A Scalable Big Data-Driven Distributed Deep Learning Framework for Breast Cancer Diagnosis Using Big Data Analytics

Muhammad Babar
Robotics and Internet-of-Things Laboratory, Prince Sultan University, Saudi Arabia
mbabar@psu.edu.sa

Sarah Kaleem
EIAS Data Science Lab, Prince Sultan University, Saudi Arabia
skaleem@psu.edu.sa (corresponding author)

Mohammed El-Affendi
College of Computer and Information Sciences, Prince Sultan University, Saudi Arabia
affendi@psu.edu.sa

Zahid Khan
Robotics and Internet-of-Things Laboratory, Prince Sultan University, Saudi Arabia
zskhan@psu.edu.sa

Received: 1 June 2025 | Revised: 9 July 2025 and 28 July 2025 | Accepted: 1 August 2025

 DOI: https://doi.org/10.48084/etasr.12485

ABSTRACT

The accurate and early detection of breast cancer remains a significant challenge in medical diagnostics, primarily due to the complexity of histopathological images and the large volume of data involved. This paper presents a novel hybrid deep learning framework that leverages Big Data Analytics (BDA) and Convolutional Neural Networks (CNNs) to enhance the accuracy of breast cancer detection. The proposed system integrates three robust deep learning architectures (VGG16, VGG19, and ResNet50) trained in parallel across distributed nodes using Apache Spark, thereby accelerating computation and enabling scalable learning. This study used the BreakHis dataset, which contains 15,918 original images collected at four magnifications. To enhance generalization and class balance, extensive data augmentation and patch extraction were applied, which expanded the dataset to approximately 275,000 training samples. The hybrid model demonstrated high performance in classification tasks, achieving high precision, recall, and F1-scores compared to existing benchmarks. Key performance indicators, such as accuracy, specificity, and sensitivity, confirm the effectiveness of the model in distinguishing between benign and malignant cases. Unlike traditional monolithic CNN approaches, the proposed system leverages distributed processing to reduce training time while efficiently handling massive datasets.

Keywords-big data; breast cancer; distributed learning; deep CNN

I. INTRODUCTION

Breast cancer is the leading cause of death for both males and females. Its five-year survival rate is only 17% for lung cancer, but if diagnosed early, survival increases to 54%. Low-dose Computed Tomography (CT) is a method used to identify and diagnose breast cancer [1]. The National Lung Screening Trial (NLST) indicates that the number of people undergoing low-dose CT (LDCT) is 20% lower compared to plain chest radiography [2]. CT provides more resolution, volumetric data, and detects even the most minor bumps. However, several hurdles exist in the implementation of LDCT in this setting, which in turn hinder accurate detection and dynamic testing. Test systems generate extensive volumetric data that is time-consuming and labor-intensive for radiologists to review carefully [3]. In addition, the concerns about radiation exposure, drug overdose, and excessive stress are intense.

CNN is a robust image processing technique that implements deep learning to perform both reproductive and descriptive tasks, often utilizing machine vision that incorporates image and video visualization, as well as complementary programs [4]. CNNs are revolutionizing computer vision across various fields, including autonomous

vehicles, drones, security, and healthcare. They excel at feature extraction and are widely applied in areas such as biomedical science, where they address challenges such as image analysis. Their success lies in their ability to extract deep features efficiently, making them a powerful tool for complex tasks. The detection of breast cancer on digital histopathological images is an area of in-depth study [5, 6].

Big data analytics is a strategy to distribute data across different computers [7]. This approach is commonly used to handle large datasets efficiently and support faster decision-making, improving automation, healthcare, and smart technologies [8]. Parallel processing across multiple systems enables faster analysis and unified results. HDFS treats distributed storage as a single system, providing scalable resource management, while Hadoop ensures flexible handling of big data [9]. Apache Spark improves performance for streaming, AI, and SQL tasks. This study aimed to use this parallel distributed framework to train Deep DCNNs. Unlike previous approaches that use either single CNN models or perform training on standalone systems, this work presents a novel hybrid ensemble of three CNNs trained in parallel across multiple Spark nodes. This distributed setup significantly reduces training time while maintaining high classification performance. Moreover, the proposed system implements a late fusion strategy for result aggregation, enhancing diagnostic reliability across heterogeneous data. The key contributions of this study are:

- Proposes a hybrid CNN framework that combines VGG16, VGG19, and ResNet50.
- Combines big data with deep learning for scalable processing and training on large medical datasets.
- Implements parallel and distributed training to significantly reduce computational time.
- Utilizes over 275,000 histopathological images, creating a comprehensive dataset for training.
- Demonstrates the advantage of distributed learning, showing how to scale across multiple machines.

