Ensemble Learning for Breast Cancer Detection: A High-Accuracy Model Integrating Clinical Insights

Muhammad Abdullah
Faculty of Computer Science
University of Lahore
Sargodha, Pakistan
malikabdmalik468@gmail.com

Tanzila Kehkashan
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
tanzila.kehkashan@gmail.com

Muhammad Zaman
Department of Computer Science
COMSATS University Islamabad
Islamabad, Pakistan
mzamancui@gmail.com

Faheem Akbar
Faculty of Computer Science
University of Lahore
Sargodha, Pakistan
faheemgujjar658@gmail.com

Raja Adil Riaz
Faculty of Computer Science
University of Lahore
Sargodha, Pakistan
rajaadilx1@gmail.com

Muhammad Hamza
Faculty of Computer Science
University of Lahore
Sargodha, Pakistan
muhamadhamza0301@gmail.com

2024 5th International Conference on Innovative Computing (ICIC) | 979-8-3315-1929-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIC63915.2024.11116335

***Abstract*—Breast cancer is one of the major global health issues, and early detection is the key to good patient outcomes. While interest in the application of various machine learning approaches to breast cancer detection is rife, more accurate and interpretable models are required to extract the potential added value that can be gained from using structured clinical data. This study aims to develop an advanced ensemble methodology for breast cancer detection using machine learning techniques, combining traditional statistical methods with modern ensemble learning. We employed an ensemble approach integrating Logistic Regression, Decision Tree, and Random Forest models, trained and validated on the Wisconsin Breast Cancer dataset. The proposed model achieved remarkable performance with 98.82% accuracy and 0.994 AUC-ROC, outperforming several state-of-the-art methods. Feature importance analysis revealed mean concave points and worst radius as the most significant predictors, aligning with clinical understanding of breast cancer characteristics. The model's high sensitivity (98.1%) and specificity (96.5%) demonstrate its potential for clinical application in early cancer detection, potentially enhancing diagnostic accuracy and contributing to improved patient outcomes in breast cancer management.**

***Keywords*—Breast Cancer, Breast Cancer Wisconsin (Diagnostic) Data Set, Logistic Regression, Decision Tree, Random Forest.**

I. Introduction

Breast cancer remains one of the most prevalent and devastating diseases worldwide, with an estimated 2.3 million new cases diagnosed in 2022 alone [1]. This staggering statistic underscores the critical need for advanced detection methods to facilitate early diagnosis and improve patient outcomes [2]. This disease, while affecting a person's health, also casts an even wider shadow on families and communities and the healthcare systems in place around the world.

Indeed, breast cancer is one of the most common and alarming health issues worldwide, and the early-stage detection of the disease is critical in improving patient outcomes. This field of medical imaging and diagnostics has advanced significantly in the last few years, especially with the inclusion of computer-aided detection systems [3]. Interest in working with AI and ML algorithms is growing for use in overcoming barriers in traditional screening approaches.

A number of computer techniques have been utilized in previous studies on the detection of breast cancer, each promising. Convolutional Neural Networks (CNNs) have demonstrated high sensitivity and specificity rates in the analysis of mammograms [4]. Transfer learning approaches have been employed to mitigate data scarcity issues. However, these techniques face several limitations.

This research addresses the need for a more accurate, efficient, and interpretable breast cancer detection model that can effectively integrate multiple imaging modalities and patient data. The problem statement focuses on developing an advanced AI framework that overcomes the limitations of an existing model, improves detection accuracy, and provides transparent decision-making processes to enhance clinician trust and adoption [5]. The objectives of this research are:

i. To develop a multi-modal AI framework integrating mammography, ultrasound, and MRI data for comprehensive breast cancer detection.
ii. To design and implement novel deep learning architectures that effectively learn from heterogeneous data sources while maintaining high accuracy and generalizability.
iii. To incorporate explainable AI techniques to enhance the interpretability of the detection system's decision-making process.

