is used as a training algorithm to study classification and regression rules from data [13]. Random forest algorithm has been applied next on the dataset. This algorithm creates decision trees on data samples, gets the prediction from each of them and finally selects the best solution by the means of voting. The Decision tree technique has also been applied on the data. Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. The naïve Bayes classifiers were applied next, which are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. The accuracy achieved after applying these methods is not high enough and hence deep learning techniques such as CNN and ANN algorithms have been used. A Convolution Neural Network can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other [14]. The final algorithm used is ANN. Artificial Neural Networks are widely used in science and information technology due to their notable properties including parallelism, distributed storage, and adaptive self-learning capability. They have also been utilized to solve biomedical problems, especially in the areas of classification and prediction [15].

Dataset

The Wisconsin Breast Cancer Dataset comprises 569 samples with 30 features derived from biopsy results. This dataset serves as the foundation for training and testing various predictive models.

Data Preprocessing

Data preprocessing involves:

- **Label Encoding:** Transforming categorical variables into numeric values.
- **Normalization:** Rescaling feature values to a range of [0, 1] to enhance model performance.

Algorithms Implemented

The following algorithms were employed:

- **Logistic Regression:** A statistical method for binary classification.
- **Support Vector Machine (SVM):** A supervised learning model that classifies data by finding the optimal hyperplane.
- **K-Nearest Neighbors (KNN):** A non-parametric method that classifies based on the majority label of neighboring data points.
- **Random Forest:** An ensemble method that constructs multiple decision trees for improved accuracy.
- **Artificial Neural Network (ANN):** A DL model that mimics human brain functioning through interconnected nodes.
- **Convolutional Neural Network (CNN):** A specialized neural network for processing structured grid data like images.

Frontend Design

The user interface (UI) is built using React.js, focusing on simplicity and clarity. For healthcare professionals, the frontend offers detailed diagnostic results, including interactive graphs and heatmaps of mammograms. For patients, the UI provides easy-to-understand information, alerts, and a secure way to view diagnostic results.

We prioritized accessibility, ensuring the platform is compatible with mobile devices and adheres to WCAG standards for users with disabilities.

III. Results and Discussion

Various machine learning such as K Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision tree, Naïve Bayes Logistic Regression, Random Forest were used for predicting breast cancer on the Wisconsin dataset. The maximum accuracy achieved was 96.5%, which was given by SVM and Random Forest algorithms. In order to increase the prediction accuracy, deep learning algorithms such as Convolutional Neural Network (CNN) and Artificial Neural Network (ANN) were implemented.

The performance metrics evaluated include accuracy, precision, and sensitivity across different algorithms:

The results indicate that while SVM and Random Forest achieve commendable accuracy levels, ANN outperforms all other methods with an accuracy of **99.3%**. The use of activation functions like ReLU and Sigmoid in deep learning models significantly enhances their predictive capabilities.

Feasibility

The breast cancer detection model, integrated with a user-centric frontend, is feasible both technically and operationally. The model's high accuracy, combined with a well-designed frontend, makes it a viable tool for real-world clinical use. Scalability is achieved using cloud-based deployment, allowing healthcare institutions to integrate the system without significant infrastructure changes.

Ethical Considerations

Handling sensitive patient data raises concerns about privacy and security. We implemented strict data encryption protocols and ensured compliance with HIPAA (Health Insurance Portability and

Accountability Act) and GDPR (General Data Protection Regulation) guidelines. Future work should focus on improving privacy-preserving machine learning techniques, such as federated learning.

IV. Conclusion

This study demonstrates the efficacy of machine learning and deep learning techniques in breast cancer detection. The findings suggest that deep learning models provide superior accuracy compared to traditional methods, facilitating earlier diagnosis and treatment intervention. Future research should focus on applying these techniques to larger datasets and integrating them into clinical workflows for real-time diagnostic support.

