International Journal of Current Research and Review
DOI: http://dx.doi.org/10.31782/IJCRR.2021.13904

Scopus

Research Article

IJCRR
Section: Healthcare
ISI Impact Factor (2019-20): 1.628
IC Value (2019): 90.81
SJIF (2020) = 7.893

Mammogram Classification with Forest Optimization using Machine Learning Algorithms

Kanya Kumari L[1], Jayaprada S[2], Ranga Rao J[3]

[1]Assistant Professor, Andhra Loyola Institute of Engineering and Technology, Department of Information Technology, Vijayawada, Andhra Pradesh, India; [2]Associate Professor, Department of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India; [3]Assistant Professor, Department of Computer Science & Engineering, VR Siddhartha Engineering College, Kanuru, Vijayawada, Andhra Pradesh, India.

ABSTRACT

Introduction: The deadly disease in Indian women is Breast Cancer (BC). A mammogram is used for identifying the tumours in the breast in the early stages which is efficient and cost-effective.

Objective: The main objective is to predict BC in the early stages using image processing and machine learning techniques.

Methods: Our proposed methodology is 6 step process which includes preprocessing, feature extraction, feature selection, splitting the data into training and testing, classification and performance measure.

Results: The experiments are done on MIAS (Mammogram Image Analysis Society) dataset. As more noise in the images of this dataset, filters are applied to get more clarity in images. Features are extracted by Local Binary Patterns (LBP) and optimized by Forest Optimization Algorithm (FOA). These features are divided into 70% training and 30% testing data for classification. The classifiers used are K- Nearest Neighbor (KNN), Naïve Bayes (NB) and Random Forest (RF).

Conclusion: The experiments show that LBP based FOA with RF classifier achieved good accuracy in classifying the mammograms.

Key Words: Breast cancer, Local Binary Patterns, Forest Optimization, Random Forest, K-Nearest Neighbor and Naïve Bayes

INTRODUCTION

In recent years, most people are suffering from cancer. Not all cancer cells are dangerous. There are nearly many types of cancers. They are pancreatic cancer, cervical cancer, lung cancer, breast cancer, colorectal cancer, thyroid cancer, kidney cancer and melanoma, etc.[1]. Normally, the patient data is called Electronic Health Record (EHR's). These records may be structured, unstructured or semi-structured data. If the particular format for the patient's record is available then it is structured data. If there is no particular format for the data then it is called unstructured data. The combination of structured and unstructured data is called semi-structured data.[2] Abnormal cancer cells are detected by using the number of modalities by the radiologist which is unstructured. These modalities are X-ray, CT-Scan (computed tomography), ultrasound, Thermogram imaging, PET (Positron Emission Tomography), and MRI (Magnetic Resonance Imaging).

Breast cancer (BC) is the commonest disease in Indian women. The screening for breast cancer is the digital X-ray called mammogram imaging modality which is cost-effective, efficient, and fewer side effects of radiation.[3] The categories of breast cancer are from 0-5 which is given by the American college of radiology called BI-RADS (Breast Imaging- Reporting and Data System) in 1986. According to BI-RADS, category-0 is an incomplete evaluation and requires additional imaging techniques have to consider. BI-RADS category-1 is negative BC, category-2 is the benign means less probability of severity, category-3 is also benign category but <2% malignancy and continuous monitoring are required. Category 4 is a suspicious abnormality that has >2% malignancy and category 5 is >=95% malignancy.[4]

Machine learning (ML) is a subset of Artificial Intelligence (AI) which gives the results based on the learning experience. There are different types of machine learning techniques known as supervised learning, unsupervised learning and semi-supervised learning. A new class label is predicted

Corresponding Author:
Kanya Kumari L, Assistant Professor, Andhra Loyola Institute of Engineering and Technology, Department of Information Technology, Vijayawada, Andhra Pradesh, India, Email: kanyabtech@yahoo.com

ISSN: 2231-2196 (Print) **ISSN:** 0975-5241 (Online)