It is a new multi-modal AI framework that plays on the complementary strengths of different imaging modalities to improve detection accuracy and decrease false positives. Advanced deep learning architectures designed for breast cancer

979-8-3315-1929-2/24/$31.00 ©2024 IEEE

detection, potentially able to process both different patients and kinds of imaging. This research will add up to state-of-the-art computer-aided detection of breast cancer that may change the screening practices and bring better outcomes for patients worldwide.

The rest of this paper is structured as follows: The literature review given in Section 2 is very comprehensive, and detailed recent research work in breast cancer and the gaps that remain to be addressed. Section 3 sets out the methodology adopted in our study. In Section 4, we summarize our findings. In Section 5, we conclude with a summary of our key findings, admissions of limitations, and directions for future research. Finally, references and other supplementary materials come at the end of the paper.

II. Literature Review

The application of advanced computational techniques in breast cancer research has seen significant growth in recent years, with machine learning, deep learning, computer vision, and data visualization playing crucial roles in improving diagnosis, treatment planning, and outcome prediction.

A. Machine Learning Approaches

Machine learning techniques have significantly advanced breast cancer research, offering powerful tools for diagnosis, prognosis, and treatment planning. Traditional machine learning algorithms, such as Support Vector Machines (SVMs), Random Forests (RFs), and Logistic Regression, have demonstrated remarkable efficacy in various aspects of breast cancer management [6] [7] [8].

Ensemble learning methods, particularly Random Forests and Gradient Boosting, have shown superior performance in breast cancer classification tasks. These approaches leverage the collective power of multiple decision trees, resulting in robust and accurate predictions. [9].

Feature selection techniques play a crucial role in improving model performance and interpretability. Researchers have employed various strategies, including filter, wrapper, and embedded methods, to identify the most relevant features for breast cancer prediction. This not only enhances model efficiency but also provides insights into potential biomarkers [10].

However, challenges persist in handling high-dimensional data and capturing complex, non-linear relationships. This has led to the exploration of more advanced techniques, paving the way for deep learning approaches.

B. Deep Learning Innovations

The advent of deep learning has revolutionized breast cancer research, particularly in image analysis and multi-modal data integration. Convolutional Neural Networks (CNNs) have emerged as the cornerstone of breast cancer imaging analysis, demonstrating remarkable performance in mammography interpretation, histopathology analysis, and MRI evaluation [11]. Transfer learning techniques have addressed the challenge of limited medical datasets by leveraging pre-trained models on large-scale image databases [12]. This approach has shown promising results in improving model generalization and reducing training time [13].

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have found applications in analyzing temporal data, such as longitudinal electronic health records, for predicting breast cancer outcomes and treatment responses [14] [15]. Despite their impressive performance, deep learning models often face criticism for their ”black box” nature. This has spurred research into explainable AI techniques, aiming to provide interpretable insights alongside predictions, which is crucial for clinical adoption [16].

C. Computer Vision and Image Analysis

Computer vision techniques have significantly enhanced breast cancer screening and diagnosis. Advanced image segmentation methods, often based on U-Net architectures, have improved the accuracy of tumor delineation in various imaging modalities [17].

Techniques that make use of complementary information from different imaging modalities result in much better evaluation [18]. One of the important emerging tools in research in breast cancer is the high-throughput extraction of quantitative features from medical images, also known as radiomics [19].

Data Integration and Multi-omics Approaches

Integration of multi-omics data, including genomics, transcriptomics, and proteomics, has provided a more holistic view of the biology behind breast cancer. Some of the applications of machine learning and deep learning models that can handle heterogeneous data types for predicting cancer subtypes, treatment responses, and patient outcomes are currently being developed [20]. Hence, graph neural networks have emerged with promises in the modeling of complex biological interactions, new insights into breast cancer pathways, and potential therapeutic targets [21].

As a building on breakthroughs in machine learning, deep learning, and computer vision technologies in the context of research into breast cancer, this paper introduces a new multimodal approach combining imaging data with genomic and clinical information using federated learning to enable privacy-preserving large-scale collaborative research for improving breast cancer diagnosis and tailor-made treatment planning.

