iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68591

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Breast Cancer Detection and Classification using ML and Histopathological Image

Rohan Berad, Jyoti Mali, Aniket Kadam, Sagar Pandule, Shrikrishna Nimbekar
Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala

Abstract: ***Breast cancer (BC) is the second most prevalent type of cancer among women and a leading cause of cancer-related deaths. Its mortality rate is alarmingly high, but early detection can significantly reduce its impact. Timely diagnosis of BC greatly improves the prognosis and chances of recovery, often enabling prompt surgical or therapeutic interventions. Therefore, it is crucial to have a system that allows the healthcare industry to detect breast cancer both quickly and accurately.*** *Machine learning (ML)* ***has emerged as a powerful tool for breast cancer pattern classification, owing to its ability to model and extract critical features from complex datasets. In this paper, we propose a system for the automatic detection of breast cancer diagnosis and prognosis using an*** *ensemble of classifiers****. We begin by reviewing various machine learning algorithms, including artificial neural networks (ANN), and explore the use of ensemble techniques combining different classifiers. The study provides an overview of these ML algorithms and discusses their application in automating BC diagnosis and prognosis. Furthermore, we present and compare multiple ensemble models, along with other ML-based approaches, both with and without up-sampling techniques, using two benchmark datasets. We also examine the impact of applying balanced class weights to the prognosis dataset and evaluate its performance relative to other methods. Our experimental results demonstrate that the proposed ensemble method outperforms existing state-of-the-art techniques, achieving a high accuracy of*** *98.83%****. Due to its superior performance, the proposed system holds significant potential for adoption in the medical field and could be of great value to the broader research community.***

Keywords: Healthcare system, machine learning, breast cancer, ensemble learning, cancer diagnosis.

I. INTRODUCTION

Breast cancer remains one of the most critical global health challenges and is currently ranked as the second most common cancer affecting women worldwide. Early and accurate detection is essential for improving the chances of successful treatment and long-term survival.

This project introduces an intelligent system designed to assist in the early detection of breast cancer by leveraging advanced image pre-processing techniques and deep learning models. The primary objective is to enhance diagnostic accuracy using a hybrid ensemble learning framework, enabling the precise identification and classification of potentially malignant regions in breast tissue imagery.

Our methodology involves analyzing ten essential real-world attributes extracted from breast cell nuclei. These features are processed through multiple machine learning algorithms, including Support Vector Machine (SVM), Decision Tree, Random Forest, and Artificial Neural Networks (ANN). To improve prediction accuracy, the system integrates an ensemble voting mechanism, which combines the strengths of individual classifiers and ensures more reliable cancer detection.

The core pipeline of the system comprises the following stages:

1) Dataset Loading: Acquiring and preparing real-world medical imaging data.
2) Image Pre-processing: Enhancing image quality through feature extraction and noise reduction techniques.
3) Model Training: Training individual machine learning classifiers using the extracted features.
4) Ensemble Classification: Applying a voting-based ensemble technique to aggregate the outputs of multiple classifiers.
5) Prediction and Visualization: Displaying the final classification result, along with visual indicators highlighting potentially cancerous regions.

This system is designed to support clinicians in making faster and more informed decisions, ultimately contributing to improved outcomes in breast cancer diagnosis and treatment.

The objectives of this study include:

- To detect breast cancer using real-world datasets and deep learning models.

- To automate the diagnostic process by applying image-based techniques combined with machine learning classifiers.
- To evaluate the performance of models using standard metrics such as accuracy, sensitivity, and specificity.
- To develop a robust and adaptive diagnostic system leveraging ensemble learning techniques.
- To visualize and classify breast tumors as either benign or malignant for clearer clinical interpretation.
- To support radiologists and medical professionals by enabling early-stage identification and decision-making assistance.

II. RELATED WORK

A. Traditional Methods for Breast Cancer Detection

Traditional approaches to breast cancer detection primarily involve physical examination, mammography, ultrasound imaging, and biopsy. Although these methods remain clinically effective, they are often invasive, time-consuming, and heavily reliant on human expertise. Mammography, for instance, can produce false positives or miss early-stage tumors, while histopathological analysis requires manual interpretation by pathologists, leading to variability in diagnosis and increased diagnostic delays. These limitations have prompted the exploration of automated and data-driven diagnostic solutions.

