# Prediction of breast cancer using neural network analysis is effective

# Prediction of breast cancer using neural network analysis is effective.

**Inas Saleh Almazari[1], Kawther Faisal Amawi[2], Mohammad Abu Assab[1], Maisa MA AL-QUDAH[3], Ala Abu Helo[4], Ahed J Alkhatib[5,6]***

[1]Department of Clinical Pharmacy, Zarqa University, Zarqa, Jordan

[2]Medical Laboratory Sciences, Zarqa University, Zarqa, Jordan

[3]Department of Medical Laboratory Sciences, AL-Balqa Applied University, As-Salt, Jordan

[4]Department of Medical Laboratory Sciences, Zarqa University, Zarqa, Jordan

[5]Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science and Technology, Irbid, Jordan

[6]Department of Medicine and Critical Care, Department of Philosophy, Academician secretary of Department of Sociology, International Mariinskaya Academy, Zarqa, Jordan

## Abstract

**Breast cancer is one of the most frequent cancers among females internationally. The main objectives of the present study were to predict the breast cancer using neural network analysis. A dataset posed on Kaggle was analyzed to predict breast cancer by neural network analysis. A model of the study consisting of three layers was constructed. Three layers included input layer of 5 covariates: perimeter mean, radius mean, smoothness mean, texture mean, and area mean. The second layer was the hidden layer, and the third layer was the output layer. The correct prediction percent of the model was 95%. The results of the model showed that the strongest predictors were arranged in the following order: perimeter mean, area mean, smoothness mean, texture mean, and radius mean. Taken together, breast cancer can be effectively diagnosed using neural network analysis that produced a model with 95% prediction correction.**

**Keywords:** Breast cancer, Malignant, Benign, Neural network analysis, Smoothness, Perimeter, Kaggle.

*Accepted on December 15, 2021*

## Introduction

### *Overview of breast cancer*

According to global statistics, Breast Cancer (BC) is one of the most frequent malignancies among women globally, accounting for most new cancer cases and cancer-related deaths, making it a major public health issue today [1].

The breast cells grow into a tumor, which can only be detected by feeling lumps in the breast or *via* an X-ray. Breast cancer is a disease that affects both men and women. Many males are being diagnosed with breast cancer all around the world. According to the American Cancer Society, 279,100 new instances of breast cancer will be diagnosed in men and women in 2020, with 42,690 fatalities, whereas the number of men will be 2,620 and deaths will be 520, and the number of women will be 276,480 and deaths would be 42,170. From 2014 to 2016, the American Cancer Society published the incidence and mortality rates of breast cancer in women of various ages.

### *Genetic involvement of breast cancer*

Tumorigenesis in humans is a multistep process involving genetic changes in the germline DNA that predispose to cancer, as well as somatic DNA changes that initiate and promote disease progression [2]. The "cancer phenotype" is determined by how the cancer genome manifests. A cancer phenotype is a way of classifying a group of qualities that result from a genotype's interaction with its environment [3]. Cancer is more than just a tumor cell; it's a local ecosystem involving a variety of physiologic changes that collectively determine malignant growth: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis [4]. All these cancer hallmarks may now be investigated and evaluated in depth in order to develop a specific cancer signature that can be used not only for diagnostic but also therapeutic purposes [5]. Breast cancer is an example of how genomic and genetic discoveries have significantly altered the diagnosis, prognosis, and treatment of this illness, which is one of the most common in the world [6]. Many factors, including age (>50), early menarche age, late pregnancy, late menopause, obesity and nutritional issues, and high-dose radiation, have been proven to predispose women to the development of breast cancer. In fact, every phenotypic manifestation, including cancer, is based on the interaction between inherited genes and exogenous stimuli (gene-environment interaction). Environmental exposure to toxicants–many of which can be found in daily products and byproducts–has been linked to an elevated risk of developing breast cancer in epidemiological studies [7]. The discovery of selected mutations in important genes such as breast cancer 1, BRCA1 (OMIM *113705) and breast cancer 2, BRCA2 (OMIM *600185) allowed researchers to better understand the molecular basis of breast cancer and establish screening regimens for high-risk women [6]. Several studies have found that women with BRCA1 or BRCA2 mutations have an increased risk of developing breast cancer after being exposed to medical radiation, such as mammography or radiation therapy [8]. According to other studies, a combination of several alleles in genes related with DNA repair pathways increased the mammography-associated risk of developing breast cancer [9]. Furthermore, epigenetic mechanisms such as altered DNA methylation, histone modifications, and differential miRNA expression have been shown to have significant effects on the structure and function of the developing mammary gland when combined with chemical and radiation exposures [10].

