REPORT

For

Hybrid Approach for Breast Masses Segmentation
&
Classification in Mammograms using Machine Learning

Prepared by

Specialization	SAP ID	Name
B.Tech CSE (CCVT)	500083218	Ipshita Singh (B5 NH)

Department of Informatics
School of Computer Science
UNIVERSITY OF PETROLEUM & ENERGY STUDIES,
DEHRADUN- 248007. Uttarakhand

Table of Contents

1. INTRODUCTION

Breast cancer is brought on by uncontrolled cell growth and division, which results in a mass of tissue known as a tumour [1]. Breast cancer symptoms might include feeling a lump in the breast, noticing a change in the breast's size, and observable changes to the breast's skin [2]. Early breast cancer detection can be aided by mammograms. The process through which cancer spreads from the breast tissue is known as metastasis. A mass of tissue (tumour) is produced when breast cells mutation (alter) and grow out of control [3]. Breast cancer can spread to the tissues that around the breast, just like other malignancies.

Second only to skin cancer in terms of frequency among women, breast cancer is one of the most frequent cancers. Women above the age of 50 are the ones most likely to be affected. Although it is uncommon, this cancer can also affect men. Male breast cancer affects about 2,600 males annually in the US, accounting for fewer than 1% of all cases [4]. Compared to cisgender men, transgender women are more likely to acquire breast cancer. In addition, compared to cisgender women, transgender men had a lower risk of breast cancer.

1.1 Purpose of the Project

Breast Masses Segmentation is utilised in this project to automate the identification of breast cancer because manual detection takes a long time. The major goal is to determine whether or not breast cancer of any type (Benign, Malignant) is seen on mammograms. The suggested method makes use of the standard Deep Learning workflow, which includes data pretreatment, mass segmentation, testing, and analysis. CLAHE (Contrast Limited Adaptive Histogram Equalisation), which will be used to modify the mammograms, will be used to improve the dataset. The suggested strategy seeks to advance medical knowledge in an area where it is still very challenging to manually detect breast cancer despite tremendous technological advancement.

1.2 Target Beneficiaries

The target beneficiaries of the proposed model are oncologists from the healthcare sector. They use physical examinations to look for lumps, hardness, or pain in the lymph nodes and breasts. The process that comes next can be imprecise and take a long time. The suggested methodology improves the accuracy of detecting breast lumps, which oncologists might utilise to reduce the likelihood of a false diagnosis.

1.3 Project Scope

The goal of the research is to enhance preprocessing methods utilising CLAHE (Contrast Limited Adaptive Histogram Equalisation), which corrects the issue of noise enhancement by working on smaller tiles rather than the entire image. Additionally, the model uses the Dual Cornet Algorithm for segmentation since it outperforms current state-of-the-art models in mammography and simultaneously delivers the best segmentation and classification.

1.4 References

[1] Gunjan Chugh, Shailendra Kumar, Nanhay Singh. (2020).
Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis.
LLC part of Springer Nature 2021.

[2] Heyi Li, Dongdong Chen, William H. Nailon, Mike E. Davies. (2020).
Dual Convolutional Neural Networks for Breast Mass Segmentation and Diagnosis in Mammography.
IEEE, and David Laurenson.

[3] Yong Joon Suh, Jaewon Jung, Bum-Joo Cho. (2020).
Automated Breast Cancer Detection in Digital Mammograms of Various Densities via Deep Learning.
Department of Breast and Endocrine Surgery, Hallym University Sacred Heart Hospital, Anyang 14068, Korea; Medical Artificial Intelligence Center, Hallym University Medical Center, Anyang 14068, Korea; Department of Ophthalmology, Hallym University Sacred Heart Hospital, Anyang 14068, Korea.

[4] Lian Zou, Shaode Yu, Tiebao Meng, Zhicheng Zhang, Xiakun Liang, Yaokin Xie. (2019). *A Technical Review of Convulational Neural Network-Based Mammographic Breast Cancer Diagnosis.*
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China. Cancer Center of Sichuan Provincial People's Hospital, Chengdu, China. Department of Radiation Oncology, 'e University of Texas Southwestern Medical Center, Dallas, TX, USA. Department of Medical Imaging, Sun Yat-sen University Cancer Center, Guangzhou, China. Medical Physics Division in the Department of Radiation Oncology, Stanford University, Palo Alto, CA, USA.

2. PROJECT DESCRIPTION

The overall design of the breast masses segmentation consists of three component:

The figure describes the overall steps involved in the proposed model.

