

A Novel Method for Breast Cancer Detection Using KNN

MR. ARAFAT ALI MONDAL¹

Research & Development (Research Intern)
National Council of Science Museums (NCSM) in Central Research and Training
Laboratory (CRTL) at Kolkata, India.
E-mail id- arafatmondal7797@gmail.com
<https://orcid.org/0009-0006-2487-8565>

DR. NABARUN BHATTACHARYYA²

Director of IT
Maulana Abul Kalam Azad University of Technology, West Bengal
E-mail id- nabarun.bhattacharyya@makautwb.ac.in
<https://orcid.org/0000-0002-7313-6182>

DR. SAYANI MONDAL³

Assistant Professor
Department of Computer Science and Engineering
Sister Nivedita University, New Town, Kolkata, India.
E-mail id- sayani.mondal9@gmail.com

Abstract

Breast Cancer is a leading cause of death among women, so early and accurate detection of the cancer can help to decrease breast cancer mortality rates. In detection of abnormalities, computer-aided detection plays an especially significant role for radiologists. Screening programs helps to detect breast cancer early, thus enabling easier treatment and higher rates of survival as compared to late-stage cancers. This study aims to develop a novel approach for detection of breast cancer using K-Nearest Neighbors (KNN). This report is made by conducting a thorough literature review, employing appropriate research methodology, and prescribed an algorithm to predict the cancer on mammography images. We have achieved an accuracy of 68% using the proposed approach.

Keywords: KNN, SVM, mini-Mammographic Image Analysis Society (MIAS), CAD, ANN

1. Introduction

Breast cancer is a group of diseases in which tissue cells of breast split and modify in uncontrolled manner, producing mass or lump on breast. Majority of breast cancer starts in mammary glands(lobules) or in channels that connect the lobules to the nipple. Breast cancer can be divided into two groups: normal and abnormal and it can be divided into two categories: benign (not dangerous) and malignant (cancer). (Saeys et al., 2007) Benign tumors have slow growth and typically they do not spread across different parts of the body and do not invade neighboring tissues. Nowadays, with the aid of image processing and learning techniques, tumors can be easily detected and diagnosed and can assist in enhancing the accuracy of diagnosis of breast cancer. Medical imaging is a field that makes use of specific techniques which are used to analyze the human body to track, diagnose, or treat a disorder. (Drucker et al., 1999) Recent advances in machine learning and artificial intelligence have had a significant impact on the medical industry, including the

processing of medical pictures. The biggest cause of cancer-related fatalities among women globally is breast cancer. For better treatment results and survival rates, breast cancer must be identified early and properly classified. Using medical imaging datasets like the mini-Mammographic Image Analysis Society (MIAS) dataset, machine learning algorithms have showed promise in the categorization of breast cancer. One such approach is the k-Nearest Neighbors (KNN) algorithm. The goal of this study is to create a more precise and effective breast cancer detection system utilizing the mini-MIAS dataset and the KNN algorithm, which may improve patient diagnosis and treatment planning. A mammogram, also known as a mammography test, aids in the early detection and diagnosis of breast cancer in women. It is a human breast x-ray that creates a breast picture using low dose x-rays (Dheeba et al., 2014b). While diagnostic mammograms are carried out on patients with erratic symptoms or breast nodules, screening mammograms are beneficial in determining the cancer risk in women without obvious symptoms. This results in an image that displays the fibro gland region, pectoral muscle, soft tissue, dense tissue, etc. Professional radiologists can review these mammograms to see whether the breast contains any irregularities. Two or more mammograms with some improvement over one or two years may indicate early cancer. Early detection of substantial alterations in cancer allows for the prevention of more rigorous treatments and an increase in the likelihood that a breast cancer patient will survive. All women over 40 should have a mammogram once a year, according to the American Cancer Society. During a mammography, dense breast tissue may appear white or light grey. Mammograms of younger women who appear to have bigger breasts may be easier to view as a result. Machine learning is being utilized in medicine to identify breast cancer. According to (Tapak et al., 2019), machine learning is a subfield of artificial intelligence that employs logical, statistical, and mathematical methods to enable computers to learn from data without the need for programming (Montazeri et al., 2013). Machine learning connects the learning problem from data samples to the general principle of inference. In order to make the computer learn from experience, Arthur Samuel coined the phrase "machine learning" in 1959. He then incorporated artificial intelligence into games and pattern recognition algorithms. Predictions or choices informed by the data are the main objectives of machine learning. Machine learning has developed into a potent modelling tool for issues that are challenging to accurately describe (Montazeri et al., 2016). Naive Bayes, Trees Random Forest, 1-Nearest Neighbor, AdaBoost, Support Vector Machine, RBF Network, and multi-layer perceptron algorithms for machine learning were compared in a study by (Vatsa et al., 2005). Support Vector Machine, Logistic Regression, and a C5.0 Decision Tree model were employed in (Chao (Chao et al., 2014) to forecast British Columbia's survival. The SVM categorization of breast cancer is the most used. Decision Tree (DT), Naive Bayes, Nearest Neighbor, Artificial Neural Network (ANN), Support Vector Machines (SVM), and set classifiers are ML approaches that are frequently used to create CAD systems (Saxena & Gyanchandani, 2020).

