

RESEARCH

Open Access



# Breast cancer prediction using gated attentive multimodal deep learning

Safak Kayikci<sup>1\*</sup> and Taghi M. Khoshgoftaar<sup>1</sup>

\*Correspondence:  
skayikci@fau.edu

<sup>1</sup> Department of Electrical  
Engineering and Computer  
Science, Florida Atlantic  
University, Boca Raton, USA

## Abstract

Women are prone to breast cancer, which is a major cause of death. One out of every eight women has a lifetime risk of developing this cancer. Early diagnosis of this disease is critical and enhances the success rate of cure. It is extremely important to determine which genes are associated with the disease. However, too many features make studies on gene data challenging. In this study, an attention-based multimodal deep learning model was created by combining data from clinical, copy number alteration and gene expression sources. Attention-based deep learning models can analyze mammography images and identify subtle patterns or abnormalities that may indicate the presence of cancer. These models can also integrate patient data, such as age and family history, to improve the accuracy of predictions. The objective of this study is to help breast cancer prediction tasks and improve efficiency by incorporating attention mechanisms. Our suggested methodology employs multimodal data and generates insightful characteristics to improve the prediction of the prognosis for breast cancer. It is a two-phase model; the first phase generates the stacked features using a sigmoid gated attention convolutional neural network, and the second phase uses flatten, dense and dropout processes for bi-modal attention. Based on our findings, the proposed model produced successful results and has the potential to significantly improve breast cancer detection and diagnosis, ultimately leading to better patient outcomes.

**Keywords:** Breast cancer prediction, Attention mechanism, Deep learning, Multimodality

## Introduction

Breast cancer is among the most common types of cancer worldwide. Although it occurs mostly after the age of 40, some women with high-risk features may develop breast cancer at a younger age. Breast cancer is usually seen in women, but it can also be found in men, although it is rare. In countries with a low or medium Human Development Index, the breast cancer mortality rate is 48%. This rate is four times higher than in countries with high or very high human development indicators [1]. Early diagnosis of breast cancer with effective screening programs positively affects the breast cancer treatment process. Mammography screening has significantly reduced breast cancer mortality rates in high-income countries

Cancer is a malady where cells become abnormal and make more cells uncontrollably. Breast cancer begins in the tissues that make up the breasts. Tumors are classified into two groups: benign or malignant. Benign tumors are not considered cancerous and do not spread throughout the body. Malignant tumors divide cells and damage surrounding tissues. When cancerous cells spread outside the breast, they usually settle in the lymph nodes under the armpit. From there, they are distributed throughout the body through lymph and blood circulation.

It is very important to identify cancerous cells in advance and to take the necessary precautions in the initial stages. In addition to methods such as mammography, magnetic resonance imaging (MRI), ultrasound and tomography, a biopsy by taking a piece of the cancerous cell gives definite results. Pathologists make decisions to try to classify benign, malignant, and normal tissues by examining these cells. Since the analysis of these histopathological images is tedious and time-consuming, computer-aided decision support mechanisms will be very helpful in this area. The knowledge and care of pathologists throughout the examination of these samples are of great importance for the diagnosis to be made correctly. Computer-aided software can reduce the risk of pathologists making wrong decisions due to factors such as possible fatigue and carelessness. This allows experts to focus on difficult-to-diagnose cases.

Decision support, expert, computer-aided design (CAD) soft computing and decision based systems developed with machine learning can help doctors diagnose diseases at early stages [2]. In addition to shortening the cost and waiting time, they also prevent incorrect decisions made by medical personnel. The main purpose of CADs is to combine human experience and technological knowledge to obtain more precise diagnoses. As a result, doctors can provide the necessary treatment by using these technologies.

### **Literature overview**

In his study, Poyraz [3] applied data mining methods to the Wisconsin breast cancer dataset and compared the results according to performance criteria. The J48 algorithm, Decision Tree algorithm, Naive Bayes, logistic regression and K-Star methods were used in the Waikato Environment for Knowledge Analysis (WEKA) working environment. He stated that the best accuracy value was obtained with logistic regression.

Iseri [4] worked on diagnosing breast cancer by applying machine learning methods to mammogram images. The study was carried out in two stages: the detection of microcalcification regions in the mammogram images and the classification of these regions according to whether they are malignant or benign. A software called Breast Cancer Detection System (BCDS) was developed in the MATLAB environment. The software was able to detect breast cancer by using four feature extraction methods, multilayer neural network and support vector machines as classifiers.

