INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Classification of Breast Cancer Patients using Deep Learning Techniques

Tukaram Navghare[1]*, Aniket Muley[1], Vinayak Jadhav[2]

1 School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Nanded, 431606, Maharashtra, India
2 Shivaji College, Udgir, 413517, Maharashtra, India

OPEN ACCESS

Received: 01-11-2023

Accepted: 22-11-2023

Published: 30-12-2023

Citation: Navghare T, Muley A, Jadhav V (2023) Classification of Breast Cancer Patients using Deep Learning Techniques. Indian Journal of Science and Technology 16(47): 4612-4619. https://doi.org/10.17485/IJST/v16i47.2758

* **Corresponding author**.

navgharetukaram@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: Breast cancer is one of the most ubiquitous cancers among women in the world and early exploration of the disease can be lifesaving. Finding breast cancer at an early stage enables quicker initiation of treatment, thereby enhancing the prospects of a positive outcome. Our aim is to identify the deep learning neural network model to classify breast cancer patients. Here, secondary open source data is considered to classify malignant and benign patients suitably**. Methods**: Deep learning neural network model, Artificial Neural Network and simulation approach is used to identify the more precise model. **Findings:** It is observed that, our proposed neural network model specified 97.5% accuracy. Efficiency of the proposed model is evaluated with the performance measures viz., MSE, RMSE etc. **Novelty:** The results of the study obtained through the proposed model express the efficiency of the model itself and also the superiority is demonstrated by comparing it with SVM, ANN, linear regression, 3DCNN deep models and existing works using various case studies. In the future, this model can be applicable in similar studies and it will give better results.

Keywords: Deep learning; Artificial Neural Network; Breast cancer; Classification; Wisconsin data set

1 Introduction

Breast cancer is one of the major wide spread malignancies observed in women throughout the world [(1)]. Breast cancer is life menacing and is observed to be the second highest cause of deaths after cancer in women.In 2018, it is observed that among overall cancer patients there are 24.2 percent of them having breast cancer. It is metastatic in nature hence, if it spreads into the organs it becomes tedious to cure [(2)]. Therefore, if the diagnosis is made earlier then the high percentage of survival can be achieved. Normally, cancers can be detected in the fibrous connective tissue or the fatty tissue inside the breast and is able to invade other healthy breast tissues. The cells of the cancer can be detected either in the ducts or the lobules of the breast [(3)]. Accurate diagnosis is one of the most important processes in breast cancer treatment [(4)]. Mammography is often used as an inspection method for identifying breast cancer to reduce significant mortality. There are number of causes of cancer viz., gender, age, estrogens, genetic and were treated as the most important risk aspects [(5)]. The images are divided into multiple

areas employing similar attributes in neighbouring characteristics viz., coarseness, softness, blush, volume, disparity, form and mass. Apart from the classification based on crucial part of the image in place of focusing on whole image. As there are various views of classification some researchers focused on deep learning models owing to their flexibility, sturdiness, and consistency that eliminate the conventional steps of Machine learning[6]. These methods automatically extract elevated abstract features from images and it can deal with the association and nonlinearity between variables[7].

In[8] the authors proposed model breast cancer Convolutional Neural Network (CNN) that help balancing pre-processing and dataset boosting and it played a good role in enhancing the detection and classification of breast cancer. In[9] the authors applied deep learning tools for breast cancer evaluation based on histopathological images of it. They have used CNN including Inception V3 and Inception ResNet V2 combined with transfer learning techniques to classify it. Further, they have proposed a new autoencoder network structure for applying non linear transformations to characteristics in the image datasets that has been extracted with Inception ResNetV2 network. Thereafter, it is being used as input for classical K means clustering algorithm on image dataset. In[10] the authors applied with UCI dataset and diagnose it with various deep learning technique algorithms viz., neural network (NN), K nearest neighbour (KNN), random forest (RF), support vector machine (SVM) and CNN.