Received: 11.11.2020 **Revised:** 08.01.2021 **Accepted:** 16.03.2021 **Published:** 07.05.2021

based on the labelled training data is called supervised learning whereas the class label is predicted based on the clusters (doesn't contain labelled data) is called unsupervised learning. The combination of supervised and unsupervised learning is called semi-supervised learning. Nowadays ML is helping physiologists and radiologists in the diagnosis, prognosis, and prediction of diseases in early stages.[5] For example, to predict whether the tumour is benign or malignant, ML estimates the class label as 0 or 1 for benign or malignant respectively. These techniques are not only helping the doctors in the prediction of the diseases but also involved in the correct medication to the patients.[6] ML is mainly helping cancer patients to predict the disease in its early stages. The mail ML techniques used in disease prediction are Support Vector Machine (SVM), Artificial Neural Networks (ANN), Bayesian Networks (BN's), K-Nearest Neighbor, and Decision Trees (DTs).[2] The paper is organized as follows: related work is in section 2, the proposed methodology is in section 3 experimental results and discussion is in section 4.

The author's in7 proposed a methodology to detect tumours by reducing the noise in MIAS mammograms using 2D median filters. Then the images are segmented by region growing approach. From these segmented images, features are extracted based on texture and features are selected by using the rough set. The classifier used is ANN to classify the mammograms. In8, the features are extracted for the DDSM mammogram image dataset by the root mean square slope, circularity, fractal dimension. SVM classifier has given better results than other classifiers. Polynomial classifier is used by using the features obtained from curvelet transformations and Linear Binary Patterns (LBP) to classify the DDSM images.[9]

Images are enhanced using chain code and a rough set. These enhanced images are segmented using vector field convolution and features are extracted by using shape, texture, and intensity.[10] The classifier RF has given better results which are measured by using 5-fold cross-validation. Median filtering; harmonic filtering and logarithmic transformations are applied.[11] Then features are extracted by Fourier transformations and weighted Fourier transformations. The features are selected by principal component analysis. The classifiers used are SVM and KNN in which performance is measured by 10-fold cross-validation. They investigated that SVM will give better results for classification from GLCM features and PCA as feature selection.

With aimed to classify the mammogram images into benign or malignant it was proved that genetic programming helps to select the best features from the WDBC dataset.[12] Deep learning is playing an important role in the classification of mammograms.[13] They have used the Convolutional Neural networks (CNN) model for feature extraction which is an improved version of AlexNet. The classifier used is SVM which gave better results.

MATERIALS AND METHODS

Our proposed methodology is a 6-step process. The steps are image acquisition, image preprocessing, feature extraction, feature selection, classification, and performance evaluation. The flow chart of the proposed methodology is represented in Figure 1.

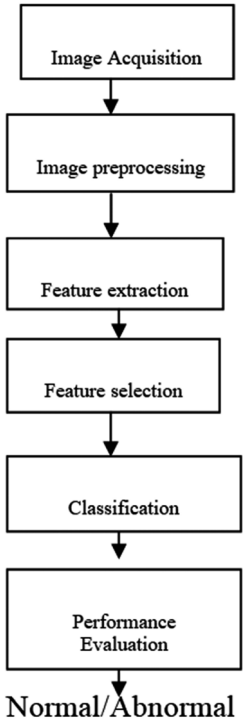

Figure 1: The architecture of proposed methodology: The methodology is a 6 step process that includes acquiring data, preprocessing the images, feature extraction, feature selection, splitting the data into training and testing, classification and performance evaluation.

The main objective of our methodology is to classify the mammogram images into either benign or malignant. To classify the mammograms first we have to acquire the image dataset. Then these images are preprocessed to get clarity in the intensities and pixel values by using preprocessing techniques. These preprocessed images are given as input to the feature extraction technique which extracts the features from the image. All the features are may not be useful for the classification. So, useful features are selected by using feature selection methods. The selected features are divided into training and testing as 75% and 25% respectively. The classifiers classify the images into either benign or malignant. The performance is measured by the confusion matrix and accuracy is also calculated to identify the good classifier.

Image Acquisition

Several imaging techniques are available to detect the tumours in mammogram images such as digital mammograms, CT-Scan (computed tomography), MRI (Magnetic Resonance Imaging), ultrasound, and PET(Positron Emission Tomography). Among these modalities, cost-effective, efficient, and less radiation is for mammograms. To our research, the mammogram image dataset is considered called MIAS (Mammogram Image Analysis Society). This dataset consists of 322 images of 1024X1024 size each and contains a combination of benign and malignant.[14] Among these 322 images, 112 and 210 are normal and abnormal images respectively. The sample images from MIAS are shown in the following Figure 2.