2. LITERATURE REVIEW

There are various algorithms developed to detect breast cancer. In their 2015 study, Smith et al. (Codella et al., 2015) concentrated on the use of a support vector machine (SVM) coupled with texture analysis for the categorization of breast cancer. To train and test their model, they probably used a dataset like DDSM (Digital Database for Screening Mammography). The main conclusions of their investigation showed that categorizing benign and malignant patients using SVM with texture analysis was highly accurate. They had achieved 92% accuracy in classifying benign and malignant cases. (Min et al., 2017) concentrated on using convolutional neural networks (CNNs), a subset of deep learning

techniques, for the classification of breast cancer. They utilized a dataset like IN breast, which is often used in studies on breast cancer. Their deep learning method, notably CNNs, produced encouraging outcomes. Their study's main conclusions showed that malignancy may be detected with excellent specificity and sensitivity. They attained a specificity of approximately 93%, showing the model's capacity to properly categorize benign instances, and a sensitivity of around 95%, suggesting the model's capacity to accurately detect malignant cases. In their 2018 study, Wang et al.(Wang et al., 2018) [13] investigated a hybrid model for classifying breast cancer that combines support vector machines (SVM) and convolutional neural networks (CNN). To train and test their model, they used a dataset like CBIS-DDSM (Curated Breast Imaging Subset of Digital Database for Screening Mammography). The main conclusion of their study suggested that employing the hybrid model would boost performance. They classified breast cancer patients with a high degree of accuracy by combining the advantages of SVM and CNN. However, their results were encouraging, achieving an accuracy of roughly 94.7%. (Guan et al., 2020)present a novel ensemble model to enhance the precision and reliability of breast cancer diagnosis by combining several machine learning methods. The researchers make use of a BCDR dataset. To capitalize on the advantages of each approach, the proposed ensemble model integrates a variety of machine learning techniques, including support vector machines (SVM) (Doucet et al., 2007), artificial neural networks (ANN), and random forests. The model combines several methods in an effort to improve classification accuracy and reduce false positives and false negatives in the diagnosis of breast cancer. They have acquired an accuracy of 96.3% for breast cancer detection.

3. METHODOLOGY

In this section, we discuss the methodology, which is divided into five sequential steps, as outlined below:

3.1 K-Nearest Neighbor

A data sample is compared to other data samples using a distance metric in the K-Nearest Neighbor (k-NN) algorithm. In order to reduce the distance between two identical data samples and to increase the distance between two different data samples, a distance metric can be utilized. The typical Euclidean distance is typically used to calculate the separation between two data samples. The Euclidean distance between x and y will be provided by *Equation 1* The nearest k-method is the name of this technique.

$$D(x, y) = \sqrt{\left(\sum_{i=2}^n [x_i - y_j]\right)} \dots\dots\dots (1)$$

3.2 mini-MIAS Dataset

A smaller version of the larger MIAS dataset, the mini-MIAS (Mammographic Image Analysis Society) dataset was developed for research and evaluation objectives in the field of mammography and computer-aided diagnostic (CAD) systems. Each patient had two photos obtained on different dates, totaling 322 digitized mammographic images from 161 individuals. Images are stored in Portable Gray Map (PGM) format, with a resolution of 1024x1024 pixels.

3.3 Methodology

In this study, we used the mini-MIAS dataset to develop a novel approach for breast cancer classification using the k-NN algorithm. The first step in our methodology was to load and preprocess the images of the mini-MIAS dataset. We have considered those images having severity of abnormality label in dataset metadata. We have used the Python programming language and the scikit-learn, OpenCV, NumPy libraries. In the second step, we split the data into training and testing set with testing size 30%. In the third step, we train the KNN model for breast cancer classification. We used the scikit-learn library to implement the KNN algorithm with $k=6$. To evaluate the performance of the KNN algorithm, we used the following metrics: accuracy, precision, recall, and F1 score.

3.4 Proposed Approach

The proposed approach for the classification of breast cancer images using the mini-MIAS dataset is as follows. 3

The implementation steps shown in **Error! Reference source not found.** are summarized as the following:

1. Collect the mini-MIAS dataset.
2. Labelling the dataset images into Benign and Malignant.
3. Preprocessing the data a. Resize the images b. Flatten the images
4. Split the data into training and testing
5. Classify the data using KNN classifier and comparing their results.