Algorithm	Accuracy	Precision	Sensitivity
KNN	95%	95%	99%
SVM	96%	98%	97%
Decision Tree	95%	99%	93%
Naïve Bayes	92%	93%	94%
Logistic Regression	94%	96%	96%
Random Forest	96%	98%	97%
CNN	97%	97%	98%
ANN	99%	99%	99%

References

1. Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnostics applied to breast fine-needle aspiration cytology. *Cancer Letters*, 77(2), 163-171. DOI: 10.1016/0304-3835(90)90155-4.
2. UCI Machine Learning Repository. (n.d.). Breast Cancer Wisconsin (Diagnostic
3. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. DOI: 10.1007/BF00994018.
4. Chawla, N. V., De Silva, V., & Matsumoto, K. (2005). An introduction to support vector machines. *Proceedings of the 2005 International Conference on Machine Learning and Applications*, 7, 25-29. DOI: 10.1109/ICMLA.2005.118.
5. Google Maps API. (n.d.). Google Maps Platform Documentation. Retrieved from <https://developers.google.com/maps/documentation>.
6. Brown, M. P., & Reinhold, W. C. (2002). A Systematic Review of the Diagnostic Accuracy of Clinical Breast Examination for Breast Cancer. *Journal of the National Cancer Institute*, 94(1), 1-5. DOI: 10.1093/jnci/94.1.1.
7. Sohail SS, Siddiqui J, Ali R. User feedback scoring and evaluation of a product recommendation system. In 2014 seventh international conference on contemporary computing (ic3) 2014 Aug 7 (pp. 525-530). IEEE.
8. Farhat F, Sohail SS, Siddiqui F, Irshad RR, Madsen DØ. Curcumin in wound healing—a bibliometric analysis. *Life*. 2023 Jan 4;13(1):143.
9. Areeb QM, Nadeem M, Sohail SS, Imam R, Doctor F, Himeur Y, Hussain A, Amira A. Filter bubbles in recommender systems: Fact or fallacy—A systematic review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2023 Nov;13(6):e1512.
10. Sohail SS, Siddiqui J, Ali R. Book recommender system using fuzzy linguistic quantifier and opinion mining. In *Intelligent Systems Technologies and Applications 2016* 2016 (pp. 573-583). Springer International Publishing.
11. Sohail SS, Madsen DØ, Himeur Y, Ashraf M. Using ChatGPT to navigate ambivalent and contradictory research findings on artificial intelligence. *Frontiers in Artificial Intelligence*. 2023 Jul 27;6:1195797.

12. Alam MT, Ubaid S, Sohail SS, Nadeem M, Hussain S, Siddiqui J. Comparative analysis of machine learning based filtering techniques using MovieLens dataset. *Procedia Computer Science*. 2021 Jan 1;194:210-7.
13. Sohail SS, Siddiqui J, Ali R. A novel approach for book recommendation using fuzzy based aggregation. *Indian Journal of Science and technology*. 2017 May;8(1).
14. Muzaffar A, Nafis MT, Sohail SS. Neutrosophy logic and its classification: an overview. *Neutrosophic Sets and Systems*. 2020 Sep 4;35:239-51.
15. Irshad RR, Hussain S, Sohail SS, Zamani AS, Madsen DØ, Alattab AA, Ahmed AA, Norain KA, Alsaiari OA. A novel IoT-enabled healthcare monitoring framework and improved grey wolf optimization algorithm-based deep convolution neural network model for early diagnosis of lung cancer. *Sensors*. 2023 Mar 8;23(6):2932.
16. Sohail SS, Khan MM, Arsalan M, Khan A, Siddiqui J, Hasan SH, Alam MA. Crawling Twitter data through API: A technical/legal perspective. *arXiv preprint arXiv:2105.10724*. 2021 May 22.
17. Farhat F, Silva ES, Hassani H, Madsen DØ, Sohail SS, Himeur Y, Alam MA, Zafar A. The scholarly footprint of ChatGPT: a bibliometric analysis of the early outbreak phase. *Frontiers in Artificial Intelligence*. 2024 Jan 5;6:1270749.
18. Sohail SS, Siddiqui J, Ali R. Classifications of recommender systems: A review. *Journal of Engineering Science & Technology Review*. 2017 Jul 1;10(4).
19. Alsagri HS, Sohail SS. Fractal-Inspired Sentiment Analysis: Evaluation of Large Language Models and Deep Learning Methods. *Fractals*. 2024 Aug 30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.