Deep Learning Approaches for Multi-Class Classification of Breast Cancer Images

Sharon G, Konkathi Rithin Kumar, Anupama P Biddargaddi, Achyut Padaki, Reemak Dawe, Meena S M
School of Computer Science and Engineering
KLE Technological University
Hubballi, India
sharonhere777@gmail.com, rithinkumar.k111@gmail.com, anupamab@gmail.com,
achyutpadaki@gmail.com, dawereemak@gmail.com, msm@kletech.ac.in

***Abstract*—Breast cancer, a complex health concern for women, relies on histopathological images for diagnosis. The classification of these images is pivotal in clinical treatment and computer-aided diagnosis. This study addresses the intrinsic challenge of imbalanced class distribution in such image data by employing resampling techniques, including Synthetic Minority Oversampling Technique (SMOTE) for minority class augmentation and Tomek Links elimination for refining decision boundaries, to achieve dataset balance. In the realm of medical image classification, our approach leverages state-of-the-art deep learning models, such as CNN, VGG-19, and DenseNet-121, for multi-class classification. Training on both imbalanced and balanced datasets facilitates a comprehensive comparison of model performance. To further enhance prediction accuracy, diverse ensembling techniques, including average and weighted average ensemble, are employed. Notably, our approach, encompassing CNN, VGG-19, DenseNet-121, Average Ensemble, and Weighted Average Ensemble, attains remarkable accuracy scores, with the Weighted Average Ensemble achieving a notable accuracy score of 93.20%.**

***Index Terms*—Deep Learning, Ensemble Learning, Medical Image Classification.**

I. Introduction

Breast cancer, the most prevalent global malignancy in women, claims one in three lives, according to the World Health Organization (WHO) [1]. Essential diagnostic tools, including mammography, MRI, and pathology tests, play a pivotal role in early detection. Histopathology images, considered the gold standard, offer comprehensive diagnostic information, particularly following mammograms [2]. The creation of digital histopathological images from breast cancer tissues involves the application of hematoxylin and eosin staining by laboratory technicians, capturing intricate details of cellular structures and abnormalities [3]. While these images provide valuable insights, their large size poses challenges, leading current research to explore the transformative potential of deep learning for improved diagnostic accuracy [4]. This study uniquely delves into deep learning applications in breast cancer histopathological image classification, rigorously addressing challenges and proposing innovative solutions to enhance diagnostic precision.

This research is dedicated to harnessing the power of deep learning for the nuanced analysis of breast cancer histopathological images. As the volume of digital pathology data grows, conventional diagnostic approaches face limitations, necessitating a transformative shift towards advanced methodologies. In addition to navigating the intricacies of breast cancer pathology through deep learning algorithms, notably convolutional neural networks (CNNs), our study places a significant emphasis on mitigating the challenges posed by imbalanced datasets. Recognizing the importance of balancing techniques, such as Synthetic Minority Oversampling Technique (SMOTE) and Tomek Links elimination, we seek to ensure a more equitable representation of diverse cases. Furthermore, our investigation incorporates ensemble learning techniques to enhance model robustness and predictive accuracy. By addressing these unique challenges in large-scale histopathological datasets, our research aims to contribute to the continuous advancement of both the accuracy and efficiency of breast cancer diagnosis through innovative technological applications.

The upcoming section of the paper is organized as follows: Section II provides a comprehensive review of previous studies on machine learning applications in breast cancer image classification and explores diverse deep learning model architectures. Section III outlines the methodology, covering dataset specifications, balancing methods, and ensembling approaches. Section IV delves into the results yielded by the proposed system which encompasses balancing the dataset using SMOTE and Tomek Links, analysis of individual model performance, comparative model analysis, and ensembling outcomes. Section V concludes the study and outlines future avenues for further research.

II. Background Study

In this section, we explore the prior research on the categorization of breast cancer images through machine learning and deep learning approaches, along with an in-depth examination of the classifiers utilized in our study, namely CNN, VGG-19, and DenseNet-121.