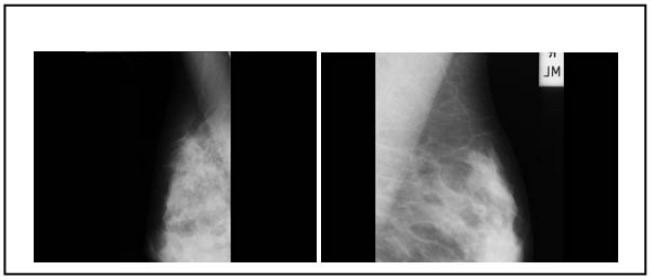

Figure 2: Images from MIAS database: mdb033, mdb116 - sample images considered from MIAS dataset.

Image preprocessing

The images are preprocessed to eliminate noise, reduce redundancy, and smoothening the edges so that efficient features are extracted. There is several preprocessing techniques are available. The image can be enhanced to get more clarity in the images. The enhancement techniques are: filtering with morphological operators, histogram equalization, noise removal using wiener filter, linear contrast adjustment, median filtering, unshaped masking, contrast-limited adaptive histogram equalization, and decorrelation stretch.[15] Filtering is one of the fastest and simple techniques used for smoothening the image or enhancing or detecting edges in the images. There is several filtering techniques are available like mean/ average filter, Gaussian filter, median filter, adaptive mean filter Weiner filter, and Laplacian filter. Edges in the image are detected by using differential operator, Robert's operator, and Sobel operator.[16] Among these filters, we have applied Gaussian, Weiner, and median filters to reduce noise and smoothing images.[17]

Feature Extraction

After preprocessing step, features are to be extracted. Extracting the features from images is called feature extraction. Several feature extraction techniques are available. Mainly, the features are divided into two types. They are local and global features.[18] Local features are concentrate on the patches of the images where are global features concentrate on the entire image. These features may be texture, structural and statistical features. Texture-based features give information about smoothness, coarseness, and regularity. So, we have used Local Binary Patterns (LBP) which is a texture-based feature extraction technique used for extracting the features. These labels each pixel in the image based on thresholding the neighbour of each pixel. The important characteristic of this is simple and efficient for grayscale images.[19] LBP gives contrast information of surrounding pixels.[18] If a middle pixel mp has 8 neighbouring pixels denoted by N= { n1, n2…n8 } the binary pattern(BP) can be denoted as follows:

$$BP[k] = \begin{cases} 0 \ if \ nk < mp \\ 1 \ otherwise \end{cases} \quad (1)$$

Likewise, BP is calculated for all the pixels of an image. These LBP features are normalized and prepare as a feature vector.

Feature Selection

All the extracted features may be useful for the classifier. So the optimal features are only useful for classification. To select optimal features, heuristic algorithms are used. These algorithms are divided into two groups. They are Evolutionary Algorithms (EA) and Swarm Intelligence (SI). Example algorithms for EA are Genetic algorithms (GA), Evolutionary Programming (EP), Forest Optimization (FO), etc. Examples of SI are Particle Swarm Optimization (PSO), Ant Colony Optimization, Firefly Optimization, etc.[20]

In our research, we have used Forest Optimization Algorithm (FOA).[21] It is an Evolutionary Optimization Algorithm used to select useful features for classification. Mainly, it works in 3- steps: 1. Local Seeding on Trees (LO) 2. Limitation of population and 3. Global Seeding (GS). Parameters like a lifetime(LI), local seeding changes(LSCH), area limit of the forest(AL), transfer learning rate(TL), global seeding changes(GSCH) are given as input. We have initialized the forest with the selected number of trees. Each tree is a 0/1 tree which is having N+1 dimensions where N is the dimensions of the feature vector and tree age is also initialized to 0. If LO operation is done then age is incremented by 1 except newly generated one.

To generate the tree children LO operation is done. This is done by a parameter called LSCH. The population of the trees is limited based on the parameters area limit (AL) and lifetime (LI). To form the candidate population (CP) some trees are removed based on age. Then the remaining trees are placed in sorted order depending on their fitness values. If the number of trees is exceeded then the trees are removed from the forest and are added to CP. A GSCH is performed on CP and is obtained by TL. Some of the bits are selected

from the selected CP trees depending on GSC. The randomly chosen bits are changed from 0 to 1 and 1 to 0. In order o select the best tree from the forest, the fitness value is determined. Then make the selected best tree age as 0 and repeat this process until any one of the specific criteria reaches. The termination criteria are 1. Several iterations 2. No difference between the fitness values in the successive iterations 3. Given accuracy measure. In our methodology, we have used number iterations as the stopping criteria. In this way, the optimal features are selected.

