Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021)
IEEE Xplore Part Number: CFP21N67-ART; ISBN: 978-0-7381-4627-0

2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) | 978-1-6654-3877-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIRCA51532.2021.9544687

Breast Cancer Detection and Prediction using Machine Learning

Prof. Alok Chauhan*1, Harshwardhan Kharpate*2, Yogesh Narekar*3, Sakshi Gulhane*4, Tanvi Virulkar*5, Yamini Hedau*6

[1,3]Department of Information Technology, Rajiv Gandhi College of Engineering and Research Nagpur

[2]Department of Computer Engineering, Cummins College of Engineering for Women, Nagpur

441110, Maharashtra, India

(Affiliation to Rashtrasant Tukadoji Maharaj Nagpur University)

Abstract - *Cancer death is one of humanity's major problems in the developing world. Even though there are numerous ways to prevent it from occurring in the first place, some cancer types remain unrepeatable. Due to the absence of adequate forecasting, clinicians are unable to devise a treatment plan that will improve patient mortality rate. Hence, the requisite of time is to develop the technique which gives minimum error to increment precision. Three algorithms SVM, CNN and KNN which prognosticate the breast cancer outcome have been compared in the project utilizing different datasets. All experiments are executed within a simulation environment and conducted in PyCharm, Anaconda platform. Aim of research categorizes in three domains. First domain is presage of cancer and second domain is presage of diagnosis and treatment and third domain fixates on outcome during treatment. The proposed work can be used to predict the outcome of different techniques and suitable techniques can be used depending upon requirement. This project is carried out to predict, detect and analyze the accuracy of breast cancer. Future research can be done to predict other variables, and breast cancer research can be categorized based on these variables. It is possible to draw inferences from the facts obtained in this study about whether or not the patient has a breast tumor.*

Keywords : **Breast Cancer, Support Vector Machines, K-Nearest Neighbor, Convolutional Neural Network, Prediction, Analysis, Magnetic Resonance Imaging, Tumor, Diagnosis**.

I. Introduction

Breast Cancer: A malignant cancer that has developed from cells in the breast is referred to as "breast cancer." Breast cancer usually starts in the cells of the milk-producing glands, known as lobules or ducts that drain milk from the lobules to the nipple. Breast cancer can also develop in the breast's stromal tissues, encompassing fatty and fibrous connective tissues.

SVM: Support Vector Machines (SVM) is a very powerful algorithm, mainly when we're talking about Classification Problems. In this post, we'll implement SVM Algorithm to Classify Breast Cancer into Malignant or Benignant.

KNN: Clustering is denoted by the letter K, or we can say it informally. The KNN algorithm is a machine learning algorithm that is supervised and solves both relegation and regression problems. It's simple to use and comprehend.

CNN: Breast cancer develops when cells begin to proliferate uncontrollably. These cells usually grow into a tumor, which can be seen on x-rays or felt as a lump. If the tumor's cells can invade surrounding tissues or spread to other parts of the body (metastasize), the tumor is considered malignant.

Prediction: The intention of this study is to create a prediction system that can foretell the onset of breast cancer at an early stage by analyzing the smallest set of attributes from a clinical dataset. The proposed experiment was done utilizing the Wisconsin breast cancer dataset (WBCD).

Analysis and MRI: Breast MRIs are most commonly used in women who have been diagnosed with breast cancer. To assist in measuring tumor volume, searching for further cancers in the breast, and looking for tumor cells in the opposite breast. For those women at high risk of breast cancer, a screening MRI, in addition to a yearly mammogram, is indicated.

Tumor: Breast cancer is a type of cancer that develops in the breast cells. Breast cancer is the second most common cancer diagnosed in women in the United States, after skin cancer. Breast cancer can strike both men and women, but it affects women far more frequently.

Diagnosis: The cells are categorized based on how they appear under a microscope. Breast cancer is classified into two types: (1) invasive ductal carcinoma (IDC) and (2) ductal carcinoma in situ (DCIS), the latter of which advances slowly and has little influence on patients' daily life.

