ECRONICON OPEN ACCESS

EC GYNAECOLOGY
Review Article

The Application of Artificial Intelligence for Determining the Predictors of Breast Cancer

Saif A Jabali[1]*, Abeer S Alshamali[2] and Raghda A Jabali[2]

[1]*Department of Preventive Medicine, JRMS, Jordan*

[2]*Ministry of Health, Jordan*

***Corresponding Author:** Saif A Jabali, Department of Preventive Medicine, JRMS, Jordan.

Received: August 09, 2023; **Published:** August 25, 2023

Abstract

The most prevalent kind of cancer in women is found in the breasts, and it is called breast cancer. The method of research begins with a step that is necessary: the identification of breast cancer risk factors. The key approach for attaining the basic aims of this study, which were to determine the predictors of breast cancer risk variables and the relative value of each predictor, was to make use of neural network analysis. These were the primary goals of this study. The present investigation was based on an analysis of neural networks that was carried out on data that was presented on Kaggle [1]. The breast cancer risk variables that can be utilized to predict the disease are the primary focus of this collection. The only dependent variable was the result, which was either no disease (1) or disease (2). A total of eight independent factors were included for this study. In total, there were 116 different cases contained within the dataset. While the category of disease comprised 37 cases, the category of no disease included 79 cases, which accounted for 68.1% of the total, and the category of disease contained 37 cases, which accounted for 31.9% of the total. The architectural model was built using a number of characteristics, such as the training component, which had the following values: the gross entropy error was 23.7884, and the proportion of erroneous predictions was 10.1%. One or more consecutive steps that showed no improvement in errora was used as the halting rule for the experiment. Over the entire testing session, the total gross entropy error came to 14,327, and 13.5% of the predictions that were made turned out to be incorrect. In descending order of importance, the factors that were found to have the greatest influence on the risk of developing breast cancer were: resistin (21.5%), insulin (18.2%), HOMA (17.5%), BMI (16.2%), glucose (14.7%), leptin (4.6%), adiponectin (4.1%), and age (3.30%). Age was found to have the least amount of influence on the risk of developing breast cancer. In conclusion, artificial intelligence is a valuable tool that can be used to determine the factors that will have an impact on the future.

Keywords: *Artificial Intelligence; Breast Cancer; Resistin; Insulin; HOMA; BMI; Glucose; Leptin; Adiponectin*

Introduction

It would appear that breast cancer is the most common form of the disease that affects women all over the world, placing just behind lung cancer as the most common form of cancer in industrialized countries [2,3].

In 154 different countries, breast cancer is the most common type of the disease identified in females, and it is the main cause of death from cancer in 103 of those countries. Over 2.1 million women were newly diagnosed with breast cancer in 2018, accounting for 24.2% of the overall number of breast cancer cases. In addition, the death rate associated with breast cancer was almost 15% [4,5].

The survival rate of patients may depend on the time of detection of breast cancer, as an example, in Malaysia is one of the lowest compared to other countries in the region [6]. This is mostly due to the fact that between 50 and 60 percent of breast cancer cases in Malaysia are detected at advanced stages. Because of this, it is absolutely necessary for research to be carried out in order to determine the various factors that play a part in deciding the percentage of breast cancer patients who are able to survive their disease [3].

The amount of time that an individual lives after receiving a diagnosis of an illness is what is indicated by the phrase "survival." Both the standardization of reporting and the judgment of whether or not anything can survive are dependent on reaching the five-year milestone. Because it takes at least 5 years to mark a patient record as having survived or not survived, several earlier studies used a 5-year threshold to evaluate the survivability of the cohort [7]. Despite the fact that overall survival rates for the disease have been gradually improving over the course of the past few years, the 5-year survival rate for breast cancer differs substantially from individual to individual [8]. This is due to the fact that breast cancer is such a complicated disease. Accurately predicting breast cancer survival could help medical professionals make more informed decisions regarding the planning of medical treatment interventions, the prevention of excessive treatment, which would lead to a reduction in economic costs [9] the more effective inclusion and exclusion of patients from randomized trials [10] and the development of palliative care and hospice care systems [11]. Because of this, one of the key goals of the most recent breast cancer research is an effort to improve the ability to predict how long patients will live after being diagnosed with the disease [12].