In his study, Şık [5] investigated the effect of data mining applications on the early diagnosis of cancer. Various classification methods such as Bayesian Networks, Naive Bayes, multilayer perceptron, logistic regression, probabilistic gradient descent, Sequential Minimal Optimization, IB1, K-Star, PART, Logistic Model Trees and random forests were applied to the Wisconsin Breast Cancer dataset in the WEKA environment. When comparing the classification results, parameters such as Kappa statistic, accuracy, precision, sensitivity,

F-measure and receiver operating characteristic (ROC) area were taken into account. He achieved 97.40% accuracy.

Bazazeh and Shubair [6] applied support vector machines, random forest and Bayesian network methods to the Wisconsin breast cancer dataset in their study on the early detection of breast cancer. In this study, in which WEKA software was used, Bayesian Networks showed the best performance according to specificity and sensitivity values. Considering the ROC curve parameter, the random forest method gave the best results. In terms of accuracy, originality and precision, support vector machines showed the best performance.

In their study, Sherafatiyan [7] used miRNA datasets of breast cancer and utilized tree-based classification models to identify minimal biomarkers. In addition to the suggested biomarkers, the ultimate significant microRNAs in breast cancer prediction have been described.

On the other hand, Turgut et al. [8] performed data classification by applying various machine learning methods on two different microarray breast cancer datasets. The authors aimed to diagnose cancer with high accuracy by using random logistic regression and iterative feature elimination feature selection methods. Support vector machines performed best in two microarray breast cancer datasets after applying two different feature selection methods.

Dhahri et al. [9] worked on machine learning algorithms for automatic detection of breast cancer. They explained that combining feature-based preprocessing methods and classification algorithms can give better results in the diagnosis of breast cancer.

Tseng et al. [10] conducted a study on determining breast cancer metastasis using machine learning technologies. They determined that the random forest-based machine learning model is the most appropriate method to predict breast cancer metastasis at least three months in advance.

Magna et al. [11] studied the use of machine learning, deep learning and word insertion applications in the classification of breast cancer by using the medical history of the patients. They tried to put forward a recommendation system that supports the physician's decision making.

Reddy et al. [12] used the deep neural networks (DNNs) method with support value for the diagnosis of breast cancer. The experimental results show that the proposed DNNs got better results than the existing methods.

Saxena and Gyanchandani [13] examined machine learning methods to make computer-assisted breast cancer diagnoses using histopathology. After examining many different approaches, it was seen that machine learning studies on breast cancer generally focused on deep learning.

Kayikci and Khoshgoftaar [14] previously studied the same breast cancer dataset. In their previous work, they created a multimodal structure for data from multiple sources. Later, this model was tested on different machine learning methods. They achieved 82% accuracy in decision trees, 90% in random forests and 88% in support vector machines.

## Materials and methods

### Dataset

There are 1980 patient records in the METABRIC dataset [15]. These data come from three different sources: clinical data, copy number alteration (CNA) data, and gene

expression data. As the output classification data, the value of long-term survivors was defined as 1 (491 records), and the data of short-term survivors was defined as 0 (1489 records). The threshold value between short-term and long-term is five years.

During the preprocessing of genetic and CNA data, unknown and null values were arranged with a weighted neighbor algorithm. In gene data, output values are assigned as  $-1$  for under-expressed genes,  $0$  for baseline genes, and  $1$  for over-expressed genes. CNA attributes have five discrete values ( $-2, -1, 0, 1, 2$ ). Normalization was applied to clinical data and scaling was done in the range of  $[0, 1]$ .

Since the feature number of CNA data is 26,000 and the feature number of genetic data is 24,000, the most important ones among these features are selected and reduced. For this purpose, the Maximum Relevance—Minimum Redundancy (MRMR) algorithm was used. At every iteration, the goal is to choose the most relevant property relating to the objective variable and the least redundant property that has been chosen at prior iterations. The affinity of an attribute  $f$  at the  $i$ -th loop is calculated as the F-statistic among the attribute and the objective variable. The redundancy is the mean (Pearson) correlation among the attribute and all the attributes that have been chosen at previous loops defined in Eq. 1 [14]. In the end, the number of genetic features was reduced to 400 and the number of CNA features to 200. In clinical data, 25 features were used.

$$score_i(f) = \frac{F(f, target)}{\sum_{s \in \text{features selected until } (i-1)} |corr(f, s)| / (i - 1)} \quad (1)$$

### Attention mechanism

Attention is a cognitive function for humans. An important feature of human perception is that it can focus only on the necessary part, rather than processing all the information completely at once [16]; for example, when human perception looks at a scene, it focuses on a certain region, ignoring unnecessary data to extract the information it needs [17]. The attention model that emerged from this was first introduced for the machine translation problem [18] and has become an important concept in the neural networks literature for other application areas [19]. Although the principle of attention models is the same, researchers have made some changes and improvements to better adapt attention mechanisms to specific tasks [20].

