Received: November 15, 2023. Revised: January 13, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Mammography-based Computer-Aided Diagnostics for the Identification of Breast Cancer Based on Machine Learning

Nagajyothi Dimmita[1] Vanga Nagasri[2] Koppaka Achutha Jyotsna[3] Pochaboina Swapna[4] Narne Srikanth[5] Pattem Sampath Kumar[6] Atheeswaran Athiraja[7]* Gunaganti Sravanthi[8] Rajeswaran Nagalingam[6]

[1]Department of Electronics and Communication Engineering, Vardhaman College of Engineering, Shamshabad, Hyderabad, Telangana State, India
[2]Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, Telangana State India
[3]Department of Electronics and Communication Engineering, CVR College of Engineering, Hyderabad, Telangana State, India
[4]Department of IT, Malla Reddy Engineering College, Secunderabad, Telangana State, India
[5]Department of Computer Science and Engineering (AI & ML), RVR & JC Collegge of Engineering, Guntur, Andhra Pradesh, India
[6]Department of Electronics and Communication Engineering, Malla Reddy College of Engineering, Secunderabad, Telangana State, India
[7]Department of Computer Science and Engineering (AI&ML), Bannari Amman Institute of Technology, Satyamangalam, Tamilnadu India
[8]Department of Computer Science and Engineering, Malla Reddy Institute of Engineering and Technology, Secunderabad, Telangana State, India.
* Corresponding author's Email: a.athiraja4@gmail.com

Abstract: Breast cancer is the most typical cancer overall and the one that affects women the most frequently. In 2022, there were 2.26 million or more brand-new cases of breast cancer in women. To identify the breast cancer in early stage, it's easy to cure. The apparent structure is examined using computer-aided diagnostic (CAD) technologies. The majority of existing diagnostic approaches rely on mammography mass features; however, the suggested strategy relies on MicroCalcification (MC) characteristics. In this research cancer cells are detected utilizing mammography images, pre-processing methods, segmentation, multi-layer perceptron with artificial algae algorithm and receiver operating characteristics (ROC) curve analysis. The national cancer institute (NCI) of provides mammogram scans. Normalization and median filtering were used as part of the pre-processing procedure. Segmentation is used to detect the location of MicroCalcification present cells using the local threshold approach. The MicroCalcification present cells and MicroCalcification missing cells are classified using the multi-layer perceptron (MPL) with artificial algae algorithm technique and also used to divide MicroCalcification-affected cells into the following categories: initial, very small, small, medium, high, and very high. ROC curve analysis was used to assess system performance. According to experimental findings, for the NCI, University of California Irvine (UCI), nathan kline institute (NKI), investigation of serial studies to predict your 1 (ISPY1), ISPY2, ISPY3, and ISPY4 datasets, the classification using the multi-layer perceptron with artificial algae method has the best accuracy of 96 percent when compared to random forest (RF), MPL with genetic algorithm (GA), hierarchical clustering random forest (HCRF), and convolutional neural network (CNN).

Keywords: CAD, MC, ROC, Multi-layer perceptron with artificial algae algorithm, Mammogram.

1. Introduction

The most prevalent disease in the world, cancer, affects everyone. The most common kind of cancer in women is specifically breast cancer [1]. As a result, any advancement in the diagnosis and prognosis of cancer sickness is crucial for maintaining good health. The early detection and prediction of cancer can both benefit greatly from machine learning approaches. Breast cancer is a cancer that develops in the breast tissue, most usually in the inner lining of milk ducts or the lobules that feed milk to the ducts [2]. Cancers that start in ducts are called ductal carcinomas, whereas cancers that start in lobules are called lobular carcinomas. Humans and other species both get breast cancer. While female breast cancer accounts for the vast majority of human occurrences, male breast cancer can also develop [3]. The visual interpretation of histopathological pictures is the gold standard for identifying breast cancer, but it is a challenging procedure that needs years of training and a high level of pathologist expertise. Sometimes it is impossible to diagnose breast cancer because of the visual interpretation's limitations and a lack of experience. Therefore, the computer-aided diagnostic (CAD) system may be considered as a helpful instrument to lower the inaccuracy of breast cancer diagnosis.In this research work we utilized the machine learning algorithm and microcalcification parameter to improve the accuracy of the proposed system.

2. Related work

According to a novel method presented by Mohiuddin Ahmed [4], breast cancer may be identified from histological scans of breast tissues using convolutional neural networks. Pathologists evaluate the histopathological pictures of the breast tissue at various magnification levels as part of the clinical procedure for diagnosing breast cancer. A single convolutional neural network (CNN) model is employed in this study to simultaneously accept input from the same image at four distinct magnification levels. The suggested method significantly outperforms current state-of-the-art techniques. In recent years, the use of random forest (RF) has become widespread as a machine learning technology that may be used to accurately diagnose a variety of illnesses [5]. However, throughout the training process, decision trees with low classification performance and high similarity may be produced, which has an impact on the model's overall classification performance. A model known as hierarchical clustering random forest (HCRF) is created [6]. The hierarchical clustering approach is used to perform clustering analysis on decision trees by evaluating the similarity between all of the decision trees [7]. In order to build the hierarchical clustering random forest with low similarity and high accuracy, sample trees are chosen from split clusters. Additionally, we optimize the chosen feature number for the prediction of breast cancer using the variable importance measure (VIM) technique. In this work, two databases from the UCI (University of California Irvine) machine learning repository were used: Wisconsin diagnosis breast cancer (WDBC) and wisconsin breast cancer (WBC). Accuracy, precision, sensitivity, specificity, and AUC (Area under ROC Curve) [8] are used to assess the performance of the suggested approach.