### *Early detection of breast cancer*

Early detection of BC improves the prognosis and chances of survival by allowing patients to receive timely clinical treatment. Patients may avoid unneeded therapies if benign tumors are classified more precisely [9]. As a result, accurate BC diagnosis and classification of individuals into malignant or benign groups is a hot topic of research. Machine Learning (ML) is widely regarded as the approach of choice in BC pattern classification and forecast modeling due to its unique benefits in detecting essential characteristics from complex BC datasets. The use of classification and data mining technologies to classify data is quite effective. Particularly in the medical industry, where such procedures are commonly employed in diagnosis and decision-making.

Voulgaries et al., focused on a classification problem and proposed a modified closest neighbor algorithm that took advantage of dataset structural information, resulting in improved classifier performance over the classic KNN [11]. The k-nearest neighbor and the Nave Bayes binary classification algorithms were compared [12], with the k-nearest neighbor having a higher accuracy and a lower error rate in the classification of breast cancer. The classification datasets were taken from the UCI repository.

Rodriguez et al. created a K-Nearest Neighbor method that aids in the detection of breast cancer without the need for a surgical biopsy [13]. The KNN method was applied with a variety of parameters and distance measures to produce accurate results, including 81.67% accuracy, 80.05% precision, 83.72% recall, 79.61% specificity, and 81.80% F1-score. On the WEKA platform, Kumar et al. [14] presented a comparative analysis and implementation of machine learning classification methods such as KNN, Nave Bayesian network, support vector machine, decision tree, and Nave Bayes to gain accuracy. They compared the results of all of the methods in this research. The Nave Bayesian Network provided the best accuracy on datasets with fewer features, whereas SVM produced excellent accuracy on datasets with more features, according to the results. With a novel idea called incremental radius approach, Aher et al. [15] provided an enhanced model of the KNN algorithm to categorize the two groups. Rather of using the Euclidian distance between the two reference points and their nearest points, this method chooses a certain radius and assesses the decision. Prasad et al. [16] presented a categorization of NNS techniques by examining, processing data structures with various methodologies and algorithms. This research compared the categories of additive, reductional, weighted, continuous, reverse, and principal axis, as well as the complexity of various data structures used in various NNS algorithms.

### *Study objectives*

The main objective of this study was to predict the breast cancer using neural network analysis.

## Methodology

This study is based on analyzing dataset posted on Kaggle [17]. There are 6 parameters in the dataset. All these criteria can be used to classify cancer; if they have substantially high values, it could indicate malignant tissue. The first parameter is ID, which is a number used to identify the user [18]. The second parameter is membrane diagnostic, of which there are two types of tissue diagnoses: malignant and benign. In cases where both membranes have distinct therapies, it is vital to determine the right diagnosis of tissue for different cancer kinds. Estimated means, standard errors, and radius mean represent a range between the center and a point on the perimeter after these two. The estimated standard error is shown by the radius se. For the estimated range, the radius worst has the highest value of the center. Because surgery is dependent on the size, it is critical to know the distance between the center and the tip. With large tumors, surgery is not an option. The standard deviation of the gray-scale values is represented by the texture mean. The standard error of the estimated standard deviation for gray-scale data is represented by texture se. Texture worst is the greatest mean value of standard deviation for gray-scale values. The standard deviation is used to identify the variation in the data and to explain how to spread out the numbers. Gray-scale is widely used to find the tumor site, and the standard deviation is needed to find the variation in the data and to explain how to spread out the numbers. The perimeter mean represents the core tumor's mean value, whereas the perimeter se represents the core tumor's standard error of the mean. On the perimeter worst column, the maximum value of the core tumor is written. As previously indicated, the area mean, area se, and area worst point are all at similar values related to the mean of the cancer cell areas. Smoothness mean is the average of regional differences in radius range, smoothness se is the standard error of the mean of local radius length variations, and smoothness worst is the greatest mean value [19,20].

