EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

Multimodal Data-Driven Intelligent Systems for Breast Cancer Prediction

Shanmugavadivu Pichai[1,*], G. Kanimozhi[2], M. Mary Shanthi Rani[3], and N. K. Riyaz[4]

[1,2,3] The Gandhigram Rural Institute, Tamil Nadu, India.
[4] Birla Institute of Technology and Science (BITS, Pilani), Rajasthan, India.

Abstract

Cancer, a malignant disease, results from abnormalities in the body cells that lead to uncontrolled growth and division, surpassing healthy growth and stability. In the case of breast cancer, this uncontrolled growth and division occurs in breast cells. Early identification of breast cancer is key to lowering mortality rates. Several new developments in artificial intelligence predictive models show promise for assisting decision-making. The primary goal of the proposed study is to build an efficient Breast Cancer Intelligent System using a multimodal dataset. The aim is to to establish Computer-Aided Diagnosis for breast cancer by integrating various data.

This study uses the TCGA "The Cancer Genome Atlas Breast Invasive Carcinoma Collection" (TCGA-BRCA) dataset, which is part of an ongoing effort to create a community integrating cancer phenotypic and genotypic data. The TCGA-BRCA dataset includes: Clinical Data, RNASeq Gene Data, Mutation Data, and Methylation Data. Both clinical and genomic data are used in this study for breast cancer diagnosis. Integrating multiple data modalities enhances the robustness and precision of diagnostic and prognostic models in comparison with conventional techniques. The approach offers several advantages over unimodal models due to its ability to integrate diverse data sources. Additionally, these models can be employed to forecast the likelihood of a patient developing breast cancer in the near future, providing a valuable tool for early intervention and treatment planning.

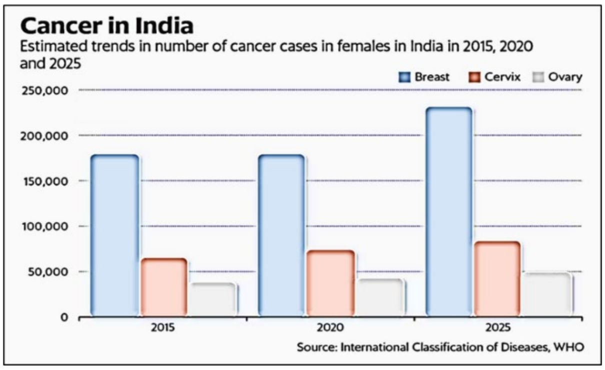
Keywords: Deep neural networks, Multimodality, Breast cancer, Genome data.

Received on 11 02 2024, accepted on 28 05 2024, published on 25 06 2024

doi: 10.4108/eetpht.10.6424

1. Introduction

Cancer is one of the leading causes of death worldwide, primarily due to late diagnosis and inadequate treatment options. It is characterized by the abnormal and uncontrolled development of cells in the body which can spread from one region to another [1]. Figure 1 illustrates the projected number of cancer cases in India for 2015, 2020, and 2025 by the World Health Organization. [2]



Cancer in India
Estimated trends in number of cancer cases in females in India in 2015, 2020 and 2025

Source: International Classification of Diseases, WHO

Figure 1. Number of estimated cancer cases in India

[*] Corresponding author. Email: psvadivu67@gmail.com

Breast cancer (BC) is the most sever and fatal ailment afflicting women. It has recently surpassed other cancer incidences as a significant cause of malignancy, particularly in women. Alarmingly, younger age groups are experiencing a higher prevalence than the worldwide average [3]. By 2023 it is anticipated that there will be 55,720 new cases of ductal carcinoma in situ (DCIS), 297,790 new instances of invasive BC, and 43,170 BC fatalities among women in the United States. Nearly 10% of BCs are hereditary or caused by inherited DNA mutations, with most hereditary cases linked to defective BRCA1 and BRCA2 genes [4,5].

Options for BC treatment have increased in both complexity and effectiveness. Improvements in machine learning (ML) and deep learning (DL) have facilitated the development of automated computer-aided diagnosis (CAD) systems which deliver precise results, increasing the efficiency of malignant tumor identification and saving time through optimal utilization. [6,7].

Numerous studies have been conducted with data based on multimodal and unimodal sources to predict BC prognosis using clinical data, imaging biomarkers, and genetic markers. However, traditional BC prediction approaches primarily rely on unimodal data, which fails to capture the full spectrum of BC characteristics. Though conventional unimodal methods have proven effective in predicting BC, they are insufficient for accurate diagnosis. To minimize medical errors, developing a multimodal approach is essential to accurately and precisely predict BC using multiple imaging modalities. This approach facilitates a more precise and reliable diagnosis. Multimodal deep learning provides a comprehensive understanding of data, improving accuracy and efficiency [9]. This powerful technique allows for the extraction of meaningful information from large datasets by combining multiple modalities.

