International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979; Impact Factor: RJIF 5.72
Received: 24-01-2020; Accepted: 25-02-2020; Published: 19-03-2020
www.allsubjectjournal.com
Volume 7; Issue 3; 2020; Page No. 97-105

Breast cancer tumor classification using machine learning algorithms

Nouman Fazal
Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan

Abstract
Breast cancer is one of the most deadly diseases, even in this modern era. However, it can be cured if it is detected in the early stages. The main issue is that breast cancer does not have very prominent symptoms in the initial stages. These cancers can be screened by different modalities and then manually screened by radiologists. The manual screening of cancer tests by radiologists is prone to error and time consuming, therefore it is highly advantageous to have a network of machine learning analysis support systems to assist radiologists' decision. Mammography is a common screening modality used for screening of breast cancer. In this paper, textural features are extracted from digital mammograms collected from the Digital Database for Screening Mammography (DDSM). These features are used for the classification of the tumor using different machine learning algorithms. The performance of the classifiers is evaluated using Accuracy, Precision, Recall, F1- Score, Training time, Receiver Operating Curve (ROC), Area under the Curve (AUC) and different Errors.

Keywords: classification, Breast, modern, cancer

1. Introduction

Breast cancer is very common disease among women [1] classified into separate groups based on distinctive methods and diverse decisive factors [2], which helps in applying different treatments. The major categories of classification are based on the pathological kinds, tumour's location, and tumour phase as well as protein and gene presence [3]. Each of these characteristic manipulates tumor handling response and analysis. In addition to all these features, symptoms found on physical exam can also be depicted in the breast cancer classification.

The complete classification of breast cancer mammogram must comprise of pathological form, status, phase, receptor condition, and based on DNA testing, the occurrence or nonappearance of genes. We have emphasized on the tumor state, because the state of breast cancer cells is structured. They can be distinguished from the texture of normal breast tissue. Normal cells have a controlled growth rate and they grow in a certain pattern. In an organ like the breast, it can be distinguished, indicating that they acquire on specific shape and appearance that specify their functioning as part of the breast [4]. This distinction of cells is lost in Abnormal (cancerous) cells. In cancer, the cells growth is disorganized, which would usually grow in a controlled manner to structure the milk ducts. The Cells splitting upturns into uncontrolled growth. Cell nuclei turns into irregular shape. The classification task is carried out in a various human practices [5].

The classification problem is related to the development of a mechanism that is helpful in the grouping of instances in which each new item has to be allocated to being one of the predetermined groups [6] based on statistical attributes or characteristics. The breast cancer diagnosis and prognosis can be significantly improved with the help of Information and Communication Technologies (ICT) [7]. The data mining techniques have advanced not only the size of data but also creating meaningful interpretation from raw data. It has made a big change in healthcare from reporting and decision to prediction results [8]. Implementation of data mining techniques for medical science area rise rapidly due to their high effectiveness and correctness in predicting outcomes, procurement of low cost medicines, increasing the survival rate from fatal diseases such as breast cancer and in making real time decision to save patient's lives.

There are many different algorithms for classification and prediction of breast cancer. They are mostly implemented individually for classification. The present paper gives an implementation of eight classifiers: Support vector Machine (SVM) [9], Decision Tree [10], K-Nearest Neighbors (KNN) [11], Logistic Regression (LR) [12], Multi-Layer Perceptron (MLP) [13], AdaBoost Classifier (ABC) [14], Gradient Boosting Classifier (GBC) [15], and Random Forest (RF) [16] which are among the top data mining algorithms and are widely used by research community [17, 18]. The aim of this work is to evaluate the efficiency and effectiveness of each algorithm in terms of accuracy, sensitivity, F1 Score, precision, ROC and AUC. The rest of this paper is categorically organized as follows. Section II is about literature review and related work. Section III presents the experimentation. Section IV deals with the tumor detection by algorithms. Finally, section V concludes the paper.

Related work

This section describes related work of the implementation of various Machine learning algorithms by researchers, for the classification of breast cancer.

Ahmad *et al.* [19], worked on the classification of clinical dataset, obtained from the Iranian Center for Breast Cancer. The performance of the classifiers (C4.5, SVM and ANN) was evaluated. Their simulation results indicate that SVM was performing better than ANN and C4.5.

Nematzadeh *et al.* [20], classified the well-known Winconsin Prognostic Breast Cancer (WPBC) and Wisconsin Breast Cancer (WBC) using the Decision Tree, Naives Bayes, ANN and SVM. The performance of the classifiers was used with three distinct kernel functions. The performance

of classifiers was assessed for both of the datasets using 10-fold cross validation technique. The experiments indicated that SVM performs better than other classifiers for WPBC dataset, having highest accuracy in Cross validation of 98.32%. However, ANN has accuracy of 98.09% in 10-folds for WBC, which is better than other classifiers for this dataset.

Hasan and Tahir [21], suggested the dataset preprocessing using Principal Component Analysis (PCA) for conversion of correlated features to uncorrelated features and classification of the preprocessed dataset using Artificial Neural Network (ANN). The PCA implementation have indicated the improved classification rate between benign and malignant tumor in WBC dataset.

