AI-Enabled Breast Cancer Diagnosis for Health Networks Using Salp Swarm and Sparrow Search Optimized Adaboost

Akshat Gaurav
Ronin Institute, Montclair, NJ, USA
akshat.gaurav@ieee.org

Brij B. Gupta*
Department of Computer Science and Information Engineering,
Asia University, Taichung 413, Taiwan,
bbgupta@asia.edu.tw

Kwok Tai Chui
Hong Kong Metropolitan University (HKMU),
Hong Kong,
jktchui@hkmu.edu.hk

***Abstract*—Breast cancer is one of the most common and deadly diseases worldwide, making early and accurate detection critical for improving patient outcomes. Machine learning has shown significant promise in automating breast cancer diagnosis, but selecting the right features and optimizing model performance remain challenging. In this context, this work provides an AI-enabled method employing an AdaBoost model optimized using Salp Swarm Optimization (SSO) for feature selection and Sparrow Search Algorithm (SSA) for hyperparameter tuning for breast cancer detection. Our proposed model attained an accuracy of 99.12%, precision (99.13%), recall (99.12%), and F1-score (99.12%). The findings show that the suggested method offers a dependable and efficient way to identify breast cancer, therefore greatly raising the diagnostic accuracy in medical systems.**

***Index Terms*—Breast Cancer Diagnosis, AdaBoost, Salp Swarm Optimization (SSO), Sparrow Search Algorithm (SSA), Machine Learning Optimization**

I. Introduction

Breast cancer remains one of the most prevalent malignancies among women globally, characterized by the uncontrolled growth of breast cells leading to tumor formation. Early detection is crucial as it significantly enhances survival rates and treatment outcomes. The role of technology, particularly in diagnostic imaging and artificial intelligence (AI), has become increasingly vital in improving the accuracy and efficiency of breast cancer diagnosis[1, 2].

Diagnostic imaging techniques, such as mammography and ultrasound, are foundational in the early detection of breast cancer. Mammography, a specialized X-ray imaging technique, is particularly effective in identifying abnormalities that may not be palpable during physical examinations. It allows for the visualization of microcalcifications and other subtle changes in breast tissue, which are often early indicators of cancer [3, 4]. Ultrasound complements mammography by providing real-time imaging and is especially useful in differentiating between solid and cystic masses [5, 6]. The integration of these imaging modalities enhances the diagnostic process, enabling healthcare providers to make informed decisions regarding further intervention or treatment.

A. Contribution

This paper introduces an approach for breast cancer diagnosis by optimizing the AdaBoost model using Salp Swarm Optimization (SSO) for feature selection and Sparrow Search Algorithm (SSA) for hyperparameter tuning.

B. Organization

The remainder of this paper is organized as follows: Section 2 reviews the state-of-art models. Section 3 describes the proposed methodology. Section 4 presents the experimental setup, the evaluation metrics used, the results obtained, and a comparative analysis with other state-of-the-art models. Section 5 concludes the paper with a summary of the findings and suggests possible directions for future work.

II. Related Work

Vijayasarveswari et al. [7] propose a Statistically Modelled Feature Selection (SMFS) method for breast cancer detection using microwave technology. The contribution lies in combining the best feature extraction method (Statistical features) and feature selection method (Neighbour Component Analysis) to improve accuracy. The model achieves 85% accuracy. Gupta et al. [8] presents the W-RLG Model, a novel deep learning approach combining Whale Optimization, RNN, LSTM, and GRU to enhance cyber threat detection in healthcare IoT systems.

Suhiman et al. [9] propose evaluating feature selection methods (IG, ReliefF, mRMR) on mRNA expression data for breast cancer diagnosis, achieving the highest accuracy with mRMR and RF classifier using the top 25 genes. They also suggest that a hybrid approach (mRMR + SVM) improved accuracy with only the top 3 ranked genes. Kaushik and Gandhi [10] proposes an Access Control-based Trust Model for Healthcare Systems, ensuring only trusted and authorized users can access cloud-based EHRs. The model enhances the accuracy and efficiency of data access in cloud-integrated

*Corresponding Author

979-8-3315-0694-0/25/$31.00 ©2025 IEEE

healthcare systems. Onyebuchi et al. [11] constructs an enterprise cloud data warehouse for e-healthcare organizations, integrating medical/clinical workflows and enabling centralized storage of patient information. It supports medical software automation, hardware integration, and improved e-healthcare information management.