The main focus of this study is to formulate a computer aided diagnosis (CAD) for BC by integrating various data modalities. Combining different data sources enables more reliable and accurate models for diagnosing and predicting outcomes than traditional techniques. The development of a highly generic and high-performing BC prediction system using different modalities is projected to give viable solutions for BC prognosis with high accuracy. The research aims to investigate the potential of Artificial Intelligence tools, increasingly achieving significant advancements in various research fields. The results will confirm that the proposed system is a feasible alternative to existing computational systems.

Section 2 provides a literature review summarizing unimodal and multimodal dataset predictions. Section 3 details the dataset and proposed methodology of the research work. Section 4 presents results and discussion and Section 5 concludes the paper with observations and future directions.

2. Literature Review

This review serves as a foundation for the study of existing solutions for BC prediction. Van't Veer *et al.* [10] analyzed 117 primary breast carcinomas using DNA microarrays and supervised classification algorithms to identify 70 genetic prognostic signatures. These signatures were used to establish prognostic markers for detecting carcinoma. The study found that inadequate signatures were linked to metastatic, invasion, and angiogenic pathways, resulting in improved predictive performance for disease outcomes.

Yap *et al.* [11] explored the use of Deep Learning (DL) methods to detect breast lesions in ultrasound images, experimenting with U-Net, LeNet, and a pretrained AlexNet. Their experiments, conducted on two custom datasets with 306 and 163 images, demonstrate that pre-trained AlexNet-based models outperformed all other models, achieving F-measures of 0.91 and 0.89, respectively.

An integrated deep learning architecture has been proposed by Antari *et al.* [12] capable of categorizing, segmenting, and detecting breast tumors. The authors employed a Full Resolution Convolutional Network (FRCN) for tumor segmentation, a Deep Convolutional Neural Network (CNN) for classification, and a YOLO-based system for tumor detection. The dataset size was increased 8-fold through application of the YOLO algorithm to expand the dataset size synthetically. The researchers tested the model against the digital database of digital mammograms from the INbreast dataset, which produced a detection accuracy of 98.96%, and a dice score of 92.69%.

In Sun D *et al.* [13] initiated BC prediction by combining genome data with pathology images. A multiple kernel learning method was used and compared with various independent models that used genome data only. Their findings suggested that combining clinical images with a 10-fold cross-validation contributed to the robustness of the prediction. Gevaert *et al.* [14] integrated clinical and 70 gene data using three strategies: full integration, decision integration, and partial integration on Bayesian networks. The results showed that methods that use clinical and microarray data have better or comparable results to those that do not use clinical or microarray data.

Sun D *et al.* [15] enhanced BC prediction prognosis using a multimodal deep neural network (NN). They combined multi-dimensional data, including gene expression, copy number alteration profiles, and clinical data, using novel deep learning techniques. This approach outperformed single-dimensional prediction methods. By combining two independent models of microarray and clinical data, Khademi *et al.* [16] developed a Probabilistic Graphical Model (PGM) for BC prediction and detection. They began by reducing the dimensionality of microarray data with Principal Component Analysis (PCA), and then built an in-depth belief network to extract data feature representations. The clinical data was then processed through a structural learning algorithm before merging with SoftMax nodes to calculate BC prognosis.

Qian *et al*. 2021 [17] employed comparable modalities for diagnosing BC, including multimodal, multiview ultrasound imaging. The deep-learning framework was employed to construct the model based on US B-mode multiview ultrasound images and view-level multiview US images as inputs for each potential clinical test lesion. The model then analyzed the suspected lesion from multiple perspectives and provided an overall likelihood of malignancy. The model's performance was tested with each bimodal and multimodal combination to predict the malignancy risk and establish the Breast Imaging Reporting and Data System (BI-RADS) category.

In Binder *et al*. [18], a novel comprehensive machine-learning approach was proposed for the identification and prediction of morphologic and molecular features from histologic BC imaging datasets. The predictions are derived from a morphologic feature training database, which contains manually annotated types of breast cells in a variety of data modes, as well as histological images from the TCGA database. Liu *et al*. [19] designed a deep learning prediction model to predict molecular subtypes of BC. The framework is based on gene and image data, employed with Image Filtration. Validation was performed by combining deep NN with convolutional networks to achieve a high level of accuracy. Arya *et al*. [20] proposed a sophisticated, multimodal-based deep learning model. This model is further enhanced to develop the generation of convoluted feature maps and the extraction of stacked features using Sigmoid CNN algorithm and Random Forest Classifier.

In BC prediction, the focus has shifted to multimodality from unimodality approaches. This study assesses the efficacy of these two approaches and identifies areas for further research and its advancement. Deep Learning models utilizing multimodal datasets are recommended due to their ability to provide richer information than their unimodal counterparts. Additionally, multimodal models are capable of processing multiple data sources, which is an advantage over their unimodal peers.