Ojha and Goel [22], used four clustering and four classification algorithms for classification of recurrent breast cancer cases using WPBC dataset. The comparison of the classification results indicates that SVM and Decision Tree (C5.0) predict the cancer with 81% accuracy. However, the fuzzy c-mean have the lowest accuracy of 37%. The mean accuracy of clustering and classification algorithms is 52% and 71% respectively for the dataset.

Ghosh *et al*. [23], used Multilayer Perceptron using Back Propagation Neural Network (MLP BPN) and SVM for diagnoses and analyzing breast cancer tumor. These algorithms are well known and widely used in the scientific community for classification. SVM showed lower error rates and hence, it has powerful classification and predictive capability in terms of medical and bioinformatics. The experiments indicated that SVM is better than MLP BPN in terms of accuracy and error rate.

Osareh and Shadgar [24], evaluated the implementation of feature ranking, feature elimination, feature selection and feature transformation in combination with SVM, KNN and Probabilistic Neural Network (PNN) for the diagnosis and prognosis of breast cancer. Recursive feature elimination (RFE) is used for elimination of features from the dataset and selecting features based on their qualitative importance for modeling error reduction. Correlation based feature selection (CFS) is used as feature selection technique with correlation filter of 0.7.The Linear Discriminant Analysis (LDA) and PCA is implemented for relevant features extraction. The SVM-RBF have accuracy of 98.80%, which is the best among the overall accuracies.

Bazazeh and Shubair [25], implemented SVM, Random Forest (RF) and Bayesian Network (BN) for the diagnosis of breast cancer and carried out detailed and systematic observation on them. Waikato Environment for Knowledge Analysis (WEKA) framework is used for implementation of the algorithms, it is easy to use and have high portability. To prevent overfitting of the classifiers, 10 fold cross validation is used. The accuracy, Precision, Recall and Area under the Curve is used to access the classification ability of the algorithms. Their Results indicate that RF have high chances of correctly classifying the tumor instances. However, SVM is better in terms of Accuracy, Precision and Recall.

Azmi and Cob [26], employed feed-forward back propagation Artificial Neural Network algorithm to classify breast cancer tumor. The dataset contains 699 cases with 241 malignant and 458 benign cases. Each of the case have 9 different attributes, which are used for training and testing. The randomized dataset is spliced in 70% training and 30% testing set. The best performing network has 3 layers, where input layer have 9 nodes and hidden layer has 7 nodes and 2 output nodes. The highest accuracy achieved by network is 96.63%, when compared to other networks.

Gayathri and Sumathi [27], provided the implementation of Relevance Vector Machine (RVM) for breast cancer tumor detection and also compared the implied system with other machine learning algorithms used for the same task. The algorithm is trained and tested with dataset of 300 instances and reduced features using Linear Discriminant Analysis (LDA). The accuracy of 96% is achieved by the classifier with sensitivity and specificity of 98% and 94% respectively.

Sonavane *et al* [28], implemented Linear Vector Quantization neural network (LVQ-NN) for the classification of the MRI images of brain tumor and Mammogram for Breast Cancer for normal and abnormal cases. The dataset for breast cancer consists of 289 images, out of which, 228 images were used for training and 61 images for testing. The images were preprocessed using different image processing techniques. Textural features were extracted from the preprocessed dataset and used for classification. The approach is novel however; the accuracy of classification is only 69%. The sensitivity and specificity is 82% and 43% respectively.

Sejong Yoon and Saejoon Kim [29] have classified mammograms from DDSM using ensemble. The Recursive Feature Elimination based SVM (RFE-SVM) is used for feature selection. Ideally infinite features can be extracted from a mammogram. However, features suitable for better classification need to be determined. The cross-validation in RFE-SVM is used to rank the features best suited for the machine learning algorithms to more accurately classify the mammograms. The datas*et al*so contains the clinical mammograms.

Methodology

The goal throughout this work would be to explore the issue of categorization and to demonstrate with image analysis the presence of breast cancer. Figure 1 describes the architectural description of the method suggested for classifying breast cancer as normal or abnormal.

