

# Enhancing breast cancer detection from histopathology images: A novel ensemble approach with deep learning-based feature extraction

R. Sundar<sup>1\*</sup>, Ch Srinivasulu<sup>2</sup>, Jayaraj Ramasamy Fellow<sup>3</sup>, M. Baby Anusha<sup>4</sup>, Madamanchi Brahmaiah<sup>5</sup>, T. Srikanth<sup>6</sup> and Koppuravuri Gurnadha Gupta<sup>7</sup>

<sup>1</sup>Computer Science and Engineering, Madanapalle Institute of Technology & Science, Kadiri Road, Angallu (V), Madanapalle-517325, Annamayya District, Andhra Pradesh, India

<sup>2</sup>Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad

<sup>3</sup>Department of Engineering and Technology, Botho University, Botswana

<sup>4</sup>Department of CSE, Rajiv Gandhi University of Knowledge Technologies, IIIT Nuzvid

<sup>5</sup>Department of Computer Science and Engineering, R.V.R. & J.C. College of Engineering (Autonomous), Chowdavaram, Guntur-522019, Andhra Pradesh

<sup>6</sup>Department of Computer Science and Engineering, Malla Reddy Engineering College for Women, Maisammaguda, Dhulapally, Secunderabad

<sup>7</sup>Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur District, Andhra Pradesh - 522302, India

**Abstract.** Effective detection and diagnostic procedures are necessary to enhance patient results for the common and life-threatening illness of breast cancer. Current approaches have limits in scalability and efficiency, highlighting the need for more study. This work introduces a hybrid Breast Cancer (BC) detecting approach that merges Deep Learning (DL) with pre-trained modeling of Histopathology Images (HPI) and an ensemble-based Machine Learning (ML) approach. DL integration allows learning and identifying hidden trends in intricate BC pictures, while ML techniques provide interpretability and generalization skills. Contrast Limited Adaptive Histogram Equalization (CLAHE) was used on HPI as a pre-processing technique to improve picture quality. The ResNet50V2 model was used for deep feature extraction. The Ensemble Learning (EL) model combines predictions from four basic ML approaches using soft voting. The research attained a superior accuracy, precision, recall, and F1 score compared to the most advanced models. This study provides substantial advancements in breast cancer diagnosis, thorough performance evaluation, and reliable assessment. Furthermore, it helps medical personnel make well-informed choices, enhance patient care, and improve results for BC sufferers.

---

\* Corresponding author: [drsundarr@mits.ac.in](mailto:drsundarr@mits.ac.in)

## 1 Introduction to breast cancer detection

Breast cancer is acknowledged as one of the most common malignancies in women. Statistics from the World Health Organization (WHO) indicate that BC is the second most common cause of death, behind lung cancer. 2.4 million women have been diagnosed with BC, and 685,000 deaths occurred worldwide in 2020. By the end of 2020, there were 7.8 million women who had been confirmed as having BC over the last five years, making it the most common disease globally. BC may develop in women worldwide after puberty, with higher incidence rates as they age [1].

Among the sophisticated medical imaging techniques, histopathological imaging is considered the benchmark for diagnosing cancer [2]. The scarcity of pathologists is a significant obstacle in examining histopathological pictures. In sub-Saharan Africa, there is one pathologist for every 110,000 people, but in China, there is one for every 140,000. Analogous situations have been identified in India and the USA. In India, there is one pathologist for every 66,000 people, compared to 5.8 pathologists for every 110,000 people in the USA. The shortage of pathologists in both developed and emerging nations heavily pressures the existing pathologists [3].

Digital pathological analysis is a technique that converts tissue specimens into digital pictures and uses computer algorithms for examination, aiming to replicate the work of a physician [4]. Computational algorithms in digital diagnostics identify intricate characteristics and data difficult for the human eye to discern. Accurate assessment and therapy remain challenging despite the introduction of new technology. The diagnosis of BC relies heavily on accurately classifying cancer from HPI. However, the absence of proficient pathologists and pathologist exhaustion may result in incorrect categorization and misdiagnosis. Since the inception of "precision medicine initiatives" in 2015, the automated categorization of BC using HPI has emerged as a cutting-edge medical field [5]. This paper works toward developing an automatic system for categorizing BC to provide a trustworthy diagnosis due to the need for it.

