Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

Breast Cancer Detection using Machine Learning Techniques

Sweta Bhise
Computer Engineering SRIEIT, Goa University
Shiroda, India

Simran Bepari
Computer Engineering SRIEIT, Goa University
Shiroda, India

Shrutika Gadekar
Computer Engineering SRIEIT, Goa University
Shiroda, India

Deepmala Kale
Computer Engineering SRIEIT, Goa University
Shiroda, India

Aishwarya Singh Gaur
Computer Engineering SRIEIT, Goa University
Shiroda, India

Dr. Shailendra Aswale
Computer Engineering SRIEIT,Goa University
Shiroda, India

***Abstract*—Breast cancer is the most common reason for deaths due to cancer. It is very necessary to detect cancer at early stages. There are various Machine Learning techniques available for the purpose of diagnosis of breast cancer data. This paper presents a Machine Learning model to perform automated diagnosis for breast cancer. This method employed CNN as a classifier model and Recursive Feature Elimination (RFE) for feature selection. Also, five algorithms SVM, Random Forest, KNN, Logistic Regression, Naïve Bayes classifier have been compared in the paper. The system was experimented on BreaKHis 400X Dataset. The performance of the system is measured on the basis of accuracy and precision. Activation function such as ReLu have been used to predict the outcomes in terms of probabilities.**

Keywords*—Breast Cancer, Dataset, CNN, KNN, Naïve Bayes, Random Forest, SVM, Logistic Regression

I. INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC)Trusted Source, breast cancer is the most common cancer in women. Breast cancer survival rates vary widely supported by many factors. Two of the most important factors are the type of cancer women have and the stage of cancer at the time they receive a diagnosis. Breast cancer is cancer that develops in breast cells. Typically, the cancer forms in either the lobules or the ducts of the breast. Cancer also can occur within the adipose tissue or the fibrous connective tissue within your breast. The uncontrolled cancer cells often invade other healthy breast tissue and may visit the lymph nodes under the arms.

Doctors say that breast cancer happened due to abnormal growth of cells in the breast and these cells spread in size like Meta Size from breast to lymph nodes or the other parts of the body also. Hence it is necessary to detect and stop the growth of these unwanted cells as early as possible to avoid the next phase consequences. If a tumor is diagnosed then the first step taken by the doctor is, they check whether the tumor is Benign or Malignant. Because the treatment and prevention methods of both the tumors are different. Benign cells are neither cancerous and nor spread but Malignant cells are cancerous and can spread to other parts of bodies. The problem with this disease is, there is no such proper diagnostic machine is present to detect cancer in the early phase so the person can start the treatment as early as possible and try to stop the growth of unwanted cells or tumors.

Early diagnosis of any disease is often curable with a touch amount of human effort. Most people fail to detect their disease before it becomes chronic. It leads to an increase in the death rate around the world. Breast cancer is one of the diseases that could be cured when the disease is identified at earlier stages before it is spreading across all the parts of thebody.

The lack of prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, time requires developing the technique which gives minimum error to increase accuracy. The available tests to detect breast cancer such as mammogram, ultrasound, and biopsy were time-consuming, so there was a need for a computerized diagnostic system in which Machine Learning methodology was used. This methodology includes algorithms that help for the classification of the tumor and detect the cells more accurately and take less time as well.

II. LITERATURE REVIEW

In this section, some of the related works previously done on breast cancer diagnosis by researchers using different machine learning approaches are discussed.

Arpita Joshi and Dr.Ashish Mehta [1], compared the classification results obtained from the techniques i.e. KNN, SVM, Random Forest, Decision Tree (Recursive Partitioning and Conditional Inference Tree). The dataset used was Wisconsin Breast Cancer dataset obtained from

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

UCI repository. Simulation results showed that KNN was thebest classifier followed by SVM, Random Forest and Decision Tree.

David A. Omondiagbe, Shanmugam Veeramani, Amandeep
S. Sidhu [2], investigated the performance of Support Vector Machine, Artificial Neural Network and Naïve Bayes using the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset by integrating these machine leaning techniques with feature selection/feature extraction methods to obtain the most suitable one. The simulation results showed that SVM-LDA was chosen over all the other methods due to their longer computational time.