Attention mechanisms are so important that they are ubiquitous and a necessary component of neural machine learning systems [21, 22]. Neural processes that require attention have been extensively studied in neuroscience and computational neuroscience. One issue that has been studied in particular is visual attention: many animals focus on certain parts of their visual input to calculate adequate responses. This principle has a huge impact on neural computation, as it is necessary to select the most important information rather than using all available information, much of it unrelated to calculating the neural response. An example with the creation of image captions can be given to explain the attention mechanism. The Convolutional Neural Network (CNN), which can also be considered as the classical method, was the method used for this method. In the next stage, the Recurrent Neural Network (RNN) further developed this phenomenon. But the problem with these methods is that when the model tries to construct the next word of the caption, that word will usually only describe a part of the image. But with the attention network or mechanism, the

image is first divided into  $n$  pieces and the  $h_1, \dots, h_n$  of each piece is calculated with a CNN representation. When the RNN creates a new word, the attention mechanism focuses on the relevant part of the image, so the decoder uses only certain parts of the image.

**Proposed model**

The data first passes through two separate CNNs simultaneously. Kernel sizes in these layers are 3 and 2, and initial values are assigned with the Glorot normal initializer.

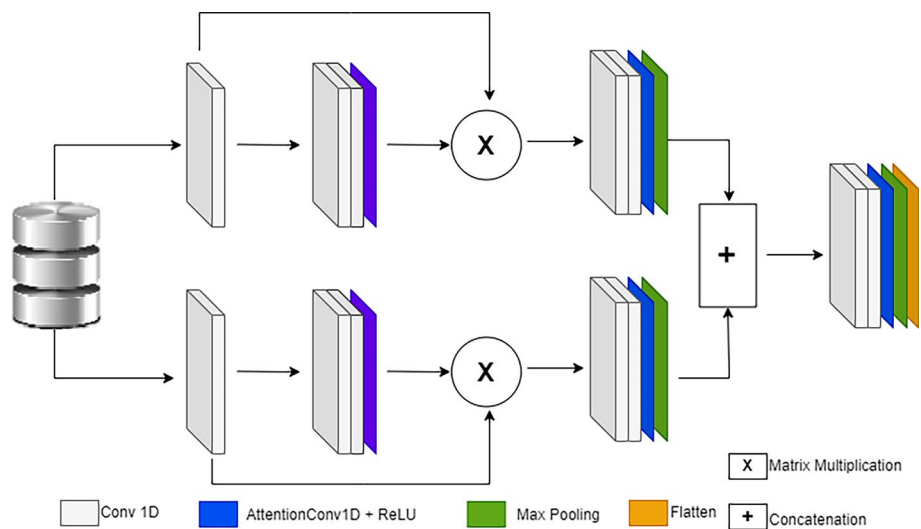
The number 1 is used for the stride value for fixed values of bias. At the end of the convolution, two different feature maps are created. The Rectified Linear Unit (ReLU) activation function was used as the activation function. After the convolutional layer, attention, max-pooling and flatten layers are applied as shown in Fig. 1.

The attention layer creates the attention-weighted matrix over the values from the biconvolutional layer. Cross-modality matrices are computed over CNA and genetic data as an example below. Max pooled features for CNA are represented by  $cna$  and for genetic expression data is  $gen$  respectively.

$$\begin{aligned} matrixe_1 &= cna.gen^T \\ matrixe_2 &= gen.cna^T \end{aligned} \tag{2}$$

After these matrices are created, probability distribution scores ( $pd_1$  and  $pd_2$ ) are calculated using the softmax activation function among these matrices. Again, the max pooled feature of each source and the probability distribution matrices are used to calculate attentive features ( $f_1$  and  $f_2$ ).