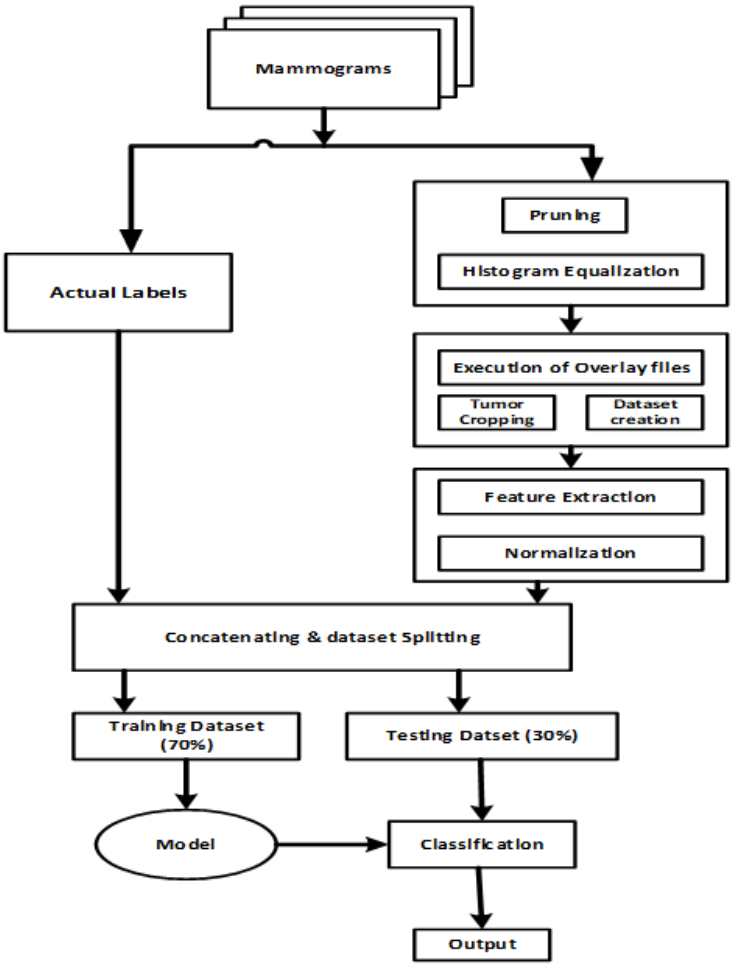

Fig 1: The suggested approach to the detection of breast cancer tumor

In this proposed architecture, Figure-1 indicates that, the Mammography images from DDSM are obtained. To enhance image quality and eliminate distortions, the image processing approaches on the images are used. Then, textural features are obtained from the enhanced image by means of GLCM. The categorization framework is established utilizing simplified and marked characteristics training data. Ultimately the algorithm categorizes the images into the usual ordinary or unusual category using the testing data.

Breast cancer database

The database used in this work for training, testing and experimentation is the Digital Database for Screening Mammography (DDSM).The database is managed by the University of South Florida. The database contains approximately 2600 studies containing Malignant, Benign and Normal cases. Each case study has 2 images of each breast. The Figure-2 is a mammogram with an abnormality.

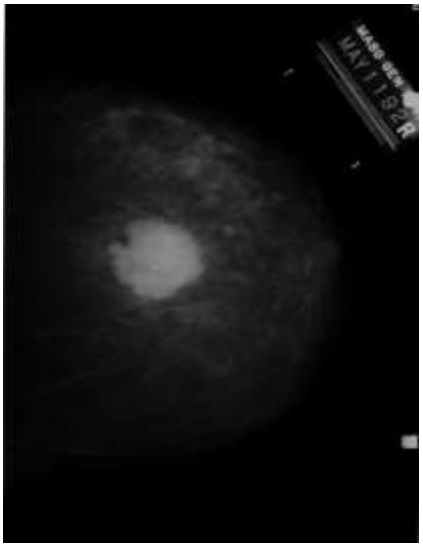

Fig 2: Original Mammogram

Image Acquisition

To inspect the human breast for diagnosis and screening, one of the well-known methods is mammography. Mammography (also known as mastography) is the procedure of utilizing low-energy X-rays (typically about 30 kVp). The mammogram is the image obtained from the mammography, which is the breast's X-ray image. Mammograms are often used in early stage breast cancer screening by medical practitioners. To increase the quality of the mammogram, the breast is compressed in parallel plates to level out the thickness of the breast, by means of a sophisticated mammography unit. When the breast density is reduced, X-rays go deeper into the breast, reducing the quantity of dispersed radiation (spread out reduce the image quality). It also reduces the dose of the required radiation, and Helps maintain the breast motionless (Stop blurring due to motion). In mammography screening, both the up-to-down view (craniocaudal, CC) and the view of the breast from a certain angle (mediolateral oblique, MLO) are displayed. The radiologists look for tumors in the mammograms. Micro-Calcifications (MCs) and masses are the main causes of formation of these tumors. The MCs are the microscopic calcium sediment built up within the breast tissues. Figure 3 shows the Mammogram along with its specific sections.