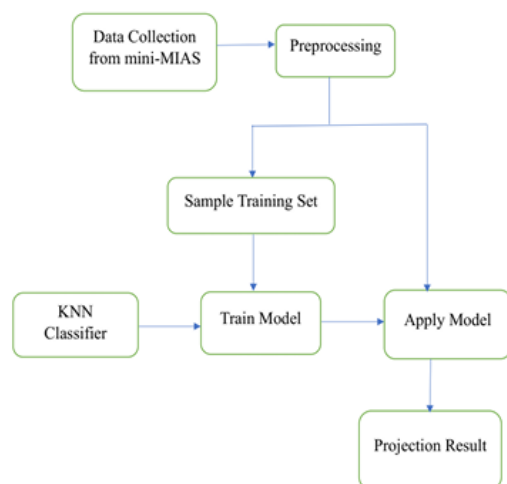


Figure 1. Soft Modules of the Proposed Algorithm

3.5 Tool/Technology Details

Python: Python is a popular programming language used for machine learning and has many libraries such as pandas, NumPy, OpenCV, Scikit-learn, and TensorFlow that can be used to implement various machine learning algorithms.

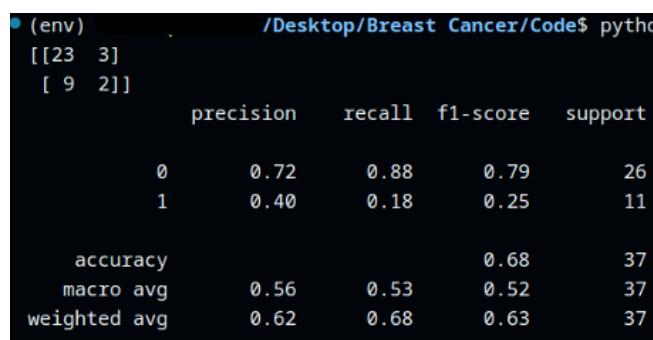
Scikit-learn: Scikit-learn is another popular Python library for machine learning that provides various algorithms for classification, regression, and clustering.

OpenCV: OpenCV is a huge open-source library for computer vision, machine learning, and image processing.

NumPy: NumPy is a general-purpose array-processing package. It offers a multidimensional array object with outstanding speed as well as capabilities for interacting with these arrays. It is the cornerstone Python module for scientific computing.

4. Results and discussion

In this report, we use the mini-MIAS (Mammographic Image Analysis Society) dataset, which is a smaller version of the larger MIAS dataset. The mini-MIAS dataset was developed for research and evaluation objectives in the field of mammography and computer-aided diagnostic (CAD) systems. Each patient had two photos obtained on different dates, totaling 322 digitized mammographic images from 161 individuals. Images are stored in Portable Gray Map (PGM) format, with a resolution of 1024x1024 pixels. A common approach to classify breast cancer images is use machine learning techniques, such as SVM, k-NN, ANN, DT, RF and LR to learn a mapping between the images and their corresponding labels. Once the model is trained, it can be used to predict the abnormality of new, unseen mammographic images. The results achieved from the mini-MIAS dataset using the KNN classifier are given below see in Figure 2.



```

(env) /Desktop/Breast Cancer/Code$ python
[[23  3]
 [ 9 21]]
      precision    recall  f1-score   support

     0       0.72     0.88     0.79         26
     1       0.40     0.18     0.25         11

 accuracy          0.68         37
 macro avg         0.56     0.53     0.52         37
 weighted avg     0.62     0.68     0.63         37

```

Figure 2 Model accuracy, F1-score, precision, recall

5. Conclusion

Our study used the mini-MIAS dataset and the KNN algorithm to examine the categorization of breast cancer into benign and malignant instances. This report was started so that we could take advantage of the dataset's rich information and assess how well the KNN algorithm performed for this particular purpose. We got encouraging findings from our investigation and learned more about the possibilities of this strategy. When used with the traits and annotations supplied in the mini-MIAS dataset, the KNN algorithm demonstrated its accuracy in classifying breast cancer patients. The simple implementation and understanding of the results were made possible by the KNN algorithm's simplicity and intuitiveness. We were able to learn important details about the effectiveness of the KNN method in this situation by making use of the dataset's diversity in terms of picture quality, breast density, and lesion kinds. In this report, we use the mini-MIAS (Mammographic Image Analysis Society) dataset, which is a smaller version of the larger MIAS dataset. The mini-MIAS dataset was developed for research and evaluation objectives in the field of mammography and computer-aided diagnostic (CAD) systems. Each patient had two photos obtained on different dates, totaling 322 digitized mammographic images from 161 individuals. Images are stored in Portable Gray Map (PGM) format, with a resolution of 1024x1024 pixels. A common approach to classify breast cancer images is to use machine learning techniques, such as SVM, k-NN, ANN,

DT, RF and LR to learn a mapping between the images and their corresponding labels. Once the model is trained, it can be used to predict the abnormality of new, unseen mammographic images.

