**Project Report**

1. **INTRODUCTION**

1.1 Project Overview

Breast cancer is one of the main causes of cancer death worldwide. Computer-aided diagnosis systems showed potential for improving the diagnostic accuracy. But early detection and prevention can significantly reduce the chances of death. It is important to detect breast cancer as early as possible. The goal is to classify images into two classifications of malignant and benign. As early diagnostics significantly increases the chances of correct treatment and survival. In this application we are helping the doctors and patients to classify the Type of Tumour for the specific image given with the help of Neural Networks.

1.2 Purpose

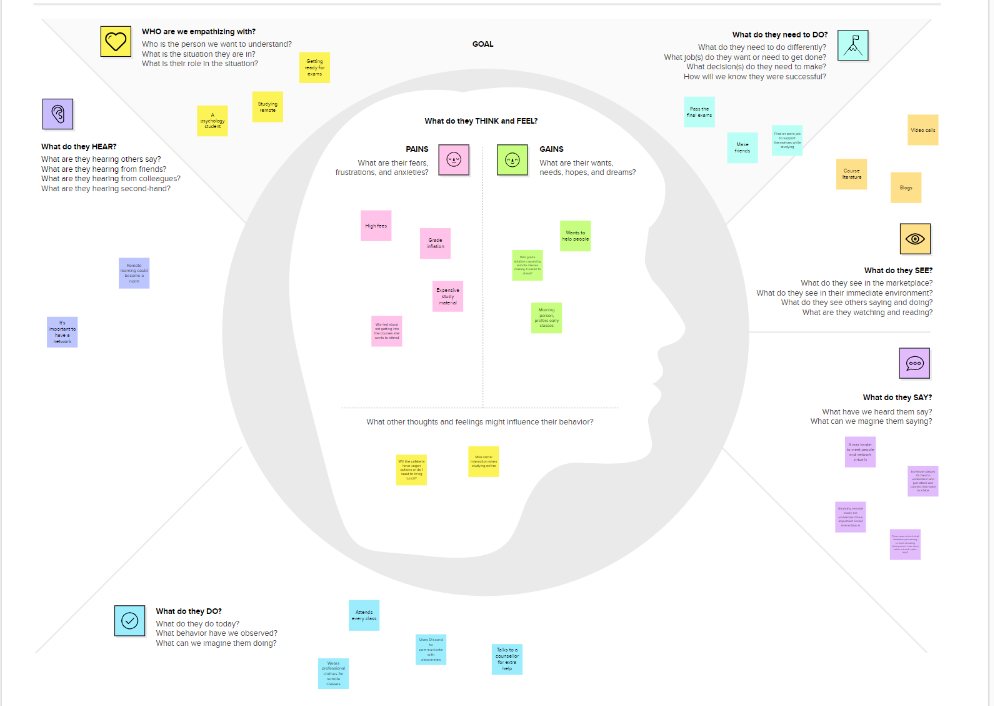
Artificial Intelligence (AI) assisted CAD systems have shown great potential applications in different fields of healthcare industry. However, there have been very few reports of clinical benefits that have come up from the usage of AI in clinical practice, meaning that the potential of AI in healthcare has not yet been realized. This makes it difficult to evaluate the efficacy of AI in this field.