Breast cancer can be effectively treated if detected early. As a result, having access to appropriate means of screening is critical for detecting the first signs of breast cancer. To screen for this disease, many imaging modalities are performed with mammography, ultrasound, and thermography. Mammography is one of the most important early detection methods for breast cancer. Since mammography is inadequate for solid breasts, ultrasound or diagnostic sonography methods are widely used. Radiations from radiography can avoid small masses, and thermography may be more effective than ultrasonography in identifying smaller malignant masses considering these issues.

The term "breast cancer" refers to a big tumor formed from breast cells. Breast cancer typically begins in the cells of the lobules, or milk-producing glands, or the ducts, which are passageways that drain milk from the lobules to the nipple. Breast cancer can also start in the stromal tissues of the breast, which include the functional areas that are rich and rigid. Cancer cells have the ability to penetrate nearby healthy breast tissue and go to the underarm lymph nodes, which are microscopic structures in the body that filter out alien pathogens.

For so many years there was only one option to detect breast cancer and that is an X-Ray from which medicos used to check the current situation of the patient suffering from breast cancer and then give treatment to them. Now as time has passed many emerging technologies have taken partin medical science and solved many of our quandaries through their technical solutions like MRI, CT scan,etc.

The proposed work can be used to predict the outcome of different techniques and suitable techniques can be used depending upon requirement. This work done is carried out to predict, detect and analyze the accuracy of breast cancer. Future research can be done to predict other variables, and breast cancer research can be divided into categories based on these variables. Based on the findings of this study, it is reasonable to state that if a patient has a breast cancer tumor, it can be identified.

II. Literature Survey

There are different methods and researches of many medicos, edifiers in the field of breast cancer and they are working as well. In this paper, we have discussed it briefly.

By utilizing Ultra Wide Band (UWB) Zhang[1] introduced a breast tumor detector that uses gelatin-oil technology to obtain experimental results and, as a result, to demonstrate the efficacy of microwave images in the prostate cancer diagnosis.

To detect breast cancer diagnosis in an expeditious manner the author [2] proposed a Stochastic pulse synthesizer in UWB application that requires utilizing a static inverter with phase detector to expedite the system detection process.

By utilizing odds-ratio curves author [3] has introduced a logistic Generalized Additive Model (GAM) along with linear Kernel smoother.

The authors [4] utilized the Back propagation Neural Network to detect tumors who have reached the age of 40. Additionally compared the results to another model that used the circular sector function network.

For breast cancer presage the authors [5] have utilized direct subtraction beam composing imager, utilizing numerical simulation and electromagnetism in their model and discovered great resolution and robustness in diagnosing breast cancer.

The authors[6] have suggested malignant lump must be excised once a patient has been diagnosed with breast cancer. Physicians must determine the disease's prognosis during this procedure. This is a forecast of the disease's expected progression. Prognosis is important because the type and intensity of the medications are based on it. The term "analysis of survival or lifetime data" is also used to describe the prognosis problem. Since the data is censored, it presents a more difficult problem than diagnosis. That is, there are only a few cases where a disease recurrence has been observed. We can

classify the patient as recurrent in this case, and we know when they will recur (TTR). On the other hand, most patients do not experience recurrence.

Although most cancer research is clinical or biological in nature, data-driven statistical research has become a popular complement, according to the author [7]. Among the most fascinating and difficult tasks for which data mining applications can be developed is predicting the outcome of a disease. The goal of this study is to compile a list of review and technical articles on breast cancer diagnosis and prognosis. It provides an overview of current research using data mining techniques to improve breast cancer diagnosis and prognosis on various breast cancer datasets.

The author [8] has made use of Histology analysis allows for the distinction of benign and malignant tissue. Furthermore, this analysis aids in the performance of a prognostic evaluation. Changes in normal breast tissue structures are referred to as benign lesions. Invasive ductal carcinoma is one type of benign lesion. Algorithms for machine learning (DL) are being introduced to the mainstream audience in order to tackle a variety of Image analysis and image recognition difficulties.

