STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 11, Issue 1, 2024
January–April
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

Comparison of K-nearest Neighbor and Artificial Neural Network Classifiers for the Detection of Breast Cancer

Jhumi Thapa[1], Anshu Ghimire[2,*]

Abstract

Breast cancer is the most common type of cancer seen in women in the present day, which is also considered a life-threatening disease. If this cancer can be detected in its early stage it can be a lifesaver for many people around the world. Machine Learning techniques have become one of the hotspots for predicting the early diagnosis of breast cancer. This research work experiments with the two most popularly used supervised machine learning algorithms, K-nearest neighbor, and artificial neural network which will detect breast cancer by training its attributes and find out the most effective with respect to confusion matrix, accuracy, and precision. The findings indicate that the artificial neural network machine achieved superior performance, outperforming all other classifiers with an accuracy rate of 98%. All the work is done in Python programming language using Scikit-Learn library and tensor flow.

Keywords: K-nearest neighbors (KNN), artificial neural network (ANN), breast cancer, machine learning, prediction, Fine Needle Aspirate (FNA)

INTRODUCTION

Globally, breast cancer stands as the predominant illness impacting women. It is a type of cancer that is formed on the cells/tissues of the breast, though rare this cancer also affects men. As many years of technical and scientific research are taking place around the world, breast cancer is still the most common and the second leading reason for life taker of women (Ankit Verma, 2019) [1]. There are mainly two types of cancer in the world: (i) Invasive Ductal Carcinoma (IDC) the most dangerous which surrounds the entire breast tissue where approximately 80% resides in this category and (ii) Ductal Carcinoma in Situ (DCIS) which is the low percentage of all cases between 20% and 53% breast cancer patients, approximately 80%, are in this category (Saad Awadh Alanazi, 2021) [2].

There were approximately 7.8 million women who were diagnosed with breast cancer by late 2020 making it the most prevailing disease (World Health Organization, 2024) [3], and further in the year 2020, it was concluded that 2.3 million women were diagnosed with cancer and 685000 took place.

Diagnosis is the procedure of investigating the stage of cancer as either benign or malignant cases in which benign is non-cancerous 80% of those biopsied – are benign (non-cancerous) and the rest malignant are cancerous (Stony Brook Cancer Center, 2023) [4]. To diagnose breast cancer, several methods and tests are used, and they are (a) a Breast Exam in which a self-check from patients

***Author for Correspondence**
Anshu Ghimire
E-mail: anshug@nec.edu.np

[1,2]Assistant Professor, Department of Computer Science and Engineering, Nepal Engineering College, Bhaktapur, Nepal

Received Date: January 14, 2024
Accepted Date: April 09, 2024
Published Date: April 24, 2024

Citation: Jhumi Thapa, Anshu Ghimire. Comparison of K-nearest Neighbor and Artificial Neural Network Classifiers for the Detection of Breast Cancer. Journal of Artificial Intelligence Research & Advances. 2024; 11(1): 78–83p.

themselves/a clinical check from doctors takes place (b) a Mammogram in which x-rays of the breast are taken to evaluate any abnormality. (c) Breast Ultrasound: this is to check lump is either a solid mass or a fluid-filled cyst. (d) Biopsy entails extracting a portion of breast tissue for analysis. (e) Breast MRI employs radio waves and a magnet to generate internal breast images, aiding in the detection of malignancies.

In today's technological era if we want to gain better results for breast cancer with a low mortality rate then we have to move towards a computerized diagnostic system in which machine learning algorithms are used. These algorithms help for efficient prediction of cancer cells and detect the cells more accurately This research focuses on the prediction of breast cancer through the use of two supervised learning techniques: Artificial neural networks and K-nearest neighbor (KNN) for breast cancer detection.

In the Kaggle repository, we have used the Wisconsin (Diagnostic) Dataset to complete our research work (Kaggle, 2016) [5]. Here we trained the dataset for identifying breast cancer. To determine the most accurate classifier, the accuracy of both machine learning models is calculated and compared.