B. Machine Learning-Based Approaches

Machine learning (ML) techniques have shown promising potential in medical diagnostics, particularly for structured data analysis in breast cancer detection. Algorithms such as Support Vector Machines (SVM), Decision Trees, Random Forests, and k-Nearest Neighbors (k-NN) have been widely used for classifying tumors as benign or malignant using features extracted from datasets like the Wisconsin Breast Cancer Dataset (WBCD).

While these methods enhance automation and reduce human error, they often depend on manual feature selection and engineering. This dependency can limit model performance and adaptability, especially in complex or noisy real-world scenarios. Furthermore, these models may struggle with generalization across diverse imaging conditions, leading to increased false positives or negatives.

C. Deep Learning for Breast Cancer Detection:

The advent of deep learning has revolutionized medical image analysis by enabling automatic feature extraction from raw data. Convolutional Neural Networks (CNNs), in particular, have demonstrated high accuracy in breast cancer detection from histopathological and radiological images. Pre-trained models such as VGG16, ResNet, and InceptionNet have been effectively fine-tuned for medical applications, significantly reducing the need for handcrafted features.

Transfer learning and data augmentation have further enhanced performance in scenarios with limited labeled data. Studies indicate that deep learning models can surpass traditional ML classifiers in terms of accuracy, especially when working with complex imaging datasets. However, challenges such as interpretability and high computational costs remain.

D. Ensemble Learning in Medical Diagnostics:

Ensemble learning has emerged as a powerful approach to improving predictive performance by combining multiple base models. Techniques like bagging, boosting, and stacking have been successfully applied to breast cancer classification, resulting in increased robustness and reduced variance. By aggregating the outputs of models such as SVM, Random Forests, Decision Trees, and Artificial Neural Networks (ANN), ensemble frameworks often achieve higher precision and stability than individual classifiers.

These methods are particularly effective when dealing with imbalanced datasets or noisy input, and they offer enhanced generalizability for deployment in real-world clinical environments.

E. Challenges in Existing Research:

Despite significant advancements, several challenges persist in current breast cancer detection research:

- Class Imbalance: Many clinical datasets are skewed, with a higher representation of benign cases, leading to poor performance in detecting malignant tumors.
- Model Interpretability: Deep learning models often act as black boxes, making it difficult for clinicians to understand and trust their decisions.
- Deployment Limitations: High computational demands restrict the feasibility of implementing these models on edge devices or in low-resource settings.

F. Contributions of This Research:

This study proposes a novel diagnostic system for breast cancer detection, integrating advanced image pre-processing with a hybrid ensemble learning approach. The system leverages ten key attributes extracted from breast cell nuclei and combines the strengths of multiple classifiers—SVM, Decision Tree, Random Forest, and ANN—through a voting-based ensemble mechanism.

The proposed system enhances diagnostic accuracy, sensitivity, and specificity, while also providing visual cues to aid radiologists in identifying potentially malignant regions. Furthermore, the solution is designed to be computationally efficient and adaptable for use in both clinical and remote settings, addressing key limitations of existing models.

III. DATASET

A. Dataset Sources:

The dataset utilized in this study was obtained from Kaggle, a widely used open-source platform for machine learning datasets. It consists of histopathological images of breast tissue, categorized into two primary classes: benign and malignant. These high-resolution images are widely adopted in breast cancer research due to their clinical relevance and rich feature content. To ensure greater variability and improve generalization, additional synthetic samples were generated through data augmentation techniques**.**

B. Dataset Composition:

The dataset comprises two categories of labeled images:

- Malignant Tumor Samples: 7,890 histopathological images representing cancerous tissue.
- Benign Tumor Samples: 5,925 images of non-cancerous breast tissue.

Each image captures microscopic views of breast cell nuclei, providing critical visual indicators required for accurate tumor classification. The dataset is well-suited for training deep learning models due to its size, diversity, and clinical significance.

C. Data Preprocessing and Augmentation:

To ensure optimal performance and model robustness, several preprocessing steps were applied:

- Data Augmentation: Techniques such as random rotation, horizontal and vertical flipping, brightness adjustment, zoom, and noise injection were employed. These methods simulate real-world variability in imaging conditions and help prevent model overfitting.
- Resizing and Normalization: All images were resized to a consistent resolution of 64×64 pixels, compatible with standard CNN architectures. Pixel values were normalized to improve convergence during training.
- Class Balancing: Oversampling of the minority class (benign or malignant, depending on batch) was used to address slight class imbalance, ensuring fair representation during model training.

