Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Smart Medicine
DOI: 10.54254/2753-8818/2026.HZ36311

Breast Cancer Risk Prediction and IDC Tumor Classification in Different Stages Based on Machine Learning Models

Yunze Li[1*], Nor Samsiah Sani[2]

[1]*Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia*
[2]*Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia*
**Corresponding Author. Email: P132306@siswa.ukm.edu.my*

Abstract. Based on the medical statistical analyses of relevant information, in recent years, breast cancer is one of the leading causes of cancer-related deaths in women worldwide. Hence, the early detection and accurate diagnosis of breast cancer patients is significant to improve their prognosis. In this study, we propose a multi-modal, data-driven intelligent framework especially designed to breast cancer analysis the Multi-Dimensional Feature Refinement Cancer Network (MDFR-Cancer-Net) that can be used to analyse breast cancer at various clinical stages based on multi-modal data.The framework of this study will be to combine tabular clinical data for risk stratification, pathological imaging data of tumors and mi-RNA molecular data for biomarker-assisted analysis at different times in the study. MDFR-Cancer-Net will first be used to reduce the number of features and improve the representation ability of features. A few are Naive Bayes and Deep U-Net; the rest are other types of models for classification and segmentation. Five-fold cross-validation and ablation tests were carried out to obtain the above results. Based on the above experimental results, the accuracy of the Naive Bayes model on the Wisconsin Breast Cancer dataset was 99%, and that of the baseline SVM was 92%. Deep U-Net reached an accuracy of 96% for the 277,524 IDC pathological image patches and exceeded the baseline CNN's accuracy of 87%. Naive Bayes had 100% accuracy in the mi-RNA molecular analysis and outperformed the baseline Random Forest (95%). In short, the framework presented here shows that all kinds of data can be used together to help diagnose breast cancer.

Keywords: MDFR-Cancer-Net, Transformer, cross-attention, breast cancer tumors, multi-modal data

1. Introduction

Breast cancer is the leading cancer among women worldwide and is often referred to as the "pink killer." According to the International Agency for Research on Cancer (IARC), the age-standardized incidence rate among women is 46.3 cases per 100,000, while the incidence in men is only 0.5 cases. Approximately 2.3 million new cases are diagnosed annually, and the incidence continues to rise globally [1]. In China, breast cancer incidence has more than tripled during the past three

DOI: 10.54254/2753-8818/2026.HZ36311

decades, while mortality has more than doubled. The prognosis of breast cancer patients is influenced by multiple factors, including age, pathological type, tumor stage, metastasis, and treatment timeliness . Besides threatening survival, breast cancer also causes psychological stress, body-image changes, and long-term treatment burden [2].

Cancer cells at this early stage are usually hidden in healthy breast tissue and can be hard to find unless they are noticed in a change in the breast tissue, or by a lump that is felt. Because malignant cells in the breast are intrinsically hard to detect in the early stages of disease, early screening and accurate diagnosis, are essential for increasing survival rates and enhancing treatment effectiveness [3]. While targeted therapy and immunotherapy have greatly enhanced patient outcomes in recent years, significant individual variation in clinical outcomes, lack of data to address, and the lack of integration between studies remain as challenges to accurate prediction of prognosis [4]. Thus, this research is targeting to combine clinical risk factors, imaging biomarkers and molecular information to facilitate precision medicine and personalized treatment strategies [5]. The specific research interests are in the field of multi-modal data analysis and intelligent diagnostic assistance technologies in breast cancer. The background section summarizes the current breast cancer data-sets and the developments in machine learning models and points out the problems in model applications in the existing research. The methods section explains data sources, complementary relationships between multi-modal data-sets, and the MDFR (Multi-modal Data Feature Reduction) data dimensionality reduction method, as well as the principles behind implementing cross-attention mechanisms in the Transformer framework and the feature extraction workflow of MDFR-Cancer-Net. The model is tested and evaluated on three different data-sets, i.e., data-set 1 is used for parameter optimization, data-set 3 is used for cross-validation, and data-set 2 is used for ablation tests to validate the effectiveness of each component of the model. The experimental results are analyzed from various aspects such as data processing, model fitting and hyper-parameter optimization, and the value, limitations and future directions of the system are further discussed from the perspectives of model innovation, artificial intelligence research and clinical diagnostic assistance. Finally, a complete overview of the results of the research is given.

2. Research background

At present, the traditional methods for medical diagnosis are mainly based on the pathological diagnosis, biochemical marker detection, mammography and the doctor's experience in determining imaging results. But, these methods are highly subjective, lack incorporation of related information, and rely on the experience of the radiologists with likely misdiagnosis, delayed diagnosis, and inconsistent diagnosis results. In large scale screening programs, the workload and anatomical complexity of breast tissue is a major challenge for subjective breast image interpretation, which also decreases the detection sensitivity and the diagnostic efficiency [3]. Furthermore, traditional approaches can be insufficient for capturing tumor heterogeneity, micro-environmental variations, and the progression of molecular changes in the context of multiple imaging modalities and challenging data integration tasks. As a result, the diagnostic and risk-stratification results are not as specific as needed for early risk-stratification and prognosis.

2.1. Literature review

The analysis of huge breast cancer data-sets is showing new possibilities in the field of artificial intelligence (AI) models. AI-based models can be used to extract features from high dimensional multi-modal data, including combining relevant information from clinical, imaging, and molecular

DOI: 10.54254/2753-8818/2026.HZ36311

data-sets; which can be used to accurately localize regions of cellular carcinogenesis and improve screening, discovery of biomarkers, prediction of treatment response and personalized treatment [4]. AI minimizes operator dependence, increases the repeatability of the results, raises the sensitivity and specificity of the results and automates the diagnostic processes by combining complex information together, compared to traditional methods [5]. Additionally, multi-modal AI systems enable the synergistic processing of diverse data sources, such as clinical data, pathological images, and molecular expression patterns, which can be leveraged to train AI models that consider all these factors, potentially providing a more holistic approach to developing targeted treatment strategies. As such, the role of AI in precision medicine and clinical decision support has come to the forefront.