The following is how the paper is set up: The prior associated works are described in section 2. The procedures (methods) and materials are explained in section 3. Section 4 of the study's findings presents the findings i.e. the result obtained. Finally, section 5 provides the conclusion.

RELATED WORKS

With the evolving technology there have been a large number of machine learning algorithms for predicting and diagnosing breast cancer. The artificial neural network (ANN), Convolution Neural Network (CNN), K-nearest neighbors (KNN Network), Support Vector Machine (SVM), Logistic Regression, and Decision Tree, among others, are a few examples of machine learning techniques. Many researchers have conducted studies on breast cancer utilizing a variety of datasets, including those from Wisconsin, SEER, Kaggle, mammography pictures, and different hospitals.

Exploring different datasets authors extract and select various essential features and complete their research work (Vattsal Singhal, 2022) [6]. The research works related to breast cancer have been outlined shortly as follows.

M. Tahmooresi et al. [7] have introduced a distinctive method for the early detection of breast cancer. This approach has used several machine learning techniques like SVM, ANN, KNN, and Decision Tree (DT) for effective breast cancer detection. The findings of these researchers suggest that SVM is the most popular method used for cancer detection applications. To enhance its efficacy, SVM was utilized either alone or in combination with another method. The maximum achieved accuracy of SVM (single or hybrid) was 99.8% which can be improved to 100%.

The authors (Md. Milon Islam, 2017) [8] demonstrated a model that uses 10-fold cross-validation to get an accurate outcome. The performance of the proposed system is assessed using accuracy, sensitivity, specificity, false discovery rate, false omission rate, and Matthews' correlation coefficient; the method produces better results in both the training and testing phases.

Additionally, by using the SVM and KNNs separately, the approaches achieved accuracy of 98.57% and 97.14% as well as specificity of 95.65% and 92.31% during testing.

However, using the University of California Irvine (UCI) Machine Learning Repository's Breast Cancer Data Set (BCD), the authors of (M.D. Bakthavachalam, 2020) [9] experimented with KNN and Naïve Bayes, the two most widely used Supervised Machine Learning models, to predict breast cancer. A comparative analysis between the two approaches was made in terms of its performance metrics using CV techniques. The KNN algorithm was used in the proposed study to get the best accuracy, 97.15%, while the NB classifier was used to reach the lowest error rate, 96.19%.

Authors from (Shler Farhad Khorshid, 2021) [10] used a different classification technique, among which accuracy, K-NN gave the most accurate prediction (99.12%) whereas authors from (Shagun Chawlaa, 2018) [11] have found a significant improvement in the accuracy when different combination of normalizations techniques and distance were tested out against the dataset. The best accuracy achieved was 98.24% when used decimal scaling normalization technique was used along with when the Manhattan distance at K = 14 was taken.

MATERIALS AND METHODS

Dataset

The dataset used in this study was obtained from the Wisconsin Breast Cancer Diagnostic dataset (WBCD) from the Kaggle data repository. The data was collected from 569 patients (with 357 instances labeled as benign and 212 labeled as malignant) and for each patient. There are 12 attributes which are ID number, diagnosis (M = malignant, B = benign), and 10 real-valued features were computed from digitized images of their FNA.

We employed various features such as texture, radius, area smoothness (variation in local radius lengths), concave points (count of concave parts in the contour), symmetry, fractal dimension, compactness (perimeter^2 area - 1.0), and concavity (degree of concave portions in the contour) (Kaggle, 2016) [5]. The dataset was divided into training and testing sets, with an 85/15 split.

For the classification of benign and malignant diagnoses, we trained and tested ANN, and K-nearest neighbor (KNN) algorithms. The algorithms were developed using Python programming language along with related libraries like Scikit-Learn and TensorFlow. Performance evaluation of the algorithms was conducted utilizing metrics including accuracy, precision, recall, and F1-score.

