SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH In cooperation with CIMNE

Prediction of distant recurrence in breast cancer using a deep neural network

Balqis Mohd Azman[12], Saiful Izzuan Hussain[12], Nor Aniza Azmi[3], Muhammad Zahin Athir Abd Ghani[3], Nor Irfan Danial Norlen[3]

1 Department of Mathematical Sciences, Faculty of Science and Technology,

2 Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

3 Diagnostic Imaging and Radiotherapy Program, School of Diagnostic and Applied Health Sciences, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia

Abstract

Breast cancer is the most common cancer diagnosed in women, and it is ranked as the second highest cancer with high mortality rate. Breast-cancer recurrence is the cancerous tumor that returned after treatment. Cancer treatments such as radiotherapy are performed mainly to kill cancer cells; however, some cells may have survived and multiply themselves at the same area as the original cancer (local recurrence) or to any other part (distant recurrence). Distant recurrence occurs when cancer cells spread to other parts of the body, most commonly to bone, breast, liver, and lungs. This study employed an Artificial Neural Network of the deep learning approach to predict distant recurrence of breast cancer. Factors that contribute to the risk of recurrence are age, type of surgery performed, tumor size, breast subtype, estrogen receptor, progesterone receptor, undergoing chemotherapy or not, and lymph node involvement. The actual value of distant recurrence is also considered to be a variable. Principal Component Analysis using five and three principal components was conducted. The outcome indicates that the model has accuracy of up to 0.80 using three principal components.

OPEN ACCESS

Published: 22/03/2022

Accepted: 09/03/2022

DOI: 10.23967/j.rimni.2022.03.006

1. Introduction

World Health Organization announced breast cancer as the most common form of cancer in 2020 and that it has overtaken lung cancer based on statistics by the International Agency for Research on Cancer. According to Global Cancer Observatory [1], breast cancer has surpassed lung cancer, with 11.7% new cases in 2020 worldwide for both sexes of all ages. Breast cancer represents one in four common cancers diagnosed among women globally and is the leading cause of cancer death. The American Cancer Society mentioned in [2] that approximately 1 in 8 women will be diagnosed with invasive breast cancer in their lifespan and that 1 in 39 women will die from breast cancer. The incidence and mortality rate in female breast cancer in South-Eastern Asia are 41.2% and 15.0%, respectively.

Breast cancer has been recorded as the highest occurrence of all cancers in women worldwide [3-5]. Feng et al. [6] define breast cancer as a compilation of distinct cancerous tumors called malignancies that present in the breast region. Breast cancer is an abnormal and rapid growth of the tissue cells in the breast region. Sainsbury et al. [7] stated that breast cancers are derived from the epithelial cells that line the terminal duct lobular unit. Sariego [8] specified that breast cancer mainly involves mammary glands, which are glands and ducts that deliver the milk. Breast cancer can be categorized into two groups: non-invasive and invasive cancer. A study by Becker [9] demonstrated that the most significant factors for breast cancer are age and low parity. The prevalence of breast cancer is staged according to the size of a tumor that requires different combination of treatments. Surgery, chemotherapy, radiotherapy, and hormone therapy are few common treatments that have been utilized in clinical practice for many years.

Radiotherapy is the treatment of choice and has been extensively used to treat different stages of breast cancer. It uses high-energy rays and specific particles to kill cancerous cells. There are two categories of radiotherapy: external beam radiation therapy (EBRT) and internal radiation (brachytherapy). EBRT directs high doses of radiation from outside the body at the cancerous tissues located inside the body. A special X-ray machine delivers radiation from multiple angles. The EBRT technique includes three-dimensional conformal radiation therapy (3D-CRT), intensity modulated radiation therapy (IMRT), and volumetric modulated arc therapy (VMAT). Abo-Madyan et al. [10] recommended that 3D-CRT is better in comparison to IMRT and VMAT because it decreases chances of relapse after completion of treatment. Radiotherapy is also an effective way of reducing the risk of having recurrent breast cancer

Correspondence: Saiful Izzuan Hussain (sih@ukm.edu.my).

and relieves metastatic symptoms. In a meta-analysis involving 10,801 patients from 17 randomized trials, Waks and Winer [11] proved that radiotherapy lessens the recurrence risk from 35% to 19.3%. Nonetheless, there are high possibilities of cancer's distant recurrence, even after completion of radiotherapy.

Recurrent breast cancer is a condition that the breast-cancer relapses after primary treatment. Typically, patients will be followed up for many years after undergoing a potentially curative treatment. Few studies have justified that recurrence is likely to occur between 1 and 2 years after surgery [12,13], which is relatively early, but a substantial proportion of cancer cells that appear cured before can relapse as local or distant tumor recurrences approximately 10 or 20 years after surgery. Nicolò et al. [14] stated that roughly around 20% to 30% will develop recurrence, although the patient has received primary treatment aimed at eliminating all cancer cells. These undetected cancer cells (micro-size metastasis) may have survived, becoming recurrent breast cancer. Moody et al. [15] highlighted that the principal cause of breast cancer-related death is recurrence and metastasis. To reduce the mortality rate, is by an accurate and early prediction on available features.

Many studies proposed the use of Artificial Intelligence (AI) in the field of medicine [16,17,18]. In fact, a data mining algorithm plays a crucial part for early-stage breast cancer. AI utilizes computer models and algorithms to build smart machines that perform human-like tasks and offer a vast potential to healthcare. Artificial neural networks (ANNs), fuzzy expert systems, Bayesian networks, and hybrid intelligent systems are techniques in artificial intelligent that have been utilized in different clinical settings. Amisha et al. [19] pointed out that the largest proportion of investments in AI research throughout 2016 were in healthcare applications compared to other sectors. Recently, different classifier algorithms are compared by using various a range of medical datasets to perform predictive analysis for more accurate diagnosis [20,21].

