CANCER DETECTION USING MACHINE LEARNING.

1. Introduction.

Medical imaging analysis plays a huge and important role in detecting certain unusual changes in the body such as breast cancer, blood cancer , brain tumour and many more. According to the statistics done by world health organisation (W.H.O), approximately 18 million new cases of cancer were reported as of 2019 which resulted in 10 million casualties.

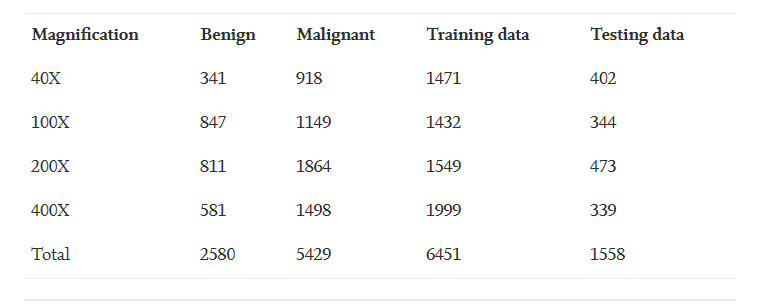
This enough proves how catastrophic the disease (cancer) can become and calls for immediate attention to put an end to this menace.

Cancer is a combination of infections which involves abnormal cell growth with the potential to invade or spread to other parts of the body. The disease itself is curable if detected early. The purpose of this project is to incorporate machine learning in detecting early signs of cancer to be able to eliminate it before it turns out to be costly and catastrophic. Research has proved early detection can increase the survival rate to 95% signifying the importance of this project. In this problem we will be focusing on one of the dangerous type of cancer called breast cancer.

* 1. Dataset

A good dataset is an important prerequisite for a novel and robust BC detection method to be developed.  due to the lack of sample availability as well as the confidentiality of patients information, of the patients, obtaining a dataset is extremely difficult. The data used in this is taken from Cancer Hospital & Institute of Research. The dataset consists of 10000 image samples from more than 800 patients with different magnifications.To expand the dataset, data augmentation is used to reduce the issue of small data size. In a bottom-up technique, natural photos are analysed. By a top-down method, most of the diagnostic images are resolved. Techniques of augmentation applied to images would not be appropriate for medical images. Because of this, it is complex to a data augmentation strategy on the dataset.

Table below summarizes the dataset.



Magnification modifies size of sample images that would increase the program efficiency. Since there are multiple tissues in the histology images, the study is very complicated at low magnification. To make a differential diagnosis, it is difficult for a machine to learn distinct characteristics from the said images with varying levels of magnification. Several stages of preparation are taken in tandem with previous experience of the magnification factors in the proposed device implementation.

1. Problem analysis

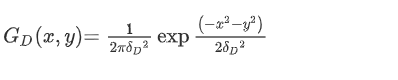
Pre-processing

The foremost task is pre-processing the image dataset. Consider, D is the picture database for breast cytology,

1.PNG

where n is defined as the count of images and D as the vector function.

Initially, the images are processed for noise reduction with an efficient Gaussian filtering technique. Channels are used to extend splendor and complexity in order to turn it into a picture of surfaces, tones and embellishments. It is characterized by the Gaussian channel,



where (x, y) is defined as pixel of the input images.

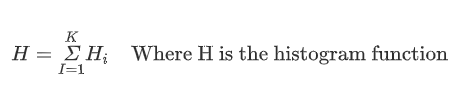
Feature extraction

From the pre-processed images, the entropy, geometrical and textural characteristics are extracted. Entropy estimate of the hazardness used to characterize texture of the dataset image. Shape characteristics is an important feature in distinguishing the features of normal and malignant cells. Each image is divided to 'm’ sub squares and then it is quantized for textual characteristics.

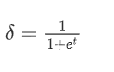
**Histo-sigmoid fuzzy clustering**

Histogram is defined as the accurate representation of numerical data distribution while Fuzzy clustering is a technique that makes one bit of data into two additional clusters.The Histo-Sigmoid Fuzzy Clustering can be further developed and checked using the various mathematical formulas in the following steps.

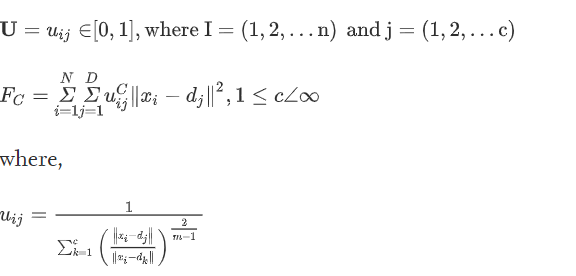
Step-1: You can measure the image histogram value by using Eq.