The identification of traits is a crucial stage in developing the breast cancer classifier for early diagnosis [9]. A collection of actual breast cancer cases typically contains both discrete and continuous data. In such a medical field, more focus is placed on the receiver operating characteristic's area under the curve. There is not enough research available right now to take into account both the hybrid feature characteristic and the particular categorization goal. For the feature selection of breast cancer datasets, Q. Wuniri [10] suggested a wrapper technique, or integrated framework, that incorporates Bayesian classifiers. They handle both discrete and continuous features, handling discrete features with the naive method for discrete features and continuous features with the kernel probability density estimation; respectively. The result of this is feature-type-aware hybrid Bayesian classifiers. The feature subsets that each classifier is given vary, and the AUC values are used as fitness indices to gauge how well each classifier performs. As a result, utilising the evolutionary technique, we may produce a feature subset that is almost optimal and develop classifiers with high AUC values. Additionally, the one-class F-score is applied to improve convergence of the method. Both the real breast cancer dataset for Chinese women and the continuous Wisconsin diagnostic breast cancer dataset were used in the experiments.

Increasing the number of data sets used is the main objective of the proposed system in order to improve the accuracy of the machine learning process. The present methods for diagnosing breast cancer have several parameters; in our suggested method, we used the MicroCalcification parameter. It provides better results when compared to existing techniques with various conditions. The majority of currently used methods just categorise items as

normal or abnormal; however, this research endeavour further divides the abnormal category into six additional categories. The confusion matrix is used to assess the proposed system. In this work, we employ the Multi-layer Perceptron with artificial algae technique as a supervised classification strategy. There are two varieties of CAD used in mammography [11].

(i)CADe (computer aided detection).
(ii)CADx (computer aided diagnosis).

Based on the features of MicroCalcifications, these two approaches are put into practise. Microcalcifications, which appear as bright spots on mammograms, are microscopic calcium deposits on human cells [12]. A mammogram's is pictorial representation of breast MicroCalcification cells are detected using a Multi-layer Perceptron with artificial algae algorithm [13]. We used threshold-based segmentation method to determine location of MicroCalcification cells. The current MicroCalcification cell was further categorised using the case based reasoning (CBR) method. The present investigation diagnoses cancer cells using microcalcifications found in mammography pictures. [14]. Using the multi-layer perceptron with artificial algae algorithm, categories MC present cells and MC absence cells [15]. The initial, very tiny, small, medium, high, and very high classes are then separated out of the MC present cell. ROC curve analysis is a tool for evaluating system effectiveness. [16].

3. Design and methodology

To diagnosis the breast cancer at early stage, we proposed method to diagnosis the cancer based on MicroCalcification characteristics. The major five modules of computer aided diagnosis of breast cancer.

- Mammogram image
- Soft coring filter
- Segmentation
- Multi-layer perceptron with artificial algae algorithm
- ROC curve analysis

3.1 Mammography image

Mammography is a diagnostic and screening procedure that uses low-energy X-rays (typically approximately 30 kVp) to inspect the human breast [17]. Mammography is used to diagnose breast cancer in its early stages, usually by detecting

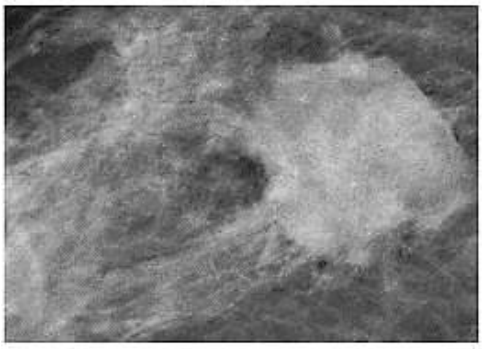

Figure. 1 Mammography image

distinctive lumps and/or MicroCalcifications. The picture below is an example of a mammography image as shown in Fig. 1. An x-ray image of the breast is called a mammogram. Breast cancer screening mammography are used to look for the illness in women who have no signs or symptoms. After a lump or other sign or symptom of breast cancer has been discovered, diagnostic mammography is performed to check for the illness [18]. Mammography screening can help lower the number of breast cancer deaths in women aged 40 to 70. False-negative findings, false-positive results, overdiagnosis, overtreatment, and radiation exposure are all possible side effects of screening mammography. Women over the age of 40 should get screening mammograms every 1 to 2 years, according to the national cancer institute. The data set was gathered by the national cancer institute in the United States and can be downloaded at http://www.cancer.gov/statistics/tools. In this research work we are using image width as 650 pixel and height as 420 pixels.

3.2 Preprocessing

Pre-processing is a method used to remove the noise from the input image. In this research work, normalization method and soft coring filter were used. Normalization method was used to convert the RGB image into grey scale image using the following Eq. (1) [19].

$$I(x, y, z) = \frac{(Gx+Gy+Gz)}{3} \quad (1)$$

A nonlinear technique called soft coring filtering is used to remove extraneous data from normalised images. Based on kernel functions that can be executed in the frequency domain, a gaussian high pass filter is employed to graphically depict the data [20]. With the aid of the sliding windowing technique, the Gaussian high pass filter quickly produces a Fourier transform in two dimensions of

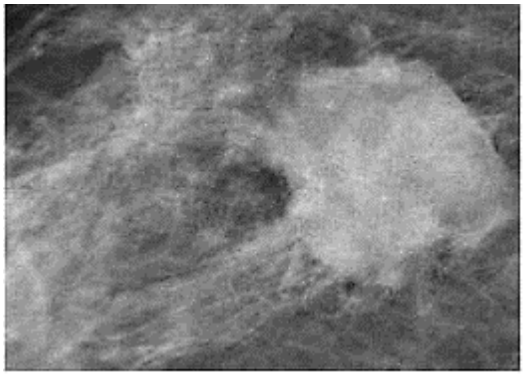

Figure. 2 Preprocessing output

convolutions. The output of the soft coring function was applied after the normalised output picture had been passed through the high pass filter α (.).

$$P(x,y) = Ih(x,y) + \alpha\big(I(x,y)\big) \quad (2)$$

Where $P(x,y) -$ Preprocessed output image
$Ih(x,y) -$ Highpassfiteroutputimage

$$Ih(x,y) = I(x,y) - Z\big(e^{jwx}, e^{jwy}\big) \quad (3)$$

$Z(e^{jwx}, e^{jwy})$- High pass filter co-efficient
$\alpha(I(x,y))$- Soft coring function

$$\alpha\big(I(x,y)\big) = m\,.\,I(x,y)(1 - e^{\left|\frac{I(x,y)}{\tau}\right|} \quad (4)$$

m, τ – random variables ranges between 0 to 1.