## Results

### Network information

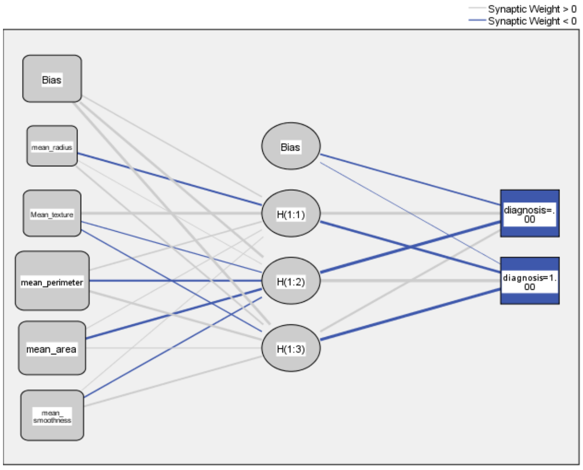
As seen in Table 2, network information included three layers, input layer, hidden layer, and output layer. Input layer included the covariates, 5 variables: mean radius, mean texture, mean perimeter, mean area, and mean smoothness. There 5 units, with standardized rescaling method for covariates. For the hidden layers, there was one hidden layer, with three units. The activation function was hyperbolic tangent. The out layer included one dependent variable, the diagnosis with two units: benign and malignant. Softmax was the activation function, and cross-entropy was the error function.

| | | | |
|---|---|---|---|
| Input layer | Covariates | 1 | Mean radius |
| | | 2 | Mean texture |
| | | 3 | Mean perimeter |
| | | 4 | Mean area |
| | | 5 | Mean smoothness |
| | Number of Unitsa | | 5 |
| | Rescaling Method for Covariates | | Standardized |
| Hidden Layer(s) | Number of Hidden Layers | | 1 |
| | Number of Units in Hidden Layer 1a | 3 | |
| | Activation Function | Hyperbolic tangent | |
| Output Layer | Dependent Variables | 1 | Diagnosis |
| | Number of Units | | 2 |
| | Activation Function | | Softmax |
| | Error Function | | Cross-entropy |

**Table 2.** Network information; a: Excluding the bias unit.

### Study model construction

As seen in Figure 1, the construction of study model showed how the covariates in the input layers were interacting with hidden layers to end with the final diagnosis. Two lines, blue and grey, were used to show the synaptic weight of interactions. Each line has different thicknesses to show the level of interactions.



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Softmax

**Figure 1.** Study model construction.

### Model summary

As indicated in Table 3, the model is summarized for training and testing parts. For training part, cross-entropy error was 59.08% incorrect prediction was 6.4%, stopping rule used was 1 step without decreased error, and training time was 0:00:00.14 second. For testing part, cross-entropy error was 21.735% incorrect predictions were 5.0%.

| | | |
|---|---|---|
| Training | Cross entropy error | 59.080 |
| | Percent incorrect predictions | 6.4% |
| | Stopping rule used | 1 consecutive step(s) with no decrease in errora |
| | Training time | 0:00:00.14 |
| Testing | Cross entropy error | 21.735 |
| | Percent incorrect predictions | 5.0% |
| Dependent variable: diagnosis | | |

**Table 3.** Model summary; a: error computations are based on the testing sample.

### *Classification of the diagnosis*

As seen in Table 4 and Figure 2, in training part, malignant cases included 148 cases among which 15 cases were classified as benign, and the percent correction was 89.9%. There were 261 benign cases, among which there were 11 cases classified as malignant, and percent correction was 95.8%. The overall percent correction in training part was 93.6% in testing part, malignant cases included 64 cases, among which 7 cases were classified as benign, and the percent correct was 89.1%. benign cases included 96 cases, among which 1 case was classified as malignant, and the percent correction was 99%. The overall percent in testing part was 95%.

| **Sample** | **Observed** | **Predicted** | | |
|---|---|---|---|---|
| | | **Malignant** | **Benign** | **Percent correct** |
| Training | Malignant | 133 | 15 | 89.9% |
| | Benign | 11 | 250 | 95.8% |
| | Overall percent | 35.2% | 64.8% | 93.6% |
| Testing | Malignant | 57 | 7 | 89.1% |
| | Benign | 1 | 95 | 99.0% |
| | Overall percent | 36.3% | 63.8% | 95.0% |
| Dependent Variable: diagnosis | | | | |