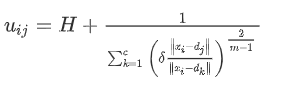
Step-2: It is possible to measure the Sigmoid function by using Eq



Step-3: In the fuzzy clustering algorithm, substitute the histo-sigmoid values.



Step-4: By applying the fuzzy clustering algorithm in Eq, measure the histo-sigmoid function



Algorithm.

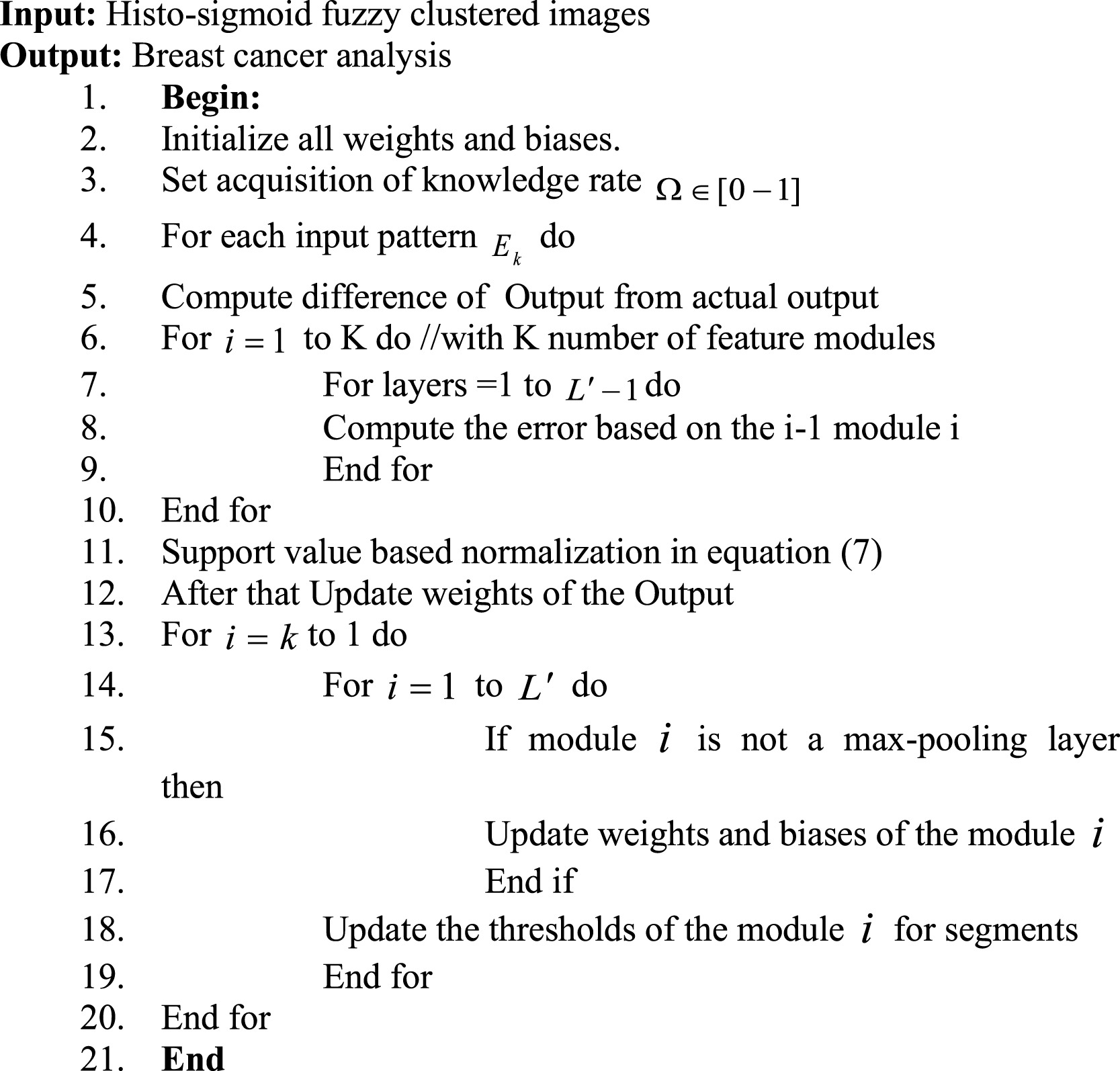


Table 4. Performance analysis of most popular BC detection methods.