$$\begin{aligned} f_1 &= pd_1.cna \\ f_2 &= pd_2.gen \end{aligned} \tag{3}$$



**Fig. 1** Operations on single datasource

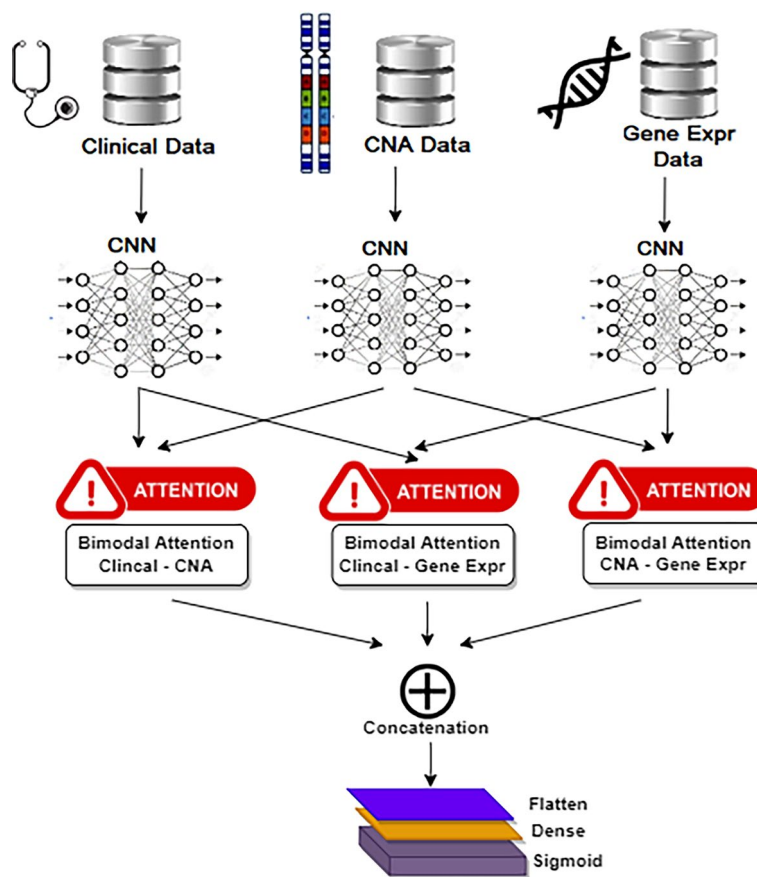
Finally, the multiplicative gating function [23] is applied to get the related factors from each source which is an element-wise matrix multiplication of attentive features and max pooled features.

$$\begin{aligned}
 attention_1 &= f_1 \odot cna \\
 attention_2 &= f_2 \odot gen
 \end{aligned}
 \tag{4}$$

The concatenation of  $attention_1$  and  $attention_2$  matrices forms the bi-modal attention between CNA and gene expression data. Bi-attention between CNA-clinical and gen-clinical is also computed in the same way. The flowchart of bimodal attention mechanism is shown in Fig. 3.

$$biattention_{(cna,gen)} = concatenation[attention_1, attention_2]
 \tag{5}$$

In the last stage, flatten, sigmoid and 50% dropout processes were applied. The general architecture is shown in Fig. 2.