Kalyani Wadkar, Prashant Pathak and Nikhil Wagh [3], did a comparative study on ANN and SVM and integrated various classifiers like CNN, KNN and Inception V3 for better processing of the dataset. The experimental results and performance analysis concluded that ANN was a better classifier than SVM as ANN proved to have a higher efficiency rate.

Anji Reddy Vaka, Badal Soni and Sudheer Reddy K. [4], presented a novel method to detect breast cancer by employing techniques of Machine Learning such as Naïve Bayes classifier, SVM classifier, Bi-clustering Ada Boost techniques, RCNN classifier and Bidirectional Recurrent Neural Networks (HA-BiRNN). A comparative analysis was done between the Machine learning techniques and the proposed methodology (Deep Neural Network with Support Value) and the simulated results concluded that the DNN algorithm was advantageous in both performance, efficiency and quality of images are crucial in the latest medical systems whilst the other techniques didn't perform as expected.

Monica Tiwari, Rashi Bharuka, Praditi Shah and Reena Lokare [5], presented a novel method to detect breast cancer by employing techniques of Machine Learning that is Logistic Regression, Random Forest, K-Nearest Neighbor, Decision tree, Support Vector Machine and Naïve Bayes Classifier and techniques of Deep Learning that is Artificial Neural Network, Convolutional Neural Network and Recurrent Neural Network. The comparative analysis between the Machine Learning and Deep learning techniques concluded that the accuracy obtained in the case of CNN model (97.3%) and ANN model (99.3%) was more efficient than the Machine Learning models.

Abdullah-Al Nahid and Yinan Kong [6], presented a novel method to detect breast cancer by image classification using Machine Learning techniques such as Convolutional Neural Network (CNN) method for breast image classification, conventional Neural Network (NN), Random Forest (RF) algorithm, Support Vector Machines (SVM) and Bayesian methods. The CNN method proved to be the best for the Breast Cancer detection as Convolutional Neural Network (CNN) techniques generally extract the features globally using kernels and these Global Features have been used for image classification.

K.Anastraj, Dr.T.Chakravarthy, K.Sriram [7], have performed a comparative analysis between differentmachine learning algorithms: back propagation network, artificial neural network (ANN), convolutional neural network (CNN) and support vector machine (SVM) on the Wisconsin Breast Cancer (original) datasets. Deep and convolutional neural network with ALEXNET was used for feature extraction are analysis the benign and malignant tumor. The simulation results concluded that support vector machine is the best approach and had given better results (94%).

S.Vasundhara, B.V. Kiranmayee and Chalumuru Suresh [8], have proposed instinctive classification of mammogram images as Benign, Malignant and Normal using various machine learning algorithms. A comparative analysis is performed between Support Vector Machines, Convolutional Neural Network and Random Forest. The simulation results concluded that CNN is the best classifier as it results in instinctive classification of digital mammograms using filtering and morphological operations.

Muhammet Fatih Ak [9], has used the dataset from Dr. William H. Walberg of the University of Wisconsin Hospital. Data visualization and machine learning techniques including logistic regression, k-nearest neighbors, support vector machine, naïve Bayes, decision tree, random forest, and rotation forest were applied to this dataset. R, Minitab, and Python were chosen to be applied to these machine learning techniques and visualization.A comparative analysis was performed amongst the all the techniques. Results obtained with the logistic regression model with all features included showed the highest classification accuracy (98.1%), and the proposed approach revealed the enhancement in accuracy performances.

Sivapriya J, Aravind Kumar V, Siddarth Sai S and Sriram S [10], have performed a comparative analysis between SVM, Logistic Regression, Naïve Bayes and Random Forest. The Wisconsin Breast cancer dataset is used to perform the comparison. Based on the result of performed experiments, the Random Forest algorithm showed the highest accuracy (99.76%) with the least error rate. ANACONDA Data Science Platform was used to execute all the experiments ina simulated environment.

Hiba Asria, Hajar Mousannif, Hassan Al Moatassime and Thomas Noel [11], conducted a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) dataset. Experimental

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

results showed that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed within a simulation environment and conducted in WEKA data mining tool.