Simple computer programs such as Microsoft Excel, SPSS, and STATA have been used in the past by medical professionals to do analyses of the factors that influence the breast cancer survival rate [13]. These conventional statistical methods are not fully adaptable in the sense that they cannot discover new variables or generate creative and integrative visualizations [14]. Because of the limitations of these traditional statistical investigations, numerous machine learning (ML) approaches have become increasingly popular in this industry [15]. These methods of machine learning (ML) include decision tree (DT), random forest (RF), neural networks, extreme boost, logistic regression, and support vector machine (SVM), just to name a few instances of each. A decision tree is a form of supervised learning strategy that organizes the results of an investigation into a tree-like structure that can be comprehended with relative ease [16]. When conducting data analysis on massive amounts of information, visualization is an essential factor that should be taken into consideration. Random forest (Breiman's algorithm), which is a derivative of DT, is able to work in both supervised and unsupervised modes, handle continuous and categorical data, and perform classification or regression tasks [17]. Random forest is a form of Breiman's algorithm. Because it possesses this ability, it is able to perform in the same manner as DT. Because neural networks accomplish modelling by first learning from data with a known outcome and then optimizing their weights in order to generate more accurate predictions in situations where the outcome is unknown [18], it is popular to think of neural networks as black boxes. This is because neural networks learn from data with a known outcome and then optimize their weights. This reputation has been earned by the highly sophisticated systems known as neural networks. The classification and regression trees that make up extreme boost are arranged in an ensemble. It is straightforward to use, it can be run in parallel, it can attain an effective level of prediction accuracy, and it has outscored other algorithms in a number of competitions including machine learning [19]. There has been some discussion on whether or not "extreme" boost should be used. Logistic regression utilizes the Gaussian distribution and is capable of handling all types of variables, including continuous, discrete, and dichotomous data; as a result, it does not require an assumption of normality [20]. Logistic regression also employs data that can be either dichotomous or continuous. The Gaussian distribution is also utilized in the process of logistic regression. A form of machine learning known as support vector machines, or SVMs, are implemented for the purpose of supervised classification. This method of machine learning first identifies the optimal decision boundary that separates data points into distinct groups, and then, using this boundary as a guide, it forecasts the type of new observations that will be made [21].

According to the findings of Ganggayah., *et al*. [3] who used artificial intelligence to predict risk factors for the prediction of breast cancer, the significant variables that emerged included the classification of the stage of cancer, the size of the tumour, the number of total axillary lymph nodes removed, the number of positive lymph nodes, the forms of initial treatment, and the techniques of diagnosis.

In another study, Almazari., *et al*. [22] conducted a study to predict breast risk factors using artificial intelligence. Breast cancer is one of the most prevalent forms of cancer that strikes women all over the world is breast cancer, which accounts for its high incidence

rate. The utilization of neural network analysis was one of the key components of this particular research endeavour, the primary focus of which was the forecasting of breast cancer. In order to make predictions regarding breast cancer, a dataset that was made available on Kaggle was analyzed with the help of neural network technology. It was planned to construct a model of the research that would have three separate layers. The input layer was made up of three different layers and five different variables. These included the mean of the area, the mean of the perimeter, the mean of the smoothness, and the mean of the texture. On the second layer was where you'd find the concealed layer, and on the third layer would be where you'd find the output layer. When it came to producing accurate predictions, the model had an accuracy rate of 95%. According to the results of the model, the predictors that were the most accurate were sorted as follows, beginning with those that were considered to be the most significant: radius mean, smoothness mean, area mean, perimeter mean, and smoothness mean. When everything is taken into account, neural network analysis, which resulted in the creation of a model with a prediction accuracy of 95%, can be used effectively to identify breast cancer. The model was created as a result of the study.