II. LITERATURE REVIEW

The diagnosis of breast cancer through histopathological imaging remains one of the most effective, yet labor-intensive, methods in clinical practice. Deep learning techniques, particularly CNNs, have emerged as powerful tools for automating the classification and detection of breast cancer from histopathological images. In [10], a DCNN model was proposed to classify breast cancer histopathological images with improved accuracy and training efficiency. This model incorporated several optimization techniques, including Adam, RMSprop, and Nesterov acceleration, and utilized CUDA-enabled GPUs to accelerate training. In [11], an approach for the early detection of breast cancer through mammography images was presented, using CNNs trained on large public datasets such as DDSM and CBIS-DDSM.

Recent advances in deep learning have significantly enhanced breast cancer detection through various imaging modalities and optimization techniques. In [12], a deep learning model integrated metaheuristic optimization for effective segmentation and classification of breast cancer using mammogram images, achieving promising diagnostic performance by refining the model's learning process. Another approach focused on combining deep learning with the Internet of Medical Things (IoMT) to enable early and efficient breast cancer diagnosis through optimized data collection and processing across medical devices [13]. In [14], a deep learning-based system was developed to support early detection by learning intricate patterns in histopathological images, providing high accuracy and reduced diagnostic delay. In [15], a robust classification framework utilized DICOM images, addressing challenges in real-world clinical scenarios by enhancing feature extraction and classification stability.

In [16], a hybrid model combined CNN-based feature extraction with traditional classifiers, such as Logistic Regression (LR) and Support Vector Machines (SVM). This two-stage framework processes image features at various magnification levels and compares classification performance. The experimental results showed that CNN+LR outperformed other combinations, providing a balance between computational simplicity and classification accuracy. In [17], a parallel CNN architecture, called 3PCNNB-Net, was introduced to classify breast cancer images. The model consisted of three identical CNN branches that independently extracted features and then merged them for final classification. Tested on the BreakHis dataset, the model achieved an accuracy of 97.14% and outperformed conventional CNN configurations. In [18], a multi-task deep learning framework was developed for fine-grained classification and grading of breast cancer histopathological images. This architecture performed simultaneous image classification and verification, leveraging shared representations to improve generalization. In [19], a four-stage mathematical model was presented for detecting breast cancer. In [20], a CNN-TQN model achieved 98.8% accuracy on breast cancer detection. These studies emphasize the growing potential of integrating deep learning with medical imaging to improve breast cancer diagnostics.

III. PROPOSED SYSTEM

Machine learning approaches rely on feature engineering and background information to identify and extract values from raw data to readable presentations. Learning determines the optimal representation of raw data and features to simplify tasks related to classification, prediction, or detection. In deep learning, the learning process aims at converting raw data into structured representations by utilizing multiple layers of processing modules. These compound layers of features are analyzed in the raw data using a particular learning process, rather than being physically designed by a project engineer.

Figure 1 shows an overview of the proposed system. Initially, the big data sets are provided to the proposed system. Preprocessing algorithms are applied to the data for validity checks. The major preprocessing includes the normalization and 3D conversion of the image datasets. Afterward, the data is merged and visualized for verification. The prepared data are loaded into the processing module and split. A specific model is selected for training and applied to multiple data splits.

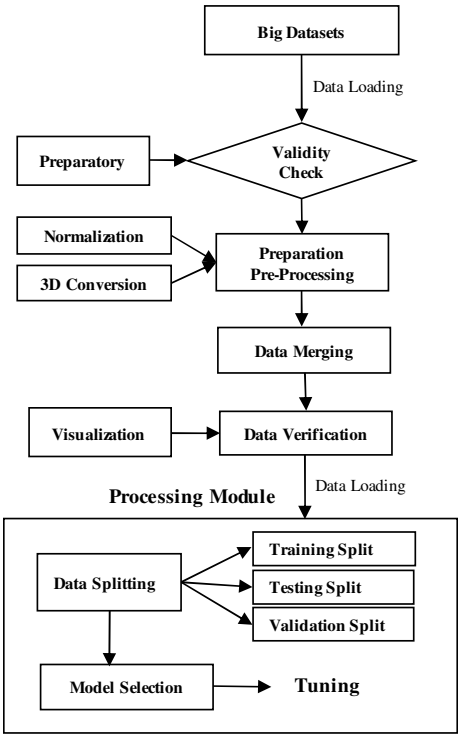

Fig. 1. System overview.