Algorithms

K-nearest Neighbors

K-nearest neighbors (KNN) is a machine learning technique that is used for classification and regression. The algorithm identifies the k-nearest data points in the training set for each test data point and determines the label of the test data point by selecting the majority class among its k-nearest neighbors. We must first establish a distance metric to assess the degree of similarity between data points before we can apply the KNN algorithm. The Euclidean distance, which determines the separation between two points in a d-dimensional space, is the most widely used distance metric (Istanbul Commerce University, 2018). Selecting a particular number of data points that are most similar to the newly categorized data point is the next stage in the KNN algorithm.

This is typically achieved by choosing an odd number of nearest neighbors, especially when there are only two classes. The classification of the new data point is determined based on the selected data points, with the category being defined as the one with the highest count among the nearest neighbors.

Artificial Neural Network

An artificial neural network is a computational model designed to imitate the functionality of the human brain. Numerous neurons in the brain exchange signals with one another chemically and electrically.

The synapses facilitate the transfer of signals between neurons. Analogously, an ANN is a method of information processing that processes data in a manner like to that of the human brain. (Wadkar, Kalyanai 2019) [12]. In an ANN, a large number of units are interconnected to process information and provide accurate results. In addition to classification, ANNs can also be used for the regression of continuous target attributes. Since ANNs have a lot of processing power, they will probably control the market in the future.

Implementation Procedure

The dataset is collected from Wisconsin and implementation is done using Python. In this methodology, we applied supervised algorithms and feature selection techniques like KNN and CNN. The study aims to develop a machine learning model for detecting breast cancer in individuals, and the implementation process is outlined in Figure 1. The process involves several stages, starting with data acquisition and preprocessing, followed by attribute selection. The model is then trained using the training data, and the results obtained are represented according to the study's objectives. Overall, the study utilizes a systematic approach to develop a predictive model for identifying breast cancer in individuals.

RESULT AND DISCUSSION

This study has used different machine learning algorithms on Breast Cancer in Wisconsin

Diagnostic dataset and later evaluated their performance using various performance metrics such as accuracy, confusion matrix, precision, F1 score, and sensitivity. The confusion matrix serves as a tool for evaluating the performance of classification tasks by comparing the predicted outcomes with the actual ones. Accuracy, the prevailing performance metric, quantifies the ratio of accurate predictions to the total number of predictions made. Precision is used to evaluate the number of correct documents and positives, respectively, returned by the Machine Learning (ML) model. The F1 score is utilized to compute the harmonic mean of precision and sensitivity.

Using the Breast Cancer Wisconsin Diagnostic Dataset (WBCD), we employed a K-nearest neighbor and artificial neural network in this study. We assessed their performance using a variety of performance measures, including confusion matrix, accuracy, precision, and F1 score. The actual and predicted outcomes of the classification task are compared using the confusion matrix to see how well it performed [13].