979-8-3503-8343-0/24/$31.00

A. Related Work

In this section, we discuss the works related to breast cancer image classification. The article [5] emphasizes neural networks, transfer learning, and CNN tuning for breast cancer image classification. The authors of [6] utilized a CNN model based on handcrafted features for breast cancer image classification. In [7], researchers conducted a comprehensive analysis, comparing the performance of various CNN models, including CaffeNet, an AlexNet variant, GoogleNet, and ResNet50. The study emphasized the significance of data augmentation coupled with fine-tuning. Exploring deep learning capabilities, [8] employed a transfer learning strategy using a pre-trained AlexNet, extracting features through DeCAF [9]. Meanwhile, [10] adopted a pre-trained CNN as a feature extractor, fine-tuning select final layers for classification. In [11], efforts were made to enhance the robustness and generalization of deep learning-based transfer learning models, particularly to address challenges associated with data acquisition and annotation in medical image analysis. The articles [12], [13] introduces a novel classification method for weak medical data, addressing the limitations of deep learning models. Unlike conventional recognition-based approaches, this method integrates complex images into ensemble learning, excelling over four alternatives in extensive testing on medical datasets. The results underscore the method's superiority, particularly in handling complex medical images, highlighting the potential of deep learning to advance research in clinical settings. The study [14] assesses different machine learning techniques using the Wisconsin Cancer Dataset. Notably, ANN achieves an accuracy of 98.57%, emphasizing its significance in early cancer detection. The primary goal of this research, as stated in [15] is to compare several machine learning techniques (XGBoost, Random Forest, etc.) for breast cancer detection, aiming for early diagnosis. Finally, [16], [17] proposed a CNN-based general framework for learning features in breast cancer histopathology images, incorporating magnification-independent classification.

B. Classifiers

Deep learning models that are used as classifiers are CNN, VGG-19, and DenseNet-121. The softmax activation function is used in the output layers of these models for multi-class classification which transforms raw outputs into interpretable probabilities. It is given by the equation 1 where $P(class_i)$ denotes the probability for class i, e is Euler's number (exponential component), z_i represents the raw score or logit for class i, and N signifies the total number of classes, influencing the normalization process. Softmax is pivotal in normalizing raw outputs, generating a probability distribution for clarity in multi-class classification, and guiding the model's decision-making process.

$$P(class_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{1}$$

1) CNN: The CNN model employs a series of convolutional layers (ranging from 32 to 128 filters) for effective feature extraction, incorporating ReLU activation and batch normalization. To prevent overfitting, max-pooling with a 2x2 window is applied, complemented by dropout layers to enhance generalization. The architectural design concludes with a dense layer featuring 128 neurons and a softmax activation function tailored for eight-class classification. As depicted in Fig. 1, this deep CNN model is specifically crafted for image classification, proficiently extracting hierarchical representations to make predictions across diverse classes.

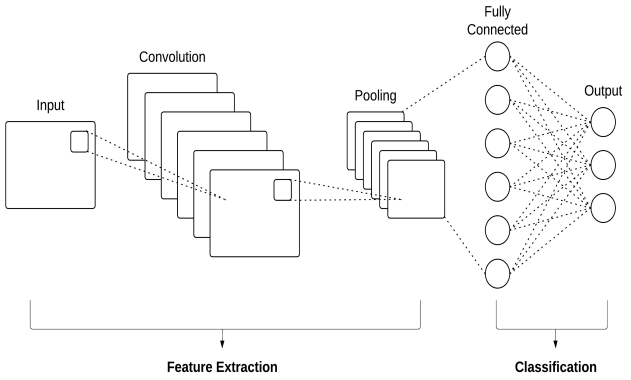

Fig. 1. CNN architecture

2) VGG-19: The application of VGG-19 involves transfer learning, initializing with pre-trained ImageNet weights. This process excludes the initial layers, primarily utilized for feature extraction. The model incorporates additional layers to construct a customized classification head, comprising Flatten, fully connected layers, ReLU activation, dropout, and batch normalization. The ultimate dense layer, featuring softmax activation, generates class probabilities for an eight-class classification task. The model is compiled using the RMSprop optimizer and categorical cross-entropy loss, adapting the strategy to the specific classification requirements. Fig. 2 provides a visual representation of the architecture.

