MULTI-MODAL IMAGING AND AI FOR AUTOMATED BREAST CANCER DIAGNOSIS SUPPORT SYSTEM

[1]Vidhya Shenigaram,[2]Susheel Kumar Thakur,[3]Choul Praveen Kumar,[4]Laxman Maddikunta
[1,2,3]AssistantProfessor,[4]Associate Professor
Department of CSE

Kshatriya College of Engineering

ABSTRACT

Breast cancer is a prevalent disease worldwide, with over 1.15 million cases diagnosed annually. Currently, the clinical management of breast cancer relies on a limited number of accurate prognostic and predictive factors. Early detection plays a crucial role in reducing mortality rates and improving the survival period of breast cancer patients. Mammography is the primary screening and diagnostic test used, and the analysis and processing of mammograms are key to enhancing breast cancer prognosi.In this project, the detection of breast cancer in mammograms is achieved through image segmentation using the Fuzzy C-means (FCM) technique. The FCM algorithm is applied to segment the mammogram into distinct regions. Following segmentation, features are extracted from these segmented regions. The extracted features are then utilized to train a classifier capable of accurately categorizing different classes in mammograms. Texture features, which provide important information about the texture patterns within the mammogram, are extracted using techniques such as multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA), and Gray-level Cooccurrence Matrix (GLCM).To distinguish masses and microcalcifications from the background tissue, morphological operators are employed. These operators help in identifying and separating tumor-affected regions from the surrounding healthy tissue. The K-nearest neighbors (KNN) algorithm is utilized as a classification technique for assigning mammogram images to their respective classes based on the extracted features.The boundaries of the tumor-affected regions in the mammogram are marked and displayed to the doctor for further examination. Additionally, the area of the tumor is provided to assist in evaluating the size and extent of the tumor. This information aids in the diagnosis and treatment planning for breast cancer patients. It’s important to note that while this description provides an overview of the project, the implementation details and specific algorithms used may vary. It is crucial to conduct further research and refer to academic papers or specific sources to gain a comprehensive understanding of the techniques and methods employed in breast cancer detection using mammography.

I. INTRODUCTION

Breast cancer is a significant cause of mortality among women worldwide, emphasizing the importance of early detection and accurate diagnosis for effective treatment. Traditional diagnostic approaches heavily rely on the expertise of physicians and visual inspections, which can be limited by human error and subjectivity. Automatic diagnosis of breast cancer has thus been the subject of extensive research, aiming to improve diagnostic accuracy through computer-aided tools.Machine learning techniques have demonstrated potential in enhancing the accuracy of breast cancer diagnosis. In a study conducted by Brause, it was found that the accuracy of diagnosis by the most experienced physician was 79.97%, whereas machine learning achieved a correct diagnosis rate of 91.1%. This highlights the potential of machine learning in improving diagnostic outcomes.

The performance of backpropagation artificial neural networks (ANN) was evaluated to enhance the accuracy of breast mass classification as benign or malignant. Radial basis function neural networks (RBFNN) have also shown excellent accuracy in microcalcification detection tasks, owing to their fast learning rates and generalization capabilities. RBFNN's advantages include a simple structure, good performance in handling nonlinear functions, and fast convergence. However, the performance of RBFNN can be affected by the increase in network structure when the input dimension grows, as well as the presence of irrelevant input components.

Support Vector Machines (SVM) is an effective statistical learning method for classification. It works by finding an optimal hyperplane that can separate different classes by mapping the input data into a higher-dimensional feature space. SVM exhibits fast training techniques, even with a large number of input data, making it suitable for recognition problems such as object recognition and face detection. Principal Component Analysis (PCA) is a technique used to reduce dimensionality based on second-order statistical information. Independent Component Analysis (ICA), on the other hand, relies on higher-order statistics to extract independent components that contain richer information compared to PCA. ICA can be employed to reduce dimensionality before training classifiers such as k-NN, ANN, RBFNN, and SVM. This reduction in complexity can lead to increased convergence velocity and performance.

The objective of the proposed study is to analyse the impact of feature reduction using ICA on the classification of tumors as benign or malignant. The WDBC dataset's dimension is reduced to a single feature using ICA. The reduced data is then divided into test and training sets using techniques like 5/10-fold cross-validation and 20% partitioning. Performance measures including accuracy, specificity, sensitivity, F-score, Youden's index, and discriminant power are computed, and the classifiers are compared using the receiver operating characteristic (ROC) curve.