These algorithms were used by the authors [9], and they produced good classification results, encouraging many researchers to use them to solve difficult problems. In breast histology pictures, a convolutional neural network (CNN) was employed to predict and categories invasive ductal cancer in [21], with a granularity of nearly 88 percent. Aside from that, data mining is commonly used in the medical field to predict and classify abnormal events in order to gain improved knowledge of serious diseases like cancer. The findings of employing text mining to detect and classify breast cancer are promising.

In [10] authors stage of breast cancer (I–IV) indicates cancer spread level in a patient. Stages are determined using statistical indicators such as tumor size, lymph node metastasis, and distant metastasis, among others. Patients must undergo breast cancer surgery, chemotherapy, radiotherapy, and endocrine therapy to prevent cancer from spreading. The research's goal is to identify and classify malignant and benign patients, as well as to figure out how to parameterize our classification techniques for high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer.

The author utilized [11] Clustering as a type of difficult optimization issue. Researchers have used a variety of ML algorithms to solve this categorization challenge. The following sections provide an overview of the two primary breast cancer categorization techniques. Scientists try to develop the optimal algorithm for achieving the most statistical basis outcome, however the classification outcome will also be influenced by data of variable quality. Furthermore, the scarcity of data will have an impact on the number of algorithm applications. The majority of machine learning algorithms are initially evaluated in open source datasets.

The author [12] suggests breast cancer instances are classified as either cancerous or innocuous. The objective of this research is to determine the number of hidden layers, the number of neurons in each hidden layer, and the kind of activation functions in hidden levels in order to build an Artificial Neural Network (ANN) with a high degree of accuracy. The Wisconsin Breast Cancer Database (WBCD), an open-access dataset, provided samples for ANN training and test. There are 699 samples in the dataset, which were split into two groups: The trained model has 599 samples, whereas the test data contains 100 samples. As inputs to the network, each sample has nine attributes that represent nine characteristics of breast fine-needle aspirates (FNAs). When three transfer functions were used, the mean square error (MSE) achieved was equated in this experiment.

The author [13] have utilized the most common indicator of breast cancer is the growth of a painless, firm mass in the breast. In the absence of a mass, however, around 10% of patients experience pain. Breast cancer, like other forms of cancer, can be treated more successfully if caught early. Early identification of melanoma improves not only the no of options, but also the likelihood of treatment success and survival [4, 5].

The author [14] has to utilize several active signal processing approaches for prostate cancer imaging, including scanning and sensor techniques. For generating the dielectric

characteristics of the breast, the tomography technique is generally conducted It is done continuously and can be represented by a complex inverse problem that needs substantial computer utility. Inverse scattering methods by assessing the absorbed and reflected microwave signals, the intrinsic properties of the breast tissues may be estimated, allowing imaging of the breast tissues to be generated from the retrieved microwave data file.

III. Methodology

There are many algorithms available for breast cancer detection and they are working well. We have utilized some of them like KNN, CNN, and SVM. We have divided the work task into 3 phases, the first one is to prognosticate cancer making utilization of MRI of the person and the second phase is to detect cancer or the tumor at the exact position in the MRI and the last phase is of analysis.