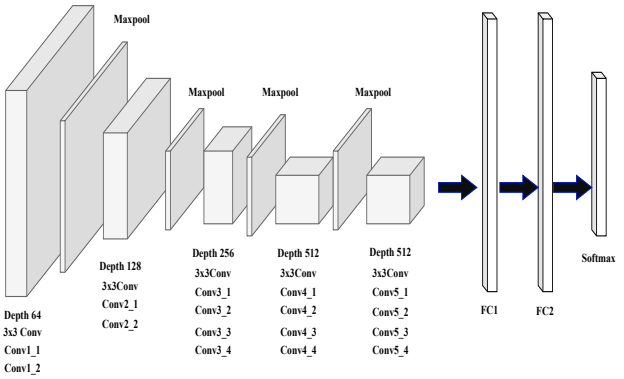

Fig. 2. VGG-19 architecture

3) DenseNet-121: DenseNet-121 employs transfer learning with ImageNet pre-training for eight-class classification, incorporating pre-trained weights, custom layers (Flatten, Dense with ReLU activation, Dropout), and batch normalization. The final Dense layer uses softmax activation, and the model is compiled with Adam optimizer and categorical cross-entropy loss to maximize accuracy. The training involves progressive unfreezing of layers for fine-tuning lower-level representations tailored to the eight-class problem. Model evaluation on a test set assesses overall performance and generalization beyond training and validation data. Fig. 3 illustrates the architectural configuration, combining transfer learning with fine-tuned top-layer adjustments for the specific classification task.

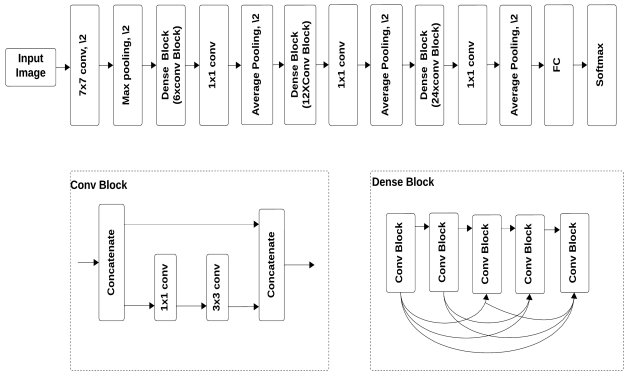

Fig. 3. DenseNet-121 architecture

III. Methodology

This portion encompasses the proposed system, features of the dataset, approaches to achieving balance, and ensemble techniques namely average ensemble and weighted average ensemble.

The proposed methodology of the system involves three key steps: balancing data through resampling, training classification models (CNN, VGG-19, DenseNet-121) on both balanced and unbalanced data, and ensembling all the models. Fig. 4 illustrates the proposed methodology of the system.

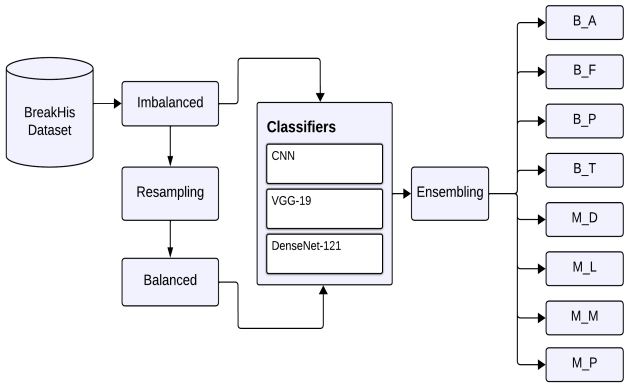

Fig. 4. Proposed methodology of the system

A. Dataset Details

The BreakHis dataset used in this study consists of images portraying benign and malignant breast cancer from a clinical investigation at the P and D Laboratory in Brazil [18]. Pathologists extract samples from the surface of biopsy breast tissues through surgical biopsy, label them, and employ the standard paraffin process for microscopic examination, adhering to routine clinical procedures.

Doctors initially identify tumors and define the Region of Interest (ROI) at a minimum of 40× magnification. Magnification is then manually increased to 100×, 200×, and 400×, capturing different aspects of tissue changes. We are using images of only 40× magnification for our work. The images are in 24-bit RGB format with a resolution of 700 by 460 pixels. Table I shows the statistics of the dataset.