Various research analyses have shown the remarkable performance of AI models in single domain prediction and classification of breast cancer data-sets. The Assigie team used the KNN model in Wisconsin Breast Cancer data-set which resulted in 94.15% of classification accuracy in the case of tumored breast and high precision in risk assessment [6]. Haque et al. created a supervised learning model which used biomarkers including ER, PR and HER2, and it was able to predict survival with a rate of 94.64% [7]. The work by Zhang's team further showed that machine learning-based bioinformatics systems are capable of accurately discovering immune prognostic markers for overall survival (OS) in colorectal cancer (CRC) patients, indicating the incorporation of survival prediction mechanism with molecule level feature analysis is of significance [8]. These findings give important references of using prognostic machine learning systems to evaluate breast cancer survival and predict breast cancer personalized treatment strategy.

In image analysis, deep learning neural networks are used in various ways, such as for tumor localization analysis and for tumor characterization. The study results show that the methods based on convolutional neural network (CNN) give a better classification accuracy of pathological images of invasive ductal carcinoma (IDC) with increase from 78% to 87% [9]. Abdelrahman et al. made a systematic review of CNN applications in mammography, where the Deep CNN architecture plays a major role, and the capability of lesion detection, feature extraction and automatic classification of breast cancer in the mammographic image, thus, deep neural networks are important [10]. Furthermore, classical CNN architectures like VGG16, Res-Net, Alex-Net and ZF-Net are known to have a steady performance of nearly 90% accuracy in particular cellular tumor identification even after optimization [11]. There are however still some difficulties for existing CNN frameworks, including fine-grained extraction of tumor borders and modelling the global context, especially for high density breast tissues or noisy imaging conditions. To overcome these problems, this work proposes the MDFR mechanism based on the Transformer with the cross-attention module and a contrastive-learning strategy to enhance the interaction of global features and make pathological image interpretation more precise.

Microcosmic research of breast cancer has become a major research direction – mi-RNA molecular analysis. The RNN based analytical framework and circulating tumor DNA (ctDNA) analysis methods have been widely used in genetic testing and prognosis assessment at the level of sequence data studies [12]. Dubey et al. used statistical methods to analyse the molecular changes associated with ER, PR, and HER2 [13] and Zhou et al. developed a glycosylation-related prognostic prediction model for triple-negative breast cancer using data from The Cancer Genome Atlas (TCGA) and identified key genes associated with patient prognosis [14], while Zeng's team further proposed a recurrence prediction framework based on LSTM with an overall gene recurrence prediction accuracy rate of 89% and an AUC value of 0.98 ± 0.01 [15]. All these studies suggest that combining molecular markers with AI algorithms can dramatically improve the precision of the prognosis and the prediction of recurrence risk.

DOI: 10.54254/2753-8818/2026.HZ36311

2.2. Breast cancer research gaps

The review of the literature indicates that while the AI models have witnessed remarkable progress in processing a single set of data and show significant improvement in practical applications, there are still a number of research gaps in terms of comprehensive information integration. Firstly, there are enormous difficulties in the unified analysis of multi-modal data. Breast cancer studies are conducted in various stages, and these studies employ different data-sets such as clinical tables, pathological images and mi-RNA molecular data which have significant structural differences. Traditional approaches are not capable of carrying out an integrated analysis in a single framework and therefore information integration is ineffective. Second, most of the previous research has centered on the association of individual mi-RNAs with breast cancer, and have not included comprehensive modeling of multiple features of mi-RNAs, resulting in less stable and reliable predictive outcomes. Furthermore, the pathological image and mi-RNA data-sets are high dimensional, sparse and have noise, which makes processing them more complex and over-fitting problems unsolved. Finally, traditional breast cancer prediction models are not accurate enough and have poor generalization performance. Conventional diagnostic methods are significantly dependent on the skill of the physician or on one single statistical model, which can be subject to bias from subjective interpretation, distribution of data and sample heterogeneity. Specifically, the breast tumor images are frequently irregular in shape, indistinct in edges, low in contrast and noisy, which is difficult for traditional CNN models to effectively capture image details and global features.

To overcome these limitations, this research presents a model that is tailored for multi-modal data integration—MDFR-Cancer-Net. Combining multi-modal information from tabular clinical data, pathological imaging data and mi-RNA molecular data, this model can be used at every stage of breast cancer screening and diagnosis. Firstly, the model uses Multi-domain Feature Optimization (MDFR) method to reduce the dimensions and optimize features. In the model design phase, Transformer-based cross-attention mechanism and contrastive learning strategy are adopted to further improve the global context awareness and fine-grained feature extraction. Machine learning and deep learning models are then used to predict the risk, classify the tumor type and make prognosis assessment. Lastly, an information layer is added, allowing for multi-modal collaborative modeling and so overcomes the limitations of single-data-source approaches, greatly increasing prediction accuracy, model robustness and clinical interpretability. MDFR-Cancer-Net's effectiveness in breast cancer diagnosis and personalized treatment support is definitively confirmed by systematic validation using cross validation and ablation experiments.

3. Methodology

The main deep learning framework used for the development and testing of models in this study is PyTorch, and the programming language is Python. The RAM of the experimental platform is 32GB and it has an NVIDIA GeForce RTX 4090 GPU. Based on the above, an AI-based breast cancer prediction analysis framework has been developed to jointly model histopathological and clinical data at different stages of the disease, as shown in Figure 1.

DOI: 10.54254/2753-8818/2026.HZ36311

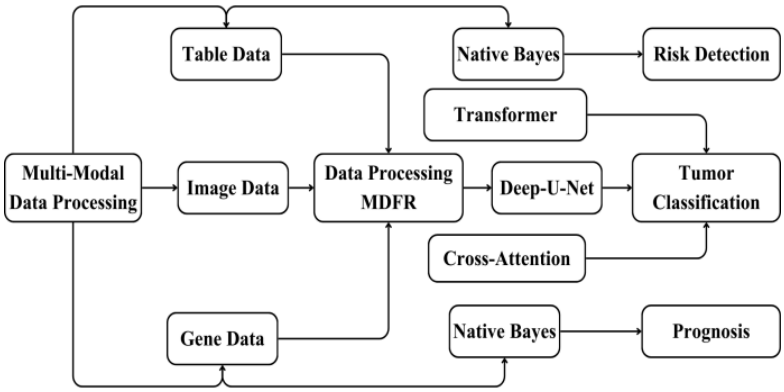

Figure 1. Workflow of the proposed research method

According to different types of data, the multi-modal inputs are organized into three categories—tabular, histopathology images, and molecular (mi-RNA) profiles—and each modality is aligned with the corresponding clinical stage in the screening-to-treatment continuum, as illustrated in Figure 2.