The noise in the input image was eliminated using a Gaussian high pass filter. To extract the input image's line and edge information, the soft coring kernel function was employed. Two-dimensional pictures were filtered using the soft coring approach, which had significantly less data loss than median filtering [21-22].

As demonstrated in Fig. 2, a two-step pre-processing strategy improves the image's quality while reducing processing time, compensating for lighting, reducing the shaded backdrop, and maintaining image contrast and brightness.

3.3 Segmentation

Threshold based segmentation method used to classify the MC present cell and MC absent cell [23]. Based on the pixel colour value the cancer cells are segmented. Colour feature is extracted by using the following Eq. (5).

$$E(x,y) = \begin{cases} \text{oif} P(x,y) < T1 \\ 1 \text{ if} T1 \le P(x,y) \le T2 \\ 0 \text{ if} P(x,y) > T2 \end{cases} \quad (5)$$

Where,
T1 –Lower threshold value
T2- Upper threshold value.

4. Materials and methods

4.1 Multi-layer perceptron with artificial algae algorithm

Each algal colony provides a potential solution (i.e. MLP), which is represented as a vector in the proposed model based on AAA [24]. Three components make up the potential solution vector. The weight values from the input layer to the hidden layer are shown in the first section, the bias values of the nodes are shown in the second section, and the weight values from the hidden layer to the output layer are shown in the third section [25]. The whole weight and bias number in the network are represented by the length of the vectors.

Fig. 3 depicts the vectorized MLP, which has n input nodes, h hidden nodes, and m output nodes. The weights from the input to the hidden layer are represented by red arrows, the weights from the hidden layer to the output layer by green arrows, and the bias values are represented by blue arrows. The solution to the vector length problem is shown in Eq. (6).

$$\text{VectorLenth} = (n * h) + (h * m) + h + m \quad (6)$$

In this instance, n denotes the quantity of input nodes, h the quantity of hidden nodes, and m the quantity of output nodes. The fitness values of the generated vectors were calculated using the mean square error (MSE) tool [26]. The square of the many calculations between the actual and anticipated values is the foundation of the MSE. The speed of the MLA with AAA is very faster and accuracy of the proposed system also very heigh because n number of hidden layers used depends on the input image.

4.2 Algorithm

Step 1. Set the weights and the threshold. Note that weights can be set to 0 or a tiny random value by setting each weight node wi(0) to 0. We'll use the former in the example below.

Step 2. Perform the following steps over the input Xj and desired output dj for each sample j in our training set D.:

2a. Calculate the actual output using Eq. (7) :

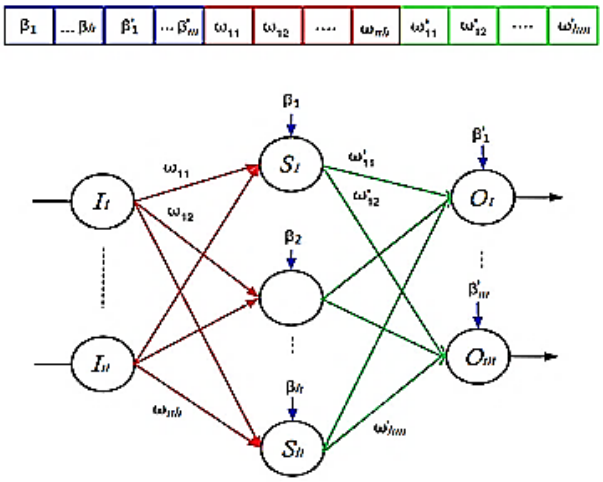

Figure. 3 MLP with AAA

$$Yj(t) = f[w(t).Xj] = f[w0 * (t) + w1(t) * xj,1 + w2(t) * xj,2 + \cdots + wn(t) * xj,n] \quad (7)$$

w(t) – weight value with respect to time

Xj – Input image

2b. Adapt weights calculated using Eq. (8):

$$wi(t+1) = wi(t) + \propto * \left(dj - yj(t)\right) * xj, \quad \text{iforallnodes } 0 \le i \le n \quad (8)$$

wi(t+1) – New wright value

$\propto$ – learning rate

The user-specified error threshold is reached or the predetermined numbers of repetitions have been reached before moving on to Step 2 [27]. The algorithm updates the weights immediately rather than waiting until all couples in the training set have used steps 2a and 2b [28-29]. This is a crucial distinction to make. Fig. 4 displays the flowchart for the algorithm.

4.3 Execution of the multi-layer perceptron with artificial algae algorithm following steps

Step 1: The mammography picture is first transformed into a two-column vector in step one (x1, x2).

Step 2: Based on the following premise, the value of two columns is converted to a binary value. '1' represents the cells with existing MicroCalcification if (x1 and x2) are more than 180. '-1' stands for the absence of cells during MicroCalcification.

Step 3: In order to determine the target value indicated in the following Table 1.

Here, the target value is determined by employing the OR logic function.

Step 4: Applying the equation, determine the net output value (8).