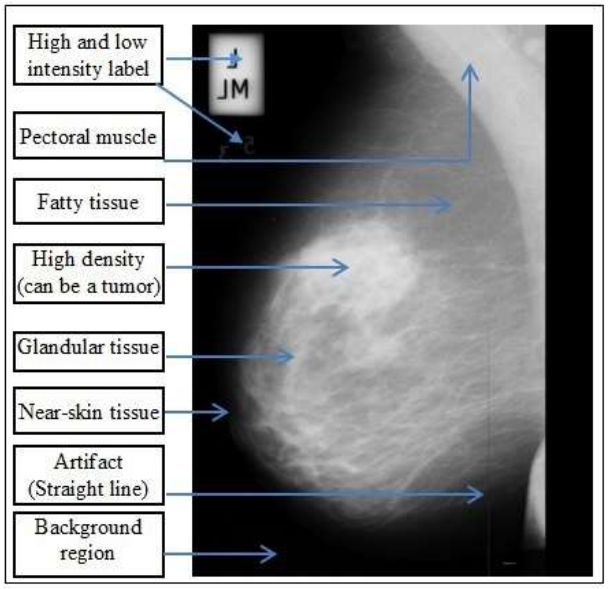

Fig 3: Various regions of a mammogram

- High and low intensity label are mammogram identifiers used to identify the mammogram. It includes the patient ID and tumor identification.
- Pectoral muscle is usually present in a mammogram. It might be pectoral major or pectoral minor muscle. They are not part of the breast but it is difficult to avoid them in mammography due to their location.
- Fatty tissues are normal non dense breast tissues.
- High density region in a mammogram might be a tumor or normal dense supportive tissues of the breast.
- Glandular tissues include lobules (which produce milk during lactation) and ducts that take milk from the lobules to the nipple during breastfeeding.
- Near Skin tissues usually have low intensity than other parts of the breast.
- Artifacts are the defects in the mammogram. It might be patient related such as movement during mammography and skin line artifacts or Hardware associated, for instance, X-ray tube filter imperfection or else grid artifacts.
- The background is the non-breast region of the mammogram.

Image Processing

Digital mammograms have abundance of unessential information and artifacts which makes it challenging to interpret it for correct diagnosis. Preprocessing improve the quality of the mammogram images and removes unnecessary segments from it. Preprocessing is essential for a consistent and precise feature extraction stage. It requires two main stages. Both phases work simultaneously to enhance the quality of the mammogram at the same time. The first stage is to make important changes to the area of the picture that draws the attention of the radiologist. The second stage is to delete the excess of the unwanted information and additional unwanted parts of the image. The pectoral muscle tissues, which were captured by the mammogram and they are not the part of breast, are removed and only breast tissues are left for further processing.

The mammogram images are of very large size and approximately 50% of the area contains the unwanted background area and mammogram description tags. As illustrated in figure 4.

The mammogram images with some abnormality have related information on the location and category of the doubtful area in the pixel level "ground truth." The tumors information are recorded in the overlay files. This overlay file information was extracted using the ' GET DDSM GROUNDTRUTH ' command from the matlab. As shown in Figure 3.5, the last image is that of the mammogram with a precisely marked boundary of a tumor.

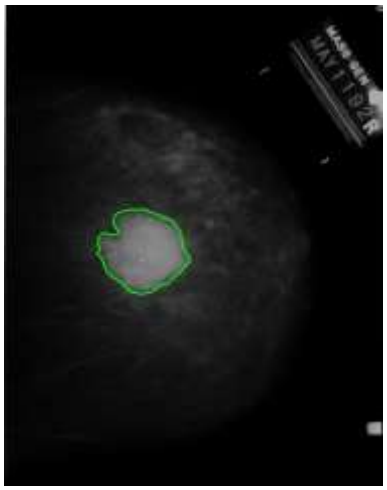

Fig 4: Tumor marked from overlay file

We then use crop operation to get rid of the unwanted portion of the image. This deletes approximately all background areas, Non-Breast muscular portion and the description labels on the images. This is required to decrease the size of the image for saving and further processing. Figure 5 illustrates the image pruning operation.

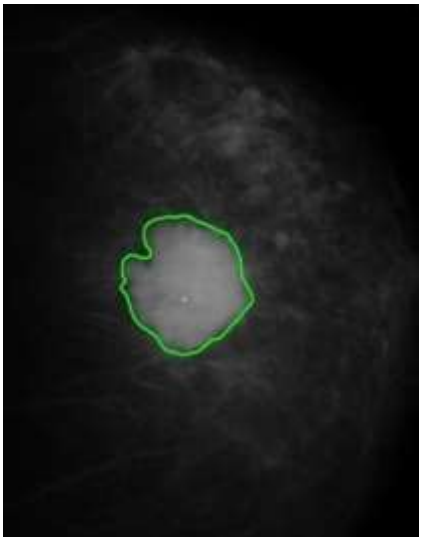

Fig 5: Image pruning

Tumor ROI cropping

The malignant and benign cases are combined into tumor category. The tumors outlined with the overlay file information and a parenchyma region from Normal mammograms are then cropped using 150x150 pixels window. The window size is selected based on the smallest tumor in the database. There are a total of 4711 tumor and 4711 Normal crops to be used for further processing. Some mammograms are not cropped and kept separate for tumor detection testing.

