International journal of innovation in Engineering

journal homepage: www.ijie.ir

Research Paper

Classification of Mammogram Images by Using SVM and KNN

Anuradha Reddy[a1], Vignesh Janarthanan[b], G Vikram[a], K Mamatha[a]

[a] *Department of Computer Science and Engineering, MRITS, Maisammaguda, Secunderabad, India*

[b] *Department of Information Technology, MRITS, Maisammaguda, Secunderabad, India*

ARTICLE INFO

Received: 12 July 2022

Reviewed: 24 July 2022

Revised: 7 August 2022

Accepted: 20 August 2022

Keywords:

Breast Cancer, Image Classification, Mammogram, SVM, KNN.

ABSTRACT

Breast cancer is a fairly diverse illness that affects a large percentage of women in the west. A mammogram is an X-ray-based evaluation of a woman's breasts to see if she has cancer. One of the earliest prescreening diagnostic procedures for breast cancer is mammography. It is well known that breast cancer recovery rates are significantly increased by early identification. Mammogram analysis is typically delegated to skilled radiologists at medical facilities. Human mistake, however, is always a possibility. Fatigue of the observer can commonly lead to errors, resulting in intraobserver and interobserver variances. The image quality affects the sensitivity of mammographic screening as well. The goal of developing automated techniques for detection and grading of breast cancer images is to reduce various types of variability and standardize diagnostic procedures. The classification of breast cancer images into benign (tumor increasing, but not harmful) and malignant (cannot be managed, it causes death) classes using a two-way classification algorithm is shown in this study. The two-way classification data mining algorithms are utilized because there are not many abnormal mammograms. The first classification algorithm, k-means, divides a given dataset into a predetermined number of clusters. Support Vector Machine (SVM), a second classification algorithm, is used to identify the optimal classification function to separate members of the two classes in the training data

[1] **Corresponding Author**
anuradhareddy.anu@gmail.com

1. Introduction

The term CANCER describes the unchecked proliferation of a collection of cells in a specific area of the body. Cancer is a dangerous condition in which the body develops growths of cells known as cancers that destroy healthy body cells (Patan et al., 2020). A lump or mass of additional tissue can develop from a collection of cells that divide quickly. Tumors are the common name for these lumps (Krishna et al., 219). Malignant tumors are cancer cells. Any malignant tumor that arises from breast cells is considered to be breast cancer. Architectural deformities, masses, and clusters of micro calcifications are significant indicators to watch for in the case of breast cancer (Ghantasala et al., 2020a). Breast cancer incidence has significantly increased in recent years. Parallel to these advancements in treatment methods and diagnostic tools, the survival rate for breast cancer has also increased over the past few years (Bhowmik et al., 2021).

Through X-ray mammography, which requires the breast tumor to have advanced to a point where it is much denser than healthy tissue, breast cancer screening has generally taken an anatomical approach (Chandana et al., 2020). Because these no palpable breast tumors are not significantly denser than healthy tissue, mammography misses 5%–15% of them. Additionally, higher density is not always associated with the presence of cancer: when dense tissue lesions are further examined via biopsy, benign lesions are frequently discovered (Ghantasala & Kumari, 2021a). Early molecular signals can be used to identify cancer in addition to density alterations (CADe, 2020; Ghantasala & Kumari, 2021b).

In 2011, the American Cancer Society anticipated that there would be around 230,480 new instances of invasive breast cancer, 57,650 new cases of noninvasive breast cancer, and about 39,520 new breast cancer-related deaths in the United States (Ghantasala et al., 2021a). Mammography, the most widely used diagnostic method, makes use of low-dose X-rays, high-contrast and high-resolution detectors, and an X-ray system specifically intended to picture the breasts (Kishore et al., 2021; Rupa et al., 2022; Ghantasala et al., 2021b). Mammography is now used for both breast cancer screening and diagnosis. There are two different mammography systems: full-field digital mammography (FFDM), which employs digital detectors as the recording media, and screen film mammography (SFM), which uses a film screen as the final recording device (Ghantasala et al., 2021c; Reddy et al., 2021). In terms of simplicity of image processing and enhancement, the digital images offered by FFDM have a number of advantages over their film counterpart.