2. **IDEATION & PROPOSED SOLUTION**

2.1 Problem Statement Definition

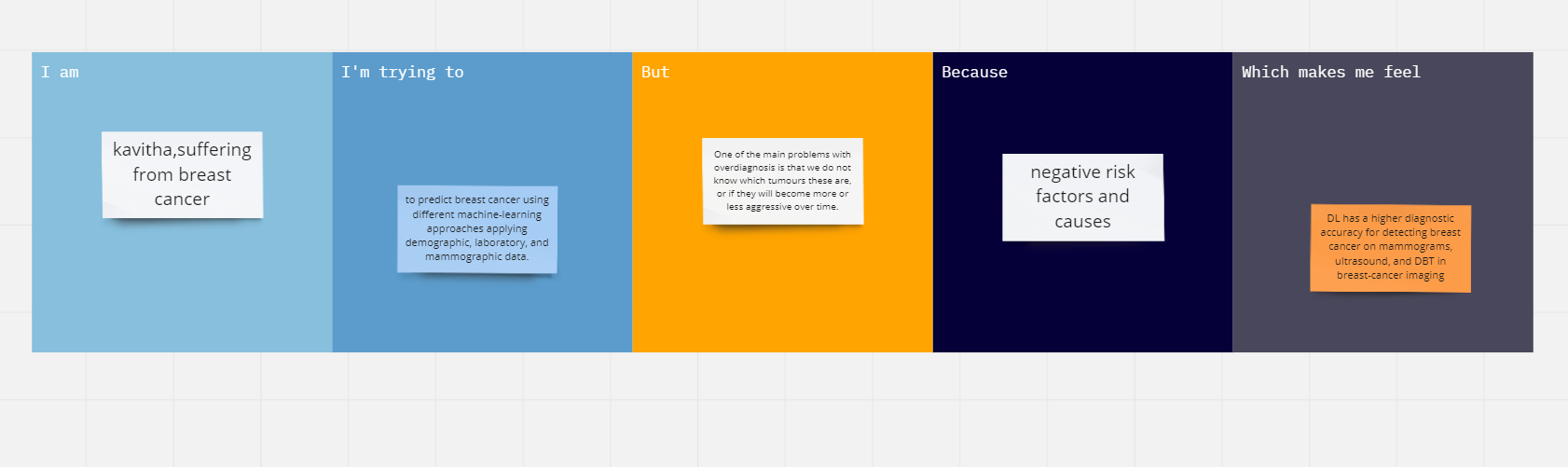
Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists.

2.2 Empathy Map Canvas



An empathy map is a collaborative tool teams can use to gain a deeper insight into their customers. Much like a user persona, an empathy map can represent a group of users, such as a customer segment. The empathy map was originally created by Dave Gray and has gained much popularity within the agile community.

2.3 Ideation & Brainstorming



Ideation is often closely related to the practice of brainstorming, a specific technique that is utilized to generate new ideas. A principal difference between ideation and brainstorming is that ideation is commonly more thought of as being an individual pursuit, while brainstorming is almost always a group activity.

2.4 Proposed Solution

The paper uses microarray breast cancer data for classification of the patients using deep 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. **REQUIREMENT ANALYSIS**