TABLE I
Dataset Statistics

Classes	40×	100×	200×	400×	Total
Adenosis	114	113	111	106	444
Fibroadenoma	253	260	264	237	1014
Tubular Adenoma	109	121	108	115	453
Phyllodes Tumor	149	150	140	130	569
Ductal Cancer	864	903	896	788	3451
Lobular Cancer	156	170	163	137	626
Mucinous Cancer	205	222	196	169	792
Papillary Cancer	145	142	135	138	560
Total	1995	2081	2013	1820	7909

BreakHis dataset includes 7,909 breast cancer images, categorized into benign and malignant classes. Benign comprises Adenosis (B_A), Fibroadenoma (B_F), Phyllodes Tumor (B_P) and Tubular Adenoma (B_T), with 2,440 images. Malignant has 5,429 images, categorized into Ductal Cancer (M_D), Lobular Cancer (M_L), Mucinous Cancer (M_M) and Papillary Cancer (M_P). Fig. 5 displays images of each type of benign and malignant cancer.

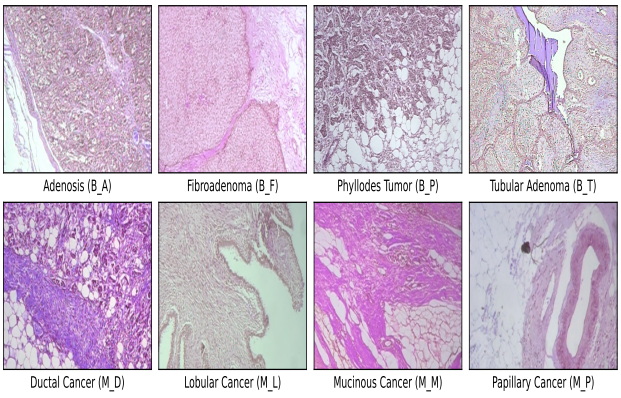

Fig. 5. Types of benign and malignant cancer

B. Balancing

In addressing imbalanced multi-class data, the utilization of resampling techniques proves essential for achieving class balance. Employing SMOTE, which generates synthetic samples for minority classes, effectively alleviates the scarcity issue, while Tomek Links, by identifying and eliminating noisy majority samples, aids in refining decision boundaries. Combination of these techniques is used to balance the class sizes which improves overall classifier performance. Through the strategic implementation of SMOTE and Tomek Links, the model becomes more adept at providing equitable predictions across all classes, ensuring unbiased learning framework.

C. Ensembling

Ensembling is a potent method that improves a classification system's overall performance and robustness by combining the predictions from several models. Models like CNN, VGG-19, and DenseNet-121 each contribute unique perspectives, capturing varied patterns and features within image data. The ensembling techniques that we are using are average ensemble and weighted average ensemble. Combining the predictions of models through these techniques significantly improves overall accuracy. The average ensemble consolidates predictions equally across models, while the weighted average ensemble assigns varying weights to models' predictions, leveraging the strengths of influential models.

1) Average Ensemble: The average ensemble combines outputs from multiple models for a collective prediction based on the highest sum. The ensemble prediction is denoted as $\hat{y}_{\text{avg\_ensemble}}$, is derived by finding the argument that maximizes the sum of individual predictions $\hat{y}_i$, where i ranges from 1 to N, representing the total number of individual predictions. This method is encapsulated by equation 2, leveraging diverse model outputs enhances accuracy by consolidating contributions into a unified decision.

$$\hat{y}_{\text{avg\_ensemble}} = \arg\max\left(\sum_{i=1}^{N} \hat{y}_i\right) \tag{2}$$

2) Weighted Average Ensemble: The weighted average ensemble combines predictions with designated weights: 0.6 for CNN, and 0.2 each for VGG-19 and DenseNet-121. The computation, as outlined in equation 3, derives the prediction $\hat{y}_{\text{weighted\_avg\_ensemble}}$ by summing the products of individual model predictions $\hat{y}_i$ and their corresponding weights w_i, where i ranges from 1 to N. The weighted average ensemble, with a higher weight for CNN, combines insights from diverse models, signifying the total number of predictions.

$$\hat{y}_{\text{weighted\_avg\_ensemble}} = \arg\max\left(\sum_{i=1}^{N} w_i \cdot \hat{y}_i\right) \tag{3}$$

IV. Results and Discussions

In this section, we discuss the results yielded by the proposed system. It delves into the distribution of image classes before and after balancing, assesses the performance of deep learning models, compares the model performance on both balanced and imbalanced datasets, and evaluates the effectiveness of ensemble techniques in enhancing predictions.