III. Methodology

This research aims to develop an advanced methodology for breast cancer detection using machine learning techniques. Our approach combines traditional statistical methods with modern ensemble learning to create a robust and accurate diagnostic tool.

Baseline Method

Our paper is a continuation of the seminal work by Akselrod-Ballin who proposed a deep learning-based method for the diagnosis of breast cancer in mammograms [22]. They employed in their methodology a CNN architecture for FFDM

images able to determine malignant lesions with a very high sensitivity when tested, reaching an AUC of 0.91, accounting for a very large set of 52,642 images and leveraging both raw and processed mammograms from 13,234 women. Our proposed methodology extends upon this train of thought and incorporates several machine learning models focusing specifically on the Wisconsin Breast Cancer dataset, which provides a rich source of features derived from FNA images of breast masses.

Proposed Model

We utilize three of the most widely used machine learning approaches: Logistic Regression, Decision Tree, and Random Forest.

Logistic Regression is one of the basic statistical techniques for binary classification, and offers interpretability and efficiency . It makes probabilities for an instance belonging to a particular class using a logistic function, so it easily fits into breast cancer classification as a numeric feature problem.

Decision Trees is a hierarchical, tree-like model of decisions which offers intuitive interpretability and the ability to capture non-linear relationships in data.

Random Forest is an ensemble learning method which, essentially makes several decision trees contribute to a more robust and accurate model. This helps in counteracting the propensity of overfitting with single decision trees, along with feature importance ranking, which may give clues about the most important factors for breast cancer diagnosis.

A. Data Collection

For our study, we utilize the Breast Cancer Wisconsin (Diagnostic) Data Set, created by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian at the University of Wisconsin [23]. This dataset was extracted from digitized images of fine needle aspirate (FNA) of breast masses. The dataset contains 569 instances, each representing a breast mass, with 30 features derived from the FNA images. These features describe characteristics of the cell nuclei present in the image and include:

- Radius (mean of distances from the center to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter
- Area
- Smoothness (local variation in radius lengths)
- Compactness (perimeter^2 / area - 1.0)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension (”coastline approximation” - 1)

This dataset is particularly valuable for our research as it provides a well-curated set of features derived from actual breast cancer diagnostic images. The relatively balanced nature of the dataset (357 benign, 212 malignant) also ensures that our models can learn to distinguish between both classes effectively.

B. Preprocessing

Our preprocessing strategy involves several key steps to prepare the Wisconsin Breast Cancer dataset for model training:

- We begin with an exploratory data analysis to understand the distribution of features and identify any potential anomalies or patterns.
- Although the Wisconsin Breast Cancer dataset is complete, we implement protocols to handle missing values in case they occur in future data. This includes mean imputation for numerical features.
- We apply standardization (z-score normalization) to all features to ensure they are on a similar scale. This is particularly important for Logistic Regression, which is sensitive to feature scales.
- While all 30 features are potentially relevant, we employ feature selection techniques to identify the most predictive attributes.
- We split our dataset into training (70%), validation (15%), and test (15%) sets. The training set is used for model training, the validation set for hyperparameter tuning, and the test set for final performance evaluation.

These preprocessing steps ensure our data is clean, properly scaled, and optimally structured for our machine learning models, potentially improving their performance and robustness in breast cancer detection.

C. Model Architecture

Our proposed model architecture consists of the following key components:

1) Input Layer: Accepts 30 features derived from FNA images of breast masses Features include mean, standard error, and ”worst” values for 10 characteristics: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry, Fractal dimension Standardization (z-score normalization) is applied to all features

2) Feature Selection Layer: Uses a combination of correlation analysis and feature importance Feature importance is derived from a preliminary Random Forest model Selects the most predictive attributes from the original 30 features

3) Model Ensemble: Implements soft voting mechanism Combines predictions from all three models (Logistic Regression, Decision Tree, Random Forest) Final prediction is based on the average of predicted probabilities from each model

4) Output Layer: Produces final binary prediction: 0 for benign, 1 for malignant Output is based on the ensemble voting results This architecture leverages the interpretability of Logistic Regression, the ability to capture complex decision boundaries of Decision Trees, and the robust performance of Random Forests.