[11] The proposed model has used some effective features of GoogLeNet and ResNet architectures and has added some new features such as granular computing, activation functions with learnable parameters, and attention layer to the new architecture.[12] Performed breast cancer patient classification study for investigating with mammographic images using CNN and SAE, and observed that CNNs are suitable techniques for it.[13] The authors proposed transfer learning architecture that consists of combination of pre trained DenseNet121 and ResNet50 models. In[14] the authors dealt with neural network related techniques use for development validation of SEER breast cancer survival dataset prediction models. In[15] the authors performed comparison of ML tools viz., SVM, KNN, RF, ANN and LR for breast cancer data. In[16] applied K-Nearest Neighbors, Logistic Regression and Ensemble Learning and classify the breast cancer patient in very excellent manner.[17] the authors proposed deep learning modelling for case control cohort regarding whether the result is in terms of breast cancer or cancer free status. They have analyzed the data with CNN tool with GoogLeNet and Linear Discriminant analysis (LDA) and observed that both of these models have superior performance than mammographic breast density. In[18] they employed traditional CNN a support CNN approaches to overcome change and size with that of blurred mammogram images. Thereafter, the flipped rotation based approach (FRbA) is implemented to boost the precision of MIAS medical data of 200 mammogram breast images classification.[19] The authors proposed fusion of deep learning model with the combination of mammogram image information and risk factor that has been based on LR model. This will help to deal with large sets of labeled data and their design extract features automatically.

In this study, our objective is to classify breast cancer patients, using deep-learning algorithm, and find out the most effective one based on the performance of each classifier in terms of confusion matrix, accuracy, precision and recall. One of the major challenges after detection of cancer in breast is to classify malignant or benign cells.

In this article, subsequent sections contain the methodology, result and discussion as well as conclusion mentioned in detail.

2 Methodology

2.1 Deep Learning

The concept initially proposed in the 1980s. Mostly of these methods use neural network architectures and are treated as deep neural networks. In general, traditional ANN contains only Two to Three invisible levels whereas deep neural networks having number of possible layers.[20] This will help to deal with large sets of labelled data and their design extract features automatically. The actual observed models of the proposed work are represented in Figures 2 and 3.

2.2 Dataset

Here, breast cancer data set has been extracted from kaggle repository (https://www.kaggle.com/datasets/ninjacoding/breast-cancer-wisconsin-benign-or-malignant). The samples arrived periodically towards Dr. Wolberg, University of Wisconsin Hospitals, Madison, Wisconsin, USA clinic last five years then he has reported his clinical cases. The data represented in ascending time span group format. Table 2 explores the parameters that has been taken for the study purpose. Attributes 2 to 10 represented in terms of case and further it has been classified in one of two possible classes as either benign or malignant. Total 699 observations were considered for this study and detailed split of instances (See Table 1).

Table 1. Database grouping information

Group No.	Number of Instances	Period
1	367	January 1989
2	70	October 1989
3	31	February 1990
4	17	April 1990
5	48	August 1990
6	49	Updated January 1991
7	31	June 1991
8	86	November 1991
Total	**699**	**as of the donated database on 15 July 1992**

Table 2. Breast cancer data set attributes information

Attribute	Domain
Sample code number	id number
Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10
Class	2 for benign, 4 for malignant
Missing attribute values	16
Class distribution	Benign: 458 (65.5%), Malignant: 241 (34.5%)

In this study, data is analyzed with the free and open source software R 4.1.2 version. The dataset is divided in to two parts- 80% and 20% dataset that is used for training and testing respectively and dataset is split using H2O R Language package. To optimize the result with R program, some packages were utilized viz., mlbench and neuralnet.

2.3 Proposed algorithm

Step 1: Steps that have been used for analysing data.
Step 2: Collect the data set.
Step 3: Identify decision variable.
Step 3: Identify or apply deep learning neural network.
Step 4: Calculate the performance measures Equations (1), (2), (3), (4), (5), (6), (7), (8), (9), (10), (11), (12) and (13).

