Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 5, July, 2021: 4710 - 4718

Research Article

Breast Cancer Detection Using Various Classification Algorithms

Battula.Asha, M.tech,
Dept of CSE, VRSEC
battula.asha99@gmail.com

D.Rajeswara Rao

Dept of CSE, VRSEC

Hodcse@vrsiddhartha.ac.in

Abstract:

Machine learning algorithms are playing an important role, nowadays, to extract the features of any application. Here used these mechanisms to detect breast cancer by considering the standard dataset. Various classification techniques are considered to find the performance of the detected results. Initially removed the noise over a data set that is missed data and then extracted the features and next applied the classifiers such as SVM, KNN, Random Forest, decision tree, etc. The classifier efficiency was evaluated by considering the characteristics like true positive, false positive, ROC curve, etc. The main strength of this work is experimental results which are shown by considering the standard dataset from Kaggle.

Keywords: Machine learning, Breast Cancer, Kaggle, SVM, Random forest, KNN, Classifier

1.Introduction

Machine learning and data mining are popular tools for knowledge discovery. Machine learning mechanisms used to detect tumor and characteristics of it. One problem here is that class in balancing while doing training the data [1]. One of the major diseases is breast cancer in women. In general, most researchers considering images for detecting breast cancer. Sometimes false detection might happen over mammogram images [2]. Hence here considered the alternative input such as dataset which can help to find out the reliable predictions [3] of breast cancer.

Machine learning, Data mining, and Deep learning techniques are important for making classification and clustering and analyzing the data such as either image data or raw data. To analyze the data, generally considered either structure or unstructured data. In the proposed work, we

considered structured data for detecting breast cancer characteristics.

To achieve the proposed objectives here we have used Naïve biased, Support vector Machine, KNN, Decision tree, Random forest approaches. Analyzed the results and shows the performance of various approaches. The main initial objective in the structured data is to identify the noise and remove
the noise if it has existed. Next required to classify the data by considering the characteristics of the specified attributes.

In 2019 according to WHO, bosom malignant growth represents 2.09 million cases and 627000 passings internationally. It is the most widely recognized malignant growth in ladies in India and records for 14% of all diseases in ladies. The extended frequency of patients with malignant growth in India among guys was 679,421 (94.1 per 100,000) and among females 712,758 (103.6 per 100,000) for the year 2020.

The next part of the paper is arranged as follows. The literature survey is presented in section2 2. Section 3 consists of the proposed methodology. Section 4 presented the results and performance analysis.

2. Literature Survey:

Suresh, et.al [1] proposed a framework to recognize if the tumor is there in the MRI picture. To accomplish the target the framework utilizes AI calculations, for example, K-means and backing vector machines.

Here they referenced that tumors have been ordered in two different ways, for example, dangerous and kindhearted.

Anji Reddy, et.al [4] present a strategy to identify bosom disease by considering AI strategies. The proposed strategy shows the exactness of the framework by thinking about different components.

Yusif, et. al[6] introduced an article where utilizes profound learning procedures to finding the cerebrum tumor by considering the UCI dataset. In this work, the creators have considered the low differentiation pictures for identifying the edges of the tumors. The creators have referenced that high assignment targets of profound learning are Medical, PC vision, Natural language preparing, and neural organization.

Li Shen et. al [11] utilizes mammography screening to recognize bosom malignant growth utilizing profound learning methods. In this work, the creators have distinguished the ROI of the picture that is the required bit that identified with the illness. To accomplish this they utilized profound learning strategies for distinguishing the necessary items from the picture.

G. Czamota et al [15] examined the Quantitative Ultrasound with Texture Predictors of Breast Tumor. They have broken down the Quantitative Ultrasound Predictors. The impediment is that the proposed framework is exceptionally lethargic. In the clinical oversight of patients having bosom malignant growth (BRCA) a fundamental part is played by imaging biomarkers (IBS). It very well may be utilized in every one of the phases of malignancy. Examining the patient, distinguishing proof of infection and its treatment assessment are the capacities that are completed with the assistance of imaging biomarkers.

