

Supervised Learning-Based Prognostic Modelling For Early Detection of Breast Cancer

Priyanka Mohan¹, Kusuma P²

¹Surana College (Autonomous), Kengeri Satellite Town, Bangalore, India

²BNM Institute of Technology, Banashankari II Stage, Bangalore, India

Email: priyankaarjun2808@gmail.com¹, kusuk85@gmail.com²

Abstract

Breast cancer continues to be among the leading cause of cancer-related mortality in women globally, and therefore there is the need to have more accurate diagnostic techniques for both diagnosis and treatment of the condition. Manual diagnoses that are currently being carried out using technologies like mammography, MRI, and ultrasound tend to be relatively slow, subjective and with some level of inter-observer variability. Therefore, this project proposes to come up with an advanced prognosis system that makes use of supervised machine learning techniques to ensure accuracy of breast cancer diagnosis as well as minimize the cost involved. Some of the machine learning algorithms like Random Forest and Decision Tree will be utilized along with computer aided diagnosis tools to determine various features that can include the tumor size, shape and texture as well as nuclear features of the cancer. Preprocessing steps included normalization, handling of missing values and feature selection before carrying out model training and testing. From the experimental results, it can be noted that the Random Forest model performed better than the decision tree model with higher accuracy and lower number of false positives and negatives.

Keywords

Breast cancer, MRI, Supervised-learning, Random forest, Decision tree

Introduction

One area of artificial intelligence is machine learning, which deals with creating models that can learn on their own and improve over time without requiring any programming from software engineers. Supervised, unsupervised, and reinforcement learning approaches are among the domains that deal with machine learning. These approaches have been applied in a number of industries, including banking, healthcare, and robotics to carry out particular tasks.

In machine learning, it is essential to create models that are both accurate and comprehensible. Decision Trees and Random Forests are two models that have these characteristics. The Decision Tree creates a hierarchical tree-like architecture that resembles a flowchart and allows data to branch out according to features. Although this model's predictions

Submission: 4 May 2026; **Acceptance:** 25 June 2026; **Available online:** July 2026



Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the web [site: https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/)

are rather simple to comprehend, decision trees are prone to overfitting and have trouble with challenging datasets.

Random Forest, an ensemble learning technique, is used to address these issues. On the basis of many sets of data samples, it builds numerous decision trees and aggregates their results to produce a final forecast. This method lowers variation and improves accuracy. Both methods are widely used in classification and regression applications, as well as in medical diagnosis, economic trend forecasting, and fraud detection.

Decision tree:

- a) Entropy (Information Gain criterion)
Entropy measures impurity or uncertainty in the dataset
- b) Information Gain (IG)
Used to decide the best feature for a split
- c) Gini Index (alternative to Entropy) Used in CART (Classification and Regression Trees)

Random Forest

- a) Bootstrap Aggregating (Bagging)
Random subsets of data are selected (with replacement) to train each tree.
- b) Voting (for Classification)
Final prediction is the majority vote from all trees:
$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_n(x)\}$$
- c) Averaging (for Regression)

Literature Survey

[1] The incidence of breast cancer is still one of the major health issues that affect women, which calls for early diagnosis. In this paper, Random Forests classification algorithm is used to detect breast cancer patients, and its effectiveness is verified via the prediction accuracy rate of 98%. Also, predictions indicate that by 2030, more female fatalities due to lung cancer than breast cancer will occur in many countries because of increasing rates of cigarette use.[2] Despite a declining trend in mortality from breast cancer, the above predictions indicate that preventive methods should be developed against lung cancer, which will become more life-threatening for females. Furthermore, breast cancer is still a frequent type of cancer in Chinese women. [3] The results of the analysis carried out for the period from 2011 to 2020 in Fujian Province, with subsequent forecasting until 2025, can provide useful information.

This study [4] makes use of ResNet architectures for the purpose of breast cancer image classification with an accuracy rate of 98% and lightweight design consisting of 3.6 million parameters through contrastive learning and architectural adjustments. On the other hand, another study [5] explores machine learning algorithms such as SVM, Decision Tree, Naive Bayes, and KNN using the Wisconsin breast cancer dataset, wherein performance is measured based on accuracy and precision. Among these machine learning algorithms, SVM was able to produce the highest accuracy rate of 96.7%, showing its capability as a method for early diagnosis of the disease. Study [6] uses breast cancer histopathology image classification using state-of-the-art feature extraction methods together with machine learning, wherein the LGBM model that utilized data augmentation produced an impressive accuracy rate of 99.87%. Another interesting result found in study [7] involves FSL versus WSLMIL as diagnostic methods in breast cancer

diagnosis. The results showed that transformer and MLP architectures were superior to traditional CNN architecture in patch-level classifications. WSLMIL with its domain-specific backbone and efficient feature aggregation technique can be used effectively in clinical applications.