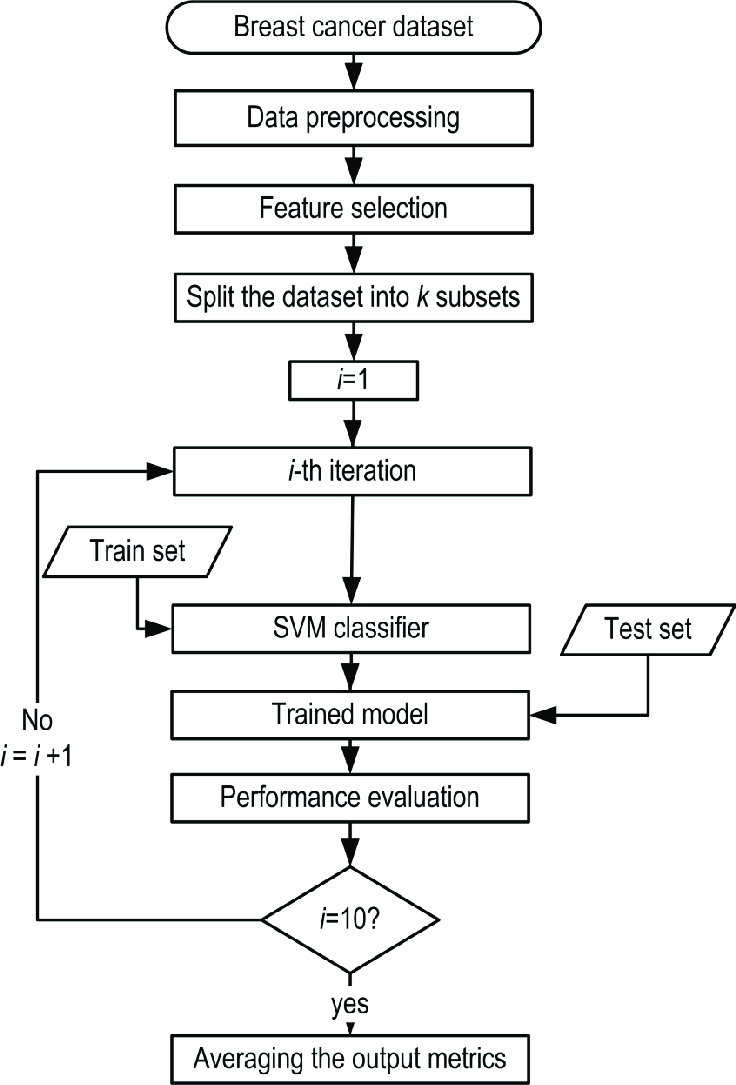
3.1 Functional requirement

In breast cancer it is not necessary that symptoms will be show every time thus helping by taking proper precautions. So, early detection and its proper classification is the only way to lessen the cancer fatality and it is a major task in medical field. The fundamental problems like ineffectiveness in capturing textural information as well as low retrieval performance caused by poor discrimination of capabilities of features. Mammography is being used for early-stage detection diagnosis and screening. Key elements here are processing and analysis for better prognosis results. Using FCM technique, image segmentation is performed here. Further, certain features are extracted through these segmented images and trained. Now, the trained images are being classified by an efficient and accurate classifier. FCM algorithm also known as Fuzzy-C-means, every data point belongs to multiple clusters which vary in degree of membership which is based on objective function. Now the multilevel Discrete Wavelet Transform is used here, to analyze the segmented region. Textual features like pixel values and cancer cells are stored in database in form of matrix. Once the extraction of feature is done and the system is trained, then it classifies the image into either Benign (Initial Stage), Malignant (Harmful) or Normal. For this, it uses KNN algorithm which classifies it depending on the shape of cancer cells [15]. After performing certain morphological operation, the system provides certain region properties like Area, Euler Number etc. and shows the detected boundary along with tumor area. Several techniques like Multi-level Discrete Wavelet Transform, PCA, GLCM are used for extracting textural features. The boundary of tumor is marked properly and shown to doctor. 19 4.1 Clustering based on the FCM algorithm Fuzzy c-means (FCM) is a clustering method that allows each data point to belong to multiple clusters with varying degrees of membership. FCM is based on the minimization of the following objective function where • D is the number of data points. • N is the number of clusters. • m is fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with m > 1. Fuzzy overlap refers to how fuzzy the boundaries between clusters are, that is the number of data points that have significant membership in more than one cluster. • xi is the ith data point. • cj is the center of the jth cluster. • μij is the degree of membership of xi in the jth cluster. For a given data point, xi, the sum of the membership values for all clusters is one. FCM performs the following steps during clustering: 1. Randomly initialize the cluster membership values, μij. 2. Calculate the cluster centers: 3. Update μij according to the following: 20 4. Calculate the objective function, Jm. 5. Repeat steps 2–4 until Jm improves by less than a specified minimum threshold or until after a specified maximum number of iterations.

3.2 Non-Functional requirements

**4. PROJECT DESIGN**

4.1 Data Flow Diagrams



4.2 Solution & Technical Architecture

This project on Breast Cancer Prediction Using Machine Learning is divided into following tasks:

Task 1: Introduction and Import Libraries.

Task 2: Download dataset directly from Kaggle.

Task 3: Load & Explore the Dataset.

Task 4: Perform Label Encoding.

Task 5: Split the data into Independent and Dependent sets and perform Feature Scaling.

Task 6: Building Logistic Regression Classifier.

Task 7: Evaluate performance of the model.

### Developing patch and whole image classifiers on CBIS-DDSM

### Setup and processing of the dataset

The DDSM[37](https://www.nature.com/articles/s41598-019-48995-4#ref-CR37) contains digitized film mammograms in a lossless-JPEG format that is now obsolete. We used a later version of the database called CBIS-DDSM[41](https://www.nature.com/articles/s41598-019-48995-4#ref-CR41) which contains images that are converted into the standard DICOM format. The dataset which consisted of 2478 mammography images from 1249 women was downloaded from the CBIS-DDSM website