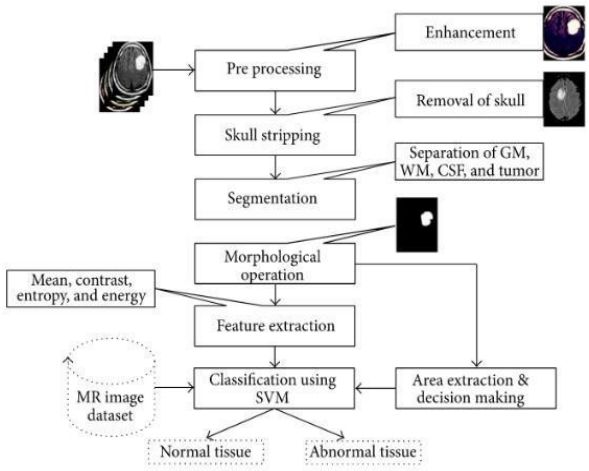


Fig. 1 Block Diagram

Here we have discussed all the steps and methods we have utilized for this researcher. First, we will take input from the utilizer and that is an MRI of the person for the first process. And the first step of the entire block diagram is to pre-process the image.
Pre-processing is required to convert data into the format in which it can directly be input into the network. This step involves multiple channeling of images, then the segmentation is done (only if required, e.g. if there is a desideratum to disunite regions of interest from the background or omit components that are no longer needed for training). After following this first step now the data is to be utilized, either in a supervised or an unsupervised manner.

The next step is Feature extraction. Features represent the visual content of the MRI image. In the case of supervised feature extraction, features are kenned and different strategies are applied to find them, but in the case of unsupervised feature extraction methods, features are not Kennedy and acquired implicitly in proposed solutions through the Convolutional Neural Network (CNN).

The last step is relegation utilizing Support Vector Machine commonly called as SVM which puts an image into the respective class and with a plenarily connected layer utilizing an activation function such as Softmax.

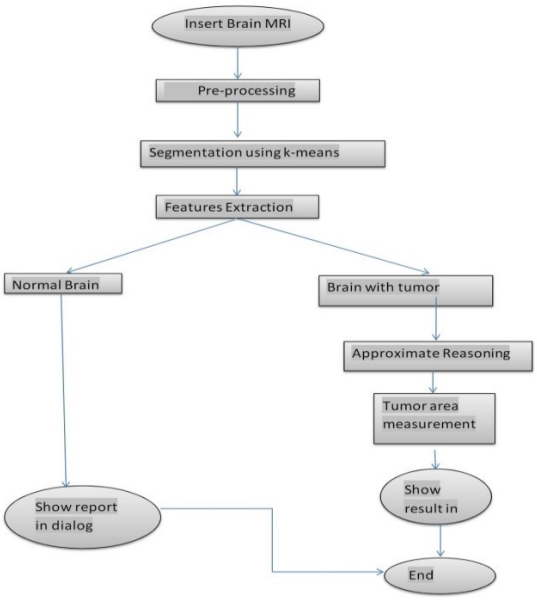


Fig. 2 Flowchart

MRI: Magnetic Resonance Imaging (MRI) displays all of the detailed pictures of the body's organs and tissues.

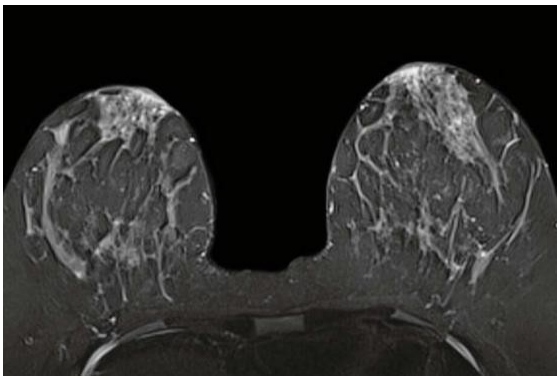
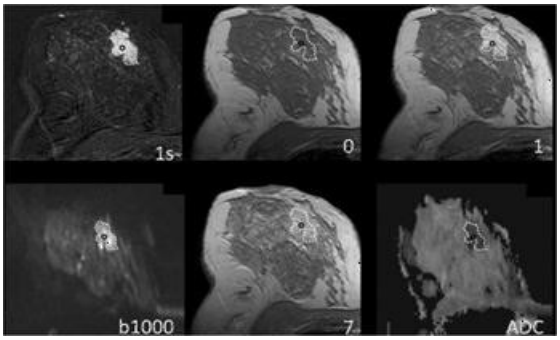