- **MSE:** Mean Square Error represents the average of the square difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \widehat{y})^2}{N} \tag{1}$$

Where, MSE= Mean square error, where n=number of data points, y_i= Observed values, $\widehat{y}$ = predicted values

- **RMSE:** It is the square root of MSE, also it has the same units as the quantity being estimated; for an unbiased estimator, i.e.

$$RMSE = \sqrt{MSE} \tag{2}$$

- **Log Loss:** It is the most important classification metric based on probabilities. It's hard to interpret raw log-loss values, but log-loss is still a good metric for comparing models. For any given problem, a lower log loss value means better predictions.

$$\log\ loss = -\frac{1}{N}\sum_{i=1}^{N} y_i\ *\ \log\left(p(y_i)\right)\ +\ (1-y_i)\ *\ \log\left(1-p(y_i)\right) \tag{3}$$

- **Mean Per-Class Error :** Mean per Class Error (in Multi-class Classification only) is the average of the errors of each class in multi-class data set. Per Class Error is defined as in Equations (4) and (5):

$$G_1 = \frac{FP}{TP+FP} \tag{4}$$

$$G_2 = \frac{FN}{FN+TN} \tag{5}$$

$$The\ Mean\ per\ class\ error = \frac{G_1+G_2}{2} \tag{6}$$

It deals with the misclassification of the type of class and if its value is least then its better for classification.

- **AUC :** AUC or ROC curve is a plot of the proportion of true positives (events correctly predicted to be events) versus the proportion of false positives (non-events wrongly predicted to be events) at different probability cut-offs.

$$AUC = \int_0^1 \frac{TP}{TP+FN} d\frac{FP}{FP+TN} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} \tag{7}$$

Where TP=true positive, TN=true negative, FN=false native, FP=false negative

$$Specificity\ = \frac{TN}{FP+TN} \tag{8}$$

- **AUCPR :** The AUPRC proves invaluable as a performance measure for skewed data when the primary concern is detecting positive instances in a given problem scenario.

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

- **Gini:** The Gini coefficient can be calculated using the formula:

$$Gini\ Coefficient\ = \frac{A}{A+B} \tag{10}$$

Where A is the area above the Lorenz Curve and B is the area below the Lorenz Curve. Also, calculated by formula

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|}{2n\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|}{2n^2\bar{x}} \tag{11}$$

Relation between Gini and AUC is:

$$Gini = 2*AUC-1 \tag{12}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{13}$$

Step 5: Identification of suitable hidden layers and hidden neurons.
Step 6: Increase hidden layers 1,2,3 layers used.
Step 7: Select optimum number of hidden layers and hidden neurons.
Step 8: comparative approach to check the performance of the model.

Figure 1 flowchart explores the visualization of the proposed study algorithm in the simplified manner.

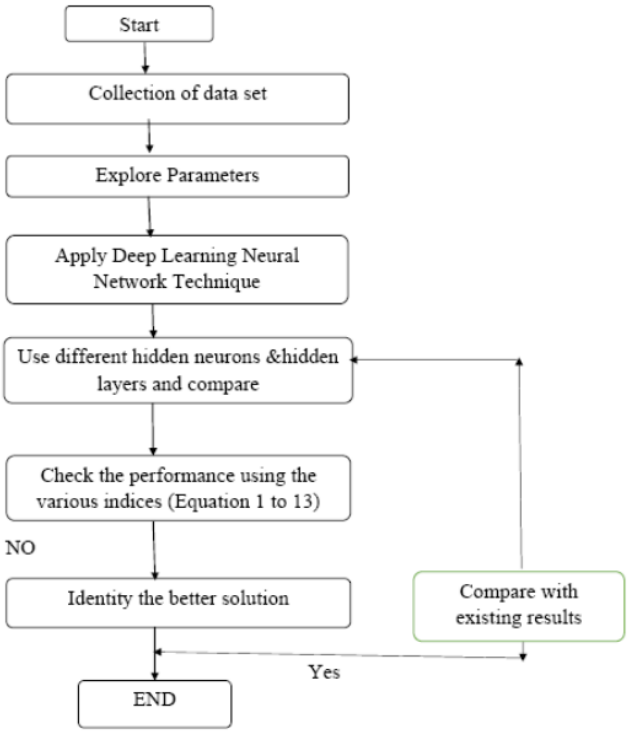

Fig 1. Flow chart of the study

3 Results and Discussion

In this section, the results obtained through proposed algorithm are discussed in detail.