Study Objectives

The main objectives of the present study were to determine the predictors of breast cancer risk variables and the relative value of each predictor.

Methodology

The foundation of this work was an examination of a dataset that was made available for public consumption on Kaggle [1]. There are a total of six distinct parameters included in the dataset. If any of these features have considerably high values, there is a possibility that cancerous tissue is present; on the other hand, none of these qualities are required for the classification of cancer. The first parameter is a number that is referred to as the ID, and it is employed in the process of identifying the user [23]. The second sort of tissue diagnosis is called the membrane diagnostic, and there are two categories of tissue diagnoses: malignant and benign. The membrane diagnostic is the type of tissue diagnosis that is most commonly used. In situations in which the two membranes require different treatments, it is vital to acquire an exact identification of the tissue in order to treat the right form of cancer. In these situations, it is also essential to treat the appropriate form of cancer, it is essential to get an accurate identification of the tissue. Estimated means, standard errors, and the mean of the radius all reflect a range that extends from the centre out to a point on the perimeter that follows after these two. This range can be thought of as beginning at the centre and ending at the mean of the radius. The calculated standard error can be shown as an illustration in the form of the radius se. The possibility that the radius will have a value that is equal to or greater than the centre constitutes the range's worst-case scenario. It is crucial to have an accurate measurement of the distance between the centre and the tip due to the fact that the size has an effect on the manner in which surgery is performed. When dealing with massive tumours, surgery is not a treatment option that can be considered. The standard deviation of the grayscale values can be thought of as being equivalent to the mean of the texture. The estimated standard deviation of the grayscale data is referred to as the "texture se" in this article. The word "texture se" refers to the standard error associated with this estimation. The worst possible texture has a mean value that is far larger than the standard deviation for the grayscale values. This means that the texture is really poor. The standard deviation is a statistical metric that is applied to data in order to determine the amount of variation that is present and to provide an explanation for how the numbers should be distributed. This is done by comparing the actual distribution of the numbers to how they should be distributed. Grayscale is frequently used in tumour localization, and the standard deviation is vital for evaluating the degree of variance in the data and offering an explanation of how to adequately space out the values. Grayscale is also frequently utilized in the process of determining whether or not a cancer has spread. The perimeter mean is the mean value of the core tumour, and the perimeter se represents the standard error of the mean. The perimeter mean and the perimeter se together make up the perimeter. The term "perimeter" refers to the area bounded by the two values taken together. On the graph, the "perimeter worst" column contains the greatest value of the core tumour, which can be found in the "maximum value" column.

As was said before, the area mean, the area standard error, and the area worst point are all at values that are comparable to one another in reference to the mean of the cancer cell areas. This is because all three of these values are centred around the mean of the area occupied by the cancer cells. The smoothness mean is the average of regional variances in radius range, the smoothness se is the standard error of the mean of local variations in radius length, and the smoothness worst is the mean value that is higher than the other two [24,25].

Results

As shown in table 1, the summary of artificial model included two parts, training and testing. Training part percentage was 72.4% and testing part percentage was 27.6%.

		N	Percent
Sample	Training	84	72.4%
	Testing	32	27.6%
Valid		116	100.0%
Excluded		0	
Total		116	

Table 1: *Case processing summary.*

As seen in table 2, network information included three layers: input layer, hidden layer and output layer. Input layer is composed of covariates such as age, BMI, glucose, and resistin. Hidden layer is composed of one hidden layer, with four units in each layer. The activation function is Hyperbolic tangent. The output layer involved the dependent variable, classification, with Softmax as the activation function.

