Cancer Detection using Machine Learning

Pushkar Sathe, Moiz Bombay, Gayathri Sudalai Mani, Dilna Kalathil, Avadhut Phadtare

[1]*Pushkar Sathe, Professor, Dept. of EXTC Engineering, SIES Graduate School of Technology, Maharashtra, India*
[2]*Moiz Bombay, Student, Dept. of EXTC Engineering, SIES Graduate School of Technology, Maharashtra, India*
[3]*Gayathri Sudalai Mani, Student, Dept. of EXTC Engineering, SIES Graduate school of Technology, Maharashtra, India*
[4]*Dilna Kalathil, Student, Dept. of EXTC Engineering, SIES Graduate school of Technology, Maharashtra, India*
[5]*Avadhut Phadtare, Student, Dept. of EXTC Engineering, SIES Graduate school of Technology, Maharashtra, India*

Abstract - *In this day of modern science and age where scientific and technological accomplishments are touching new heights with every second that is passing, the main step in cancer detection is how to classify tumours into malignant or benign which is a challenging task. Machine learning techniques can enormously improve the accuracy of diagnosis. We aim to classify tumour into malignant or benign tumour using different features from several cell images. Machine learning uses the computer data to learn and then use this data to learn a particular pattern or trend in the data. The increasing cancer rate all over the world in today's date is alarming and there is an increased need for efficient cancer detecting techniques. This is possible using Machine learning. This technique provides early detection of tumour which eventually helps in early diagnosis which plays an important role in the treatment of tumour patients. According to global statistics breast cancer is a significant public health problem in today's society because of its widespread increase in cancer rates. Because of its unique advantages in critical features detection from complex data sets, machine learning (ML) is widely recognized as the methodology of choice in cancer pattern classification. This project aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid in accurate cancer detection.*

Key Words: **Machine Learning, Neural Network, Support vector machine, Data-set.**

1. INTRODUCTION

Cancer cells are present in a tumour and may spread to other parts of the body. It is the leading cause of death and breast cancer is one of the most common and life-threatening malignant tumours among women in the world. Lymphoma: Cancer that begins in the cells of the immune system. Cancer cells are present in a tumour and may spread to other parts of the body. This affects both the physical and mental health of women. Breast cancer incidence has been increasing ever since the 1970s all over the world. As in any disease early detection and starting the right treatment is always key. The survival rate increases significantly with early detection of cancer. Early detection is beneficial for both, impaired as they do not need to go under expensive treatment of benign tumors and for the doctors as early detection will help them in giving proper treatments to their patients and necessary care would be taken of them. For such societal benefits cancer detection is a massive requirement. Detection of cancer is done using various techniques such as image processing, deep learning, artificial intelligence etc. and uses various parameters as data such as text data, mammograms, cell images, cell radius. Machine learning is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being programmed in a detailed manner. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. This process of learning begins with the observation of data in order to establish a certain pattern in the data provided and to make better and more reliable decisions in the future. The main aim is to allow the systems to learn automatically and provide better results without human intervention and alter the actions accordingly. Machine learning plays a major role in the analysis of huge amounts of data. Machine learning has a wide range of applications in today's world. Computer systems use machine learning to perform a particular task based on trends and specific patterns without relying on a detailed instruction set. It focuses on making predictions and thus producing an output based on that prediction. It is safe to say that machine learning can be used in various fields to achieve efficient and faster results with the help of available data. Use of these techniques for detection of cancer rules out the human errors and helps the doctor to get to a result much faster. In this project we use images of cancerous cell and feed them to the module for training and testing in order to classify them into malignant and benign cells and aim to achieve best results in this process.We have selected the Breast Histopathology Data-set for our project .This data-set consist of classified images of cancerous and non cancerous cells. Hence it is appropriate for the training and testing part of the project which classifies the image into benign and malignant[7] i.e positive and negative images.The fig1 shown below.

