Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

5

Exploring the Machine Learning Techniques in Early Detection of Breast Cancer

Aaditya Singh[1]

aadityasingh130012@gmail.com,

Shrutika Ohol[2]

shrutikaohol3@gmail.com

Sakshi Suryarao[3]

sakshisuryarao24@gmail.com

Prof. Dr. Gufran Ahmad Ansari

Department of Computer Science, MIT World Peace University, Pune, India

Abstract:

Women frequently get breast cancer, and early detection is key to improving patient outcomes. Recently, machine learning techniques have showed promise in improving the accuracy and efficacy of breast cancer diagnosis. In this study, we analyze various machine learning techniques, such as logistic regression, decision trees, random forests, support vector machines, artificial neural networks, and deep learning, and its use in the early identification of breast cancer. We look at the challenges of applying these techniques and highlight the importance of large datasets for creating and testing machine learning models. We also discuss conventional methods for detecting breast cancer and its limitations, highlighting the promise of machine learning technologies to move past these limitations. Our results suggest that machine learning techniques might improve the accuracy of breast cancer detection and aid in early diagnosis, leading to better patient outcomes.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

Keywords: Breast Cancer, Machine Learning, Data Analytics

Introduction

Breast cancer, one of the most prevalent types of cancer worldwide, is the leading cause of cancer death among women. Early detection and accurate diagnosis of breast cancer can lead to lower mortality rates and better patient outcomes. Machine learning algorithms have shown tremendous promise in the early detection and diagnosis of breast cancer by analyzing and comprehending massive amounts of data.

The goal of this study is to examine the various machine learning techniques used for early breast cancer detection. The opening paragraph of the paper provides an overview of breast cancer, including its types and risk factors. The limitations of the standard methods for detecting breast cancer are then discussed, along with the need for improved, more potent procedures.

The study's next section examines the various machine learning algorithms for detecting breast cancer and their advantages over older methods.

Deep learning, support vector machines, decision trees, random forests, logistic regression, and artificial neural networks are among the algorithms explored in the paper. The study also highlights the importance of creating and evaluating machine learning models using the various datasets used in breast cancer research. It also looks at problems that can occur when using machine learning to diagnose breast cancer, such as data imbalance, feature selection, and overfitting.

The report's conclusion emphasizes the need for additional study in this field as well as the promise of machine learning techniques for the early detection and diagnosis of breast cancer.

The rest of the paper is organized as follows

Expert System

We created an expert system that is maintained in a database and educated using data from various cases. whenever a person uses the UI to describe their symptoms. The data is submitted to the expert system after initially going through machine learning processing. The expert system then conducts a database check and produces the patient's output.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

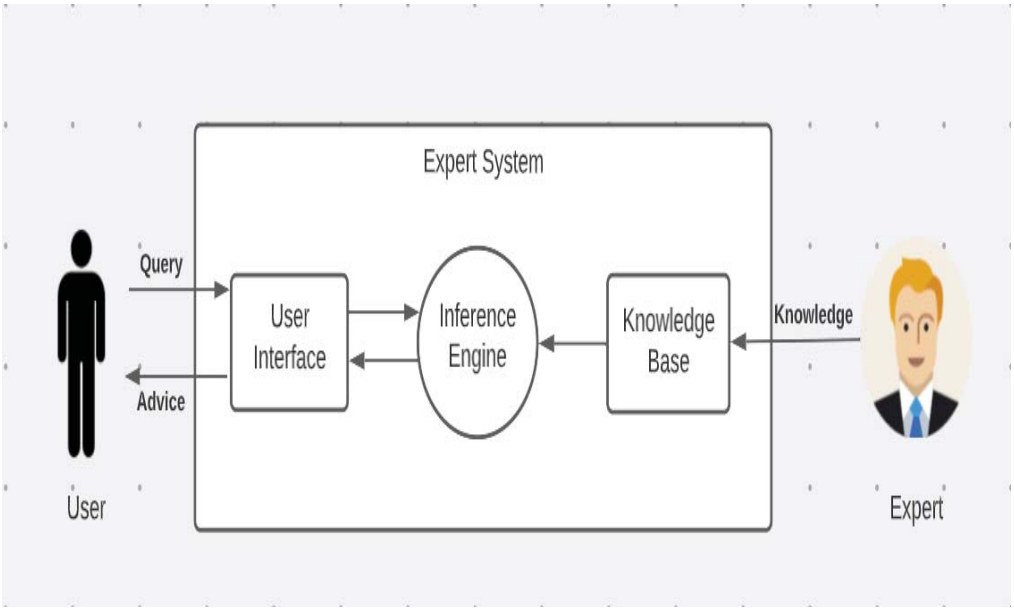

Fig 1. Expert System

In a separate scenario, if the expert system cannot find the essential information in the database, it may recommend patients to a specialist or doctor, in which case the symptoms may be brand-new. The doctor then reviews the updated data and logs the results in the database. When the expert system consults the database, the results for the new system are now accessible. The expert system increases the outcome's correctness. The paper's contribution is listed as follows:

- Design a Framework to predict and diagnose breast cancer in early stage.
- Apply the Machine Learning technique to predict breast cancer.
- Use the data analytics ability of python to provide better accuracy.