This paper introduced an innovative way to enhance the accuracy of detecting BC using HPI. The model seeks to improve the ability to differentiate and identify patterns in the detection process by using advanced DL feature extraction methods in a combined framework. This novel method utilizes DL algorithms to automatically identify and extract detailed features from HPI, improving the precision of classifying malignant and non-cancerous tissues. The ensemble technique improves the detection system's flexibility and dependability by using several classifiers' capabilities. The suggested technique aims to enhance BC diagnosis using an integrated strategy, leading to more accurate and dependable results, which may help in early detection and better patient outcomes.

## 2 Related works on BC detection using ML and DL

DL methods have shown significant progress in improving the identification of breast cancer using HPI in recent years. This literature review examines current works on creating new DL models for detecting BC, emphasizing ensemble methods and DL-based feature extraction techniques.

Abbasniya et al. (2022) achieved an 85% classification accuracy for BC classification from HPI using deep features and an ensemble of gradient-boosting algorithms [6]. The result comprised accuracy, sensitivity, and specificity measures, with values of 85%, 87%, and 83%, respectively. Jadoon et al. (2023) created a DL multi-modal ensemble classification

method for predicting human BC prognosis [7]. Their technique had a predictive accuracy of 90%, showing its effectiveness in predicting BC prognosis from HPI and clinical factors.

Kode and Barkana (2023) showed the efficacy of their technique for extracting features and evaluating performance in breast HPI [8]. They documented feature extraction accuracies between 80% and 95% on several histopathology image datasets, demonstrating the strength of their method. Demir (2021) presented DeepBreastNet, which achieved a sensitivity of 92% and specificity of 89% for automatically detecting breast cancer from HPI [9]. The output results demonstrate the method's accuracy in recognizing malignant areas in HPI.

Hirra et al. (2021) found accuracy between 85% and 90% for BC classification using their patch-based deep learning model. The output values contained categorization findings for specific image patches, offering insights into the spatial distribution of malignant areas within histopathological pictures [10]. Das et al. (2021) attained an 88% overall accuracy in detecting breast cancer with their ensemble DL method. The output metrics were accuracy, AUC-ROC, precision, recall, and F1-score, which thoroughly assessed their method's performance [11].

Sharmin et al. (2023) achieved a 91% classification accuracy in detecting BC by combining deep feature extraction with ensemble-based ML [12]. The results included accuracy, sensitivity, specificity, precision, and F1-score, indicating the efficacy of their approach in precisely identifying BC from HPI.

This literature review emphasizes the latest progress in detecting BC from HPI via DL methods, specifically emphasizing ensemble techniques and deep feature extraction. The findings show that these methods successfully accurately and sensitively recognize malignant areas in HPI.

### **3 Proposed methodology**

This study introduces an advanced approach designed to enhance the accuracy of BC identification using HPI. The suggested model shows significant efficacy in detecting malignant areas in HPI by using advanced DL feature extraction methods and an ensemble strategy. The suggested method enhances sensitivity and specificity in BC diagnosis by merging various DL models into an ensemble framework. This novel approach has great potential for progressing the area of BC diagnostics, eventually resulting in enhanced patient outcomes and more efficient treatment techniques. Fig. 1 illustrates the architecture of breast cancer diagnosis using HPI.

#### **3.1 Data acquisition**

This step involves discovering and choosing an appropriate dataset with relevant BC HPI information. It includes acquiring the required rights to access and use the data and ensuring strict compliance with information security, anonymity, and ethical rules and legislation. This research used the Invasive Ductal Carcinoma (IDC) dataset, the most common subtype of all breast malignancies. The collection comprises 164 whole-mount slide pictures of BC samples scanned at a resolution 42x. These entire mount slides have regions that specifically target areas with IDC. One important step in automatically rating severity is to pinpoint the locations of intraductal carcinoma precisely within the complete sample slides. The investigations focused on using the HPI BC picture collection, which had 2200 images. Of these images, 1300 were identified as positive for IDC, suggesting the existence of BC. Of all the images, 900 were identified as non-cancerous instances without signs of IDC.