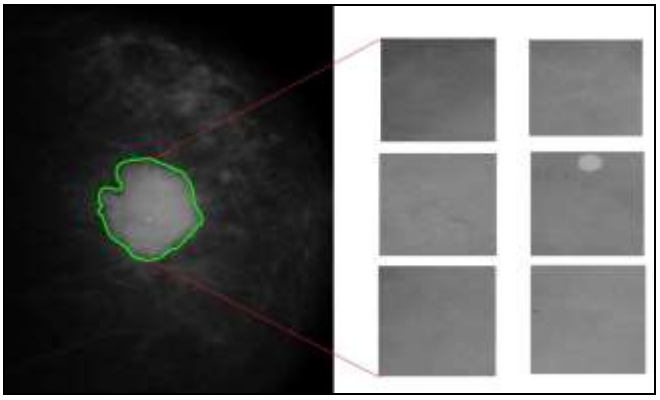

Fig 6: Tumor cropping

Gray Level Co-Occurrence Matrix (GLCM)

The tumor area from the mammogram is extracted and the GLCM is computed. For an image, the co-occurring values at a defined offset or the distance and angular spatial relationship over an image sub-region of a specific size, a co-occurrence matrix, also known as the co-occurrence-distribution is defined. From the gray-scale images, GLCM is computed. This measures the frequency of the horizontal, vertical or diagonal occurrence of a pixel with Gray level (gray intensity or Tone) relative to the neighboring pixels.

GLCM describes the relationship among two pixels at a moment, termed the reference and the pixel of the neighbor. From the gray scale values, GLCM is prepared. This takes into consideration the number of horizontally, vertically and diagonally neighboring pixels to a reference pixel with a certain gray-level (gray scale intensity or gray tone) values.

The directions of GLCM are:

1. Horizontal (0º)
2. Vertical (90º)
3. Diagonal

a. Bottom left to top right (-45º)

b. b)Top left to bottom right (-135º)

The directions of the GLCM are shown in Figure 7. The Figure 3.9 is an example of an image's GLCM.

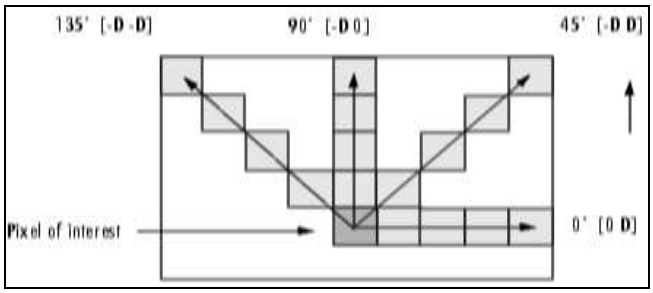

Fig 7: GLCM directions

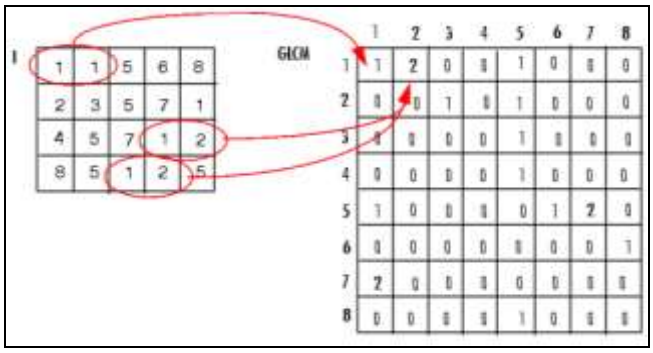

Fig 8: Gray Level Co-Occurrence Matrix showing the number of occurrences

The properties of GLCM are

- GLCM has the same number of columns and rows because of the similar range of values in the points of reference and adjacent pixels.
- The total number of columns and rows of the matrix is equal to the input image quantization plane.

For instance, a given test image comprises of four gray level intensities that is 0, 1, 2 and 3. Four-bit data contains $256=2^8$ probable values, 8x8 matrix would be computed, which will have 64 cells. A 16 gray-level image will have 65536x65536 matrixes, having 429, 496, 720, 3 cells. The co-occurrence matrix is consistent along crosswise. No gray-level distinction is found in the diagonal element set (0-0, 1-1, 2-2, 3-3). If the majority of the pixels match their adjacent cells, there will be very little contrast in the image. There is one-degree gray difference, if there is a variance of 1 cell away from the crosswise (0-1, 1-2, 1-3 etc). The disparity in gray levels will be greater if there is more distance from the diagonal.

Features Extraction

Features Extraction is known as the technique of extracting higher-level information from images. Texture is a key component of human visual perception. Analytic texture methods take into account the statistical variation of gray level values, by calculating localized characteristics on each and every location of the image and estimating a set of information from local characteristic distributions [37]. This approach is commonly used in the image analysis,

particularly in the biotech field. Feature mining involves two stages. In the first stage the GLCM is calculated, while in the second phase the GLCM-based texture features are defined.

The 6 texture descriptors are determined. The following are the details of these features. The number of gray levels is Ng, pd is a standardized symmetric GLCM of size Ng x Ng, and Pd (i, j) is the normalized GLCM (i j) th element.

The difference in intensity or gray level between the reference pixel, and the adjacent pixel, gives us contrast. The intensity differences in GLCM will be greater if the contrast is high:

$$Contrast = \sum i \sum j (i-j)^2 \, pd(i,j)$$

$$Homogeneity = \sum i \sum j \frac{1}{(1+(i-j)^2 \, pd(i,j)}$$

Energy is obtained from the Angular Second Moment (ASM).The ASM measures the adjacent gray levels ' uniformity.