In summary, this project aims to explore the effect of feature reduction using ICA on the classification of benign and malignant tumors. Various classifiers such as k-NN, ANN, RBFNN, and SVM are evaluated using performance measures, and their effectiveness is compared using ROC curves. The study encompasses background knowledge on the dataset, ICA, and the classifiers, followed by a detailed methodology, experimental results, and discussions.

II. LITERATURE SURVEY

[1] Siya bend Turgut et al., "Microarray Breast Cancer Data Classification Using Machine Learning Methods" [IEEE 2018] The paper uses microarray breast cancer data for classification of the patients using machine learning methods. In the first case, eight different machine learning algorithms are applied to the dataset and the results of classification were noted. Then in the second case, two different feature selection methods such as Recursive Feature Elimination (RFE) and Randomized Logistic Regression (RLR) were applied on the microarray breast cancer dataset and 50 features were chosen as stop criterion. Again, the same eight machine learning algorithms were applied on the modified dataset. The results of the classifications are compared with each other and with the results of the first case. The methods applied are SVM, KNN, MLP, Decision Trees, Random Forest, Logistic Regression, Ad boost and Gradient Boosting Machines. After applying the two different feature selection methods, SVM gave the best results. MLP is applied using different number of layers and neurons to examine the effect of the number of layers and neurons on the classification accuracy [3].

[2] Varalatchoumy M et al., "Four Novel Approaches for Detection of Region of Interest in Mammograms - A Comparative Study" [ICISS 2017] The paper compares Four Novel approaches used for detection of Region of Interest in Mammographic images based on database and Real time images. In Approach I histogram equalization and dynamic thresholding techniques were used for preprocessing. Region of Interest (ROI) was partitioned from the preprocessed image by using particle swarm optimization and k means clustering methods. In Approach II preprocessing was done using various morphological operations like erosion followed by dilation. For the identification of ROI, a modified approach of watershed segmentation was used. Approach III uses histogram equalization for preprocessing and an advanced level set approach for performing segmentation. Approach IV, which is considered to be the most efficient approach that uses different morphological operations and contrast limited adaptive histogram equalization for image preprocessing. A very novel algorithm was developed for detection of Region of Interest. Approaches I and II were applicable for Mammographic Image Analysis Society (MIAS) database images alone. Approaches III and IV were applicable for MIAS and Real time hospital images. The various graphs presented in the comparative study, clearly depicts that the novel approach that used a novel algorithm for detection of ROI is proved to be the most efficient, accurate and highly reliable approach that can be used by radiologists to detect tumors in MRM images [4].

[3] Ammu P K et al., "Review on Feature Selection Techniques of DNA Microarray Data" [IJCA 2013] This paperreviews few major feature selection techniques employed in microarray data and points out the merits and demerits of various approaches. Feature selection from DNA microarray data is one of the most important procedures in bioinformatics. Biogeography Based Optimization (BBO) is an optimization algorithm which works on the basis of migration of species between different habitats and the process of mutation. Particle Swarm Optimization (PSO) is an algorithm which works on the basis of movement of particles in a search space. Redundancy based feature selection approaches can be used to remove redundant genes from the selected genes as the resultant gene set can achieve a better representation of the target class. A two-stage hybrid filter wrapper method where, in the first stage a subset of the original feature set is obtained by applying information gain as the filtering criteria. In the second stage the genetic algorithm is applied to the set of filtered genes. Gene selection based on dependency of features where the features are classified as independent, half dependent and dependent features. Independent features are those features that doesn't depend on any other features. Half dependent features are more relevant in correlation with other features and dependent features are fully dependent on other features [5].

III. PROPOSED SYSTEM

In gene analysis, it is important to select relevant genes that play a significant role in determining various biological processes. Gene selection

techniques based on feature dependency have been explored to identify independent, half dependent, and dependent features.
Independent features refer to those genes that do not depend on any other genes. These genes exhibit their influence on biological processes without being influenced by other genes. They provide unique information and insights into specific characteristics or functions.

Half dependent features are considered to have a moderate level of dependency on other genes. These genes exhibit correlation or association with certain other genes, indicating their relevance in specific biological pathways or interactions. While they may have some level of dependence, they also possess individual importance and contribute significantly to the overall understanding of gene behavior.

Dependent features, as the name suggests, are fully dependent on other genes. These genes rely on the expression or behavior of other genes to manifest their impact. They do not provide independent information but rather act as downstream indicators or markers of other genes' activities or variations.

The categorization of features into independent, half dependent, and dependent groups helps in understanding the interplay and relationships among genes. It assists in identifying key genes that drive biological processes independently, as well as those that are strongly influenced by other genes.

It's worth mentioning that the citation [5] is provided to acknowledge the source from which this categorization of features based on dependency is derived. However, without access to the specific source, it is not possible to provide further details or context regarding the citation.