Fig. 3 MRI of Breast Cancer



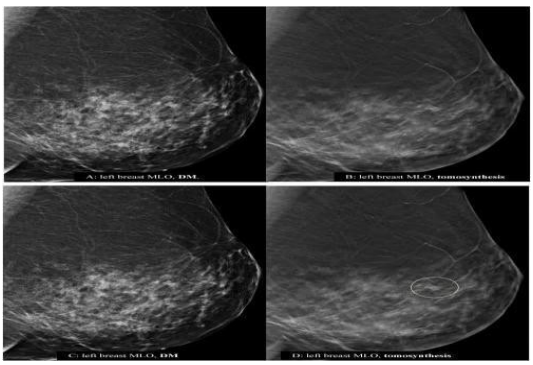


Fig. 4 De-noising of MRI

Fig 5. Feature Extraction Process

Pre-processing:

Pre-processing is a method for performing operations on visuals at the lowest possible level of complexity. The objective of pre is to increase the picture quality.

Feature Extraction:

Feature extraction is a component of the dimensionality truncation process. In this, an immensely colossal number of pixels of the image is efficiently represented in a way that only fascinating components of the image are captured.

Classification:

The task of assigning a label to an image is known as relegation. The relegation technique is used to classify every pixel in a digital image into one of several categories.

Segmentation:

The technique of segmenting a digital image into many pieces is known as image segmentation. The term "image segmentation" refers to the process of locating objects and boundaries in images.

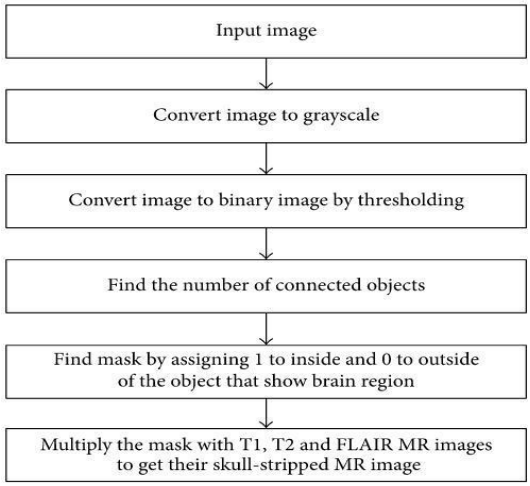


Fig. 6 Process of Breast Cancer Detection

CNN:

CNNs are used to look for patterns in images. This is accomplished by pattern recognition using an image. In the few front layers of CNNs, the network can recognize lines and corners. However, as we go further we may transmit these patterns down and begin to recognize more intricate qualities by distorting an image and looking for patterns in our neural net. Because of this characteristic, CNNs are highly good at spotting object in images. The suggested technique employs Neural network to detect breast cancer in scans of breast tissue.

As demonstrated in Figure 5, a CNN contains three primary layers: a convolutional layer, a

pooling layer, and a fully connected layer. In the first layer, the response of neurons associated to local regions is computed. Each one is derived using a dot product of weights and region. Typical image input filters are modest in area, such as [3 3], [5 5], or [8 8]. These filters using a scaling factor on the image to scan the image while learning repeating correlations that appear in any section of the image. The stride is the distance between two filters. If the filter component is lower than the stride hyper component, the convolution is extended to overlapping windows.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

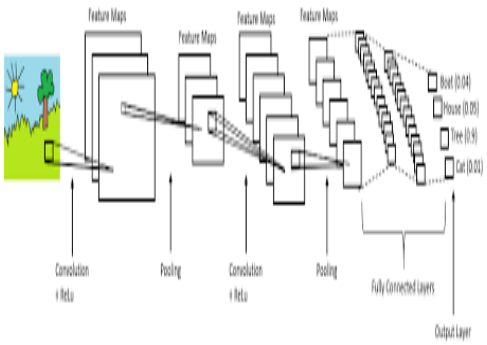


Fig. 7 CNN Algorithm Working

KNN:

Clustering is denoted by the letter K, or we can say it verbally. The KNN method is a supervised machine learning technique that solves both relegation and regression problems. It's simple to use and comprehend.