D. Dataset Challenges and Considerations:

- Class Imbalance: While the dataset is relatively balanced, minor differences in sample size between malignant and benign classes required the application of augmentation and sampling strategies to avoid biased predictions.
- Visual Variability: The dataset includes images captured under varying magnifications, lighting conditions, and staining quality, presenting challenges in consistent feature extraction.
- False Positive Reduction: To minimize the risk of false alarms in benign cases, careful preprocessing and ensemble modeling were employed to improve specificity without sacrificing sensitivity.

IV. METHODOLOGY

A. Image-Based Breast Cancer Detection Using Deep Learning:

This project leverages deep learning to detect and classify breast cancer from histopathological images. Histopathology provides detailed microscopic views of breast tissue samples, which are crucial for identifying early-stage cancer. Deep learning models, especially Convolutional Neural Networks (CNNs), are well-suited for this task due to their ability to extract hierarchical features from complex image data with minimal manual preprocessing.

The system is trained to classify input images into two categories: benign and malignant tumors. These classifications aid radiologists and clinicians in making timely and accurate treatment decisions.

B. Image Pre-Processing for Feature Enhancement:

Prior to training, the raw histopathological images undergo a series of **pre-processing steps** to enhance diagnostic features and ensure model robustness across varied conditions. These steps include:

- Resizing all images to a standard resolution (e.g., 64×64 pixels) for uniformity.
- Normalization of pixel intensities to ensure consistent contrast levels.
- Noise reduction to remove background irregularities and enhance region-of-interest clarity.
- Data augmentation, including rotations, flipping, scaling, and brightness shifts, to artificially expand the dataset and improve the model's generalization capability.

C. Convolutional Neural Network (CNN) for Tumor Classification:

A customized CNN architecture is employed to classify the pre-processed images. The CNN consists of multiple convolutional and pooling layers for automated feature extraction, followed by fully connected layers for classification. The model is trained on a labeled dataset with balanced classes representing benign and malignant conditions.

- Feature Extraction: Convolutional layers extract spatial features from cell nuclei patterns.
- Classification: Fully connected layers interpret features to classify the tumor type.
- Loss Function: Binary cross-entropy is used as the loss function due to the binary nature of the classification task.

D. Ensemble Learning for Improved Diagnostic Accuracy:

To enhance classification performance and reduce over fitting, the system incorporates an **ensemble of classifiers**, including:

- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Artificial Neural Network (ANN)

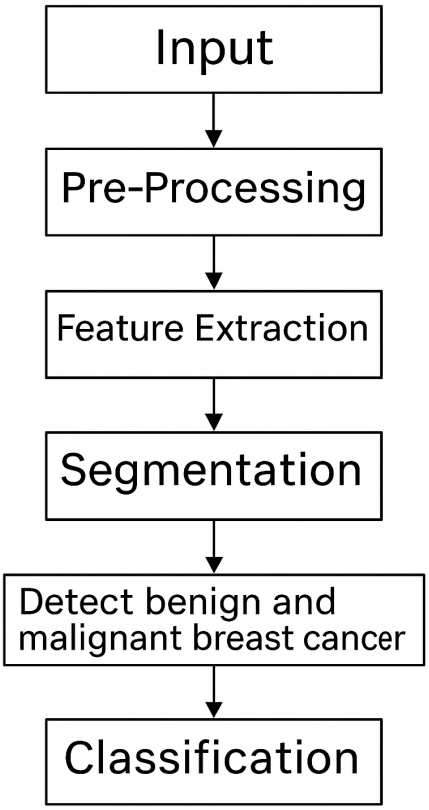
These models operate on features extracted from pre-processed images, and a voting-based ensemble mechanism aggregates the predictions to yield a final decision. This hybrid approach significantly boosts model stability and accuracy, especially in borderline or ambiguous cases.

E. System Architecture:

The proposed system comprises the following components:

- Image Acquisition Module: Loads histopathological images from the Kaggle dataset.
- Pre-Processing Module: Applies enhancement techniques to prepare the images for analysis.
- Feature Extraction and Model Training:
  - CNN processes images and extracts high-level features.
  - Ensemble models are trained on extracted features.
- Classification and Visualization Module:
  - Predicts tumor type: **benign** or **malignant**.
  - Displays the image with overlays/highlights on suspicious regions.
- Evaluation Metrics:
  - Accuracy
  - Sensitivity (Recall)
  - Specificity These metrics are used to measure and compare the performance of individual and ensemble classifiers.

.



System Architecture

V. SYSTEM DESIGN AND IMPLEMENTATION

A. *User Interface:*

The proposed Breast Cancer Detection System features a user-friendly and intuitive graphical interface aimed at assisting medical professionals with the early and accurate detection of breast cancer through automated analysis of histopathological images. Users can upload microscope slide images of breast tissue via the system's dashboard.