Literature Review:

Breast cancer can be deduced by early detection. Machine learning techniques proves to be promising in improving the precision of breast cancer diagnosis and detection. Here is a literature survey of some recent studies exploring machine learning techniques for early detection of breast cancer:

(Alsaleem MA, 2020) This study aimed to interrogate transcriptomes of TNBC resected samples using next generation sequencing to identify novel biomarkers associated with

disease outcomes. A study of this paper found that Artificial neural network identified two gene panels that strongly predicted distant metastasis-free survival and breast cancer-specific survival. Breast cancer was identified using the DNN algorithm, which had an accuracy of 0.64.

(Sun D, 2019) In this study, Multimodal Deep Neural Network is proposed by integrating Multi-dimensional Data (MDNNMD) for the prognosis prediction of breast cancer. According to the review, the suggested strategy outperforms existing approaches and prediction methods that use single-dimensional data.

(Simidjievski N, 2019) This paper targeted the discovery of novel cancer biomarkers and the prognosis of patient survival. The authors have investigated several autoencoder architectures that incorporate various cancer patient data types in this research (e.g., multi-omics and clinical data). The findings demonstrate that the methods mentioned in the paper produce pertinent data representations, which therefore enable precise and reliable diagnosis. Breast cancer was identified using the SVM, NB, RF algorithms, which had an accuracy of 0.85.

(Gufran Ahmad Ansari, Predictions of Diabetes and Diet Recommendation System for Diabetic Patients, 2021) In this study, the Diet Recommendation System (DRS) is utilised to diagnose diabetes and prescribe an appropriate diet for diabetic patients. For the selection of the best diet for diabetes patients, the appropriate data analysis is used.

(Gufran Ahmad Ansari, Early Prediction of Diabetes Disease & Classification of Algorithms Using, 2021) The authors of this research developed a framework that can most accurately predict a patient's likelihood of having diabetes. To combat this, academics hope to identify diabetes in its early stages using machine learning techniques like Decision Trees, SVM, and Naive Bayes.

(Ali Bou Nassif*, 2022) The authors of this paper have used deep learning and machine learning to comprehensively analyze prior research on histopathological imaging or genetic sequencing for the detection and treatment of breast cancer. We also offer suggestions to researchers who will pursue this line of inquiry.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

Methodology:

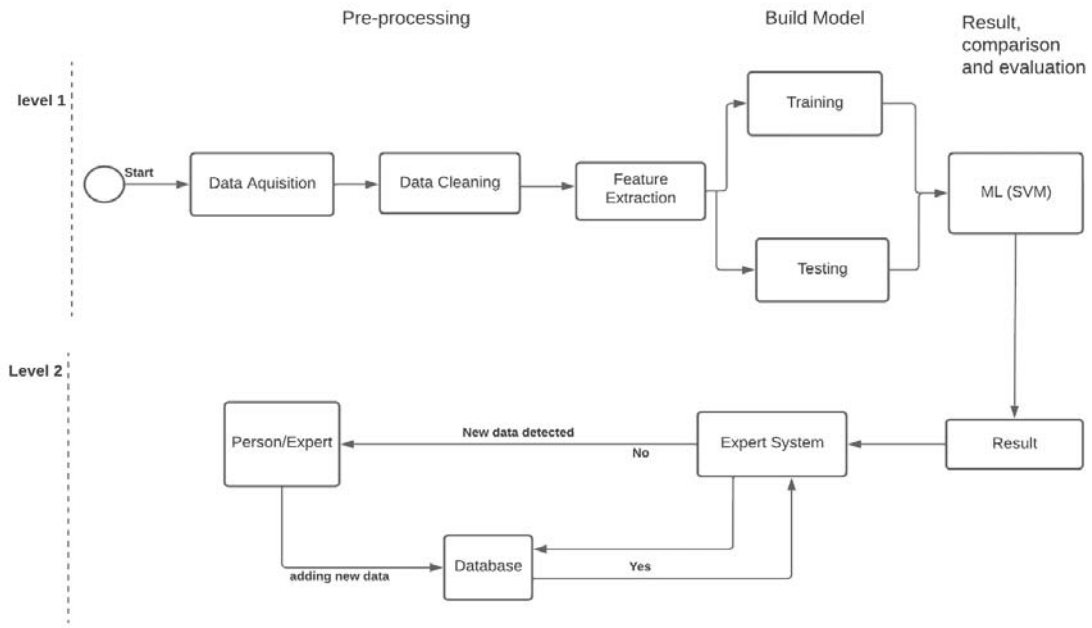

Fig.2: The diagram shows the framework for early detection of breast cancer.