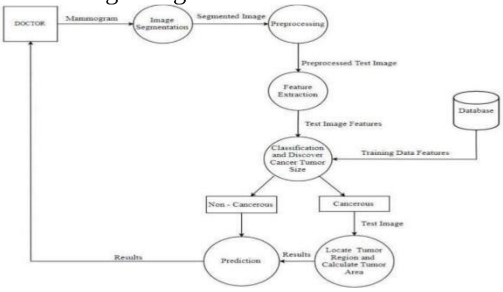
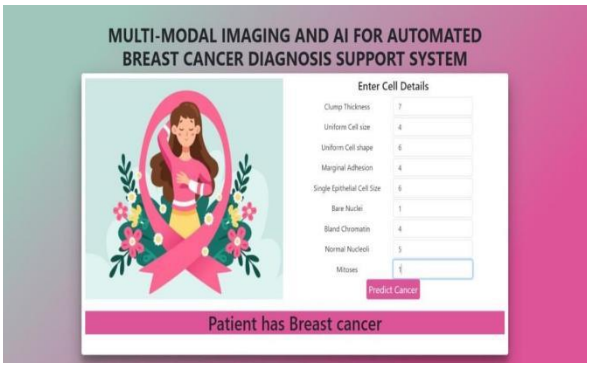


Figure 1. Proposed breast cancer prediction and Tracking flow diagram

IV. RESULTS


MULTI-MODAL IMAGING AND AI FOR AUTOMATED BREAST CANCER DIAGNOSIS SUPPORT SYSTEM

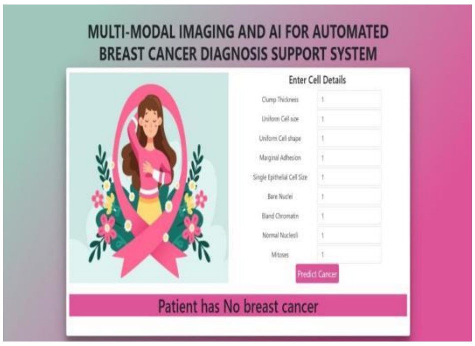
Enter Cell Details

Clump Thickness 7
Uniform Cell size 4
Uniform Cell shape 6
Marginal Adhesion 4
Single Epithelial Cell Size 6
Bare Nuclei 1
Bland Chromatin 4
Normal Nucleoli 5
Mitoses 1

Predict Cancer

Patient has Breast cancer

Figure 2: Output screen of predicting cancer with random values


MULTI-MODAL IMAGING AND AI FOR AUTOMATED BREAST CANCER DIAGNOSIS SUPPORT SYSTEM

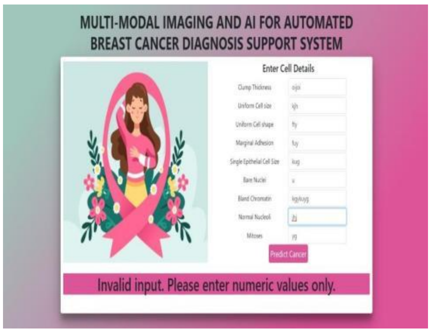
Enter Cell Details

Clump Thickness 1
Uniform Cell size 1
Uniform Cell shape 1
Marginal Adhesion 1
Single Epithelial Cell Size 1
Bare Nuclei 1
Bland Chromatin 1
Normal Nucleoli 1
Mitoses 1

Predict Cancer

Patient has No breast cancer

Figure 3: Output screen of patient with no breast cancer


MULTI-MODAL IMAGING AND AI FOR AUTOMATED BREAST CANCER DIAGNOSIS SUPPORT SYSTEM

Enter Cell Details

Clump Thickness ojoi
Uniform Cell size kjh
Uniform Cell shape fty
Marginal Adhesion fuy
Single Epithelial Cell Size kug
Bare Nuclei v
Bland Chromatin kgykuyg
Normal Nucleoli jhj
Mitoses yg

Predict Cancer

Invalid input. Please enter numeric values only.