Classification is a crucial and essential task in machine learning and data mining. Rahman et al. [16] concluded from past research that there are numerous algorithms for classification and prediction, specifically in cancer care and performance evaluation of Support Vector Machine (SVM), NM, C4.5 and k-NN, which are among the most influential algorithm in the research community. However, Danaee et al. [22] discovered noisy high-dimensional data that manifest a challenge to conduct prediction. This leads to a critical need to improve accuracy and to identify attributes that play crucial roles in cancer. Owing to a massive volume of data, high amount of noise, and the complexity of biological networks, many researchers have proposed Principal Component Analysis (PCA) for dimensionality reduction [23].

In the meantime, this research sought to develop a prediction model using Deep Neural Network (DNN) and to evaluate the performance in terms of accuracy by comparing the number of attributes used in predicting the distant recurrence of breast cancer in patients after its treatment. DNN is a supervised learning algorithm, which is grouped under neural network. Deep learning classification algorithms can also additionally assist the oncologist to forecast the likelihood of the patient with breast cancer to relapse. Thus, early action can be taken to improve the quality of life of the patient and lower the rate of morbidity. There are a few factors that may cause cancer to relapse. According to Ditsatham et al. [24], the age, tumor size, and lymph node involvement were one are among the risk factors for recurrence. Other factors include type of surgery (mastectomy or lumpectomy), histological type, Hormone Receptor (HR), human epidermal growth factor receptor-2 (HER2), Estrogen Receptor (ER) and Progesterone Receptor (PR), and chemotherapy treatment. A study by Arvold et al. [25] highlighted the increasing risk of recurrence with the presence of gene expression biomarkers, which are HR, HER2, ER, and PR. Thus, this study helps to predict distant recurrences based on the important features of recurrence in breast cancer.

2. Literature review

Breast-cancer recurrence may occur in the same site as that of the original cancer (local recurrence), or it may spread to other parts of the body through distant recurrence [26]. Yates et al. [26] also found that distant recurrence for breast cancer commonly occurs in bone, brain, liver, lung, and distant lymph nodes. Patients with cancer have the risk of breast-cancer relapse even after treatment completion. Dowsett et al. [27] discovered that patients with positive ER have 50% probability of breast-cancer relapse post initial treatment. This can lower the quality of the patients' life and increase the need for further treatment and a longer period of treatment to improve the result and is associated with an increased rate of morbidity. Nowadays, the increment of cases of distant recurrence in breast cancer has become tremendous and robust. According to Coles et al. [28], the death that is related to this pathology is mostly related to metastasis and recurrence. To reduce the mortality rate, in this case, is by an accurate prediction on available features.

Radiologist could achieve a better result performance with the assistance of AI. Rodríguez-Ruiz et al. [29] conducted a study to evaluate breast-cancer detection supported by machine learning and analysis by radiologist. This study aimed to observe the Area Under the Receiver Operating Characteristic (AUROC), specificity, sensitivity, and reading time per second between the conditions. On average, AUROC with AI support is higher than unaided reading with 0.89 and 0.87, respectively, whereas sensitivity increased with AI support (86%). In terms of time, the reading time per case was similar between two conditions. The result proved that AI system could enhance analysis performance in diagnosing

breast cancer. Advancement in software processing and high imaging quality are effective in increasing accuracy of the treatment given, allowing treatment, specifically radiotherapy, to be delivered effectively, safely, and effectively.

A number of researchers have attempted to predict breast-cancer recurrence in some sort of pattern, and very few manage to get a high-accuracy model with a small error. Hussain et al. [30] emphasized that a suitable algorithm and architecture of a model are the key to a critical task in prediction. There may be the existence of redundant or highly correlated attributes in data with high dimension that can severely degrade classification accuracy. This poses a compulsive challenge to machine learning. Hence, Howley et al. [31] investigated the use of PCA to resolve the issue. Skittides and Früh [32] acknowledged PCA to reduce the original datasets by applying orthogonal transformation and aborted extra attributes.

PCA is a classical statistical method that transforms the original dataset into a new uncorrelated attribute dataset called principal component (PC). Classifiers and PCA are applied to a dataset to improve its accuracy. The performance of numbers of supervised classifiers using PCA has been monitored by Mushtaq and Yaqub [23]. Five different classifiers such as SVM, Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (Gaussian NB), and K-Nearest Neighbor (K-NN) were used. For each sample with different classifiers, PCA functionality for dimensionality reduction was applied in the preprocessing, and a confusion matrix was used for performance evaluation. The results indicate performance variations according to each classifier in which Naïve Bayes with sigmoid PCA has the highest accuracy of 99.2%, whereas SVM with a sigmoid function had the worst performance with 83.5% accuracy. With this breast-cancer dataset, KNN performs best among other five supervised machine learning algorithms. A similar study by Jhajharia et al. [33], a prognostic study, detected breast cancer using PCA that had been coupled with a classifier, ANN. The result indicated that applying both PCA and a classifier can lead to an increase in performance. Machine learning is rapidly advancing through deep learning and continues to boost the interest in applying these techniques to improve the accuracy of cancer diagnosis.

