Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220084

Classification Model for Tumor Detection in Breast Cancer Patients Using K-Neighborhood and Decision Tree

R.Sai Monika [a,1] and P.Sriramya [b]
[a] *Research Scholar, Dept. of CSE, Saveetha School of Engineering,*
[b] *Professor, Dept. of AI&DS, Saveetha School of Engineering,*
[a,b] *SIMATS, Chennai, Tamilnadu, India.*

Abstract. Breast cancer is a malignancy affecting many women worldwide and is associated with a high fatality rate. The paper's main objective is to detect the breast cancer tumor using a K-neighborhood and compare it with Decision Tree classification to evaluate accuracy using the Machine Learning technique. The k-neighborhood algorithm is applied to 10 images from a dataset of more than 300. For the same, the accuracy values are evaluated. It consists of breast cancer images in the research study of K-neighborhood machines and a decision tree with 20 sample sizes. Based on the statistical analysis, the significance value for calculating accuracy was $p<0.05$. Breast cancer detection is performed using a K-neighborhood, which means 87.5% and 80.5% in the Decision Tree classification. The performance of the K-neighborhood is considerably improved than the Decision Tree classification in terms of accuracy.

Keywords: K-neighborhood, Decision tree, Machine learning, Innovative Prediction, Breast cancer.

1. Introduction

Breast cancer is under high and low magnification; the form of the nuclei and the architectural pattern of the tissue is analyzed for breast cancer diagnosis creation of the classification prototype for evaluating breast cancer [1]. The resources that could be better spent examining patients and focusing on cases where disease classification is difficult to define are supplied with features that are usually the image processing computer-aided diagnostic that can be used [2]. Imaging diagnosis is hypothetical, and it entails radiologists or computer-aided detection systems interpreting various medical picture testing, including nipple aspirate fluid analysis [3]. Recurrence of breast cancer is one of the most common anxieties that women predict can help alleviate recent advances in data mining techniques [4][1]

Breast cancer is a severe disease that develops when the cell grows out of control and approaches breast cancer testing [5]. A mammogram is a low screening performance on women with breast cancer symptoms and detects disease early in other

[1] P.Sriramya, Dept. of AI&DS, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, *Chennai, India.* E-mail: sriramyap@saveetha.com

body parts. Mammography is performed on patients [6]. The subject of image mining is the extraction of implicit knowledge and images with data relationships or different patterns not explicitly in photographs [7]. Breast Cancer Healthline, Beating cancer Advisor, and Cancer therapy advisor are some applications that can apply some cancer detection techniques to solve. The KNN algorithm presented a mechanism for evaluating alternative KNN algorithm settings and choosing the best one. Our selection process is based on test data results and establishes good scores [8]. Breast cancer mammography is an important imaging technique for diagnosing breast cancer early [9]. The studies are based on the publicly available dataset, containing around 8000 microscopic biopsy pictures of benign and malignant breast cancers from 82 people [10]. The work looks at how may use parallel programming with the K-Nearest Neighbors (KNN) technique, which is used to classify and forecast big datasets, is a non-parametric method that classifies data using a similarity measure [11]. The textural qualities of ROI, as well as the efficacy of the KNN classification algorithm in identifying digital mammograms as normal or abnormal, are investigated in this research. The wavelet decomposition coefficient energy is utilized as input [12]. Most selected studies preprocess their input photos by decreasing noise and normalizing colors, while some use segmentation with thresholding approaches to extract the region of interest [13]. The enhancement of structure and the color are three types of image preprocessing, and contrast improvement is currently being improved [14]. The disadvantage will make it more informal for additional image processing to meet the desired goal of image resizing and noise reduction, and cancer part removal technique. [15].

The drawback of the existing system is to identify the cell where the cell is present and detect the tumor with a reduced power image. There are sometimes error diagnoses. The study's main aim is to improve the detection accuracy of tumors in breast cancer images by using a K-neighbourhood regression tree compared with Decision tree classification.