Figure 4: Output screen of entering invalid inputs

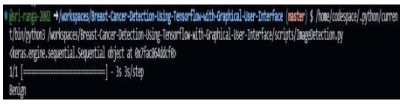

```
@sri-ranga-2002 →/workspaces/Breast-Cancer-Detection-Using-Tensorflow-with-Graphical-User-Interface (master) $ /home/codespace/.python/curren
t/bin/python3 /workspaces/Breast-Cancer-Detection-Using-Tensorflow-with-Graphical-User-Interface/scripts/ImageDetection.py
<keras.engine.sequential.Sequential object at 0x7fac854ddcf0>
1/1 [==============================] - 3s 3s/step
Benign
```

Figure 5: Output screen of predicting benign or malignant tumour

V. CONCLUSION

Breast cancer is a highly fatal disease, causing a significant number of deaths every year. Currently, the clinical management of breast cancer relies on a limited number of accurate prognostic and predictive factors. However, by utilizing a clustering approach with Level Set, it is possible to achieve high accuracy in detecting affected cell shapes and accurately marking the detected contours.

In this proposed system, image segmentation is performed using the Fuzzy-C-means (FCM) clustering algorithm. FCM assigns each data point to multiple clusters with varying degrees of membership based on an objective function. The segmented region is then thoroughly analyzed using a combination of Multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA), and Gray Level Cooccurrence Matrix (GLCM) features. A total of 13 features are extracted from the segmented regions, and their pixel values in the form of a matrix are stored in a database.

Once the features are extracted and the system is trained, the image is classified into three categories: Benign, Malignant, and Normal. This classification is achieved using the K-nearest neighbors (KNN) classifier technique, which heavily relies on the shape of the cancer cells in the image.Suitable morphological operations are performed to compute region properties such as area and Euler number, enabling the system to display the boundary-detected image along with the tumor area. These techniques significantly improve the accuracy in tracking breast cancer cells.

To evaluate the correctness and effectiveness of each algorithm, the system assesses data classification in terms of accuracy, precision, sensitivity, and specificity. The goal of this design is to provide high accuracy and maximum efficiency in predicting and tracking breast cancer. The proposed combination of the Multi-Level Wavelet Conversion strategy, PCA with 13 extracted features, and classification yields an average accuracy of approximately 92%.

As a future improvement, the system could incorporate additional features such as recommending medicines or treatments based on the severity of the patient's condition. This prediction and recommendation system would assist doctors in diagnosing and treating the disease more efficiently.

REFERENCES

[1] I. Christoyianni, E. Dermatas, and G. Kokkinakis, “Fast detection of masses in computer-aided mammography,” *IEEE Signal Processing Magazine*, vol. 17, no. 1, pp. 54–64, 2000.

[2] N. Salim, *Medical Diagnosis Using Neural Network*, Faculty of Information Technology University, 2013, http://www.generation5.org/content/2004/MedicalDiagnosis. asp.

[3] A. Tartar, N. Kilic, and A. Akan, “Classification of pulmonary nodules by using hybrid features,” *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 148363, 11 pages, 2013.

[4] N. Kilic, O. N. Ucan, and O. Osman, “Colonic polyp detection in CT colonography with fuzzy rule based 3D template matching,” *Journal of Medical Systems*, vol. 33, no. 1, pp. 9– 18, 2009.

[5] A. Mert, N. Kilic,, and A. Akan, “Evaluation of bagging ensemble method with time-domain feature extraction for diagnosing of arrhythmia beats,” *Neural Computing and Applications*, vol. 24, no. 2, pp. 317–326, 2014.

[6] R. W. Brause, “Medical analysis and diagnosis by neural networks,” in *Proceedings of the 2nd International Symposium on Medical Data Analysis (ISMDA ’01)*, pp. 1–13, Madrid, Spain, October 2001.

[7] T. S. Subashini, V. Ramalingam, and S. Palanivel, “Breast mass classification based on cytological patterns using RBFNN and SVM,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5284– 5290, 2009.

[8] M. N. Gurcan, H.-P. Chan, B. Sahiner, L. Hadjiiski, N. Petrick, and M. A. Helvie, “Optimal neural network architecture selection: improvement in computerized detection of microcalcifications,” *Academic Radiology*, vol. 9, no. 4, pp. 420– 429, 2002.

[9] A. P. Dhawan, Y. Chitre, C. Bonasso, and K. Wheeler, “Radialbasis-function based classification of mammographic microcalcifications using texture features,” in *Proceedings of the 17th IEEE Engineering in Medicine and Biology Annual Conference*, pp. 535–536, September 1995.

[10] A. T. Azar and S. A. El-Said, “Superior neuro-fuzzy classification systems,” *Neural Computing and Applications*, vol. 23, no. 1, supplement, pp. 55–72, 2012.

[11] M. Jia, C. Zhao, F. Wang, and D. Niu, “A new method for decision on the structure of RBF neural network,” in *Proceedings of the 2006 International Conference on Computational Intelligence and Security*, pp. 147–150, November 2006.

