Acta Polytechnica **65**(2):136–142, 2025
https://doi.org/10.14311/AP.2025.65.0136

Published by the Czech Technical University in Prague

A COMPARATIVE STUDY OF BREAST CANCER DETECTION AND RECURRENCE PREDICTION USING CATBOOST CLASSIFIER

Rana Dhia'a Abdu-aljabar[a,*], Khansaa Dheyaa Aljafaar[a], Zinah Jaffar Mohammed Ameen[b], Hala A. Naman[c]

[a] *Al-Nahrain University, Information Engineering College, Al-Jadryia, Baghdad, Iraq*

[b] *University of Technology, Computer Engineering Department, Baghdad, Alsinaea, Iraq*

[c] *University of Wasit, College of Engineering, Kut, Wasit, Iraq*

[*] corresponding author: ranadhiaa1@nahrainuniv.edu.iq

Abstract. In 2019, breast cancer accounted for over one-third of all cancer cases in women in Iraq. It affects both men and women, though it is more common in women. This study delves into advanced machine learning techniques – CatBoost, XGBoost, Random Forest, SVM, KNN, and Naive Bayes – to improve the detection and prediction of breast cancer recurrence after healing. The goal is to evaluate models using key metrics (sensitivity, specificity, precision, F1 score, accuracy, ROC, and AUC score). Among all algorithms examined, CatBoost stood out, showcasing AUC values above 98 %, 90 %, and 83 % on different datasets. This research demonstrates how machine learning techniques can significantly improve the accuracy of breast cancer detection and recurrence prediction, steering healthcare providers towards better patient care outcomes and more effective treatment plans.

Keywords: CatBoost classifier, breast cancer, machine learning, bioinformatic.

1. Introduction

The occurrence rate of new cases of cancer in Iraq grew between 2000 and 2019 (from 52.00 to 91.66 incidences per 100 000 people). In 2019, breast cancer ranked first among the top ten types cancers in terms of both percentage and occurrence (34.08 % and 35.95/100 000, respectively), and it also had the highest fatality rate (22.58 % and 6.22/100 000) among all types of cancers [1].

Due to developments in breast cancer screening, medical professionals can now detect breast cancer at an early stage. Early detection greatly increases the likelihood of a successful cancer cure. Even in situations where breast cancer is incurable, there are several strategies to extend the life of the patient. Research on breast cancer has led to discoveries that are helping medical practitioners select the best treatment plans [2].

Numerous papers illustrate the continuous efforts to use machine learning techniques for breast cancer detection, diagnosis, risk assessment, and prognosis prediction. This includes different types of medical imaging data and clinical information. Deniz et al. (2018) delved into the application of convolutional neural networks (CNNs) with transfer learning for automated classification of breast cancer histopathological images – showcasing a notable high classification accuracy [3]. Adam Yala et al. (2019) studied deep learning performance in interpreting digital mammography images specifically for breast cancer screening [4]. Similarly, Scott M. McKinney et al. (2020), investigated using artificial intelligence for predicting breast cancer risk through digital mammography images showing a promise for improved risk assessment [5]. A review paper by J. Kim et al. (2021), summarised new developments in the use of ultrasound imaging and deep learning for automated identification and categorisation of breast cancer [6]. In another paper, Meenalochini et al. (2021) discussed the results of various machine-learning methods for automating the classification of mammography images [7]. In a different research, S. Joo et al. (2021) presented a multimodal deep learning model designed to predict the pathologic complete response (pCR) to neoadjuvant chemotherapy (NAC) in breast cancer patients by integrating clinical information and pre-treatment MR images [8]. Similarly, treading the path of innovation, Xiang Li et al. (2021) investigated a fascinating direction by examining the data taken from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) to predict molecular subtypes of breast cancer using machine learning algorithms [9]. J. Li et al. (2021) wrote a review research to evaluate the present machine learning models and studying models for predicting breast cancer survival, highlighting challenges and opportunities for future research [10]. Mahmood, Ali A. et al. (2023) created a model that trains large-scale MWSIs using the MATLAB platform using sample-based processing. The model uses transfer learning techniques and an Inception-v3-based architecture to detect cancer in different samples [11]. This study demonstrates how important it is to select the most suitable machine learning model for medical prediction tasks and validates CatBoost's reliability.