**Fig. 2** General architecture

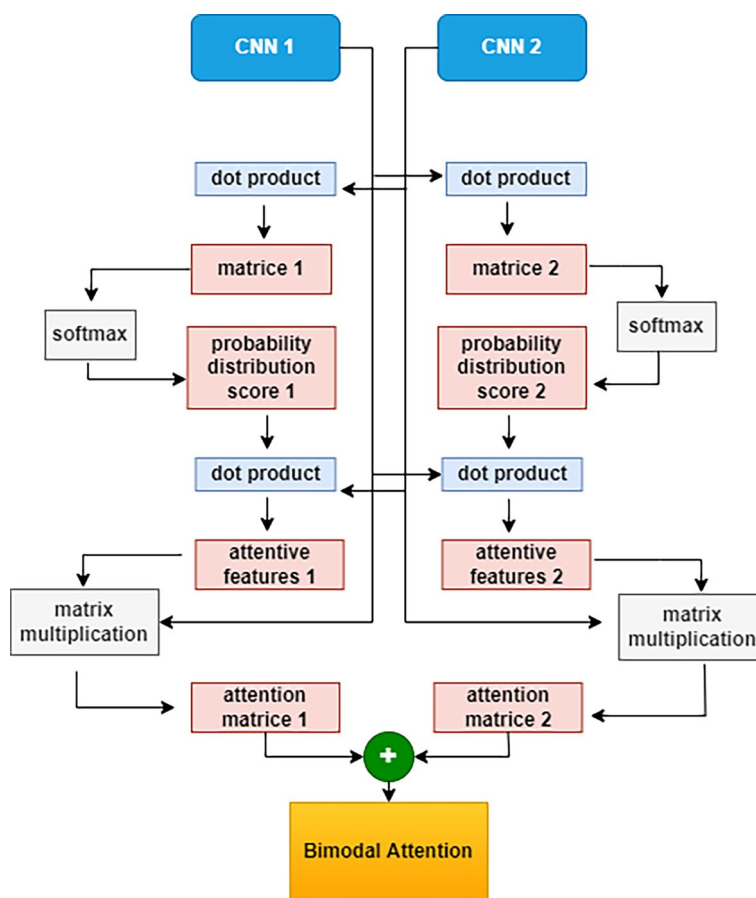
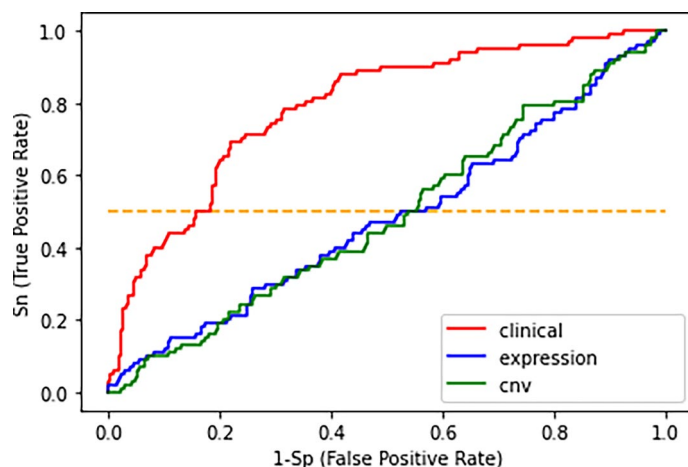


Fig. 3 BiModal attention flowchart

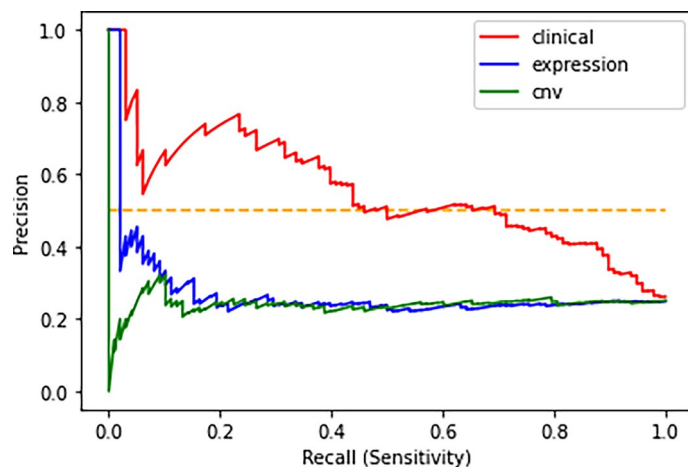
### Results and comparison

The receiver operating characteristic (ROC) defines the diagnostic success of a test by means of high specificity and sensitivity values. Also, it is used to compare these values obtained at different cut-off points. On the ROC curve, true positive values (sensitivity) are represented as a function of false positive values (1-specificity) concerning different intersection points. Each point on the curve represents the sensitivity/specificity ratio relative to a given threshold. The area under the curve (AUC) shows how successfully a parameter is separated between the two groups. The ROC curve for the datasets is shown in Fig. 4. The best performance is shown in clinical data with 0.82, expression data with 0.69 and CNA data with 0.68.

The precision–recall curve can be used when the data is unbalanced. Precision shows the relevance, while recall shows how many actual relevant values. Also, this curve shows the tradeoff between precision and recall for different thresholds. If the area under the curve is high, it means that both recall and precision are high. High precision is related to a low false positive rate while a high recall value is related to a low false negative rate. This indicates that the model returns consistent values. The precision recall curve for the datasets is shown in Fig. 5.