2. Materials and Methods

The proposed research was carried out in the Image Processing Lab at SIMATS' Saveetha School of Engineering's Department of Computer Science and Engineering. This procedure used a sample of more than 300 photos to perform the testing. Two groups were tested to see how accurate they were in detecting images. Each group of machine learning algorithms was subjected to a total of ten iterations to improve accuracy. In Group 1, a K-neighbourhood regression was utilized, while in Group 2, a Decision tree technique was used. G power calculation determines the required samples for this analysis (Abbas et al., 2021). The analysis' minimum power is set at 0.8, while the maximum allowed error is set to 0.5. Breast cancer detection is carried out using a K-neighbourhood regression tree with a sample size of 20 people divided into two groups and assessed. The dataset was downloaded from Kaggle, a subsidiary of Google LLC. The parameters include tumor size, histologic type, grade, axillary lymph node status, and tumor growth function.

The independent variables are texture_mean, area_mean, smoothness_mean, concave_mean, concavity_mean, compactness_mean, perimeter_mean, fractal_dimension_mean and the dependent variables are texture_se, area_se, smoothness_se, concave_se, concavity_se, compactness_se, perimeter_se,

fractal_dimension_se. The training dataset is 70%, and the testing dataset is 30%. The Malignant, Benign, and Normal Images are taken. These are the different types of intensity values (Malignant, Benign, or Normal) for 5 five sample images taken, based on which the detection of tumor is determined, and this is shown in Table 1 for the sample images considered.

Table 1. Different types of intensity value (Malignant, Benign or Normal) for the 5 sample images taken

Image No	Malignant	Benign	Normal
Image 1	0	1	0
Image 2	0	0	1
Image 3	1	0	0
Image 4	1	0	0
Image 5	0	1	0

a. *Decision tree Classification :*The Decision Tree technique generates classification rules based on trees. It organizes cases or predicts the values of a dependent (target) variable based on the importance of independent (predictor) variables. Validation tools for exploratory and confirmatory classification analyses are included in the approach. Decision trees use a top-down approach to data. Tumors can be detected by looking for suspicious areas with low contrast to their surroundings and how the extract looks in tumors. The tumor detection system employs breast cancer novel preprocessing segment and filtering, with only edges identified during segmentation. In the decision tree algorithm, the image level is low.

b. *K-neighborhood regression* The K-nearest neighbor algorithm is used for pattern recognition and grouping. It's a popular tool in predictive analysis. The K-NN algorithm finds current data points closest to new data as it arrives. The data samples are assumed to be represented in a metric space by K-NNs. The calculation will find the new data samples from K adjacent neighbors at that point. The data points are in metric space; how the distance will be measured is a significant concern. Euclidean distance is the most widely used among many options. The calculation will then check the quantity of knowledge that focuses on each class among these K neighbors and relegate. Software specifications aim to identify the resources that must be installed on a device for an application to run. Before the software can be installed, it must install its particular prerequisites. The following are the software requirements at a bare minimum: Matlab version 2019 for Windows.

2.3. Statistical Analysis

In addition to experimental research, the work was statistically analyzed using IBM's Statistical Package for Social Sciences (SPSS). The software calculates the average, standard deviation, and standard error mean. To compare the two groups, an independent sample T-test analysis was used. We have used "IBM SPSS Independent T-test Analysis" to do statistical analysis for two independent variables utilizing a support vector machine and a decision tree.

3. Results

Figure 1(a) shows the image from the input database for the Novel preprocessing segment. Fig. 1(b) gives the Novel preprocessing segment to remove the noise with an adaptive median filter from the image in the input image. Image is segmented threshold value converted into new pixels and formed the tumor with the line formation shown in Fig. 1 (c). The tumor will be detected in the breast and can identify cancer with that image, whether malignant, benign, or average, as represented in Figure 1(d).