Akbar et al. [12] propose a breast tumor segmentation technique that combines contextual mapping using Swin Transformer with advanced edge analysis from DCE-MRI scans. Yu and Reiff-Marganiec [13] introduces a Causal Probability Description Logic Framework that integrates NLP, causality analysis, and extended knowledge graph technologies to enable machines to learn and infer causal relationships among diseases, symptoms, and other health facts. It demonstrates the framework's effectiveness by processing 801 diseases.

He et al. [14] propose a novel one-class classification approach combining double kernel mapping and a modified autoencoder based on the Broad Learning System (DKVBLS-AE) to enhance anomaly detection, particularly in medical datasets with imbalanced classes. Xiao et al. [15] proposes the PCE-CF service recommendation framework, which uses an embedded user portrait model to provide personalized recommendations for senior care services. It incorporates dynamic behavior modeling, location context, and deep learning for improved efficiency and feasibility. Zhang et al. [16] propose a category-weight instance fusion learning model for unsupervised domain adaptation in breast cancer diagnosis. Their key contribution is the integration of a category-weighted contrast knowledge distillation module to align domains at a category level and an instance-aware feature mixing module to enhance image style consistency across domains.

III. Proposed Work

This section presents the details of proposed model. As Figure 1 shows, the procedure starts with data preprocessing—including cleaning and balancing—then proceeds with Salp Swarm Optimization's (SSO) feature selection. The SSO method repeatedly generates fitness values and updates local and global best solutions, thereby optimizing the feature set. Training and testing datasets then separate the chosen characteristics. Using the Sparrow Search Algorithm (SSA), which modifies AdaBoost's hyperparameters for enhanced model performance, hyperparameter tweaking helps to better optimise the training dataset. Following training, the AdaBoost classifier produces test data output predictions that provide a consistent breast cancer diagnosis tool.

A. Data Preprocessing

After collecting the breast cancer dataset, data preprocessing is performed, which includes label encoding and normalization.

a) Label Encoding:: Categorical labels, such as 'Benign' and 'Malignant', are converted to numeric values as follows:

$$L = \begin{cases} 0 & \text{for Benign} \\ 1 & \text{for Malignant} \end{cases} \tag{1}$$

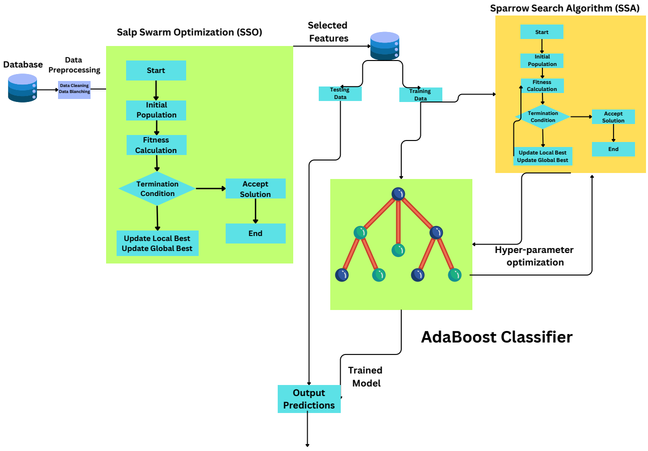


Fig. 1: Proposed Model

Normalization: The features are normalized using min-max normalization to bring them to a similar scale:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

where x_{norm} is the normalized value, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of feature x.

B. Feature Selection using Salp Swarm Optimization (SSO)

The Salp Swarm Optimization (SSO) algorithm is used to select the most relevant features from the dataset.

- Position update for the leader salp:

$$x_1^i(t+1) = x_1^i(t) + c_1 \times \left(c_2 \times \frac{ub_i - lb_i}{2} + lb_i \right) \tag{3}$$

  where:
  - $x_1^i(t)$ is the position of the leader salp in the search space,
  - ub_i and lb_i are the upper and lower bounds,
  - c_1 and c_2 are random coefficients.
- Position update for the follower salps:

$$x_j^i(t+1) = \frac{x_{j-1}^i(t) + x_j^i(t)}{2} \tag{4}$$

- Fitness function: The fitness function evaluates the quality of the selected features based on classification accuracy:

$$F(x) = \frac{1}{1 + A(x)} \tag{5}$$

  where $A(x)$ represents the accuracy of the classifier for feature subset x.

C. Hyperparameter Optimization using Sparrow Search Algorithm (SSA)

After selecting the features, the Sparrow Search Algorithm (SSA) is used to optimize the hyperparameters of the AdaBoost classifier.