3. Methodology

3.1. Dataset

The data was collected from The Cancer Genome Atlas (TCGA), the world's most extensive repository of genomic data. The Center for Cancer Genomics of the National Cancer Institutes (NCI) with National Human Genome Research Institute initiated TCGA in 2006. From the TCGA repository, this research specifically used the TCGA Breast Invasive Carcinoma Collection (TCGA-BRCA) dataset [21], a repository of over 10,000 patient profiles and related genomic data from BC patients. This data set contains information related to the pathological and molecular features of BC tumors, as well as information about other demographic factors such as age, race, and gender.

3.2. Data Preprocessing

Data preprocessing is an important element of machine learning projects, allowing the clean up and modification of data to enhance its usability and resolve issues. Before executing any algorithms, it is necessary to verify the representation and quality of the data. This dataset contained no missing values, and values with no counts were treated as 0. The primary objective is to preprocess the data into the format required by the models to ensure a reasonably accurate representation of the data [22].

Clinical Data Preprocessing

The TCGA-BRCA clinical data mainly include data from 1,097 patients, covering variables such as age at diagnosis, vital status, days to death, days to last follow-up, tumor status, pathologic stage, gender, and race. Vital status refers to whether the patient is alive, deceased or unknown. Days to last follow-up is calculated as the number of days between the last follow-up and the initial diagnosis. Table 1 provides an overview of the dataset.

Table.1 Overall Analysis of the cardiovascular disease dataset

Dataset Summary	Analysis
Dataset	TCGA-BRCA
#Patients	1097
#Alive	945
#Deceased	152
Median Age	59
Age Range	26-90
Gender (M:F)	12:1085

Data imputation (DI) is the process of filling in missing values within a dataset. This step can improve data quality and accuracy by replacing missing values with estimates based on existing information [23]. DI can help prevent bias and skew when analyzing data and reduce errors due to incomplete datasets. It can also reduce noise in a dataset and make it easier to identify meaningful patterns in the data. The DI techniques for the various features were carefully chosen depending on the nature of the data and its value [24]. In this work, DI is applied if the attributes have 'not available' values based on the count and importance of the feature(s).

One-hot encoding, also known as binary encoding, is one of the data encoding techniques during preprocessing. Each category is represented as a separate column with binary values (1 or 0). Additionally, it helps reduce overfitting by providing features that can be used to train the model [25]. In this study, one-hot encoding was performed on the three attributes gender, tumor status, and vital status. After one-hot encoding, attribute data contains values of mostly 0s and 1s, which may lead to inefficient pattern during the training phase. Other attributes such as race, margin status, histological type, and pathologic stage are converted into categorical data as per the category found in the dataset. Table 2 consists of the clinical characteristics of the TCGA-BRCA clinical Dataset.

Table 2. Clinical characteristics of TCGA-BRCA clinical Dataset

Attribute	Description
PID	Patient ID
Days to Birth	No. of days to birth
Gender	Female 1; Male-0
Race	Race Category (white-0, black or African American-1, Asian or American Indian-2)
Tumor_Status	Neoplasm disease stage (tumor free-0; with tumor-1)
Days_to_last_followup	Last contacted days
Days_to_Death	No. of days to death
Age	Patient's age in years (26 - 90 years)
Margin Status	Close-0; positive-1; negative-2
Pathologic Stage	Pathologic stage I to stage V and stage X
Histological Type	Infiltrating Ductal Carcinoma-1; Infiltrating Lobular Carcinoma-2; Metaplastic Carcinoma-3; Mucinous Carcinoma-4; Mixed and Others-5
lymph_node_examined_count	Count of Lymph node examination
Initial_Diagnosis_Year	Initial diagnostic year (range from 1988 – 2013)
Vital Status	Predictor class (0- Alive; 1- Deceased)

Datasets are typically represented by a distribution of values that helps in understanding the data. The data distribution can reveal the underlying structure, such as its range, outliers etc. Figure 2 illustrates the data distribution frequency of the TCGA BRCA dataset. Analyzing the frequency can provide valuable insights into the prevalence and characteristics among clinical parameters.