**Table 4.** Classification of the diagnosis.

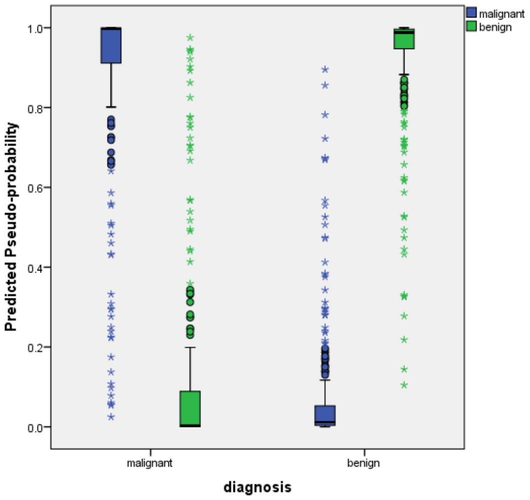


**Figure 2.** Prediction pseudo-probability of the diagnosis.

### *The importance of independent variables*

As illustrated in Table 5 and Figure 3, the following covariates as predictors of breast cancer came in the following order: mean perimeter (0.32, 100%), mean area (0.249, 77.8%), mean smoothness (0.205, 63.9%), mean texture (0.151, 47.2%), and mean radius (0.074, 23.2%).

| **Covariate** | **Importance** | **Normalized importance** |
|---|---|---|
| Mean radius | .074 | 23.2% |

| | | |
|---|---|---|
| Mean texture | .151 | 47.2% |
| Mean perimeter | .320 | 100.0% |
| Mean area | .249 | 77.8% |
| Mean smoothness | .205 | 63.9% |

**Table 5.** The importance of independent variables.

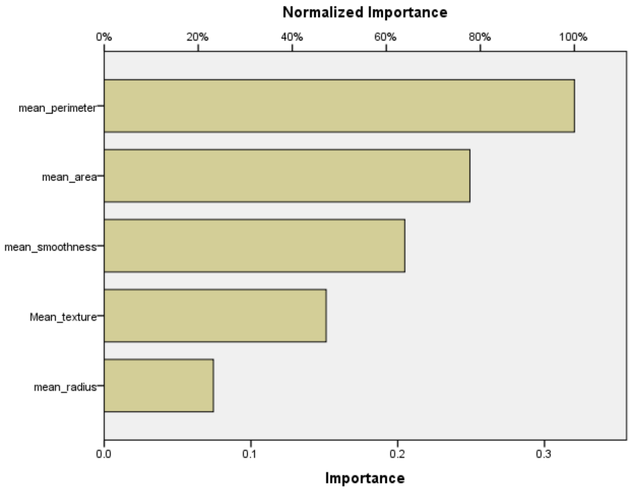


**Figure 3.** The relative importance of covariates.

## Discussion

The present study was conducted to construct a prediction model that can show how some covariates related to breast cancer to be useful. The overall prediction model correction percent was 95%.

According to this study, mean perimeter was the strongest predictor of breast cancer. The results confirmed other studies in which models were constructed to predict breast cancer.

Mean area of the tumor came in the second rank as a predictor of breast cancer. As tumor increases, it implies malignant diagnosis more than benign diagnosis. In their study, Sharma et al reported the importance of tumor size in the progression of breast cancer [21]. The choice of therapeutic treatment of breast cancer also concerns with tumor size [22].

Mean smoothness ranked the third predicting factor of breast cancer. MRI can be used to evaluate the tissue of breast cancer [23]. Recently, Tazeoğlu et al. reported smooth breast thickness as one of breast lesion variables that can be investigated by MRI [24].

Mean texture was the 4th predicting factor of breast cancer. Texture of the breast implies heterogeneity of the breast [25]. Mammographic descriptions of breast density texture hold promise as an independent risk factor for breast cancer risk and as a means of distinguishing between cancer subtype risks.

Mean radius was the 5th predicting factor of breast cancer. Radius of breast tissue was considered in several studies as a predicting factor of breast cancer [26,27].