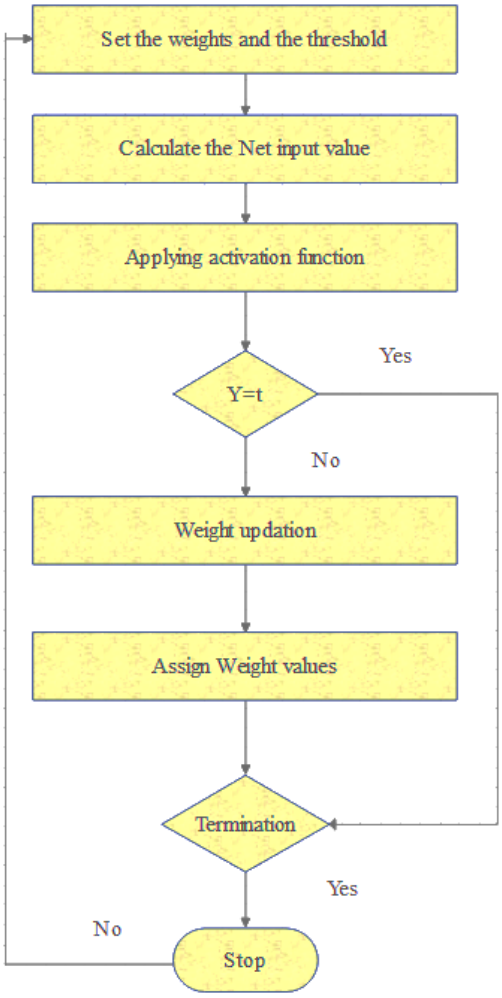

Figure. 4 Flow chart for perceptron algorithm

Table 1. OR logic function

X1	X2	T
1	1	1
1	-1	1
-1	1	1
-1	-1	-1

Step 5: Step 4 is carried out till the condition of termination. These requirements are as follows:

• There is no weight change; Actual value always equals desired value.

Step 6: Sort the cells into those that have MicroCalcification and those that don't, as indicated in Fig. 4.

5. Results and discussion

To diagnosis the cancer is very sensitive because if we provide false report, it gives wrong perception to the medical experts. This application is used to diagnosis the cancer at early stage so that we evaluate our proposed system with various aspects

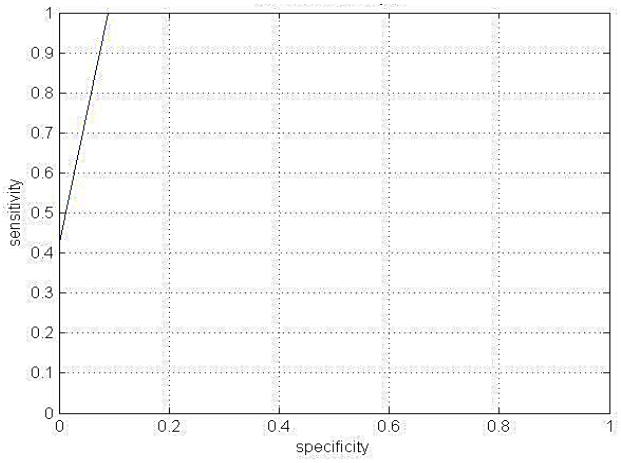

Figure. 5 ROC Curve for breast cancer diagnosis

Table 2. Accuracy value for various databases

Database	MLP with AAA	CNN	HCRF	MLP with GA
NCI	97	92	81	78
UCI	96	93	85	79
NKI	93	91	82	78
ISPY1	94	92	79	81
ISPY2	97	91	78	80
ISPY3	92	90	81	79
ISPY4	98	91	82	78

like accuracy, specificity, sensitivity and F1 score value. In this section, we are discussing ROC curve analysis, true positive rate, false positive rate, accuracy, specificity, sensitivity, F1 score value and error rate.

5.1 ROC curve

A binary classifier system's performance as its discrimination threshold is changed is graphically represented by a receiver operating characteristic (ROC) curve, or simply ROC curve [30]. It is created by plotting, at various thresholds, the percentage of true positives out of positives (TPR = true positive rate) with the percentage of false positives out of negatives (FPR = false positive rate) [31]. TPR stands for sensitivity, and FPR, or true negative rate, stands for one less than specificity.

The ROC is often referred to as a relative operating characteristic curve [32] because it

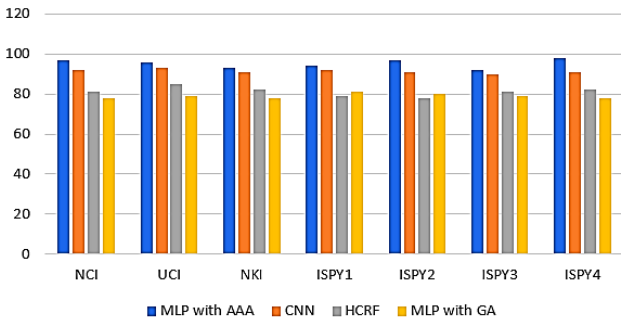

Figure. 6 Cancer diagnosis accuracy value for various databases

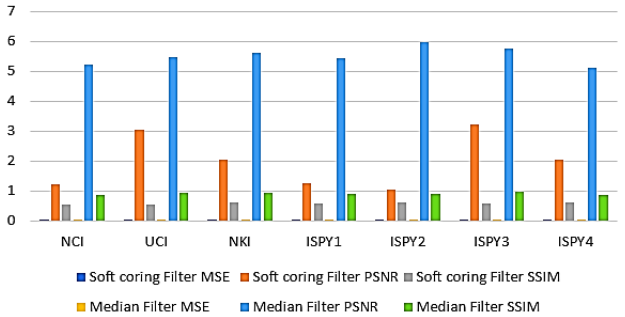

Figure. 7 Performance analysis filter using MSE, PSNR & SSIM

contrasts two operating characteristics (TPR and FPR) when the criteria vary. We are using the different data sets like NCI, UCI, NKI, ISPY1, ISPY2, ISPY3 and ISPY4.

Each data set contains 100 mammogram images, in that 60 images used for training data sets and remaining used for testing data set. The output of the roc curve is depicted in Fig. 5 of the graph below. To determine the accuracy score for every image (shown in Fig. 6), similarly. It is illustrated in the following Table 2.