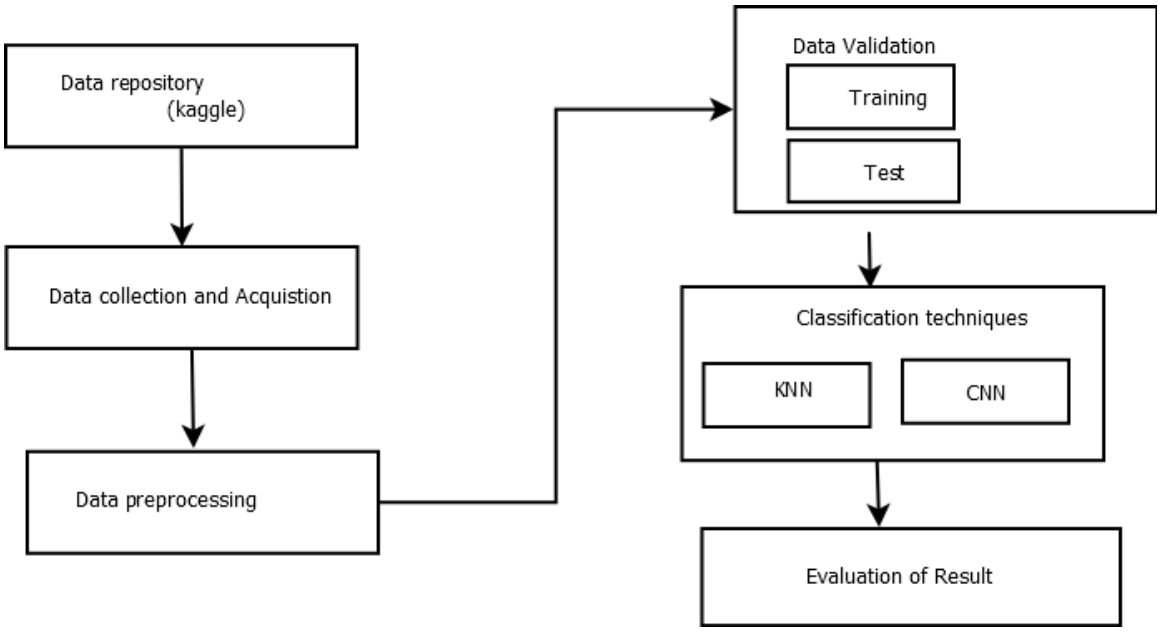

Figure 1. Proposed system.

Table 1. Comparison of performance metrics.

Algorithm	Precision	Recall	F1-score	Support	class
Artificial neural network	0.98	0.98	0.98	54	malignant
	0.97	0.97	0.97	32	benign
K-nearest neighbor	0.97	0.97	0.97	98	malignant
	0.99	0.95	0.97	73	benign

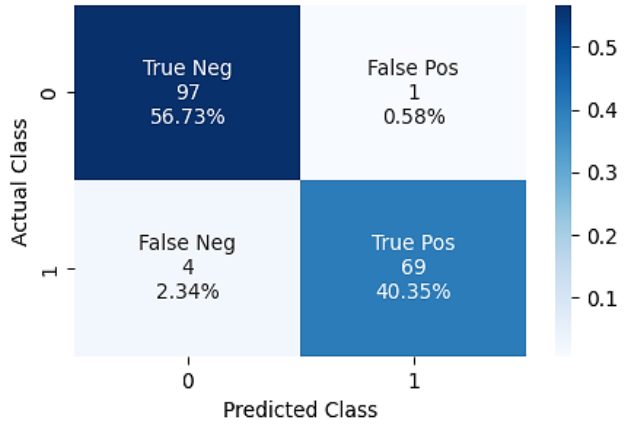

Figure 2. Confusion matrix for k-nearest neighbor.

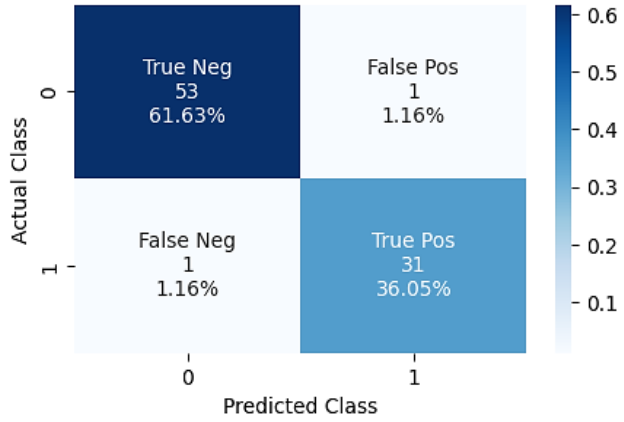

Figure 3. Confusion matrix for artificial neural network.