A. Balancing

The analysis of the dataset revealed a significant imbalance, particularly with the Ductal Cancer (M_D) class having a larger number of images compared to others as depicted in Fig. 6. This imbalance could potentially affect the efficacy of the model.

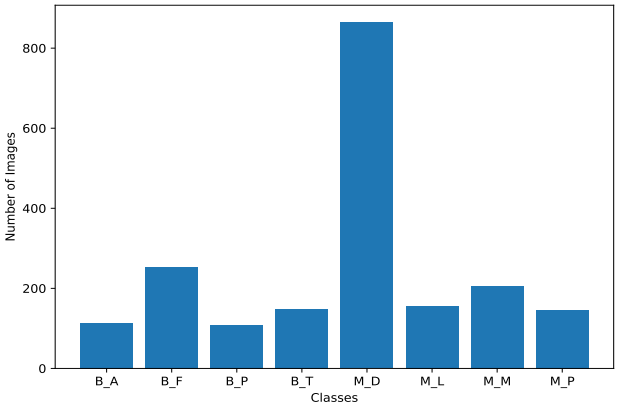

Fig. 6. Distribution of different classes before balancing

The dataset is balanced using SMOTE and Tomek links, ensuring an even distribution of classes for robust and unbiased learning during subsequent model training and evaluation. Fig. 7 illustrates the uniform distribution of images across each class.

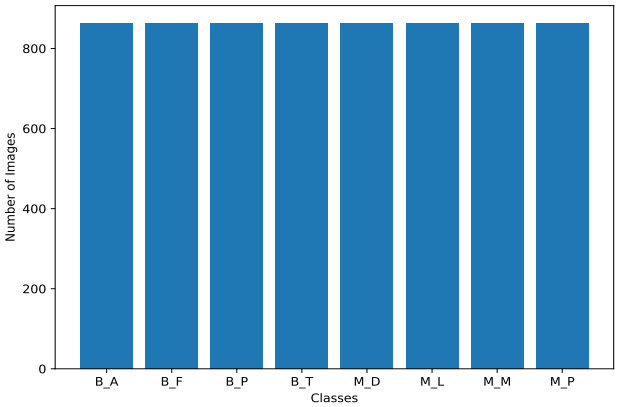

Fig. 7. Distribution of different classes after balancing

B. Individual Model Performance Analysis

The evaluation metrics used to analyze the performance of the individual deep learning models is accuracy score which is calculated as the ratio of correctly predicted instances to the total number of instances assessed, given by the equation 4. Additionally, the loss function employed is categorical cross-entropy, a standard choice for multi-class classification problems.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (4)$$

1) CNN: After training for 50 epochs, the CNN achieved an accuracy of 91.29%. This result highlights the model's effective pattern recognition, ensuring reliable classifications. Refer to Fig. 8 for the corresponding accuracy and loss graph, providing a visual overview of the CNN's performance.

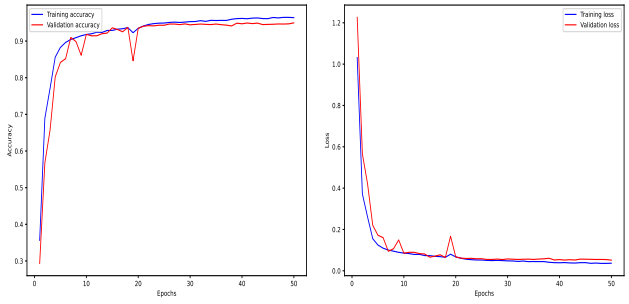

Fig. 8. CNN accuracy (left) and loss (right) graph

2) VGG-19: VGG-19 achieved 81.78% accuracy after 50 training epochs, showcasing adaptability in validation and unseen data scenarios. This underscores robust generalization, emphasizing reliability in diverse data challenges. Fig. 9 shows the accuracy and loss graph, providing a visual overview of VGG-19's performance.