K-nearest neighbour (k-NN) is a pattern recognition technique that uses training datasets to find k's closest relatives in subsequent samples. The theory of the adjacent method is used to describe a set of training datasets near the new point, which are then used to anticipate the breakpoint.

As in k-nearest neighbor (k-NN) learning, the user can choose the specimen or change it based on the local point density.

K-denotes clustering or we can verbalize KNN algorithm is a machine learning supervised algorithm which solves both relegation and regression quandaries. It's easy to implement and understand. K-nearest neighbour (k-NN) is a pattern recognition technique that uses training datasets to find k's closest relatives in subsequent samples. Any indicator standard can be used to calculate the distance. A likely candidate is Euclidean distance. The closest neighbor is also usable for a large number of datasets due to its simple structure and can produce better results for complex datasets.

for the vector $\mathbf{p} = (p_1,\ldots, p_n)$ has n scalar components.

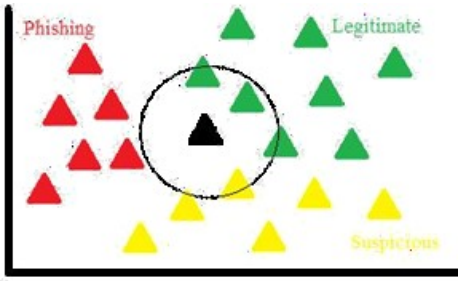


Fig. 8 KNN Algorithm Working

SVM:

Long-term support for SVM Vector Machine is a supervised machine learning system that may be used to solve both regression and relegation problems. But due to its high precision and results, it is widely utilized in the relegation process. It is additionally kennedy as a binary relegation algorithm

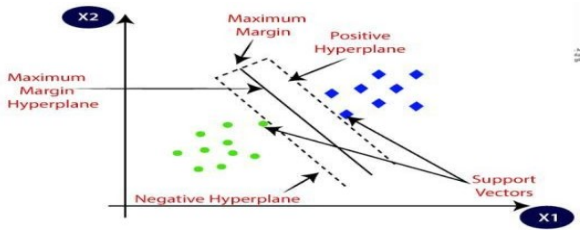


Fig. 9 SVM Algorithm Working

Table presents the comparative results on technique SVM, KNN and accuracy of the technique is up to 95% and in the second column the author name is Chaurasia el.at and with the dataset of Wisconsin Breast Cancer Detection using tool PyCharm Navigator and technique SVM and accuracy of the

technique is 97.13% and next and next the third column the author name is Saad Awadh Alanazi and with the Boosting Breast Cancer Detection using tool PyCharm Navigator and technique CNN and accuracy of the technique is 93.1%.

Author	Dataset	Tool	Technique	Accuracy
Keles, M Kaya	Breast Cancer Wisconsin dataset	Python	SVM, KNN	Upto 95%
Chaurasia el. At	Wisconsin Breast Cancer (Original) dataset	Pycharm Navigator	SVM	97.13 %
SaadAwadhAlanazi	Boosting Breast Cancer Detection	Pycharm Navigator	CNN	93.1%

IV. Result

- Click on the Detect button, Then Image show the cancer is detected or not.

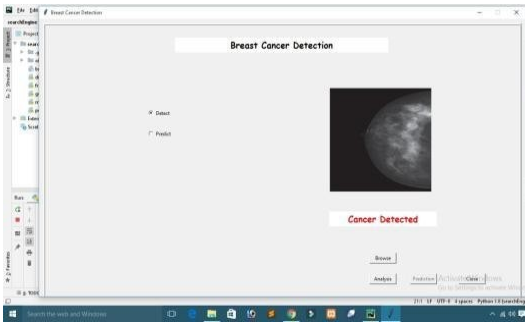


Fig. 10 Cancer Detected

- If the Cancer is detected click on the predict button, then the image is converted into the gray scale.