The ASM value is higher if the pixels are highly identical. Consider

$$Energy = \sqrt{ASM}$$

$$ASM = \sum i \sum j \, pd(i,j)^2$$

The variation of the gray level pixel pairs is Dissimilarity:

$$Dissimilarity = \sum i \sum j \, pd(i,j) |i-j|$$

Correlation feature illustrate the linear dependency of gray level values in the co-occurrence matrix:

$$Correlation = \sum i \sum j \, pd(i,j) \frac{(i-\mu y)}{\sigma x \sigma y}$$

Where, μy are the means and σx, σy are the standard deviations. They are mentioned as follows

$$\mu x = \sum i \sum j \, pd(i,j)$$

$$\mu y = \sum i \sum j \, pd(i,j)$$

$$\sigma x = \sqrt{\sum i \sum j (i-\mu x)^2 \, pd(i,j)}$$

$$\sigma y = \sqrt{\sum i \sum j (i-\mu y)^2 \, pd(i,j)}$$

Features Representation

Characteristics or attributes are qualities calculated from the mammography image scale. We have collected textural properties of a tumor crop image, as we outlined in the previous section. For the mammography picture description, we have picked 6 textural characteristics. In this study two key categories were chosen from the mammographic pictures (Normal (0) and Abnormal (1)). Therefore, the quantities for each mammographic picture are determined. In this scenario, the total number of samples from the data set comprises of two classes of 9400 crop files. The class attributes were supplied during the learning process, because we use a controlled model of training. Evaluation features samples not within the learning set of data are used to determine the classification performance of the proposed system. This is chosen randomly according to the position of the defined indiscriminate stream which induces arbitrary separation

Cancer	contrast	dissimilarity	Homogenity	energy	correlation	ASM
1	5.510459947	1.904653905	2.904653905	0.089142152	0.837014868	0.007946159
1	4.796607094	1.714508083	2.714508083	0.105659836	0.814739004	0.011164001
1	5.295209943	1.808309897	2.808309897	0.080834801	0.933120494	0.006534285
1	9.427106958	2.412500134	3.412500134	0.046037998	0.979390038	0.002119497
1	6.533792978	2.010523451	3.010523451	0.088442221	0.8840587	0.007822028
1	12.91808033	2.856460095	3.856460095	0.034568385	0.983160117	0.001194968
1	5.858749887	2.478907738	3.478907738	0.037804747	0.983581376	0.001429199
1	9.231243763	2.39018406	3.39018406	0.04655725	0.579943649	0.002167103
1	12.53940758	2.816383925	3.816383925	0.037099198	0.976744234	0.001376385
1	7.293340195	2.107094257	3.107094257	0.063555651	0.939352603	0.004039425
1	7.895219087	2.185612982	3.185612982	0.058259951	0.949968923	0.003394222
1	8.492851728	2.305316157	3.305316157	0.066652683	0.899534083	0.00444258
1	7.876848408	2.205615531	3.205615531	0.062416162	0.956301171	0.003895777
1	9.380803774	2.39018406	3.39018406	0.051135345	0.958328206	0.002614823
1	42.79700533	2.217409054	3.217409054	0.066600081	0.895224188	0.004435785
1	6.883425585	2.051437903	3.051437903	0.051579836	0.983709604	0.002660479

Classifiers Training and Testing

The classifiers were trained and evaluated using standard 70%-30% split, as suggested by Patricia S. Crowther [30]. The previously extracted features dataset is split into 70 percent training data and 30 percent test data. The training set contains 6595 instances while the testing set contains 2827 instances. The classifiers are then individually trained and tested using this training and testing set. The results of classification of each classifier are discussed below:

1. Effectiveness of Classifiers

In This section, we evaluate the effectiveness of all classifiers in terms of Accuracy, AUC, Training time, correctly classified instances and incorrectly classified instances:

Table 1: Classifiers Performance

Evaluation criteria	Classifiers							
	SVM	DT	KNN	LR	MLP	ABC	GBC	RF
Accuracy (%)	81	83	78	84	81	85	87	91
AUC (%)	78	70	79	92.3	87.2	92.3	93.4	93.3
Training time (sec)	0.9	0.05	0.07	0.01	0.2	2.06	1.2	0.4
Correctly classified instances	2345	2381	2350	2383	2412	2399	2408	2425
Incorrectly classified instances	482	446	477	444	415	428	419	402

In order to better measure the execution of classifiers, simulation miscue is also considered. To do so, we evaluate

the effectiveness of our classifiers in terms of:

- **Mean Absolute Error (MAE):** Measures how close forecasts or predictions are to the eventual outcomes.
- **Kappa statistic (KS):** Measures the inter-annotator agreements a chance-corrected measure of agreement between the classifications and the true Classes.
- **Root Mean Squared Error (RMSE):** It measures the average magnitude of the error.
- **F-Beta Score:** It is the weighted harmonic mean of precision and sensitivity, reaching its optimal value at 1 and its worst value at 0.
- **Hamming loss (HL):** Average of the minimum number of *errors* that could have transformed one wrong prediction into the other, between two sets of samples.
- **Matthews Correlation Coefficient (MCC):** Describing Confusion Matrix by a single number ranging from -1 to 1. Where 1 represents perfect prediction, 0 means no better than random and -1 shows total disagreement between prediction and observation.
- **Jaccard Index (JI):** Gives similarity score by comparing set of predicted labels for a sample to the corresponding set of labels in true labels.