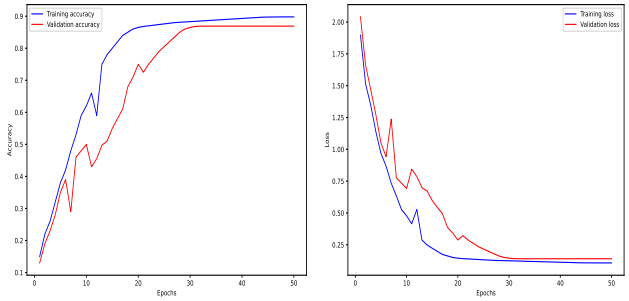

Fig. 9. VGG-19 accuracy (left) and loss (right) graph

3) DenseNet-121: After 50 epochs, DenseNet-121 achieved 83.64% accuracy, consistently performing across all metrics. It demonstrated precise predictions and a balanced precision-recall, highlighting reliability in predicting outcomes for unseen data. Fig. 10 displays the accuracy and loss graph of DenseNet-121.

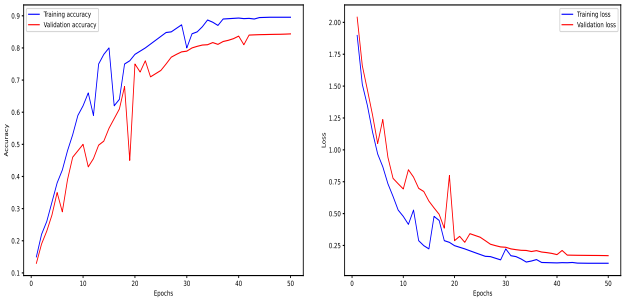

Fig. 10. DenseNet-121 accuracy (left) and loss (right) graph

C. Comparative Analysis of Models

Models trained on imbalanced data achieved accuracies of 80.28% for CNN, 65.49% for VGG-19, and 68.21% for DenseNet-121. Balancing the data improved accuracy to 91.89% for CNN, 81.78% for VGG-19, and 83.64% for DenseNet-121 as shown in the Fig. 11. The significant increase underscores the importance of addressing class imbalance to fully utilize the predictive power of these models, emphasizing the crucial role of balanced training for enhanced CNN, VGG-19, and DenseNet-121 performance and reliability.

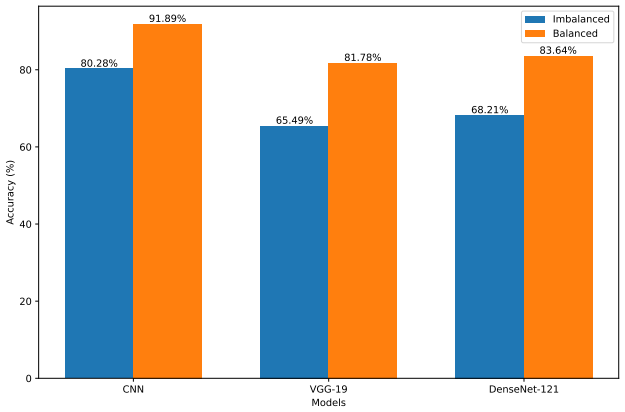

Fig. 11. Accuracy of different models on imbalanced and balanced data

D. Ensembling

Ensemble techniques, applied to balanced data, achieved 92.91% accuracy for the average ensemble and slightly higher at 93.20% for the weighted average ensemble. Refer to Table II for details across all models and ensembles.

TABLE II
ACCURACY SCORES OF DIFFERENT MODELS

Sl. No.	Model	Accuracy Score
1	CNN	91.89%
2	VGG-19	81.78%
3	DenseNet-121	83.64%
4	Average Ensemble	92.91%
5	Weighted Average Ensemble	93.20%

V. Conclusion and Future Scope

In this study, after training different deep learning models like CNN, VGG-19, and DenseNet-121 and using different ensembling techniques like average ensemble and weighted average ensemble we can say that the ensembling techniques give better accuracy than the individual deep learning models. Also, we saw that the deep learning models yield better accuracy when trained on balanced data than on imbalanced data. So, we can conclude that the resampling techniques are necessary to balance the data and solve the class imbalance problem, and the ensembling techniques are necessary to improve the overall accuracy.