Table 3. Performance measure of various ANN models

Hidden Neuron	MSE	RMSE	LogLoss	Mean Per-Class Error	AUC	AUCPR	Gini	Error	Steps
2	**0.022**	**0.150**	**0.0874**	0.021616	0.9944	0.9873	0.9888	5.9631	731
1,1,1	0.0366	0.1914	0.1791	**0.019541**	0.9946	**0.9893**	0.9893	5.7939	171
4,3,1	0.0262	0.1620	0.1083	0.022708	0.9903	0.9779	**0.9807**	1.4827	260
4,4,1	0.0247	0.1573	0.1061	0.020633	0.9925	0.9838	0.9850	2.9267	**326**
5,3,2	0.0252	0.1588	0.0972	0.023800	0.9938	0.9864	0.9876	**0.4981**	786
5,5,5	0.02649	0.16277	0.1269	0.021616	**0.9947**	0.9890	0.9895	1.483	365

Table 4. Accuracy levels of ANN models

Hidden Neuron	Confusion Matrix			Error rate	Precision	Recall	Accuracy
		Benign	Malignant				
2	Benign	442	16	0.0349	0.97424893	0.9954955	0.9742
	Malignant	2	239	0.0083			
	Total	444	255	0.0256			
1,1,1	**Benign**	**442**	**16**	**0.0349**	**0.97567954**	**0.99774266**	**0.9756**
	Malignant	**1**	**240**	**0.0042**			
	Total	**443**	**256**	**0.0243**			
4,3,1	Benign	441	17	0.0371	0.97281831	0.99548533	0.9728
	Malignant	2	239	0.0083			
	Total	443	256	0.0272			
4,4,1	Benign	441	17	0.0371	0.97424893	0.99773756	0.9742
	Malignant	1	240	0.0041			
	Total	442	257	0.0257			
5,3,2	Benign	440	18	0.0393	0.9713877	0.99547511	0.9713
	Malignant	2	239	0.0082			
	Total	442	257	0.0286			
5,5,5	Benign	442	16	0.0349	0.97424893	0.9954955	0.9742
	Malignant	2	239	0.0082			
	Total	444	255	0.0257			

Table 5. Comparison of different used model with proposed model

	Method Used	Data	Language	Accuracy in (%)
Fang et al. [21]	3DCNN	Primary	-	71.0
Ragab et al. [22]	SVM,ANN ANN	DDSM	Python	94.0
Wadkar et al. [23]	SVM,KNN,CNN	Wisconsin	_	97
Naji et al. [24]	SVM, LR KNN,	Wisconsin	Python	97.2
Proposed Model	**ANN with Simulation**	**Wisconsin**	**R**	**97.56**

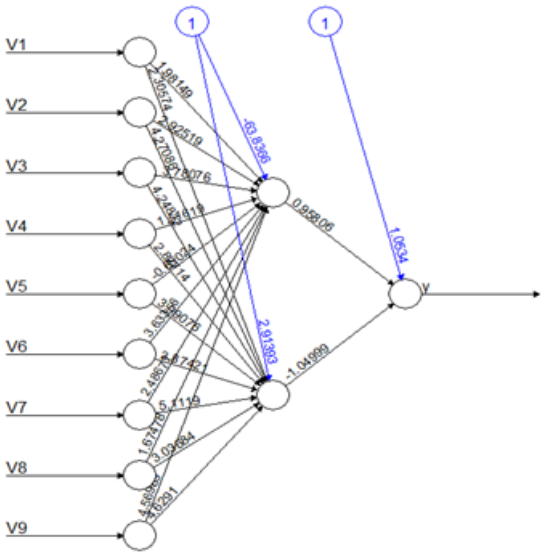

Fig 2. 9-2-1 NN model

After applying Artificial neural network model on Breast Cancer Wisconsin dataset, we used Confusion Matrix, Error rate, Mean square error, Root Mean square error, Accuracy, LogLoss, Precision, Mean Per-Class Error, Sensitivity Gini coefficient, recall, AUC steps as performance measures to evaluate and compare the models and identify the best model for the breast