2. Dataset description

This project used three different types of datasets. The first two databases are classified the patients depending on whether to detect breast cancer or not. One method relies on data from digitised images, while the other uses microarray gene expression. The third one classified the patients depending on the recurrence of the disease after the healing. The data is computed from clinical information.

The first dataset is called WDBC (Wisconsin Diagnostic Breast Cancer), which is taken from the UC Irvine Machine Learning Repository [12]. 32 features are computed from a digital image of a fine needle aspirate (FNA) of a breast tumour. The explanation of the features of the cell nuclei seen in the image is as follows: the first two features are the ID and the diagnosis (M for malignant, B for benign), while the others are input features. There are 10 main real-valued features calculated for each cell nucleus: radius, texture, smoothness, concavity, concave spots, compactness, perimeter, area, symmetry, and fractal dimension. Each image has 30 features calculated from the standard error, mean, and "worst" (the average of the three largest values) of the previous 10 features. The total number of instances is 569 (357 benign, and 212 malignant). In the pre-processing stage of this dataset, one instance was deleted from this dataset because of incomplete information. Due to the imbalance in the dataset, with more benign than malignant instances, the number of malignant instances is increased to improve learning outcomes.

The second dataset is called GSE42568. This dataset has been submitted to the GEO (Gene Expression Omnibus) data repository represented by gene expressions. This dataset is from the type of gene expression microarray with its clinical information. There are 17 normal instances and 104 cases of breast cancer samples (removed before tamoxifen or chemotherapy agent treatment) from individuals who were diagnosed between the ages of 31 and 89 (mean age = 58 years). At the time of diagnosis, twenty of the women were under fifty and seventy-seven women were exactly or over fifty years. The size of the tumours ranges (from 0.6 cm to 8.0 cm) with a mean of 2.79 cm. Eighteen tumours were smaller than 2 cm (T1) in the maximal dimension, while 83 tumours were between 2–5 cm (T2), and 3 tumours were larger than 5 cm (T3). The remaining 54 666 features are dedicated to gene expression [13]. The pre-processing step involves duplicating normal instances several times to reduce the gap between normal and breast cancer cases, leading to better learning outcomes.

The third dataset is called BCRD (Breast Cancer Recurrence Data) [14], which was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. There are 286 cases total in this dataset, 201 of which belong to the no-recurrence class and 85 to the recurrence class. The instances are described by 10 attributes. Attribute Information: the first attribute is the class, no-recurrence-events or recurrence-events, and the remaining 9 attributes are [14]:

- Age: 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99.
- Menopause (pre-or postmenopausal status at the time of diagnosis): lt40, ge40, premeno.
- Tumour size: Every 5 sequence units represented a group. For example: (0–4, 5–9, ..., 55–59)
- Inv-nodes (Axillary lymph nodes with breast cancer metastases that are obvious): Every three sequence numbers of nodes represented one group. For example: (0–2, 3–5, ..., 36–39)
- Node-caps: yes, no.
- Degree of malign: 1, 2, 3.
- Breast: right, left.
- Breast-quad: central, left-up, left-low, right-up, right-low.
- Irradiance: no, yes.

During pre-processing, the dataset duplicates instances of the recurrence class multiple times to balance it with the no-recurrence class, improving learning outcomes.

This study used three datasets to offer various perspectives on breast cancer: tumour cell characteristics (WDBC), gene-level insights (GSE42568), and patient recurrence risk factors (BCRD). They help create strong predictive models for breast cancer, covering early diagnosis, treatment planning, and recurrence prediction.

3. Materials and methods

Predicting breast cancer recurrence is a critical task in medical research, with the goal of accurately identifying patients at risk of disease recurrence for timely intervention and improved prognosis. This comparative analysis study used several state-of-the-art machine learning algorithms to explore their effectiveness in detecting and predicting breast cancer recurrence. It used CatBoost, XGBoost, KNN, Random Forest, Naive Bayes, and LIBSVM to ensure a comprehensive comparison of different machine learning techniques. This diversity allows us to identify the best-performing algorithm for breast cancer classification tasks and better understand how different methods handle the nuances of different datasets.

3.1. CatBoost

CatBoost (Categorical Boosting) is a gradient boosting algorithm for decision trees. It is particularly useful for categorical features. It was developed by Yandex researchers and engineers [15]. Thus, datasets including a range of data types can benefit from it.

