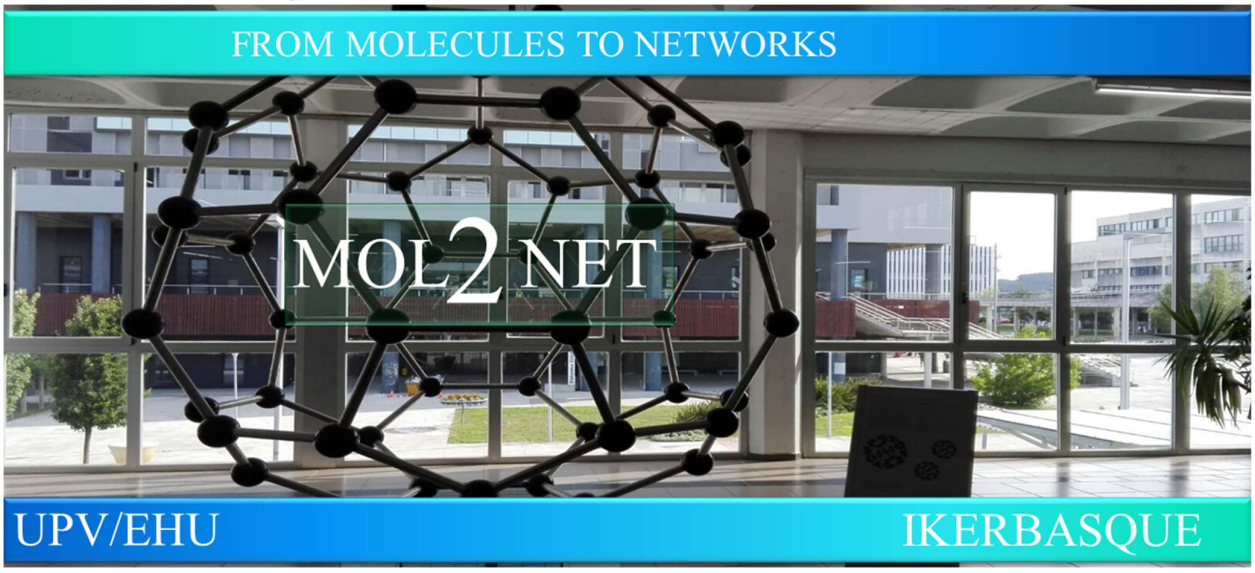
MOL2NET'22, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 8th ed.



Breast Cancer Diagnosis Using Machine Learning Techniques

Samreen Naeem [*, a], *and Aqib Ali* [a]