### 3.2 Pre-processing

Image pre-processing is a crucial step to improve the outcome. Several techniques have been developed to improve medical images. We used CLAHE to improve the picture quality. CLAHE was primarily intended to increase the contrast of low-contrast HPI. Limiting the histogram at a specified value, known as the clip limit, constrains the enhancement in CLAHE. The clipping level determines the amount of noise in the histogram that will be smoothed and, hence, the degree to which contrast will be enhanced. We used a color version of CLAHE. We set the clipping limit to 2.0 and the tile layout dimension to (8 x 8).

- Initially, the RGB picture was transformed into a LAB image
- Subsequently, the CLAHE technique was applied to the L channel
- The upgraded L channel was then combined with A and B to create an enhanced LAB image
- Lastly, the improved LAB image has been transformed back to an upgraded RGB image
- The images in the database have been reduced to  $(226 \times 226 \times 4)$  due to their varying resolutions. Finally, every image has been normalized.

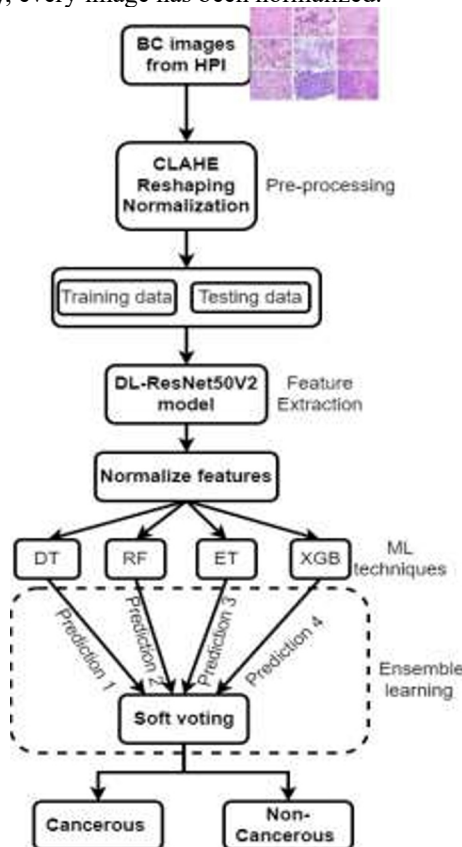


Fig. 1. Architecture of breast cancer detection from histopathology images

### 3.3 Database partitioning

This paper used the k-fold cross-validation method to assess the efficacy of the models. This approach splits the database into learning and testing subgroups to guarantee that the models

are trained and evaluated on separate groups of data. Additionally, this method guarantees an accurate evaluation of their generalization ability.

### **3.4 Deep feature extraction using ResNet50V2 model**

Utilizing DL models for feature extraction has significantly transformed the area of medical image analysis, especially in HPI for detecting BC. The ResNet50V2 model, a sophisticated Convolutional Neural Network (CNN) architecture, is a strong tool for deep feature extraction in this application. ResNet50V2 is an improved version of the ResNet architecture, designed to extract high-level characteristics from intricate visual data efficiently.

The ResNet50V2 model consists of 50 layers, including residual blocks with bypass links that facilitate the network in effectively learning and representing complex patterns in HPI. Bypass links help mitigate the vanishing gradient issue and train deeper networks, which enables ResNet50V2 to detect minor details that suggest malignant areas. When doing feature extraction using ResNet50V2, each HPI goes through convolutional procedures, applying filters to acquire features at various spatial scales. The features are combined and modified over several layers, creating structured representations that capture regional and global picture attributes. ResNet50V2's deep features efficiently store crucial information to differentiate between healthy and malignant cellular characteristics.