It makes use of an advanced technique called ordered boosting, which performs better than traditional

gradient boosting algorithms and provides shorter training times [16]. With its ability to reduce overfitting and limit the need for hyperparameter adjustments, CatBoost is a desirable choice for practical uses. This method uses the same framework of gradient boosting to increase the model accuracy, which focuses on poorly classified samples. Additionally, it is highly suitable for real-world data as it handles categorical information without requiring extra encoding or pre-processing.

3.1.1. CatBoost work

The CatBoost approach uses gradient descent to decrease the loss function, iteratively constructing an ensemble of trees. Each time, a new tree is added to minimise the loss function. Once the negative gradient of the loss function for the current predictions has been found, a new tree is fitted to the gradient. The gradient descent step size is determined by the learning rate. The method is repeated until the convergence conditions are met or a predefined number of trees has been added. CatBoost combines its predictions from every tree in the ensemble to generate its final output.

3.1.2. Catboost math

The following is a representation of CatBoost: CatBoost attempts to learn an $F(x)$ function that predicts the target, which is the y variable, from a training dataset containing N samples and M features. Each sample is represented as (x_i, y_i). Where X_i is the M features vector and y_i is its target variable.

$$F(x) = F_0(x) + \sum_{m=1}^{M} \sum_{i=1}^{N} f_m(x_i), \qquad (1)$$

where

$F(x)$ is what CatBoost aims to define as the general prediction function. It predicts the appropriate target variable, y, given an input vector, x.

$F_0(x)$ is the first estimation or baseline forecast. It is frequently set equal to the training dataset's target variable's mean.

$\sum_{m=1}^{M}$ represents the total collection of trees. The ensemble's total number of trees is M.

$\sum_{i=1}^{N}$ represents the total collection of training sets. The ensemble's total number of training samples is N.

$f_m(x_i)$ represents the m^{th} tree's expected value for the i^{th} training set. Every tree in the ensemble contributes its own forecast to the final prediction for every training sample.

The overall prediction $F(x)$ can be calculated by adding the forecasts of each tree, $f_m(x_i)$, for each training sample and the initial guess $F_0(x)$, according to the equation. Every tree (m) and training sample (i) passes through the summing procedure [16].

3.2. K-nearest neighbours (KNN)

KNN is a straightforward, yet effective technique that can be applied to regression and classification problems. It works on the principle of similarity, where the average value or majority class of an instance's closest neighbours in the feature space is used to make predictions for that instance. KNN is computationally efficient for small to medium-sized datasets because it is non-parametric and doesn't require a training phase [17].

3.3. XGBoost

The scalable and high performance gradient boosting system known as XGBoost is well optimised. The goal function incorporates a regularisation term to manage model complexity and avoid overfitting. With its many hyperparameters, XGBoost's great degree of customisation enables fine-tuning for specific datasets and goals [18].

3.4. Random forest

It is a classification technique based on an ensemble learning technique that builds several decision trees during training, which leads to the prediction of the class mode for each tree. Using feature sampling and bootstrap aggregating (bagging) adds unpredictability, which lessens variance and improves generalisation. Random Forest can effectively handle high-dimensional data and is resistant to noise and outliers [19].

3.5. Naive Bayes

Based on the Bayes theorem and the supposition of conditional independence of features given the class label, the Naive Bayes classifier is probabilistic. In fact, Naive Bayes frequently performs remarkably well, especially for text classification and other high-dimensional datasets, despite its seeming simplicity and naive assumptions. Large-scale applications can benefit from its low training data requirements and processing efficiency [20].

3.6. LIBSVM

A popular library for implementing support vector machines (SVMs), LIBSVM (Library for Support Vector Machines) can handle problems involving both regression and classification. In order to achieve strong generalisation performance, the SVM tries to identify the ideal hyperplane that separates multiple classes in the feature space with the greatest margin.

Modelling nonlinear interactions can be done with flexibility because of the LIBSVM's range of kernel functions, which include radial basis function (RBF), polynomial, and linear kernels [21].

In order to determine the best strategy for detecting and predicting the recurrence of breast cancer, we carefully assessed and analysed the distinct qualities and trade-offs of each of these algorithms.