4.3 User Stories

**5. CODING & SOLUTIONING (Explain the features added in the project along with code)**

5.1 Feature 1

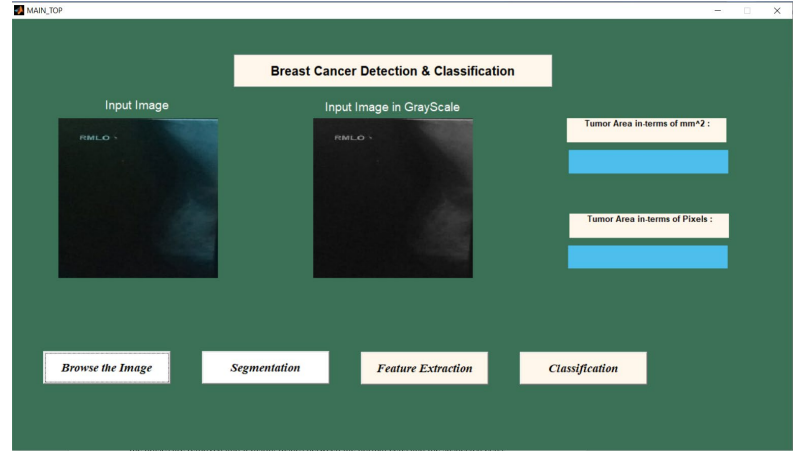
5.2 Feature 2

5.3 Database Schema (if Applicable)

**6. RESULTS**

* 1. Performance Metrics .

This section describes the screens of the “Breast cancer prediction and tracking using machine learning methods”. The figures shown below for each module. The initial stage is to preprocess the image and then convert it to grey image. This helps the system to perform segmentation in future steps. In Figure 5. There is a transformation of the image form the system to grey scale. The system converts any dimension of image to a 2d image. For example, if the image is in 3d it will automatically convert it to the 2D image and if it is 2D the grey scale image will remain as 2D image



**7. ADVANTAGES & DISADVANTAGES**

Breast cancer screening every 2 years over a 20-year period:

* **Reduces the risk of dying from breast cancer**  
  Of 1,000 women who have a mammogram every 2 years for 20 years, 7 deaths are prevented
* **Reduces the risk of having to undergo chemotherapy**  
  Screening often allows for the detection of cancers at an early stage of development. Treatment is then possible without chemotherapy.
* **Allows women to know the health of their breasts**  
  The vast majority of women (nearly 98 %) will not have breast cancer if their mammograms and additional examinations do not reveal cancers.

Breast cancer screening every 2 years for 20 years can lead to:

* **Periods of waiting and anxiety when additional examinations are required**  
  Almost half the women who participate in the screening for 20 years (453 in 1,000) have at least one additional examination. This represents 156 more women than in the 1,000 who do not participate in the screening.
* **Possible overdiagnosis**  
  Of 77 breast cancer diagnoses, 10 would be cases of overdiagnosis.

**Overdiagnosis** is the discovery of a cancer that would never have been detected without screening. It can happen that a woman receives a diagnosis for cancer that would never have had an effect on her health or consequences on her life – like a cancer that develops very slowly or a benign cancer. This could happen to participants in the screening program because a mammogram detects breast cancer in the early stages of development.

However, given that it is still impossible to differentiate between harmless cancers and deadly ones, all cancers are treated. As such, women in the screening program could:

* Receive treatment that would not be necessary
* Suffer the side effects of these treatments
* Have to live with the experience of having been diagnosed with cancer
* Have frequent medical appointments to ensure that the cancer does not return

## Limitations

Breast cancer screening every 2 years for 20 years does not guarantee:

* **That all breast cancers will be detected**  
  Of 1,000 women who have a mammogram every 2 years for 20 years, 77 will be diagnosed with breast cancer. Of these, 21 will be diagnosed with cancer even though their mammography results were normal. Such a situation can occur if:
  + The cancer was not visible on the mammogram
  + The cancer was not yet developed at the time of the mammography
* **That all participants with breast cancer will survive**  
  Of 1,000 women who have a mammogram every 2 years for 20 years, it is estimated that 13 will die of breast cancer.

Also, a mammogram is an X-ray. Like all X-ray examinations, radiations are emitted. According to several studies, the risk of breast cancer due to radiation emitted during a mammogram is very low in women aged 50 to 69 who participate in screening.

**8. CONCLUSION**

Breast Cancer represents one of the diseases that makes highest number of deaths every year. At present, only few accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer. Here, by making use of Clustering with Level Set approach, high accuracy can be achieved in detection of effected cell shapes with exact marking on detected contours. The proposed system helps to enhance the performance of mammogram retrieval by selecting optimal features. The Fuzzy-C-means (FCM) clustering has been used for Image segmentation. Each data point belongs to multiple clusters with varying degrees of membership, and it is based on the objective function. The segmented region is completely analyzed by using the Multi-level Discrete Wavelet Transform, Principal Component Analysis (PCA) along with Gray Level Cooccurrence Matrix (GLCM) features. Totally 13 features are extracted and their pixel values in the form of matrix is stored in database. After the features are extracted & completely trained the system classifies the image into Benign, Malignant and Normal using the KNN classifier technique which mainly depends on the shape of the cancer cells in the image. By performing suitable morphological operations, system computes the suitable region properties such as Area, Euler number etc., and displays the boundary detected image along with the tumor area

