Breast Cancer Detection Using Image Processing and Machine Learning

Akshaya A[1] , Manjula Sanjay Koti[2,] Priyadarshini S[3]

[1,2,3]Dayananda Sagar Academy of Technology and Management, Karnataka, India

Email: akshayaa6778@gmail.com[1], hodmca@dsatm.edu.in[2], priyadarshini-mca@dsatm.edu.in[3]

Abstract

As the outlines picturize, one driving reason for death in women across the entire world is breast cancer. It is an often-occurring disease in women, affecting approximately 2.1 million women annually. Studies indicate it generally affects women more in developed regions, although rates are increasing globally. While prevention may not be a feasible option, improving the outcomes and survival rates of breast cancer is a viable goal. Breast cancer mortality can be considerably decreased by more efficient treatments, which are made possible by early discovery of the disease. Many researchers and scientists are working on methods to facilitate early detection of breast cancer. Using the K-Nearest Neighbors (KNN) algorithm is one such technique. KNN is a straightforward machine learning technique that works well for regression and classification. In order to categorize an input according to the majority class of its neighbors, it first finds the k-nearest data points to the input. Using features taken from medical imaging, KNN can be utilized to determine a tumor's malignancy or benignity in the context of breast cancer detection. This algorithm is a useful tool for creating precise and dependable diagnostic systems since it can adjust and get better with additional data.

Keywords

Benign and malignant, Cancer Detection, Image Processing, K-NN

Introduction

The MIAS (Mammographic Image Analytic Society) is an exploration group in the Unified Realm, which intrigues itself in concentrating on mammograms & has made a computerized data set unidentified to the examination performed. The dataset consists of 322 breast images of 161 subjects. It is this kind of dataset that is used in building up the model for the purpose of arriving at a legitimate expectation. After that, with several electronic picture processes, the images are pre-processed right away. Meanwhile, they meet what company needs. The first step in the image processing is resizing which to ensure that all images have the same purpose. After resizing, the image is dim scaled & a gaussian channel is used that eliminate any foundation noise that can amplify the image. The smooth image which is now filtered is then passed to change division process. Highlight extraction & grouping are more advancements more. By employing the K-NN classifier, the order is done and the idea of the malignancy is received. Similarly, the project to predict the severity of the disease using involuntary brain organisation employs the specific attributes of carcinoma.

Submission: 16 October 2024; **Acceptance:** 15 November 2024

Bosom malignant growth begins in the cells of the covering or outer lining (epithelial) of the channels or ducts (85 percent) or clusters referred to as lobules (15 percent) of the bosom gland. Initially, the malignant change is tied to the channel or lobule that is 'in situ', that is, for the most part non-invasive, & has little potential to metastasize. Breast cancer treatment can be very effective if the disease is detected early. Treatment of breast ailment mainly involves surgery, radiation, and medications (endocrine therapy, chemotherapy, in addition to a type of natural therapy) to treat the metastasize spread of a cancerous tumor from the breast cancer through the bloodstream. Such therapy, which can prevent the growth & spread of the disease, eventually serves as a life-saving intervention.

Literature Review

Early-stage breast cancer detection system using glcm feature extraction and k-nearest neighbor (k-NN) on mammography image (Htay & Maung, 2018) and describes to design an automated tool to detect early-stage breast cancer in mammograms, through median filtering for noise reduction and Otsu's thresholding for image segmentation, with feature extraction via first-order statistical and GLCM methods. The k-Nearest Neighbor (k-NN) classifier is then used to classify the abnormalities, and the performance of the algorithm is assessed on mini-MIAS database.

Texture analysis of mammogram for the detection of breast cancer using LBP and LGP: to present a neural-based model for the classification of early-stage breast cancer using mammography scans (Ponraj & Mercy, 2017). The method consists of a preprocessing step where the image undergoes median filtering to eliminate noise, image slicing for cropping and finally, applying Otsu's binarization for segmentation to isolate the region of breast interest. Feature extraction encompasses first order statistical as well as GLCM predominantly textural descriptors. Lastly, the k-Nearest Neighbor (k-NN) classifier is applied to differentiate between the extracted objects by classifying them as normal or abnormal with the help of mini-MIAS database for the mini-MIAS database.