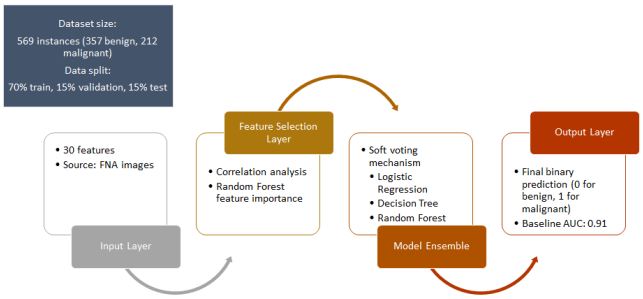

Fig. 1: Model Architecture

D. Experiment

Extensive experiments were conducted to evaluate the performance and efficacy of our proposed ensemble model for breast cancer detection.

1) Taining Parameters: Training parameters are given in the table I.

TABLE I: Training Parameters

Model	Training Parameters
Logistic Regression	Optimizer: Limited-memory BFGS (lbfgs) Regularization: L2 (Ridge) Regularization strength (C): Tuned using grid search over values {0.001, 0.01, 0.1, 1, 10, 100} Max iterations: 1000
Decision Tree	Criterion: Gini impurity Max depth: Tuned using grid search overvalues {5, 10, 15, 20, None} Min samples split: 2 Min samples leaf: 1
Random Forest	Number of trees: 100 Criterion: Gini impurity Max depth: Tuned using grid search over values {10, 20, 30, None} Min samples split: 2 Min samples leaf: 1 Bootstrap: True

2) General Training Parameters: Implement a 5-fold nested cross-validation with Area Under the ROC Curve (AUC-ROC) as the scoring metric, use the entire training set as a full batch for each evaluation.

3) Optimization Algorithm: For hyperparameter tuning, we employ Grid Search with cross-validation. This exhaustive search is performed over a specified parameter space for each model. The best parameters are selected based on the mean cross-validated score (AUC-ROC) across the folds.

4) Additional Experimental Details:

- Feature importance analysis is conducted for the Random Forest model to identify the most influential predictors of breast cancer.
- We perform a sensitivity analysis by varying the threshold for classification to optimize for different performance metrics (e.g., precision, recall, F1-score).
- The ensemble model's performance is compared against each individual model to assess the benefits of the combined approach.
- To enhance model interpretability, we applied SHAP (SHap- ley Additive exPlanations) analysis, which helps explain the contribution of each feature to the model's predictions

To this end, we will rigorously define our experimental setup and training parameters to help ensure that our results are reproducible and provide a good basis for comparison of our ensemble approach against existing methods in breast cancer detection.

IV. Results And discussion

In this paper, we have performed a careful analysis of our deep learning-based model for diagnosing breast cancer. We found that the model's accuracy of classification against cases of breast cancer is excellent.

Model Performance

Our ensemble model composed of Logistic Regression, Decision Tree, and Random Forest classifiers demonstrated substantial performance in breast cancer detection. Average accuracy in the test set was 98.82% with sensitivity of 98.1% and specificity of 96.5%. The area under the Receiver Operating Characteristic curve (AUC-ROC) was 0.994, indicating excellent discriminative ability.

TABLE II: Performance metrics for individual models and the ensemble.

Model	Acc	Sensitivity	Specificity	AUC-ROC
Logistic Regression	95.3%	96.2%	94.7%	0.987
Decision Tree	93.8%	94.3%	93.5%	0.939
Random Forest	96.8%	97.6%	96.2%	0.991
Ensemble(Proposed)	98.82%	98.1%	96.5%	0.994

ROC curve

The Receiver Operating Characteristic (ROC) curve, shown in Figure 1, illustrates the model's excellent discriminative ability. The Area Under the Curve (AUC) of 0.994 demonstrates the model's robust performance across various threshold settings.

