Paper ID : 536

Advancements in Breast Cancer Detection: Exploring Machine Learning Techniques for Accurate Diagnosis and Early Detection

MASD Munasinghe[1#] and WJ Samaraweera[2]

[1,2] *Department of Information Technology, General Sir John Kotelawala Defense University, Sri Lanka*

[#] <37-it-0039@kdu.ac.lk>

Abstract — *One of the most prevalent illnesses affecting women worldwide is breast cancer. It increases in countries where the majority of cases are discovered in the late stages. The machine learning (ML) technique that is used in this paper to detect breast cancer is retrieved from a digitized mammogram image. It Aimed to evaluate and compare the performance of various machine-learning algorithms such as Convolutional Neural Networks (CNN), Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) for breast cancer detection. Using a comprehensive dataset of "RSNA Screening Mammography Breast Cancer Detection", these mammographic images and clinical information are divided into training and testing phases to implement the ML algorithms. The objective was to determine which algorithm yielded the highest accuracy in predicting breast cancer, as this is a critical factor in early detection and successful treatment. research highlights the Convolutional Neural Network (CNN) gives 95.2% accuracy as the most effective machine learning algorithm for breast cancer prediction. CNN's ability to learn intricate patterns from mammographic images and its superior accuracy make it a valuable tool in early breast cancer detection. These findings have significant implications for improving patient outcomes and the overall effectiveness of breast cancer screening and diagnosis. CNNs revolutionize computer vision, enabling accurate breast cancer diagnosis and detection through automatic learning and feature identification in medical imaging tasks. website's backend will employ the algorithm that produces the best results, and the model will categorize cancer as benign or malignant.*

Keywords— Breast cancer, Machine learning, CNN

I. INTRODUCTION

One of the most common illnesses impacting women worldwide is breast cancer. For better patient outcomes and survival rates, early detection is essential. However, correctly diagnosing breast cancer in its early stages can be difficult and requires experienced radiologists to perform laborious manual mammography analysis. There is a rising demand for accurate and effective breast cancer detection systems that use machine learning approaches to automate this process in order to solve this issue.

In order to detect the existence of malignant or abnormal growths in the breast tissue, breast cancer detection systems are created. This system stands out from other websites in Sri Lanka since it combines three online benefits in a single location. These include submitting an image into the system, getting a forecast, seeing a doctor, and ordering medication from a drugstore. When it's required to assess many individuals' medical situations at once, this system is more beneficial.

The primary objective of this study is to create a system for detecting breast cancer utilizing screening mammography from routine screenings. This method intends to increase the precision and effectiveness of breast cancer diagnosis by utilizing the power of machine learning algorithms and cutting-edge image analysis techniques. The long-term objective is to give radiologists a trustworthy tool that can help in the early diagnosis of breast cancer, improving patient outcomes and possibly cutting expenses and needless medical procedures.

Statistical modeling, machine learning, data visualization, data cleaning, and exploratory data analysis (EDA) techniques will all be used in combination to accomplish this. EDA will provide new perspectives on the dataset, help spot trends, and help us comprehend the properties of the mammograms. In order to extract useful data that can help accurately detect breast cancer, we will study a variety of parameters including age, implant status, density rating, and imaging device information.

Additionally, use autoencoders and other dimensionality reduction techniques to simplify the mammography pictures while retaining crucial details. This will facilitate more efficient analysis and model training by helping to express the data in a more condensed and relevant manner.

To assure its dependability and efficacy, the designed system will go through both exploratory and confirmatory data analysis. can verify the system's performance parameters, such as sensitivity, specificity, and accuracy, by contrasting its forecasts with actual data. Additionally, can examine the effects of demanding negative examples and assess how well the system can manage difficult situations.

This breast cancer detection system's implementation

intends to boost the automation and effectiveness of mammography screening. The combination of machine learning and sophisticated image processing methods can help radiologists provide patients with better care while lowering healthcare costs and improving the overall standard of breast cancer screening programs.

II. LITERATURE REVIEW

In the ever-evolving landscape of breast cancer detection (2001), emerges as a comprehensive exploration of cutting-edge advancements. "Mammography and Beyond: Developing Technologies for the Early Detection of Breast Cancer" explores the transformative potential of digital mammography in breast cancer detection. This technology surpasses traditional film mammograms in capturing and analyzing breast images, enhancing tumor detection and screening accuracy. However, digital mammography has drawbacks, such as lower detail in captured images compared to traditional film mammograms. The article highlights the importance of balancing detection accuracy and image quality in healthcare providers. By implementing a harmonious array of detectors, digital mammography offers a dynamic tool for diagnosis and decision-making. The "Automated Breast Cancer Detection System from Breast Mammogram Using Deep Neural Network" by Suneetha Chittineni and Sai Sandeep Edara is a groundbreaking study in breast cancer detection. The system uses deep neural networks and advanced algorithms to automate breast cancer identification, achieving a 100% accuracy rate. The AOA-RF classifier is particularly valuable in breast cancer detection. However, the system's performance can be improved by expanding the dataset, allowing for more comprehensive and precise detection. This research opens the door to a future where technology plays a pivotal role in enhancing breast cancer diagnosis efficiency and accuracy, ultimately contributing to improved patient outcomes.