Dana Bazazeh and Raed Shubair [12], performed a comparative analysis on three of the machine learning techniques, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN). The Wisconsin original breast cancer data set was used as a training set. Simulation results obtained has proved that classification performance varies based on the method that is selected. Results have showed that SVMs have the highest performance in terms of accuracy, specificity and precision. However, RFs have the highest probability of correctly classifying tumor.

III. METHODOLOGY

DATA SET

The data used for the experiments was acquired from Kaggle. This dataset is BreakHist_Dataset consisting of four directories representing the magnification of the images respectively i.e. 100X, 200X, 400X and 40X. The dataset consists of 7,858 instances in total which are divided into the four magnification directories. Each magnification directory consists of two directories representing the tumours i.e. Benign and Malignant.

PREPROCESSING

Feature Selection

The importance of feature selection in a machine learning model is inevitable. It turns the data to be free from ambiguity and reduces the complexity of the data. Also, it reduces the size of the data, so it is easy to train the model and reduces the training time. It avoids over fitting of data. Selecting the best feature subset from all the features increases the accuracy. Some feature selection methods are wrapper methods, filter methods, and embedded methods.

Recursive Feature Elimination

RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. This is in contrast to filter-based feature selections that score each feature and select those features with the largest (or smallest) score. Technically, RFE is a wrapper-style feature selection algorithm that also uses filter-based feature selection internally. RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains.

Segmentation

Splitting operation performed on images in 2X2, 3x3 up to 10X10 patches we called it as segmentation. In this segmentation process we train to the system to identify the close regions of interest which are important to detect the BC. By eliminating unrelated data from the image, it's easy to identify the tumor as early as possible. K-mean clustering algorithm is a method of groups it means similar objects combine in same group. Segmentation operation rely on it for better results and it gives better results when similar objects present in one group. It processes fastly as compare scattereddata [1].

CLASSIFIERS

Support Vector Machine (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane.

SVM clearly is the most effective classifier of all as it works really well with clear margin of separation and high dimensional data, but is not suitable for large data sets because the required training time is higher and also, underperforms when the data set has more noise.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. The working of KNN is based on finding the points in the data which is close to the new point enter into the machine. Then algorithm sort it separately that closet points in terms of distance frame arrival point. This particular distance in point is measured by using different methodologies but Euclidian distance is mostly used by experts. In next phase, take a particular number of points

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

whose distance is less as compare to other points and classify it into different category. Points are chosen in KNN is in the form of odd number, like number of classes is 2 similarly the highest number point is categories separately as a new data point [6].

KNN algorithm is very simple to implement and is capable of handling large data sets but, the computation cost is high because of calculating the distance between the data points for all the training samples and also, there is always the need to determine the value of K which may increase the algorithmcomplexity.

Random Forest

Random forest is a supervised learning algorithm. It is a collection of Decision Trees. Decision Tree is hierarchical in nature in which nodes represent certain conditions on a particular set of features, and branches split the decision towards the leaf nodes. Leaf determine the class labels. Decision Tree can be constructed either by using Recursive Partitioning or by Conditional Inference Tree. Recursive Partitioning is the step-by-step process by which a Decision Tree constructed by either splitting or not splitting each node. We can say that the tree is learned by splitting the source set into subsets based on an attribute value test. The recursion is completed when the subset at a node has all the same value of the target variable. Conditional Inference Tree is a statistical based approach that uses non parametric tests as splitting criteria that is corrected for multiple testing to avoidover fitting.
Random Forest is suitable for high dimensional data modeling as it can handle missing values, continuous, categorical and binary data but for very data sets, the size of the trees can take up a lot of memory. It can tend to over-fit, so there is a need to tune the hyper-parameters [1].

Logistic Regression

In linear regression, the linear regression hyperplane that is obtained cannot be used to predict the dependent variable by using the independent variable. Hence, when there is categorical data, logistic regression is used. Logistic Regression predicts whether something is true or false instead of predicting something continuous. It is used for classification.
The sigmoid function is used to convert the independent variable into an expression of probability which ranges from 0 and 1 concerning the dependent variable. The ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular Machine Learning algorithm. A drawback of Logistic Regression is the assumption of linearity between the dependent and independent variables.