[a] *College of Automation, Southeast University, Nanjing, China.*

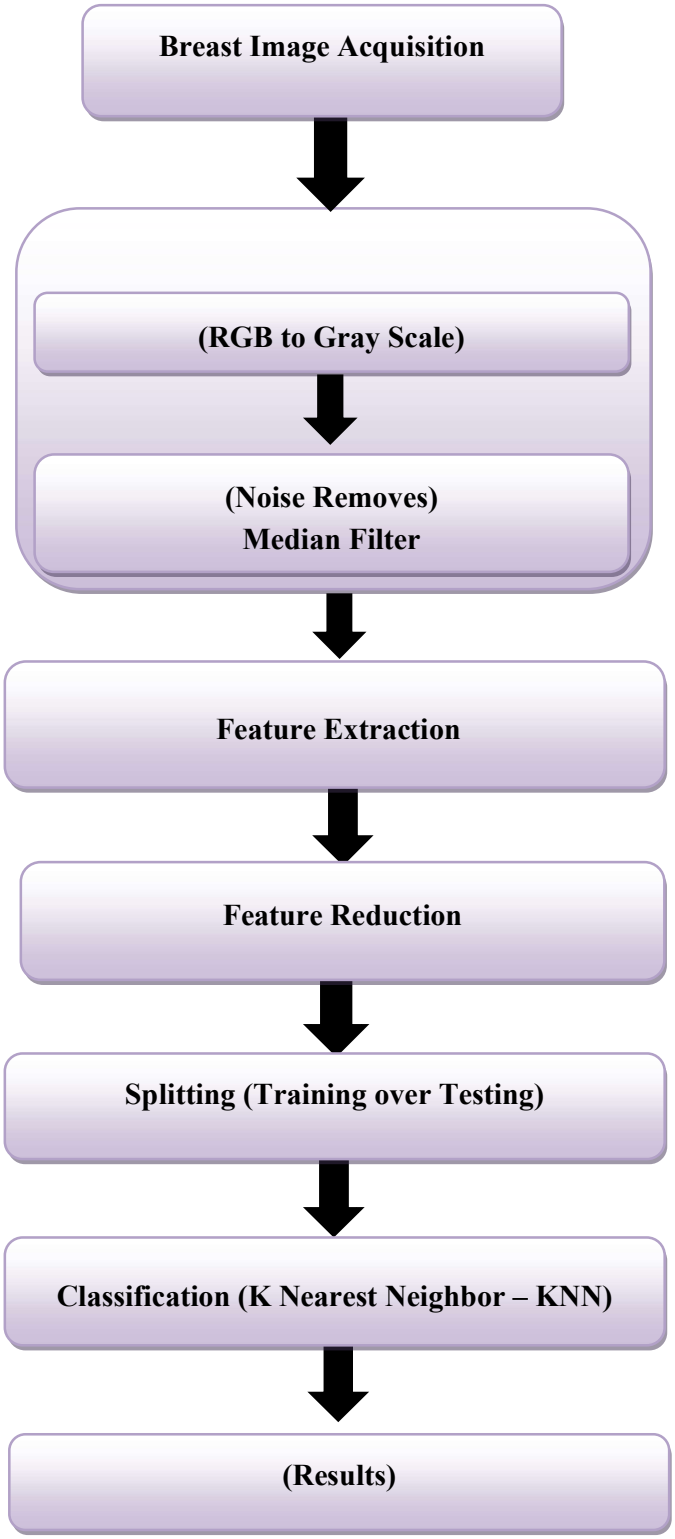
. * *Corresponding author: samreencsit@gmail.com*

Abstract.

Computers can analyze information faster than people but cannot make decisions. Computers of today are acquiring machine learning techniques to enhance analysis and prediction. These techniques enable expert assistance systems and enhance computer decision-making. Machine learning algorithms are assisting medical professionals in making rapid diagnoses thanks to their successful classification and diagnostic capabilities. Machine learning may be effective and is used more frequently in cancer diagnosis. The second most common disease in the world and the main reason for death among women is breast cancer. Like other malignancies, early identification of breast cancer lowers mortality. Machine learning techniques assist in diagnosing breast cancer, which calls for specialized human knowledge. With machine learning, computers can swiftly identify patterns in complex and large data sets. Due to these qualities, machine learning is frequently used to detect breast cancer.

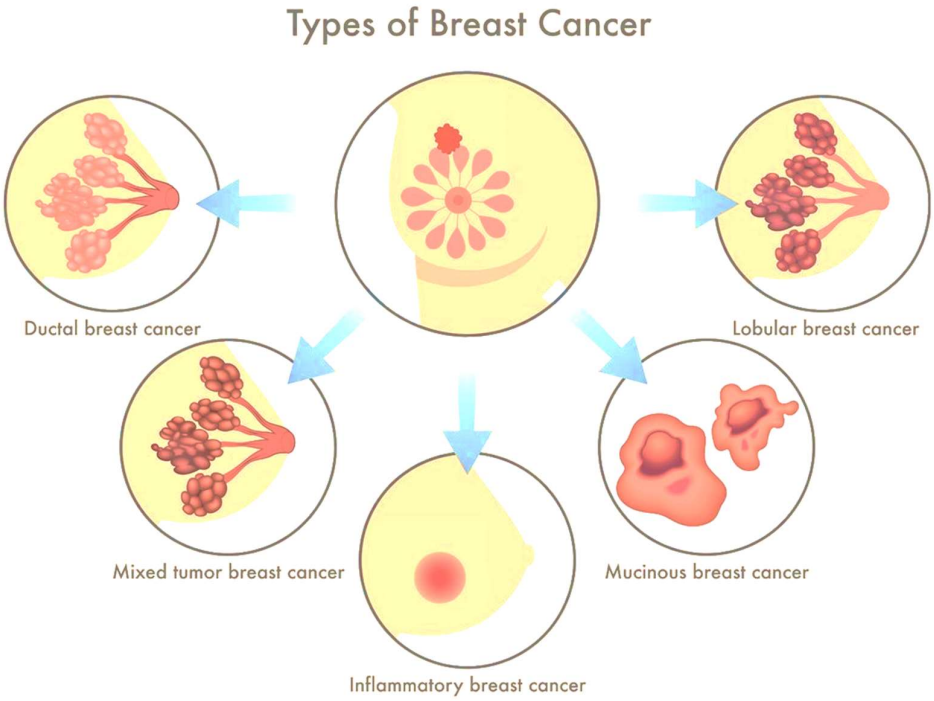
Keyword: Breast Cancer, Classification, KNN, Machine Learning.