Position update in SSA

$$x_j^i(t+1) = x_j^i(t) + \alpha \times (x_{\text{best}}^i - x_j^i(t)) + \beta \times (x_j^i(t) - x_{\text{worst}}^i) \quad (6)$$

where:

- x_{best}^i and x_{worst}^i are the best and worst positions in the population,
- α and β are random step size control variables.

D. AdaBoost Classifier

Finally, the optimized AdaBoost classifier is used for training and making predictions.

- Weight update: For incorrectly classified samples:

$$w_{i+1} = w_i \times \exp\left(\alpha_t \cdot I(y_i \neq h_t(x_i))\right) \quad (7)$$

where:

  - w_i is the weight of the i-th sample,
  - α_t is the weight of the t-th weak learner,
  - $I(y_i \neq h_t(x_i))$ is the indicator function that checks if the prediction was incorrect.
- Final Prediction:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (8)$$

where T is the total number of weak learners, and α_t is the contribution of each weak learner to the final prediction.

IV. Results and Discussion

A. System Information

Operating on a Linux-based environment with an x86_64 architecture, our system was utilized to create the proposed model with two physical core and four logical core parallel processing capabilities. With 32 GB accessible and 34 GB of RAM overall, it guarantees plenty of memory for machine learning chores. The system boasts 2.29 TB of free data storage out of 8.65 TB of disk capacity. With almost all of its 16 GB RAM accessible, a Tesla P100-PCIE-16GB GPU offers model-training high-performance capability.

B. Dataset Representation

We used a Kaggle breast cancer dataset for this study, which has thirty features characterizing different traits of cell nuclei seen in digitalized pictures of breast masses. The dataset has two loabels : M (Malignant) for malignant cells and B (Benign) for non-cancerous cells.

As shown in Figure 2, the collection comprises 357 benign and 212 malignant tagged samples. The uneven character of this distribution—more benign instances than malignant ones—may affect the learning process of the model. In healthcare uses like breast cancer diagnosis, where misclassification may have major effects, ensuring balanced model performance on both classes is very vital.

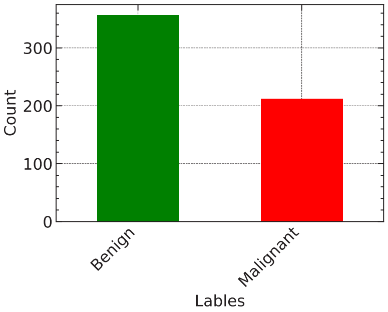


Fig. 2: Class Distribution

C. Performance of Salp Swarm Optimization (SSO)

In this work, we used the Salp Swarm Optimization (SSO) method for feature selection to enhance the efficiency and performance of the AdaBoost model in breast cancer diagnosis. Designed on the swarming behavior of salps in the water, SSO is a bio-inspired method used for navigation and searching for the ideal collection of characteristics in high-dimensional datasets. Out of the initial thirty characteristics, we found the sixteen most important ones by using SSO on our dataset. These characteristics significantly help to differentiate benign from malignant tumors precisely, hence lowering the computational complexity of the model and still maintaining good prediction accuracy.

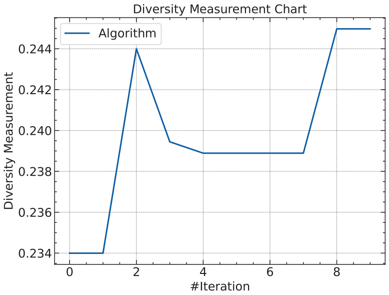


Fig. 3: Diversity Measurement Chart

Figure 3, the Diversity Measurement Chart, shows how population variety changed throughout many rounds of the optimization process. This variety guarantees the algorithm's exploring capacity, therefore preventing it from being caught in local minima throughout the feature-selecting phase. With peaks at iterations 2 and 9 representing the dynamic behavior of the algorithm during the search for the best features, the diversity displays oscillations across the iterations, as shown in the image.

Furthermore, shown on the Runtime Chart (Figure 4) is the iteration time taken for optimization. The runtime exhibits minor fluctuations among runs, with a clear rise at iteration five, peaking at around 4.6 seconds before steadying in following iterations. This runtime efficiency emphasizes that the SSO method is computationally controllable, which qualifies for

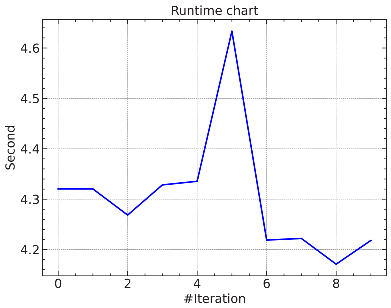


Fig. 4: Runtime Measurement Chart

feature selection in machine learning applications, including big datasets.