In 2019, Y. Jiang, et al [17] explored the relationship of histopathological pictures to distinguish chest illness. Convolution neural association becomes possibly the most important factor here. They also used the little SE-ResNet module. The fundamental point behind this work was the advancement of another convolution neural association model.

E. Kontopodis, et al [18] investigated the limit played by setup-based regular engravings despite plan less natural engravings which are open in graphical construction. Evaluation results spread out arrangement less natural engravings transforms into another and strong substitute.

A. Rampun, et al [19] assessed, how the chest illness was facilitated from the picture of the chest. For the satisfaction of assessment, the maker uses gathered design convolution neural association. Here, quantitative results of starting examination were gotten the message out about for individuals overall. It was by then obtained as decision help in the association of information structure settings which is identified with the association of Breast Cancer.

M. Gupta [20] set up a blended plan for aiding the appraisal of sickness which happened in the chest. For its arrangement, the examiner completes the advancement of progressive least squares. Disease-like danger turns out to be notable as the contamination which makes considering the blend of different disorders. Dangerous development is a kind of disorder wherein the body cells become uncontrolled and reach out in an uncontrolled manner. It might be shipped off other body parts. Chest threat has four stages.

B. Dai, et al [21] perform research in the ID of sickness which is related to the chest. They execute methodologies related to discretionary woods in the affirmation association of chest illness. In current, the presence of enormous data and AI gives a basic obligation to the space of the clinical benefits division.

3. Proposed method

The proposed system uses to detect breast cancer from the considered dataset. In the process, we have considered some properties which are related to cancer diseases such as clump thickness, size of clump, shape, marginal adhesion, epithetical size, bare nucleoli, etc. At the first stage of the process, needs to do preprocessing. In this module, identify the missed data in the data fields of a dataset and update that field. The next module is required to extract the feature of the dataset and then make classifications. Fig 1 shows the proposed system process.

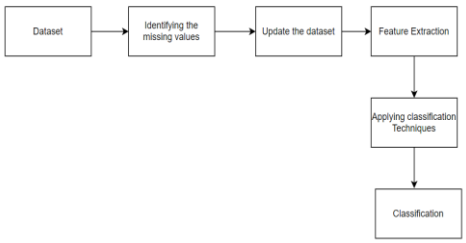

Fig 1. Proposed system process

The growth in machine learning causes to increase in the role of deep learning in all applications. This medical community also showing the importance to use machine learning. Feature extraction is done from the preprocessed data as shown in fig 3. Next for doing classification here used K means results on different techniques such as SVM, Linear regression, Random Forest, K nearest Neighbour, etc [12].
Here considering the K Nearest Neighbor, where KN value is 16 based on the parameter calculating the Mean Absolute Error values and then updating the dataset.he proposed system process is as follows.

Step 1: Read Dataset

Step 2: identified the noise i.e missed data
Step 3: Update the dataset by handling the Noise data Considered the random variable function to update the dataset
Step 4: Feature Extraction by considering properties such as clump thickness, size of clump, shape, marginal adhesion, epithetical size, bare nucleoli, etc.
Step 5: Apply various classification Techniques like SVM, Random Forest, Naïve Bias, KNN, Decision Tree, etc
Step 6: Analyze the system results

Pseudocode of the ROC for all the considered classifiers was shown here

```
print("Naive Bayes AUC = ",roc_auc_nb)
print("K Nearest Neighbors AUC = ",roc_auc_knn)
print("Logistic Regression AUC = ",roc_auc_lr)
print("Support Vector Machine AUC = ",roc_auc_svm)
print("Decision Trees AUC = ",roc_auc_dt)
print("Random Forest AUC = ",roc_auc_rfc)
print("MLP AUC = ",roc_auc_mlp)
```

```
Naive Bayes AUC =  0.9740030843798193
K Nearest Neighbors AUC =  0.9833663802599693
Logistic Regression AUC =  0.9889843577880059
Support Vector Machine AUC =  0.9892046706322979
Decision Trees AUC =  0.9622163472130427
Random Forest AUC =  0.9841374752148049
MLP AUC =  0.9881031064111038
```

4. Results and Analysis

The proposed approach helps to compare the performance evaluation by considering various machine learning approaches. To implement this application uses, Python, Jupiter platform and considered the standard dataset from Kaggle. As a part of preprocessing, initially identified the noise data. Fig 1. Shows the missing data in the considered dataset. Bare nucleoli attribute values are missing that as highlighted in Fig 1. The considered data set has various attributes regards breast cancer such as clump size, thickness, uniformity, the difference between normal nucleoli and bare nucleoli, etc.