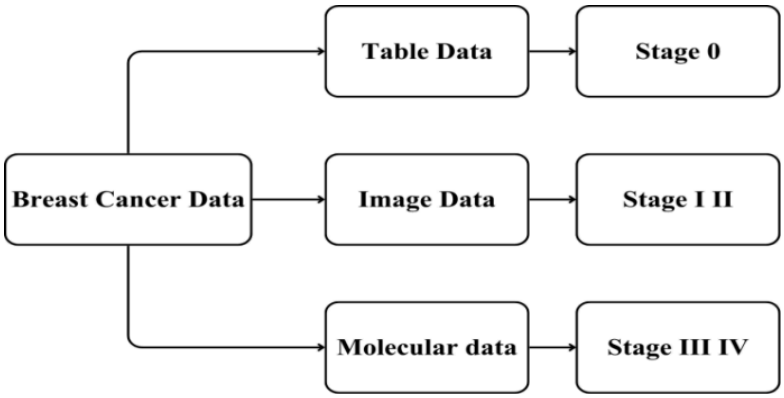

Figure 2. Distribution of data-sets across clinical stages

3.1. Data-set

This research utilizes three publicly available data-sets, with task definitions based on clinical workflow steps. During the specific data processing procedure, the feature information expressed by breast cancer-related multi-modal data is projected into a shared latent space using the MDFR method. This approach extends the information across modalities, effectively mitigating issues such as multi-modal data imbalance. Detailed data-set information is provided in Table 1.

Table 1. Details of the data-sets

data-set	Data type	data-set size	Features	Classification
1	Tabular data	569	30	Binary
2	Image data	277,524	50*50	Binary
3	mi-RNA data	1928	133	Binary

Dataset 1: Tabular Data for Detection and Analysis of Breast Cancer Risk Stratification. The above table-data is from the Wisconsin Breast Cancer dataset and can be obtained publicly on

DOI: 10.54254/2753-8818/2026.HZ36311

Kaggle. The total number of samples in the dataset is 569; there are 357 benign samples (62.7%) and 212 malignant samples (37.3%). The 30 items in the table are biomolecular data, genes and other daily life factors. There are no missing values in the dataset and no category imbalance. Principal Component Analysis (PCA) was employed at the start of the data analysis to reduce the dimensionality of the 30-dimensional feature set to 16, and then tumor risk was calculated for the purpose of predicting disease risk.

Dataset 2: The main image data for tumor size recognition and classification/detection of lesions (segmentation) are taken from the IDC pathological image dataset. The dataset is from the public Kaggle dataset, and it contains 162 whole-slide images of breast cancer (BCa) specimens from all types of patients, all scanned at a magnification of 40x. A total of 277,524 blocks of 50x50 pixel RGB image samples (non-continuous slices) were randomly selected, and among them, 198,738 samples (71.6%) were determined to be malignant and 78,786 samples (28.4%) were benign. It is a set of blocks, and each block contains several images for comparison. MDFR was used to reduce the dimensionality of the input image blocks in the analysis of the feature model to reduce the dimensions, and then the designed deep U-Net model was used to extract and output IDC classification results (malignant/benign).

Dataset 3: miR-143 Molecular Data for Molecular Marker-Assisted Prognosis. The dataset is the public mi-RNA breast cancer dataset from Kaggle, and in total across all populations, it has 1,928 samples and 133 features. The above are the attributes obtained from the expression profile data of the 11 miRNAs, including miR-21 and miR-155. The number of benign and malignant samples in this category of the dataset is about the same, so no further sample enrichment needs to be conducted in the preprocessing stage to ensure model stability during training, and the sub-type classification results will be more in line with clinical reality.

3.2. Data processing

Data preprocessing is a process for EDA analysis and exploration of data features. In this research, modality-specific preprocessing was performed to standardize inputs and reduce spurious variation. For the tabular data-set (data-set 1), completeness (no missing values) was verified and stratified splitting was applied; when class imbalance was present within a training fold, oversampling was used only on the training portion to avoid information leakage. For the histopathology patches (data-set 2), images were resized to a consistent resolution and intensity-normalized, and on-the-fly augmentation (random flips/rotations, scaling, translation, and color jitter in brightness/contrast/saturation) was applied to improve invariance to staining and acquisition differences. For the mi-RNA profiles (data-set 3), expression values were log-transformed and scaled; sparse and low-variance features were filtered, and class-weighted training (focal loss) was used when needed. The overall preprocessing workflow is summarized in Figure 3.

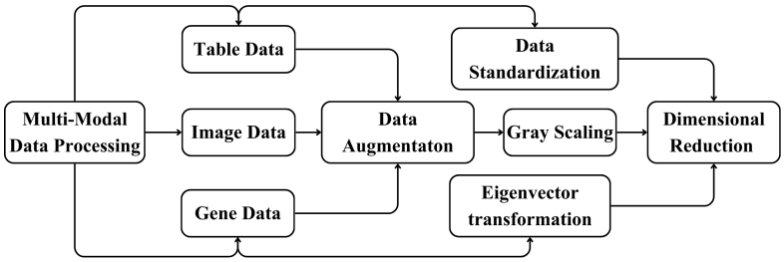

Figure 3. Data preprocessing workflow

DOI: 10.54254/2753-8818/2026.HZ36311

Comprehensive analysis of both the macro and micro aspects of the dataset is carried out to find the latency period and location of breast cancer. The above preprocessing will improve the uniformity of the analysis of risk factors, tissue morphology and molecular markers. The distribution of class 1 in the dataset is shown in Figure 4.