Future research could focus on refining ensemble methods with adaptive strategies for varying dataset complexities, aiming to bolster model performance across diverse real-world scenarios.

References

[1] Breast Cancer. Available at https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., & Zwiggelaar, R. (2018). Deep learning in mammography and breast histology, an overview and future trends. Medical Image Analysis, 47, 45-67. https://doi.org/10.1016/j.media.2018.03.006

[3] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455-1462, July 2016, doi: 10.1109/TBME.2015.2496264.

[4] Veta, Mitko, Josien PW Pluim, Paul J. Van Diest, and Max A. Viergever. "Breast cancer histopathology image analysis: A review." IEEE transactions on biomedical engineering 61, no. 5 (2014): 1400-1411.

[5] Aloyayri, Abdulrahman, and Adam Krzyżak. "Breast cancer classification from histopathological images using transfer learning and deep neural networks." In Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I 19, pp. 491-502. Springer International Publishing, 2020.

[6] Daoud, Mohammad I., Samir Abdel-Rahman, Tariq M. Bdair, Mahasen S. Al-Najar, Feras H. Al-Hawari, and Rami Alazrai. "Breast tumor classification in ultrasound images using combined deep and handcrafted features." Sensors 20, no. 23 (2020): 6838. https://doi.org/10.3390/diagnostics13193113

[7] Y. Yari, T. V. Nguyen and H. T. Nguyen, "Deep Learning Applied for Histological Diagnosis of Breast Cancer," in IEEE Access, vol. 8, pp. 162432-162448, 2020, doi: 10.1109/ACCESS.2020.3021557.

[8] A. Titoriya and S. Sachdeva, "Breast Cancer Histopathology Image Classification using AlexNet," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2019, pp. 708-712, doi: 10.1109/ISCON47742.2019.9036160. https://doi.org/10.1155/2023/6530719.

[9] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean and L. Heutte, "Deep features for breast cancer histopathological image classification," 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 2017, pp. 1868-1873, doi: 10.1109/SMC.2017.8122889.

[10] Liu, Xiaoqi, Chengliang Wang, Jianying Bai, and Guobin Liao. "Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images." Neurocomputing 392 (2020): 253-267.

[11] Wang, Jian, Hengde Zhu, Shui-Hua Wang, and Yu-Dong Zhang. "A review of deep learning on medical image analysis." Mobile Networks and Applications 26 (2021): 351-380.

[12] N. Liu, X. Li, E. Qi, M. Xu, L. Li and B. Gao, "A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data," in IEEE Access, vol. 8, pp. 171263-171280, 2020, doi: 10.1109/ACCESS.2020.3014362.

[13] Ganaie, Mudasir A., Minghui Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan. "Ensemble deep learning: A review." Engineering Applications of Artificial Intelligence 115 (2022): 105151.

[14] Alshayeji, Mohammad H., Hanem Ellethy, and Renu Gupta. "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach." Biomedical Signal Processing and Control 71 (2022): 103141.

[15] Ozcan, Irem, Hakan Aydin, and Ali Cetinkaya. "Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer." Asian Pacific Journal of Cancer Prevention: APJCP 23, no. 10 (2022): 3287.

[16] Kumar, Sumit, and Shallu Sharma. "Sub-classification of invasive and non-invasive cancer from magnification independent histopathological images using hybrid neural networks." Evolutionary Intelligence 15, no. 3 (2022): 1531-1543.

[17] N. S. Patil, S. D. Desai and S. Kulkarni, "Magnification independent fine-tuned transfer learning adaptation for multi-classification of breast cancer in histopathology images," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 1185-1191, doi: 10.1109/ICAC3N56670.2022.10074159.

[18] Pereira, Mayke (2023), "BreakHis - Breast Cancer Histopathological Database", Mendeley Data, V1, doi: 10.17632/jxwvdwhpc2.1