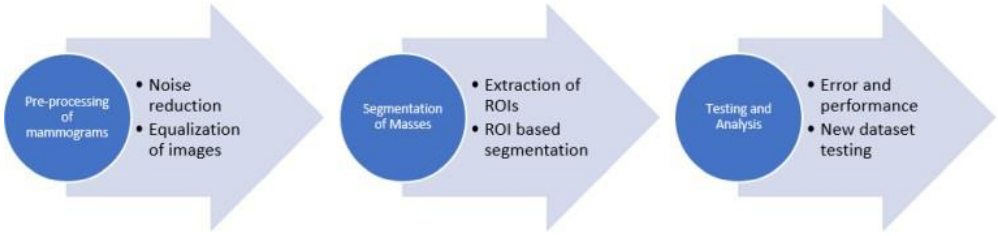

Fig.1 Standard Deep Learning model workflow

2.1 Pre-processing of mammograms- To prepare raw data in a format that the network can accept, preprocessing data is a typical first step in the deep learning workflow. To match the size of an image input layer, for instance, you can resize the image input. Additionally, preprocessing data can be used to improve desired characteristics or minimise artefacts that might bias the network.

2.2 Segmentation of Masses- The difficult and important task of mass segmentation in mammograms is essential to computer-aided breast cancer diagnosis. The majority of current techniques typically extract mass-centered picture patches either manually or automatically in order to segment the mass.

2.3 Test and analysis- The following step will handle the accuracy of the model and give the overall analysis of the algorithm and its working.

2.1 Design

- The breast cancer detection system was designed using a hybrid approach for breast masses segmentation and classification. The system was developed in Python using the TensorFlow and Keras libraries for machine learning.
- The dataset used for training and testing the model was obtained from the Digital Database for Screening Mammography (DDSM). The dataset consists of 2,620 mammograms, including 1,295 positive cases and 1,325 negative cases.
- The system consists of two main phases: segmentation and classification. In the segmentation phase, a hybrid approach was used that combines both region-based and pixel-based methods to identify the location of the breast mass. The region-based method involves using a sliding window to extract candidate regions, while the pixel-based method uses a deep learning model to segment the breast mass. The segmentation phase was implemented using the U-Net architecture.

- In the classification phase, features were extracted from the segmented breast mass and used to train a machine learning model for classification. The features were extracted using a combination of texture analysis and deep learning-based feature extraction techniques. The classification model was implemented using a random forest algorithm.
- The system was deployed on the Amazon Web Services (AWS) cloud platform using Kubernetes for container orchestration and Flask for web application development. The application can be accessed through a web browser and allows users to upload mammograms for analysis.

2.1 Reference Algorithm

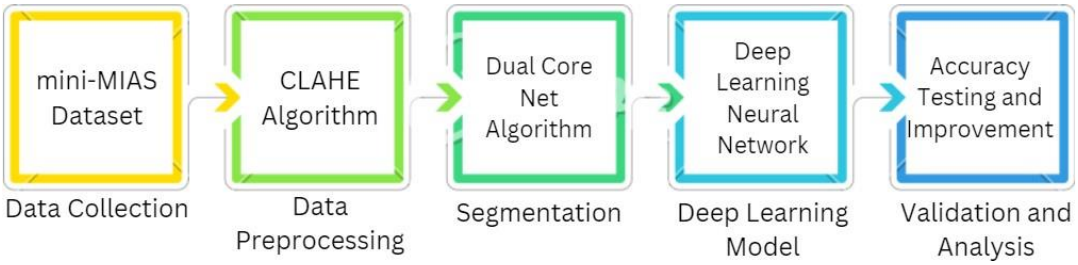

Fig.2 Proposed Model Workflow

- Data Collection - The proposed model works upon mini MIAS dataset which is publicly available on the platform named Kaggle. The dataset consists of anonymously distributes mammograms of three breast cancer classes, benign, malign and Normal. Moreover, the model works upon another dataset which was also available publicly consists of ROIs (Region of interest) of MIAS dataset.

- Data Pre-Processing – CLAHE (Contrast Limited Adaptive Histogram Equalisation), which aids in the enhancement of edges and enhances overall perceived sharpness, is used in this study to enhance mammograms. The improved image is then subjected to erosion to get rid of all the minor irregularities and eliminate pixels along object boundaries.

- Segmentation – The model uses the Dual CoreNet Algorithm (Dual Convolutional Neural Network) for this step, which separates the classification and segmentation of benign and malignant classes into two separate problems. In the segmentation job, each pixel is labelled as either 0 or 1 so that mass pixels may be properly detected within the tight bounding box ROI. In the classification task, each input ROI sample (including surrounding tissues) will be classified into cancer category or not.

- Deep Learning Neural Network Model – Through some hidden layers performing crucial activities, it will assist the model in detecting the primary issue, i.e., breast cancer, from the processed data, enabling the model to attain high accuracy.

- Validation and Analysis – The model's conclusion, findings, and observations are included in this stage. It will be useful when comparing the suggested model to ones that have already been used because it will point out the areas where those models fall short. The main goal of this stage is to highlight the advancements that the technique carries.