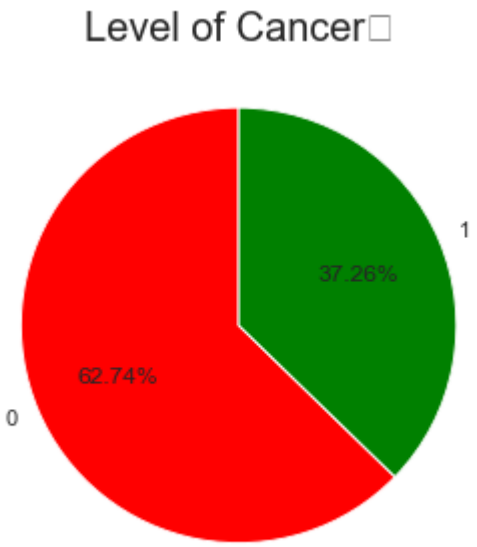

Figure 4. Class distribution of the breast cancer data-set

As shown in Figure 4, malignant cases account for 37.25% of the samples in the tabular data-set. According to the correlation matrix analysis, the proportional weights of the microscopic mi-RNA factors that have the greatest impact on breast cancer tumors can be obtained, as shown in Figure 5.

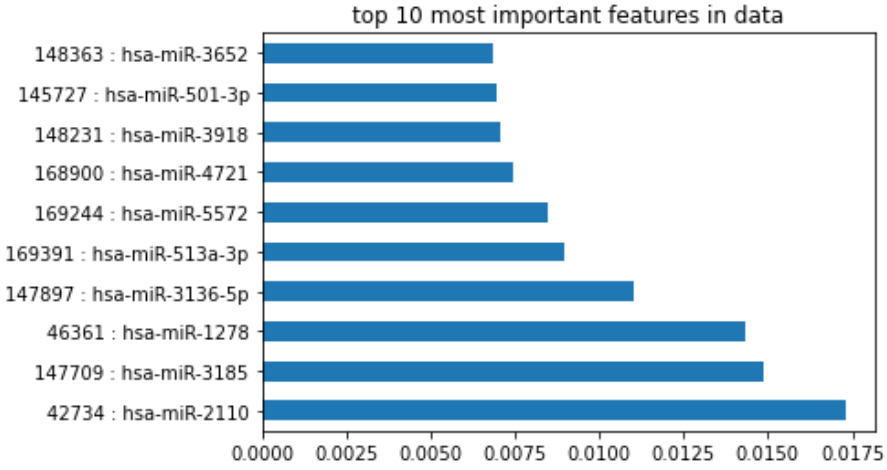

Figure 5. Top 10 mi-RNAs associated with breast cancer

Five-fold cross-validation is employed to divide the dataset into five equal parts for cross-validation in the training process of this study, and thus the clarity of the final experimental results has been improved. An 80/20 Train-Test Split was Used. I used 5-fold cross-validation on the training set to select a model and perform all preprocessing (scaling and resampling) within each training fold to avoid data leakage.

DOI: 10.54254/2753-8818/2026.HZ36311

3.3. MDFR

The dimensionality reduction method MDFR (multi-modal Dimensionality Reduction Framework) proposed in this research is designed to address the high-dimensional characteristics of multi-modal breast cancer data (clinical data in tabular format + imaging data + genetic molecular data). Its core objective is to effectively prevent over-fitting through stepwise, modality-adaptive dimensionality reduction. The following is a concise and logically structured summary of the overall approach, highlighting the purpose, methodology, and over-fitting prevention mechanisms of each step. The specific steps are illustrated in Figure 6:

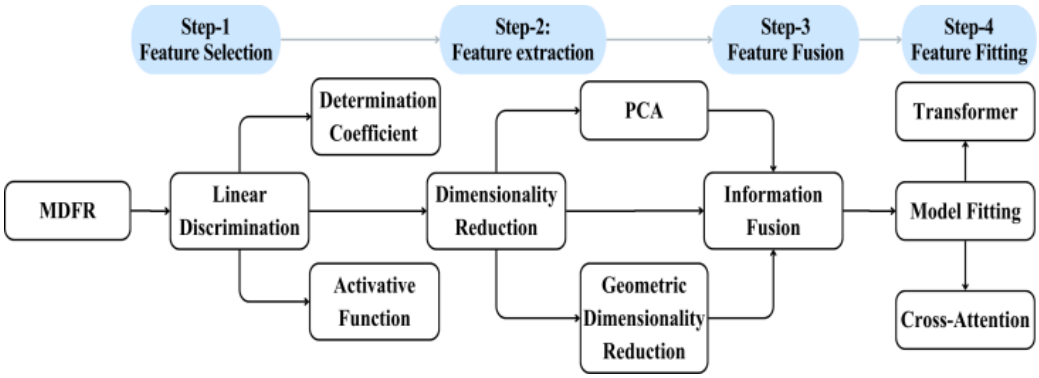

Figure 6. Steps in the MDFR framework

Step 1: Selection of Modal Features and Preliminary Dimensionality Reduction (Reduce noise and dimensionality in the early stage of disaster risk mitigation). Linear Discriminant Analysis is applied to structured/serialised data (such as tabular data and genomic molecular data), and based on the goodness-of-fit criterion R^2 and a threshold of 0.9, it is determined whether there is a significant linear relationship between the independent variables (features) and the dependent variable (label), such as breast cancer subtypes/prognosis. Activation Functions (ReLU/Softmax) are employed to introduce non-linearity in the features obtained after processing the pixels of an image.

Step 2: The system performs adaptive feature extraction and dimensionality reduction based on correlations among data features. In this research, data-sets 1 and 3 exhibit a linear relationship, so principal component analysis (PCA) was directly employed for dimensionality reduction; whereas for data-set 2 (which demonstrates a nonlinear relationship), geometric dimensionality reduction techniques were applied—including horizontal pixel compression (reducing the number of columns along the width/column direction) and vertical pixel compression (halving the number of rows through maximum pooling and stepwise convolution) to compress the feature maps along the height-row direction.