A DNN has ANN characteristics, with a more complex network structure that has several hidden layers aiming to set new records in accuracy, specifically for image recognition and sound recognition [34]. These hidden layers are used for learning different attribute maps for a diagnosis. Liang et al. [35] explained that medical data are more complex nowadays, and traditional models, for instance, LR and random forest (RF), are thus not efficient in forecasting medical diagnosis. DNN is a powerful pixel classifier and is listed as supervised learning [36]. Many discussions and experiments have been made regarding this issue and the results were observed. Deep learning has been proved to bring a revolution in machine learning, showing improved performance over traditional approaches.

Gerazov and Conceicao [37] classified tumor type for patients with breast cancer using the advanced deep learning methods. Preprocessing starts with feature extraction that was done using PCA for the objective of dimensionality reduction that transformed the feature vector to a 2D one. The system was assessed using DNN and CNN. DNN considered the number of neurons per layer, and CNN used a convolutional layer instead. The number of neurons per layer for each DNN model differing in the range of 10 to 1,000 neurons, whereas the number of convolutional layers of CNN was between 2 and 4, and the number of filters per layer was from 20 to 100. The networks were trained using Stochastic Gradient Descent using a constant learning rate, a linearly changing learning rate or the Adaptive Moments (Adam) algorithms. By using the PCA-DNN model, the experiment has achieved the best result of 92.81% classification accuracy.

Several authors have been trying to ensemble machine learning (ML) systems intended to yield a better performance. Mambou et al. [38] developed a DNN and SVM as an intelligent textile to perform the classification of the data. In this context, a well-trained DNN was combined with an SVM model as a classifier. A set of thermal imaging analyzing a breast is used as the data. This process will take the thermal images of the breast in, and the output will then categorize the images as a cancerous or healthy breast. DNN was modified at the last fully connected layer in such a way as to obtain a powerful binary classification (cancerous breast or healthy breast), whereas SVM was coupled with DNN in the case of uncertainty in the output of DNN.

Sun et al. [39] on breast-cancer risk analysis conducted a preliminary study in which two experiments were compared for DNN. Eight hundred forty samples were used and grouped as training and testing sets randomly in the first experiment, whereas 84,000 samples were extracted according to a region of interests (ROIs) in the second. The method of 10-fold cross validation was used for all sample experiments, and the risk score for each case cancer was predicted based on the percentage of the ROIs, with its score surpassing the threshold. Based on different thresholds, the ROC curve was plotted to find the best threshold that maximizes the area of containing rectangular under the curve. Under this threshold, the case-based accuracy is 0.6972, and ROI-based accuracy is 0.6707. A related study by Badré et al. [40] focusing on polygenic risk for breast cancer aimed to compare a deep learning model of DNN and a ML alternative. LR, DT, RF, AdaBoost, gradient boosting, SVM, and Gaussian naive Bayes were implemented and tested for comparison. With a GWAS dataset, DNN was then found to surpass performance of alternative ML techniques with AUROC curve (AUC) of 67.4% for DNN, whereas BLUP rank was second with 64.2%, BayesA had the rank third (64.5%), and 62.4% was for LDpred. Kim et al. [41] compared ANN, SVM, and Cox regression classification models in predicting

recurrent cancer cases for patients with breast cancer. The research produced results in which SVM outperformed other models in terms of accuracy (84.6%), whereas ANN has higher accuracy (81.84%) compared to the Cox regression (72.6%) method.

The study of Jhajharia et al. [33] used an ANN design for breast-cancer prognosis models. Datasets are sorted into a data matrix consisting of 462 observations and 9 attributes after PCA is carried out. The study was conducted by comparing PCA-ANN, SVM, Naive Bayes, IBK, DT, and OneR models. Average classification accuracy is 90% with the highest model accuracy, PCA-ANN (98.39%), and models with the lowest accuracy of OneR (92.7%). So, when the PCA is carried out before building a model, the accuracy of the forecast will be increased. Almost all of the ANNs used nowadays are fully connected. Belciug and El-Darzi [42] recommended use of a partial connected artificial neuronal network (PCNN) model to detect breast cancer and recurrent breast cancer. Based on analysis carried out against two sets of Wisconsin Prognostic Breast-Cancer data, accuracy is measured based on statistical analysis, a confusion matrix, and length of time the analysis is conducted. PCNN performs similarly to MLP neuronal network models. With the use of such comparative tests, there are no statistically significant differences with p-value < 0.005.

Overall, the prediction of breast-cancer recurrence is one of the interesting problems to be explored. There has been little quantitative analysis in this field using the deep neural network method together with PCA. The evidence for this relationship is also inconclusive and much uncertainty still exists. This paper intends to examine the accuracy of the deep neural network together with PCA. This paper provides an important opportunity to advance the understanding of the breast-cancer recurrence model from deep learning perspective.

3. Materials and methods

3.1 Data

Data of patients with breast cancer who were treated with radiotherapy were accessed from the Surveillance, Epidemiology, and End Results (SEER) database [43]. Features that contributed to the risk of distant recurrence are age at diagnosis in 2012 (≤35, 36 to 64, ≥65 years old), tumor size (<2 cm or ≥2 cm), lymph node involvement (positive or negative), type of surgery (no surgery, breast-conserving surgery, or mastectomy), ER (positive or negative), PR (positive or negative), chemotherapy (yes or no), and breast subtype (HR+/HER2+, HR+/HER2−, HR−/HER2+, or HR−/HER2−). Distant recurrence (yes/no) was also one of the attributes and the output of this research.