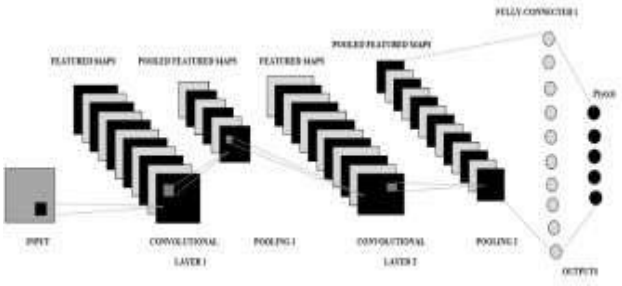
Naïve Bayes

Naïve Bayes classifier is a supervised learning algorithm that is used for classification. It is based on the Bayes theorem that is finding the probability of an event after an event has already occurred. It is one of the simplest yet powerful ML algorithms in use and finds applications in many industries. Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. This limits the applicability of this algorithm in real-world use cases. This algorithm faces the 'zero-frequency problem' where it assigns zero probability to a categorical variable whose category in the test data set wasn't available in the training dataset. We can overcome this issue by using a smoothing technique. Also, another drawback of Naïve Bayes is that it requires large data sets to attain its best accuracy.

IV. PROPOSED METHODOLOGY

The proposed methodology will help us to distinguish between malignant and benign tumor at a faster rate. CNN being a complex and complicated classifier can extract vital features automatically without depending on preprocessing. It is more proficient because it filters the important parameters and also is flexible being capable to work exceptionally well on image data.

A schematic structure of CNN is given below:



The main focus of our project is to differentiate between malignant and benign tumor using Convolution Neural Network with Keras in the backend and then analyze the result to see how the model can be useful in practical scenario.
The following steps are performed for model building and evaluation:

i. Importing all the libraries which are essential.
ii. Making dictionary of images and labels.
iii. Labels are based on image category.
iv. Normalization of image set.
v. Splitting data into training and testing sets.
vi. Building Architecture of model (CNN).
vii. Cross Validating Model (comparing with different classifiers).
viii. Testing Model.

Building of CNN:

Layers in Convolutional Neural Network:

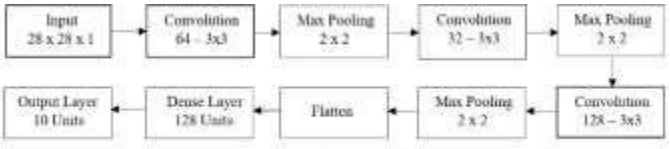


Fig. Building of CNN

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

Layer 1: Convolutional Layer

Convolution is the first step in the process of extracting valuable features from an image. Convolutional operations are performed by various filters present in the convolutional layers. Every image is considered as a matrix of pixel values.

Consider a 5x5 image whose pixel values are either 0 or 1. Consider filter matrix with dimension of 3x3. The dot product is computed to obtain the convolved feature matrix by sliding the filter matrix over the 5x5 image.

The first Convolutional layer of CNN is essentially standard image filter (+ ReLu). It aims to take a raw image and extract basic features from it. These are referred to as low level features. The second Convolutional Layer, instead of the raw image, accepts the features extracted by the first as its input. This allows it to combine these basic shapes into more complex features. The features extracted become more and more complex as we go further down the network. Layers near the middle of the network extract the so-called mid-level features, while the final layers extract thehigh-level features.

ReLU layer

ReLU stands for the rectified linear unit. Once the feature maps are extracted, the next step is to move them to a ReLU layer.

ReLU performs an element-wise operation and sets all the negative pixels to 0. It introduces non-linearity to the network, and the generated output is a rectified feature map. The original image is scanned with multiple convolutions and ReLU layers for locating the features.

Layer 2: Pooling Layer

A pooling layer is another building block of CNN. They are used to reduce the dimensions of the feature maps. Hence, it reduces the number of parameters to learn and the amount of computation performed in the network.

The Pooling Layer summarizes the features present in a region of the feature map generated by a convolution layer. So, further operations are performed on summarized features instead of precisely positioned features generated by the convolution layer. This makes the model more robust to variations in the position of the features in the input image.

Layer 3: Dropout Layer

Dropout Layer is used to set the weights of few randomly selected nodes to zero. Which means it drops few nodes and allows CNN model to learn the parameters from this new distributed set. It helps the model to overcome the problem of over- fitting.