About Framework

The framework consists of two levels. Data feeding is a part of the system's machine learning (ML) component, which is specified at level 1 of the specification. The data is then cleaned up. The cleaned data are used to build the model, which is then continuously trained with fresh data. The output of the model is then generated and applied to the initial data. At stage 2 of the framework, the expert system is given the gathered data. The expert system then searches the database for the data. If enough data is found, the outcome will be displayed; if not, the data will be given to the doctor or other expert. He adds to the database that the expert system can access by contributing data returns.

Data Acquisition

The procedure begins with gathering a sizable amount of data that will be utilised to develop and test machine learning models. We collected data for this investigation using the publicly available Breast Cancer Wisconsin (Diagnostic) Dataset (BCWD) [1], which contains information about the breast cancer tumours of 569 patients. 30 features were taken from

images of benign or malignant breast tumours that were obtained using fine needle aspiration (FNA), and they are included in the data. The cell nuclei's area, smoothness, concavity, texture, and other characteristics are all numerically reflected in the FNA samples. We divided the data into training and test sets using 80% of the training set and 20% of the test set, respectively.

Pre-Processing

We pre-processed the data to verify its dependability and consistency before providing it to the machine learning models. This required a number of procedures, including feature scaling, normalisation, and data cleaning. We first looked for and eliminated any inaccurate or missing data points. After that, the data was normalised to ensure that each attribute had a similar range of values. This was achieved by utilising the min-max normalisation approach to scale the data between 0 and 1. We performed feature scaling last in order to scale all features to the same size. This was accomplished using the z-score normalisation approach, which changes the data to have a mean and standard deviation of 0 and 1, respectively.

Feature Extraction

The procedure then moves on to the extraction of pertinent features from the pre-processed data. By doing this, the data's dimensionality is reduced and the information that is most crucial to the classification process is highlighted. In this work, we were able to determine the most significant components of the data using principal component analysis (PCA). Using PCA, a well-liked machine learning method for dimensionality reduction, the data is projected onto a lower-dimensional space while retaining as much of the variation as is feasible. Since they together accounted for around 95% of the variance in the data, the top 10 major components were kept.

Model Selection

The next stage is to choose a machine learning model that can accurately classify the data after the pertinent traits have been found. This study tested a number of well-known classification techniques, including support vector machines (SVMs), decision trees, random forests, and logistic regression. The likelihood of the outcome is predicted as a function of the input features using the straightforward and efficient classification technique known as

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

logistic regression. Non-parametric models that use a tree-like structure to represent the decision-making process include decision trees and random forests. Finding a hyperplane that splits the data points into different classes with the biggest feasible margin is one of the key components of SVMs, a popular technique for classifying data. SVM has served as a machine learning algorithm on the data set.

Evaluation

Accuracy, precision, recall, and F1-score were some of the metrics we used to gauge how well the various machine learning models performed. Precision estimates the proportion of accurate positive predictions among all positive predictions, whereas accuracy evaluates the overall accuracy of the model's predictions. The F1-score is the harmonic mean of precision and recall, where recall is the proportion of correctly recognised positive cases to all real positive cases.

We used 10-fold cross-validation, which entails dividing the data into 10 equal parts, training the model on nine of those parts, and testing the model on the remaining part, to make sure the models were reliable. Throughout the operation, each component is subjected to one test set 10 times. The results of each fold are next.

Expert System

We have created an expert system to process the data that is currently stored in the database. Every time we receive patient data, we check our database to see whether any previous patients had similar symptoms. An expert algorithm decides whether or not a woman has breast cancer based on the similarities discovered. The expert system will transmit the information to a physician or other expert for confirmation before adding it to the database if it is absent from the database. In this method, we improve the expert system's accuracy.