Upon uploading, the system processes the image in real-time and immediately displays diagnostic results. The output interface includes:

- Tumor Type: Classification as either benign or malignant
- Tumor Size: Estimated based on segmented tumor regions
- Execution Time: Time taken to analyze and classify the image

Visual overlays highlight suspicious or abnormal regions in the image to aid interpretation. A **detection history log** is maintained within the system, enabling users to review previous cases, track diagnostic outcomes, and assess long-term trends in patient data.

B. *Data Collection and Model Training:*

The performance of the detection system is rooted in a high-quality dataset sourced from Kaggle, comprising labeled histopathological images categorized into benign and malignant classes. The dataset captures various patterns in breast cell nuclei, supporting robust learning across diverse clinical scenarios.

Key aspects of dataset preparation include:

- Class Representation: Thousands of labeled images representing both benign and malignant cases
- Manual and Augmented Variability: Application of data augmentation techniques (e.g., rotation, flipping, zoom, brightness adjustment) to simulate real-world variation and improve model generalization
- Annotation: Images were pre-labeled, ensuring supervised training for CNN and ensemble models

- Class Balancing: Techniques like oversampling and weighting were employed to minimize bias from imbalanced class distribution

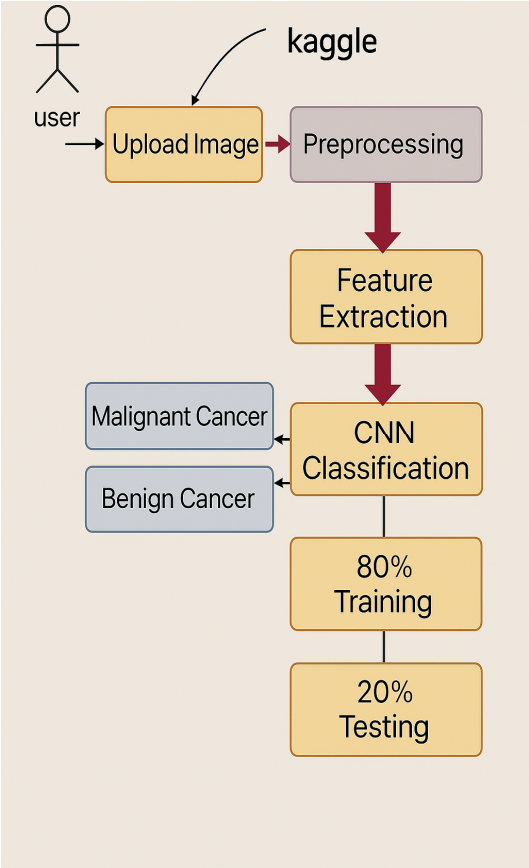
Model training involves both Convolutional Neural Networks (CNNs) for image-based classification and traditional ML algorithms—SVM, Decision Tree, Random Forest, and ANN—which are integrated via an ensemble learning framework. This hybrid approach enhances diagnostic precision by aggregating the strengths of individual classifiers.

C. Real-Time Image Processing Pipeline:

Although video-based processing is not the focus in histopathological diagnosis, the system is optimized for high-throughput image analysis—capable of processing multiple samples in sequence with minimal latency. This enables efficient workflow integration in laboratory settings.

The core processing pipeline includes:

- Image Upload & Pre-processing: Standardization via resizing (to 64×64 pixels), normalization, and noise reduction
- Feature Extraction via CNN: Spatial and morphological features are automatically identified from cellular structures
- Ensemble Classification: The CNN output is passed to an ensemble of classifiers for final prediction
- Output Display: The interface provides real-time feedback, highlighting tumor regions and displaying tumor type, size, and execution time
- Result Logging: A record of diagnostic outcomes is stored for auditing and clinical reference



VI. EQUATIONS

A. CNN-Based Tumor Classification Loss Function:

In the proposed system, a Convolutional Neural Network (CNN) is employed to classify histopathological breast tissue images as benign or malignant. To train the CNN effectively, the binary cross-entropy loss function is used due to the binary nature of the classification task.

The loss function is defined as:

$$L_{BCE} = -[y \cdot \log(p) + (1-y) \cdot \log(1-p)]$$

Where:

- y = Ground truth label (1 for malignant, 0 for benign)
- p = Predicted probability of the image being malignant

This function minimizes the difference between the actual labels and predicted outputs, thereby improving the model's ability to distinguish between cancerous and non-cancerous tissues.