The Fig. 6 indicates that our proposed methods having better diagnosis. The area under curve of the proposed system is more than 95 % it is shown in the above figure. The proposed algorithms are applied through the various data sets, the result an analysis of the system evaluated by accuracy value. It can be calculated with aid of true positive rate and false positive rate. Confusion matrix used to calculate the true positive, false positive, true negative and false negative [33]. We were used different four algorithm used to diagnosis the breast cancer.

In Table 3, the performance of median filtering technique based on the MSE, PSNR and SSIM may be inferred from this graph (Shown in Fig. 7). It indicates that's low MSE and SSIM and high PSNR value.

The average accuracy value for the proposed

Table 3. Performance analysis filter using MSE, PSNR & SSIM

Data Set	Soft coring Filter			Median Filter		
	MSE	PSNR	SSIM	MSE	PSNR	SSIM
NCI	0.01452	1.234	0.5264	0.02345	5.236	0.8569
UCI	0.01245	3.032	0.5267	0.03125	5.479	0.9214
NKI	0.024785	2.031	0.6214	0.03214	5.627	0.9457
ISPY1	0.021456	1.245	0.5867	0.03546	5.421	0.8964
ISPY2	0.012456	1.025	0.6223	0.03424	5.987	0.9123
ISPY3	0.014856	3.231	0.5623	0.03278	5.748	0.9746
ISPY4	0.012458	2.034	0.5978	0.03892	5.124	0.8597

algorithms MLP with AAA, CNN, HCRF and M LP with GA respective percentage is 96%, 93%, 85% and79%.

We are considering the median filter and soft coring fitter for the result analysis. Filtering method is important module of the proposed system, performance analysis done by using MSE, PSNR and SSIM parameters [34-36]. When compare to the median filter and soft coring filtering method, soft coring filter provide the better performance. For NCI data sets soft coring filter MSE, PSNR and SSIM respectively 0.01452, 1.234 and 0.5264 and similar to the median filer values respectively 0.02345, 5.236 and 0.8569.

The evaluation of our proposed algorithms' performance, including MLP with AAA, CNN, HCRF, MLP with GA, and RF.The following characteristics were used to evaluate the classification method's performance: sensitivity, specificity, F1 score, and accuracy. The average sensitivity value for MLP with AAA, CNN, HCRF, MLP with GA and RF respectively 95, 91, 81,78 and 74. The specificity value for MLP with AAA, CNN, HCRF, MLP with GA and RF respectively 94, 90, 82,76 and 75. The F1 score value for MLP with AAA, CNN, HCRF, MLP with GA and RF respectively 96, 92, 84,75 and 78. The accuracy value for MLP with AAA, CNN, HCRF, MLP with GA and RF respectively 95, 91, 84,75 and 73.We inferred from Table 4 and Fig. 8 that MLP+AAA has greater sensitivity, specificity, F1 score, and accuracy values of 95 percent.

To assess the suggested system according to its execution time. We are used CPU based computer to run our application. We were used different data sets

Table 4. Performance analysis of classification method

Classification Techniques	SEN %	SPEC	F1 Score	Acc %
RF	74	75	78	73
MLP+GA	78	76	75	75
HCRF	81	82	85	84
CNN	91	90	92	91
MLP+AAA	95	94	96	95

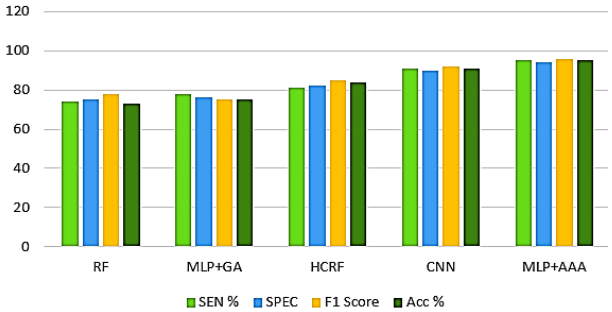

Figure. 8 Performance analysis of classification method

Table 5. Performance analysis based on execution time

Data set	Execution Time (seconds)				
	RF	MLP+GA	HCRF	CNN	MLP+AAA
NCI	190	185	178	176	112
UCI	192	184	182	165	132
NKI	192	186	179	163	124
ISPY1	191	183	185	162	115
ISPY2	196	185	176	168	116
ISPY3	193	186	175	167	118
ISPY4	192	183	179	165	113

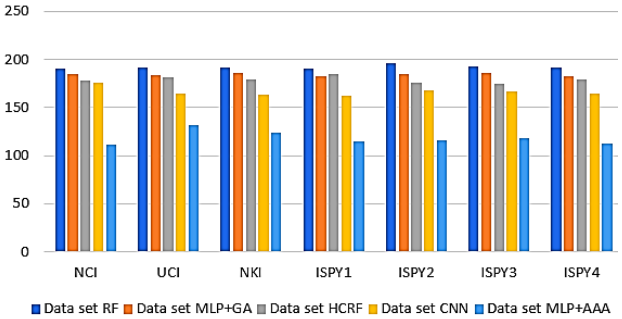

Figure. 9 Performance analysis based on execution time

Table 6. Performance analysis based on error rate

Data set	Error Rate				
	RF	MLP+GA	HCRF	CNN	MLP+AAA
NCI	0.321	0.213	0.152	0.052	0.016
UCI	0.345	0.265	0.164	0.058	0.012
NKI	0.328	0.248	0.174	0.056	0.017
ISPY1	0.324	0.218	0.191	0.057	0.011
ISPY2	0.389	0.298	0.182	0.053	0.015
ISPY3	0.347	0.276	0.172	0.057	0.013
ISPY4	0.365	0.256	0.163	0.051	0.014