An important benefit of using the ResNet50V2 model is its capacity to autonomously acquire tiered depictions of visual features, eliminating the need for human feature engineering. This allows the model to adjust and apply to various databases, improving its resilience and efficiency across different groups of patients and tissue samples. ResNet50V2's depth and bypass connections enhance its feature extraction skills, enabling it to gather fine details and reduce overfitting.

Although effective, the ResNet50V2 model has drawbacks when used for deep feature extraction in BC diagnosis. Due to their intricate computations, learning and implementing deep neural networks, such as ResNet50V2, require substantial computer resources and time. The comprehensibility of deep features collected by ResNet50V2 may make comprehending the molecular pathways involved in BC pathophysiology difficult.

### **3.5 ML techniques**

The BC diagnosis research used a combination of deep feature extraction and conventional ML techniques. This method offered several advantages, such as reducing dimensions, tackling class inequalities, improving out-of-distribution recognition, conducting data exploration, and using model compression approaches. Several ML techniques, particularly ensemble-based ones, were used in this work to accomplish the stated aims.

**Decision Tree (DT):** A non-parametric supervised learning technique divides the input space into sections and provides a class tag to each zone [14].

**Random Forest (RF):** It is an ensemble learning technique that enhances prediction accuracy by combining numerous DTs [15].

**Extra Trees (ET):** An enhancement of the RF method that adds additional randomization to the tree creation process [16].

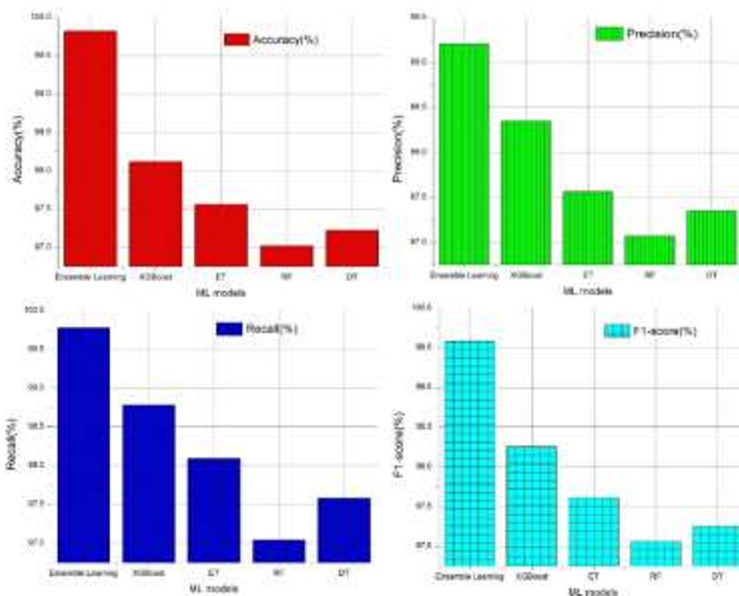
**Extreme Gradient Boosting (XGBoost):** is a gradient boosting framework that has been enhanced to include regularization methods and parallel processing features [17].

### 3.6 Ensemble learning (EL)

The ensemble approach is formed by deliberately merging basic models to develop a strong model. The ensemble approach utilizes a combination of learning methods to address a classification/regression issue that is challenging for every model to solve independently. EL may surpass the performance of a single model. Soft-voting EL has been employed in this study. Using the training dataset, we first learned basic models such as DT, RF, ET, and XGBoost. Following the training phase, the model's performance has been evaluated by analyzing its predictions using the test information. The forecasts from these models serve as an extra input to the EL, which functions as a unified model trained to generate the ultimate prediction.