An intelligent decision for breast cancer diagnosis, develop an intelligent decision support system (IDSS) for breast cancer diagnosis, consisting of pre-processing, segmentation, feature extraction, and classification stages (Al-Salman & Almutairi, 2019). Preprocessing eliminates noise, segmentation uses the K-means algorithm to identify regions of interest, and feature extraction employs discrete wavelet transform (DWT) and gray level co-occurrence matrix (GLCM) methods. Classification with an artificial neural network (ANN) achieved 96.563% accuracy using the MIAS dataset and two-fold cross-validation.

Darapuredd and team work on implementation of optimization algorithms on Wisconsin Breast cancer dataset using deep neural network (Darapuredd et al., 2019). In this work, a model is developed and different optimization algorithms were implemented to access the correctness of classifying data with respect to accuracy which is feasible for computer-aided diagnosis. Machine learning can assist and alert expert radiologist more effectively than current screening techniques. In this paper RMS propagation (Root Mean Square Propagation) and SGD (stochastic gradient descent) optimization algorithms were implemented on a deep neural network with sigmoid neurons and accuracy is compared.

Qayyum and Basit present automatic breast segmentation and cancer detection via SVM in mammograms (Qayyum, & Basit, 2016). This work aims at recognizing breast cancer in digital mammography using segmentation by Otsu method in the inclined area, elimination of pectoral muscle using canny edge detection and classification using SVM. Texture analysis was conducted on the chosen images using Gray Level Cooccurrence Matrices (GLCM) and their features were extracted. Another fact is that the methodology tested successfully on the Mini-MIAS database which speaks about the reliable application of this technique to recognize the breast cancer cases.

Addeh introduced early detection of breast cancer using optimized ANFIS and features selection (Addeh et al., 2017). In this method, ANFIS is used as intelligent classifier and association rules (AR) technique is used as feature selection algorithm. In ANFIS, the value of radius has significant effect on system accuracy. Therefore, in the proposed method we used cuckoo optimization algorithm (COA) to find the optimal value of radius. The proposed method is applied on Wisconsin Breast Cancer Database (WBCD) and the results show that the proposed method has high detection accuracy.

The second common cause of death in women is breast cancer, yet it is a reversible disease if the symptoms are detected on time (Sangeetha & Murthy, 2017). Compared to other studies that diagnose carcinomas at the tumor stage with a moderate degree of certainty, this paper seeks to develop a new approach using high level digital image processing methodologies for the detection of early-stage breast asymmetry and micro calcification cancer cells. The proposed method intends to achieve high accuracy and also has the added advantage of offering a complete approach in handling issues related to false positive and false negative results.

Methodology

The current approach to the detection of breast cancer through image processing and pattern recognition involves the use of different techniques for detection of anomalies in mammograms. These usually require two or more processing steps, such as preprocessing, segmentation, feature extraction and classification steps. Some preprocessing techniques include contrast enhancement and image normalization with an aim of minimizing image noises, thereby enhancing the quality of the mammograms.
At this step, they applied the segmentation methods like Otsu's bill-local thresholding and the region growing method to extract the ROI from the breast images. There are 2 methods involved as stated below:

a. ROI Extraction Method 1

The ROI was removed at 598x598 at its unique size. The whole ROI was resized to 598x598, with cushioning to give setting. the ROI had the size of one aspect more than 1.5 times the other aspect; it was extricated as two tiles focused on the focal point of an amount of ROI along its biggest aspect.

b. ROI Extraction Method 2

In the remote possibility that the ROI was more modest than a 598x598 tile, it was removed with 20% cushioning on one or the other side. In the unlikely event that the ROI was bigger than a 598x598 tile it was removed with 5% cushioning. Every ROI was then arbitrarily edited multiple times, utilizing irregular flipping & turning. In the first phase, the Image processing stage and KNN using image as mammography images applied as input to detect Benign or malignant. In the second stage was using attribute based KNN to detect the severity.

The proposed method, is a result of the high levels of accuracy required and the consequent need to ensure extensive coverage of all potential and probable cases, it is usually possible to use ML in the diagnosis of medical conditions, specifically cancerous tissue. Mixed research strategies and algorithms used in cancer detection models provide promising accuracy. This is important since in clinical data, the outcome certainty is paramount in the decision-making process. While some models work better when the inputs are accurate, other models can perform well when there is a variation in the inputs. Hence, certain criteria such as stability, precision, and complexity analysis must be considered when making the selection for modelling.