Research paper [8] addresses the problem of the application of both deep learning and classical machine learning algorithms for analyzing infrared thermography data for early diagnosis of breast cancer. The combined ResNet152-SVM model demonstrated 97.62% accuracy, suggesting a high level of diagnostic efficiency. Research paper [9] considers methods for fine-tuning the parameters of the machine learning algorithms used for breast cancer diagnosis. KNN classifier showed a perfect recall and, in conjunction with logistic regression and multilayer perceptron, provided 99.42% accuracy, confirming the high efficiency of optimization methods. In addition, SelfAdaptNet [10] integrates self-supervised learning algorithms BYOL and SimCLR with adversarial techniques for improving the efficiency of detection in conditions of low amounts of available data.

Proposed Methodology

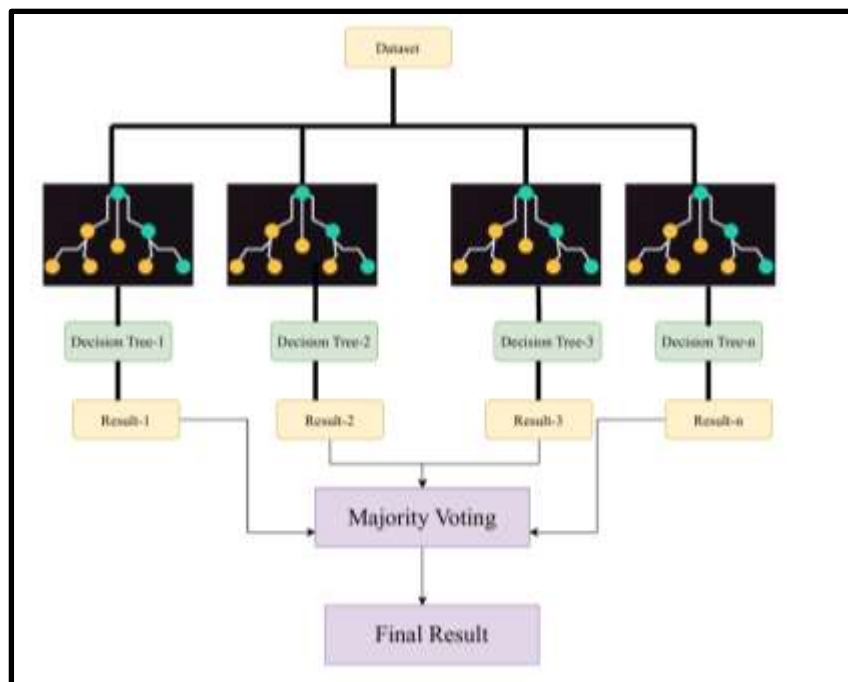


Figure 1. Proposed system Architecture

Figure 1 shows the proposed system Architecture. For the suggested approach, it was decided to use the Random Forest algorithm as a way to increase the accuracy and reliability of classification models. The workflow will start from acquiring and preparing a data set with labels, which includes normalization, feature encoding, and the handling of missing values. Then, after splitting the data set into several samples, bagging is used to train a different tree on each sample. At each node while constructing trees, some subset of randomly selected features is considered for making decisions, which ensures that trees have a low correlation, and overfitting is decreased. This is the primary advantage of Random Forest, since one tree classifier tends to overfit the training data. When each of the decision trees predicts the output for the input sample, then the predicted value is voted on for making the final prediction. The data is analyzed using scatterplots that show correlations between various features of breast cancer, with different colors

being used to denote Malignant samples (357) and Benign samples (212). This clearly shows how the variables are correlated among each other, in each of the diagnostic classes.