Figure 3, Confusion matrix shows that ANN predicts correctly 84 instances constituted of 31 malignant cases that are malignant and 53 benign cases that are benign, and 2 cases incorrectly predicted including 1 case of malignant class predicted as benign and 1 case of benign class predicted as malignant. Whereas from Figure 2, the confusion matrix of KNN shows 166 correctly classified instances out of which 69 malignant cases are malignant, 97 benign cases that are benign, and 2 cases are incorrectly predicted including 1 case of malignant class predicted as benign and 1 case of benign class predicted as malignant.

Table 1 shows the accuracy percentages for the Wincson Breast Cancer Diagnostic dataset. The results indicated that all classifiers have varying accuracies, but ANN has the highest accuracy on the testing set (98%).

CONCLUSION

In our study, we have tested Artificial Neural Networks and K-NN, on the WBCD using the Scikit-Learn library. We evaluated the results based on various metrics such as confusion matrix, accuracy, precision, and F1-score to identify the most reliable and precise algorithm with the highest accuracy. After thorough analysis, we found that Artificial Neural Networks have higher accuracy than K-NN, achieving an accuracy of 98.0%. Based on the WBCD dataset, we thus determined that ANN is the best machine learning method for breast cancer detection and prediction. However, it is important to note that our study's results are limited to the WBCD database, and future research should confirm the findings on other datasets. Moreover, we plan to apply our and other machine learning algorithms on larger datasets with more disease classes to achieve even higher accuracy in future works.

REFERENCES

1. Verma A, Kumar A, Kumar MS. Breast cancer prediction using support vector machine. Int Res J Eng Technol (IRJET). 2019.
2. Alanazi SA, Kamruzzaman MM, Islam Sarker MN, Alruwaili M, Alhwaiti Y, Alshammari N, Siddiqi MH. Boosting breast cancer detection using convolutional neural network. J Healthc Eng. 2021;2021:5528622. DOI: 10.1155/2021/5528622. PubMed: 33884157.
3. World Health Organization (WHO). Breast cancer. Who.int. 2024. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
4. Different kinds of breast lumps. Stony Brook Cancer Center. Stonybrookmedicine.edu. 2023. Available from: https://cancer.stonybrookmedicine.edu/breast-cancer-team/patients/bse/breastlumps
5. UCI machine learning. Breast cancer Wisconsin (diagnostic) data set. Kaggle.com. 2016. Available from: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.
6. Singhal V, Chaudhary Y, Verma SK, Agarwal U, Sharma MP. Breast cancer prediction using KNN, SVM, logistic regression and decision tree. Int J Res Appl Sci Eng Technol. 2022;10:1877-81. DOI: 10.22214/ijraset.2022.42688.
7. Tahmooresi M, Afshar A, Rad BB, Nowshath KB, Bamiah MA. Early detection of breast cancer using machine learning techniques. J Telecommun Electron Comput Eng (JTEC). 2018 Sep 26;10(3-2):21-7.
8. Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-nearest neighbors. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). 2017. p. 226-9. DOI: 10.1109/R10-HTC.2017.8288944.
9. Bakthavachalam MD, Raj DS. A study of breast cancer analysis using K-nearest neighbor with different distance measures and classification rules using machine learning. Eur J Mol Clin Med. 2020;7:4842-51.
10. Khorshid SF, Mohammed A. Breast cancer diagnosis based on K-nearest neighbors: A review. PalArch's J Archaeol Egypt/Egyptol. 2021;18:1927-51.
11. Chawla S, Kumar R. Breast cancer detection using K-nearest neighbour algorithm. In: Proceedings of the International Conference on Computational Intelligence and Internet of Things. New Delhi: G.B. Engineering College; 2018. p. 799-805.
12. Wadkar KP. Breast cancer detection using Ann network and performance analysis. Int J Comput Eng Technol (IJCET). 2019;10:75-86.
13. Eyupoglu C. Breast cancer classification using k-nearest neighbors algorithm. Online J Sci Technol. 2018;8:29-34.