One of the key ways to some extent to detect breast cancer at an early stage is the digital mammogram. The lack of ionizing radiation, non-invasiveness, relatively small instrumentation, and cost-effectiveness of digital mammography are some of its benefits. Additionally, as indicated by Wroblewska et al. (2003), abnormalities are frequently concealed by and entrenched in different densities of breast tissue structures when viewing a mammographic image, leading to high rates of missed breast cancer cases

2. Literature Review

A neural network approach was utilized by Youssry et al. (2002) to identify potential confined lesions in digitalized mammograms. Back propagation algorithms were used to train the neural network. The process mostly depends on the size of the discrepancy between the histograms of normal and malignant tissue. Digital mammograms with single and multiscale masses were explored by te Brake & Karssemeijer (1999). Due to the wide range of sizes that masses might have, scale is a crucial factor in the automated detection of masses in mammograms. In this study, it was investigated whether mass detection could be accomplished at a single scale or whether it would be better to employ the output of the detection method at many scales in a multiscale system. A computer-based technique for detecting micro calcification in digital mammograms was examined by Chan & Doi (1988). The methodology relies on a difference image technique to eliminate structured background from the mammography by subtracting a signal suppressed image from a signal enhanced image.

Following that, possible micro calcification signals are extracted using global and local thresholding approaches. A statistical technique for detecting micro calcifications in digital mammograms was developed by Karssemeijer (1993). The technique is based on the application of statistical models and the broad Bayesian image analysis framework.

A filter bank was employed by Nakayama & Uchiyama (2005) to find nodular and linear patterns. At each resolution level, the filter bank is constructed so that the sub images produce the components of a Hessian matrix. The following three characteristics of a new filter bank are determined by computing the small and large eigenvalues. (A) Different-sized nodular patterns can be improved. (b) Different-sized nodular and linear patterns can both be improved. (c) By removing these patterns, the original image can be recreated. The filter bank is used to improve mammogram microcalcifications.

3. Applied Algorithms

3.1. K-means algorithm

The procedure works with a set of vectors in the d-dimensional space D = xi | I = 1,..., N, where xi stands for the ith data point. The method is started by selecting the first k cluster representatives, or "centroids," at k sites (Mandal et al., 2020; Ghantasala et al., 2020b). Techniques for choosing these initial seeds include randomly selecting samples from the dataset, using them as the clustering solution for a small section of the data, or disrupting the data's global mean k times.

The algorithm then repeats these two phases until convergence:

- *Data assignment is the first step: Each data point is matched to its nearest centroid, with ties being broken at random. The data is divided as a result.*
- *Relocating the "means" is step two: The mean (middle) of all the data points allocated to each cluster representative is moved there. The relocation is to the weighted mean of the data partitions if the data points have a probability measure (weights).*

Since Euclidean distance is the standard measure of proximity, it is simple to demonstrate that the non-negative cost function

3.2. Support vector machines

Support vector machines (SVM) are regarded as a must-try in today's machine learning applications as it provides one of the most reliable and accurate approaches among all well-known algorithms (Kumari & Ghantasala, 2020). It has a strong theoretical underpinning, can be trained with just a dozen examples, and is not sensitive to the number of dimensions. Additionally, effective SVM training techniques are being created at a rapid rate.

A maximum margin hyperplane between the two classes is found using Support Vector Machines and data from two classes (Reddy et al., 2022). The hyperplane is chosen so that the support vectors—the closest data points on either side—are as far away from it as possible (Gadde et al., 2022). By applying a kernel function to the data to make it linearly separable; support vector machine (SVM) classifiers can be extended to nonlinearly separable data (Pradeep Ghantasala et al., 2022). The linear kernel, polynomial kernels of orders 1, 2, and 3, and the radial basis function kernel were all employed in this study. It was mentioned in how to use a similar strategy using wavelet SVM.