	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bare_nucleoli	bland_chromatin	normal_nucleoli	mitoses	class
0	8	4	5	1	2	?	7	3	1	4
1	6	6	6	9	6	?	7	8	1	2
2	1	1	1	1	1	?	2	1	1	2
3	1	1	3	1	2	?	2	1	1	2
4	1	1	2	1	3	?	1	1	1	2
5	5	1	1	1	2	?	3	1	1	2
6	3	1	4	1	2	?	3	1	1	2
7	3	1	1	1	2	?	3	1	1	2
8	3	1	3	1	2	?	2	1	1	2
9	8	8	8	1	2	?	6	10	1	4
10	1	1	1	1	2	?	2	1	1	2
11	5	4	3	1	2	?	2	3	1	2
12	4	6	5	6	7	?	4	9	1	2
13	3	1	1	1	2	?	3	1	1	2
14	1	1	1	1	1	?	2	1	1	2
15	1	1	1	1	1	?	1	1	1	2

Fig 1. Missed data in the dataset

Fig 2 shows the updated missed data column with random values. The updated the dataset will become an input to the process for evaluating the performance.

	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bland_chromatin	normal_nucleoli	mitoses	class	bare_nucleoli
0	8	4	5	1	2	7	3	1	4	7
1	6	6	6	9	6	7	8	1	2	8
2	1	1	1	1	1	2	1	1	2	1
3	1	1	3	1	2	2	1	1	2	1
4	1	1	2	1	3	1	1	1	2	1
5	5	1	1	1	2	3	1	1	2	1
6	3	1	4	1	2	3	1	1	2	1
7	3	1	1	1	2	3	1	1	2	1
8	3	1	3	1	2	2	1	1	2	1
9	8	8	8	1	2	6	10	1	4	8
10	1	1	1	1	2	2	1	1	2	1
11	5	4	3	1	2	2	3	1	2	1
12	4	6	5	6	7	4	9	1	2	8
13	3	1	1	1	2	3	1	1	2	1
14	1	1	1	1	1	2	1	1	2	1
15	1	1	1	1	1	1	1	1	2	2

Fig 2. Updated Data set

Fig 3. Shows the classification of the data set by considering the K nearest neighbors. Here dataset has been classified into two cases i.e class 0 and class 1 Where the KN value is 16. Here considered the dump thickness and size uniformity factors.

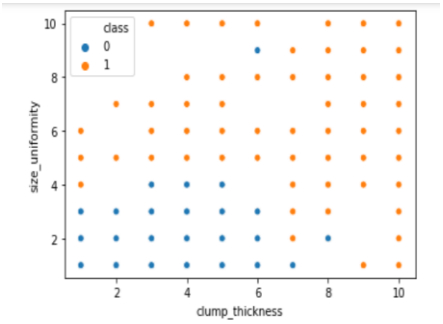

Fig 3. Class specification to the dataset

Fig 4 shows the performance of the system by representing a confusion matrix. True positive and false positive factors have been considered to present the confusion matrix[7].

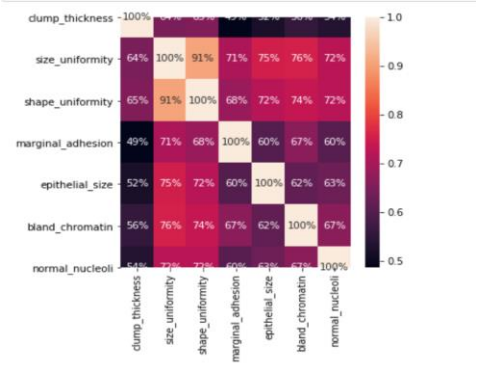

Fig 4. Correlation Matrix

Fig 5 gives information about the efficiency of the application by considering the entropy and Gini values. By considering these values evaluated the accuracy of the system results.