Classification

The optimal feature vectors of the MIAS dataset is given as input to the different classifiers to predict the tumour is normal or abnormal. We have used KNN and RF classifiers.

a. K- Nearest Neighbor Classifier

A simple and efficient supervised learning algorithm called K-Nearest neighbour (KNN) is used to classify the mammogram into normal or abnormal. This classifier depends on the distance metric[22]. Several distance measures are available such as Manhattan distance, Euclidean distance, Mahalanobis distance, and Minkowski distance. Mostly used distance measure is Euclidean distance[23]. Based on the value of K, the classifier works. In our proposed methodology we have chosen k=4.

b. Naïve bayes classifier

It works on the theorem called the Bayesian theorem. It outperforms well for classification problems[24]. It works on the given equation(2)

$$\text{Posterior probability} = \frac{prior\ probability X likelihood}{Evidance} \quad (2)$$

The posterior probability is a sample with specific characteristics in a class. It will be calculated by multiplying prior probability and likelihood where prior probability is the probability of class appearance and likelihood is the probability of emergence of sample like characteristic in a class[25].

c. Random Forest Classifier

It is a supervised learning machine learning algorithm used for both classification and regression. The working is based on the decision trees. Decision trees are constructed for the randomly selected samples and class labels are generated. Then voting is done for the predicted labels. The majority voted label is the final predicted class label.

Experimental results and Discussion

Our methodology is implemented using python on a personal computer which an i5 processor and 4GB RAM with Windows 10 OS. The experiments are done on the MIAS dataset. These images are preprocessed by using Gaussian, wiener, and median filters to reduce noise and smoothing the edges. These preprocessed images are given as input to the feature extraction technique called Local Binary Pattern to get a feature vector; the features are reduced by applying the Evolutionary Algorithm called Forest Optimized Algorithm (FOA). These optimized features are given as input to the classifiers KNN, NB, and random forest. The performance is measured based on the confusion matrix represented in which calculates the sensitivity[26], specificity, precision, and f1-score are given in equations 3 to 6 as follows. The performance measures used for the three classifiers are represented below.

$$\text{Sensitivity/ Recall} = \frac{TPS}{TPS+FNS} \quad (3)$$

$$\text{Specificity} = \frac{TNS}{FPS+TNS} \quad (4)$$

$$\text{Precision} = \frac{TPS}{TPS+FPS} \quad (5)$$

$$\text{f1-score} = \frac{2XPrecisionXRecall}{Precision+Recall} \quad (6)$$

True Positives means Sick people are predicted correctly, False Positives means healthy people are wrongly predicted, True Negatives mean healthy people predicted correctly, False Negatives mean sick people predicted incorrectly. The performance measures are represented in the following table 1 and also in figures 3 where x-axis values are performance measure values in percentages and y-axis is the performance measure. The sensitivity, specificity, precision and f1-score values for LBP+FOA+KNN methodology are 94.6%, 94.8%,94.8%, and 94.4% respectively. These values for LBP+FOA+NB methodology obtained are 95.3%, 95.4%, 95.5%, and 95.8% respectively. Similarly, for LBP+FOA+RF methodology are 96.9%, 96.4%, 96.5%, and 96.1% respectively.

Table 1: Confusion Matrix parameters. This table indicates the specificity, sensitivity, precision and f1-score in % for the experiments such as LBP+FOA+RF, LBP+FOA+NB and LBP+FOA+KNN

Dataset	Proposed Methodology	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
MIAS	LBP+FOA+KNN	94.6	94.8	94.8	94.4
	LBP+FOA+NB	95.3	95.4	95.5	95.8
	LBP+FOA+RF	96.9	96.4	96.5	96.1

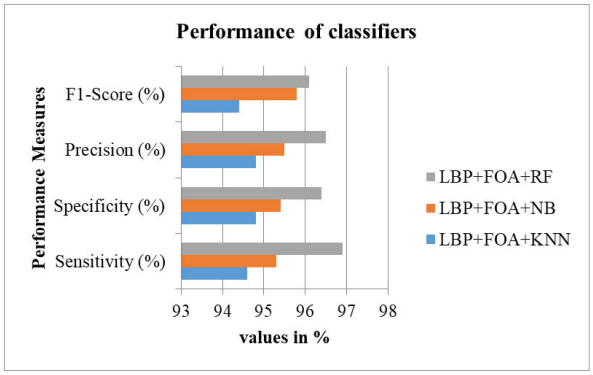

Figure 3: Performance measures of classifiers – Graph represents the specificity, sensitivity, precision and f1-score performance measures in %. This figure shows the performance measures of the experiments such as LBP+FOA+RF, LBP+FOA+NB and LBP+FOA+KNN.