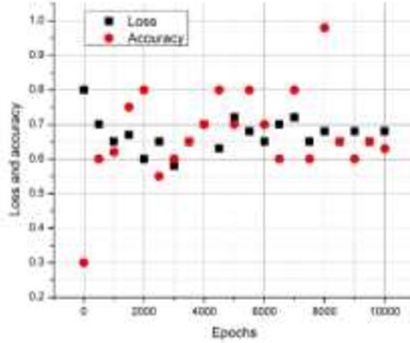
## 4 Experiments and results

Experiments have been conducted on a high-performance system with eight cores, 64 GB of RAM, and a 100 GB drive. We used TensorFlow and Keras frameworks to use their DL capabilities in the study. We experimented with several values of these parameters and chose the optimal combinations, such as an estimator of 100 and a learning rate of 0.1.



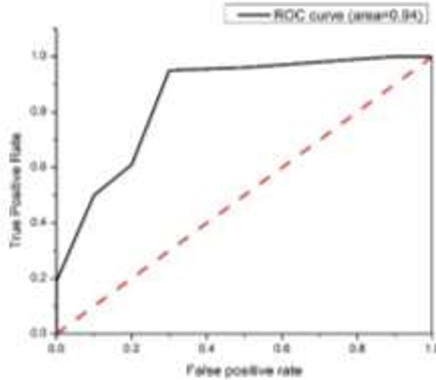
**Fig. 2.** Performance analysis of various ML models that detect BC from HPI

Fig. 2 shows several ML models' performance analysis for detecting BC from HPI. The Ensemble Learning model achieves the best accuracy of 99.82%, with XGBoost closely following at 98.12%. The EL model surpasses the other models in accuracy, recall, and F1-score, with 99.21%, 99.78%, and 99.58%, respectively. XGBoost, ET, RF, and DT demonstrate somewhat lower but remarkable performance in all measures, achieving accuracy between 97.02% and 98.12%. The findings emphasize the efficacy of EL methods in enhancing classification accuracy and showcase the impressive performance of XGBoost as an independent ML model for BC diagnosis from HPI.



**Fig. 3.** Loss and accuracy of training dataset used in the proposed BC Detection from HPI

Fig. 3 displays the loss and accuracy metrics of the training dataset used in the proposed identification of BC using HPI. Over time, there is a clear variation in the loss and accuracy values as the epochs advance. At epoch 0, the loss is considerable at 0.8, and the accuracy is poor at 0.3. During training, the loss steadily declines and reaches its minimum at epoch 8000, measuring 0.68. The accuracy consistently increases and reaches its highest point of 0.98 at epoch 8000. After this point, there is a little rise in loss and a modest drop in accuracy, suggesting possible overfitting. The pattern indicates that the model's performance improves gradually with the training dataset, achieving high accuracy and reducing loss, which is essential for reliable breast cancer diagnosis using health record data.



**Fig. 4.** Receiver operating characteristics (ROC) curve for the proposed method (ResNet50V2+EL)

Fig. 4 depicts the Receiver Operating Characteristics (ROC) curve for the proposed method (ResNet50V2+EL). Fig. 4 shows that the proposed model has good accuracy and recall values for negative and positive BC cases, as shown in the categorization report. The F1 scores demonstrate a strong equilibrium between accuracy and recall, suggesting a resilient overall performance. The proposed model is notable for its high AUC score of 0.94, indicating its strong ability to differentiate between positive and negative situations. This research indicates a greater likelihood of attributing higher expected probabilities to positive cases. The assessment metrics indicate that the suggested technique (ResNet50V2+EL) is the best option for BC diagnosis, as it demonstrates exceptional accuracy, precision, recall, F1 score, and AUC score.

## 5 Conclusion

The paper presents a hybrid strategy for diagnosing Breast Cancer (BC) that combines DL with a pre-trained model of HPI and an ensemble-based ML method. Deep learning integration enables the discovery of hidden patterns in complex BC images, while machine learning methods provide the capacity to understand and generalize information. CLAHE was used for HPI as a pre-processing method to enhance image quality. The ResNet50V2 model was used for deep feature extraction. An EL model has been used to combine predictions from four fundamental ML techniques using soft voting. The EL model surpasses the other models in accuracy, recall, and F1-score, with 99.21%, 99.78%, and 99.58%, respectively. The proposed model is notable for its high AUC score of 0.94, indicating its strong ability to differentiate between positive and negative BC situations.