Graphical Abstract



Introduction

Although modern computers can digest information more quickly than people, they are less capable of making decisions. Because of this, several machine-learning approaches have been created and are still being created to help computers do better analyses and reach better conclusions. The data may be used to extract meaning and make predictions using a variety of approaches, including clustering, classification algorithms, decision trees, and artificial neural networks [1]. With their adequate capacities for categorization and diagnosis, machine learning methods are used more often to assist medical practitioners in diagnosing illnesses.



The leading cause of cancer-related mortality among women globally, particularly those between the ages of 40 and 49, is breast cancer [2]. With 12.6% of the 1*.1 million cancer cases discovered worldwide in 2020, it is the second most common cancer kind after lung cancer. The breast tissue, particularly in the milk ducts and glands, manifests as tiny tumors or lumps. The mass is benign if it is smooth and has distinct edges, whereas it is malignant or cancer-risky if it has rough borders and a rough structure [3]. Breast cancer early identification is crucial for lowering the mortality rate, just as it is for all other forms of cancer. Breast cancer diagnosis and diagnosis based on test findings need specialist human understanding. A breast cancer diagnosis has been successfully studied using evolving machine-learning approaches [4].

Machine learning is a subfield of artificial intelligence that includes a variety of statistical, probabilistic, and optimization approaches. Learning from past data helps computers rapidly find

patterns in complicated and massive data sets. Machine learning is often employed in cancer detection and treatment because of this capability [5]. Every year, 1.38 million brand-new instances of breast cancer are reported. While several research is being undertaken to aid in identifying breast cancer, the application of machine learning methods in the clinical area, particularly for cancer diagnosis, is growing. Clustering, artificial neural networks, support vector machines, fuzzy and artificial fuzzy logic, and hybrid approaches are often utilized in breast cancer detection, even though various machine learning algorithms are used [6].

Materials and Methods

.

Cancer develops when cells divide without proper regulation, resulting in the formation of lumps that are referred to as tumors. Both benign and malignant types of tumors exist. Malignant tumors have a high growth rate, which causes them to occupy surrounding tissue and cause harm to it. A possible indicator of breast cancer is an abnormality in breast tissue and a change in breast form or skin color. Early detection is essential in treating breast cancer, as it is with all other forms of the disease. In this investigation, which took place at the public dataset "CBIS-DDSM: Breast Cancer Image Dataset – Kaggle" for research purposes that included 600 samples with breast cancer results. These samples were shared with other researchers [7]. The information was then categorized using five distinct machine learning models and put to the test after being arbitrarily separated into a training set of 70% and a test set of 30%. An experiment was made between the test achievements of the classifiers K Nearest Neighbor – KNN, using Python, models were developed and put through their paces [8].

The mean, standard deviation, worst-case scenario, and maximum value are additional properties that may be inferred from these features. Based on these findings, a diagnostic class with the letters B (benign) and M (malignant) is established, designating whether the tumor is benign or malignant. The 600 data points, 300 falls into the benign category and 300 into the malignant category.