D. Performance of Sparrow Search Algorithm (SSA)

Following feature selection using the SSO method, we used the Sparrow Search Algorithm (SSA) to hyperparameter optimization of the AdaBoost model.

Figure 5a, the Diversity Measurement Chart, demonstrates how population variation falls with successive rounds. This slow decrease in variety suggests that the SSA method is heading toward a best answer. While subsequent rounds emphasize the exploitation of the most promising areas in the search space, early iterations exhibit more variety, suggesting the study of a broad range of possible solutions.

During the optimization phase, the Exploration vs. Exploitation Chart (Figure 5b) shows even more the harmony between exploration and exploitation. The algorithm moves from exploration (blue line) to exploitation (green line) as the count rises. This change guarantees that the SSA focuses on improving the best solutions discovered throughout the process.

Furthermore shown in Figure 5c is the Global Objectives Chart, which demonstrates how the objective value—that of the AdaBoost model's error or loss—reaches a steady minimum after few iterations. This suggests that SSA effectively tuned the hyperparameters to reduce the error rate of the model, hence enhancing its predictive ability.

At last, the runtime chart (Figure 5d) displays SSA's iteration-time consumption. More iterations cause a little increase in runtime; it peaks at iteration 8. The runtime stays within a reasonable range despite occasional variations, hence SSA is a time-efficient technique for hyperparameter tuning.

E. Performance of Proposed Model

We trained the AdaBoost model for breast cancer detection following the optimal hyperparameter acquisition using the Sparrow Search Algorithm (SSA). We computed the classification report, which comprises measures of precision, recall, and F1-score for every class, to assess its performance (Table I).

With an overall accuracy of 99%, the model performed rather well—that is, it correctly identified 99% of the samples. The model showed an accuracy of 0.99, a recall of 1.00, and an F1-score of 0.99 for the benign class (label 0), hence effectively identifying almost all benign instances free from numerous false positives. Reflecting its capacity to precisely identify malignant instances, with just a tiny number of false negatives, the model attained an accuracy of 1.00, a recall of 0.98, and an F1-score of 0.99 for the malignant class (label 1).

TABLE I: Classification Report

	Precision	Recall	F1-score	Support
0	0.99	1.00	0.99	71
1	1.00	0.98	0.99	43
accuracy			0.99	114
macro avg	0.99	0.99	0.99	114
weighted avg	0.99	0.99	0.99	114

Plotting the confusion matrix (Figure 6) helped us evaluate the improved AdaBoost model even further. The matrix offers an unambiguous picture of the forecasts of the model. With no false positives (top-left cell), the program properly identified all 71 benign instances. With only one false negative (bottom-left cell), the model correctly recognized 42 out of 43 instances for malignant cases, therefore misclassifying just one malignant case as benign. This validates the great accuracy and potency of the model in differentiating benign from malignant breast cancer patients.

F. Comparative Analysis

We evaluated our suggested method's performance against several different machine learning models, including SVM, Logistic Regression, Gradient Boosting, Extra Trees, K-Nearest Neighbors, Naive Bayes, XGBoost, CatBoost, LightGBM, and Random Forest. Key measures including accuracy, precision, recall, and F1-score formed the basis of the assessment (Table II).