**Fig. 4** Receiver operating characteristics (ROC)



**Fig. 5** Precision recall curve

**Table 1** *T*-test values

	AUC	Accuracy	Sensitivity
METABRIC			
<i>t</i> -value	22.45	18.42	26.01
<i>p</i> -value	1.23E-14	3.26E-13	2.13E-15
TCGA_BRCA			
<i>t</i> -value	- 0.46	54.21	59.59
<i>p</i> -value	1.28	4.27E-21	7.83E-22

We used the TCGA-BRCA dataset [24] to compare the method’s validity. This dataset contains the same features as METABRIC (clinical, CNA, gene expression data) There are 250 records for long-term survivor patients and 830 for short-term survivor patients. We used a two-sample *T*-test and a one-way ANOVA test for comparison. The AUC, accuracy, sensitivity, *p*-values and *t*-values are shown in Tables 1 and 2.



**Table 2** Anova test values

	AUC	Accuracy	Sensitivity
METABRIC			
t-value	545.26	374.89	671.03
p-value	1.23E-14	3.26E-13	2.13E-15
TCGA_BRCA			
t-value	0.221	2938.22	3551.52
p-value	1.28	4.27E-21	7.83E-22

**Table 3** Performance indicators for benchmarking MDNNMD

Model	AUC	Accuracy	Precision	Sensitivity
MDNNMD	0.845	0.826	0.749	0.450
Proposed model	0.950	0.912	0.841	0.798

The *T*-test is based on the theory that samples of two independent variables have the same equivalent means. In these measurements, it is assumed that the inhabitants have the same variance. It is possible to use this test when two independent variables are from the same or different groups. If a high *p*-value is taken (like > 0.05 or > 0.1), we have to accept null hypotheses of the mean scores. If this *p*-value is less than the threshold value, we can ignore the null hypothesis for equal means.

On the other hand, the ANOVA test is used to measure the null hypothesis where two or more groups are in the same population mean. It is mostly applied in groups of different sizes. The requirements to be considered in this test are that the samples are independent, each sample comes from a normally distributed group, and the standard deviations of the groups are equal (also called homoscedasticity).

For the *T*-test, we achieved *t*-values of 22.45 (AUC), 18.42 (accuracy) and 26.01 (sensitivity) and for the ANOVA-test, values of 545.26 (AUC), 374.89 (accuracy) and 671.03 (sensitivity). On the other hand, we considered *p*-values on both tests as 0.00 because they were very small enough to be neglected. Both *t*-values and *p*-values show that our recommended architecture is based on statistically sound results and will be helpful in breast cancer prediction methods.

Sun et al. [25] proposed a multimodal deep neural network by integrating multi-dimensional data (MDNNMD) with using the same METABRIC and TCGA-BRCA datasets. The AUC values and supplementary performance measures are listed in the Table 3. The comparisons of the models that were used in [25] are performed at a threshold of 0.45 and a specificity of 95%. The results of proposed gated attentive method enhance the assessment of breast cancer survival by 10.5%, 8.6%, 9.2% and 34.8% in regards to AUC, accuracy, precision, sensitivity values respectively. Further experimental benchmarking of the proposed method with other methods will be made as a future work to provide empirical evidence that supports the superiority or effectiveness of our proposed method over existing ones. The results of the experimental benchmarking analysis will provide valuable insights and guidance by identifying gaps in current methods or suggesting areas for improvement.

## Conclusions

Along with the development of artificial intelligence technologies, namely deep learning techniques and CNNs, there have been important developments in the diagnosis of diseases. Such techniques do not require a feature domain description and can achieve classification performances that can outperform even human experts. The results obtained with the deep learning approach for mammogram-based breast cancer risk assessment are promising. Moreover, the use of big data and machine learning provides new opportunities to improve the accuracy of screening tests and better guide scanning protocols, especially for image-based scanning.

## Abbreviations

AUC	Area under curve
BCDS	Breast cancer detection system
CAD	Computer aided design
CNA	Copy number alteration
CNN	Convolutional neural network
DNN	Deep neural network
MDNNMD	Multimodal deep neural network multi-dimensional data
MRMR	Maximum relevance minimum redundancy
MRI	Magnetic resonance imaging
RELU	Rectified linear unit
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SVM	Support vector machines
TCGA	The cancer genome atlas
WEKA	Waikato environment for knowledge analysis

## Acknowledgements

Not applicable.

## Author contributions

SK designed the research, ran experiments, analyzed results, and wrote the manuscript. TMK guided the direction of the research and helped to finalize the work. Both authors read and approved the final manuscript.