Layer 4: Flatten Layer

Flattening is used to convert all the resultant 2-dimensional arrays from pooled feature maps into a single long continuous linear vector. This step of flattening is very essential in order to use the fully connected network after the layers of convolutional or maxpool. It further combines all the local features found from the previous layers. The flattened matrix is fed as input to the fully connected layer to classify the image.

Layer 5: Dense Layer

Dense layer is the regular deeply connected neural network layer. It is basically the out layer of CNN which can be seen as a fully connected layer.

I*mage recognition by CNN:*

The pixels from the image are fed to the convolutional layer that performs the convolution operation. It results in a convolved map. The convolved map is applied to a ReLU function to generate a rectified feature map. The image is processed with multiple convolutions and ReLU layers for locating the features. Different pooling layers with various filters are used to identify specific parts of the image. The pooled feature map is flattened and fed to a fully connected layer to get the final output.

CONCLUSION

In this paper we examined different machine learning techniques for breast cancer detection. We performed a comparative analysis of CNN, KNN, SVM, Logistic regression, Naïve Bayes and Random forest. It was observed that CNN outperforms the existing methods when it comes to accuracy, precision and also size of the data set.

REFERENCES

[1] Arpita Joshi and Dr. Ashish Mehta “Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer” (2017).
[2] David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu “Machine Learning Classification Techniques for Breast CancerDiagnosis” (2019).
[3] Kalyani Wadkar, Prashant Pathak and Nikhil Wagh “Breast Cancer Detection Using ANN Network and Performance Analysis with SVM” (2019).
[4] Anji Reddy Vaka, Badal Soni and Sudheer Reddy “Breast Cancer Detection by Leveraging Machine Learning” (2020).
[5] Monika Tiwari, Rashi Bharuka, Praditi Shah and Reena Lokare “Breast Cancer Prediction using Deep learning and Machine Learning Techniques”.
[6] Abdullah-Al Nahid and Yinan Kong “Involvement of Machine Learning for Breast Cancer Image Classification: Asurvey” (2017).
[7] K.Anastraj , Dr.T.Chakravarthy and K.Sriram " Breast Cancer detection either Benign Or Malignant Tumor using Deep Convolutional Neural Network With Machine Learning Techniques " (2019).
[8] S.Vasundhara , B.V. Kiranmayee and Chalumuru Suresh "Machine Learning Approach for Breast Cancer Prediction" (2019).
[9] Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications" (2020).
[10] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S "Breast Cancer Prediction using Machine Learning" (2019).
[11] Hiba Asria, Hajar Mousannifb, Hassan Al Moatassime, Thomas Noeld "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" (2016).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

[12] Dana Bazazeh and Raed Shubair "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis" (2016).
[13] Ramik Rawal “Breast Cancer Prediction Using Machine Learning” (2020).
[14] S. Karthik, R. Srinivasa Perumal and P. V. S. S. R. Chandra Mouli “Breast Cancer Classification Using Deep Neural Networks” (2019).
[15] Abdullah-Al Nahid, Aaron Mikaelian and Yinan Kong “Histopathological breast-image classification with restricted Boltzmann machine along with backpropagation.” (2018).
[16] Syed Jamal Safdar Gardezi, Ahmed Elazab, Baiying Lei and Tianfu Wang "Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review" (2019).
[17] Sebastien Jean Mambou , Petra Maresova , Ondrej Krejcar , Ali Selamat and Kamil Kuca.
[18] "Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model" (2018).
[19] Hannah Le "Using Machine learning models for breast cancer detection" (2018).
[20] Saleem Z. Ramadan “Methods Used in Computer- Aided Diagnosis for Breast Cancer Detection Using Mammograms: A Review” (2020).
[21] M. Tahmooresi , A. Afshar, B. Bashari Rad , K. B. Nowshath and M. A. Bamiah "Early Detection of Breast Cancer Using Machine Learning Techniques".
[22] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza and Nikahat Kazi “Breast Cancer Diagnosis and Recurrence Prediction Using Machine Learning Techniques” (2015).
[23] Shubham Sharma , Archit Aggarwal and Tanupriya Choudhury "Breast Cancer Detection Using Machine Learning Algorithms" (2018).
[24] Ram MurtiRawat ,Shivam Panchal ,Vivek Kumar Singh ,Yash Panchal "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning" (2020).