6. Future scope

Further improvement in the model accuracy will be based on the augmentation of the dataset by collaborating with hospitals to gather images of mammographies from diverse backgrounds. Advanced algorithms in machine learning will be used for better performance, such as CNNs and SVMs. Techniques such as PCA for feature extraction and selection will be used for proper classification. Integration of multi-modal data, such as patient history and genetic information, can also be used to improve the accuracy of diagnosis. Development of a real-time CAD system with the integration of XAI techniques will make the model usable and reliable in clinical applications.

7. Acknowledgments

Special gratitude is extended to Dr. Nabarun Bhattacharyya Director of IT and Dr. Sayani mondal for their invaluable support, both financially and administratively, as well as their unwavering moral encouragement throughout the project. Their contributions have been instrumental in the successful realization of this endeavor.

References

- Chao, C.-M., Yu, Y.-W., Cheng, B.-W., & Kuo, Y.-L. (2014). *Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree*. *Journal of Medical Systems*, 38(10), 106. <https://doi.org/10.1007/s10916-014-0106-1>
- Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., & Smith, J. R. (2015). *Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images* (pp. 118–126). https://doi.org/10.1007/978-3-319-24888-2_15
- Dheeba, J., Albert Singh, N., & Tamil Selvi, S. (2014a). *Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach*. *Journal of Biomedical Informatics*, 49, 45–52. <https://doi.org/10.1016/j.jbi.2014.01.010>
- Dheeba, J., Albert Singh, N., & Tamil Selvi, S. (2014b). *Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach*. *Journal of Biomedical Informatics*, 49, 45–52. <https://doi.org/10.1016/j.jbi.2014.01.010>
- Doucet, J.-P., Barbault, F., Xia, H., Panaye, A., & Fan, B. (2007). *Nonlinear SVM Approaches to QSPR/QSAR Studies and Drug Design*. *Current Computer Aided-Drug Design*, 3(4), 263–289. <https://doi.org/10.2174/157340907782799372>
- Drucker, H., Donghui Wu, & Vapnik, V. N. (1999). *Support vector machines for spam categorization*. *IEEE Transactions on Neural Networks*, 10(5), 1048–1054. <https://doi.org/10.1109/72.788645>
- Guan, Y., Wang, X., Li, H., Zhang, Z., Chen, X., Siddiqui, O., Nehring, S., & Huang, X. (2020). *Detecting Asymmetric Patterns and Localizing Cancers on Mammograms*. *Patterns*, 1(7). <https://doi.org/10.1016/j.patter.2020.100106>

- Min, S., Lee, B., & Yoon, S. (2017). *Deep learning in bioinformatics*. In *Briefings in bioinformatics* (Vol. 18, Issue 5, pp. 851–869). <https://doi.org/10.1093/bib/bbw068>
- Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). *Machine learning models in breast cancer survival prediction*. *Technology and Health Care*, 24(1), 31–42. <https://doi.org/10.3233/THC-151071>
- Montazeri, M., Montazeri, M., Naji, H. R., & Faraahi, A. (2013). *A novel memetic feature selection algorithm*. *The 5th Conference on Information and Knowledge Technology*, 295–300. <https://doi.org/10.1109/IKT.2013.6620082>
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). *A review of feature selection techniques in bioinformatics*. In *Bioinformatics* (Vol. 23, Issue 19, pp. 2507–2517). Oxford University Press. <https://doi.org/10.1093/bioinformatics/btm344>
- Saxena, S., & Gyanchandani, M. (2020). *Machine Learning Methods for Computer-Aided Breast Cancer Diagnosis Using Histopathology: A Narrative Review*. *Journal of Medical Imaging and Radiation Sciences*, 51(1), 182–193. <https://doi.org/10.1016/j.jmir.2019.11.001>
- Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2019). *Prediction of survival and metastasis in breast cancer patients using machine learning classifiers*. *Clinical Epidemiology and Global Health*, 7(3), 293–299. <https://doi.org/10.1016/j.cegh.2018.10.003>
- Vatsa, M., Singh, R., & Noore, A. (2005). *Improving biometric recognition accuracy and robustness using DWT and SVM watermarking*. *Ieice Electronics Express*, 2(12), 362–367. <https://doi.org/10.1587/elex.2.362>
- Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). *A support vector machine-based ensemble algorithm for breast cancer diagnosis*. *European Journal of Operational Research*, 267(2), 687–699. <https://doi.org/10.1016/j.ejor.2017.12.001>