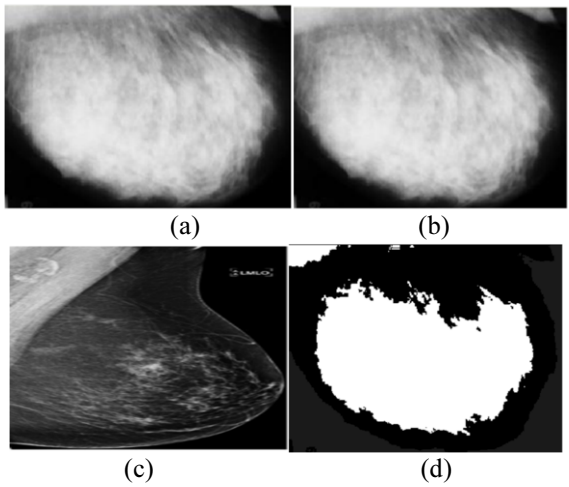

Figure 1. Simulation results of Regression tree based breast cancer detection system (a) Input image (b) Preprocessing to remove the noise (c) Segmented image (d) Tumor detected image

The breast cancer detection accuracy of the K-neighbourhood regression was much superior to the Decision Tree classification, as shown in Table 2. The detection accuracy of the k-nearest method was much superior to that of the Decision tree 0.000 in all iterations. The K-neighbourhood regression algorithm performed substantially better than decision tree regression, as evidenced by the above values. Table 3 shows that k-neighbourhood has higher accuracy than decision tree regression based on group statistics. Compared to the decision tree 80.5 and std, the mean of k-neighborhood is 87.5. Error The decision tree has a standard of 0.15590, and the k-neighborhood has a mean of 0.2252.

Table 2. Experimental Results of the accuracy achieved for Decision tree and K-neighbourhood regression

	Accuracy	
Image No	**Decision Tree**	**K-neighborhood regression**
1	86.5374	87.5964
2	86.5322	87.5912
3	86.5460	87.5964
4	86.5350	87.5912
5	86.5395	87.5945
6	86.5370	87.5992
7	86.5310	87.5983
8	86.5297	87.6112
9	86.5310	87.6011
10	86.5356	87.6112

Table 3. Comparison of mean accuracy, Std Deviation, Std error of the K-neighbourhood Regression with the Decision Tree.

Algorithm	N	Mean	Std.Deviation	Std.Error Mean
Decision Tree	10	80.535	.0049	.0015
K-neighbourhood Regression	10	87.599	.0071	.0022

An Independent Sample T-test was performed with the value obtained from the iterations. Significance values and group statistics of proposed and existing algorithms are shown in Table 4, whereas t-test equality is calculated. A confidence interval of the difference as lower and upper values range as shown in Table 4. Table 4 specifies an independent sample T-test comparing significance level with value $p<0.05$ and fixed confidence interval (k-neighborhood regression appears to perform significantly better than decision tree regression with $p = 0.000$).

Table 4. Statistical analysis of mean and standard deviation for Decision tree and K-neighbourhood regression

	F	Sig.	T	f	Sig.(2 - tailed)	Mean Differ	Std.Error Differ	95% Confidence interval of the lower	95% Confidence interval of the upper
Equal Variance assumed	875	362	2578.478	8	.000	7.06362872	.0027394566	7.0693841	7.05787334
Equal Variance not assumed			2578.478	6.013	.000	7.06362872	.0027362872	7.0694357	7.05782172

Figure 2 shows a bar chart that compares the mean accuracy of breast cancer diagnosis using decision tree regression against k-neighborhood regression. The mean

accuracy of the Decision tree is better than KNN, and the standard deviation of the Decision tree is slightly better than KNN. The decision tree offers the most consistent results with the slightest variance. With its standard deviation, k-neighborhood regression looks to yield the most varied results; nonetheless, there is a statistically significant difference between decision tree and k-neighborhood regression. The graph shows that K-neighbourhood regression outperforms decision tree regression in accuracy.