Another common machine learning algorithm that can be applied both for classification and regression problems is an algorithm called K-Nearest Neighbors (KNN). This technique of classification or regression is a type of non-parametric method which tries to predict the output by using the ‘k’ nearest neighbors to the given data point. This is because, breast cancer is a disease that affects many women, and timely diagnosis is important when it comes to treatment since it is associated with better results. Mammography is one of the common clinical imaging processes that are used for breast cancer screening. However, differentiation of patterns and components within breast tissues is not easy because breast tissues are complex and display variations in patterns.

Results and Discussions

While CNN is triumphant when it comes to classification of pictures, the weaknesses of this approach that was adopted in this paper are overshadowed by the benefits of KNN algorithm that is particularly appropriate for breast cancer detection for the following reasons: Unlike most other classification algorithms, KNN exercises instance-based learning and does not have any constraints on the distribution of data thus is non-parametric, it gives a high level of accuracy in situations where simplicity and interpretation are preferable. Table 1 below shows the algorithms, accuracy and descriptions.

Table 1: The algorithms, accuracy and descriptions.

Algorithm	Accuracy	Description
K-Nearest Neighbors (KNN)	95%	Non-parametric method used for classification and regression.
Convolutional Neural Networks (CNN)	94.5%	Deep learning algorithm effective for image classification tasks.

The table 1 highlights two popular machine learning algorithms, K-Nearest Neighbours (KNN) and Convolutional Neural Networks (CNN), with their respective accuracies and descriptions.

KNN, achieving an accuracy of 95%, is a non-parametric algorithm widely used for both classification and regression tasks. It operates by identifying the 'k' nearest data points in the feature space to a given input and classifying the input based on majority voting among its neighbours. Its simplicity and effectiveness make it suitable for smaller datasets or problems with clear boundaries, though it may struggle with larger, high-dimensional datasets due to computational inefficiency.

CNNs, with a slightly lower accuracy of 94.5%, excel in image classification and recognition tasks. As a deep learning technique, CNNs utilize convolutional layers to extract spatial features from images, followed by pooling and fully connected layers for classification. Their ability to capture intricate patterns and relationships in data makes them the go-to choice for complex visual recognition tasks, such as object detection and facial recognition.

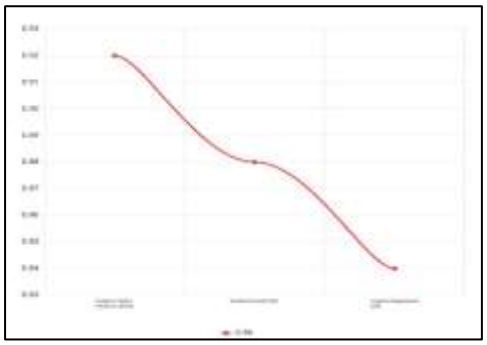

Figure 1: Represents the decreasing trend or performance metric over time or iterations

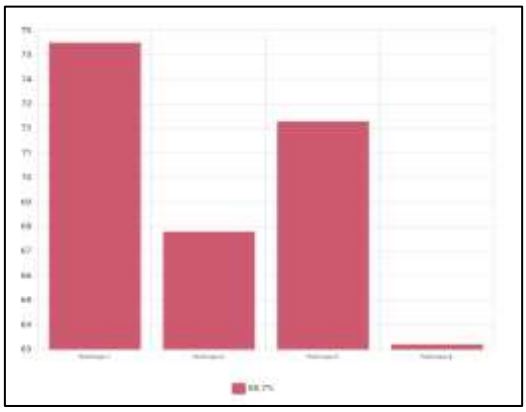

Figure 2: A bar chart comparing distinct categories,

The figure 1 appears to depict a line chart, possibly representing a decreasing trend or performance metric over time or iterations. The x-axis could represent variables such as time, epochs, or iterations, while the y-axis could measure accuracy, loss, or another performance indicator. The steady decline suggests diminishing values, which might signify the behaviour of a training process or some form of decay. This type of graph is commonly used to monitor trends in machine learning, optimization, or statistical analysis, especially when observing how parameters change over time.

The figure 2 appears to be a bar chart comparing distinct categories, likely representing performance metrics, accuracies, or frequencies for different models or methods. Each bar represents a unique category or group, with the y-axis indicating the corresponding quantitative measure. The height of the bars provides a clear visual comparison, highlighting differences or patterns among the categories. This visualization is ideal for showing categorical data in experiments or evaluations, making it easier to identify which category performs better or stands out.