Figure 2 details the architecture of the proposed system. Figure 3 shows a schematic diagram of the hybrid framework showing the data flow and integration.

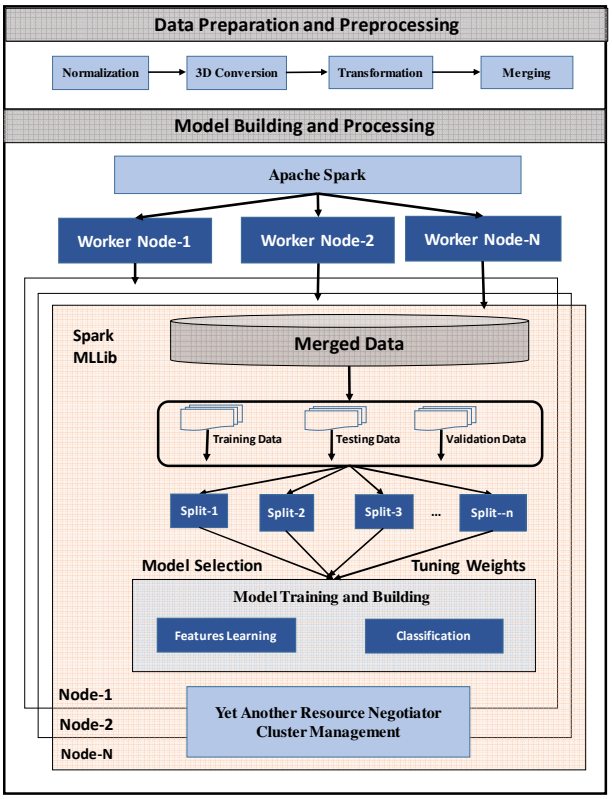

Fig. 2. Proposed system.

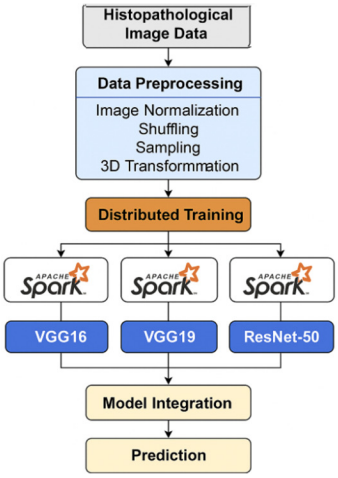

Fig. 3. Flow of the proposed system.

A. Big Data Preparation and Pre-Processing

The dataset used is the BreakHis Breast Cancer Histopathological Image Dataset [21], containing 15,918 H&E-stained images from 82 patients at 40×, 100×, 200×, and 400× magnifications, with a native resolution of ~700×460 pixels. To enhance diversity and balance, augmentation (rotation, flips, scaling, brightness/contrast) and patch extraction (resized to 224×224) were applied, expanding the dataset to ~275,000 samples for training, validation, and testing.

Standard pre-trained ResNet-50, VGG-19, and VGG-16 models from the Keras library were utilized. The VGG16, VGG19, and ResNet50 architectures were used as baseline models with modifications tailored to the dataset. Specifically, the top classification layers were removed, and new dense layers were added to adapt the networks for the four-class breast cancer classification task. Learning was employed by freezing the convolutional base in initial experiments and then fine-tuning deeper layers. Dropout layers (rate = 0.5) were added to reduce overfitting. All models were trained using the Adam optimizer with a learning rate of 0.0001.

TABLE I. TRAINING PROCESS DETAILS

Attribute/Models	VGG16	VGG19	ResNet50
Trainable layers	Last 5 layers	Last 6 layers	Last 10 layers
Input size	224×224×3	224×224×3	224×224×3
Dropout	0.5	0.5	0.5
Optimizer	Adam	Adam	Adam
Learning rate	0.0001	0.0001	0.0001
Epochs	50	50	50
Batch size	32	32	32

This study used two classification schemes. The first was a binary setup distinguishing Benign from Malignant samples. The second was a four-class setup: Benign, In Situ Carcinoma, Invasive Carcinoma, and Other Malignant Variants (grouping the remaining subtypes), maintaining clinical relevance while improving the balance between classes for training. Machine learning approaches are common in their capacity to process natural data in its raw form, relying on feature engineering and background information to identify and extract reasonable values from raw data to readable presentations.