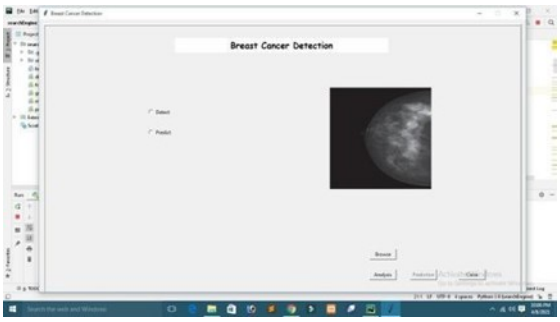


Fig. 11 Gray Scale Conversion

- For More Clarity,Againclick on the predict button then the image is converted from gray scale to HSV It show the clear tumor present in the breast.

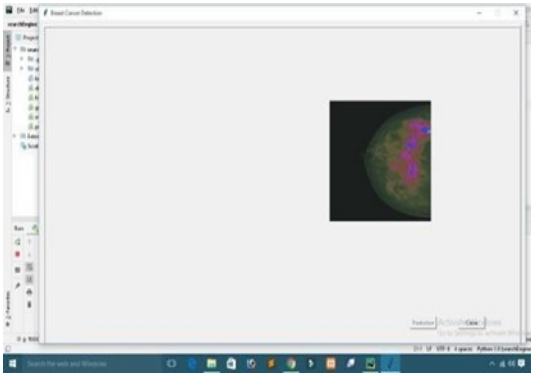


Fig. 12 HSV conversion

- From the image graph it shows the size of block or segment And the No. Of pixels also as image is change No. Of pixels is also change.

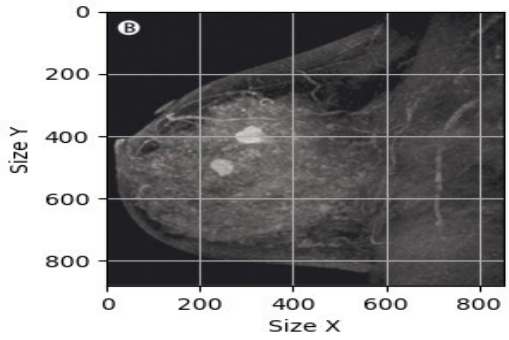


Fig. 13

- From the graph 1, graph shows the stage 1 of cancer it can be verified from the image that total No. Pixels present in the graph shows the stage of the cancer. Here the accuracy of graph is low because the cancer is at first stage. No. of pixels is also minimum.

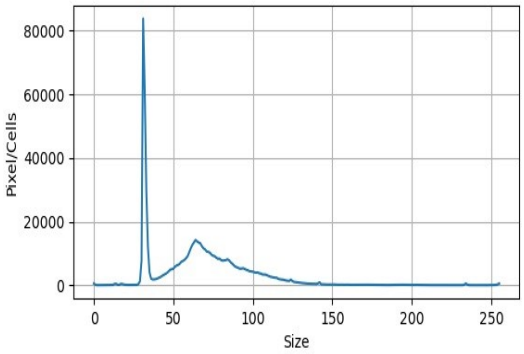


Fig. 14

- From the graph 2, graph shows the stage 2 of cancer the yellow line in the graph indicated that the accuracy of graph is high.

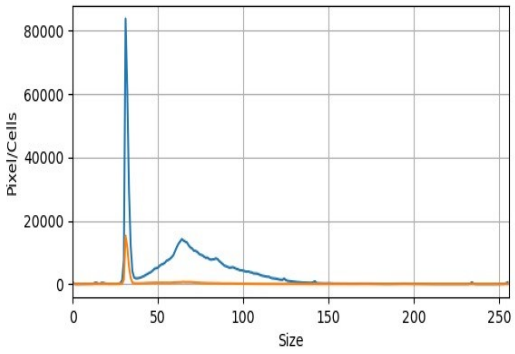


Fig. 15

- In the graph 3,No. Of pixels are change because the cancer is on stage 3 and the accuracy of graph is very high.