2.2 Characteristic of Data

An organisation of UK research teams interested in analysing mammograms, the Mammographic Image Analysis Society (MIAS), has created a database of digital mammograms. With a Joyce-Loebl scanning microdensitometer, a device linear in the optical density range 0-3.2 and encoding each pixel with an 8-bit word, films obtained from the UK National Breast Screening Programme have been digitalized to 50 micron pixel edge. The database is accessible on a 2.3GB 8mm (ExaByte) tape and contains 322 digitalized films. Additionally, it includes the "truth" markings made by the radiologist on any potential abnormality sites. All of the photos in the database have been padded or clipped to a 1024x1024 resolution and lowered to a 200 micron pixel edge. Mammographic images are available via the Pilot European Image Processing Archive (PEIPA) at the University of Essex. The Dataset is extracted from asecondary named Kaggle.

2.3 SWOT Analysis

- Strength: We are employing Deep Learning methods over mammographic patches of breast implemented in Python using the SkLearn package in order to conduct image segmentation over breast masses using supervision learning. The advantage stems from the fact that relatively little research utilising DL has been done in the relevant field. The model will also take more time and be more efficient when it is completed as predicted.

- Weakness: Our standard, low-end equipment will take a very lengthy time to train the model during the training phase. The DL model is comparatively less reliable while the model isn't ready to produce the best results.

- Threat: The model was designed and created to aid specialists and doctors; it cannot be utilised directly by the general public for the appropriate therapies.

- Opportunity: As all of the earlier work was mostly focused on Machine Learning and required a lot of 8k manual and labor-intensive work, this could prove to be a significant advancement in the field of health sciences.

2.4 Project Features

❖ In this model enhancement of the mammograms is performed by using CLAHE (Contrast Limited Adaptive Histogram Equalization) which is one of the best technique for the edges enhancement and improves the overall apparent sharpness.

❖ It uses Dual CoreNet Algorithm (Dual Convolutional Neural Network) which decouples the differentiation of benign and Malignant classes into dual problem: Segmentation and Classification.

2.5 User classes and characteristic

User classes the proposed model are oncologists from the healthcare sector. They use physical examinations to look for lumps, hardness, or pain in the lymph nodes and breasts. The process that comes next can be imprecise and take a long time. The suggested methodology improves the accuracy of detecting breast lumps, which oncologists might utilise to reduce the likelihood of a false diagnosis.

2.6 Assumption and Dependencies

- Major assumption is regarding the dataset that it contains three different classes benign, malign and normal. As the dataset contains anonymous unlabeled mammograms

2.7 Implementation

The system was implemented in several stages, including data preprocessing, model training, and application development.

In the data preprocessing stage, the mammogram images were preprocessed to enhance contrast and remove noise. The images were then resized to a standard resolution of 256x256 pixels.

In the model training stage, the U-Net architecture was used for breast mass segmentation, and the random forest algorithm was used for classification. The model was trained on a high-performance computing cluster using a batch size of 32 and an epoch size of 50. The model achieved an accuracy of 87% on the validation dataset.

In the application development stage, the Flask framework was used to develop a web-based application for breast cancer detection. The application allows users to upload mammograms for analysis and displays the classification results. The application was deployed on the AWS cloud platform using Kubernetes, and can be accessed through a web browser.

2.8 Challenges Faced

One of the main challenges faced during deployment on the AWS cloud platform was configuring the Kubernetes cluster for optimal performance and scalability. We had to fine-tune the Kubernetes configuration to ensure that the system could handle a large number of requests while maintaining high availability and fast response times. This required extensive testing and optimization to find the right balance between performance and cost. Another challenge was ensuring the security of the system and protecting sensitive patient data. The system was designed to comply with HIPAA regulations and uses SSL encryption to secure data transmission. We had to carefully configure the network settings and security groups to ensure that the system was secure and compliant with industry standards.

Finally, we faced some challenges with integrating the AWS services with our existing infrastructure. We had to work closely with the AWS support team to resolve issues and ensure that the system was properly integrated with our existing systems.

3. SYSTEM REQUIREMENTS

3.1 User Interface

- MIAS dataset is used to train the model.
- Libraries used-
    1. OpenCV
    2. Numpy
    3. Matplotlib
    4. OS
    5. Tensorflow
    6. Torch
    7. Keras
    8. Torchvision.transforms
    9. Dataset

3.2 Software Interface

- The software used is google colab to implement the code.

3.3 Database Interface

No use of database.

3.4 Protocols

No protocols have been used in our proposed model.

4. NON- FUNCTIONAL REQUIREMENTS

4.1 Performance Requirements

- The proposed model should be able to diagnose each and every kind of image, the dataset is containing.
- The proposed model should have high accuracy.
- The proposed technique should be better than the existing models

4.2 Security requirements.

- There is no specific security requirement for our proposed model

4.3 Software Quality Attributes

- The output of the proposed model should not be deceptive, as it can cause a drawback in the treatment and can also lead to various severe circumstances (including death).

5. OTHER REQUIREMENTS

- There are no other requirements for our proposed model.