1) Image Shuffling

Rearranging strategies involve reorganizing data. However, they maintain consistent connections between sections by haphazardly rearranging data from a dataset into a quality or a group of characteristics. Data can be replaced with similar attributes from different records, preserving the overall features. This approach works well for analytical use cases, ensuring that key metrics remain valid. It also allows the generated data to be used safely for testing and training, as the distribution remains consistent.

2) Sampling

This step analyzes the data to separate positive and negative training samples. Next, the dataset is divided into training, validation, and test sets.

3) Data Normalization and Conversion

The pixel values were normalized to the range [0, 1]. Before inputting the data into the DCNNs, it must be ensured that it is compatible with the model requirements, specifically converting them into a 3D input format, to train each model.

B. Model Building and Processing

In the proposed hybrid framework, each histopathological image is simultaneously processed by three pre-trained CNNs: VGG16, VGG19, and ResNet-50. VGG16 emphasizes low-level textures and edges, VGG19 captures deeper hierarchical structures with its extended depth, and ResNet-50 leverages residual connections to extract highly abstract features without gradient degradation. The resulting feature maps are flattened and concatenated into a joint framework that integrates local details, structural semantics, and deep contextual cues for robust representation. The joint embedding is passed through a fully connected layer to balance the contributions of all three networks. The final classification is achieved through late fusion, where the softmax outputs of VGG16, VGG19, and ResNet-50 are combined using a weighted average based on their validation accuracies. This approach improves robustness to class imbalance and takes advantage of the complementary strengths of VGG (local and mid-level features) and ResNet-50 (deep abstract patterns), resulting in a more stable and accurate classifier.

All models were trained for 50 epochs with a batch size of 32. The dataset was divided into 70% training, 15% validation, and 15% testing. Early stopping with patience (5) and model checkpointing based on validation loss were used. The training was performed on distributed nodes using Apache Spark MLlib, enabling efficient parallelism of the three CNN models across worker nodes. After training, softmax probabilities from each model were collected and aggregated using a weighted average ensemble strategy. A late fusion ensemble approach was adopted, where the softmax outputs of VGG16, VGG19, and ResNet50 were combined using a weighted averaging method. Weights were dynamically assigned based on the validation accuracy of each model, allowing higher-performing models to contribute more heavily to the final prediction. This method improved robustness and minimized the bias introduced by any single model.

The final component of the model involves the data processing pipeline integrated with Apache Spark. In our Spark-based framework, the driver partitions the dataset and transmits the weights of VGG16, VGG19, and ResNet-50 to all worker nodes. Each worker executes all three CNNs on its assigned data, fuses their outputs through weighted averaging (weights proportional to validation accuracy), and sends the fused results back to the driver for global aggregation. To evaluate scalability and performance efficiency, the system was initially deployed on two nodes, with subsequent trials using nodes 3 through 6. This distributed training approach enabled the division of large datasets across multiple computing nodes, facilitating parallel execution. As a result, training times were reduced and system performance improved. This architecture demonstrates the ability of the proposed CNN model, when combined with big data tools like Apache Spark, to efficiently handle massive image datasets and deliver high accuracy across different classes while optimizing resource usage and reducing computational delays.

IV. RESULTS AND DISCUSSION

A. Data Source and Collection

This study used the dataset in [19], comprising 72% (198,737 images) of negative results and 28% (78,786 images) with positive results. The dataset contains no sensitive personal data, meeting all ethical standards for medical imaging data.

B. Performance Indicators and Results

Various performance indicators were employed to evaluate the performance, including accuracy, precision, recall, and F-measure. Accuracy refers to the proportion of positive predicted results that are correct. Precision is a ratio between correctly predicted positive images and the total predicted positive images. This indicator will help identify when the rate of False Positives (FP) is high. Recall (Sensitivity) is a ratio of correctly predicted positive images over the total positive images. F1 score is the harmonic mean of Precision and Recall, with a higher F1 score meaning fewer FP and FN, leading to more reliable results.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total Predictions}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (4)$$

Figures 4-8 illustrate the results, showing that the proposed approach achieved great results in terms of accuracy, precision, recall, and F1 score. In addition, specificity was calculated to measure the model's accuracy in identifying individuals without cancer. This metric accurately reflects the proportion of truly negative cases correctly classified as negative among all truly negative cases. High specificity means that the test effectively rules out the condition in healthy individuals, lowering the chance of FP.