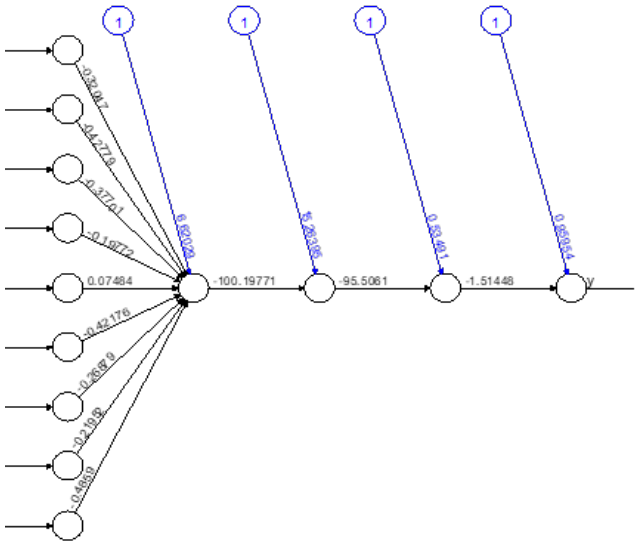

Fig 3. 9-(1,1,1)-1 NN model

cancer patient classification. Confusion Matrix is the way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is a table with two dimensions viz. "Actual" and "Predicted" and furthermore, both the dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", and "False Negatives (FN)". Accuracy is most common performance measure for breast cancer patient classification.

Here, we have performed sensitivity analysis and obtained 25 neural network models with different neuron and layers. We found six neural network optimal models that are represented in Figures 2 and 3. The result of performance measures of these models are summarized in Table 3. From Tables 3 and 5, results reveal that the single layer with hidden neurons model is found to be the more suitable and the results are compared with other ANN models and previously studied models.

4 Conclusion

In this study our aim is to classify the breast cancer disease patients of i.e. Benign and Malignant using Artificial Neural Network with simulate hidden neurons and various layer. This model gives us more precise results in terms of classification and to predict cancer disease classification in terms of malignant or benign with higher accuracy. i.e. (97.56%) as compared to State-of-the-art method as shown in Table 5. Wisconsin breast cancer data set is extracted from kaggle repository. The dataset consists of 699 patients' details. In the future more data would be added to the database and increase in the hidden neurons with layers will be helpful for classification problem which would increase help in better result, it would work more accurately and to evaluate cancer risk assessment.

References

1) Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209–249. Available from: https://doi.org/10.3322/caac.21660.
2) Nasser M, Yusof UK. Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. *Diagnostics*. 2023;13(1):1–26. Available from: https://doi.org/10.3390/diagnostics13010161.
3) Ghoushchi SJ, Ranjbarzadeh R, Najafabadi SA, Osgooei E, Tirkolaee EB. An extended approach to the diagnosis of tumour location in breast cancer using deep learning. *Journal of Ambient Intelligence and Humanized*. 2023;14:8487–8497. Available from: https://doi.org/10.1007/s12652-021-03613-y.
4) Ak MF. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. *Healthcare*. 2020;8(2):1–23. Available from: https://doi.org/10.3390/healthcare8020111.
5) Gupta S, Gupta MK. Computational Model for Prediction of Malignant Mesothelioma Diagnosis. *The Computer Journal*. 2023;66(1):86–100. Available from: https://doi.org/10.1093/comjnl/bxab146.
6) Ranjbarzadeh R, Saadi SB. Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering. *Measurement*. 2020;150:107086. Available from: https://doi.org/10.1016/j.measurement.2019.107086.