## Conclusion

The present study showed that breast cancer can be effectively diagnosed using neural network analysis that produced a model with 95% prediction correction.

## References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71(3): 209-249.
2. Low SK, Zembutsu H, Nakamura Y. Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer Sci. 2018; 109: 497–506.
3. Camp JG, Platt R, Treutlein B. Mapping human cell phenotypes to genotypes with single-cell genomics. Science. 2019; 365: 1401–1405.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011; 144: 646–674.
5. Campbell PJ, Getz G, Korbel JO, et al. Pan-cancer analysis of whole genomes. Nature. 2020; 578: 82–93.
6. Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clin Oncol. 2018; 15: 353–365.
7. Kortenkamp A. Breast cancer and environmental risk factors: An appraisal of the scientific evidence. Breast Cancer Res. 2008; 10: 45.
8. https://towardsdatascience.com/breast-cancer-cell-type-classifier-ace4e82f9a79
9. Gray JM, Rasanayagam S, Engel C, et al. State of the evidence 2017: an update on the connection between breast cancer and the environment. Environ Health. 2017; 16(1): 94.
10. Luzhna L, Kutanzi K, Kovalchuk O. Gene expression and epigenetic profiles of mammary gland tissue: insight into the differential predisposition of four rat strains to mammary gland cancer. Mutat Res Genet Toxicol Environ Mutagen. 2015; 779: 39–56.
11. Voulgaris Z, Magoulas GD. Extensions of the k nearest neighbour methods for classification problems. In Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications, AIA. 2008; 8: 23-28.
12. Amandeep K, Prabhjeet K. Breast cancer detection and classification using analysis and gene-back proportional neural network algorithm. Int J Innov Technol Explor Eng. 2019; 8(8).
13. Rodriguez V, Sharma K, Walker D. Breast cancer prediction with K-nearest neighbor algorithm using

different distance measurements. Software Engineering Project. 2018.

14. Ajay K, Sushil R, Tiwari AK. Comparative study of classification techniques for breast cancer diagnosis. Int J Comput Sci Eng. 2019; 7(1): 234-240.
15. Aher P, Raut P. Incremental radius approach for classification. In 2017 International Conference on Advances in Computing, Communication and Control (ICAC3). 2017;1-6.
16. Rajendra PM, Partha SC. Comparative analysis of nearest neighbor query processing techniques international conference on recent trends in computing 2015 (ICRTC-2015). Procedia Comput Sci. 2015; 57: 1289–1298.
17. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.
18. Wolberg H. Wisconsin breast cancer database. University of Wisconsin Hospitals; Madison, WI, USA. 1991.
19. Alickovic E, Subasi A. Breast cancer diagnosis using GA feature selection and rotation forest. Neural Comput Appl. 2017; 28: 753–763.
20. Ak MF. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. Healthcare. 2020; 8(2): 111.
21. Sharma GN, Dave R, Sanadya J, et al. (2010). Various types and management of breast cancer: An overview. J Adv Pharm Technol Res. 2010; 1(2): 109–126.
22. Rath GK. Radiation therapy in the management of cancer. 50 years of cancer control in India. 2010.
23. Osborne MC, Boolbal SK. Diseases of the breast (4th edn). Philadelphia: Lippincott Williams & Wilkins. 2009; 1–11.
24. Tazeoğlu D, Dağ A, Arslan B, et al. Breast hamartoma: Clinical, radiological, and histopathological evaluation. Eur J Breast Health. 2021; 17(4): 328-332.
25. Malkov S, Shepherd JA, Scott CG, et al. Mammographic texture and risk of breast cancer by tumor type and estrogen receptor status. Breast Cancer Res . 2016; 18: 122.
26. Vishali S, Anita B. An analysis on prediction of breast cancer using radius nearest neighbor algorithm over other classification algorithms. Materials Today: Proceedings. 2021.
27. de Gonzalez AB, Berg CD, Visvanathan K, et al. Estimated risk of radiation-induced breast cancer from mammographic screening for young BRCA mutation carriers. J Natl Cancer Inst. 2009; 101: 205–209.

## *Corresponding to:

Ahed J Khatib

Department of Legal Medicine

Toxicology and Forensic Medicine

Jordan University of Science and Technology,

Zarqa, Jordan

E-mail: ajalkhatib@just.edu.jo