K Nearest Neighbor (KNN): Class is a model that enables classifying a point from those that have already been categorized following the estimated number of K nearest points. The Euclidean distance is often used while determining the locations nearest to one another. Depending on the data being examined, a different ideal K value may be chosen. Large K values lessen the separation of the borders between the classes while lessening the noise impact on the classification [9].

The data set had 600 samples, of which 70% were used for training and the remaining 30% for testing. Random selection was used to choose the samples for each purpose. After the training process was finished, the classification accuracy was assessed using the test data. The proportion of it

correctly predicted when the test classes and the classes created by the system are compared indicates the overall accuracy of the categorization performed by the system. There is a thorough analysis of four possible outcomes in the classified dataset: *True positives* are defined as the first positive sample being correctly recognized as positive (TP) [10-11].

On the other hand, a false negative (FN) is when the initial positive sample is mistakenly labeled as unfavorable while the original negative sample is damaging. When correctly classified, it is known as a true negative (TN); when the first negative sample is mistakenly labeled as positive, it is known as a false positive (FP). The confusion matrix is the matrix that shows all of these various circumstances [12].

Results and Discussion

- K Nearest Neighbor – KNN
- Time taken to build the model: 0.38 seconds
- Test mode: 70 %Training, and 30% Testing

Table 1: KNN Classifier Summary

Total Number of Instances	600	
Correctly Classified Instances	568	**94.6667 %**
Incorrectly Classified Instances	32	**5.3333 %**
Kappa statistic	0.8933	
Mean absolute error	0.053	
Root mean squared error	0.2275	
Relative absolute error	10.5958 %	
Root relative squared error	45.5021 %	

Table 2: KNN Classifier Detailed Accuracy

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	Class
0.953	0.060	0.941	0.953	0.947	0.893	0.974	**Normal**
0.940	0.047	0.953	0.940	0.946	0.893	0.977	**Abnormal**
0.947	**0.053**	**0.947**	**0.947**	**0.947**	**0.893**	**0.975**	**Weighted Avg.**