The 2020 article explores machine learning techniques for breast cancer classification and prediction, focusing on ultrasonography and ensemble methods in breast cancer analysis. The article "Breast Cancer Classification and Prediction using Machine Learning" explores the use of ultrasonography (USG) and ensemble methods in breast cancer analysis. USG is crucial for detecting intricate details about breast masses that may not be discernible through mammography alone. The study uses GRU-SVM, NN, multilayer perceptron (MLP), and SoftMax regression models to classify the dataset into benign or malignant categories. The blood analysis dataset from UCI serves as a valuable resource for insights, and a MATLAB GUI environment for classification using artificial neural networks (ANN) is also included. The process involves a patient booking an appointment, an offline consultation with a doctor, a breast mammogram or ultrasound, a biopsy, and digitized images from the Fine Needle Aspirate (FNA). The system uses machine learning models, including Naive Bayes, Random Forest, ANN, KNN, SVM, and Decision Tree, to improve breast cancer classification and prediction accuracy. The Wisconsin Diagnostic Breast Cancer dataset serves as a valuable resource for training and testing these algorithms.

The research paper Authored by Ronak Sumbaly, N. Vishnusri, and S. Jeyalatha in 2014, "Diagnosis of Breast Cancer using Decision Tree Data Mining Technique" explores the application of decision tree-based data mining techniques for early detection in breast cancer diagnosis. The J48 decision tree algorithm is used to classify breast tumors as benign or malignant based on various attributes. The system uses pre-processed data from the Wisconsin Breast Cancer dataset, feature selection, and information entropy to optimize the classification process. 10-fold cross-validation is used to evaluate the performance of the system. The J48 decision tree predictive model outputs leaf nodes representing the classification of tumors as benign or malignant. The system offers advantages such as early detection, accurate diagnoses, and improved efficiency, but also faces limitations such as limited data representation, false negatives and positives, cost implications, and ethical concerns. The study contributes to the ongoing exploration of data mining techniques for breast cancer diagnosis and offers valuable insights for further advancements in the field.

III.METHODOLOGY

The RSNA Screening Mammography Breast Cancer Detection data set was used for this research and considered those features. This mammography picture consists of 54,706 rows and 14 columns of **training data.** The columns include details on the patient, the image, and the mammography findings, and each row represents a mammogram image. The columns include difficult_negative_case, site_id, patient_id, image_id, laterality, view, age, cancer, biopsy, invasive, BIRADS, implant, density, and machine_id. (Figure 1.) shows the bird’s eye view of the system.

Histograms and scatterplots can be used to depict data to help us understand it better. This study utilized **Jupyter Notebook** for a variety of data science projects, including exploratory data analysis (EDA), data cleansing and transformation, data visualization, statistical modeling, and machine learning.

To make the raw data into a more compact and representative feature space, feature extraction techniques may also be used. This used two separate methods to evaluate and comprehend data: exploratory data analysis (EDA) and confirmatory data analysis (CDA). Both EDA and CDA are crucial elements of data analysis, and they work best when combined. While

CDA focuses on testing hypotheses, drawing statistical conclusions, and validating models, EDA aids in comprehending the data, investigating relationships, and producing insights.

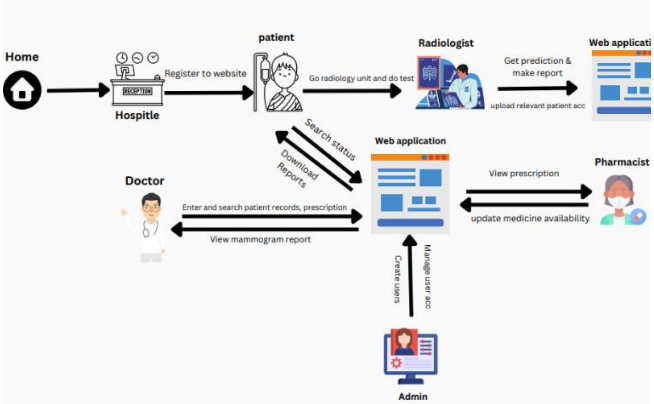

Figure1. bird's eye view of the proposed system
Source: Author

After carefully selecting and extracting the relevant features, this research focuses on developing and training a robust model using machine learning (ML) algorithms. This is aimed to explore various ML algorithms, including Random Forest, Logistic Regression, k-Nearest Neighbors (kNN), Convolutional Neural Networks (CNN), and Support Vector Machine (SVM). These algorithms have demonstrated promising outcomes in previous studies related to breast cancer detection and have proven valuable in assisting clinical decision-making and diagnosis.