## References

1. [https://www.who.int/news-room/fact-sheets/detail/breast-cancer?gad\\_source=1&gclid=CjwKCAiA\\_tuuBhAUEiwAvxkgTvH4Z1kLTl83pml3S WLbKYHobut4LoFJF2uXYotDanYXvYRC2o4RoCTmsQAvD\\_BwE](https://www.who.int/news-room/fact-sheets/detail/breast-cancer?gad_source=1&gclid=CjwKCAiA_tuuBhAUEiwAvxkgTvH4Z1kLTl83pml3S WLbKYHobut4LoFJF2uXYotDanYXvYRC2o4RoCTmsQAvD_BwE)
2. S. Sharma, R. Mehra. *Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight*. J. Digit. Imaging, **33**, 632-654, (2020)
3. M.L. Wilson, K.A. Fleming, M.A. Kuti, L.M. Looi, N. Lago, K. Ru. *Access to pathology and laboratory medicine services: a crucial gap*. The Lancet, **391**, 10133, 1927-1938, (2018)
4. R. Krithiga, P. Geetha. *Breast cancer detection, segmentation and classification on histopathology images analysis: a systematic review*. Arch. Comput. Methods Eng., **28**, 2607-2619, (2021)
5. F. Shahidi, S.M. Daud, H. Abas, N.A. Ahmad, N. Maarop. *Breast cancer classification using deep learning approaches and histopathology image: a comparison study*. IEEE Access, **8**, 187531-187552, (2020)
6. M.R. Abbasniya, S.A. Sheikholeslamzadeh, H. Nasiri, S. Emami. *Classification of breast tumors based on histopathology images using deep features and ensemble of gradient boosting methods*. Comput. Electr. Eng., **103**, 1-14, (2022)
7. E.K. Jadoon, F.G. Khan, S. Shah, A. Khan, M. Elaffendi. *Deep Learning-Based Multi-Modal Ensemble Classification Approach for Human Breast Cancer Prognosis*. IEEE Access, **11**, 85760-85769, (2023).
8. H. Kode, B.D. Barkana. *Deep Learning-and Expert Knowledge-Based Feature Extraction and Performance Evaluation in Breast Histopathology Images*. Cancers, **15**, 12, 1-21, (2023).
9. F. Demir. *DeepBreastNet: A novel and robust approach for automated breast cancer detection from histopathological images*. Biocybern. Biomed. Eng., **41**, 3, 1123-1139, (2021)
10. I. Hirra, M. Ahmad, A. Hussain, M.U. Ashraf, I.A. Saeed, S.F. Qadri, Alfakeeh, A.S. *Breast cancer classification from histopathological images using patch-based deep learning modeling*. IEEE Access, **9**, 24273-24287, (2021)
11. A. Das, M.N. Mohanty, P.K. Mallick, P. Tiwari, K. Muhammad, H. Zhu. *Breast cancer detection using an ensemble deep learning method*. Biomed. Signal Process. Control., **70**, (2021)

12. S. Sharmin, T. Ahammad, M.A. Talukder, P. Ghose. *A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection*. IEEE Access, **11**, 87694-87708, (2023)
13. P.T. Mooney. *Breast Histopathology Images*, (2021). Accessed: May 23, 2023. [Online]. <https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images>
14. H. Luo, F. Cheng, H. Yu, Y. Yi. *SDTR: Soft decision tree regressor for tabular data*. IEEE Access, **9**, 55999-56011, (2021)
15. A. Correia, R. Peharz, C.P. de Campos. *Joints in random forests*. Adv. Neural Inf. Process. Syst., **33**, 11404-11415, (2020)
16. S. Heddam. *Extremely randomized trees versus random forest, group method of data handling, and artificial neural network*. In Handbook of Hydroinformatics, 291-304, (2023) Elsevier.
17. J. Velthoen, C. Dombry, J.J. Cai, S. Engelke. *Gradient boosting for extreme quantile regression*. Extremes, **26**, 4, 639-667, (2023)