As mentioned, many open-source codes for ML were considered and applied for breast cancer detection but were found that KNN offered better results than CNN and YOLO among the others. This modern approach in the health sector gives high accuracy which is vital in early diagnosis hence, enhancing patient's survival to more than ninety percent. After detection, the type of cancer determines the approach such as neo-adjuvant chemotherapy followed by surgery with the intention of eradicating the cancer. Therefore, as a conclusion, it can be said that the KNN is a better tool in identifying breast cancer.

Conclusion

Finally, this research on Machine learning methodologies for dual-stage breast cancer detection presents a clear understanding of the effectiveness of the proposed methods and algorithms such as CNNs, SVMs, ensemble methods, and others in the identification of cancer in early and advanced stages. Setting emphasis on data collection, cleaning and ethically sound practices, it also provides a roadmap on how to build sound models with feature engineering, tuning and promotion of model biases. However, it is crucial to assimilate rather than globalize utilization of machine learning for improving detection rates since it is an aide for clinical judgment. It will be seen that the subsequent research could focus on the development of enhanced deep learning, Explainable AI, as well as integration of more than one modality data for better early detection and thus better patient outcome. To this end, this report aims at creating the context from which new, revolutionary ML models and platforms can be designed and built, in order to support health care workers and patients alike with a more efficient and honest approach to the diagnosis and combating of breast cancer.

References

Addeh, A., Demirel, H., & Zarbakhsh, P. (2017). Early detection of breast cancer using optimized ANFIS and features selection. In *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)* pp. 39-42. https://doi.org/10.1109/CICN.2017.8319352

Al Salman, H., & Almutairi, N. (2019). IDSS: An intelligent decision support system for breast cancer diagnosis. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)* pp. 1-6. https://doi.org/10.1109/CAIS.2019.8769579

Darapureddy, N., Karatapu, N., & Battula, T. K. (2019, May). Implementation of optimization algorithms on Wisconsin Breast cancer dataset using deep neural network. In *2019 4th International conference on recent trends on electronics, information, communication & technology (RTEICT)* pp. 351-355. https://doi.org/10.1109/RTEICT46194.2019.9016822

Dimmita, N., Nagasri, V., Jyotsna, K. A., Swapna, P., Srikanth, N., Kumar, P. S., ... & Nagalingam, R. (2024). Mammography-based Computer-Aided Diagnostics for the Identification of Breast Cancer Based on Machine Learning. *International Journal of Intelligent Engineering & Systems*, *17*(2). https://doi.org/10.22266/ijies2024.0430.23

Htay, T. T., & Maung, S. S. (2018, September). Early stage breast cancer detection system using glcm feature extraction and k-nearest neighbor (k-NN) on mammography image. In *2018 18th International Symposium on Communications and Information Technologies (ISCIT)* pp. 171-175. https://doi.org/10.1109/ISCIT.2018.8587920

Patil, P. P., & Kotrappa, S. (2020). A novel approach to detect microcalcification for accurate Detection for diagnosis of breast cancer. *Internet of Things, Smart Computing and Technology: A Roadmap Ahead*, 81-94. https://doi.org/10.1007/978-3-030-39047-1_4

Ponraj, N., & Mercy, M. (2017). Texture analysis of mammogram for the detection of breast cancer using LBP and LGP: A comparison. In *2016 eighth international conference on advanced computing (ICoAC)* (pp. 182-185). https://doi.org/10.1109/ICoAC.2017.7951766

Qayyum, A., & Basit, A. (2016). Automatic breast segmentation and cancer detection via SVM in mammograms. In *2016 International conference on emerging technologies (ICET)* pp. 1-6. https://doi.org/10.1109/ICET.2016.7813261

Sangeetha, R., & Murthy, K. S. (2017). A novel approach for detection of breast cancer at an early stage by identification of breast asymmetry and microcalcification cancer cells using digital image processing techniques. In *2017 2nd International Conference for Convergence in Technology (I2CT)* pp. 593-596. https://doi.org/10.1109/I2CT.2017.8226198

Varma, C., & Sawant, O. (2018). An alternative approach to detect breast cancer using digital image processing techniques. In *2018 International conference on communication and signal processing (iccsp)* pp. 0134-0137. https://doi.org/10.1109/ICCSP.2018.8524576