3.2 Data preprocessing

A total of 22,402 patients with breast cancer have been identified. All of the patients with breast cancer data and the nine attributes were inserted into an Excel spreadsheet, and the file format was then converted from an XML Spreadsheet into a comma-separated value file. Since almost half of the data were patients without distant recurrence that led to an unbalanced dataset. Thus, random undersampling was used to get a balanced dataset. The data that were obtained had attribute noise such as missing data and unknown data. The missing data were handled using the filtration and case deletion approach. Totals of patients with breast cancer after deletion were 677 with 9 attributes. Data were in a specific code (SEER code base) according to attributes that were classified according to the SEER code. All values that were in text form were changed into numeric values, whereas attributes that were in float type were transformed into integers for easy analysis.

3.3 Model development

Ways to achieve the objectives of the present study can be reflected through Figure 1. The method of determining the number of significant variables and building a more detailed and clear DNN model will be explained further on.

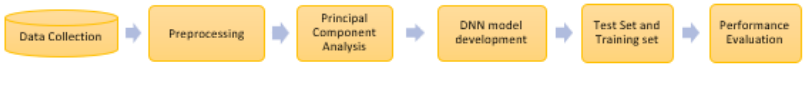

Figure 1. Model development process

PCA has proven its potential use as a variable selection method for many studies to identify cancer; several analyses were reported in previous research [44]. The goal of this objective is to determine the number of variables in building DNN models by running PCA. This subtopic will deepen the PCA process to improve the performance of ML methods in the high-dimensional data classification.

Lever et al. [45] study suggested that PCA focuses on reducing data methods by projecting data that have a large dimension to a smaller dimension called the main component (PC). PC 1 is chosen to minimize the distance between data and projection. By minimizing the distance, it can maximize the projected variance. The highest variable variants between the variables that have been projected represent the contribution of each key component. PC 2 is selected using the same way, provided that the component is interdependable. In other words, the PC is mutually orthogonal. Lubis et al. [46] explains the PCA's calculation is based on an eigen value calculation that represents the dissemination of data. By using a PCA, n variables from the data will be selected and projected into a new variable called PC. The use of the main k-components will generate the same value as when n variables is used.

DNN model design construction is carried out in Google Colaboratory and trained using the Tensorflow V2.2.0 package. In this study, DNN had several layers comprising layers of input, two hidden layers, and an output layer. The main component known as neurons can be connected in a variety of ways and a different number of hidden layers to form the design of different neuronal networks. Ritthipravat [47] explained that the design of the often used neuronal network is the Multi-Layer Perceptron (MLP). The statement was stronger based on previous studies other than [48]. Examples of a DNN framework with two hidden layers for recurrent cancer predictions are presented in Figure 2.

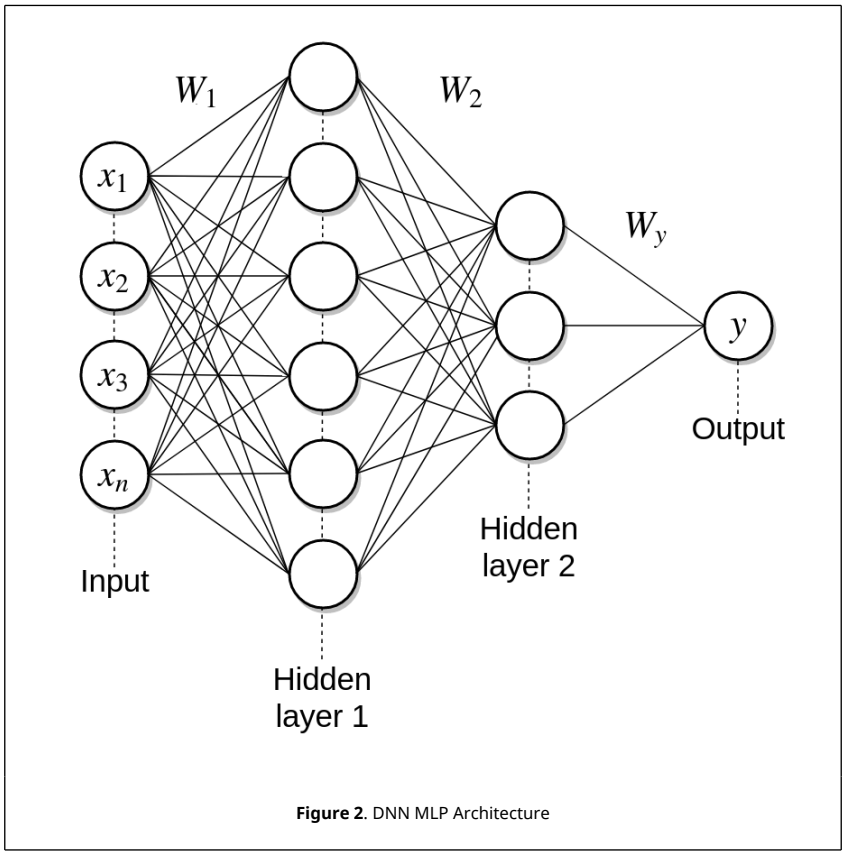

Figure 2. DNN MLP Architecture

Jerez et al. [49] explained that the neural network consists of layers of neurons that are connected consecutively. The first layer of the network is called the input layer receiving signals from data entering the network. Each neuron is connected to different neurons in different layers by signal landing pathways. Signals are sent via this route to other neurons. Each neuron summarizes the signals and alters the resulting signals as neuronal output using the activation function. The study used rectified linear units (ReLU) as an activation function on the first and second hidden layers, whereas the output function was a sigmoid function. The last layer is called the output layer. The output signal is then sent to another neuron in the next layer. The DNN model was then trained using mini-batches with a batch size of 32. The Adam optimizer (Ribeiro et al. 2016), an adaptive learning rate optimization algorithm, was used to update the weights in each mini-batch. The results of the forecast for this research are binary, 0 (negative distant recurrence) and 1 (positive distant recurrence).