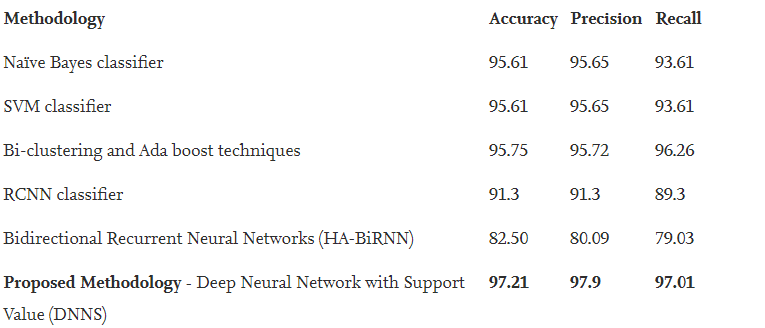
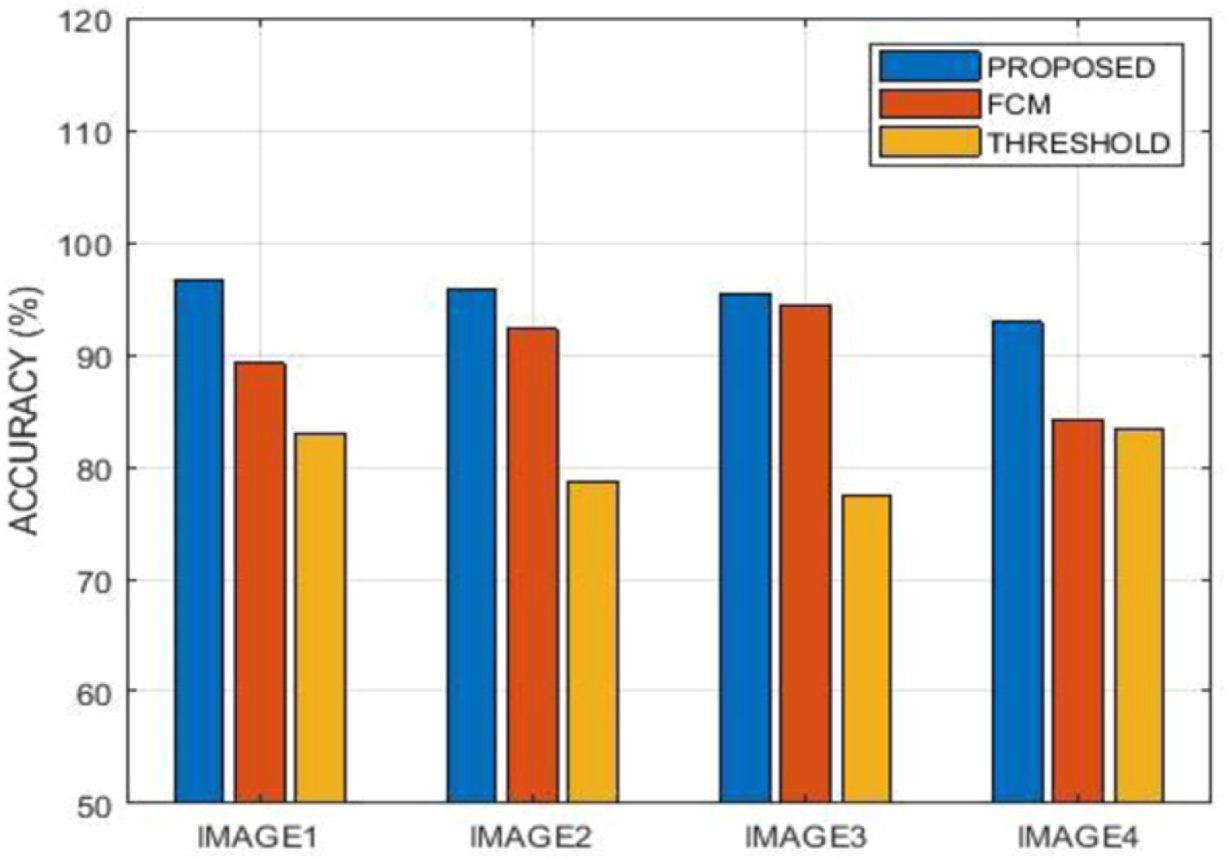


Fig. 2. Comparison of results (proposed method in (c) with the existing methods (a) and (b)).



Functional requirements.

1. The user must have the tissue images data to detect cancer for the program to work.

Non functional requirements.

1. Usability

It should be easy to use the system. The user can use the summarized text to enter if necessary, press just one button. Since time-saving is one of the software's features.