Input Layer	Covariates	1	Age
		2	BMI
		3	Glucose
		4	Insulin
		5	HOMA
		6	Leptin
		7	Adiponectin
		8	Resistin
	Number of Units[a]		8
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		4
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Classification
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy
a. Excluding the bias unit			

Table 2: *Network information.*

As shown in figure 1, a schematic representation showing how the three layers are interacting with each other to predict the importance of risk factors. The input layer covariates interacted with the hidden layer involving different lines of intensity, thick and thin, and different lines, blue and grey. The intensity of color and thickness of lines indicated the impact of covariate on the prediction of cancer. From the hidden layer to the output layer, lines took the last arrangement.

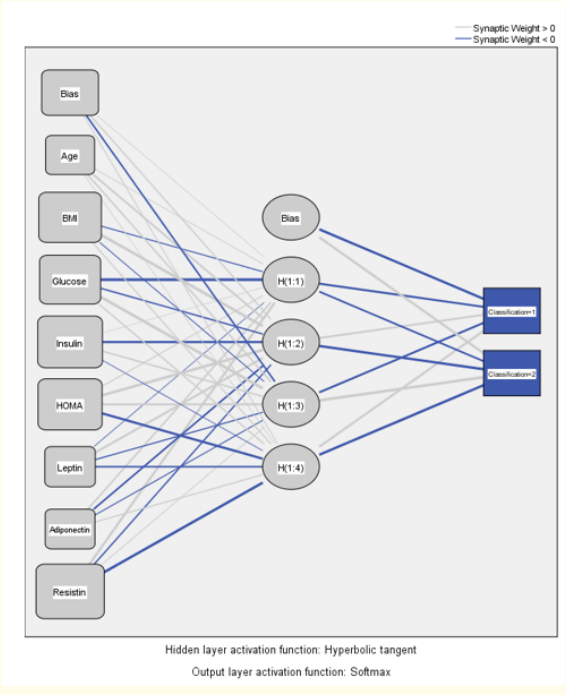

Figure 1: *A schematic representation of the model.*

As seen in table 3 and figure 2, the importance of independent variables is shown. The normalized importance of age was 15.3%, and ranked the 8th, this was followed by adiponectin 18.9% and ranked the 7th, leptin was 21.6% and ranked the 6th, glucose level was 68.5% and ranked the 5th, BMI was 75.3% and ranked the 4th, HOMA was 81.5% and ranked the 3rd, insulin was 84.7% and ranked the 2nd, and resistin was 100% and ranked the 1st.

Discussion

The results of the present study showed an analysis of an artificial model, focusing on its training and testing parts, network architecture, and the importance of independent variables in predicting the importance of risk factors for cancer.

	Importance	Normalized Importance
Age	.033	15.3%
BMI	.162	75.3%
Glucose	.147	68.5%
Insulin	.182	84.7%
HOMA	.175	81.5%
Leptin	.046	21.6%
Adiponectin	.041	18.9%
Resistin	.215	100.0%

