

Concept Paper

Not peer-reviewed version

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

[Jagrati Mathpal](#) *

Posted Date: 14 November 2024

doi: 10.20944/preprints202411.1059.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

Jagrati Mathpal

VIT Bhopal University, Bhopal, Madhya Pradesh, India; jagratimathpal@gmail.com

ABSTRACT: Breast cancer is among the most common cancers affecting women globally. Early detection is crucial in reducing mortality rates and improving treatment outcomes. This project utilizes machine learning to develop a breast cancer detection model based on patient medical data. The Random Forest Classifier was selected due to its high accuracy and capacity to handle imbalanced datasets. The project also integrates a frontend interface that allows users to input relevant data and find nearby cancer treatment centers through a location-based service. With an accuracy of over 95%, the model offers a promising tool to assist healthcare professionals and patients. Future improvements aim to enhance the dataset and user accessibility, making it a more versatile and scalable solution.

1. Introduction

Breast cancer accounts for a significant portion of cancer cases worldwide. While various diagnostic methods like mammography and biopsy are used, they often require specialized equipment and expertise. The role of machine learning in healthcare has grown due to its ability to analyze vast amounts of data and predict disease outcomes efficiently.

This project aims to develop a machine learning model capable of predicting breast cancer based on clinical data. Additionally, it incorporates a user-friendly frontend interface that assists patients in finding nearby treatment centers. By integrating technology with healthcare, the project seeks to improve early detection and streamline access to medical services.

2. Background

Breast cancer diagnosis traditionally relies on a combination of medical imaging techniques, such as mammography, ultrasound, and biopsies. However, these techniques require time and specialized equipment. With the growth of machine learning, predictive models have emerged as valuable tools to assist in the early diagnosis of diseases, including breast cancer.

2.1. Machine Learning

Machine learning models analyze structured data—such as tumor features and other clinical measurements—to detect abnormalities. Commonly used models include Decision Trees, Support Vector Machines (SVM), Logistic Regression, and ensemble methods like Random Forest. Each of these models has strengths, but ensemble methods like Random Forest are particularly effective in handling imbalanced data, which is common in medical datasets.

2.2. Role of User-Friendly Interfaces

A key challenge in medical diagnostics is making technology accessible to non-experts. A well-designed interface can enable patients to interact with predictive models by inputting relevant data and obtaining reliable insights without requiring extensive medical knowledge.

3. Methodology

3.1. Data Collection

The dataset used for this project was sourced from the UCI Machine Learning Repository, specifically the Breast Cancer Wisconsin (Diagnostic) dataset. It consists of 569 instances with 30 numerical features derived from digitized images of fine needle aspirate (FNA) of breast masses. Key

features include attributes such as clump thickness, uniformity of cell size, cell shape, marginal adhesion, and other clinical factors. The dataset is labelled as either malignant (cancerous) or benign (non-cancerous), serving as the target variable for the classification model.

3.2. Data Preprocessing

Data pre-processing steps were crucial for ensuring the quality of input to the machine learning model:

- **Handling Missing Values:** The dataset had no missing values, so no imputation was required.
- **Normalization:** All features were normalized to ensure they were on the same scale, which is essential for machine learning models like Random Forest.
- **Train-Test Split:** The dataset was split into 80% for training and 20% for testing to validate the model's performance on unseen data.

3.3. Model Selection

Support Vector Machine (SVM) was selected as the model for this project. SVM is well-suited for binary classification tasks and works by finding the optimal hyperplane that separates data points of different classes (malignant and benign) with the maximum margin. SVM was chosen over models like Decision Trees and Logistic Regression because of its high accuracy in handling high-dimensional datasets and its ability to effectively model non-linear relationships using kernels.

3.4. Evaluation Metrics

To assess the model's performance, the following metrics were calculated:

- **Accuracy:** The overall percentage of correct predictions made by the model.
- **Precision:** The proportion of true positive predictions out of all positive predictions made.
- **Recall (Sensitivity):** The proportion of true positive cases detected out of all actual positive cases.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure for the model's performance, particularly useful for imbalanced datasets. Confusion matrices were also generated to visualize the true positive, true negative, false positive, and false negative predictions.

3.5. Website Development

The website includes a React.js frontend where users input medical data to receive breast cancer risk predictions, powered by an SVM model hosted on a Flask backend. The backend processes the data and returns predictions to the frontend. Additionally, the tool incorporates a treatment center locator using the Google Maps API, helping users find nearby cancer treatment facilities based on their location.

4. Results and Discussion

The Support Vector Machine (SVM) model achieved notable performance in breast cancer classification. It recorded an accuracy of 96%, a precision of 94%, a recall of 95%, and an F1-score of 94.5%. The confusion matrix reflected a low number of false positives and false negatives, indicating that the model is highly effective in detecting both malignant and benign cases. These metrics demonstrate that the SVM model is well-suited for binary classification, especially after preprocessing steps such as feature scaling and the application of SMOTE to handle class imbalance.

Table 1. SVM Model Performance Metrics.

Metric	Value
Accuracy	96%

Precision	94%
Recall	95%
F1-Score	94.5%

In terms of practical application, the integration of the SVM model into a user-friendly interface allows individuals with no technical or medical expertise to use the system effectively. The treatment center locator feature, powered by the Google Maps API, adds additional value by providing users with nearby cancer treatment options. This tool not only predicts breast cancer risk but also bridges the gap between diagnosis and immediate medical assistance, making it highly practical for users in real-world scenarios.

The model's performance could be further enhanced by expanding the dataset, adding more diverse data points to improve generalization. Additionally, while the system provides accurate predictions, future iterations may integrate more complex medical data, such as genetic information, to increase prediction accuracy and cover more nuanced cases.

5. Conclusions

The Breast Cancer Detection project successfully developed a tool that uses a Support Vector Machine (SVM) model to predict breast cancer risk, achieving an accuracy of 96%. The model's high precision, recall, and F1-score demonstrate its effectiveness in classifying cases as malignant or benign. Coupled with a user-friendly interface and a treatment center locator, the system provides vital information and resources for users.

While the current model performs well, future enhancements such as expanding the dataset and incorporating additional clinical features could further improve accuracy. Overall, this project represents a significant step in utilizing machine learning for healthcare, offering a valuable resource for timely diagnosis and treatment access.

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. In2014 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. InIntelligent 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, Alsaiani 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.