[12] J. K. Sing, S. Thakur, D. K. Basu, M. Nasipuri, and M. Kundu, “High-speed face recognition using self-adaptive radial basis function neural networks,” *Neural Computing & Applications*, vol. 18, no. 8, pp. 979–990, 2009.

[13] R. Huang, L. Law, and Y. Cheung, “An experimental study: on reducing RBF input dimension by ICA and PCA,” in *Proceedings of the 2002 International Conference on Machine Learning and Cybernetics*, vol. 4, pp. 1941–1945, November 2002.

[14] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1989.

[15] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240–3247, 2009.

[16] B. Wang, H. Huang, and X. Wang, “A support vector machine based MSM model for financial short-term volatility forecasting,” *Neural Computing and Applications*, vol. 22, no. 1, pp. 21– 28, 2013.

[17] M. Pontil and A. Verri, “Support vector machines for 3D object recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637–646, 1998.

[18] J. Zhou, G. Su, C. Jiang, Y. Deng, and C. Li, “A face and fingerprint identity authentication system based on multi-route detection,” *Neurocomputing*, vol. 70, no. 4–6, pp. 922–931, 2007.

[19] E. Gumus, N. Kilic, A. Sertbas, and O. N. Ucan, “Evaluation of face recognition techniques using PCA, wavelets and SVM,” *Expert Systems with Applications*, vol. 37, no. 9, pp. 6404–6408, 2010.

[20] S. Kara, A. Guven, and S. Ic¸er, “Classification of macular and¨ optic nerve disease by principal component analysis,” *Computers in Biology and Medicine*, vol. 37, no. 6, pp. 836– 841, 2007.

[21] A. Hyvarinen and E. Oja, “Independent component analysis:¨ algorithms and applications,” *Neural Networks*, vol. 13, no. 45, pp. 411–430, 2000.

[22] M. P. S. Chawla, “A comparative analysis of principal component and independent component techniques for electrocardiograms,” *Neural Computing and Applications*, vol. 18, no. 6, pp. 539–556, 2009.

[23] S. D. Villalba and P. Cunningham, “An evaluation of dimension reduction techniques for one-class classification,” *Artificial Intelligence Review*, vol. 27, no. 4, pp. 273–294, 2007.

[24] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, “Machine learning techniques to diagnose breast cancer from imageprocessed nuclear features of fine needle aspirates,” *Cancer Letters*, vol. 77, no. 2-3, pp. 163–171, 1994.

[25] K. H. Liu, B. Li, Q. Q. Wu, J. Zhang, J. X. Du, and G. Y. Liu, “Microarray data classification based on ensemble independent component selection,” *Computers in Biology and Medicine*, vol. 39, no. 11, pp. 953–960, 2009.

[26] 2013, http://research.ics.tkk.fi/ica/fastica/.

[27] J. Bilski, “The UD RLS algorithm for training feedforward neural networks,” *International Journal of Applied Mathematics and Computer Science*, vol. 15, pp. 115–123, 2005.

[28] N. Sivri, N. Kilic, and O. N. Ucan, “Estimation of stream temperature in Firtina Creek (Rize-Turkiye) using artificial neural network model,” *Journal of Environmental Biology*, vol. 28, no. 1, pp. 67–72, 2007.

[29] O. A. Abdalla, M. H. Zakaria, S. Sulaiman, and W. F. W. Ahmad, “A comparison of feed-forward back-propagation and radial basis artificial neural networks: A Monte Carlo study,” in *Proceedings of the International Symposium in Information Technology (ITSim ’10)*, vol. 2, pp. 994–998, Kuala Lumpur, Malaysia, June 2010.

[30] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, July 1992.

[31] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[32] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Wiley, New York, NY, USA, 1953.

[33] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[34] L. L. Pesce and C. E. Metz, “Reliable and computationally efficient maximum-likelihood estimation of proper binormal ROC curves,” *Academic Radiology*, vol. 14, no. 7, pp. 814– 829, 2007.

[35] J. Hamidzadeh, R. Monsefi, and H. S. Yazdi, “DDC: distancebased decision classifier,” *Neural Computing & Applications*, vol. 21, no. 7, pp. 1697–1707, 2012.

[36] A. M. Krishnan, S. Banerjee, C. Chakraborty, and A. K. Ray, “Statistical analysis of mammographic features and its classification using support vector machine,” *Expert Systems with Applications*, vol. 37, no. 1, pp. 470–478, 2010.

[37] S. C. Bagui, S. Bagui, K. Pal, and N. R. Pal, “Breast cancer detection using rank nearest neighbor classification rules,” *Pattern Recognition*, vol. 36, no. 1, pp. 25–34, 2003.