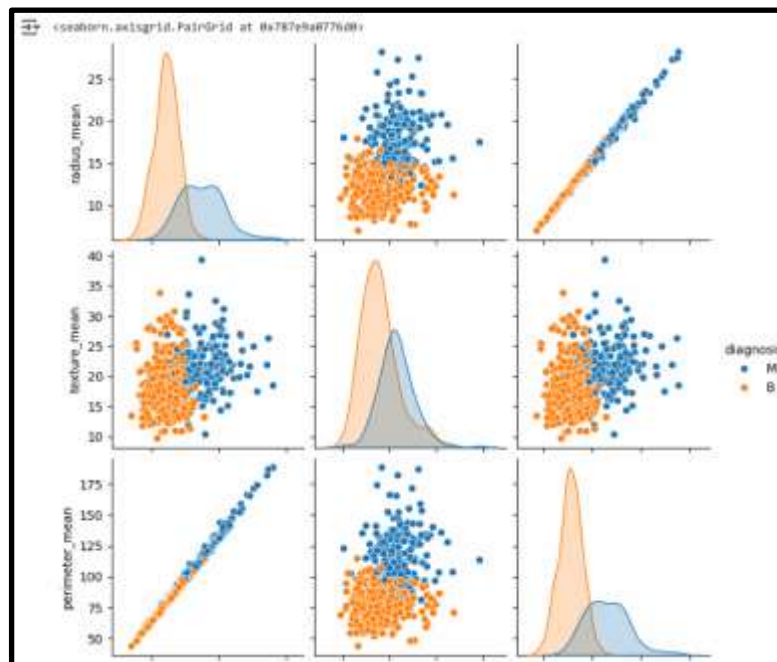


Figure 2. Heatmap visualization in breast cancer dataset.

Figure 2 shows the heatmap demonstrating the correlation between various features in the breast cancer data set. The heatmap is developed with the use of Seaborn, where each square denotes the strength of the relationship between each pair of the features in terms of percentage. It enables one to determine the presence of strong positive or negative relationships and identify which features can be ignored while selecting the important ones. Figure 3 shows the categories count.

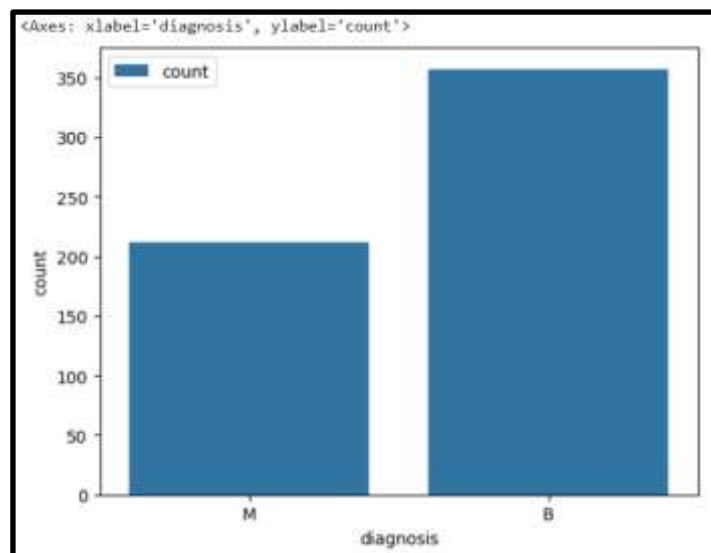


Figure 3. Categories count

The graph shows the frequency of diagnostic classes in the breast cancer dataset, namely Malignant (M) and Benign (B). The x-axis denotes the two classes, and the y-axis shows their frequencies. It is evident from the graph that there are more benign records than malignant records, with the former being more than 350 and the latter being just more than 200. Thus, it

implies that there is a slight imbalance between the two classes, with a higher representation of benign records over malignant records.