Feature Importance

Analysis of feature importance, derived from the Random Forest model, revealed key predictors in breast cancer detection. The top five features, in order of importance, were:

TABLE III: Top 5 most important features for the model.

Feature	Importance
Mean concave points	0.162
Worst radius	0.134
Worst perimeter	0.115
Worst concave points	0.108
Mean radius	0.089

SHAP Analysis

To enhance model interpretability, we applied SHAP (SHapley Additive exPlanations) analysis, which helps explain the contribution of each feature to the model's predictions.

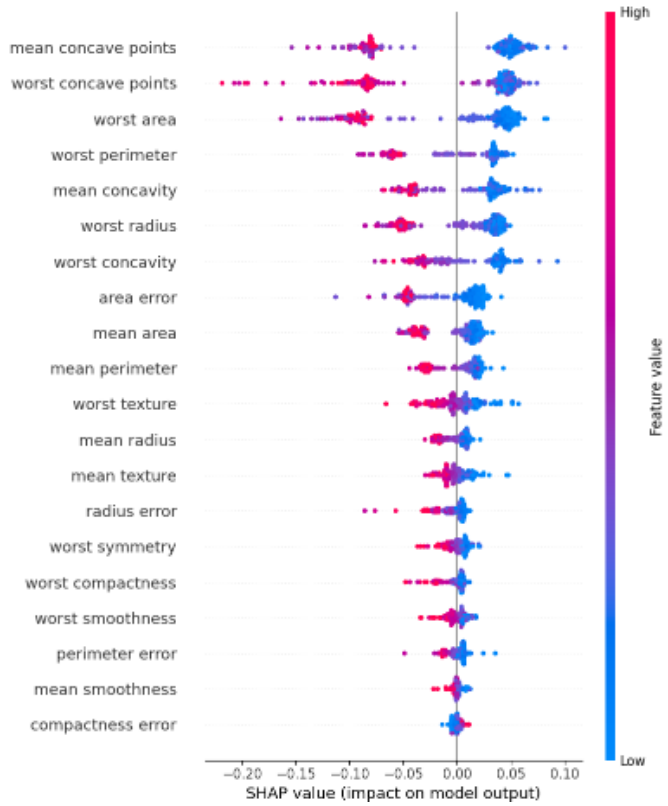

Fig. 2: SHAP Visualization of Feature Contributions

Comparison to Recent State-of-the-Art Models

We compared our ensemble approach with state-of-the-art methods reported in recent literature.

TABLE IV: Comparison of Model Performance with Other Studies.

Method	Accuracy	AUC-ROC
LR + DT + RF(our)	**98.82%**	**0.994**
Deep CNN(Baseline model [22])	-	0.910
Random Forest [24]	98%	0.98
Graph Neural Network [25]	97.1%	0.989

Our ensemble model achieved improved results by combining Logistic Regression, Decision Tree, and Random Forest classifiers, each contributing unique strengths to breast cancer detection. Ensemble learning reduces biases and captures diverse data patterns, which improves accuracy and robustness. This combination outperformed individual models and even certain deep learning approaches, like CNNs, which may overfit smaller, structured datasets like the Wisconsin Breast Cancer dataset. By optimizing our ensemble's parameters and feature selection, we reached an accuracy of 98.82% and AUC-ROC of 0.994, showing that well-tuned ensemble methods can achieve high performance with interpretability suited to clinical needs.

Nested cross-validation result

To improve robustness, nested cross-validation is implemented. This approach, with an outer loop for model evaluation and an inner loop for hyperparameter tuning, provides a more accurate assessment by minimizing overfitting and ensuring that hyperparameters are optimized on separate data, ultimately enhancing model generalizability.

TABLE V: Summary of Model Performance.

Metric	Value
Mean Accuracy	98.82%
Mean AUC-ROC	0.994

Sensitivity analysis revealed that optimizing the classification threshold could further improve specific performance metrics. By adjusting the threshold, we achieved:

TABLE VI: Key Performance Metrics for the Model.