Step 3: multi-modal feature fusion. After completing cross-modal feature extraction and preliminary dimensionality reduction, the selected features (low-dimensional vectors from tables, compressed gene representations, and image-based dimension-reduced embeddings) achieve efficient information integration.

Step 4: Dimension control and regularization at the model level. During model training, by designing a self-learning process model that integrates the Transformer + cross-attention mechanism with a learnable modal importance adaptation mechanism, dimensionality reduction is transformed from a passive constraint into an active optimization process.

DOI: 10.54254/2753-8818/2026.HZ36311

3.4. Model

Tabular data and mi-RNA molecular data were analyzed using various models such as logistic regression, decision tree, random forest, support vector machine and Naive Bayes. Of these, the Naive Bayes classifier which is based on prior probability calculation showed the best overall performance. This model in data fitting utilizes conditional likelihood along with class prior probabilities and uses the P(y) function for classification, so the prior probabilities can be used to account for the distribution of each class, and to minimize prediction biases due to class imbalance during classification. As the basic learning model, a deep U-Net model with deep learning neural networks was used for the classification task of medical imaging tumor data. A Transform encoder-decoder was used for the image feature learning process, with the addition of contrastive learning and attention mechanisms to focus on the feature extraction. In particular, the encoder is designed to learn multi-scale features by progressive down sampling and the decoder is designed to fuse the multi-level features (e.g., using transposed convolution) and then predict the value of each pixel through skip connections and up sampling.Encoder-decoder structure with gradual downsampling and upsampling is employed in the feature learning stage of the model to obtain tumour probability maps and class labels (benign or malignant). Binary cross-entropy loss is employed to improve the accuracy of tumour classification in the model, Adam and RMSprop optimisers are used to tune the parameters, and a learning rate of 0.0001 - 0.001 is set for effective optimisation of parameters.

4. Research results

The experimental procedures were conducted as follows: For data-sets 1 and 3, confidence intervals of 95%, 97.5%, and 99% were set to ensure the accuracy of Bayesian model outputs, with cross-validation using the five-fold method for comprehensive evaluation. data-set 2 was analyzed through varied hyper-parameter configurations, including learning rate settings and early stopping iterations. A comprehensive ablation experiment was performed to examine the impact of MDFR dimensionality reduction analysis on the research, followed by data visualization using the optimal solution.

4.1. Results for tabular data

The research findings on the parameter settings and confidence intervals for data-set 1 are presented in Table 2.

Table 2. Parameter settings and results for data-set 1

data-set	Confidence	var_smoothing	AUC	Accuracy
1	95%	1e-9	0.997 ± 0.03	99.0%
1	95%	1e-8	0.994 ± 0.03	98.7%
1	95%	1e-7	0.991 ± 0.03	98.4%
1	97.5%	1e-9	0.989 ± 0.01	98.0%
1	97.5%	1e-8	0.987 ± 0.01	97.8%
1	97.5%	1e-7	0.988 ± 0.01	97.9%
1	99%	1e-9	0.995 ± 0.01	98.9%

DOI: 10.54254/2753-8818/2026.HZ36311

Table 2. (continued)

1	99%	1e-8	0.996 ± 0.01	99.0%
1	99%	1e-7	0.994 ± 0.01	98.9%

In dataset 1, var_smoothing=1e−9 refined feature distributions, enabling sensitive Naive Bayes discrimination with 100% AUC and accuracy at 95% confidence.

Table 3. Five-fold cross-validation results for data-set 1

Model/accuracy	1	2	3	4	5	Mean
SVM	92.0%	91.5%	91.8%	92.3%	91.6%	91.9%
LR	94.0%	94.3%	93.8%	93.5%	93.6%	93.8%
DT	98.0%	97.5%	97.8%	96.3%	96.6%	97.6%
RF	97.5%	96.4%	96.8%	97.3%	96.6%	96.9%
NB	99%	98.5%	98.8%	98.3%	98.6%	98.7%

Overall, the model's performance remains stable between 92% and 100% across all parameter combinations, indicating that this naive Bayes framework exhibits robustness and reliability under varying data scales and levels of uncertainty. The results of the naive Bayes model are shown in Figure 7.

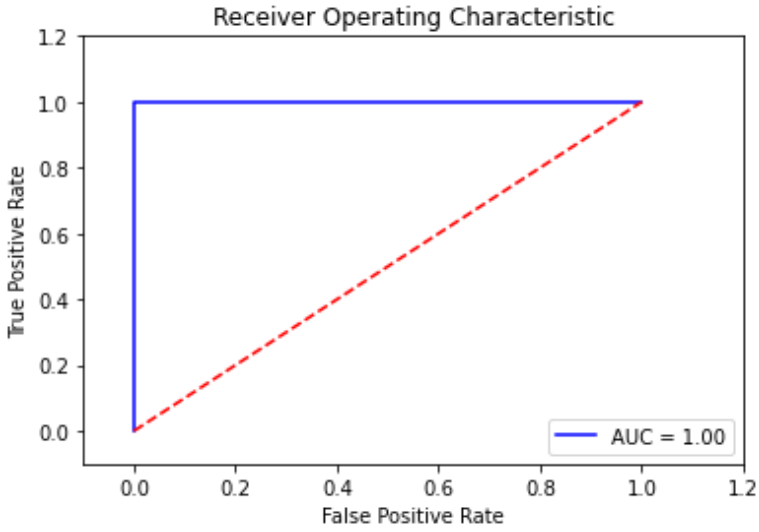

Figure 7. ROC curve of the naive Bayes model

Encoder-decoder structure with gradual downsampling and upsampling is employed in the feature learning stage of the model to obtain tumour probability maps and class labels (benign or malignant). Binary cross-entropy loss is employed to improve the accuracy of tumour classification in the model, Adam and RMSprop optimisers are used to tune the parameters, and a learning rate of 0.0001 - 0.001 is set for effective optimisation of parameters.

DOI: 10.54254/2753-8818/2026.HZ36311

4.2. Results for image data

In the deep U-Net network model for breast cancer tumor IDC patch classification (dataset 2), the 256x256 image pixels are divided into 50x50 pitches for feature learning extraction, each batch size is set to 16 images, and the total number of epochs is 50. Normalisation has been applied to the encoder and decoder blocks to stabilise the initialisation and improve the optimisation process, and early stopping is used to find hyperparameters that reduce the validation loss. The particular list of hyperparameters and the accuracy of the final results are shown in Table 4.