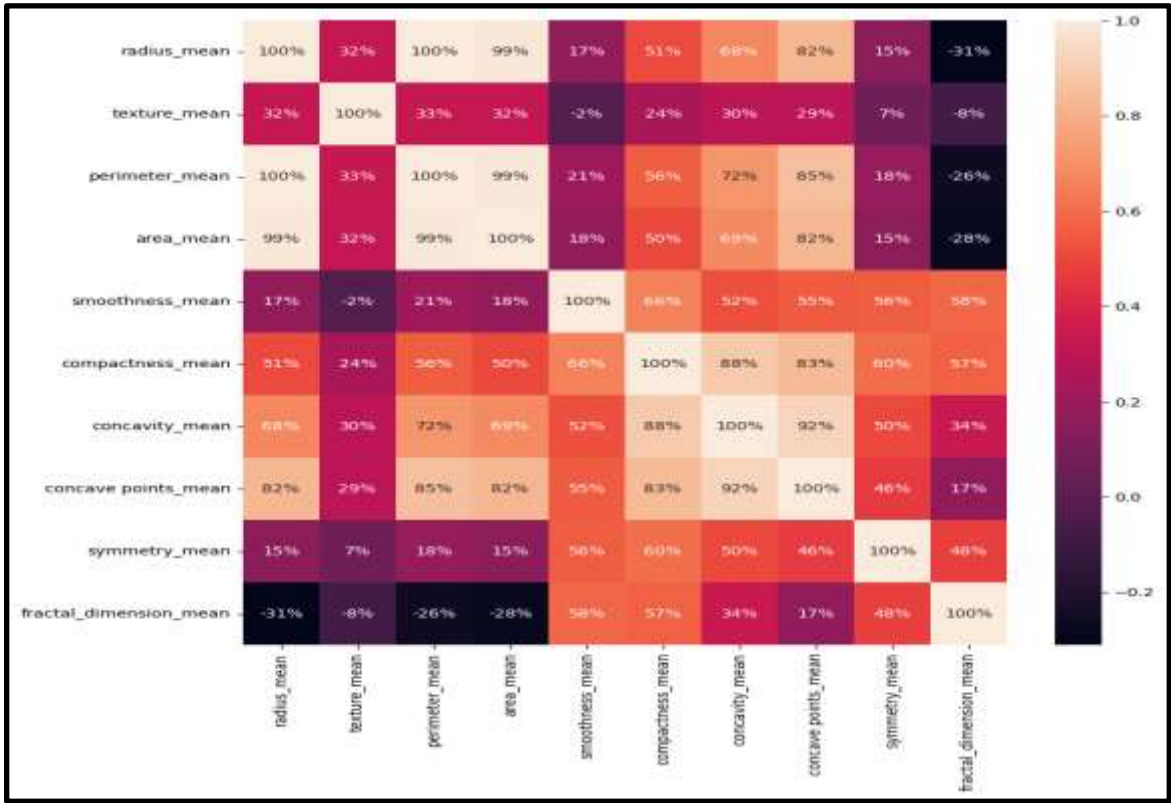


Figure 4. Features from Dataset

Figure 4 provides a correlation heatmap that shows the relationship between different mean features in the breast cancer dataset. Each cell represents a percentage value, indicating the level of correlation between two features. The higher the percentage value, the stronger the correlation, where the closer the value is to 100%, the stronger is the positive correlation. On the other hand, if the value is near 0% or even less than 0%, then it shows the weaker or inverse correlation between the features. The values on the diagonal are 100%, as all attributes are positively correlated with themselves.

In the above heatmap, the attributes like radius_mean, perimeter_mean, and area_mean are highly positively correlated, implying that the traits are related to each other, and the traits could be related to the size of the tumor. Moreover, the correlation between the concavity_mean, compactness_mean, and concave points_mean are also high, showing that these traits are related to the shape of the tumor. On the other hand, some of the attributes like texture_mean and fractal_dimension_mean are weakly correlated with other attributes. Figure 5 shows the classification performance report for Model 0 and Model 1.

Model 0					
	precision	recall	f1-score	support	
B	0.96	0.97	0.96	67	
M	0.96	0.94	0.95	47	
accuracy			0.96	114	
macro avg	0.96	0.95	0.95	114	
weighted avg	0.96	0.96	0.96	114	
0.956140350877193					
Model 1					
	precision	recall	f1-score	support	
B	0.96	0.97	0.96	67	
M	0.96	0.94	0.95	47	
accuracy			0.96	114	
macro avg	0.96	0.95	0.95	114	
weighted avg	0.96	0.96	0.96	114	
0.9385964912280702					

Figure 5. Classification performance report for Model 0 and Model 1

The comparison of the performance of two models for breast cancer diagnosis is presented in the table below. One should note that the efficiency of these two models can be assessed using the indicators of precision, recall, f1-score, and accuracy of classification, which were separately evaluated for benign (class B) and malignant (class M) tumors.

In particular, the indicator of precision for model 0 is 0.96 for both types of tumors. However, at the same time, the indicator of recall for benign cancer is somewhat larger compared to malignant one and equals 0.97 and 0.94, correspondingly. As a consequence, the resulting f1-scores are 0.96 and 0.95 for benign and malignant tumors, respectively. Overall accuracy is 0.96, which is very similar to the macro and weighted average values of 0.95 and 0.96, respectively. Finally, the final score of model 0 equals 0.9561.