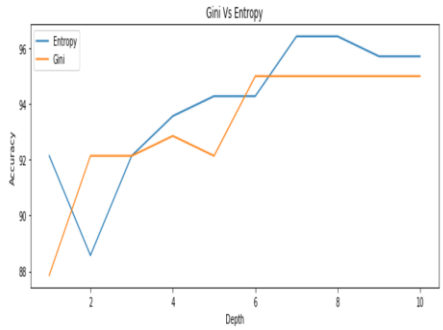

Fig 5. Accuracy Evaluation

Fig 6 snd Fig 7 shows the comparison of classification with the existing SVM, Naive Bayesian, and Random Forest, KNN, and decision tree are used. Here we considered the true positive and false positive characteristics to show the test results.

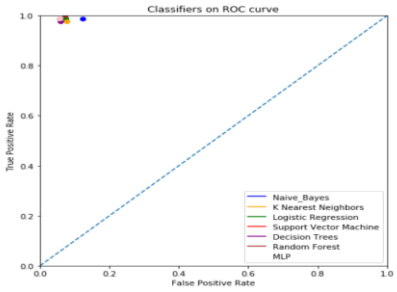

Fig 6. , False positive & True Positive graph

Fig 7 shows the test results of the proposed system application by considering the true positive and true negative values over classification techniques for breast cancer feature analysis. Fig 7 replicates our objectives.

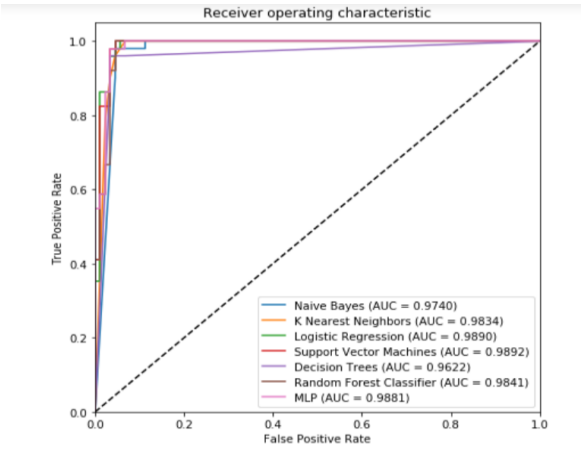

Fig 7. True positive and True negative test.

5. Conclusion

Breast cancer disease is one of the major diseases of all women. We can predict the disease at earlier stages itself by analyzing the characteristics of the disease. Hence in this work, we have evaluated the disease characteristics by considering various attributes of the disease like clump thickness, size, nucleoli, etc. Here we used a machine approached for evaluating the disease characteristics. This work shows variances among the machine learning approaches while showing the performance.

References

[1] D. Suresha, N. Jagadisha, H. S. Shrisha and K. S. Kaushik, "Detection of Brain Tumor Using Image Processing," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 844-848, doi: 10.1109/ICCMC48092.2020.ICCMC-000156.

[2] A. Gupta, D. Kaushik, M. Garg and A. Verma, "Machine Learning model for Breast Cancer Prediction," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 472-477, doi: 10.1109/I-SMAC49090.2020.9243323.

[3] S. S. Prakash and K. Visakha, "Breast Cancer Malignancy Prediction Using Deep Learning Neural Networks," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 88-92, doi: 10.1109/ICIRCA48905.2020.9183378.

[4] K. Shaukat, I. Nawaz, S. Aslam, S. Zaheer and U. Shaukat, "Student's performance in the context of data mining," *2016 19th International Multi-Topic Conference (INMIC)*, 2016, pp. 1-8, doi: 10.1109/INMIC.2016.7840072.

[5] Y. A. Hamad, K. Simonov and M. B. Naeem, "Brain's Tumor Edge Detection on Low Contrast Medical Images," 2018 1st Annual International Conference on Information and Sciences (AiCIS), 2018, pp. 45-50, doi: 10.1109/AiCIS.2018.00021.

[6] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 98-103, doi: 10.1109/ICACCCN.2018.8748777.