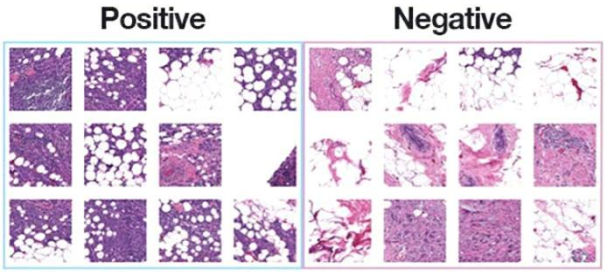

Fig -1: Positive and negative cell images

2. RELATED WORK

Constant studies have been in the field of computer aided methods that are useful in the medical field. As machine learning is new technology it gives advancement in the medical field. By reviewing the past publications and researches related to the study, the researcher will have an idea of how such a study has been done in the past. In this way, this research may be able to reflect, compare itself, learn from setbacks and produce a stronger and more efficient study. They developed a Deep Learning Model utilizing a restricted Boltzmann machine that mainly used back propagation algorithms for classification of histography images i.e they proposed an automatic BC image classifier framework which has been constructed using state-of-the art deep neural network techniques. Instead of using raw images they utilized Tamura features, as they provide textural information. As a deep-learning tool they implemented an unsupervised restricted Boltzmann machine which contains four layers and is guided by a supervised back propagation technique. For the back-propagation, scaled conjugate gradient techniques have been utilized.[1] They used K-nearest neighbor(KNN) and Support vector machine(SVM) with 150 images database. This gives us idea about that the proposed classifier that can improve the performance of recognition. SVM is a binary classifier; it is convenient for classification/recognition in high dimensional space and consequently suitable for image classification and object recognition. KNN is a method for classifying objects based on closest training examples in the feature vector. [2] It presented the various machine learning techniques for determining the breast cancer from using 3D images and SVM. It tells us about Deep Learning for Pixel Level Image Fusion using CSR Technique and automatic Document Meta-data Extraction Using Support Vector Machines. [3] They used DWT tool for image filtering and Back-Propagation Neural Network(BPNN) for processing .This paper teaches us the technique that applies BPNN for developing a novel, and robust digital image watermarking techniques in the DWT domain. The aim of this work is to develop an image watermarking algorithm, which has to satisfy two of its important Requirements, transparency and robustness. [4] Other paper proposes an automated method with a principled work-flow for diagnosing breast cancer. The data used in this research work is the Wisconsin Diagnostic Breast Cancer Data-set (WDBC).The contribution of this research is to show that machine learning approaches which include Support Vector Machine (SVM), Artificial Neural Networks (ANN) and Naive Bayes (NB) can solve pattern classification problems effectively. [5] The recent paper proposes a novel technique for breast cancer segmentation and detection from a digital mammogram. The method has been completed with the help of morphological operations and Artificial Neural Networks. [6]

3. FLOW DIAGRAM

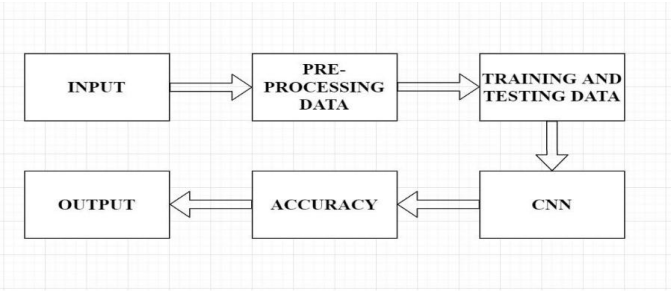

Fig -2: Flow diagram of proposed work

4. EXPERIMENT AND METHODOLOGY

4.1 Data-sets

A malignant tumour that has developed from cells of the breast is breast cancer. Although some of the scientifically known risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity) that increase a woman's chance of developing breast cancer, the exact causes or some of these risk factors that causes cells to become cancerous are not exactly known by them. Research is under way to learn more and scientists are making great progress in understanding how certain changes in DNA can cause normal breast cells to become cancerous.

This part of the project includes the research about the available dataset that can be used for the project. The initial dataset that we planned on working had certain drawbacks. CBIS-DDSM dataset: Neither the analytics nor the experimental validation of the data were sufficient to move forward with the project. There was no exact data regarding the classification of tumours into benign and malignant incase of scan images of patients. One the most important demerit of this dataset was the improper organization of data and huge redundant records which can cause glitches in the training process.