4. Architecture of the proposed system

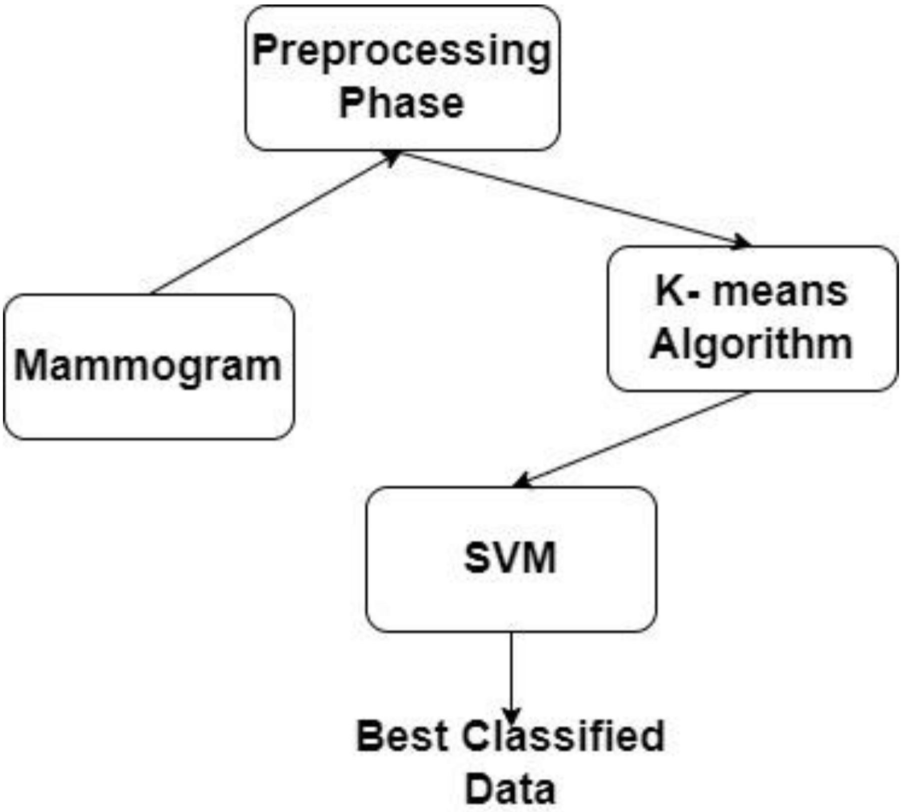

Fig. 1. Basic Architecture

Mammogram results are used as an input and are sent to the preprocessing stage for data filtering. An essential consideration in low-level image processing is pre-processing. It is possible to remove the noise from an image using filtering. A low pass filter lowers the frequent variations in an image's grey level while a high pass filter passes frequently changing grey levels. In other words, the low pass filter softens and frequently gets rid of the sharp edges. The Median filter is a unique kind of low pass filter. The Median filter observes all pixel values in an area of the image (3 x 3, 5 x 5, 7 x 7, etc.), records them, and adds them to an array known as the element array. The element array is sorted after that, and the median value is determined. This was accomplished by using bubble sort to sort the element array in ascending order, returning the middle items of the sorted array as the median value. The collection of all the median values from the element arrays collected for each and every pixel makes up the output image array. The median filter enters a series of loops that round the whole image array. The preprocessed data is given into the first classification algorithm following the preparation phase (i. e. k-means algorithm). Processed data can be transformed into specific grouped data using the k-means algorithm. The SVM algorithm then receives the clustered data as input and generates the best categorised data.

5. Conclusion

Mammographic image analysis has been given a fresh framework for two-way categorization approach. It is uncommon to use two-way classification to the issue of mammographic image analysis, according to a survey of the literature that is currently available. We are confident that by defining new functionalities that are better suited to mammography, the performance of the suggested system can be further improved. In this article, two-way categorization architecture is shown in a very general manner. It exemplifies how an abstract structure enables us to find useful classification of breast cancer imaging data. Because of its simplicity and ability to produce results that will inspire a real-time breast cancer diagnosis system, this algorithm has been built for use in future research and development

References

- Bhowmik, C., Ghantasala, G. P., &AnuRadha, R. (2021). A Comparison of Various Data Mining Algorithms to Distinguish Mammogram Calcification Using Computer-Aided Testing Tools. In Proceedings of the Second International Conference on Information Management and Machine Intelligence (pp. 537-546). Springer, Singapore.
- CADe, M. (2020). CADx for Identifying Microcalcification Using Support Vector Machine. Journal of Communication Engineering & Systems, 10(2), 9-16p.
- Chan, F. P., & Doi, K. (1988). CJ\Tyhoiny, IL. Lam, and Ri\. Schmidt. Computer-aided detection of microcalcifications in mammograms: Methodology and pre1iiiinaiy clinica. I study. Investgati've Radiology, 23(9), 664-671.
- Chandana, P., Ghantasala, G. P., Jeny, J. R. V., Sekaran, K., Deepika, N., Nam, Y., &Kadry, S. (2020). An effective identification of crop diseases using faster region based convolutional neural network and expert systems. International Journal of Electrical and Computer Engineering (IJECE), 10(6), 6531-6540.
- Gadde, S. S., Anand, D., Sasidhar Babu, N., Pujitha, B. V., Sai Reethi, M., & Pradeep Ghantasala, G. S. (2022). Performance Prediction of Students Using Machine Learning Algorithms. In Applications of Computational Methods in Manufacturing and Product Design (pp. 405-411). Springer, Singapore.
- Ghantasala, G. P., & Kumari, N. V. (2021a). Breast Cancer Treatment Using Automated Robot Support Technology ForMri Breast Biopsy. INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS, 1(2), 235-242.
- Ghantasala, G. P., &Kumari, N. V. (2021b). Identification of Normal and Abnormal Mammographic Images Using Deep Neural Network. Asian Journal For Convergence In Technology (AJCT), 7(1), 71-74.
- Ghantasala, G. P., Kallam, S., Kumari, N. V., &Patan, R. (2020a, March). Texture Recognization and Image Smoothing for Microcalcification and Mass Detection in Abnormal Region. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.
- Ghantasala, G. P., Kumari, N. V., &Patan, R. (2021c). Cancer prediction and diagnosis hinged on HCML in IOMT environment. In Machine Learning and the Internet of Medical Things in Healthcare (pp. 179-207). Academic Press.
- Ghantasala, G. P., Rao, D. N., & Mandal, K. (2021a). Machine Learning Algorithms Based Breast Cancer Prediction Model. Journal of Cardiovascular Disease Research, 12(4), 50-56.
- Ghantasala, G. P., Reddy, A. R., & Arvindhan, M. (2021b). Prediction of Coronavirus (COVID-19) Disease Health Monitoring with Clinical Support System and Its Objectives. In Machine Learning and Analytics in Healthcare Systems (pp. 237-260). CRC Press.
- Ghantasala, G. P., Tanuja, B., Teja, G. S., & Abhilash, A. S. (2020b). Feature Extraction and Evaluation of Colon Cancer using PCA, LDA and Gene Expression. Forest, 10(98), 99.
- Kishore, D. R., Syeda, N., Suneetha, D., Kumari, C. S., &Ghantasala, G. P. (2021). Multi Scale Image Fusion through Laplacian Pyramid and Deep Learning on Thermal Images. Annals of the Romanian Society for Cell Biology, 3728-3734.