CatBoost		XGBoost		LIBSVM	
Hyperparameter	**Value**	**Hyperparameter**	**Value**	**Hyperparameter**	**Value**
max_depth	6	max_depth	6	kernel	RBF
n_estimators (Trees)	100	n_estimators (Trees)	100	gamma	1
Learning rate	0.3	Learning rate	0.3	tolerance	0.001
Subsample	0.7	Subsample	0.7	C	1
grow_policy	Lossguide				

Random Forest		KNN		Naive Bayes	
Hyperparameter	**Value**	**Hyperparameter**	**Value**	**Hyperparameter**	**Value**
max_depth	6	n_neighbor	2	var_smoothing	1e-9
min_samples_leaf	1	weights	uniform	sample_weight	None
n_Trees	500	algorithm	auto		
min_samples_split	0.05	leaf_size	1		

TABLE 1. Hyperparameters settings of the comparative models.

4. EXPERIMENTAL SETUP

In this experiment, we used Python 3.7 for programming. The hyperparameter settings of all classification systems presented in this article are shown in Table 1.

We used the Sensitivity, Specificity, F1 score, accuracy, and AUC matrices for the comparison of all classification systems. The train-test split was 70 % for training and 30 % for testing.

Before going through the results, it is necessary to explain the meaning of each metric used in this study: Sensitivity measures the proportion of actual positive cases that the model correctly identifies as positive. High sensitivity indicates that the model is effective at identifying positive cases. This metric is crucial in situations where missing positive cases (false negatives) is costly, such as in disease detection. High specificity, or True Negative Rate, demonstrates the model's proficiency in accurately identifying negative cases. It's especially important when falsely identifying negatives as positives (false positives) has significant consequences. Precision measures the ratio of correct positive predictions to the total positive predictions made. High precision indicates that the model generates fewer false positive errors when identifying positive predictions. This metric is particularly important when false positives are costly or undesirable. The F1 Score metric is especially useful when there's a class imbalance, and it needs to balance both recall and precision. A high F1 score indicates a balance between correctly identifying positive cases and limiting false positives. The AUC (Area Under the ROC Curve) metric measures the area under the Receiver Operating Characteristic (ROC) curve, which plots Sensitivity against Specificity. A high AUC indicates that the model has a good balance between Sensitivity and Specificity across all threshold levels. It is a valuable measure for understanding the overall model performance, especially for datasets with imbalanced classes. The accuracy metric measures the ratio of correct predictions to total predictions. High accuracy may not mean an effective performance if one class dominates, as the model may be biased towards the majority class.

5. RESULTS AND DISCUSSION

The research used a variety of databases for comparing the CatBoost model with other machine learning models. The results of each machine learning model applied to each database are illustrated in Table 2 and Figures 1, 2, and 3.

The table and figures compare various machine-learning models on the WDBC, GSE42568, and BCRD datasets. The models evaluated are CatBoost, XGBoost, LIBSVM, Random Forest, KNN, and Naive Bayes. The performance metrics used include Sensitivity, Specificity, Precision, F1 Score, AUC, and Accuracy. Here are the key points:
CatBoost performed exceptionally well on the WDBC dataset, achieving the highest Sensitivity (0.95), perfect Specificity (1.00), Precision (1.00), F1 Score (0.98), AUC (0.98), and Accuracy (0.98). XGBoost followed closely with a similar performance, but slightly lower Sensitivity (0.94) and F1 Score (0.97). The other models, LIBSVM, Random Forest, KNN, and Naive Bayes, performed competitively but were slightly behind CatBoost and XGBoost on most metrics.

In the second dataset (GSE42568 dataset); CatBoost and LIBSVM tied for the best performance in several metrics: Sensitivity (1.00), Precision (0.97), F1 Score (0.98), AUC (0.90), and Accuracy (0.97). XGBoost showed a lower performance, particularly in Specificity (0.40) and AUC (0.67), indicating that it struggled with this dataset. Naive Bayes performed poorly in Specificity (0.0) and AUC (0.50), suggesting that it was not suitable for this dataset.