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

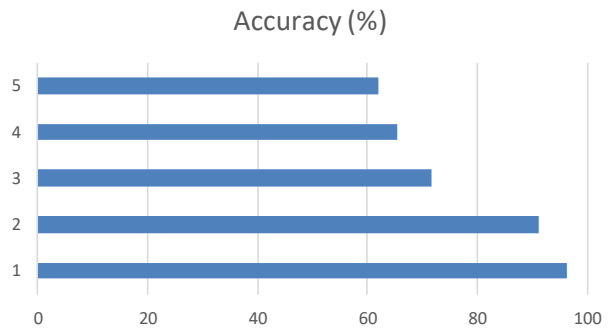

Fig. 4. Accuracy of the proposed hybrid approach.

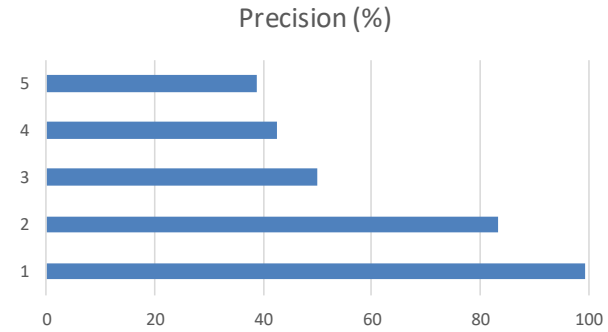

Fig. 5. Precision of the proposed hybrid approach.

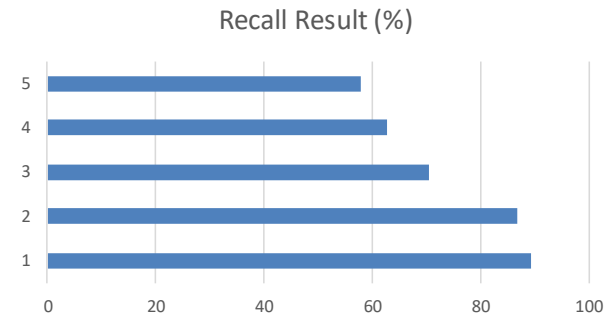

Fig. 6. Recall of the proposed hybrid approach.

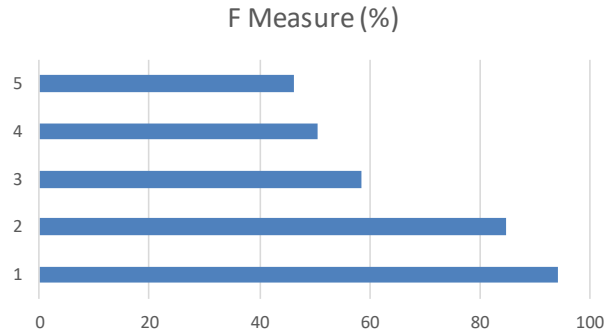

Fig. 7. F-measure of the proposed hybrid approach.

Figure 9 shows a comparison between specificity and sensitivity (Recall). Figure 10 shows a performance comparison of the proposed model for 2-class and 4-class classification using eight nodes, which offered the best balance between performance, computational efficiency, and communication overhead.

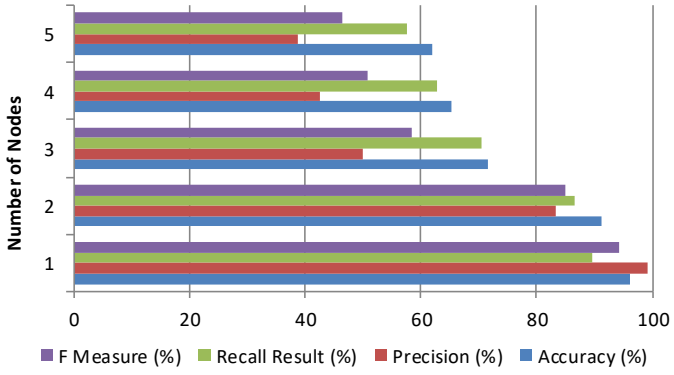

Fig. 8. Overall results.