The weightage of each variable is taken at random, i.e., the number approaching zero (non-zero). With one variable representing an input node, neurons are activated forward-propagation, and predictions are produced. DNN is trained to use Batch Gradient Descent by changing the weights to minimize the cost function. This function is an error between a real value and a predictive value. This method can help the analysis process for data with high dimensions.

When all training sets go through DNN, they are called epochs. The study set 100 epochs to observe the performance of trained DNN models.

Data samples are randomly divided into two independent sets, which are sets of training (70%) and testing (30%). The training set was used to generate the forecast model, and the remaining 30% of the data (test set) were used to estimate the accuracy of the model. Accuracy is the number of actual ratios to the number of predictions made. It is calculated based on a confusion matrix value of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). A confusion matrix is depicted in Table 1.

Table 1. Confusion matrix

		Actual Value	
		Positive (1)	Negative (0)
Predicted Value	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

4. Results and discussion

4.1 Demographic data

Subsequently, analysis and comparison of model accuracy are based on the number of key components selected. There were a total of 677 breast-cancer patient data after the abortion process was carried out with nine attributes. The age of patients with the highest breast cancer is in the range of 35 to 64 years (68.83%) and 239 (51.29%) of them are positive cases of repeated breast cancer. Patients with breast cancer undergoing lumpectomy surgery were more (54.95%) than those undergoing mastectomy surgery (37.52%). The percentage difference between patients undergoing chemotherapy and not undergoing such treatment was 19.06%. One hundred ninety-six (57.82%) patients with recurrent cancer cases have a sub-brand of HER2−/HR+ breasts. The percentage of patients with positive PR data was 67.80%, whereas the percentage of those with a positive ER was 80.06%. The size of the tumor in excess of 2 cm is higher (62.48%) compared to a size of less than 2 cm (37.52%). 86.43% of patients who were positive cases of recurrent cancer were positive lymph node engagement. The negative cases of recurrent breast cancer were fewer (49.93%) than the positive cases (50.07%). Table 2 presents the descriptive analysis based on each attribute.

Table 2. Demographic characteristics

Attributes	Total (n = 677)	Negative distant recurrence (n = 338)	Positive distant recurrence (n = 339)	p-value
Age	7.92 ± 2.63	8.10 ± 2.57	7.73 ± 2.68	0.287
<35	25 (3.69%)	11 (3.25%)	14 (4.13%)	
35 until 64	466 (68.83%)	227 (67.16%)	239 (70.50%)	
≥65	186 (27.47%)	100 (29.56%)	86 (25.37)	
Type of Surgery				0.045
No surgery	51 (7.53%)	0	51 (15.04%)	
Lympectomy	372 (54.95%)	278 (82.25%)	94 (27.73%)	
Mastectomy	254 (37.52%)	60 (17.75%)	194 (57.23%)	
Chemotherapy				0.037
No	274 (40.47%)	183 (54.14%)	91 (26.84%)	
Yes	403 (59.53%)	155 (45.86%)	248 (73.16%)	
Breast Subtype				0.063
HER2+/HR+	108 (15.95%)	36 (10.65%)	72 (21.24%)	
HER2−/HR+	440 (64.99%)	244 (72.19%)	196 (57.82%)	
HER2+/HR−	46 (6.79%)	18 (5.33%)	28 (8.26%)	
HER2−/HR−	83 (12.26%)	40 (11.83%)	43 (12.68%)	
PR				0.035
Negative	218 (32.2%)	96 (28.4%)	122 (35.99%)	
Positive	459 (67.80%)	242 (71.6%)	217 (64.01%)	
ER				0.030
Negative	135 (19.94%)	58 (17.16%)	77 (22.71%)	
Positive	542 (80.06%)	280 (82.84%)	262 (77.29%)	
Tumor Size				0.037
<2 cm	254 (37.52%)	204 (60.36%)	50 (14.75%)	
≥2 cm	423 (62.48%)	134 (39.64%)	289 (85.25%)	
Lymph Node Involvement (LN)				0.037
Negative	282 (41.65%)	236 (69.82%)	46 (13.57%)	
Positive	395 (58.35%)	102 (30.18%)	293 (86.43%)	
Distant Recurrence				
Negative	338 (49.93%)			
Positive	339 (50.07%)			

4.2 PCA Number of variables

This section explains the results of the analysis obtained based on the number of different variables, i.e., eight, five, and three. PCA is carried out using Jupyter Notebook software version 6.1.4, whereas the DNN model development process is carried out using Google Colaboratory software with available programming language on Python version 3.7.0.

The data are standardized to be normally distributed with a mean = 0 and variance = 1. This can facilitate the analysis process of PCAs as this process ensures that the PCAs are interdependently linear and orthogonal to one other. Maier and Dandy [50] insisted that normalization and standardization are normally used to ensure that each variable receives a similar assessment. The mean of each column in the data matrix is calculated and deducted from each column. Therefore, the new matrix has data with a zero mean.

The next process is the calculation of the correlation value between variables based on the covariance matrix. This value measures the strength of the relationship between variables in the data matrix. This calculation is carried out using standardized data. Based on Table 3, the attribute's covariance mapping to itself is the variance value of the data. The size of the covariance matrix is based on the number of variables for which eight variables produce an 8×8 matrix.