In the BCRD dataset, CatBoost outperformed other models with Sensitivity (0.78), Specificity (0.88), Precision (0.84), F1 Score (0.81), AUC (0.83), and Accuracy (0.83). XGBoost and Random Forest demon-

Model name	Sensitivity	Specificity	Precision	F1 score	AUC	Accuracy
WDBC dataset						
CatBoost	0.95	1.00	1.00	0.98	0.98	0.98
XGBoost	0.94	1.00	1.00	0.97	0.97	0.98
LIBSVM	0.91	0.98	0.97	0.94	0.94	0.95
RForest	0.92	0.99	0.98	0.95	0.96	0.96
KNN	0.91	0.99	0.98	0.94	0.96	0.96
Naive Bayes	0.92	0.99	0.98	0.95	0.96	0.96
GSE42568 dataset						
CatBoost	1.00	0.80	0.97	0.98	0.90	0.97
XGBoost	0.94	0.40	0.91	0.92	0.67	0.86
LIBSVM	1.00	0.80	0.97	0.98	0.90	0.97
RForest	1.00	0.60	0.94	0.97	0.80	0.95
KNN	0.97	0.80	0.97	0.97	0.88	0.95
Naive Bayes	1.00	0.0	0.86	0.93	0.50	0.86
BCRD dataset						
CatBoost	0.78	0.88	0.84	0.81	0.83	0.83
XGBoost	0.76	0.80	0.76	0.76	0.78	0.78
LIBSVM	0.29	0.83	0.58	0.38	0.56	0.58
RForest	0.53	0.80	0.68	0.60	0.66	0.68
KNN	0.43	0.85	0.70	0.53	0.64	0.66
Naive Bayes	0.33	0.81	0.59	0.42	0.57	0.59

Table 2. Comparison of results for breast cancer detection and recurrence prediction.

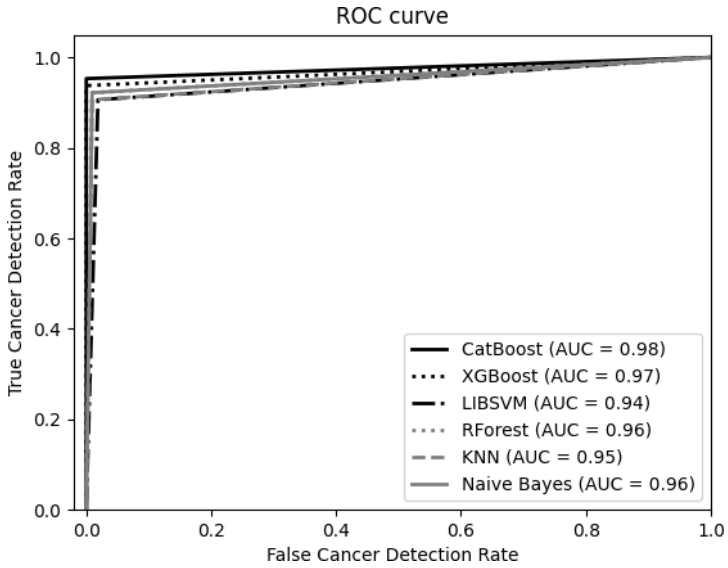

Figure 1. The ROC and AUC values of the comparative models when used with the WDBC dataset.

strated moderate performance, they fell short in effectiveness when compared to CatBoost. IBSVM and Naive Bayes had the lowest sensitivity at 0.29 and 0.33, respectively, making them less effective at predicting outcomes for this dataset.

The results show that CatBoost excels in its performance across multiple datasets, particularly in WDBC and GSE42568. This indicates that CatBoost adeptly handles various data distributions and recognises feature importance, making it a reliable option for a wide range of datasets. It uses a unique boosting method that reduces overfitting, enhancing generalisation and results in high sensitivity and specificity in most datasets. The CatBoost model is tailored for efficiently handling categorical data, which likely enhances its performance on complex datasets, such as WDBC and BCRD.

From this study, the pros and cons of the CatBoost model can be summarised in Table 3.

6. Conclusion

Based on a thorough evaluation of the results, CatBoost is an effective model for detecting and predicting breast cancer recurrence across different datasets.