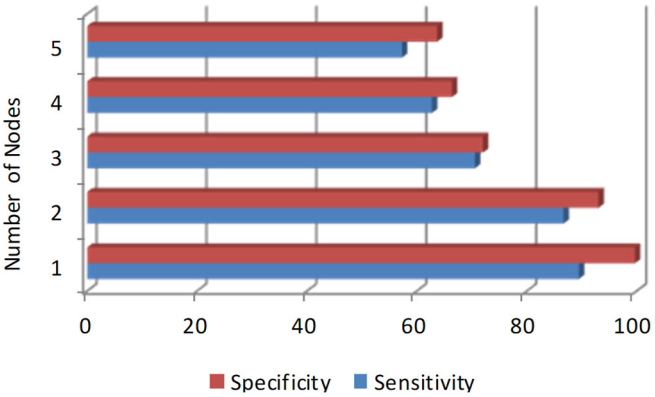

Fig. 9. Specificity and sensitivity of the proposed hybrid approach.

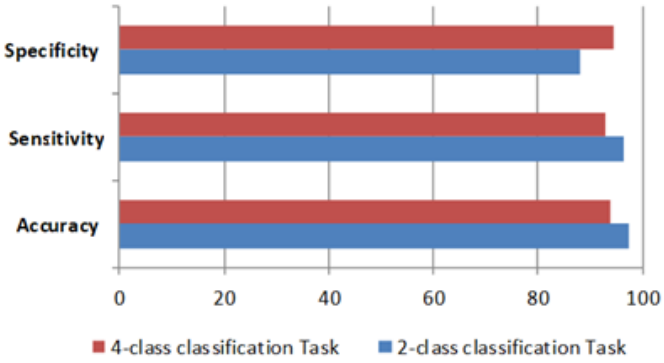

Fig. 10. Comparison of results for 2- vs 4-class classification with 8 nodes.

V. CONCLUSION

This study presented a scalable deep learning framework that combines CNN models for early breast cancer detection. Using VGG16, VGG19, and ResNet50 within an Apache Spark-based distributed setup, the system can effectively handle large histopathological datasets. Results demonstrate high accuracy, sensitivity, and specificity, surpassing those of traditional methods. Key strengths include efficient preprocessing, optimized training, and model fusion, enhancing both reliability and speed. The proposed framework offers a practical solution for real-world clinical use, enabling faster and more accurate diagnoses.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

REFERENCES

[1] C. Santucci *et al.*, "European cancer mortality predictions for the year 2025 with focus on breast cancer," *Annals of Oncology*, vol. 36, no. 4, pp. 460–468, Apr. 2025, https://doi.org/10.1016/j.annonc.2025.01.014.

[2] S. E. Robertson *et al.*, "Comparing Lung Cancer Screening Strategies in a Nationally Representative US Population Using Transportability Methods for the National Lung Cancer Screening Trial," *JAMA Network Open*, vol. 7, no. 1, Jan. 2024, Art. no. e2346295, https://doi.org/10.1001/jamanetworkopen.2023.46295.

[3] D. Mastrodicasa *et al.*, "Use of AI in Cardiac CT and MRI: A Scientific Statement from the ESCR, EuSoMII, NASCI, SCCT, SCMR, SIIM, and RSNA," *Radiology*, vol. 314, no. 1, Jan. 2025, Art. no. e240516, https://doi.org/10.1148/radiol.240516.

[4] M. A. Wahed, M. Alqaraleh, M. S. Alzboon, and M. S. Al-Batah, "Evaluating AI and Machine Learning Models in Breast Cancer Detection: A Review of Convolutional Neural Networks (CNN) and Global Research Trends," *LatIA*, vol. 3, pp. 117–117, Jan. 2025, https://doi.org/10.62486/latia2025117.

[5] D. Tsietso, A. Yahya, R. Samikannu, B. Qureshi, and M. Babar, "Computational Approach for Automated Segmentation and Classification of Region of Interest in Lateral Breast Thermograms," *Computers, Materials and Continua*, vol. 80, no. 3, pp. 4749–4765, Sep. 2024, https://doi.org/10.32604/cmc.2024.052793.

[6] D. Tsietso *et al.*, "Multi-Input Deep Learning Approach for Breast Cancer Screening Using Thermal Infrared Imaging and Clinical Data," *IEEE Access*, vol. 11, pp. 52101–52116, 2023, https://doi.org/10.1109/ACCESS.2023.3280422.