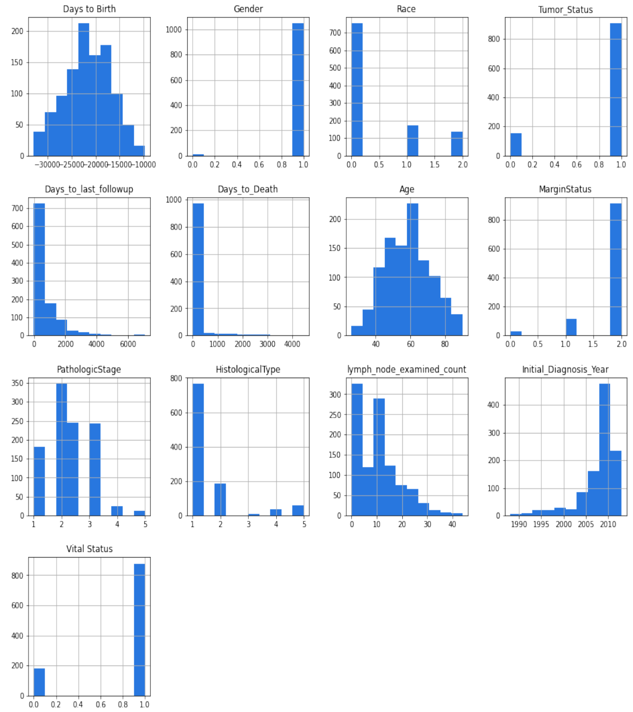


Figure 2. Data distribution of TCGA BRCA clinical data

Feature selection involves choosing the most significant features to enhance the performance of the computational models, which helps to avoid overfitting, which is essential when constructing a model. The dataset was processed using a correlation matrix and a heatmap feature selection approach.

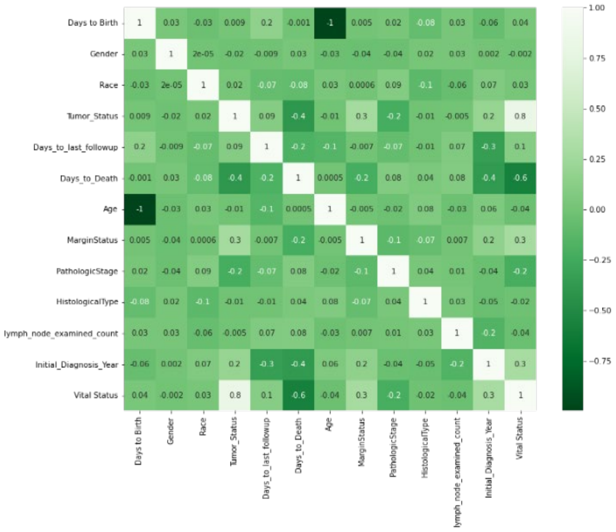


Figure 3. Correlation matrix heatmap for TCGA-BRCA clinical data

A correlation matrix can be used to evaluate the degree of similarity between independent and dependent properties. [26]. The heatmap provides a visual summary of the correlation matrix, making it easier to identify patterns and relationships between features. This heatmap is used to depict an associated feature on the resulting correlation matrix, as shown in Figure 3.

This correlation matrix was used to determine which traits were most closely related: gender, age, pathologic stage, and histological type. Table 3 represents the clinical data after the application of preprocessing methods.

Table 3. TCGA-BRCA clinical data after preprocessing

Dataset Summary	Analysis
Dataset	TCGA-BRCA
#Patients	1058
#Alive	876
#Deceased	182
Median Age	58
Age Range	26-90
Gender (M:F)	12: 1046

Feature scaling is a data preprocessing step that helps ensure all the features are on the same scale. This step is essential to guarantee optimal model performance and avoid any potential bias. Standardization and normalization are two popular techniques used for feature scaling. Standardization changes each feature so that its mean is 0 and standard deviation is 1, while normalization transforms each feature into a range between 0 and 1. For this purpose, standardization replaces the values by their z-scores [27] and is given in Eqn 1:

$$X_{stand} = \frac{\mathrm{X} - \mu(\mathrm{X})}{\sigma\ (\mathrm{X})} \qquad (1)$$

which indicates that the features are redistributed with μ - mean of 0 and σ - standard deviation of 1.

Clinical Data Deep Neural Networks Architecture

Deep Neural Networks (DNNs) are algorithms that are revolutionizing the healthcare industry by providing new techniques to analyse clinical datasets. It is increasingly used in clinical datasets to obtain meaningful insights and patterns from large amounts of data, performing complex, nonlinear computations which can be used to identify trends, correlations, and outliers in the data. DNNs are composed of several neuron layers, allowing the network to learn from data more effectively than traditional machine learning algorithms [28]. It can also be used for predictive analytics, allowing healthcare professionals to anticipate potential issues before they occur.

Initially, the proposed Optimized DNN Multimodal analytics (ODNN-MA) architecture is created using the sequential model. In DNN, data is entered into the input layer, them forwarded to the several hidden layers. Finally, the result is passed to the output layer [29]. The TCGA-BRCA has 12 input parameters. The dataset was then taken for splitting into training and testing sets with a ratio of 70:30, respectively. Subsequently, 10 hidden layers were used with ReLu as the activation function. This study examined the binary classification problem, and hence, the Sigmoid activation function was applied at the output node. Once the layers are developed, NN architecture is constructed to determine the difference of real and expected outputs. Adam is the optimizer and accuracy is the metric used to evaluate model performance. The training data was fit to the model using batch size 32, and the model went through 30 iterations to train across the entire dataset. The regularization is employed to resolve issues of overfitting or underfitting. This method inhibits learning a more sophisticated or flexible model while reducing the risk of overfit [30]. To train the suggested model, two regularizers were used. Figure 4 depicts the ODNN-MA architecture of the proposed clinical work.