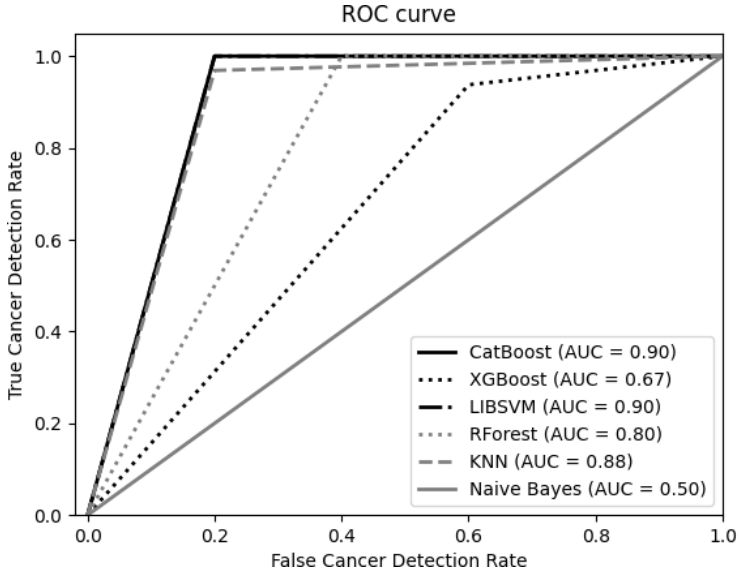

FIGURE 2. ROC and AUC values of the comparative models when used with the GSE42568 dataset.

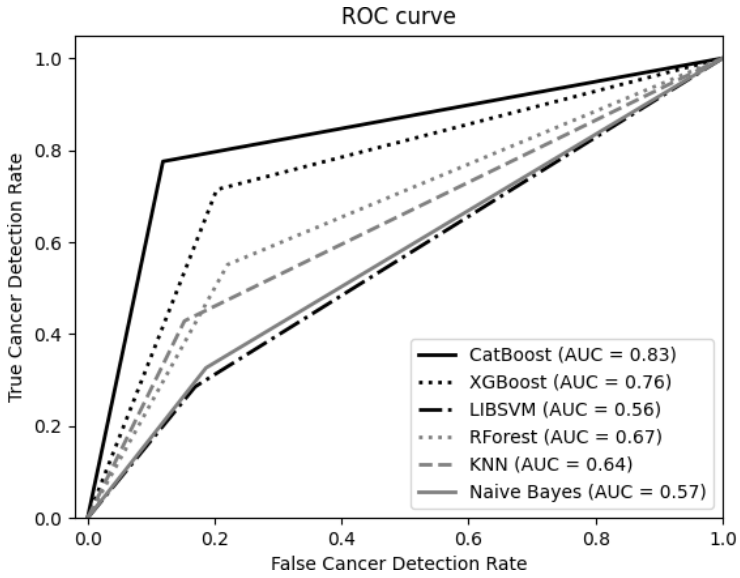

FIGURE 3. ROC and AUC values of the comparative models when used with the BCRD dataset.

The pros	The cons
It can handle categorical features. This can decrease the pre-processing time. It was useful for learning on small datasets because it reduces overfitting well. It requires little adjustment of the hyperparameters. It works well with imbalanced datasets.	Memory-intensive, particularly when working with big datasets. Takes longer time than other models to train on small datasets.

TABLE 3. The pros and cons of the CatBoost model.

This option is recommended for the study as it consistently performs well in terms of F1 Score, Accuracy, Specificity, Precision, Sensitivity, and AUC score. Although there is still great potential in other models, such as XGBoost and LIBSVM, they are not as consistent and stable as CatBoost. Therefore, the CatBoost model can be trusted to detect and predict the recurrence of breast cancer. Depending on the dataset, certain models offer benefits and drawbacks. For example, XGBoost and LIBSVM perform well but are less reliable.

This study shows that CatBoost is reliable and emphasises the importance of selecting the right machine learning model for medical predictions.

7. Future works

Based on the findings and recommendations of this work, several avenues for future research can be explored to improve the ability of CatBoost models to predict breast cancer recurrence:

(1.) Model Tuning and Optimisation.

(2.) Incorporating Additional Data.

(3.) Integrating the benefits of multiple models could lead to more accurate predictions.

By addressing these next areas, research can make a significant contribution to advancing predictive modelling in the detection of recurrent breast cancer. As a result, more accurate, reliable, and therapeutically useful instruments can be produced, improving patient outcomes.