Table 4. hyper-parameter settings, accuracy, and loss of the deep U-Net

data-set	Learning rate	Patience	Loss	Accuracy
2	0.0001	5	0.35	97%
2	0.0001	10	0.42	95.8%
2	0.0001	20	0.50	94.7%
2	0.0005	5	0.40	96.2%
2	0.0005	10	0.47	95.2%
2	0.0005	20	0.51	94.7%
2	0.00001	5	0.38	96.5%
2	0.00001	10	0.40	96.2%
2	0.00001	20	0.42	95.8%
2	0.00005	5	0.53	94.3%
2	0.00005	10	0.45	95.4%
2	0.00005	20	0.48	95.0%

Table 4 reports the hyper-parameter search over learning rate and early-stopping patience. The best configuration achieved 97% validation accuracy with a loss of 0.35. The best results of the deep U-Net model are shown in Figure 8.

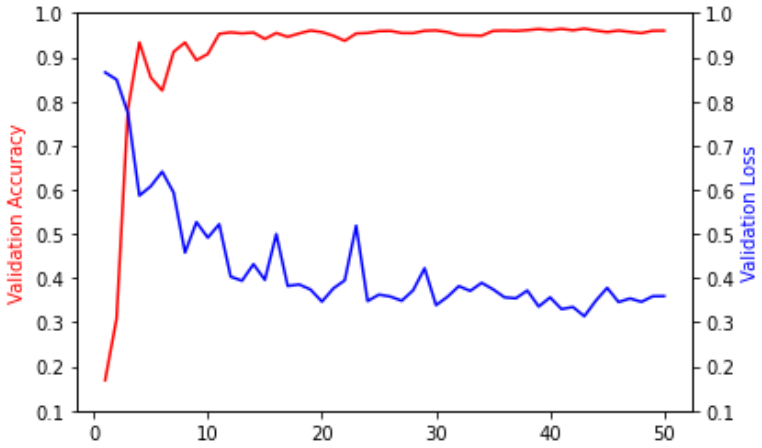

Figure 8. Accuracy and loss curves of the deep U-Net

DOI: 10.54254/2753-8818/2026.HZ36311

The value of the class label in the prediction result of the test set and the percentage value of the test set are shown in the confusion matrices in Figures 9 and 10.

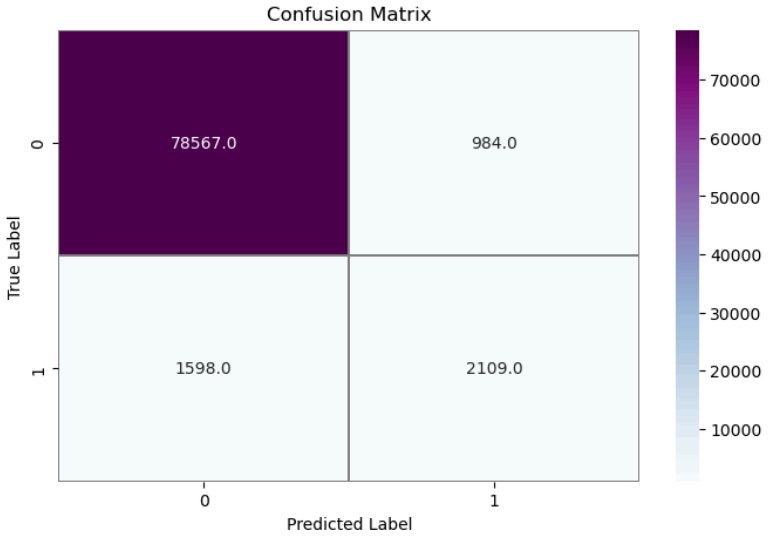

Figure 9. Confusion matrix of predicted classes on the test set

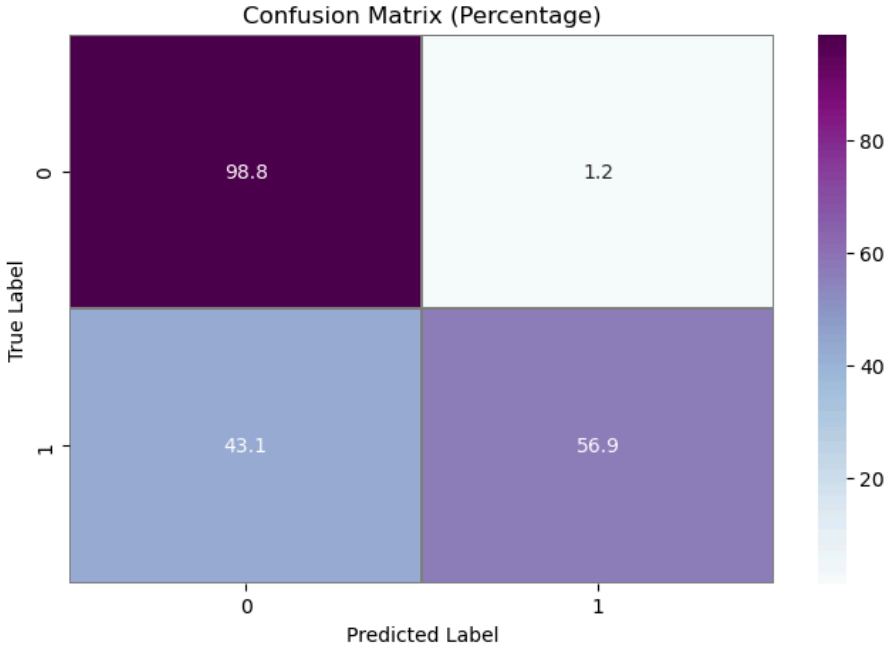

Figure 10. Percentage confusion matrix of predicted classes on the test set

Quantify the contribution of each part and then perform ablation studies. Dataset 2: The combination of MDFR, cross-attention and contrastive learning increased the accuracy of the deep U-Net to 96.5% and reduced the loss to 0.35; the baseline CNN had an accuracy of 87% and a loss of 1.5. Based on the above results, it was decided to combine MDFR, cross-attention and contrastive learning to improve the discriminative ability and stability of cross-modal learning in a model. The details are shown in Table 5.