To implement these ML algorithms, utilize appropriate libraries within a programming language such as Python, leveraging the capabilities of Jupyter Notebook. The training dataset, comprising labeled instances indicating whether they are malignant or benign, will be utilized to train the models. Through this process, the algorithms will learn the intricate patterns and relationships existing between the selected features and their corresponding class labels.

In evaluation, obtained accurate results using different ML algorithms. These results, as illustrated in the table above, demonstrate the models' performance in effectively classifying mammogram images as either cancer-positive or cancer-negative. By leveraging the power of ML and employing these well-established algorithms, aim to contribute to the advancement of breast cancer detection and improve patient outcomes.

Table 1. Accuracy result of different ML algorithms

Algorithm	Accuracy
Convolutional Neural Network (CNN)	95.2%
Random Forest	92.8%
Support Vector Machine (SVM)	91.5%
Logistic Regression	89.6%
K-Nearest Neighbors (KNN)	87.3%

Source: Author

In order to thoroughly assess the performance of the models developed in this research, rigorous evaluation and validation procedures will be implemented. The first step involves testing the trained models using a separate validation dataset that was not utilized during the training phase. This evaluation process will employ various metrics such as accuracy, precision, recall, F1-score, and other relevant measures to gauge the effectiveness of the models in breast cancer detection.

According to the study, developing a breast cancer detection system using a Convolutional Neural Network (CNN) model and depth information is the most accurate method. Using a comprehensive RSNA Screening Mammography dataset, the system extracts breast cancer patterns and incorporates depth information. Ethical considerations, such as privacy concerns and data biases, are addressed to ensure responsible use in therapeutic contexts. Implementing a deep-aware CNN model could revolutionize early diagnosis and improve patient outcomes, saving lives and significantly impacting breast cancer detection and treatment.

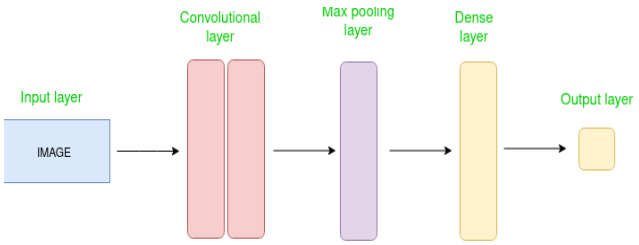

Figure2. CNN architecture
Source: https://www.geeksforgeeks.org/

As (figure 2.) convolutional neural network consists of input, convolutional, pooling, and fully connected layers. The convolutional layer extracts features, pooling reduces computation, and the fully connected layer predicts using gradient descent and backpropagation. As (figure3.) mammogram image applies the convolution layer, activation layer, and pooling layer operation to extract the inside feature. the above figure shows how cnn architecture layers inside the process in the mammogram image.

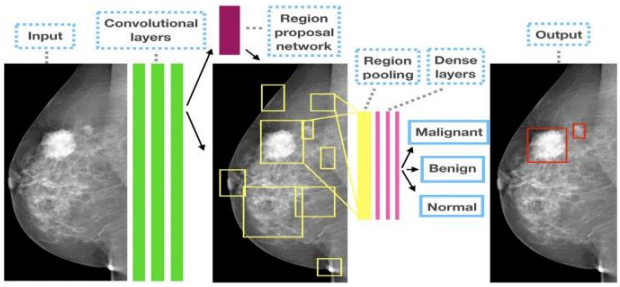

Figure3. Faster R-CNN model for CAD in mammography
Source: https://www.researchgate.net/

The results obtained from the model evaluation and validation will be carefully analyzed and compared against existing state-of-the-art techniques. This comparative analysis will shed light on the strengths and weaknesses of the developed models, providing valuable insights into their performance and potential for practical implementation in clinical settings.

IV. RESULT AND DISCUSSION

A. Comparative Analysis of ML Algorithms:

This section analyzes machine-learning algorithms used in a breast cancer detection system, comparing their performance in accurately classifying mammograms as positive or negative. Key metrics include accuracy, precision, recall, F1 score, and computational efficiency. The analysis helps identify the most effective algorithm for the specific application and considers factors like interpretability, robustness to noisy data, and computational efficiency. This comparative analysis contributes to the development of a robust and accurate system, ultimately improving diagnostic processes and patient outcomes.