References

[1] M. M. Y. Al-Hashimi. Trends in breast cancer incidence in Iraq during the period 2000–2019. *Asian Pacific Journal of Cancer Prevention* **22**(12):3889–3896, 2021. https://doi.org/10.31557/APJCP.2021.22.12.3889

[2] I. J. Mustafa, O. R. Abdullah, N. Al-Saffar, et al. Quality of life assessment in women with breast cancer in Nineveh, Iraq. *Cureus* **16**(1):e51589, 2024. https://doi.org/10.7759/cureus.51589

[3] E. Deniz, A. Şengür, Z. Kadiroğlu, et al. Transfer learning based histopathologic image classification for breast cancer detection. *Health Information Science and Systems* **6**(1):18, 2018. https://doi.org/10.1007/s13755-018-0057-x

[4] A. Yala, C. Lehman, T. Schuster, et al. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**(1):60–66, 2019. https://doi.org/10.1148/radiol.2019182716

[5] S. M. McKinney, M. Sieniek, V. Godbole, et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**(7788):89–94, 2020. https://doi.org/10.1038/s41586-019-1799-6

[6] J. Kim, H. J. Kim, C. Kim, et al. Weakly-supervised deep learning for ultrasound diagnosis of breast cancer. *Scientific Reports* **11**(1):24382, 2021. https://doi.org/10.1038/s41598-021-03806-7

[7] G. Meenalochini, S. Ramkumar. Survey of machine learning algorithms for breast cancer detection using mammogram images. *Materials Today: Proceedings* **37**:2738–2743, 2021. https://doi.org/10.1016/j.matpr.2020.08.543

[8] S. Joo, E. S. Ko, S. Kwon, et al. Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Scientific Reports* **11**(1):18800, 2021. https://doi.org/10.1038/s41598-021-98408-8

[9] Y. Zhang, J.-H. Chen, Y. Lin, et al. Prediction of breast cancer molecular subtypes on DCE-MRI using convolutional neural network with transfer learning between two centers. *European Radiology* **31**(4):2559–2567, 2021. https://doi.org/10.1007/s00330-020-07274-x

[10] J. Li, Z. Zhou, J. Dong, et al. Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS One* **16**(4):e0250370, 2021. https://doi.org/10.1371/journal.pone.0250370

[11] A. A. Mahmood, S. Sadeq, Y. I. Aljanabi, A. H. Sabry. Developing a convolutional neural network for classifying tumor images using Inception V3. *Eastern-European Journal of Enterprise Technologies* **3**(9 (123)):86–93, 2023. https://doi.org/10.15587/1729-4061.2023.281227

[12] W. Wolberg, O. Mangasarian, N. Street, W. Street. Breast cancer Wisconsin (diagnostic), 1993. https://doi.org/10.24432/C5DW2B

[13] C. Clarke, S. F. Madden, P. Doolan, et al. Correlating transcriptional networks to breast cancer survival: A large-scale coexpression analysis. *Carcinogenesis* **34**(10):2300–2308, 2013. https://doi.org/10.1093/carcin/bgt208

[14] L. Lin. *Properties and applications of biharmonic and K-harmonic distances in clustering.* Bachelor's thesis, Oregon State University, Corvallis, Oregon, USA, 2024.

[15] A. V. Dorogush, V. Ershov, A. Gulin. CatBoost: gradient boosting with categorical features support. *arXiv preprint* 2018. https://doi.org/10.48550/arXiv.1810.11363

[16] L. Prokhorenkova, G. Gusev, A. Vorobev, et al. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, vol. 31, pp. 1–11. 2018.

[17] S. Zhang, J. Li. KNN classification with one-step computation. *IEEE Transactions on Knowledge and Data Engineering* **35**(3):2711–2723, 2021. https://doi.org/10.1109/TKDE.2021.3119140

[18] R. D. Abdu-Aljabar, O. A. Awad. Improving lung cancer relapse prediction using the developed Optuna_XGB classification model. *International Journal of Intelligent Engineering and Systems* **16**(1):131–141, 2023. https://doi.org/10.22266/ijies2023.0228.12

[19] M. Bader-El-Den, E. Teitei, T. Perry. Biased random forest for dealing with the class imbalance problem. *IEEE Transactions on Neural Networks and Learning Systems* **30**(7):2163–2172, 2019. https://doi.org/10.1109/TNNLS.2018.2878400

[20] F.-J. Yang. An implementation of Naive Bayes classifier. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 301–306. 2018. https://doi.org/10.1109/CSCI46756.2018.00065

[21] C.-C. Chang, C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3):27, 2011. https://doi.org/10.1145/1961189.1961199