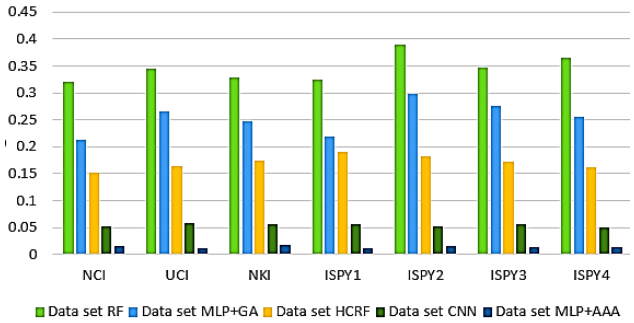

Figure. 10 Performance analysis based on error rate

such as NCI, UCI, NKI, ISPY1, ISPY2, ISPY3 and ISPY4. To apply our proposed algorithms like MLP with AAA, CNN, HCRF, MLP with GA and RF to different data sets [38-41]. Our suggested algorithms and the data sets they represent have varying execution times based on the quantity and quality of our photographs. In terms of execution time, Table 5 and Fig.9 show the overall performance of many classification algorithms used for face recognition classification. When compared to other categorization algorithms, MLP+AAA required less time to perform. The overall effectiveness of several classification techniques used for categorization of data sets is shown in Table 6 and Fig. 10. In a comparison of classification algorithms, MLP+AAA had the lowest error rate of all the systems tested. Our technique is used for detecting MicroCalcification in mammograms.

Using the MLP with AAA, locate MicroCalcification cells and MicroCalcification absence cells. This approach makes it simple to distinguish between MicroCalcification-positive and MicroCalcification-negative cells. The MLP with AAA is a supervised learning classification approach in which the output value is anticipated, eliminating the possibility of misclassifying the MicroCalcification present and missing cells. Because the MLP with AAA learning process is a supervised machine learning approach, the actual and expected outputs were always the same. Based on the MicroCalcification analysis, if the MicroCalcification was found in the mammography, the patient was diagnosed with cancer. MLP with AAA was used to divide MC present cells into the following categories: beginning, small, medium, high, and very high. It is impossible to anticipate whether or not a patient will be afflicted by cancer. If a patient is diagnosed with cancer, antibiotics or surgery may be recommended.

6. Conclusion

A lot of effort has gone towards computerized CADx approaches to help radiologists distinguish between benign and malignant MCs, which is a challenging decision. It has been demonstrated that a CADx method may accurately diagnose clustered MCs even better than radiologists. MC images that were grouped were utilized to train a feed forward neural network (FFNN), which was utilized in this technique. The ability of more recent machine learning methods, including SVM, Kernel fisher discriminant (KFD), RVM, and committee machines (including ensemble averaging and the well-known boosting method Adaboost), to categories MC clusters as cancerous or benign was examined. We chose characteristics from those have been shown to qualitatively reflect radiologists' considerations. These features are determined by the size, shape, and distribution of each individual MC within the cluster. In terms of the area under the receiver-operating characteristic curve, the assessment research discovered that the SVM, KFD, and RVM kernel techniques perform equally well, but statistically they all exceed FFNN or AdaBoost.The proposed system's limitation is that performance accuracy will automatically increase but time complexity will increase if the training data set is increased to more than 10,000 data sets. To maintain the large number of databases, more space is required. To overcome these limitations in this paper we recommended for use GPU (Graphical Processor Unit) system.

List of abbreviations

CAD- computer-aided diagnostic
MC- MicroCalcification
ROC-receiver operating characteristics

MLP+AAA- multi-layer perceptron with artificial algae algorithm technique
NCI- national cancer institute
CNN-convolutional neural network
RF-random forest
HCRF-hierarchical clustering random forest
VIM-variable importance measure
UCI -University of California Irvine
WDBC -wisconsin diagnosis breast cancer
WBC-wisconsin breast cancer
CADe -computer aided detection
CADx -computer aided diagnosis
FFNN-feed forward neural network
KFD-Kernel Fisher discriminant
SVM -support vector machine
RVM -relevance vector machine
TPR = true positive rate
FPR = false positive rate
MSE-mean square error
SSIM -structured similarity indexing method
GA- genetic algorithm
PSNR- peak signal to noise ratio

Ethics approval and consent to participate

Not applicable.

Human and animal rights

Nohumans and animals were used in this study.

Availability of data and materials

The data set was gathered by the national cancer institute in the United States and can be downloaded at http://www.cancer.gov/statistics/tools. All data generated of analyzed during this study are included in this published article.

Funding

No funding involved in this work.

Conflicts of interest

“The authors declare no conflict of interest”.

Author contributions

Conceptualization, ND and VN; methodology, A. A; software, KAJ; validation, P.SK, KAJ, and ND; formal analysis, RN; investigation, GS; resources, VN; data curation, NS; writing—original draft preparation, AA; writing—review and editing, ND; visualization, NS; supervision, RN; project administration, KAJ; funding acquisition, PSK.

References

[1] H. Zhang, L. Guo, D. Wang, J. Wang. L. Bao, S. Ying, H. Xu, and J. Shi, “Multi-Source Transfer Learning Via Multi-Kernel Support Vector Machine Plus for B-Mode Ultrasound-Based Computer-Aided Diagnosis of Liver Cancers”, *IEEE Journal of Biomedical and Health Informatics*, Vol. 10, No. 10, pp. 3874-3885, 2021, 10.1109/JBHI.2021.3073812.

[2] C. Chen, Y. Wang, J. Niu, X. Liu, Q. Li, and X. Gong, "Domain Knowledge Powered Deep Learning for Breast Cancer Diagnosis Based on Contrast-Enhanced Ultrasound Videos", *IEEE Transactions on Medical Imaging*, Vol. 40, No. 9, pp. 2439-2451, 2021, doi: 10.1109/TMI.2021.3078370.

[3] Y. Wang, Z. Ma, K. C. Wong, and X. Li, "Evolving Multi objective Cancer Subtype Diagnosis From Cancer Gene Expression Data", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 18, No. 6, pp. 2431-2444, 2021, doi: 10.1109/TCBB.2020.2974953.