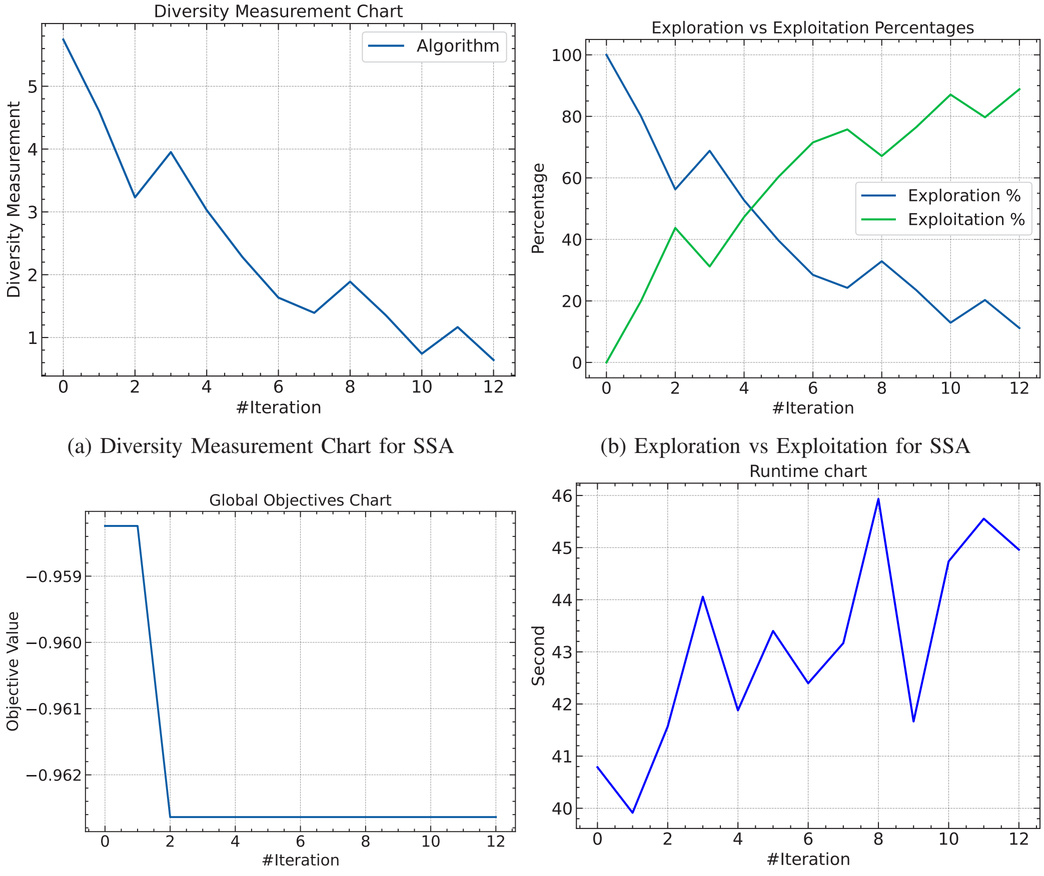
With a 99.12% accuracy level, the proposed model exceeded all others. With an accuracy of 98.25%, Extra Trees and XGBoost—the next best-performing models—also performed really well. At 97.37%, other models like Random Forest, Catboost, and LightGBM exhibited somewhat lower accuracy.

With regard to precision, the suggested model once more scored the highest—99.13%, suggesting that it generated rather few false positives. Closely behind with a 98.29% accuracy, XGBoost and Extra Trees.

With a recall of 99.12%, well above the competing models, the suggested model showed better performance for recall, which gauges the capacity to properly detect positive instances (malignant).

Reflecting its general balanced performance in both precision and recall, the F1-score—which is the harmonic mean of accuracy and recall—was greatest for the suggested model at 99.12%, surpassing other models like Extra Trees and XGBoost.

These results in Table II show that our suggested method offers superior performance across all evaluation criteria compared to other machine learning models.



(a) Diversity Measurement Chart for SSA

(b) Exploration vs Exploitation for SSA

(c) Global Objectives Chart for SSA

(d) Runtime Chart for SSA

Fig. 5: Performance of Sparrow Search Algorithm (SSA)

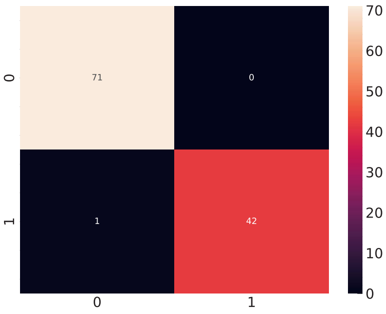


Fig. 6: Confusion Matrix

TABLE II: Comparative Analysis

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.9649	0.9652	0.9649	0.9647
Logistic Regression	0.9649	0.9668	0.9649	0.9645
Gradient Boosting	0.9737	0.9737	0.9737	0.9736
Extra Trees	0.9825	0.9829	0.9825	0.9824
K-Nearest Neighbors	0.9737	0.9737	0.9737	0.9736
Naive Bayes	0.9649	0.9668	0.9649	0.9645
XGBoost	0.9825	0.9829	0.9825	0.9824
CatBoost	0.9737	0.9737	0.9737	0.9736
LightGBM	0.9737	0.9737	0.9737	0.9736
Random Forest	0.9737	0.9748	0.9737	0.9735
Proposed Model	**0.9912**	**0.9913**	**0.9912**	**0.9912**

V. Conclusion

Using AdaBoost, enhanced by SSO for feature selection and SSA for hyperparameter tuning, we presented an AI-enabled breast cancer diagnostic model in this work. We improved the model's prediction performance and efficiency by selecting the 16 most important features with the help of SSO. Overcoming various models like XGBoost, LightGBM, and Random Forest, the suggested approach obtained an accuracy of 99.12%. In the future, we will plan to test the model or wider database and in real-world scenarios.