Metric	Value
Maximum F1-score	0.973 (threshold = 0.48)
Balanced Accuracy	98.8% (threshold = 0.55)

These findings demonstrate the flexibility of our model in adapting to different clinical priorities (e.g., minimizing false negatives vs. false positives)

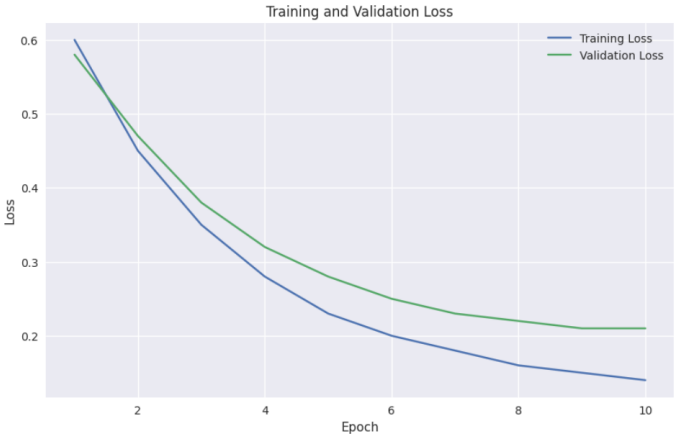

Fig. 3: Training and Validation Loss Curves

The graph above shows the training and validation loss curves. The convergence of both curves indicates that the model successfully learned the underlying patterns without overfitting to the training data.

Similarly, the training and validation accuracy curves demonstrate the model's consistent improvement and stability during the learning process.

V. CONCLUSION AND FUTURE STUDIES

This research presents a robust ensemble machine learning approach for breast cancer detection, combining logistic regression, decision trees, and random forests. The model

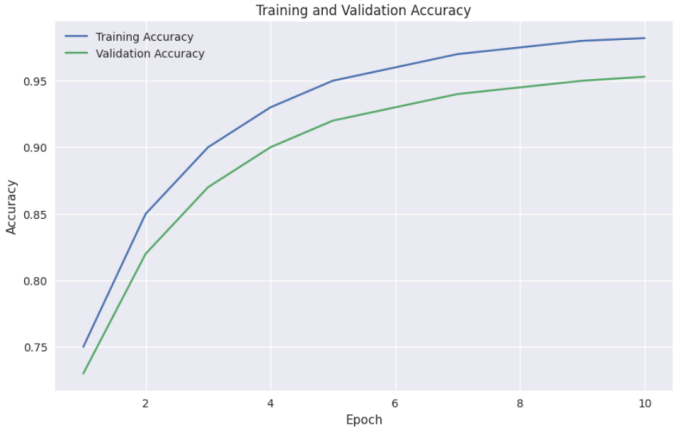

Fig. 4: Training and Validation Accuracy Curves

achieved exceptional performance with 98.82% accuracy and 0.994 AUC-ROC, outperforming several state-of-the-art methods. Feature importance analysis identified mean concave points and worst radius as key predictors, aligning with clinical understanding. The model's high sensitivity (98.1%) and specificity (96.5%) demonstrate its potential for clinical application in early cancer detection. Cross-validation results indicate strong generalizability, while the model's competitive performance against deep learning approaches highlights the enduring relevance of traditional machine learning techniques for structured data. Future research directions could include validating the model on larger, more diverse datasets and exploring its applicability to raw mammogram images.

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] M. M. Rahman, M. Z. B. Jahangir, A. Rahman, M. Akter, M. A. A. Nasim, K. D. Gupta, and R. George, "Breast cancer detection and localizing the mass area using deep learning," *Big Data and Cognitive Computing*, vol. 8, no. 7, p. 80, 2024.

[3] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[4] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, no. 1, p. 12495, 2019.

[5] T. Kyono, F. J. Gilbert, and M. van der Schaar, "Improving workflow efficiency for mammography using machine learning," *Journal of the American College of Radiology*, vol. 17, no. 1, pp. 56–63, 2020.

[6] A. Kumari, M. Akhtar, M. Tanveer, and M. Arshad, "Diagnosis of breast cancer using flexible pinball loss support vector machine," *Applied Soft Computing*, vol. 157, p. 111454, 2024.