DOI: 10.54254/2753-8818/2026.HZ36311

Table 5. Ablation results for the cross-attention and contrastive learning mechanisms

Task	data-set	Method	Accuracy	Loss	Improvement
Tumor recognition	2	cross-attention	93%	0.78	3%
Tumor recognition	2	Contrastive learning	95%	0.48	1%
Tumor recognition	2	MDFR	94%	0.65	2%

The table results show that the various necessary and sufficient factors affecting the performance of the model in the tumor recognition task of dataset 2. Without cross-attention, the accuracy of the model and its loss were 0.93 and 0.78, respectively. Add the above mechanism, and then the performance will increase by about 3%; thus, cross-feature attention modelling can improve the correlation modelling of tumor regions and contextual information. Without contrastive learning, the accuracy of the model was only 95 per cent higher than that of the baseline model without cross-attention, and the loss value was about 0.48. Although the total rise was only 2%, the drop in loss shows that the distinguishability of features and class separation have improved to some extent. The model that only uses the MDFR feature fusion module achieved an accuracy of 94% and a loss of 0.65; it is still 2% lower than that of the full model. Therefore, it can be seen that multi-scale multi-feature fusion is needed to describe tumours.

4.3. Results for mi-RNA data

For microscopic mi-RNA molecular prediction, both accuracy and AUC curve values of naive Bayes model reached 100%. This shows that the prediction results based on probabilistic models are very accurate in the case of independent assumptions. The specific results are shown in Table 6. For the mi-RNA molecular task (data-set 3), Gaussian naive Bayes achieved an AUC of 1.0 and up to 100% accuracy under the optimal variance-smoothing configuration, indicating that the refined low-dimensional representation retains highly discriminative molecular signals.

Table 6. Parameter settings and results for data-set 3

data-set	Confidence	var_smoothing	AUC	Accuracy
3	95%	1e-7	0.996 ± 0.004	100%
3	95%	1e-6	0.996 ± 0.004	99.5%
3	95%	1e-5	0.992 ± 0.004	99.2%
3	97.5%	1e-7	0.979 ± 0.001	97.0%
3	97.5%	1e-6	0.980 ± 0.001	97.2%
3	97.5%	1e-5	0.978 ± 0.001	97.1%
3	99%	1e-7	0.987 ± 0.001	97.9%
3	99%	1e-6	0.988 ± 0.001	98.0%
3	99%	1e-5	0.986 ± 0.001	97.8%

Dataset 3 has the characteristic of miRNA molecular data features, and it was found that, when compared with var_smoothing at 1e-7 and 1e-6, regularisation of variance can improve the stability of the model and effectively deal with the numerical instability caused by feature noise or minimal

DOI: 10.54254/2753-8818/2026.HZ36311

variance, thus enhancing the stability of generalization. Finally, the best model is chosen at the 95% confidence level and var_smoothing = 1e-7. The particular results are presented in Table 6. Based on the above results, it was decided that var_smoothing = 1e-7 is the optimal choice; it offers a good balance between numerical stability and efficiency (Table 6), and therefore will be used in the model.

Table 7. Five-fold cross-validation results for data-set 3

Model/accuracy	1	2	3	4	5	Mean
SVM	98.0%	97.3%	97.4%	98.2%	98.1%	98.0%
LR	96.0%	96.1%	95.7%	95.4%	95.9%	95.8%
DT	94.0%	93.5%	93.8%	93.5%	93.7%	93.6%
RF	95%	95.4%	94.8%	94.3%	94.6%	94.9%
NB	100%	99.5%	99.8%	99.3%	99.6%	99.7%

Table 7 shows the five-fold cross validation results which show that the model is able to get more than 90% accuracy for all the validation tasks in the test data-set. Remarkably, under certain circumstances, the prior probability based naive Bayes algorithm can be up to 100% accurate in detecting mi-RNA molecules compared to traditional classification algorithms, and has remarkable computational efficiency and simplicity, especially when independent variables and dependent variables have strong prior dependence in small data-sets.

5. Research discussion

To address the problem of tumor image recognition (Dataset 2), MDFR and cross-attention were combined to improve the accuracy of the deep U-Net model to 96.5% and reduce the loss function to 0.35; this was a 9% increase over that obtained by the base CNN model (accuracy 87%, loss 1.5), and the learning curves did not show significant overfitting. MDFR fusion of multi-scale expression features was used to predict the mi-RNA molecular marker (Dataset 3), and both the cross-validation and external validation sets showed that the accuracy of the Naive Bayes model was 100% and the AUC was 1.0; it exceeded the performance of the baseline model by 5%. In short, all the tests have shown that MDFR can improve both the discrimination and generalisation capabilities of the model by reducing the dimensionality of multi-modal data. The details are shown in Table 8.

Table 8. Results of the MDFR comparison experiments

Task	data-set	Model	Method	Accuracy	Baseline	Improvement
Risk stratification	1	NB	MDFR	99%	92%	7%
Tumor recognition	2	U-Net	MDFR	96%	87%	9%
Molecular markers	3	NB	MDFR	100%	95%	5%

The overall experimental results indicate that MDFR significantly enhances the discriminative ability and generalization stability of models across different modalities through multi-modal data dimensionality reduction.

DOI: 10.54254/2753-8818/2026.HZ36311

5.1. Research contribution

The three main contributions of this paper are as follows: (1) Introduction of the MDFR adaptive dimensionality reduction method to improve the processability of multi-source data; (2) Development of an image model that integrates deep U-Net and feature fusion to enhance the robustness of IDC lesion classification; and (3) Application of a Gaussian Naive Bayes model with variance smoothing and empirical category priors for probabilistic modeling of both tabular features and mi-RNA features. Based on the applied datasets, the accuracy of the framework in stratifying breast cancer risk is 99%, the accuracy of IDC lesion classification is around 96-97%, and the accuracy of mi-RNA sub-type prediction is 100%; thus, it can be concluded that the framework has good feature extraction and classification capabilities.