From the above chart, it can be observed that the performance indicators for model 1 are practically the same as those for model 0, apart from the final score, which is 0.9386.

Overall, both models demonstrate high effectiveness in distinguishing between benign and malignant cases, with Model 0 showing a slight advantage in reliability. Figure 6 shows the predictive versus actual values.



Predicted values:

```
[ 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'M' 'B' 'M'
  'M' 'M' 'M' 'M' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'M' 'B' 'M' 'B' 'M' 'B' 'M'
  'B' 'M' 'B' 'M' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'B' 'B' 'M' 'M' 'M' 'M' 'B'
  'B' 'B' 'B' 'B' 'B' 'M' 'M' 'M' 'B' 'B' 'M' 'B' 'M' 'M' 'M' 'B' 'B' 'M'
  'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'M' 'M' 'M' 'B' 'M' 'B' 'B' 'B' 'M' 'M'
  'B' 'M' 'B' 'M' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'M' 'B'
  'B' 'M' 'B' 'M' 'M' 'B' ]
```

Actual values:

```
[ 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'M'
  'M' 'M' 'M' 'M' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'M' 'B' 'M' 'B' 'M' 'B' 'M'
  'B' 'M' 'B' 'M' 'M' 'B' 'M' 'B' 'B' 'M' 'B' 'B' 'B' 'M' 'M' 'M' 'M' 'B'
  'B' 'B' 'B' 'B' 'B' 'M' 'M' 'M' 'B' 'B' 'M' 'B' 'M' 'M' 'M' 'B' 'B' 'M'
  'B' 'M' 'M' 'B' 'B' 'B' 'B' 'B' 'M' 'M' 'M' 'B' 'M' 'B' 'B' 'B' 'M' 'M'
  'B' 'M' 'B' 'M' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'M' 'B'
  'M' 'M' 'B' 'M' 'M' 'B' ]
```

Figure 6. Predictive vs Actual values.

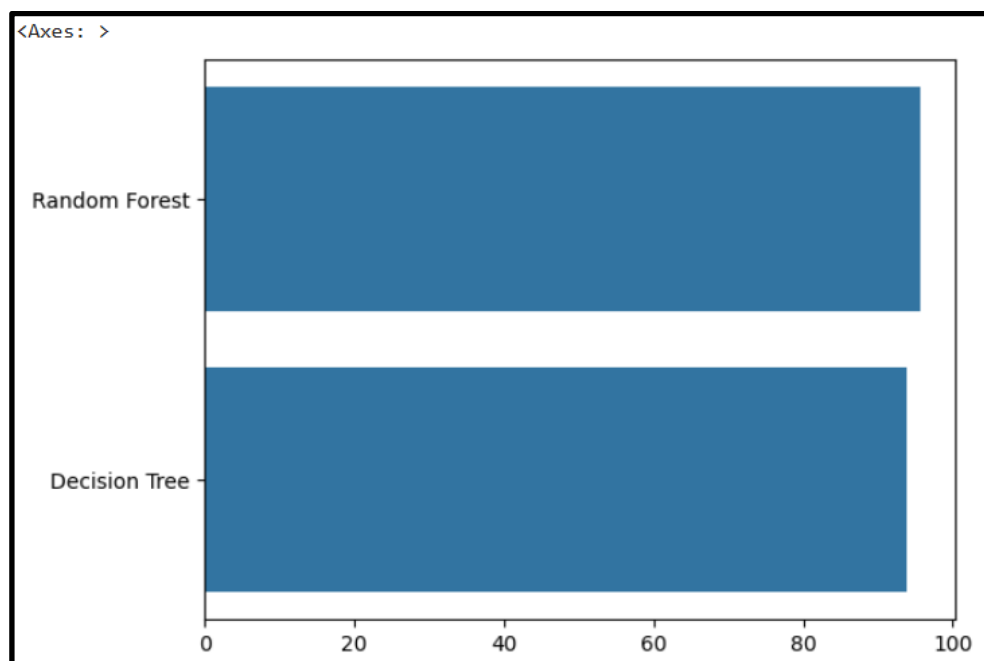


Figure 7. Bar chart comparison made between Decision Tree and Random Forest algorithms

Figure 7 shows the bar chart comparison made between Decision Tree and Random Forest algorithms. In terms of performance score, each algorithm is able to attain an almost perfect performance of up to 100 on the x-axis, thus performing well on the current data set. The advantage in Random Forest is seen from the slightly higher score attained by the algorithm compared to that of Decision Tree, which is represented by the relatively larger bar for the algorithm in the horizontal bar chart.