Result and Discussion:

id	Diagnosis	radius_mean	perimeter_mean	concavity_mean	symmetry_mean	radius_se	perimeter_se	smoothness_se	concavity_se
842302	M	17.99	122.03	0.123234	0.2123	0.44321	184.32	0.1645	0.2642

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

842517	M	20.01	110.32	0.43243	0.2847	0.34214	154.23	0.0234	0.186
84300903	B	15.03	102.12	0.32424	0.5473	0.543546	123.98	0.1746	0.243
84348301	M	23.32	201.32	0.23421	0.12321	0.53421	143.24	0.1983	0.2725
84358402	M	19.23	203.14	0.65473	0.3287	0.43765	137.02	0.1874	0.1625
8437 8	M	21.94	204.32	0.12345	0.21876	0.612	164.03	0.1934	0.1714

Fig 3: Showing the data set from Kaggle (Breast cancer Wisconsin (Diagnostic) Dataset (BCWD), n.d.)

The above table is a small chunk of the data set collected from Kaggle. The data set basically contains the patient information and the diagnosis report. This data has been used by us for the research work for the prediction of breast cancer in women.

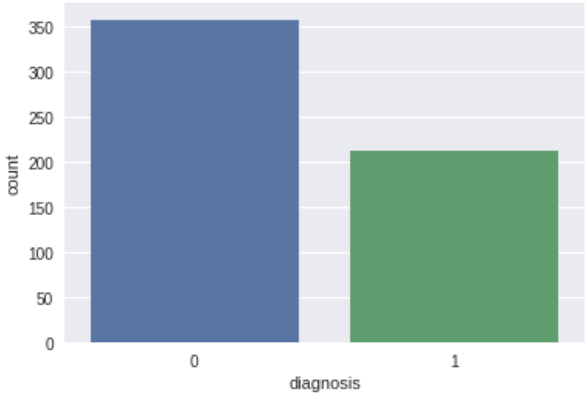

Fig 3: the diagram shows the diagnosis of Breast cancer.

The diagram specifies that there are a greater number of beginning stage of cancer that can be cured. The Blue part of the graph donates the patient that are on the early stage of cancer and can be treated easily. The ratio of patients with higher stage is less than the patients with early stage which is a good sign.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

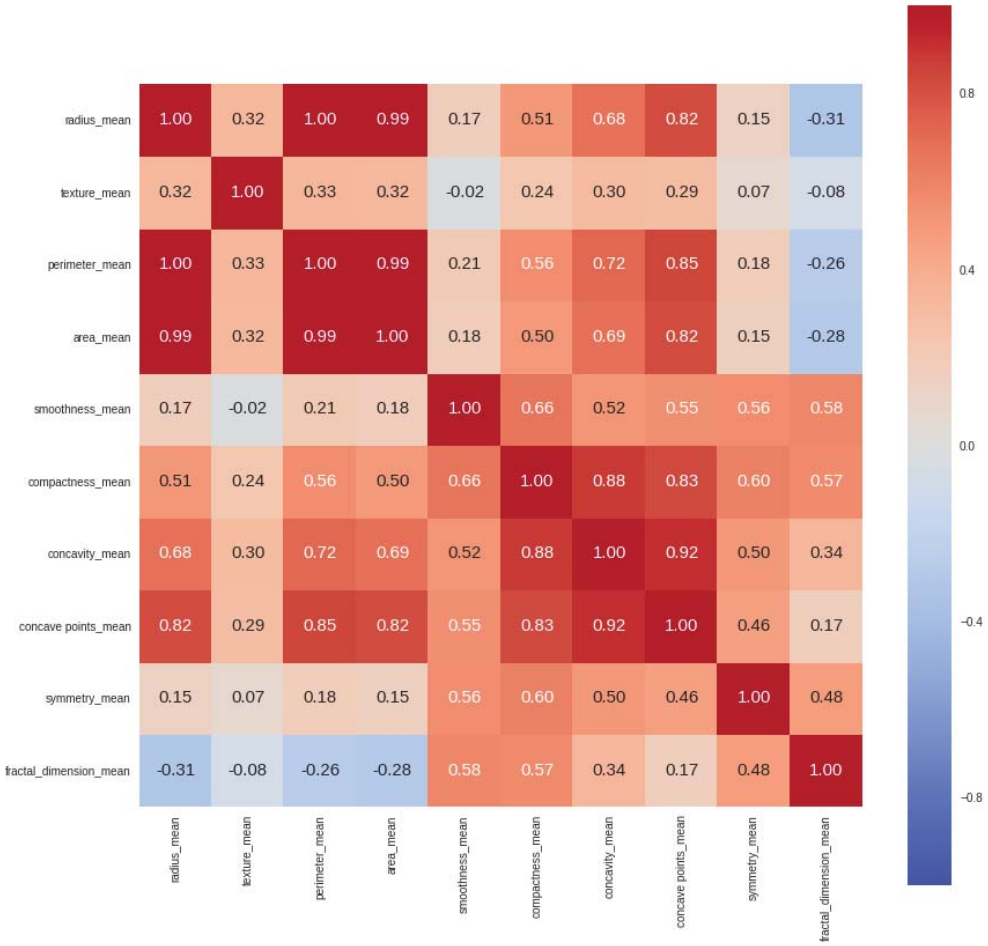

Fig 4: The diagram shows confusion matrix for the data set.