[4] M. Ahmed and M. R. Islam, "A Multiple-Input Based Convolutional Neural Network in Breast Cancer Classification from Histopathological Images", In: *Proc. of 2021 24th International Conference on Computer and Information Technology (ICCIT)*, pp. 1-5, 2021, doi: 10.1109/ICCIT54785.2021.9689856.

[5] X. Tang, L. Cai, Y. Meng, C. Gu, J. Yang, and J. Yang, "A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis with Transcriptome and Functional Proteomic", *IEEE Access*, Vol. 9, No. 1, pp. 51659-51668, 2021, doi: 10.1109/ACCESS.2021.3070428.

[6] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm", *IEEE Access*, Vol. 10, No.1, pp. 3284-3293, 2022, doi: 10.1109/ACCESS.2021.3139595.

[7] M. Li, X. Ma, C. Chen, Y. Yushuai, S. Zhang, Z. Yan, C. Cheng. C. Fangfang, Y. Bai, X. Yujie P. Zhou, and M. Mingrui “Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathological Images”, *IEEE Access*, Vol. 9, No. 1, pp. 53687-53707, 2021, doi: 10.1109/ACCESS.2021.3071057.

[8] A. Davri, E. Birbas, T. Kanavos, G. Ntritsos, N. Giannakeas, A. T. Tzallas, and A. Batistatou, “Deep Learning on Histopathological Images

for Colorectal Cancer Diagnosis: A Systematic Review", *Diagnostics*, Vol. 12, No. 1, pp. 837-879. 2022, https://doi.org/10.3390/diagnostics12040837.

[9] O. Iizuka, F. Osamu, K. Kei, R. Michael, A. Koji, and T. Masayuki. "Deep learning models for histopathological classification of gastric and colonic epithelial tumours", *Scientific Reports*, Vol. 10, No. 1, pp. 1504-1521, 2020. doi: 10.1038/s41598-020-58467-9.

[10] Q. Wuniri, W. Huangfu, Y. Liu, X. Lin, L. Liu, and Z. Yu, "A Generic-Driven Wrapper Embedded With Feature-Type-Aware Hybrid Bayesian Classifier for Breast Cancer Classification", *IEEE Access*, Vol. 7, No. 1, pp. 119931-119942, 2019, doi: 10.1109/ACCESS.2019.2932505.

[11] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm", *IEEE Access*, Vol. 10, No. 1, pp. 3284-3293, 2022, doi: 10.1109/ACCESS.2021.3139595.

[12] Y. Li, F. Zhang, and C. Xing, "Screening of Pathogenic Genes for Colorectal Cancer and Deep Learning in the Diagnosis of Colorectal Cancer", *IEEE Access*, Vol. 8, No. 1, pp. 114916-114929, 2020, doi: 10.1109/ACCESS.2020.3003999.

[13] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P. A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis", *IEEE Transactions on Cybernetics*, Vol. 50, No. 9, pp. 3950-3962. 2020. doi: 10.1109/TCYB.2019.2935141.

[14] Y. Li, J. Chen, P. Xue, C. Tang, J. Chang, C. Chu, K. Ma, Q. Li, Y. Zheng, and Y. Qiao, "Computer-Aided Cervical Cancer Diagnosis Using Time-Lapsed Colposcopic Images", *IEEE Transactions on Medical Imaging*, Vol. 39, No. 11, pp. 3403-3415, 2020, doi: 10.1109/TMI.2020.2994778.

[15] S. M. S. Hasan, D. Rivera, X. C. Wu, E. B. Durbin, J. B. Christian, and G. Tourassi, "Knowledge Graph-Enabled Cancer Data Analytics", *IEEE Journal of Biomedical and Health Informatics*, Vol. 24, No. 7, pp. 1952-1967, 2020, doi: 10.1109/JBHI.2020.2990797.

[16] D. Florensa, P. Godoy, J. Mateo, F. Solsona, T. Pedrol, M. Mesas, and R. Pinol, "The Use of Multiple Correspondence Analysis to Explore Associations Between Categories of Qualitative Variables and Cancer Incidence", *IEEE Journal of Biomedical and Health Informatics*, Vol. 25, No. 9, pp. 3659-3667, 2021, doi: 10.1109/JBHI.2021.3073605.

[17] G. Li, C. Li, G. Wu, D. Ji, and H. Zhang, "Multi-View Attention-Guided Multiple Instance Detection Network for Interpretable Breast Cancer Histopathological Image Diagnosis", *IEEE Access*, Vol. 9, No. 1, pp. 79671-79684, 2021, doi: 10.1109/ACCESS.2021.3084360.

[18] Q. Zhou, B. Yong, Q. Lv, J. Shen, and X. Wang, "Deep Autoencoder for Mass Spectrometry Feature Learning and Cancer Detection", *IEEE Access*, Vol. 8, No.1, pp. 45156-45166, 2020, doi: 10.1109/ACCESS.2020.2977680.

[19] P. Liu, B. Fu, S. X. Yang, L. Deng, X. Zhong, and H. Zheng, "Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer", *IEEE Transactions on Biomedical Engineering*, Vol. 68, No. 1, pp. 148-160, 2021, doi: 10.1109/TBME.2020.2993278.

[20] G. Huang, J. Zhu, J. Li, Z. Wang, L. Cheng, L. Liu, H. Li, and J. Zhou, "Channel-Attention U-Net: Channel Attention Mechanism for Semantic Segmentation of Esophagus and Esophageal Cancer", *IEEE Access*, Vol. 8, No. 1, pp. 122798-122810, 2020, doi: 10.1109/ACCESS.2020.3007719.

[21] G. Alexandrou, N. Moser, K. T. Mantikas, J. R. Manzano, S. Ali, R. C. Coombes, J. Shaw, P. Georgiou, C. Toumazou, and M. Kalofonou, "Detection of Multiple Breast Cancer ESR1 Mutations on an ISFET Based Lab-on-Chip Platform", *IEEE Transactions on Biomedical Circuits and Systems*, Vol. 15, No. 3, pp. 380-389, 2021, doi: 10.1109/TBCAS.2021.3094464.