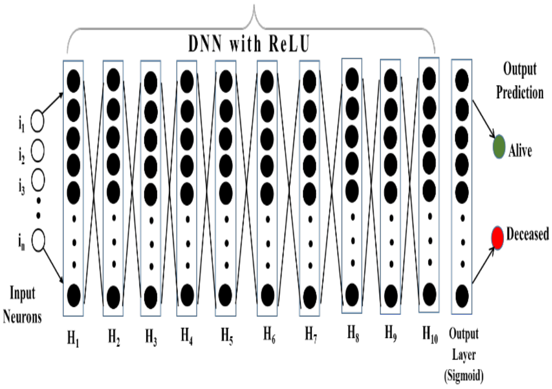
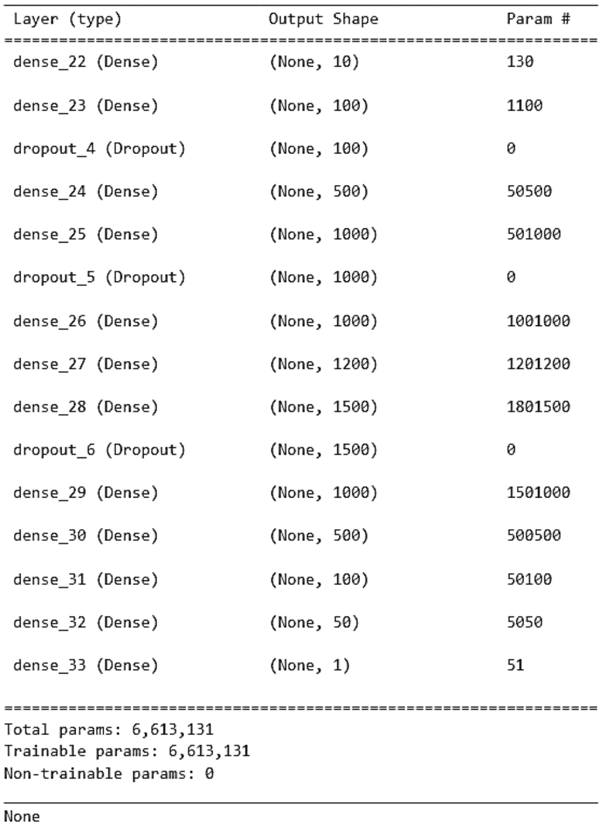


Figure 4. ODNN-MA architecture for TCGA-BRCA clinical data

L2 Parameter Regularization technique improves the accuracy of models by reducing overfitting. It is an optimization strategy that alters the loss function during training by adding a penalty term. Then, model weights are penalized, to ensure they are modest and close to their original values. This reduces the risk of overfitting, as large model weights can result in overly complex models that learn from noise instead of true underlying patterns [31].