Table 3 . Covariance matrix

	Age	Surgery	Chemotherapy	Breast Subtype	ER	PR	Tumor Size	LN
Age	1.001							
Surgery	−0.155	1.001						
Chemotherapy	−0.355	0.302	1.001					
Breast Subtype	0.072	0.079	0.073	1.001				
ER	0.025	−0.096	−0.284	−0.832	1.001			
PR	0.036	−0.088	−0.272	−0.568	0.678	1.001		
Tumor Size	−0.095	0.311	0.381	0.055	−0.143	−0.162	1.001	
LN	−0.157	0.283	0.403	−0.042	−0.077	−0.076	0.472	1.001

Eigen decomposition is the process of obtaining eigen values and eigen vectors. The symmetrical covariance matrix simplifies the process of calculating the value of eigen values and eigen vectors [51]. Eigen vectors are sorted in decreasing value based on their respective eigen values. The eigen values and symmetrical eigen vectors are orthogonal to one other. Table 4 shows eigen vectors according to PC.

Table 4. Eigen vectors

2.711	**1.946**	**0.989**	**0.133**	**0.771**	**0.420**	**0.523**	**0.520**

Referring to Table 4, we can conclude that the first column is the most significant component with a value of 2.711 and is called the first principal component (PC1).

Observations based on Table 5 and 6 indicate that the percentage of PC1 variance met 33.83% of the total covariance, whereas PC2 represented 24.29% until the percentage of variance required for the PC3 variance was 12.34%. Therefore, PC1, PC2, and PC3 concluded the maximum share of contributions from the data. In this context, researchers can ignore other PCs; however, this study takes into account predictive analysis with different numbers of PCs to observe the difference in precision rates of DNN models.

Table 5. Variance

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
33.833	24.285	12.341	9.621	6.527	6.492	5.240	1.662

Table 6. Cumulative variance

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
33.833	58.118	70.458	80.079	86.606	93.098	98.338	100

4.3 Performance evaluation

The DNN architechture uses a forward feed neuronal network with several parameters. The number of neurons used in both layers of hidden neurons is six units obtained by using activation function ReLU. Meanwhile, the number of neurons on the output layer is one unit with sigmoid as a function. This parameter can be easily referred to in Table 7.

Table 7. DNN parameter

Number of Layers	3
Hidden layer neuron	12

Output layer	1
Activation function for hidden layers	ReLU
Activation function the output layer	Sigmoid

This study focuses on DNN models for predicting positive breast cancer patients a few months after the main treatment. Through this study, the highest accuracy rate of 0.80 for the DNN model can be achieved after the PCA is carried out to determine the use of variable numbers. Table 8 presents the accuracy rate performance of the DNN model according to the number of variables, i.e., eight, five, and three variables. Based on the table, the value varies between the precision rate for the use of eight and three variables and is 0.08, whereas the precision rate difference of five and three variables is 0.01. The use of three variables representing 70.46% of contributions increased the accuracy of the model from 0.72 to 0.80.

Table 8. DNN performance value

Number of PC	Cumulative Variance (%)	Accuracy
8	100.00	0.72
5	86.61	0.79
3	70.46	0.80

5. Conclusions

In conclusion, the study used in-depth learning methods and classification model designs to develop recurrent breast-cancer prediction models using retrospective patient with breast-cancer data in 2012 from an open virtual database, SEER. Deep learning approach has great potential to produce better results than traditional ways of predicting distant recurrence. Several preprocessing processes were carried out before the model was developed to improve forecast performance. The PCA was found to be a significant contribution of each variable in explaining the variance in input data. The predicted output is in the form of binary numbers: 0 (negative distant recurrence) and 1 (positive distant recurrence). The accuracy rates of DNN models used eight, five, and three, and the variables were 0.72, 0.79, and 0.80, respectively.

In terms of research limitations, this study focuses only on the accuracy rate of DNN models based on the number of variables without regard of the rate of sensitivity, specification, and consistency. This study also focused only on one model without comparing precision rates, sensitivity, specification, and consistency of several deep learning classification models. There are several aspects of this research that can be further improved or extended in future research, such as the use of more complex models, such as RNN and CNN, involving a wider range of data available in the medical field such as pictures. Comparison of neuronal network deep learning models can be carried out to see the difference in accuracy between models to determine the best classification models. In conclusion, future improvement on the quality of the study will widen the usage of ML in clinical and hospital settings.

Acknowledgements

The research was funded by the Ministry of Higher Education(MOHE) trough Fundamental Research Grants Scheme under the grant number FRGS/1/2020/STG06/UKM/03/1.

References

[1] Global Cancer Observatory, 2020. https://gco.iarc.fr/

[2] American Cancer Society. Breast cancer facts & figures 2019-2020. Am. Cancer Soc, pp. 1-44, 2019.

[3] Chen W., Zheng R., Baade P.D., Zhang S., Zeng H. et al. Cancer statistics in China, 2015. CA: A Cancer Journal for Clinicians, 66(2):115–132, 2016. https://doi.org/10.3322/caac.21338

[4] Siegel R.L., Miller K.D., Jemal A. Cancer statistics, 2018. CA: A Cancer Journal for Clinicians, 68(1):7–30, 2018. https://doi.org/10.3322/caac.21442

[5] Torre L.A., Bray F., Siegel R.L., Ferlay J., Lortet-Tieulent J. et al. Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians, 65(2):87–108, 2015. https://doi.org/10.3322/caac.21262

[6] Feng Y., Spezia M., Huang S., Yuan C., Zeng Z. et al. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes and Diseases, 5(2):77–106, 2018. https://doi.org/10.1016/j.gendis.2018.05.001

[7] Sainsbury J.R.C., Anderson T.J., Morgan D.A. ABC of breast diseases: breast cancer. BMJ, 321(7263):745-750, 2000.