[7] S. Laghmati, A. Tmiri and B. Cherradi, "Machine Learning based System for Prediction of Breast Cancer Severity," 2019 International Conference on Wireless Networks and Mobile Communications (WINCOM), 2019, pp. 1-5, doi: 10.1109/WINCOM47513.2019.8942575.

[8] N. Kaur and M. Sharma, "Brain tumor detection using self-adaptive K-means clustering," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 1861-1865, doi: 10.1109/ICECDS.2017.8389771.

[9] S. Dencks et al., "Relative Blood Volume Estimation from Clinical Super-Resolution US Imaging in Breast Cancer," 2018 IEEE International Ultrasonics Symposium (IUS), 2018, pp. 1-4, doi: 10.1109/ULTSYM.2018.8580013.

[10] S. Bharati, M. A. Rahman and P. Podder, "Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA," 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), 2018, pp. 581-584, doi: 10.1109/CEEICT.2018.8628084.

[11] N. Kumar, G. Sharma and L. Bhargava, "The Machine Learning based Optimized Prediction Method for Breast Cancer Detection," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1594-1598, doi: 10.1109/ICECA49313.2020.9297479.
[12] M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 121-124, doi: 10.1109/ICOEI.2019.8862533.
[13] S. N. Singh and S. Thakral, "Using Data Mining Tools for Breast Cancer Prediction and Analysis," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777713.
[14] S. J. Malebary and A. Hashmi, "Automated Breast Mass Classification System Using Deep Learning and Ensemble Learning in Digital Mammogram," in IEEE Access, vol. 9, pp. 55312-55328, 2021, doi: 10.1109/ACCESS.2021.3071297.
[15]S. Dencks et al., "Relative Blood Volume Estimation from Clinical Super-Resolution US Imaging in Breast Cancer," 2018 IEEE International Ultrasonics Symposium (IUS), 2018, pp. 1-4, doi: 10.1109/ULTSYM.2018.8580013.
[16] S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.
[17] M. R. Basunia, I. A. Pervin, M. Al Mahmud, S. Saha and M. Arifuzzaman, "On Predicting and Analyzing Breast Cancer using Data Mining Approach," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1257-1260, doi: 10.1109/TENSYMP50017.2020.9230871.
[18] P. Chauhan and A. Swami, "Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1-8, doi: 10.1109/ICCCNT.2018.8493927.
[19] B. Fu, P. Liu, J. Lin, L. Deng, K. Hu and H. Zheng, "Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data," in IEEE Transactions on Biomedical Engineering, vol. 66, no. 7, pp. 2053-2064, July 2019, doi: 10.1109/TBME.2018.2882867.
[20] D. Kaushik, B. R. Prasad, S. K. Sonbhadra and S. Agarwal, "Post-Surgical Survival Forecasting of Breast Cancer Patient: A Novel Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 37-41, doi: 10.1109/ICACCI.2018.8554745.
[21] M. Ma, Y. Shi, W. Li, Y. Gao and J. Xu, "A Novel Two-Stage Deep Method for Mitosis Detection in Breast Cancer Histology Images," 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3892-3897, doi: 10.1109/ICPR.2018.8546192.
[22] R. Dhanya, I. R. Paul, S. Sindhu Akula, M. Sivakumar and J. J. Nair, "A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1049-1055, doi: 10.1109/ICCS45141.2019.9065563.
[23] A. Rampun, B. W. Scotney, P. J. Morrow and H. Wang, "Breast Mass Classification in Mammograms using Ensemble Convolutional Neural Networks," 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), 2018, pp. 1-6, doi: 10.1109/HealthCom.2018.8531154.
[24] Y. Xiao, J. Wu, Z. Lin and X. Zhao, "Breast Cancer Diagnosis Using an Unsupervised Feature Extraction Algorithm Based on Deep Learning," 2018 37th Chinese Control Conference (CCC), 2018, pp. 9428-9433, doi: 10.23919/ChiCC.2018.8483140.

[25] M. Gupta and B. Gupta, "An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP)," 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1-3, doi: 10.1109/IC3.2018.8530572.