The dropout process involves randomly eliminating a subset of neurons from a NN during training to induce the model to generate more accurate representations of the data [32]. The dropout pattern may differ depending on the layers used. Each iteration of the dropout process involves the random deletion of nodes and its connections. Thus, each iteration has its own set of nodes with its own set of outputs. The summary of a sequential DNN model can be seen in Figure 5.



```
Layer (type)                 Output Shape              Param #
=================================================================
dense_22 (Dense)             (None, 10)                130

dense_23 (Dense)             (None, 100)               1100

dropout_4 (Dropout)          (None, 100)               0

dense_24 (Dense)             (None, 500)               50500

dense_25 (Dense)             (None, 1000)              501000

dropout_5 (Dropout)          (None, 1000)              0

dense_26 (Dense)             (None, 1000)              1001000

dense_27 (Dense)             (None, 1200)              1201200

dense_28 (Dense)             (None, 1500)              1801500

dropout_6 (Dropout)          (None, 1500)              0

dense_29 (Dense)             (None, 1000)              1501000

dense_30 (Dense)             (None, 500)               500500

dense_31 (Dense)             (None, 100)               50100

dense_32 (Dense)             (None, 50)                5050

dense_33 (Dense)             (None, 1)                 51

=================================================================
Total params: 6,613,131
Trainable params: 6,613,131
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 5. Summary of the Sequential DNN model

RNA Seq Data

RNA sequencing (RNA-seq) is pivotal in cancer research for helping researchers understand tumor classification and progression by tracking changes in gene expression and the transcriptom. Preprocessing RNA sequencing data is critical to gaining meaningful insights from raw sequencing data. This process involves several steps, such as filtering, normalization, and mapping sequencing reads onto the reference genome or transcriptome [33]. Preprocessing helps to identify relevant gene expression levels, splicing variants, and alternative transcripts. In addition, it facilitates subsequent analysis, including gene set enrichment and differential expression profiling.

In this study, genomic data primarily focus on RNASeq data from TCGA-BRCA. Genomic data are typically supplemented with clinical outcomes, including general clinical information and cancer status [34]. The dataset comprises 60, 660 gene data points across tumor and normal cases. Both normalized fragments per kilobase per million (FPKM) and raw data count were utilized. Raw count data helped select genes that exhibited significant differential expression, while normalized FPKM data were employed in classification and ensemble procedures [35]. Table 4 provides an overview of the number of tumor samples for BRCA subtypes.

Table 4. Number of tumor samples for BRCA subtypes

Subtypes	**Basal Like**	**Her2**	**LumA**	**LumB**	**Normal Like**
Values	192	82	564	207	40

The BRCA data has the following phases:

- Obtain the gene data with its subtypes.
- Split up the data as training and testing sets
- Train data with ODNN-MA architecture with regularization parameters.
- Evaluate the classification result on testing data.

To filter out genes with a mean value below 0.2 and a variance value below 2 across tumor samples, 1,085 genes were selected for BRCA data upon receipt of the tumor data. Subsequently, the tumor samples were divided into five subtypes based on the clinical BRCA data: Basal_like tumor samples, Her_2 tumor samples, Lum_A tumor samples, Lum_B tumor samples, and the Normal_like tumor sample. Table 4 illustrates the specific size of the tumor sample for each subtype. The selected genes were divided into training and testing sets in a 75:25 ratio. Then, ODNN-MA architecture, as depicted in Figure 4, is applied on the entire data.

4. Results and Discussion

This chapter provides an overview of the experimental results of two alternative dataset modalities for TCGA-BRCA using the deep NN system for the prediction of breast cancer.

4.1. Performance Metrics

This work addresses the prediction problem, thus, the performance measures taken are mainly related to classification. For detecting Breast Cancer, the target variable of 1 is considered deceased, and the target variable of 0 is considered a negative instance. This negative instance indicates that the patient is free of the tumor and is still alive.

The confusion matrix evaluates the model's preciseness and completeness and is used for the classification problem, with two or more classes as output. The arrangement of the table or matrix helps to visualize the performance of the algorithm [36].

Constraints of the Confusion matrix are:

- True Positives: When the actual and projected BC instances are true.
- True Negatives: When predicted instances are false, and actual instances are precisely false.
- False Positives: When the actual BC instances are false, but the prediction is true.
- False Negatives: When the prediction is false, but actual cases are true.