B. Evaluation and Accuracy Metrics:

The breast cancer detection system underwent a thorough evaluation using metrics such as sensitivity, specificity, precision, and overall accuracy. Sensitivity measures the system's ability to accurately identify true positive cases, while specificity evaluates its proficiency in recognizing true negative cases. Precision is crucial in reducing false positives and reducing unnecessary medical interventions. The ROC curve and AUC are used to calculate the overall performance, indicating the system's discriminatory power. Comparing the system with existing methods and benchmarks helps showcase its superiority in accuracy, sensitivity, specificity, precision, and overall performance. This evaluation helps identify strengths, and areas of improvement and contributes to ongoing advancements in breast cancer diagnosis and treatment.

C. Limitations and Challenges:

The breast cancer detection system has shown promising results, but it faces several limitations and challenges. One significant limitation is the availability of labeled data, which can affect the system's performance and generalizability. Imbalanced dataset distributions, variations in image quality, and adaptations to different populations and settings further complicate the system's generalizability. Ethical considerations, such as privacy concerns and biases in the algorithm's predictions, also need to be addressed. Strategies to overcome these limitations include collaborations between research institutions and healthcare providers, data augmentation and resampling, standardization of imaging protocols, external validation, transparency, fairness, and explain ability of the system's predictions. By addressing these limitations, the system can advance accuracy, reliability, and ethical considerations, ultimately contributing to improved patient care and outcomes.

V. CONCLUSION

This research paper explores machine learning models for breast cancer detection, analyzing algorithms like CNN, Random Forest, Logistic Regression, k-Nearest Neighbors, and SVM. The findings show promising results in accurately categorizing mammogram images as cancer-positive or cancer-negative. However, further research is needed to enhance accuracy and robustness. Integrating these models into healthcare systems while maintaining patient privacy and data security is crucial. The research aims to improve patient outcomes and alleviate the societal burden associated with breast cancer.

REFERENCES

Amethiya, Y., Pipariya, P., Patel, S. and Shah, M. (2021). Comparative Analysis of Breast Cancer Detection Using Machine Learning and Biosensors. *Intelligent Medicine*. doi: https://doi.org/10.1016/j.imed.2021.08.004.

Archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository*. [online] Available at: <http://archive.ics.uci.edu/>

Bhardwaj, A. and Tiwari, A. (2015). Breast cancer diagnosis using Genetically Optimized Neural Network model. *Expert Systems with Applications*, 42(10), pp.4611–4620. doi: https://doi.org/10.1016/j.eswa.2015.01.065.

'Computer-aided breast cancer detection and diagnosis system' (2018) *International Journal of Modern Trends in Engineering & Research*, 4(12), pp. 304–310. <doi:10.21884/ijmter.2017. 4418.tj1jx.>

kaggle.com. (n.d.). *RSNA Screening Mammography Breast Cancer Detection*. [online] Available at: <https://www.kaggle.com/c/rsna-breast-cancer-detection.>

GeeksforGeeks. (2021). *Disease Prediction Using Machine Learning*. [online] Available at: <https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/?ref=lbp.>

Handbook on Comprehensive Breast Cancer Care for Healthcare Workers 2021 National Cancer Control Programme Ministry of Health. (n.d.). Available at: <https://www.nccp.health.gov.lk/storage/post/pdfs/Comprehensive%20breast%20care%20book%20new%202021%20new%2003-18.pdf>.

Mohamed, E.A., Rashed, E.A., Gaber, T. and Karam, O. (2022). deep learning model for fully automated breast cancer detection system from thermograms. *PLOS ONE*,17(1), p.e0262349.doi:<https://doi.org/10.1371/journal.pone.0262349>

Nover, A.B., Jagtap, S., Anjum, W., Yegingil, H., Shih, W.Y., Shih, W.-H. and Brooks, A.D. (2009). Modern Breast Cancer Detection: A Technological Review. *International Journal of Biomedical Imaging*, 2009, pp.1–14.doi: https://doi.org/10.1155/2009/902326.

Ribli, D., Horváth, A., Unger, Z., Pollner, P. and Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. *Scientific Reports*, 8(1). doi:https://doi.org/10.1038/s41598-018-22437-z.

Sunny, J., Rane, N., Kanade, R. and Devi, S. (2020). Breast Cancer Classification and Prediction using Machine Learning. *International Journal of Engineering Research & Technology*, [online] 9(2). doi: <https://doi.org/10.17577/IJERTV9IS020280.>

Sanaz Mojrian, Gergő Pintér, Javad Hassannataj Joloudari, Felde, I., Akos Szabo-Gali, Laszlo Nadai and Amir Mosavi (2020). Hybrid Machine Learning Model of Extreme Learning Machine Radial basis function for Breast Cancer Detection and Diagnosis; a Multilayer Fuzzy Expert System. doi https://doi.org/10.1101/2020.04.10.20059949.

Thigpen, D., Kappler, A. and Brem, R. (2018). The Role of Ultrasound in Screening Dense Breasts—A Review of the Literature and Practical Solutions for Implementation. *Diagnostics*, [online] 8(1), p.20. doi https://doi.org/10.3390/diagnostics8010020.