[8] Sariego J. Breast cancer in the young patient. American Surgeon, 76(12):1397–1400, 2010. https://doi.org/10.1177/000313481007601226

[9] Becker S. A historic and scientific review of breast cancer: The next global healthcare challenge. International Journal of Gynecology & Obstetrics, 131:S36-S39, 2015.

[10] Abo-Madyan Y., Aziz M.H., Aly M.M.O.M., Schneider F., Sperk E. et al. Second cancer risk after 3D-CRT, IMRT and VMAT for breast cancer. Radiotherapy and Oncology, 110(3):471–476, 2014. https://doi.org/10.1016/j.radonc.2013.12.002

[11] Waks A.G., Winer E.P., Breast cancer treatment: A review. JAMA - Journal of the American Medical Association, 321(3):288–300, 2019, https://doi.org/10.1001/jama.2018.19323

[12] Fischer U., Zachariae O., Baum F., Von Heyden D., Funke M. et al. The influence of preoperative MRI of the breasts on recurrence rate in patients with breast cancer.

European Radiology, 14(10):1725-1731, 2004.

[13] Weiss R.B., Woolf S.H., Demakos E., Holland J.F., Berry D.A., et al. Natural history of more than 20 years of node-positive primary breast carcinoma treated with cyclophosphamide, methotrexate, and fluorouracil–based adjuvant chemotherapy: A study by the cancer and leukemia group B. Journal of Clinical Oncology, 21(9):1825-1835, 2003.

[14] Nicolò C., Périer C., Prague M., MacGrogan G., Saut O., et al. Machine learning versus mechanistic modeling for prediction of metastatic relapse in breast cancer. BioRxiv, 259–274, 2019. https://doi.org/10.1101/634428

[15] Moody S.E., Perez D., Pan T.C., Sarkisian C.J., Portocarrero C.P., et al. The transcriptional repressor Snail promotes mammary tumor recurrence. Cancer Cell, 8(3):197–209, 2005. https://doi.org/10.1016/j.ccr.2005.07.009

[16] Rahman M.A., Chandren Muniyandi R., Albashish D., Rahman M.M., Usman O.L. Artificial neural network with Taguchi method for robust classification model to improve classification accuracy of breast cancer. PeerJ Computer Science, 7:e344, 2021.

[17] Mohammed S.A., Darrab S., Noaman S.A., Saake G. Analysis of breast cancer detection using different machine learning techniques. In Communications in Computer and Information Science: 1234 CCIS, Springer Singapore, 2020. https://doi.org/10.1007/978-981-15-7205-0_10

[18] Murtaza G., Shuib L., Abdul Wahab A.W., Mujtaba G., Nweke H.F., et al. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. Artificial Intelligence Review, 53(3):1655-1720, 2020.

[19] Amisha, Malik P., Pathania M., Rathaur V.K. Overview of artificial intelligence in medicine. Journal of Family Medicine and Primary Care. 8(7):2328–2331, 2019. https://doi.org/10.4103/jfmpc.jfmpc_440_19

[20] Amrane M., Oukid S., Gagaoua I., Ensarİ T. Breast cancer classification using machine learning. Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 1-4, 2018. doi: 10.1109/EBBT.2018.8391453.

[21] Ganggayah M.D., Taib N.A., Har Y.C., Lio P., Dhillon S.K. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Medical Informatics and Decision Making, 19(1):1–17, 2019. https://doi.org/10.1186/s12911-019-0801-4

[22] Danaee P., Ghaeini R., Hendrix D.A. A deep learning approach for cancer detection and relevant gene identification. Pacific Symposium on Biocomputing, 0(212679):219–229, 2017. https://doi.org/10.1142/9789813207813_0022

[23] Mushtaq Z., Yaqub A., Hassan A., Su S.F. Performance analysis of supervised classifiers using PCA based techniques on breast cancer. International Conference on Engineering and Emerging Technologies, ICEET 2019, 1–6, 2019. https://doi.org/10.1109/CEET1.2019.8711868

[24] Ditsatham C., Somwangprasert A., Watcharachan K., Wongmaneerung P., Khorana J. Factors affecting local recurrence and distant metastases of invasive breast cancer after breast-conserving surgery in Chiang Mai University Hospital. Breast Cancer (Dove Medical Press), 8:47–52, 2016. https://doi.org/10.2147/BCTT.S99184

[25] Arvold N.D., Taghian A.G., Niemierko A., Abi Raad R.F., Sreedhara M., et al. Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 29(29):3885–3891, 2011. https://doi.org/10.1200/JCO.2011.36.1105

[26] Yates L.R., Knappskog S., Wedge D., Farmery J.H., Gonzalez S., et al. Genomic evolution of breast cancer metastasis and relapse. Cancer Cell, 32(2):169-184, 2017.

[27] Dowsett M., Sestak I., Regan M.M., Dodson A., Viale G., et al. Integration of clinical variables for the prediction of late distant recurrence in patients with estrogen receptor-positive breast cancer treated with 5 years of endocrine therapy: CTS5. Journal of Clinical Oncology, 36(19):1941–1948, 2018. https://doi.org/10.1200/JCO.2017.76.4258

[28] Coles C.E., Moody A.M., Wilson C.B., Burnet N.G. Reduction of radiotherapy-induced late complications in early breast cancer: the role of intensity-modulated radiation therapy and partial breast irradiation: part II—radiotherapy strategies to reduce radiation-induced late effects. Clinical Oncology, 17(2):98-110, 2005.