**9. FUTURE SCOPE**

. These techniques improve accuracy in tracking the breast cancer cells. To assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity, and specificity. Hence the design is to provide high accuracy and maximum efficiency in prediction and tracking of breast cancer. The combination of Multi-Level Wavelet Conversion strategy associated to PCA with 13 features extracted and then classified gives an average accuracy of nearly 92%. 37 As a future improvement, the system can add more features such as recommendation of medicines/treatments based on the severity of the patient. This prediction and recommendation system can help doctors to diagnose and cure the disease more efficiently

**10. APPENDIX**

Source Code

1. DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. CA Cancer J Clin. learning methods.” 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (2018): 1-3.2011;61(6):409-418. doi:10.3322/caac.20134. 2. Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, ‘‘A review of breast cancer detection in medical images,’’ in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2018, pp. 1–4. 3. Turgut, Siyabend et al. “Microarray breast cancer data classification using machine Varalatchoumy and M. Ravishankar, "Comparative study of four novel approaches developed for early detection of breast cancer and its stages," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 411-416, doi: 10.1109/ICICI.2017.8365384. 4. M. Ravishankar and M. Varalatchoumy, "Four novel approaches for detection of region of interest in mammograms — A comparative study," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 261-265, doi: 10.1109/ISS1.2017.8389410. 5. Ammu P K and Preeja V. Article: Review on Feature Selection Techniques of DNA Microarray Data. International Journal of Computer Applications 61(12):39-44, January 20. 6. Bing Nan Li, Chee Kong Chui, Stephen Chang, S.H. Ong,Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation, Computers in Biology and Medicine, Volume 41, Issue 1, 2011, Pages 1-10, ISSN 0010-4825. 7. Reddy, V. Anji, and Badal Soni. "Breast Cancer Identification and Diagnosis Techniques." Machine Learning for Intelligent Decision Science. Springer, Singapore, 2020. 49-70. 8. C. Yanyun, Q. Jianlin, G. Xiang, C. Jianping, J. Dan and C. Li, "Advances in Research of Fuzzy C-Means Clustering Algorithm," 2011 International Conference on Network Computing and Information Security, Guilin, 2011, pp. 28-31, doi: 10.1109/NCIS.2011.104. 9. V. Pali, S. Goswami and L. P. Bhaiya, "An Extensive Survey on Feature Extraction Techniques for Facial Image Processing," 2014 International Conference on Computational Intelligence and Communication Networks, 2014, pp. 142-148, doi: 10.1109/CICN.2014.43. 10. Jolliffe IT, Cadima J. Principal Component Analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202. 11. Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of Livestock Research. 1. 10.5455/ijlr.20170415115235. 12. Kalist, V. & Packyanathan, Ganesan & Joseph, Maria & B.S, Sathish & Murugesan, R.. (2020). Image Quality Analysis Based on Texture Feature Extraction Using Second-Order Statistical Approach. 10.1007/978-981-15-3172-9\_48. 13. Chattopadhyay, P., Konar, P. Feature Extraction using Wavelet Transform for Multi-class Fault Detection of Induction Motor. J. Inst. Eng. India Ser. B 95, 73–81 (2014). 14. S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693, July 1989, doi: 10.1109/34.192463. 15. S. Singh, D. Srivastava and S. Agarwal, "GLCM and its application in pattern recognition," 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), Dubai, 2017, pp. 20-25, doi: 10.1109/ISCBI.2017.8053537. 16. H. Wang.: Nearest Neighbours without k: A Classification Formalism based on Probability, technical report, Faculty of Informatics, University of Ulster, N.Ireland, UK (2002) 17. G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN Model-Based Approch in Classification,” Lecture Notes in Computer Science, pp 986-996, 2003. 18. Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, ‘‘Risk factors and preventions of breast cancer,’’ Int. J. Biol. Sci., vol. 13, no. 11, p. 1387, 2017. COMPARISION 19. Islam, M.M., Haque, M.R., Iqbal, H. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 1, 290 (2020).

GitHub & Project Video Demo Link

https://github.com/Subeee28