[7] S. Kaleem, A. Sohail, M. U. Tariq, and M. Asim, "An Improved Big Data Analytics Architecture Using Federated Learning for IoT-Enabled Urban Intelligent Transportation Systems," *Sustainability*, vol. 15, no. 21, Jan. 2023, Art. no. 15333, https://doi.org/10.3390/su152115333.

[8] M. T. J. Mehedy *et al.*, "Big Data and Machine Learning in Healthcare: A Business Intelligence Approach for Cost Optimization and Service Improvement," *The American Journal of Medical Sciences and Pharmaceutical Research*, vol. 7, no. 03, pp. 115–135, Mar. 2025, https://doi.org/10.37547/tajmspr/Volume07Issue03-14.

[9] K. J. Merceedi and N. A. Sabry, "A Comprehensive Survey for Hadoop Distributed File System," *Asian Journal of Research in Computer Science*, pp. 46–57, Aug. 2021, https://doi.org/10.9734/ajrcos/2021/v11i230260.

[10] K. C. Burçak, Ö. K. Baykan, and H. Uğuz, "A new deep convolutional neural network model for classifying breast cancer histopathological images and the hyperparameter optimisation of the proposed model," *The Journal of Supercomputing*, vol. 77, no. 1, pp. 973–989, Jan. 2021, https://doi.org/10.1007/s11227-020-03321-y.

[11] G. Hamed, M. A. E. R. Marey, S. E. S. Amin, and M. F. Tolba, "Deep Learning in Breast Cancer Detection and Classification," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, 2020, pp. 322–333, https://doi.org/10.1007/978-3-030-44289-7_30.

[12] M. Sreevani and R. Latha, "A Deep Learning with Metaheuristic Optimization-Driven Breast Cancer Segmentation and Classification Model using Mammogram Imaging," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20342–20347, Feb. 2025, https://doi.org/10.48084/etasr.9406.

[13] A. Naz, H. Khan, I. U. Din, A. Ali, and M. Husain, "An Efficient Optimization System for Early Breast Cancer Diagnosis based on Internet of Medical Things and Deep Learning," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15957–15962, Aug. 2024, https://doi.org/10.48084/etasr.8080.

[14] A. Bekkouche, M. Merzoug, M. Hadjila, and W. Ferhi, "Towards Early Breast Cancer Detection: A Deep Learning Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17517–17523, Oct. 2024, https://doi.org/10.48084/etasr.8634.

[15] T. N. Nguyen, T. T. Nguyen, T. H. Nguyen, and B. V. Ngo, "A Robust Approach for Breast Cancer Classification from DICOM Images," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23499–23505, Jun. 2025, https://doi.org/10.48084/etasr.10931.

[16] K. Gupta and N. Chawla, "Analysis of Histopathological Images for Prediction of Breast Cancer Using Traditional Classifiers with Pre-Trained CNN," *Procedia Computer Science*, vol. 167, pp. 878–889, Jan. 2020, https://doi.org/10.1016/j.procs.2020.03.427.

[17] A. M. Ibraheem, K. H. Rahouma, and H. F. A. Hamed, "3PCNNB-Net: Three Parallel CNN Branches for Breast Cancer Classification Through Histopathological Images," *Journal of Medical and Biological Engineering*, vol. 41, no. 4, pp. 494–503, Aug. 2021, https://doi.org/10.1007/s40846-021-00620-4.

[18] L. Li *et al.*, "Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 14509–14528, Jun. 2020, https://doi.org/10.1007/s11042-018-6970-9.

[19] T. Abdeljawad, R. U. Din, N. Fatima, K. Shah, K. J. Ansari, and H. Alrabaiah, "Mathematical modeling of breast cancer with four stages," *International Journal of Biomathematics*, Apr. 2025, Art. no. 2550036, https://doi.org/10.1142/S1793524525500366.

[20] T. Mahmood, T. Saba, and A. Rehman, "Breast cancer diagnosis with MFF-HistoNet: a multi-modal feature fusion network integrating CNNs and quantum tensor networks," *Journal of Big Data*, vol. 12, no. 1, Mar. 2025, Art. no. 60, https://doi.org/10.1186/s40537-025-01114-9.

[21] "BreakHis - Breast Cancer Histopathological Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/waseemalastal/breakhis-breast-cancer-histopathological-dataset.