Table 3: *Independent variable importance.*

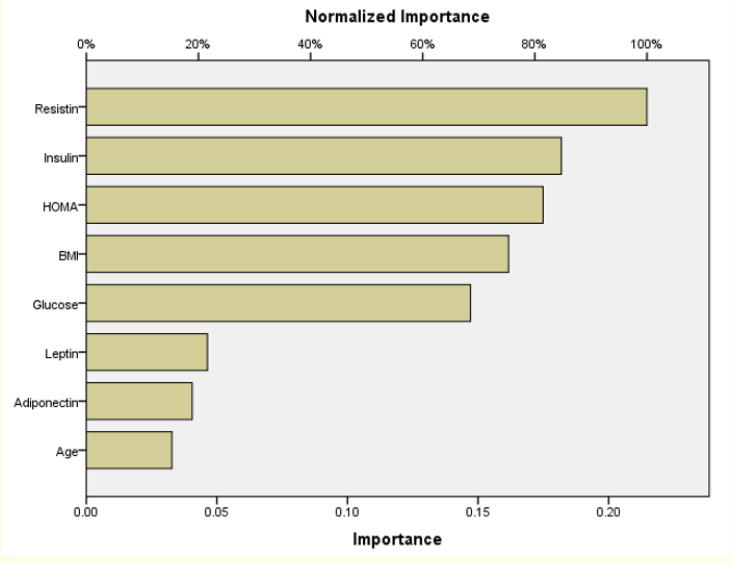

Figure 2: *Normalized importance of independent variables.*

Training and testing parts: Table 1 shows the breakdown of the dataset into training and testing parts. The training part consists of 84 samples, which accounts for 72.4% of the total dataset, while the testing part comprises 32 samples, representing 27.6% of the dataset. This division is essential in machine learning to train the model on a subset of data and evaluate its performance on unseen data. It is crucial to have a sufficient amount of data for training to ensure the model learns patterns and generalizes well to unseen instances. This approach has been described in other studies such as Almazari., *et al*. [22] and Pocrnic., *et al*. [26].

Network architecture: Table 2 provides information about the network architecture used in the artificial model. It consists of three layers: the input layer, hidden layer, and output layer. The input layer incorporates various covariates such as age, BMI, glucose, insulin, HOMA, leptin, adiponectin, and resistin. These covariates serve as inputs to the model and are standardized for rescaling. The hidden layer has one layer with four units, and the activation function used is Hyperbolic tangent. The output layer is responsible for predicting the dependent variable, which is classification in this case. It consists of two units and employs the Softmax activation function. The error function used for training is Cross-entropy, a common choice for classification tasks. In a previous study on a data set by Kaggle, Almazari., *et al*. [22] reported similar findings.

Schematic representation (Figure 1): Figure 1 provides a schematic representation of the model, illustrating how the three layers interact to predict the importance of risk factors for cancer. The input layer covariates are shown interacting with the hidden layer through lines of different intensities (thick and thin) and colors (blue and grey). The intensity and thickness of the lines indicate the impact of each covariate on the cancer prediction. Moving from the hidden layer to the output layer, the lines take on a different arrangement, representing the flow of information towards the final prediction. These schematic presentations are considered parts of artificial intelligence models [3,22,26].

Importance of independent variables: Table 3 and figure 2 present the importance of independent variables in predicting the importance of risk factors. The normalized importance values indicate the relative contribution of each variable. In this case, resistin has the

highest importance value of 100% and is ranked first. Insulin follows with an importance of 84.7% and is ranked second, while HOMA, BMI, and glucose have importance values of 81.5%, 75.3%, and 68.5% respectively, ranking them third, fourth, and fifth. Leptin, adiponectin, and age have lower importance values and ranks. These results suggest that resistin is considered the most important variable in predicting the importance of risk factors for cancer in the given model. It is worth mentioning that the arrangement of dependent variable importance may vary from one study to another and researchers have to input their maximal efforts to achieve best predicting findings [22,26-29].

Conclusion

Overall, the provided results offer insights into the artificial model's training and testing parts, network architecture, the interplay between layers, and the importance of independent variables. These findings contribute to understanding the model's performance and the relative significance of different variables in predicting the importance of risk factors for cancer.