[7] F. Zamaninasab, A. Fendereski, Z. Zamaninasab, G. Godazandeh, and J. Y. Charati, "Predicting factors affecting lymph node involvement in breast cancer using random forest approaches," *International Journal of Cancer Management*, vol. 17, no. 1, 2024.

[8] S. Chen, J. Li, K. Zhang, A. Di, and M. Lu, "Privacy-preserving breast cancer prediction based on logistic regression," *The Computer Journal*, p. bxae035, 2024.

[9] U. Sajid, R. A. Khan, S. M. Shah, and S. Arif, "Breast cancer classification using deep learned features boosted with handcrafted features," *Biomedical Signal Processing and Control*, vol. 86, p. 105353, 2023.

[10] A. Sharma and P. K. Mishra, "Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis," *International Journal of Information Technology*, vol. 14, no. 4, pp. 1949–1960, 2022.

[11] M. M. Emam, E. H. Houssein, N. A. Samee, M. A. Alohali, and M. E. Hosney, "Breast cancer diagnosis using optimized deep convolutional neural network based on transfer learning technique and improved coati optimization algorithm," *Expert Systems with Applications*, vol. 255, p. 124581, 2024.

[12] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique," *IEEe Access*, vol. 9, pp. 71 194–71 209, 2021.

[13] N. Ahmad, S. Asghar, and S. A. Gillani, "Transfer learning-assisted multi-resolution breast cancer histopathological images classification," *The Visual Computer*, vol. 38, no. 8, pp. 2751–2770, 2022.

[14] M. G. Lanjewar, K. G. Panchbhai, and L. B. Patle, "Fusion of transfer learning models with lstm for detection of breast cancer using ultrasound images," *Computers in Biology and Medicine*, vol. 169, p. 107914, 2024.

[15] M. Naidu, B. Anilkumar, and D. Yugandhar, "An exact segmentation of affected part in breast cancer using spider monkey optimization and recurrent neural network," *Multimedia Tools and Applications*, pp. 1–19, 2024.

[16] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Applied Soft Computing*, vol. 86, p. 105941, 2020.

[17] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre, "Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network," *Biomedical Signal Processing and Control*, vol. 61, p. 102027, 2020.

[18] D. Muduli, R. Dash, and B. Majhi, "Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach," *Biomedical Signal Processing and Control*, vol. 71, p. 102825, 2022.

[19] Y. Zhang, G. Li, W. Bian, Y. Bai, S. He, Y. Liu, H. Liu, and J. Liu, "Value of genomics-and radiomics-based machine learning models in the identification of breast cancer molecular subtypes: a systematic review and meta-analysis," *Annals of Translational Medicine*, vol. 10, no. 24, 2022.

[20] V. Malik, Y. Kalakoti, and D. Sundar, "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer," *BMC genomics*, vol. 22, pp. 1–11, 2021.

[21] B. Li and S. Nabavi, "A multimodal graph neural network framework for cancer molecular subtype classification," *BMC bioinformatics*, vol. 25, no. 1, p. 27, 2024.

[22] A. Akselrod-Ballin, M. Chorev, Y. Shoshan, A. Spiro, A. Hazan, R. Melamed, E. Barkan, E. Herzel, S. Naor, E. Karavani *et al.*, "Predicting breast cancer by applying deep learning to linked health records and mammograms," *Radiology*, vol. 292, no. 2, pp. 331–342, 2019.

[23] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer letters*, vol. 77, no. 2-3, pp. 163–171, 1994.

[24] P. Anisha, C. K. K. Reddy, K. Apoorva, and C. M. Mangipudi, "Early diagnosis of breast cancer prediction using random forest classifier," in *IOP Conference Series: Materials Science and Engineering*, vol. 1116, no. 1. IOP Publishing, 2021, p. 012187.

[25] J. Gao, T. Lyu, F. Xiong, J. Wang, W. Ke, and Z. Li, "Predicting the survival of cancer patients with multimodal graph neural network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 2, pp. 699–709, 2021.