The Table-2 provides the errors of individual classifier.

Table 2: Simulation and Training Error

Errors	Classifiers							
	SVM	DT	KNN	LR	MLP	ABC	GBC	RF
MAE	0.17	0.16	0.17	0.16	0.15	0.15	0.15	0.14
KS	0.66	0.68	0.66	0.69	0.71	0.70	0.70	0.72
RMSE	0.41	0.40	0.41	0.40	0.38	0.39	0.38	0.38
HL	0.17	0.16	0.17	0.16	0.15	0.15	0.15	0.14
MCC	0.67	0.68	0.67	0.70	0.71	0.71	0.71	0.72
JI	0.73	0.73	0.72	0.75	0.76	0.76	0.76	0.75

The Figure-9 provides a graphical representation of the errors.

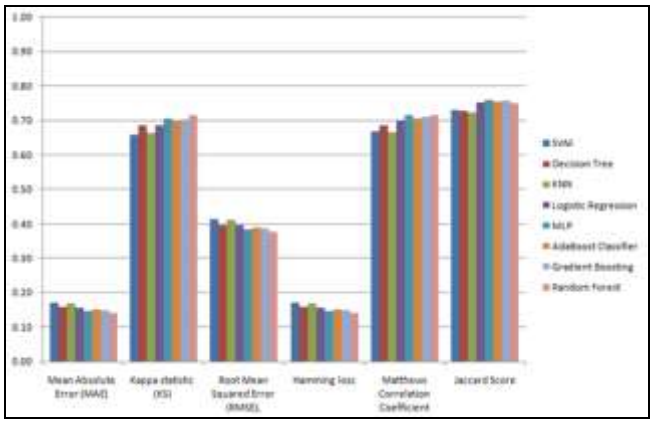

Fig 9: Comparative diagram evaluation criteria

2. Efficiency of Classifiers

Once the model is trained we can make predictions on it, we can check how efficient it is. For that, we compare the accuracy measures based on precision, Sensitivity, and F1-Score values:

Table 3: Accuracy Measures

Classifiers	Precision	Sensitivity	F1 Score
SVM	0.78	0.92	0.84
DT	0.85	0.83	0.84
KNN	0.81	0.88	0.84
LR	0.79	0.94	0.86
MLP	0.92	0.78	0.84
ABC	0.82	0.93	0.86
GBC	0.81	0.92	0.86
RF	0.87	0.85	0.86

Since Confusion matrices indicate a useful way for evaluating classifier, each row of Table 4 shows the number of instances in an actual class while each column shows predictions.

Table 4: Confusion Matrix

Classifiers	Tumor	Normal	Class
SVM	1305	118	Tumor
	364	1040	Normal
DT	1185	238	Tumor
	208	1196	Normal
KNN	1248	175	Tumor
	302	1102	Normal
LR	1342	81	Tumor
	363	1041	Normal
MLP	1322	101	Tumor
	314	1090	Normal
ABC	1323	100	Tumor
	328	1076	Normal
GBC	1341	117	Tumor
	308	1061	Normal
RF	1207	216	Tumor
	186	1218	Normal

The ROC plot gives better visualization of the skillfulness of the classifiers. The Figure-10 shows the ROC of the classifiers:

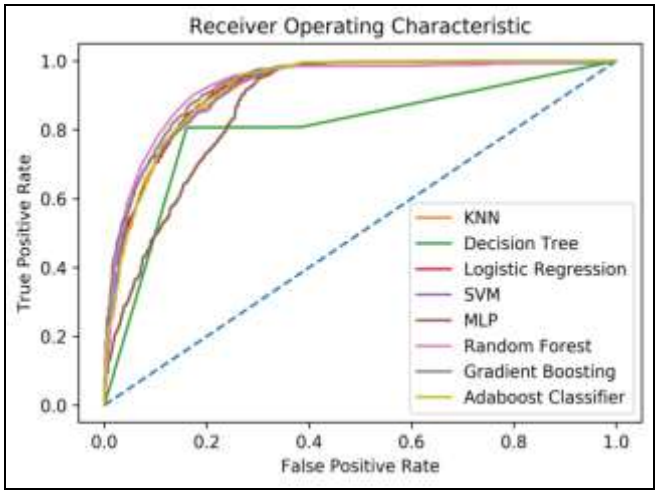

Fig 10: ROC of Classifiers

Discussion and Conclusion

Machine learning Algorithms have very high capability of

Classification and it can evidently bypass the need for human intervention in classification. The breast cancer tumor can easily be detected by Machine learning classifiers. Each algorithm has certain limitations and cannot be used for the overall detection. The results indicates that Random Forest have the highest 91% accuracy and lowest accuracy of KNN and SVM. However, the accuracy of individual algorithms is between 75% and 91%, but their amalgamation and voting system can be used to accurately detect the tumor in mammogram. The amalgamation of even number algorithms can have the problem of 50% votes, which will raise the problem of random case. The system can be further improved by introducing Deep Learning in ensemble and the prediction have certain weights based on their effectiveness.