## Funding

This research received no specific Grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Availability of data and materials

The METABRIC and TCGA data are available at references.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 16 October 2022 Accepted: 2 May 2023

Published online: 13 May 2023

## References

1. Office for National Statistics. Mortality statistics-underlying cause, sex and age; 2019. <https://www.nomisweb.co.uk/datasets/mortsa>.
2. Postalcioglu S, Erkan K. Soft computing and signal processing based active fault tolerant control for benchmark process. *Neural Comput Appl*. 2009;18:77–85.
3. Poyraz O. Tıpta veri madenciliği uygulamaları: Meme kanseri veri seti analizi. Master's thesis, Trakya Üniversitesi Fen Bilimleri Enstitüsü; 2012.
4. İşeri İ, et al. Mamogram görüntülerinden makine öğrenmesi yöntemleri ile meme kanseri teşhisi; 2014.

5. Şık MŞ. Veri madenciliği ve kanser erken teşhisinde kullanımı. Master's thesis, İnönü Üniversitesi Sosyal Bilimler Enstitüsü; 2014.
6. Bazazeh D, Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In: 2016 5th international conference on electronic devices, systems and applications (ICEDSA); 2016. p. 1–4.
7. Sherafatian M. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene*. 2018;677:111–8.
8. Turgut S, Dağtekin M, Ensari T. Microarray breast cancer data classification using machine learning methods. In: 2018 electric electronics, computer science, biomedical engineering's meeting (EBBT). IEEE; 2018. p. 1–3.
9. Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi M. Automated breast cancer diagnosis based on machine learning algorithms. *J Healthc Eng*. 2019;2019:1.
10. Tseng Y-J, Huang C-E, Wen C-N, Lai P-Y, Wu M-H, Sun Y-C, Wang H-Y, Lu J-J. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *Int J Med Inform*. 2019;128:79–86.
11. Magna AAR, Allende-Cid H, Taramasco C, Becerra C, Figueroa RL. Application of machine learning and word embeddings in the classification of cancer diagnosis using patient anamnesis. *IEEE Access*. 2020;8:106198–213.
12. Vaka AR, Soni B, Reddy S. Breast cancer detection by leveraging machine learning. *ICT Exp*. 2020;6(4):320–4.
13. Saxena S, Gyanchandani M. Machine learning methods for computer-aided breast cancer diagnosis using histopathology: a narrative review. *J Med Imaging Radiat Sci*. 2020;51(1):182–93.
14. Kayikci S, Khoshgoftaar T. A stack based multimodal machine learning model for breast cancer diagnosis. In: 2022 international congress on human–computer interaction, optimization and robotic applications (HORA). IEEE; 2022. p. 1–5.
15. The European Genome-phenome Archive (EGA). Metabarc; 2016. <https://www.ncbi.nlm.nih.gov/ega/>. Accessed August 2022.
16. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing*. 2021;452:48–62.
17. Rensink RA. The dynamic representation of scenes. *Vis Cogn*. 2000;7(1–3):17–42.
18. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Preprint; 2014. [arXiv:1409.0473](https://arxiv.org/abs/1409.0473).
19. Chaudhari S, Mithal V, Polatkan G, Ramanath R. An attentive survey of attention models. *ACM Trans Intell Syst Technol (TIST)*. 2021;12(5):1–32.
20. Kayikci S. A deep learning method for passing completely automated public turing test. In: 2018 3rd international conference on computer science and engineering (UBMK). IEEE; 2018. p. 41–44.
21. Mo L, Zhu Y, Wang G, Yi X, Wu X, Wu P. Improved image synthesis with attention mechanism for virtual scenes via UAV imagery. *Drones*. 2023;7(3):160.
22. Khoshboresh-Masouleh M, Shah-Hosseini R. Real-time multiple target segmentation with multimodal few-shot learning. *Front Comput Sci*. 2022;4:1.
23. Dhingra B, Liu H, Yang Z, Cohen WW, Salakhutdinov R. Gated-attention readers for text comprehension. Preprint; 2016. [arXiv:1606.01549](https://arxiv.org/abs/1606.01549).
24. National Cancer Institute. The Cancer Genome Atlas (TCGA). <https://portal.gdc.cancer.gov/projects/TCGA-BRCA>. Accessed August 2022.
25. Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans Comput Biol Bioinf*. 2018;16(3):841–50.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

---