The framework should also be user-friendly for administrators, as everyone can be an administrator.

1. Reliability.

Machine learning, function engineering and deep engineering is used to build this program techniques for learning. So, there is no certain reliable percentage that is observable in this phase.

User-supplied information would also be used to compare performance and calculate reliability. With the latest machine learning techniques, data collected by users should be appropriate for reliability if appropriate knowledge is collected.

The maintenance cycle should not be a matter because it is always the accurate version. If administrators want to update it, it willIt takes a long time to upload and upgrade the server runtime. It is possible to meet the consumers and use the software at any time, but maintenance should not be a major problem.

1. Performance

There should be as little calculation time and response time as possible, since one of the software 's functionality saves time. The entire duration of data analysis and detection review should not be longer over 1 minute seconds for 10 mb of data.

Design.

Architecture overview.

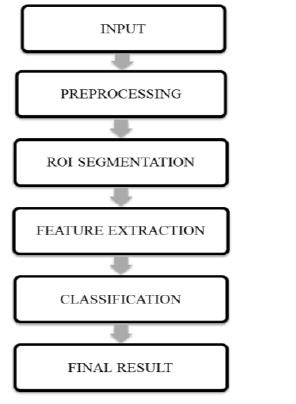


Fig: the figure above demonstrate the system architecture for breast detection system.

1. Input

The input data comprises of 1000 tissues image samples taken from more than 800 breast cancer patients with different magnifications levels.

1. Pre-processing

pre-processing strategies ensures accurate, effective, and correct analysis . This method involves resizing the dataset images and converting them into a dark-scale image over the data image and using channels. Cleaning of information alludes to strategies to discover, expel, and supplant awful or incomplete information. Sensitive knowledge patterns can be distinguished by identifying outrageous and abrupt shifts in the nearby environment. Smoothing and detruding are ways of expelling clamour and direct patterns from data, whereas the limits of knowledge are modified by scaling. Gathering and binning strategies are processes that discern ties between the factors of information.

1. Roi segmentation.

It is possible to describe the system of parcelling the image into fragments as image division. For the comparable characteristics, the division is carried out. This  proposed methodology updates the k-mean classification calculation by displaying a regurgitated division plot that examines the centroid of each set in the portion and re-fragments the information depending on the closest centroid just at end of its comparative characteristics. This technique helps to extract essential image attributes, in terms of which details are seen effectively.

1. Feature extraction.

When examining complex details one of the serious conflict that arises is the huge number of variables involved. To solve this problem feature extraction is required. For this project we use the entropy, geometrical and textural characteristics of the processed images. Entropy is the hazardness prediction required to characterize the texture of the input data image. Shape properties play an essential role in displaying the difference between the features of normal and malignant cells. Each image is divided into 'n' sub squares and quantized for textual properties.

Project management.

This project was completely written using python 3 programming language . Python is a high level object oriented programming language which is used and is popular when it comes to artificial intelligence such as machine learning. It has a powerful network of libraries which are essential in carrying out machine learning. This project used python together the powerful built libraries namely:

1. Pandas

Pandas is a python module that is used in machine learning for data cleaning and analysis and thus for this project it was used to clean and analyse the breast cancer dataset.

1. Sklearn

This is a powerful and efficient library in python that is used in machine learning for classification, regression and clustering.

It was used in this project for classification using its inbuilt algorithm such as k-neighbours/

1. Numpy

Numpy is a library in python majorly used in machine learning for manipulation of arithmetic data.

Development platform.

This project was developed in BigMl. Bigml is a product that provides a collection of machine learning tools for classification and regression. Some of the algorithm found here are : random forest, decision trees , k neighbours. It also equipped with the capabilities of conducting unsupervised learning algorithms.

References.

1. Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol, CA: O'Reilly Media, Inc.
2. Atlas L, Cole R, Connor J, et al. Performance comparisons between backpropagation networks and classification trees on three real-world applications. Advances in Neural Inf. Process. Systems. 1990;2:622–629. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Advances+in+Neural+Inf.+Process.+Systems&title=Performance+comparisons+between+backpropagation+networks+and+classification+trees+on+three+real-world+applications&author=L+Atlas&author=R+Cole&author=J+Connor&volume=2&publication_year=1990&pages=622-629&)]
3. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst. 2003;95:470–8. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/12644540)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Natl+Cancer+Inst&title=Variations+in+lung+cancer+risk+among+smokers&author=PB+Bach&author=MW+Kattan&author=MD+Thornquist&volume=95&publication_year=2003&pages=470-8&pmid=12644540&)]