7) Zhu W, Xie L, Han J, Guo X. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers*. 2020;12(3):1–19. Available from: https://doi.org/10.3390/cancers12030603.
8) Abunasser BS, Al-Hiealy MR, Zaqout IS, Abu-Naser SS. Convolution Neural Network for Breast Cancer Detection and Classification Using Deep Learning. *Asian Pacific Journal of Cancer Prevention*. 2023;24(2):531–544. Available from: https://journal.waocp.org/article_90487.html.
9) Xie J, Liu R, Luttrell J, Zhang C. Deep Learning Based Analysis of Histopathological Images of Breast Cancer. *Frontiers in Genetics*. 2019;10:1–19. Available from: https://doi.org/10.3389/fgene.2019.00080.
10) Chen H, Wang N, Du X, Mei K, Zhou Y, Cai G. Classification prediction of breast cancer based on machine learning. *Computational Intelligence and Neuroscience*. 2023;2023:1–9. Available from: https://doi.org/10.1155/2023/6530719.
11) Zakareya S, Izadkhah H, Karimpour J. A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images. *Diagnostics*. 2023;13(11):1–23. Available from: https://doi.org/10.3390/diagnostics13111944.
12) Yao H, Zhang X, Zhou X, Liu S. Parallel Structure Deep Neural Network Using CNN and RNN with an Attention Mechanism for Breast Cancer Histology Image Classification. *Cancers*. 2019;11(12):1–14. Available from: https://doi.org/10.3390/cancers11121901.
13) Yari Y, Nguyen TV, Nguyen HT. Deep Learning Applied for Histological Diagnosis of Breast Cancer. *IEEE Access*. 2020;8:162432–162448. Available from: https://ieeexplore.ieee.org/document/9186080.
14) Gupta S, Gupta MK. A Comparative Analysis of Deep Learning Approaches for Predicting Breast Cancer Survivability. *Archives of Computational Methods in Engineering*. 2022;29(5):2959–2975. Available from: https://doi.org/10.1007/s11831-021-09679-3.
15) Islam MM, Haque MR, Iqbal H, Hasan MM, Hasan M, Kabir MN. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*. 2020;1(5):1–14. Available from: https://doi.org/10.1007/s42979-020-00305-w.
16) Murtirawat R, Panchal S, Singh VK, Panchal Y. Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 02-04 July 2020, Coimbatore, India. IEEE. 2020;p. 534–540. Available from: https://ieeexplore.ieee.org/document/9155783.
17) Arefan D, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning modeling using normal mammograms for predicting breast cancer risk. *Medical Physics*. 2020;47(1):110–118. Available from: https://doi.org/10.1002/mp.13886.
18) Alfifi M, Shady M, Bataineh S, Mezher M. Enhanced Artificial Intelligence System for Diagnosing and Predicting Breast Cancer using Deep Learning. *International Journal of Advanced Computer Science and Applications*. 2020;11(7):498–513. Available from: https://thesai.org/Downloads/Volume11No7/Paper_63-Enhanced_Artificial_Intelligence_System.pdf.
19) Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*. 2019;292(1):60–66. Available from: https://doi.org/10.1148/radiol.2019182716.
20) Sharif MI, Li JP, Naz J, Rashid I. A comprehensive review on multi-organs tumor detection based on machine learning. *Pattern Recognition Letters*. 2020;131:30–37. Available from: https://doi.org/10.1016/j.patrec.2019.12.006.
21) Fang Y, Zhao J, Hu L, Ying X, Pan Y, Wang X. Image classification toward breast cancer using deeply-learned quality features. *Journal of Visual Communication and Image Representation*. 2019;64:102609. Available from: https://doi.org/10.1016/j.jvcir.2019.102609.
22) Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*. 2019;7:e6201. Available from: https://doi.org/10.7717/peerj.6201.
23) Wadkar K, Pathak P, Wagh N. Breast Cancer Detection Using ANN Network and Performance Analysis With SVM. *International Journal of Computer Engineering and Technology*. 2019;10(3):75–86. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3555041.
24) Naji MA, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Computer Science*. 2021;191:487–492. Available from: https://doi.org/10.1016/j.procs.2021.07.062.