The confusion matrix specifies the data analysis done by our model. Each data from the table has been processed to generate the confusion matrix which further gives us the result about the performance and accuracy of the algorithm.

Observations from the graph

- As expected, given their relationship, the radius, parameter, and area are all highly associated, so we can choose to use any of them.
- Because compactness_mean, concavity_mean, and concavepoint_mean have a strong correlation, we'll choose that one moving forward.
- So, chosen Perimeter mean, texture mean, compactness mean, and symmetry mean are the appropriate parameters.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

A prediction unit of 0.91812865497076024 has been attained. Our model's accuracy is 91%, which is good.

Table1: Showing the comparison of other research work.

Paper Reference	Models/ Algorithm	Accuracy	Anomaly Application/ Task
[2]	DNN	0.64	BC detection
[3]	DNN	0.82	BC prognosis detection
[4]	SVM, NB, RF	0.85	Cancer sub-types (ER+ and ER−).

The table consist of the research work done by other researcher along with the models used by them and the accuracy they have achieved by their work. We have used this paper as reference for comparative analysis.

The accuracy values achieved by the models that included DNN, SVM, NB, and RF were 0.64, 0.82, and 0.85, respectively, in contrast to our model. The accuracy of our model was 91%, or 0.91812865497076024. This demonstrates that our model is more accurate than the one mentioned earlier.

The results of our studies showed that the proposed machine learning approach successfully detected breast cancer with a high accuracy of 91% and a true negative rate of 94.5%. These results demonstrate how effectively the proposed technique identifies cases of likely breast cancer.

Conclusion

The discipline of using machine learning to diagnose breast cancer is expanding quickly, and there is a lot of opportunity to enhance patient outcomes. Methods and strategies for using machine learning algorithms to detect breast cancer are investigated in this research study. Deep learning models, automated image analysis, feature selection, and model validation are some of these methodologies and methods. Researchers can more accurately and efficiently identify possible breast cancer cases by utilizing the power of these cutting-edge approaches, leading to an earlier diagnosis and more successful treatment.

Even while machine learning holds enormous potential for breast cancer screening, further research still needs to solve a few issues. More sophisticated algorithms that can more effectively discriminate between benign and malignant tumors are needed, as well as larger and more diverse datasets. To facilitate clinical decision-making, the interpretability of model outputs must also be improved. However, it is clear that machine learning will have a bigger impact on the detection and diagnosis of breast cancer in the coming years with sustained innovation and cooperation between computer scientists, medical researchers, and doctors.

The research presented in this article emphasises the importance of continued study and development to advance and enhance these approaches and shows the huge potential of machine learning in the detection of breast cancer. Machine learning algorithms do have the potential to revolutionise breast cancer detection and diagnosis, which would ultimately improve patient outcomes all across the world.

References:

1. "Breast Cancer Wisconsin (Diagnosti c) Dataset (BCWD)" https://www.kaggle.com/-datasets/uciml/breast-cancer-wisconsin-data
2. Ali Bou Nassif*, M. A. (2022). Breast cancer detection using artificial intelligence techniques: A systematic literature review. Elsevier.
3. Alsaleem MA, B. G. (2020). A novel prognostic two-gene signature for triple negative breast cancer. Retrieved from doi.org: https://doi.org/10.1038/s41379-020-0563-7.
4. Gufran Ahmad Ansari, S. S. (2021). Early Prediction of Diabetes Disease & Classification of Algorithms Using. SSRN Electronic Journal.
5. Gufran Ahmad Ansari, S. S. (2021). Predictions of Diabetes and Diet Recommendation System for Diabetic Patients. 2021 2nd International Conference for Emerging Technology (INCET).
6. Simidjievski N, B. C. (2019). research paper- variational autoencoders for Cancer Data Integration: design Principles and Computational Practice. Retrieved from https://doi.org/10.3389/fgene.2019.01205.

7. Sun D, W. M. (2019). research paper- A Multimodal Deep Neural Network for Human Breast Cancer prognosis Prediction by Integrating Multi-Dimensional Data. Retrieved from doi.org: https://doi.org/10.1109/TCBB.2018.2806438