The best classifier can also be decided by another performance measure called accuracy given below in equation 7. The accuracy of each classifier is represented in table 2. The accuracies for KNN, NB, and RF obtained are 94.5%, 95.8%, and 96.2% respectively and are also represented in figure 4.

$$\text{Accuracy} = \frac{True\ Positives + True\ Negetives}{total} \quad (7)$$

Table 2: Accuracy comparison of proposed classifiers

Test No	Dataset	Classifier	Accuracy (%)	Existing/ Proposed
1	MIAS	KNN	94.5	Proposed approach
2		NB	95.8	Proposed Approach
3		RF	96.2	Proposed Approach

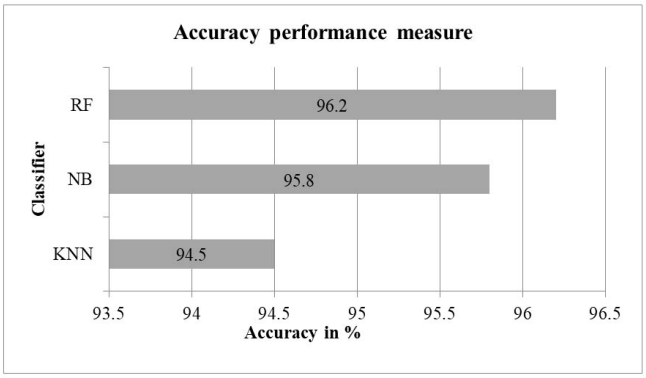

Figure 4: Accuracy of classifiers – This figure shows the accuracies obtained for the proposed methodologies such as LBP+FOA+RF, LBP+FOA+NB and LBP+FOA+KNN and observed that LBP+FOA+RF is better than the other two experiments.

By observing the above performance measures, we conclude that LBP based feature extraction with FOA is giving better classification results using random forest.

CONCLUSION

For our study of research, we considered the MIAS mammogram image dataset. These images are having more noise and are reduced by applying filters like Gaussian, wiener, and median. By applying the filters, clarity in the image is increased. Then features are extracted by LBP and the best features are selected by FOA. These optimized features are given as input to the classifiers namely KNN, NB, and RF. These classifiers are evaluated by using the performance measures like sensitivity, specificity, precision and f1-score are 94.6%, 94.8%, 94.8%, and 94.4% for the KNN classifier. For NB classifier these measures are 95.3%, 95.4%, 95.5%, and 95.8%. For RF classifier, the measures are 96.9%, 96.4%, 96.5%, and 96.1%. Including these parameters, we have also calculated the accuracy of these classifiers and obtained them as 94.5%, 95.8%, and 96.2% for KNN, NB and RF respectively. These results analyze the mammogram images using ML algorithms and found that LBP+FOA+RF shows better results than compared to other classifiers. In future, we apply our proposed methodology to detect another type of cancers and also decided to use deep learning techniques.

ACKNOWLEDGMENT

The authors acknowledge the immense help received from the scholars whose articles are cited and included in references to this manuscript. The authors are also grateful to authors/editors/publishers of all those articles, journals, and books from which the literature for this article has been reviewed and discussed.

Conflict of Interest: The authors involved in the current study does not declare any competing conflict of interest

Source of Funding: No fund or sponsorship in any form was obtained from any organization for carrying out this research work.

Authors Contribution: The authors confirm contribution to the paper as follows:

1. L Kanya kumari - Design, literature search, data acquisition, manuscript preparation, manuscript editing, and manuscript review.