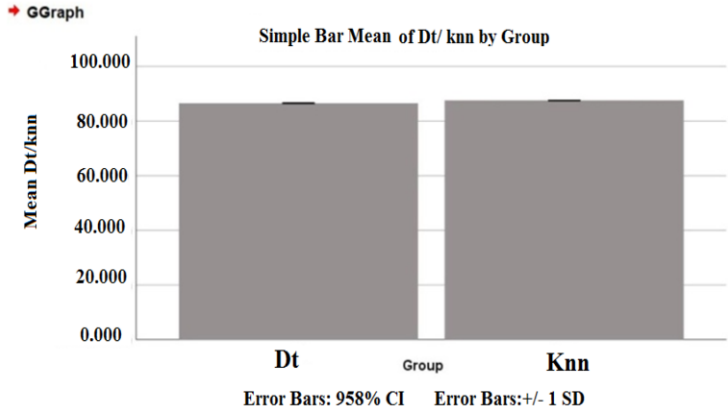

Figure 2. Comparison of Decision tree and KNN method in terms of mean accuracy. X-Axis: Decision tree vs KNN Algorithm and Y-Axis: Mean accuracy of detection ± 1 SD.

4. Discussion

Based on the significance value obtained in statistical analysis, the K-neighbourhoodregression algorithm has significantly superior accuracy than decision tree regression in this study. Table 3 presents the importance of Mean Std. Deviation and Std for the K-neighborhood regression and decision tree regression techniques. Error Mean. Group Statistics for the mean accuracy, standard deviation, and standard error mean have been calculated for the T-Test sample. The sample size of each algorithm is 10. Figure 2 indicates that K-neighbourhood regression has a lower mean accuracy error (87.5%) than decision tree regression (85.6 %).

In this research, breast cancer detection using RFSS-based features with the algorithm of thermal images the classification of accuracy is 72%.[16]. In this research, the cancer of novel genetic algorithms in component selection using the KNN method for prognosis with the classification of accuracy is 76%.[17]. The paper presented the mammography classification using SVM and KNN regression of cancer with a classification accuracy is 94%.[18]. This research was a breast cancer diagnosis, and the prognosis with the artificial intelligence technique of the classification to accuracy is 89%.[19]. In this paper, the accuracy enhanced lung cancer prognosis improved patient survivability using the Gaussian classifier system with the category of accuracy is 79%.[7]. In this research, the cancer of the detection of mammographic mass using gaussianHermite of all algorithms with the classification of accuracy is 88%.[20].

Disease-related inability to carry out typical everyday activities. Treatments are more likely to die as a result of poor overall health. The noise from the tumor will be removed by breast cancer treatment. The tumor will be detected at a particular point.

Identify the type of tumor in this; it is benign cancer. The K-neighbourhood regression has a high accuracy value and has the image quality to identify the tumor.

5. Conclusion

Even though the study's results are superior in both experimental and statistical analysis, this work has several drawbacks. Image processing is employed to improve accuracy in novel breast cancer prediction using a K-neighborhood regression compared to decision tree regression. Breast cancer survivors with functional impairments face several challenges. Hence it is concluded that K-neighbourhood Regression is significantly better than the Decision Tree algorithm in detecting tumors in breast images. This research work shows that the accuracy for the detection of breast cancer tumors using a K-neighbourhood regression is significantly better than the Decision Tree algorithm.

References

[1] Chanda PB, Sarkar SK. Detection And Classification Technique Of Breast Cancer Using Multi-Kernal SVM Classifier Approach. *2018 IEEE Applied Signal Processing Conference (ASPCA)*. Epub ahead of print 2018. DOI: 10.1109/aspcon.2018.8748810.

[2] Rastghalam R, Pourghassem H. Breast cancer detection using the spectral probable feature on thermography images. *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP)*. Epub ahead of print 2013. DOI: 10.1109/iranianmvip.2013.6779961.

[3] Bagchi S, Tay KG, Huong A, et al. Image processing and machine learning techniques used in a computer-aided detection system for mammogram screening - a review. *International Journal of Electrical and Computer Engineering (IJECE)* 2020; 10: 2336.