Acknowledgment

This research work is supported by National Science and Technology Council (NSTC), Taiwan Grant No. NSTC112-2221-E-468-008-MY3.

References

[1] P. Kumari, A. Shankar, A. Behl, V. Pereira, D. Yahiaoui, B. Laker, B. B. Gupta, and V. Arya, "Investigating the barriers towards adoption and implementation of open innovation in healthcare," *Technological Forecasting and Social Change*, vol. 200, p. 123100, 2024.

[2] B. B. Gupta, M. D. Lytras *et al.*, "Fog-enabled secure and efficient fine-grained searchable data sharing and management scheme for iot-based healthcare systems," *IEEE Transactions on Engineering Management*, 2022.

[3] T. Ball, "Diagnostic imaging of late-stage breast cancer," *Journal of Diagnostic Medical Sonography*, vol. 28, pp. 152–156, 2012.

[4] J. Yao, "Progress in the application of artificial intelligence in ultrasound diagnosis of breast cancer," *Frontiers in Computing and Intelligent Systems*, vol. 6, pp. 56–59, 2023.

[5] J. Chen, "Attention gate and dilation u-shaped network (gdunet): an efficient breast ultrasound image segmentation network with multiscale information extraction," *Quantitative Imaging in Medicine and Surgery*, vol. 14, pp. 2034–2048, 2024.

[6] A. Bhushan, A. Gonsalves, and J. Menon, "Current state of breast cancer diagnosis, treatment, and theranostics," *Pharmaceutics*, vol. 13, p. 723, 2021.

[7] V. Vijayasarveswari, N. Mahrom, R. A. A. Raof, L. A. E. K. Phak, M. A. Ab Razak, B. P. Silan, A. A. Abdul Halim, M. W. Nasrudin, N. Ramli, and Y. Rahayu, "Development of statistically modelled feature selection method for microwave breast cancer detection," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 50, no. 1, p. 250 – 263, 2025.

[8] B. B. Gupta, A. Gaurav, R. W. Attar, V. Arya, A. Alhomoud, and K. T. Chui, "A sustainable w-rlg model for attack detection in healthcare iot systems," *Sustainability*, vol. 16, no. 8, p. 3103, 2024.

[9] M. S. Suhiman, S. M. Deni, A. Z. U.-S. M. Japeri, A. Asmat, and L. Wang, "Classification of breast cancer subtypes using microarray rna expression data," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 46, no. 1, p. 75 – 85, 2025.

[10] S. Kaushik and C. Gandhi, "Capability-based access control with trust for effective healthcare systems," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 12, no. 1, pp. 1–28, 2022.

[11] A. Onyebuchi, U. O. Matthew, J. S. Kazaure, N. U. Okafor, O. D. Okey, P. I. Okochi, J. F. Taiwo, and A. O. Matthew, "Business demand for a cloud enterprise data warehouse in electronic healthcare computing: Issues and developments in e-healthcare cloud computing," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 12, no. 1, pp. 1–22, 2022.

[12] A. Akbar, S. Han, N. U. Rehman, R. Irshad, K. Ahmed, M. M. Ali, and A. A. Mazroa, "Reinforcement tokenization and graph convolution for high-precision breast tumor segmentation in dce-mri," *Biomedical Signal Processing and Control*, vol. 100, 2025.

[13] H. Q. Yu and S. Reiff-Marganiec, "Learning disease causality knowledge from the web of health data," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, no. 1, pp. 1–19, 2022.

[14] N. He, J. Duan, and J. Lyu, "Double kernel and minimum variance embedded broad learning system based autoencoder for one-class classification," *Neurocomputing*, vol. 611, 2025.

[15] J. Xiao, X. Liu, J. Zeng, Y. Cao, and Z. Feng, "Recommendation of healthcare services based on an embedded user profile model," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, no. 1, pp. 1–21, 2022.

[16] C. Zhang, P. Chen, and T. Lei, "Category-weight instance fusion learning for unsupervised domain adaptation on breast cancer histopathology images," *Biomedical Signal Processing and Control*, vol. 99, 2025.