5.2. Limitations and future work

MDFR-Cancer-Net has achieved good predictive results on the chosen data, but it is still not perfect. First, all the experiments used public datasets; thus, the sample size, collection standards and labelling definitions may differ from those in real-world multi-centre clinical practice. Second, Dataset 2 only has block-level image labels and does not contain patient-level clinical outcomes; thus, there may be an overestimation of the model's performance. Furthermore, the high accuracy observed on data-set 3 may be influenced by data-set-specific characteristics and limited distribution variations, necessitating validation with independent external data.

Most of the research is based on public data at present. Future research should include data from many countries, hospitals and devices, have a large amount of data, increase the sample size and diversity of the population, and further evaluate the stability, generalisation and clinical applicability of the model.To evaluate models, it is crucial to incorporate other information like clinical indicators, pathological images, breast imaging data, mi-RNAs, gene expression profiles and patient history to explore the correlation between modalities and to build more complete breast cancer risk prediction and tumor classification models. So far as model prediction, methods such as attention visualization, SHAP analysis and Grad-CAM can be used to interpret the prediction results to show which clinical features, image regions, or molecular markers affect the prediction results the most to promote the transparency of the prediction model and facilitate the physicians' confidence in the prediction results. Moreover, in addition to retrospective data validation, the practical efficacy of the model in early screening, auxiliary diagnosis, risk stratification and treatment decision making in real-world clinical settings should be tested in prospective studies, as well as the risk of misdiagnosis and missed diagnosis.

6. Conclusion

This research has achieved excellent results in the detection of breast cancer tabular data, tumor image IDC data, and mi-RNA molecular data. First of all, this research can better cope with the complex amount of data: in the process of breast cancer carcinogenesis, the data related to breast cancer in different periods are very different, including pathology, genomics, imaging, and other aspects of the whole process. Deep learning algorithms can process data from different periods more accurately. Compared with using only a single model to fit a single data-set, the resulting analysis is more comprehensive. Therefore, all kinds of services can be provided more comprehensively for patients with breast cancer at all stages. In the future, most research on breast cancer will be personalised, precise and comprehensive, and new advances in discovery and treatment will be

DOI: 10.54254/2753-8818/2026.HZ36311

achieved. The above way can be used for other cancers and diseases. Without a doubt, this will be a new milestone in the research of clinical breast cancer.

References

[1] Haque, M. N., et al. (2022, April). Predicting characteristics associated with breast cancer survival using multiple machine learning approaches. *Computational and Mathematical Methods in Medicine, 2022*, 1–12. https://doi.org/10.1155/2022/1249692

[2] Liu, X., et al. (2022, June). Combining machine learning with Cox models for identifying risk factors for incident post-menopausal breast cancer in the UK Biobank. *medRxiv*. https://doi.org/10.1101/2022.06.27.22276932

[3] deAlmeida, R. J., de Moraes Luizaga, C., Eluf-Neto, J., de Carvalho Nunes, H. R., Carvalho Pessoa, E., & Murta-Nascimento, C. (2022, May). Impact of educational level and travel burden on breast cancer stage at diagnosis in the state of Sao Paulo, Brazil. *Scientific Reports*. https://doi.org/10.1038/s41598-022-12487-9

[4] Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022, May). Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine, 127*, 102276. https://doi.org/10.1016/j.artmed.2022.102276

[5] Lotter, W., et al. (2021, February). Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine, 27*(2), 244–249. https://doi.org/10.1038/s41591-020-01174-9

[6] Assegie, T. A. (2021). An optimized K-nearest neighbor based breast cancer detection. *Journal of Robotics and Control, 2*(3). https://doi.org/10.18196/jrc.2363

[7] Desai, M., & Shah, M. (2020, November). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and convolutional neural network (CNN). *Clinical eHealth*. https://doi.org/10.1016/j.ceh.2020.11.002

[8] Zhang, Z., Huang, L., Li, J., & Wang, P. (2022, April). Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: A novel machine learning survival predictive system. *BMC Bioinformatics, 23*(1). https://doi.org/10.1186/s12859-022-04657-3

[9] Alanazi, S. A., et al. (2021, April). Boosting breast cancer detection using convolutional neural network. *Journal of Healthcare Engineering, 2021*(15). https://doi.org/10.1155/2021/5528622

[10] Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., & Abdel-Mottaleb, M. (2021, April). Convolutional neural networks for breast cancer detection in mammography: A survey. *Computers in Biology and Medicine, 131*, 104248. https://doi.org/10.1016/j.compbiomed.2021.104248

[11] Fahrozi, F., Hadiyoso, S., & Hariyani, Y. S. (2022, April). Breast cancer detection in mammography image using convolutional neural network. *Jurnal Rekayasa Elektrika, 18*(1). https://doi.org/10.17529/jre.v18i1.23255

[12] Lin, Z., Chou, W.-C., Cheng, Y.-H., He, C., Monteiro-Riviere, N. A., & Riviere, J. E. (2022, March). Predicting nanoparticle delivery to tumors using machine learning and artificial intelligence approaches. *International Journal of Nanomedicine, 17*, 1365–1379. https://doi.org/10.2147/ijn.s344208

[13] Dubey, A., Yadav, P., Peeyush, Verma, P., & Kumar, R. (2022, January). Investigation of proapoptotic potential of Ipomoea carnea leaf extract on breast cancer cell line. *Journal of Drug Delivery and Therapeutics, 12*(1), 51–55. https://doi.org/10.22270/jddt.v12i1.5172

[14] Zhou, H., Wang, Z., Guo, J., Zhu, Z., & Sun, G. (2024). Analysis of the potential biological significance of glycosylation in triple-negative breast cancer on patient prognosis. *American Journal of Translational Research, 16*(6), 2212–2232. https://doi.org/10.62347/PXAR3644

[15] Zeng, L., et al. (2023, March). The innovative model based on artificial intelligence algorithms to predict recurrence risk of patients with postoperative breast cancer. *Frontiers in Oncology, 13*. https://doi.org/10.3389/fonc.2023.1117420