Conflicts of Interest

The authors have no conflicts of interest.

References

1. US Cancer Statistics Working Group. United States cancer statistics: 1999–2011 incidence and mortality web-based report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute, 2014.
2. Parkin DM, Pisani P, Ferlay J. Global cancer statistics. CA: a cancer journal for clinicians. 1999; 49(1):33-64.
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018; 68(6):394-424.
4. Nitu R, Rogobete AF, Gundogdu F, Tanasescu S, Boruga O, Sas A, *et al*. microRNAs expression as novel genetic biomarker for early prediction and continuous monitoring in pulmonary cancer. *Biochemical genetics*. 2017; 55(4):281-290.
5. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine learning*. 2002; 46(1-3):389-422.
6. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, *et al*. Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry. 2005; 29(1):37-46.
7. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences. 2001; 98(26):15149-15154.
8. Asri H, Mousannif H, Al Moatassime H, Noel T. June. Big data in healthcare: Challenges and opportunities. In International Conference on Cloud Technologies and Applications (CloudTech) (pp. 1-7). IEEE, 2015.
9. Pradhan A. Support vector machine-A survey. International Journal of Emerging Technology and Advanced Engineering. 2012; 2(8):82-85.
10. Quinlan JR. C4. 5: programs for machine learning. Elsevier, 2014.
11. Larose DT, Larose CD. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, 2014.
12. Haydon MD, Stanton AL, Ganz PA, Bower JE. Goal disturbance in early-stage breast cancer survivors. Journal of psychosocial oncology. 2019; 37(4):478-493.
13. Salman NH, Ali SIM. Breast Cancer Classification as Malignant or Benign based on Texture Features using Multilayer Perceptron. International Journal of Simulation--Systems, Science & Technology, 2019; 20(1).
14. Huang Q, Chen Y, Liu L, Tao D, Li X. On combining biclustering mining and AdaBoost for breast tumor classification. IEEE Transactions on Knowledge and Data Engineering, 2019.
15. Rao H, Shi X, Rodrigue AK, Feng J, Xia Y, Elhoseny M, *et al*. Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing. 2019; 74:634-642.
16. Nicolo C, Perier C, Prague M, MacGrogan G, Saut O, Benzekry S, *et al*. Machine learning versus mechanistic modeling for prediction of metastatic relapse in breast cancer. Bio Rxiv, 2019, p.634428.
17. Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 20, 2019.
18. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, *et al*. Top 10 algorithms in data mining. Knowledge and information systems. 2008; 14(1):1-37.
19. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013; 4(124):3.
20. Nematzadeh Z, Ibrahim R, Selamat A. May. Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. In 2015 10th Asian Control conference (ASCC) (pp. 1-6). IEEE, 2015.
21. Hasan H, Tahir NM. May. Feature selection of breast cancer based on principal component analysis. In 2010 6th International Colloquium on Signal Processing & its Applications (pp. 1-4). IEEE, 2010.
22. Ojha U, Goel S. January. A study on prediction of breast cancer recurrence using data mining techniques. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 527-530). IEEE, 2017.
23. Ghosh S, Mondal S, Ghosh B. February. A comparative study of breast cancer detection based on SVM and MLP BPN classifier. In First International Conference on Automation, Control, Energy and Systems (ACES) (pp. 1-4). IEEE, 2014.
24. Osareh A, Shadgar B. April. Machine learning techniques to diagnose breast cancer. In 2010 5th International Symposium on Health Informatics and Bioinformatics (pp. 114-120). IEEE, 2010.
25. Bazazeh D, Shubair R. December. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA) (pp. 1-4). IEEE, 2016.
26. Azmi MSBM, Cob ZC. December. Breast cancer prediction based on backpropagation algorithm. In 2010 IEEE Student Conference on Research and Development (SCOReD) (pp. 164-168). IEEE, 2010.

27. Gayathri BM, Sumathi CP. December. Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-5). IEEE, 2016.
28. Sonavane R, Sonar P, Sutar S. May. Classification of MRI brain tumor and mammogram images using learning vector quantization neural network. In 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS) (pp. 301-307). IEEE, 2017.
29. Yoon S, Kim S. November. AdaBoost-based multiple SVM-RFE for classification of mammograms in DDSM. In 2008 IEEE International Conference on Bioinformatics and Biomeidcine Workshops (pp. 75-82). IEEE, 2008.
30. Crowther PS, Cox RJ. September. A method for optimal division of data sets for use in neural networks. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 1-7). Springer, Berlin, Heidelberg, 2005.