[4] Zain ZM, Alshenaifi M, Aljaloud A, et al. Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis. *International Journal of Advances in Intelligent Informatics* 2020; 6: 313.

[5] Yulianti A, Labellapansa A, Kadir EA, et al. Image Segmentation of Nucleus Breast Cancer using Digital Image Processing. *Proceedings of the Second International Conference on Science, Engineering, and Technology*. Epub ahead of print 2019. DOI: 10.5220/0009105900640067.

[6] Wu YC. Classification of Microcalcifications of the Diagnosis of Breast Cancer Using Artificial Neural Networks. Epub ahead of print 1996. DOI: 10.21236/ada326304.

[7] R K, R GR. Accuracy Enhanced Lung Cancer Prognosis for Improving Patient Survivability Using Proposed Gaussian Classifier System. *J Med Syst* 2019; 43: 201.

[8] P. AP, Palacios P. A, Suzuki Y. An Immune Algorithm that Uses a Master Cell for Component Selection for the kNN Method. *2019 Global Medical Engineering Physics Exchanges/ Pan American Health Care Exchanges (GMEPE/PAHCE)*. Epub ahead of print 2019. DOI: 10.1109/gmepe-pace.2019.8717318.

[9] Alpaslan N, Kara A, Zencir B, et al. Classification of breast masses in mammogram images using KNN. *2015 23nd Signal Processing and Communications Applications Conference (SIU)*. Epub ahead of print 2015. DOI: 10.1109/siu.2015.7130121.

[10] Vocaturo E, Zumpano E. Multiple Instance Learning approaches for Melanoma and Dysplastic Nevi images classification. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Epub ahead of print 2020. DOI: 10.1109/icmla51294.2020.00217.

[11] Athani S, Joshi S, Ashwath Rao B, et al. Parallel Implementation of kNN Algorithm for Breast Cancer Detection. *Evolution in Computational Intelligence* 2021; 475–483.

[12] Nusantara AC, Purwanti E, Soelistiono S. Classification of Digital Mammogram based on Nearest-Neighbor Method for Breast Cancer Detection. *International Journal of Technology* 2016; 7: 71.

[13] Zerouaoui H, Idri A. Reviewing Machine Learning and Image Processing Based Decision-Making Systems for Breast Cancer Imaging. *J Med Syst* 2021; 45: 8.

[14] Ali Y, Hamed S. Early Breast Cancer Detection using Mammogram Images: A Review of Image Processing Techniques. *Biosciences Biotechnology Research Asia* 2015; 12: 225–234.

[15] Cruz M, Bernardino J. Data Mining Techniques for Early Detection of Breast Cancer. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Epub ahead of print 2019. DOI: 10.5220/0008346504340441.

[16] Darabi N, Rezai A, Hamidpour SSF. BREAST CANCER DETECTION USING RSFS-BASED FEATURE SELECTION ALGORITHMS IN THERMAL IMAGES. *Biomedical Engineering: Applications, Basis and Communications* 2021; 2150020.

[17] Pawlovsky AP, Matsuhashi H. The use of a novel genetic algorithm in component selection for a kNN method for breast cancer prognosis. *2017 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*. Epub ahead of print 2017. DOI: 10.1109/gmepe-pahce.2017.7972084.

[18] Sonar P, Bhosle U, Choudhury C. Mammography classification using modified hybrid SVM-KNN. *2017 International Conference on Signal Processing and Communication (ICSPC)*. Epub ahead of print 2017. DOI: 10.1109/cspc.2017.8305858.

[19] Jain A, Jain A, Jain S. *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*. World Scientific, 2000.

[20] Eltoukhy MM, Elhoseny M, Hosny KM, et al. Computer aided detection of mammographic mass using exact Gaussian–Hermite moments. *Journal of Ambient Intelligence and Humanized Computing*. Epub ahead of print 2018. DOI: 10.1007/s12652-018-0905-1.