Bibliography

1. https://www.kaggle.com/uciml/breast-cancer-wisconsindata
2. Ponnuraja CC., *et al*. "Decision Tree Classification and Model Evaluation for Breast Cancer Survivability: A Data Mining Approach". *Biomedical and Pharmacology Journal* 10 (2017): 281-289.
3. Ganggayah MD., *et al*. "Predicting factors for survival of breast cancer patients using machine learning techniques". *BMC Medical Informatics and Decision Making* 19.48 (2019).
4. Bray F., *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians (2018).
5. Li J., *et al*. "Predicting breast cancer 5-year survival using machine learning: A systematic review". *PLoS ONE* 16.4 (2021): e0250370.
6. Islam T., *et al*. "The Malaysian breast Cancer survivorship cohort (MyBCC): a study protocol". *BMJ Open The New England Journal of Medicine* 5 (2015): e008643.
7. Delen D., *et al*. "Predicting breast cancer survivability: a comparison of three data mining methods". *Artificial Intelligence in Medicine* 34.2 (2005): 113-127.
8. Polyak K. "Heterogeneity in breast cancer". *The Journal of Clinical Investigation* 121.10 (2011): 3786-3788.
9. Altman Douglas G. "Prognostic models: a methodological framework and review of models for breast cancer". *Cancer Investigation* 27.3 (2009): 235-243.
10. Altman DG and Royston P. "What do we mean by validating a prognostic model?" *Statistics in Medicine* 19.4 (2015): 453-473.
11. Stone P and Lund S. "Predicting prognosis in patients with advanced cancer". *Annals of Oncology Official Journal of the European Society for Medical Oncology* 18.6 (2007): 971.
12. Kourou K., *et al*. "Machine learning applications in cancer prognosis and prediction". *Computational and Structural Biotechnology Journal* 13 (2015): 8-17.
13. Bhoo-Pathy N., *et al*. "Trends in presentation, management and survival of patients with de novo metastatic breast cancer in a southeast Asian setting". *Scientific Reports* 5 (2015): 16252.

14. Pearce CB., *et al*. "Machine learning can improve prediction of severity in acute pancreatitis using admission values of APACHE II score and C-reactive protein". *Pancreatology* 6 (2006): 123-131.

15. Huber M and Kurz C. "Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning". *BMC Medical Informatics and Decision Making* 19.1 (2019): 3.

16. Chen W., *et al*. "A comparative study of logistic model tree, random forest, and classi fi cation and regression tree models for spatial prediction of landslide susceptibility". *Catena* 151 (2017): 147-160.

17. Genuer R., *et al*. "VSURF: an R package for variable selection using random forests". *RJ* 7.2 (2015): 19-33.

18. Amato F., *et al*. "Artificial neural networks in medical diagnosis". *Journal of Applied Biomedicine* 11.2 (2013): 47-58.

19. Pilaftsis A and Rubio J. "The Higgs Machine Learning Challenge". *Journal of Physics: Conference Series* 664.7 (2015): 072015.

20. Erener A., *et al*. "A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM)". *Engineering Geology Journal* 203 (2016): 45-55.

21. Sacchet MD., *et al*. "Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory". *Frontiers in Psychiatry* 6 (2015): 21.

22. Inas Saleh Almazari., *et al*. "Prediction of breast cancer using neural network analysis is effective". *Journal of RNA and Genomics* (2021).

23. Wolberg H. "Wisconsin breast cancer database". University of Wisconsin Hospitals; Madison, WI, USA (1991).

24. Alickovic E and Subasi A. "Breast cancer diagnosis using GA feature selection and rotation forest". *Neural Computing and Applications* 28 (2017): 753-763.

25. Ak MF. "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications". *Health Care* 8.2 (2020): 11.

26. Ivan Pocrnic., *et al*. "Herring and Gregor Gorjanc". *Journal: Genetics Selection Evolution* 54.1 (2022).

27. Malkov S., *et al*. "Mammographic texture and risk of breast cancer by tumor type and estrogen receptor status". *Breast Cancer Research* 18 (2016): 122.

28. Vishali S and Anita B. "An analysis on prediction of breast cancer using radius nearest neighbor algorithm over other classification algorithms". *Materials Today: Proceedings* (2021).

29. De Gonzalez AB., *et al*. "Estimated risk of radiation-induced breast cancer from mammographic screening for young BRCA mutation carriers". *Journal of the National Cancer Institute* 101 (2009): 205-209.

Volume 12 Issue 9 September 2023