The confusion matrix observations represent the classifiers' performance as precision, recall, F1-score, accuracy, and specificity. When the target variable classes in the data are approximately balanced, accuracy will be a good metric [37]. Table 5 provides a quick explanation of the performance metrics.

Table 5. Performance Metrics

Metrics	Description	Formula
Precision	The proportion of correctly identified cases among positive instances.	$\frac{TP}{TP+FP}$
Recall	The proportion of total relevant results that were successfully categorized.	$\frac{TP}{TP+FN}$
F1-score	The weighted average of the precision and recall.	$\frac{2*recall*precision}{recall+precision}$
Accuracy	The fraction of real positive or true negative findings.	$\frac{TP+TN}{TP+TN+FP+FN}$
Specificity	The fraction of real negatives forecasted as negatives	$\frac{TN}{TN+FP}$

The performance of the precision, accuracy, recall, specificity, and F1-score of ODNN-MA on TCGA-BRCA clinical dataset demonstrates the importance of the preprocessing technique according to the nature of its features, as illustrated in Table 6. The confusion matrix for ODNN-MA with the TCGA-BRCA clinical dataset is depicted in Figure 6.

Table 6. Performance of ODNN-MA on TCGA-BRCA clinical data

	Precision	Recall	F1-score	Support
0	0.91	1.00	0.96	287
1	1.00	0.57	0.73	63
Accuracy			0.92	350
MacroAvg	0.96	0.79	0.84	350
WeightedAvg	0.93	0.92	0.91	350

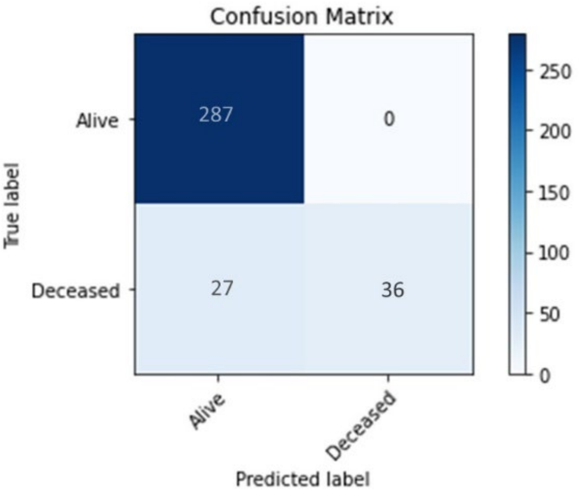


Figure 6. Confusion Matrix for ODNN-MA on TCGA-BRCA clinical dataset

It is evident from the above figure that the confusion matrix is representative of the ODNN-MA testing set classification report for the proposed clinical dataset on TCGA- BRCA. The above matrix indicates that out of the 350 testing cases studied, 287 were examined and predicted as alive cases (True Positive (TP)), 36 were observed and predicted to be deceased (True Negative (TN)), no cases were predicted to be False Negative (FN), and 27 were observed as deceased but projected to be alive (False Positive (FP)).

Table 7 illustrates the performance of RNASeq-based BRCA classification. RNASeq-based BRCA subtypes classification is based on the proposed ODNN-MA architecture. From the table, it is evident that the vascal-like subtype produces a higher accuracy score than the other types.

Table 7. Performance of ODNN-MA on TCGA-BRCA RNASeq subtypes

BRCA subtypes	Sensitivity	Specificity	Accuracy
Basal_Like	0.96	0.95	0.96
Her_2	0.78	0.88	0.88
Lum_A	0.86	0.81	0.86
Lum_B	0.81	0.83	0.84
Normal_Like	0.75	0.85	0.85

The Stratified K-fold Cross-Validation method enhances the classification accuracy of the dataset by partitioning it randomly in equal ratios for each fold. This approach assesses the quality of the classifier output on the Area Under the Receiver Operating Characteristic (ROC) curve. The ROC curve visually represents the diagnostic capacity of a binary classifier, plotting the true positive rate (TPR) against the false positive rate (FPR). Values close to one on the ROC curve indicate superior performance of the machine learning model [38].

The ROC curve illustrates the correlation between the TPR and the FPR as a function of the changing discriminating threshold. AUC (Area Under the ROC Curve) provides an aggregated performance measure across all potential classification thresholds. AUC values range from 0 to 1, with higher values indicating better classification performance [39].

Figures 7 and 8 illustrate the ROC and AUC of the proposed ODNN-MA applied to the clinical dataset of TCGA- BRCA, illustrating the comparison of TPR against FPR.

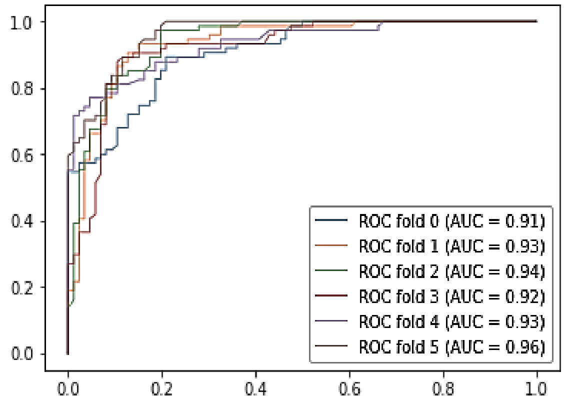


Figure 7. Receiver Operating Characteristics (ROC) Curve for ODNN-MA

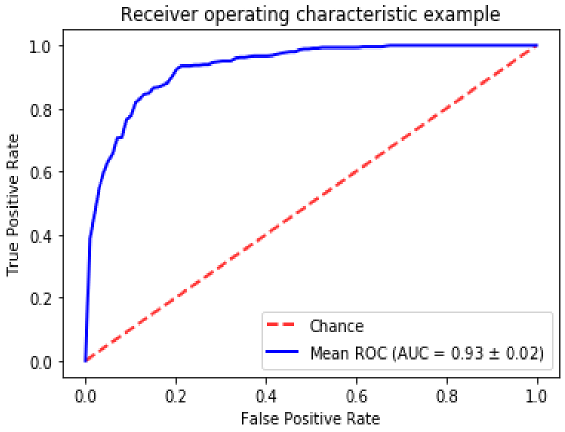


Figure 8. AUC for ODNN-MA

The above curve summarizes the results of the proposed ODNN-MA based on clinical data for TCGA-BRCA, which demonstrate a statistically significant AUC value of 0,93 ± 0,02. This indicates that the model can accurately distinguish between live and deceased cases with a probability of 93%.

5. Conclusion

Breast cancer is one of the most prevalent cancers among women, accounting for 69% of cancer-related deaths in this demographic. Early detection of breast cancer is crucial as it remains a significant health challenge today. Detecting it early can substantially improve survival rates by enabling timely treatment. Though traditional deep learning models used within detection excel with specific data types, multimodal deep learning models are even more effective due to their ability to integrate richer, more comprehensive data from multiple sources compared to traditional unimodal models. In this study, the performance model achieved 92% accuracy on clinical data, an AUC ROC score of 0.93 ± 0.0, and 96% accuracy on RNASeq data. This capability enables the model to leverage diverse data sources, offering significant advantages over unimodal approaches. The evidence of this study suggests that integration of these models would be certain to improve the estimation of the likelihood of patients' risk of developing breast cancer in future.

References

[1] World health organization cancer. (2018). Fact Sheet-Cancer. Available at: https://www.who.int/health-topics/cancer

[2] https://www.livemint.com/news/india/icmr-data-shows-unequal-toll-of-cancer-on-women-11670349329355.html

[3] https://www.industryarc.com/PressRelease/2625/Oncology-Market-Research.html