- Krishna, N. M., Sekaran, K., Vamsi, A. V. N., Ghantasala, G. P., Chandana, P., Kadry, S., ... &Damaševičius, R. (2019). An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using EEG signals. IEEE Access, 7, 77905-77914.

- Kumari, N. V., &Ghantasala, G. P. (2020). Support Vector Machine Based Supervised Machine Learning Algorithm for Finding ROC and LDA Region. Journal of Operating Systems Development & Trends, 7(1), 26-33.

- Mandal, K., Ghantasala, G. P., Khan, F., Sathiyaraj, R., & Balamurugan, B. (2020). Futurity of Translation Algorithms for Neural Machine Translation (NMT) and Its Vision. In Natural Language Processing in Artificial Intelligence (pp. 53-95). Apple Academic Press.

- Nakayama, R., & Uchiyama, Y. (2005). Development of new filter bank for detection of nodular patterns and linear patterns in medical images. Systems and Computers in Japan, 36(13), 81-91.

- NKarssemeijer, N. (1993, July). Recognition of clustered microcalcifications using a random field model. In Biomedical Image Processing and Biomedical Visualization (Vol. 1905, pp. 776-786). SPIE.

- Patan, R., Ghantasala, G. P., Sekaran, R., Gupta, D., & Ramachandran, M. (2020). Smart healthcare and quality of service in IoT using grey filter convolutional based cyber physical system. Sustainable Cities and Society, 59, 102141.

- Pradeep Ghantasala, G. S., Nageswara Rao, D., & Patan, R. (2022). Recognition of Dubious Tissue by Using Supervised Machine Learning Strategy. In Applications of Computational Methods in Manufacturing and Product Design (pp. 395-404). Springer, Singapore.

- Reddy, A. R., Ghantasala, G. S., Patan, R., Manikandan, R., & Kallam, S. (2021). Smart Assistance of Elderly Individuals in Emergency Situations at Home. In Internet of Medical Things (pp. 95-115). Springer, Cham.

- Reddy, A., Gude, V., Mamatha, K., & Rao, D. N. (2022). Smart Waste Management Systems by Using Automated Machine Learning Techniques. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172, 2(04), 16-25.

- Rupa, C., MidhunChakkarvarthy, D., Patan, R., Prakash, A. B., & Pradeep, G. G. (2022). Knowledge engineering–based DApp using blockchain technology for protract medical certificates privacy. IET Communications.

- Te Brake, G. M., & Karssemeijer, N. (1999). Single and multiscale detection of masses in digital mammograms. IEEE transactions on medical imaging, 18(7), 628-639.

- Wroblewska, A., Boninski, P., Przelaskowski, A., & Kazubek, M. (2003). Segmentation and feature extraction for reliable classification of microcalcifications in digital mammograms. Optoelectronics Review, (3), 227-236.

- Youssry, N., Abou-Chadi, F., & El-Sayad, A. M. (2002, December). A neural network approach for mass detection in digitized mammograms. In accepted to be published in the 1st Annual Conference of Biomedical Engineering.