Hence it was concluded that evaluation based on detection rate and accuracy levels with this data-set is not appropriate. Breast cancer Wisconsin data-set: This data-set was a text type data-set which had a classified set of data into malignant and benign tumours based on the cell features threshold values for various features like cell radius, perimeter, area, symmetry and other such

characters. We didn't go for this data-set since we planned on a better version of cancer detection model which focused on images

4.2 Pre-processing data

Data preprocessing is a data mining technique that is used for filter data in a usable format. Because the real-world data-set is almost available in different formats. It's not available as per our requirement so it must fit the data-set in an understandable format. Data pre processing is a proven method of resolving such issues. Data pre-processing converts the data-set into usable format for pre-processing. We have used a standardization method to pre process the data-set.

Encoder Method: Label Encoder is an efficient tool for encoding the levels of the categorical features into numeric values. Label Encoder encode labels with value between 0 and classes -1. All our categorical features are encoded. In this paper we have classified malignant and Benign in diagnosis with 0 and 1. After encoding the dataset we have applied a neural network on these dataset and achieved accuracy.

4.3 Normalizer Method

Normalization is a better technique to use when distribution of the data is unknown.. Data normalization is the process of rescaling one and more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. After applying the Label Encoder method all dataset is converted into numeric dataset. Because Label Encoder encodes all string label dataset into numeric dataset. Normalization process is known to work on numeric dataset. So after normalized data we have applied a neural network on that dataset and achieved accuracy. But again accuracy is not yet adequate. So we have applied the next process for pre-processing.

4.4. Convolutional Neural Network

Convolutional Neural Network is a type of neural network technique which has proven to be specifically efficient with image recognition and classification. The input is a tensor with shape (number of images) x (image width) x (image height) x (image depth) in the programming CNN. Then the image becomes abstracted to a feature map, with shape (number of images) x (feature map width) x (feature map height) x (feature map channels)after passing through a convolutional layer. The following attributes are expected in a convolutional layer within a neural network :

- Width and height (hyper-parameters) defined by convolutional kernels.
- The number of input channels and output channels (hyper-parameter).
- The number channels (depth) of the input feature map must be equal to depth of the Convolution filter (the input channels)

The convolution of input and passing its result to the next layer is done by Convolutional layers. This is similar to the response of a neuron in the visual cortex to a specific stimulus. Each and every convolutional neuron processes data only for its receptive field. It is not practical to apply this architecture to images, even though fully connected feed forward neural networks can be used to learn features as well as classify data. Due to the very large input sizes related with images, a very high number of neurons would be required, even in a simplistic architecture which is not very deep, where each pixel is a relevant variable. For example,in the second layer, a fully connected layer for an image of size 100 x 100 has 10,000 weights for each neuron. A way out to this issue is provided by the convolution operation by decreasing the number of free parameters, allowing the network to be even deeper with keeping the parameters few. Taking an example, irrespective of image size, tiling regions of size 5 x 5, each with the same shared weights, requires only 25 learnable parameters. If went by this, back propagation resolves the exploding or vanishing gradients difficulty in training traditional multi-layer neural networks with many layers.

ReLU is the abbreviation of rectified linear unit, which applies the non-saturating activation functionf(x)=max(0,x). It successfully removes negative values from an activation map by setting them to zero. The nonlinear properties of the overall network and the decision function without affecting the receptive fields of the convolution layer are increased by it. Other functions are also used to extend nonlinearity, for instance the saturating hyperbolic tangent f(x)=tanh(x), f(x)=|tanh(x)| and the sigmoid function sigma(x)=(1+e^-x)^-1 . Since it trains the neural network several times faster without a large kr significant penalty to generalization accuracy, ReLU is preferred a lot more than the other functions.

4.5. Graphical User Interface

The final part of the project includes a graphic user interface(GUI) which will act as a platform to feed fresh cell images for the module to detect whether the given image is cancerous or non cancerous. For this we use Flask. Flask is a web application framework which is written in python. It means by using flask,with the help of the tools, libraries and technologies provided by it , a complete web application can be built. This specific web application can be anything from wiki, a web page, a blog to something as big as a commercial website or a web-based major application. Flask also supports extensions that can add application details and characteristics as if they were implemented in Flask . It has proven to be a significant tool for learning web development basics and best techniques along with the main elements of a web

framework that are regular to almost all frameworks. Using all these components a GUI is created which will help in the classification of a new image.