2. S Jayaprada- Concepts, design, literature search, manuscript preparation, manuscript editing.

3. J Ranga Rao-Literature search, manuscript preparation, manuscript editing, and manuscript review.

REFERENCES

1. Cancer.net [Accessed on 19-07-2020].
2. Kanyakumari L, Jagadesh BN. A Review on Big Data Analytics in Multiple Levels of Health Informatics. Int J Sci Res 2017;73(6).
3. Fass L. Imaging and cancer: A review. Mol Oncol 2008;34:115–152.
4. Nam SY, Ko EY, Han BK, Shin JH, Ko ES, Hahn SY. Breast Imaging Reporting and Data System Category 3 Lesions Detected on Whole-Breast Screening Ultrasound. J Breast Can 2016;19(3):301–307.
5. Kouroua K, Exarchosab TP, Exarchosa KP, Karamouzisc KV, Fotiadisab DI. Machine learning applications in cancer prognosis and prediction. Comp Str Biotech J 2015;(13): 8-17.
6. Machine learning in cancer diagnostics, EBioMedicine 2019;(45): 1–2.
7. Peng W, Mayorga RV, Hussein EMA. An Automated Confirmatory System for Analysis of Mammograms. Comp Meth Progr Biomed 2015;125:134-44.
8. Li H, Meng X, Wang T, Tang Y, Yin Y. Breast masses in mammography classification with local contour features. Biomed Eng Online 2017;16(1):44.
9. Daniel O, Bruno1a T, Marcelo Z, Nascimentoab RP, Ramos VR, Leandro AB, et al. LBP operators on curvelet coefficients as an algorithm to describe texture in breast cancer tissues. Expert Systems with Applications 2016;55:329-340.
10. Dong M, Lu X, Ma Y, Guo Y, Ma Y, Wang K. An Efficient Approach for Automated Mass Segmentation and Classification in Mammograms. J Digit Imaging 2015;28(5):613-25.
11. Zhang YD, Wang SH, Liu G, Yang J. Computer-aided diagnosis of abnormal breasts in mammogram images by weighted-type fractional Fourier transform. Adv Mech Engi 2016;8(2):1-11.
12. Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Nagi MF. Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. J Healthcare Engg 2019.
13. Zhao X, Wang X, Wang H. Classification of Benign and Malignant Breast Mass in Digital Mammograms with Convolutional Neural Networks. Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine. 2018; 47–50.
14. J Suckling. The Mammographic Image Analysis Society Digital Mammogram Database. Excerpta Medica, Int Congr Ser 2015;375-378.
15. Mohan S, Ravishankar M. Modified Contrast Limited Adaptive Histogram Equalization Based on Local Contrast Enhancement for Mammogram Images. In: Das V.V., Chaba Y. (eds) Mobile Communication and Power Engineering. AIM 2012. Comm Comp Inform Sci 2012;29(6):718.
16. Punitha A, Amuthan K. Suresh J. Benign and malignant breast cancer segmentation using optimized region growing technique. Future Comp Inform J 2018;3(2):348-358.
17. B Bektaş, Emre IE, Kartal E, Gulsecen S. Classification of Mammography Images by Machine Learning Techniques. 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, 2018;580-585.
18. Al Nahid A, Kong Y. Involvement of Machine Learning for Breast Cancer Image Classification: A Survey. Comput Mathem Meth Med 2017;21(8):325-329.
19. http://www.scholarpedia.org/article/Local_Binary_Patterns [Accessed on 30-09-2020].
20. Rao RV. Teaching-Learning-Based Optimization (TLBO) Algorithm And Its Engineering Applications. Springer International Publishing, Switzerland, 2016. DOI:10.1007/978-3-319-22732-0
21. Alaei SH, Shahraki H, Rowhanimanesh AR, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iran J Basic Med Sci 2016;(19):476-482.
22. Kanya kumari L, Jagadesh BN. A Novel Approach for Detection of Tumors in Mammographic Images using Fourier Descriptors and KNN. Electr Engg. 2020;601:1877-1884.
23. Kourou K. Machine learning applications in cancer prognosis and prediction. Comp Str Biotech J 2015;12:8–17.
24. Rathi M. Breast Cancer Prediction using Naïve Bayes Classifier. Int J Infor Technol Syst 2012;1(2):77-80.
25. Sanjaya MD, Pradnyana MA, Putrama M. Classification of breast cancer using Wrapper and Naïve Bayes algorithms. J Physics 2017; 12: 6–7.
26. Sweta A, Sharma U. Effectiveness of Cytological Scoring Systems for Evaluation of Breast Lesion Cytology with its Histopathological Correlation. Int J Curr Res Rev 2021;13(4):33-38.