Conclusion

Finally, it can be seen that in this case, there is a great utility of using supervised machine learning models like decision trees and random forests in helping with early breast cancer diagnosis. Both models demonstrate very good effectiveness when classifying malignant and benign tumors. However, it should be noted that the second model is slightly superior because it involves ensembling and can thus generalize better. Moreover, certain steps such as data preparation, feature selection, and visualization improve the outcomes obtained. Thus, a fairly reliable technique has been developed for helping with diagnosing breast cancer.

Acknowledgement

There is no grant or funding bodies to be acknowledged for preparing this paper

References

- Ashraf, F. B., Alam, S. M., & Sakib, S. M. (2024). Enhancing breast cancer classification via histopathological image analysis: Leveraging self-supervised contrastive learning and transfer learning. *Heliyon*, 10(2). <https://doi.org/10.1016/j.heliyon.2024.e24094>
- Gupta, A., Anand, R., Pandey, D., Sindhwani, N., Wairya, S., Pandey, B. K., & Sharma, M. (2021). Prediction of breast cancer using extremely randomized clustering forests (ERCF) technique: prediction of breast cancer. *International Journal of Distributed Systems and Technologies (IJDST)*, 12(4), 1-15. <https://doi.org/10.4018/IJDST.287859>
- Hridoy, R. H., Arni, A. D., Ghosh, S. K., Chakraborty, N. R., & Mahmud, I. (2024). Performance enhancement of machine learning algorithm for breast cancer diagnosis using hyperparameter optimization. *International Journal of Electrical & Computer Engineering (2088-8708)*, 14(2). <http://doi.org/10.11591/ijece.v14i2.pp2181-2190>
- Jalloul, R., Krishnappa, C. H., Agughasi, V. I., & Alkhatib, R. (2024). Enhancing early breast cancer detection with infrared thermography: a comparative evaluation of deep learning and machine learning models. *Technologies*, 13(1), 7 <https://doi.org/10.3390/technologies13010007>
- Kanber, B. M., Al Smadi, A., Noaman, N. F., Liu, B., Gou, S., & Alsmadi, M. K. (2024). LightGBM: A leading force in breast cancer diagnosis through machine learning and image processing. *IEEE Access*, 12, 39811-39832. <https://doi.org/10.1109/ACCESS.2024.3375755>
- Kang, H., Xu, Q., Chen, D., Ren, S., Xie, H., Wang, L., ... & Chen, X. (2024). Assessing the performance of fully supervised and weakly supervised learning in breast cancer histopathology. *Expert Systems with Applications*, 237, 121575. <https://doi.org/10.1016/j.eswa.2023.121575>
- Koç, İ., Tashan, W., Shayea, I., & Zhetpisbayeva, A. (2024). Breast cancer detection based on machine learning. In *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 1-6). IEEE. <https://doi.org/10.1109/CSNT60213.2024.10545785>
- Martín-Sánchez, J. C., Lunet, N., Gonzalez-Marron, A., Lidon-Moyano, C., Matilla-Santander, N., Cleries, R., ... & Martinez-Sanchez, J. M. (2018). Projections in breast and lung cancer mortality among women: a Bayesian analysis of 52 countries worldwide. *Cancer research*, 78(15), 4436-4442. <https://doi.org/https://doi.org/10.1158/0008-5472.can-18-0187>
- Torabi, M., Rasouli, A. H., Wu, Q. J., Cao, W., & Pourpanah, F. (2024). Self-supervised adversarial adaptation network for breast cancer detection. *Engineering applications of artificial intelligence*, 133, 108489. <https://doi.org/10.1016/j.engappai.2024.108489>
- Wen, Y., Ma, J., Xiang, Z., Lin, Y., Huang, Y., Liu, L., ... & Sun, Y. (2025). Trends of Female Breast Cancer Incidence, Mortality, and Survival in Fujian Province of China: 2011–2020 and Projection to 2025. *Cancer Medicine*, 14(13), e71033. <https://doi.org/10.1002/cam4.71033>