Table 3: Confusion Matrix result using KNN Classifier

Classified as	A	B
A = Normal	286	14
B = Abnormal	18	282

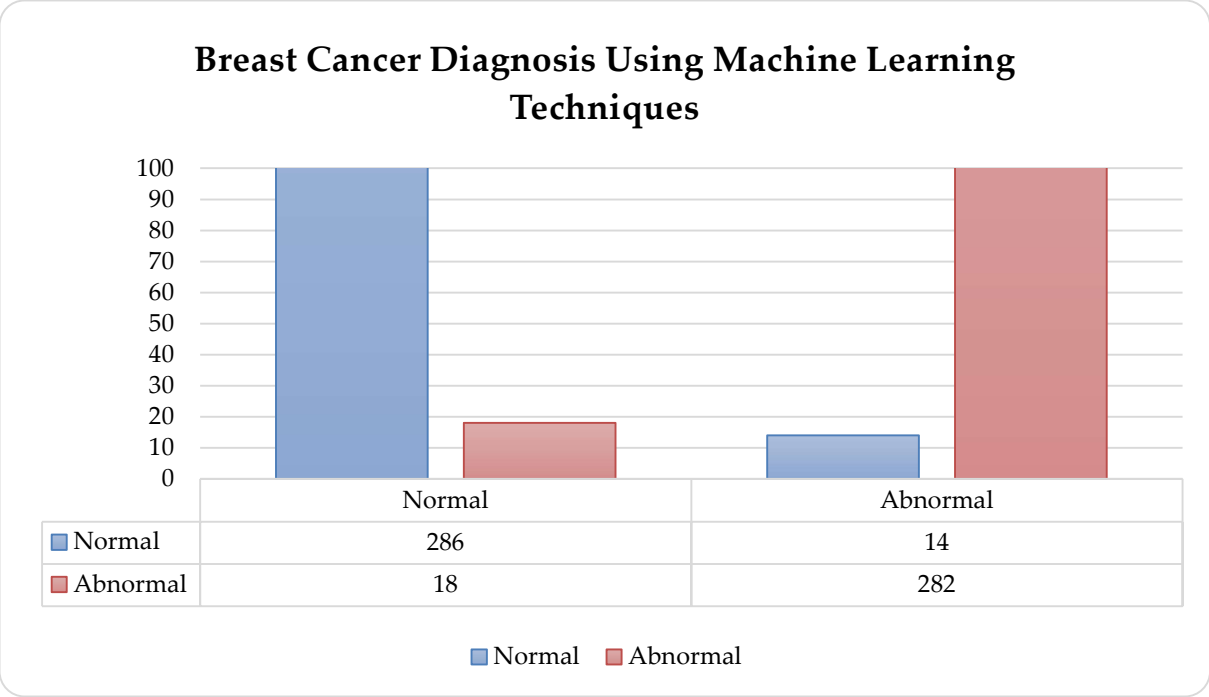


Figure 1: Accuracy of Dataset using KNN Classifier

Conclusions

In the classification produced by K Nearest Neighbor (KNN) models, 30% of the total data set, samples, were used for testing purposes. The model was trained using the remaining 70%. After going through the testing process, it was found that the logistic regression model, which had an accuracy level of 94.66%, had the best degree of success. Computers may develop the capacity to make independent decisions by analyzing human experts' information via machine learning techniques. As a result, experts begin to serve as a support system and can provide more potent results as learning data quantity and diversity increase.

References

[1]. Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., ... & Siegel, R. L. (2022). Breast cancer statistics, 2022. CA: A Cancer Journal for Clinicians, 72(6), 524-541.

[2]. Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., ... & Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast, 66, 15-23.

[3]. Gaynor, N., Crown, J., & Collins, D. M. (2022, February). Immune checkpoint inhibitors: Key trials and an emerging role in breast cancer. In Seminars in cancer biology (Vol. 79, pp. 44-57). Academic Press.

[4]. Karami Fath, M., Azargoonjahromi, A., Kiani, A., Jalalifar, F., Osati, P., Akbari Oryani, M., ... & Payandeh, Z. (2022). The role of epigenetic modifications in drug resistance and treatment of breast cancer. Cellular & Molecular Biology Letters, 27(1), 1-25.

[5]. Diamantopoulou, Z., Castro-Giner, F., Schwab, F. D., Foerster, C., Saini, M., Budinjas, S., ... & Aceto, N. (2022). The metastatic spread of breast cancer accelerates during sleep. Nature, 607(7917), 156-162.

[6]. Allugunti, V. R. (2022). Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science, 4(1), 49-56.

[7]. Yadav, R., Pande, S., & Khamparia, A. (2020, December). Breast Cancer Classification Using Convolution Neural Network (CNN). In International Conference on Advanced Informatics for Computing Research (pp. 283-292). Springer, Singapore.

[8]. Ali, A., Mashwani, W. K., Naeem, S., Uddin, M. I., Kumam, W., Kumam, P., ... & Chesneau, C. (2021). COVID-19 infected lung computed tomography segmentation and supervised classification approach.

[9]. Naeem, S., Ali, A., Qadri, S., Khan Mashwani, W., Tairan, N., Shah, H., ... & Anam, S. (2020). Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images. Applied Sciences, 10(9), 3134.

[10]. Ali, A., Qadri, S., Khan Mashwani, W., Kumam, W., Kumam, P., Naeem, S., ... & Sulaiman, M. (2020). Machine learning based automated segmentation and hybrid feature analysis for diabetic retinopathy classification using fundus image. Entropy, 22(5), 567.

[11]. ALI, A.; NAEEM, S.; ZUBAIR, M. Machine Learning Based Classification of Chronic Kidney Disease Using CT Scan Images, in Proceedings of the MOL2NET'22, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 8th ed., 26–31 December 2022, MDPI: Basel, Switzerland, doi:10.3390/mol2net-08-13903.