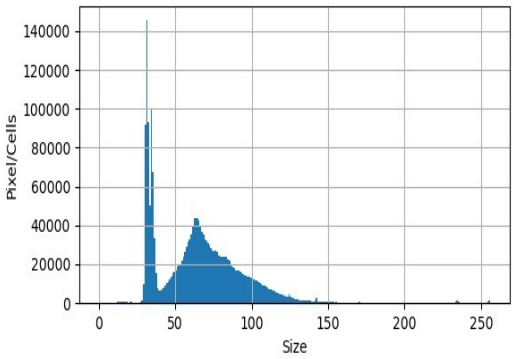


Fig. 16

V. Conclusion

All the above methods and breast cancer detection process is prosperously done with the avail of three algorithms that we have utilized are SVM, KNN, CNN which proved to be very efficacious and have vigorous precision to obtain the exact results.

SVM and CNN were utilized as a relegation algorithm while KNN is utilized to find the precision of the result and for front end design PyCharm libraryTKinter () is used.

We have use 3 Algorithms in that SVM accuracy is 98% and CNN accuracy is up to 95% and KNN accuracy is 73%. SVM and CNN these algorithms has high accuracy than other algorithms. So we have used these algorithms.

V. References

[1]. A systematic and quantitative review of "magnetic resonance imaging" J. Psych. 2020; 172:110-20 by Larie S, Abukmeil S.

[2]."A Hybrid Model to Support the Early Diagnosis of Breast Cancer," introduced by D. Carvalho, at Procedia Comput. Sci., vol. 91, pp. 927–934, Jan. 2019.

[3]."Tumor prediction system" at IJRET in 2019 by M K Agrawal, Meena Keshav Kumari.

[4].Cancer detection using histopathological image by RS Barman, in the year 2020 at IEEE.

[5].Ikeda, Debra M. Breast Imaging: The Requisites. Philadelphia, PA: Elsevier Inc.; 2020: 90-162.

[6].Clough GR, Truscott J, and Haigh I. 'Can High Frequency Ultrasound Predict Metastatic Lymph Nodes in Patients with Invasive Breast Cancer?' The Society and College of Radiographers. 2019; 12: 96-104.

[7].SarvestanSoltani A., Safavi A. A., Parandeh M.

N. and Salehi M., 2011, "Predicting Breast Cancer Survivability using data mining techniques," Software Engineering (ICSTE), 2nd International Conference, v2, 227-231.

[8].Alghodhaifi, H., Alghodhaifi, A., Alghodhaifi, and M.: "Predicting Invasive Ductal Carcinoma in breast histology images using Convolution Neural Network". In: 2019 IEEE National Aerospace and Electronics Conference (NAECON), pp. 374–378 (2019)

[9].Silva, J., Lezama, O.B.P., Varela, N., Borrero, 'L.A.: Integration of data mining algorithms and ensemble learning for predicting the breast cancer recurrence'. In: Miani, R., Camargos, L., Zarpelão, B., Rosas, E., Pasquini, R. (eds.) GPC 2019. LNCS, vol. 11484, pp. 18–30. Springer, Cham (2019).

[10]. B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4

[11]. Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

[12]. "On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset" by Abien Fred M. Agarap, 7 February 2019.

[13]. "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study" by Mumine Keles, Feb 2019

[14]. Yang L, Fu B, Li Y, Liu Y, Huang W, Feng S, et al.'Prediction model of the response to neoadjuvant chemotherapy in cancers by a Naive Bayes algorithm.' Computer methods and programs in biomedicine. 2020; 192:105458.

[15]. Sutton , Onishi N, Fehr DA, Dashevsky BZ, Sadinski M, Pinker K, et al. 'A machine learning algorithm that classifies cancer(breast) pathologic complete response on MRI post-neoadjuvant chemotherapy.' Breast cancer research: BCR. 2020;22(1):57. pmid:32466777

[16]. Phung MT, Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. BMC cancer. 2019;19(1):230. Pmid:30871490.