[22] L. D. Lopez, J. P. D. Morales, A. F. C. Martin, S. V. Diaz, and A. L. Barranco, "PROMETEO: A CNN-Based Computer-Aided Diagnosis System for WSI Prostate Cancer Detection", *IEEE Access*, Vol. 8, No. 1, pp. 128613-128628, 2020, doi: 10.1109/ACCESS.2020.3008868.

[23] A. Masood, P. Yang, B. Sheng, H. Li, P. Li, J. Qin, V. Lanfranchi, J. Kim, and D. D. Feng, "Cloud-Based Automated Clinical Decision Support System for Detection and Diagnosis of Lung Cancer in Chest CT", *IEEE Journal of Translational Engineering in Health and Medicine*, Vol. 8, No. 1, pp. 1-13, 2020, Art No. 4300113, doi: 10.1109/JTEHM.2019.2955458.

[24] W. Zhang, J. Huang, H. N. Chen, M. F. Elahe, and M. Jin, "A Cancer Diagnosis Method Combining miRNA-lncRNA Interaction Pairs and Class Weight Competition", *IEEE Access*, Vol. 8, No. 1, pp. 67059-67074, 2020, doi:

10.1109/ACCESS.2020.2985405.

[25] Z. Chen, W. Zhang, H. Deng, and K. Zhang, "Effective Cancer Subtype and Stage Prediction via Dropfeature-DNNs", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 19, No. 1, pp. 107-120, 2022, doi: 10.1109/TCBB.2021.3058941.

[26] H. Liu, Y. Xu, Z. Zhang, N. Wang, Y. Huang, Y. Hu, Z. Yang, R. Jiang, and H. Chen, "A Natural Language Processing Pipeline of Chinese Free-Text Radiology Reports for Liver Cancer Diagnosis", *IEEE Access*, Vol. 8, No. 1, pp. 159110-159119, 2020, doi: 10.1109/ACCESS.2020.3020138.

[27] P. Yin, B. Hu, Q. Li, Y. Duan, and Q. Lin, "Imaging of Tumor Boundary Based on Multielements and Molecular Fragments Heterogeneity in Lung Cancer", *IEEE Transactions on Instrumentation and Measurement*, Vol. 70, No. 1, pp. 1-7, 2021, Art No. 4006207, doi: 10.1109/TIM.2021.3102755.

[28] R. A. Welikala, P. Remagnino, J. H. Lim, C. S. Chan, S. Rajendran, T. G. Kallarakkal, R. B. Zain, R. D. Jayasinghe, J. Rimal, A. R. Kerr, and R. Amtha, "Automated Detection and Classification of Oral Lesions Using Deep Learning for Early Detection of Oral Cancer", *IEEE Access*, Vol. 8, No. 1, pp. 132677-132693, 2020, doi: 10.1109/ACCESS.2020.3010180.

[29] E. Danku, E. H. Dulf, R. P. Banut, H. Silaghi, and C. A. Silaghi, "Cancer Diagnosis With the Aid of Artificial Intelligence Modeling Tools", *IEEE Access*, Vol. 10, No. 1, pp. 20816-20831, 2022, doi: 10.1109/ACCESS.2022.3152200.

[30] M. T. Gevari, G. Aydemir, G. Gharib, O. Kutlu, H. Uvet, M. Ghorbani, and A. Koşar, "Local Carpet Bombardment of Immobilized Cancer Cells With Hydrodynamic Cavitation", *IEEE Access*, Vol. 9, No. 1, pp. 14983-14991, 2021, doi: 10.1109/ACCESS.2021.3052893.

[31] M. Li, X. Nie, Y. Reheman, P. Huang, S. Zhang, Y. Yuan, C. Chen, Z. Yan, C. Chen, X. Lv, and W. Han, "Computer-Aided Diagnosis and Staging of Pancreatic Cancer Based on CT Images", *IEEE Access*, Vol. 8, No. 1, pp. 141705-141718, 2020, doi: 10.1109/ACCESS.2020.3012967.

[32] Y. Yari, T. V. Nguyen, and H. T. Nguyen, "Deep Learning Applied for Histological Diagnosis of Breast Cancer", *IEEE Access*, Vol. 8, No. 1, pp. 162432-162448, 2020, doi: 10.1109/ACCESS.2020.3021557.

[33] X. H. Wang, B. Zheng, W. F. Good, J. L. King, and Y. H. Chang. "Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network", *Int. J. Med. Inform*, Vol. 54, No. 2, pp. 115–126, 1999.

[34] L. Jiang, H. Zhang, Z. Cai, and J. Su, "Learning Tree Augmented Naive Bayes for Ranking. In: Zhou, L., Ooi, B.C., Meng, X. (eds) Database Systems for Advanced Applications. DASFAA 2005", *Lecture Notes in Computer Science*, Vol. 3453, pp. 688-698, Springer, Berlin, Heidelberg. https://doi.org/10.1007/11408079_63.

[35] L. Jiang, H. Zhang, and Z. Cai, "A Novel Bayes Model: Hidden Naive Bayes", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 10, pp. 1361-1371, 2009, doi: 10.1109/TKDE.2008.234.

[36] L. Jiang, G. Kong, and C. Li, "Wrapper Framework for Test-Cost-Sensitive Feature Selection", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 51, No. 3, pp. 1747-1756, 2021, doi: 10.1109/TSMC.2019.2904662.

[37] L. Jiang, C. Li, and S. Wang, "Cost-sensitive Bayesian network classifiers", *Pattern Recognition Letters*, Vol. 45, No. 1, pp. 211-216, 2014,https://doi.org/10.1016/j.patrec.2014.04.017.

[38] https://www.cancer.gov/types/breast

[39] https://archive.ics.uci.edu/dataset/14/breast+cancer

[40] https://www.nki.nl/research/research-groups/marjanka-schmidt/breast-cancer-datasets-elsi/

[41] https://www.ispytrials.org/