[4] Mertz, S., Mayer, M., Paonessa, D., Papadopoulos, E., Alessandro, F., Peccatori, K. S., ... & Spence, D. (2016).

Breast Cancer Center Survey: Cancer center management, support, and perception of mBC patient needs across 582 healthcare professionals

[5] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32381-3/fulltext

[6] Bini, S. A. (2018). Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?. The Journal of arthroplasty, 33(8), 2358-2361.

[7] Karthik, S., Perumal, R. S., & Mouli, P. C. (2018). Breast cancer classification using deep neural networks. In Knowledge computing and its applications (pp. 227-241). Springer, Singapore.

[8] Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. Briefings in Bioinformatics, 23(2), bbab569. https://doi.org/10.1093/bib/bbab569

[9] https://jina.ai/news/what-is-multimodal-deep-learning-and-what-are-the-applications/

[10] Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. nature, 415(6871), 530-536.

[11] Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., ... & Marti, R. (2017). Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics, 22(4), 1218-1226.

[12] Al-Antari, M. A., Al-Masni, M. A., Choi, M. T., Han, S. M., & Kim, T. S. (2018). A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. International journal of medical informatics, 117, 44-54.

[13] Sun, D., Li, A., Tang, B., & Wang, M. (2018). Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Computer methods and programs in biomedicine, 161, 45-53.

[14] Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., & Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics, 22(14), e184-e190.

[15] Sun, D., Wang, M., & Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM transactions on computational biology and bioinformatics, 16(3), 841-850.

[16] Khademi, M., & Nedialkov, N. S. (2015, December). Probabilistic graphical models and deep belief networks for prognosis of breast cancer. In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) (pp. 727-732). IEEE.

[17] Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., ... & Shung, K. K. (2021). Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nature biomedical engineering, 5(6), 522-532.

[18] Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., ... & Klauschen, F. (2021). Morphological and molecular breast cancer profiling through explainable machine learning. Nature Machine Intelligence, 3(4), 355-366.

[19] Liu, T., Huang, J., Liao, T., Pu, R., Liu, S., & Peng, Y. (2022). A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. Irbm, 43(1), 62-74.

[20] Arya, N., & Saha, S. (2021). Multimodal advanced deep learning architectures for breast cancer survival prediction. Knowledge-Based Systems, 221, 106965.

[21] Lingle, W., Erickson, B. J., Zuley, M. L., Jarosz, R., Bonaccio, E., Filippini, J., Net, J. M., Levi, L., Morris, E. A., Figler, G. G., Elnajjar, P., Kirk, S., Lee, Y., Giger, M., & Gruszauskas, N. (2016). The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA) (Version 3) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP

[22] Kanimozhi, G., & Shanmugavadivu, P. (2021). Optimized DEEP neural networks architecture model for breast cancer diagnosis. cancer, 3, 4.

[23] Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, 41(12), 3692-3705.

[24] Jerez JM, Molina I, García-LaencinaPJ, Alba E, Ribelles N, Martín M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine. 2010 Oct 1;50(2):105-15.

[25] Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing data preprocessing in credit classification: One-hot encoding or imputation?. Emerging Markets Finance and Trade, 58(2), 472-482.

[26] Kanimozhi, G., Shanmugavadivu, P., & Rani, M. M. S. (2020). Machine Learning-Based Recommender System for Breast Cancer Prognosis. Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries, 121-140.

[27] Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. Mach Learn Tech Rep, 1(1), 1-6.

[28] Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., & Salem, A. B. M. (2018). Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal, 3(1), 68-71.

[29] Ganesan, K., Pichai, S., Kavitha, M. S., & Takahashi, M. (2022). Data imputation in deep neural network to enhance breast cancer detection. International Journal of Imaging Systems and Technology, 32(6), 2094-2106.

[30] Van Laarhoven, T. (2017). L2 regularization versus batch and weight normalization. arXiv preprint arXiv:1706.05350.

[31] Zhang, G., Wang, C., Xu, B., & Grosse, R. (2018). Three mechanisms of weight decay regularization. arXiv preprint arXiv:1810.12281.

[32] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[33] Xiao, Yawen; Wu, Jun; Lin, Zongli; Zhao, Xiaodong (2017). A Deep Learning-based Multi-model Ensemble Method for Cancer Prediction. Computer Methods and Programs in Biomedicine, 153, 1-9

[34] https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723

[35] Yu, Z., Wang, Z., Yu, X., & Zhang, Z. (2020). RNA-seq-based breast cancer subtypes classification using machine learning approaches. Computational intelligence and neuroscience, 2020.

[36] Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion Matrix-based Feature Selection. MAICS, 710, 120-127

[37] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

[38] https://www.section.io/engineering-education/how-to-implement-k-fold-cross-validation/

[39] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition. 1997 Jul 1;30(7):1145-59.