[29] Rodríguez-Ruiz A., Krupinski E., Mordang J.-J., Schilling K., Heywang-Köbrunner S.H., et al. Detection of breast cancer with mammography: Effect of an artificial intelligence support system. Radiology, 290(2):305–314, 2018. https://doi.org/10.1148/radiol.2018181371

[30] Hussain S., Quazilbash N.Z., Bai S., Khoja S. Reduction of variables for predicting breast cancer survivability using principal component analysis. In Proceedings - IEEE Symposium on Computer-Based Medical Systems, 131–134, 2015. https://doi.org/10.1109/CBMS.2015.62

[31] Howley T., Madden M.G., O'Connell M.L., Ryder A.G. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowledge-Based Systems, 19(5):363–370, 2006. https://doi.org/10.1016/j.knosys.2005.11.014

[32] Skittides C., Früh W.G. Wind forecasting using principal component analysis. Renewable Energy, 69:365–374, 2014. https://doi.org/10.1016/j.renene.2014.03.068

[33] Jhajharia S., Varshney H.K., Verma S., Kumar R. A neural network-based breast cancer prognosis model with PCA processed features. In International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, 1896–1901, 2016. https://doi.org/10.1109/ICACCI.2016.7732327

[34] Karthik S., Srinivasa Perumal R., Chandra Mouli P.V.S.S.R. Breast cancer classification using deep neural networks. Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques, 1:227–241, 2018. https://doi.org/10.1007/978-981-10-6680-1_12

[35] Liang B., Tian Y., Chen X., Yan H., Yan L., et al. Prediction of radiation pneumonitis with dose distribution: A Convolutional Neural Network (CNN) based model. Frontiers in Oncology, 9:1500, 2020. https://doi.org/10.3389/fonc.2019.01500

[36] Cireşan D.C., Giusti A., Gambardella L.M., Schmidhuber J. Mitosis detection in breast cancer histology images with deep neural networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8150 LNCS(PART 2), 411–418, 2013. https://doi.org/10.1007/978-3-642-40763-5_51

[37] Gerazov B., Conceicao R.C. Deep learning for tumour classification in homogeneous breast tissue in medical microwave imaging. In 17th IEEE International Conference on Smart Technologies, EUROCON 2017 - Conference Proceedings, 564–569, July 2017. https://doi.org/10.1109/EUROCON.2017.8011175

[38] Mambou S.J., Maresova P., Krejcar O., Selamat A., Kuca K. Breast cancer detection using infrared thermal imaging and a deep learning model. Sensors (Switzerland), 18(9):2799, 2018, https://doi.org/10.3390/s18092799

[39] Sun W., Tseng T.-L., Zheng B., Qian W. A preliminary study on breast cancer risk analysis using deep neural network. In Lecture Notes in Computer Science, 385–391, 2016. https://10.1007/978-3-319-41546-8_48

[40] Badré A., Zhang L., Muchero W., Reynolds J.C., Pan C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. Journal of Human Genetics, 66(4):359-369, 2021.

[41] Kim W., Kim K.S., Lee J.E., Noh D.Y., Kim S.W., et al. Development of novel breast cancer recurrence prediction model using support vector machine. Journal of Breast Cancer, 15(2):230-238, 2012.

[42] Belciug S., El-Darzi E. A partially connected neural network-based approach with application to breast cancer detection and recurrence. In 5th IEEE International Conference Intelligent Systems, 191-196, 2010.

[43] Surveillance, Epidemiology, and End Results (SEER), 2020. https://seer.cancer.gov/

[44] Buciński A., Baczek T., Krysiński J., Szoszkiewicz R., Załuski J. Clinical data analysis using artificial neural networks (ANN) and principal component analysis (PCA) of patients with breast cancer after mastectomy. Reports of Practical Oncology and Radiotherapy, 12(1):9–17, 2007. https://doi.org/10.1016/S1507-1367(10)60036-3

[45] Lever J., Krzywinski M., Altman N. Points of significance: Principal component analysis. Nature Methods, 14(7):641–642, 2017, https://doi.org/10.1038/nmeth.4346

[46] Lubis A.H., Sihombing P., Nababan E.B. Analysis of accuracy improvement in K-Nearest Neighbor using Principal Component Analysis (PCA). Journal of Physics: Conference Series, 1566:1, 2020. https://doi.org/10.1088/1742-6596/1566/1/012062

[47] Ritthipravat P. Artificial neural networks in cancer recurrence prediction. In Proceedings - 2009 International Conference on Computer Engineering and Technology, ICCET 2009, 2:103–107, 2009. https://doi.org/10.1109/ICCET.2009.84

[48] Gorunescu F., Gorunescu M., El-Darzi E., Gorunescu S. A statistical evaluation of neural computing approaches to predict recurrent events in breast cancer. In IEEE 2008 4th International IEEE Conference "Intelligent Systems" (IS) - Varna, Bulgaria, 11-38–11-43, 2008. https://10.1109/IS.2008.4670506

[49] Jerez J.M., Franco L., Alba E., Llombart-Cussac A., Lluch A., Ribelles N., Munárriz B., Martín M. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. Breast Cancer Research and Treatment, 94(3):265–272, 2005. https://doi.org/10.1007/s10549-005-9013-y

[50] Maier H.R., Dandy G.C. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environmental Modelling and Software, 15(1):101–124, 2000. https://doi.org/10.1016/S1364-8152(99)00007-9

[51] Prusty M.R., Jayanthi T., Chakraborty J., Velusamy K. Feasibility of ANFIS towards multiclass event classification in PFBR considering dimensionality reduction using PCA. Annals of Nuclear Energy, 99:311–320, 2017. https://doi.org/10.1016/j.anucene.2016.09.015