Fig -3: GUI

5. MODEL IMPLEMENTATION

Neural Network algorithm has been applied on the dataset in this stage. Neural network here works like a human biological method. In which we have to provide the input and get the output. But in these two layers, some hidden layers are present and it becomes mandatory that some additional process must be added before calculating the final output. These added units of bias, some of additional hidden layers, calculate the activation function and the final output is generated. The following parameters for calculating and training the model have been used in this model.

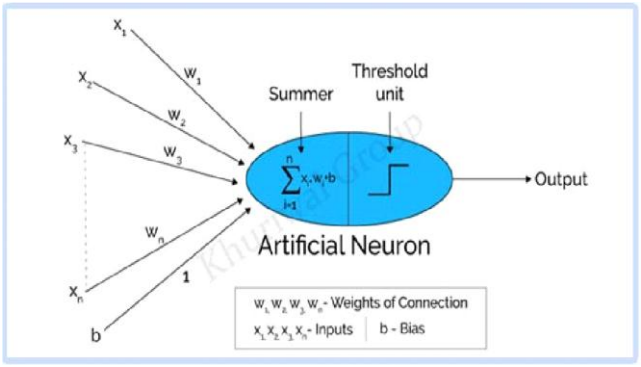

Fig -4: Neural Network Functionality

Fig shows the basic functionality of the neural network in which after the fixed position the neurons are fired that is called threshold condition.

Table -1: Parameters in CCN model

Parameters used in CCN model	
Number of inputs	512
Number of neurons	512
Activation function	Relu
Number of epochs	10

6. RESULTS AND DISCUSSION

In this paper we have proposed a Deep Learning with Neural Network algorithm for diagnosis and detection of breast cancer. For pre-processing breast cancer dataset , we have used the standardization method . Then we have implemented a neural network algorithm and achieved 87.64% accuracy. By using breast-histopathology-images we tested 1728 images and received 87.64% which shows promising results as compared to other machine learning algorithms.

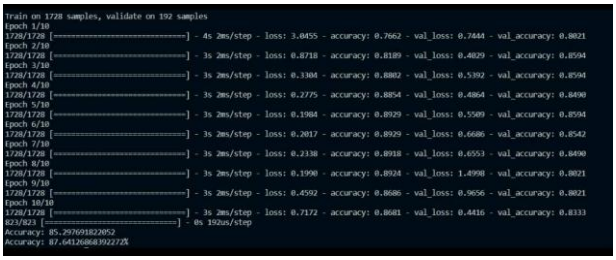

Fig -5: Result of model

7. CONCLUSIONS

Cancer detection using machine learning is a system that mainly targets the early detection of cancer. With the ever increasing surge in the number of cancer diagnosis there is a need for better technology that increases the chance of survival. Hence we aim to use machine learning for early diagnosis which eventually increases the survival rate. For this, we have fed the machine with categorized data sets of malignant and benign cell images and for better results we have also shuffled these images within themselves and then have successfully trained the machine with malignant and benign data sets after which successful testing is performed with an unknown data set.

8. REFERENCES

[1] Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong,"Histopathological breast-image classification with restrictedBoltzmann machine along with backpropagation", Biomedical Research Volume 29, Issue 10, (2018).

[2] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," 2010 5th International Symposium on HealthInformatics and Bioinformatics, Antalya, 2010, pp. 114-120.

[3] D. T. Saleh, A. Attia and O. Shaker, "Studying combined breast cancer biomarkers using machine learning techniques," 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herlany, 2016, pp. 247-251.

[4] A. I. Pritom, M. A. R. Munshi, S. A. Sabab and S. Shihab,"Predicting breast cancer recurrence using effective classification and feature selection technique,"

2016 19th International Conference on Computer and Information Technology (ICCIT), Dhaka, 2016, pp. 310-314.

[5] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras AlKhaimah, 2016, pp. 1-4. Computer and Communication Engineering. Vol. 5, Issue 3, 2017.

[6] R. D. Ghongade and D. G. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM algorithm," 2017 1st International Conference on Electronics, Materials Engineering and NanoTechnology (IEMENTech), Kolkata, 2017, pp. 1-6.

[7]https://www.kaggle.com/paultimothymooney/breast-histopathology-images.