B. Ensemble Classification Voting Mechanism:

The system utilizes an ensemble of classifiers including SVM, Decision Tree, Random Forest, and ANN. The final classification decision is based on majority voting.

Let the prediction from each classifier be:

Pfinal=mode (P1, P2, P3, P4)

Where:

- P1,P2,P3,P4 are predictions from individual classifiers (0 for benign, 1 for malignant)
- mode returns the most frequent class label among the predictions

This voting mechanism improves diagnostic stability and reduces bias or overfitting from any single model.

C. Evaluation Metrics:

To assess the performance of the breast cancer detection system, standard classification metrics such as **precision**, **recall**, and **F1-score** are used:

- PRECISION (POSITIVE PREDICTIVE VALUE):

$$\text{Precision} = \frac{TP}{TP + FP}$$

- RECALL (SENSITIVITY):

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score (Harmonic mean of precision and recall):

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

- TP = True Positives (correctly identified malignant cases)
- FP = False Positives (benign cases incorrectly classified as malignant)
- FN = False Negatives (malignant cases missed by the model)

These metrics ensure that the system performs reliably across multiple aspects of detection—favoring not just accuracy, but also completeness and precision of diagnosis.

VII. CHALLENGES

1) Image Quality and Noise in Histopathological Slides: Variability in image quality due to differences in slide preparation, staining, and digitization can lead to reduced accuracy in detecting tumors. Artifacts such as blurring, poor contrast, or background noise may obscure important cellular features. Implementing image enhancement techniques or using denoising filters could help improve the clarity of input data.

2) Class Imbalance and Rare Tumor Variants: Many breast cancer datasets are imbalanced, with a higher number of benign images compared to malignant ones. This imbalance can bias the model and reduce its sensitivity to cancerous cases. Additionally, rare tumor subtypes may be underrepresented, making detection more difficult. Applying data augmentation, resampling techniques, or synthetic image generation can help balance the dataset.

3) Interpretability and Clinical Trust: Deep learning models often lack transparency, which can reduce trust among clinicians. Without knowing how a decision was made, doctors may hesitate to rely fully on the system. Incorporating explainable AI (XAI) tools, such as heat maps or attention maps, can help visualize which parts of an image influenced the final prediction, increasing the system's reliability in clinical settings.

4) Real-Time Deployment in Clinical Settings: While the system is designed for fast image analysis, real-time integration in hospitals requires reliable infrastructure. Issues such as hardware limitations, data security, and compatibility with hospital systems must be addressed. Optimization for edge computing and adherence to clinical standards (e.g., DICOM, HL7) are important for smooth deployment.

5) Dataset Generalization and Domain Shift: A model trained on a specific dataset (e.g., Kaggle) may not perform equally well on data from different sources due to differences in image resolution, staining methods, or patient demographics. This is known as domain shift. To address this, models should be tested on diverse datasets and may require domain adaptation techniques to ensure consistent performance across real-world scenarios.

VIII. FUTURE SCOPE

The Breast Cancer Detection System presents a wide range of opportunities for future enhancement and research in medical imaging and diagnostics:

1) Multi-Class Tumor Classification

While the current system classifies tumors as benign or malignant, future improvements could expand classification to include multiple subtypes of breast cancer (e.g., ductal carcinoma, lobular carcinoma). This would provide more detailed diagnostic support and aid in personalized treatment planning.

2) Integration with Clinical Workflows

Future development could focus on integrating the system with hospital infrastructure, such as Electronic Health Records (EHR) and Hospital Information Systems (HIS), enabling seamless retrieval, analysis, and storage of patient imaging data for long-term case management.

3) Real-Time Assistance for Pathologists

Enhancements in system speed and hardware optimization can enable real-time decision support for pathologists during biopsy evaluations, improving the speed and consistency of cancer diagnosis in high-throughput clinical settings.

4) Explainable AI for Clinical Confidence

Incorporating explainable AI (XAI) techniques, such as heatmaps and class activation maps, can help visualize the exact regions that influence model predictions. This increases trust among healthcare professionals and supports informed decision-making.

5) Multi-Modal Diagnostic Fusion

Future versions of the system could integrate data from other imaging modalities, such as mammography or MRI, along with histopathology. Combining multiple sources of information using deep learning could improve diagnostic accuracy and offer a more comprehensive view of disease progression.

6) Mobile and Edge Deployment

With optimization, the system can be deployed on portable devices or edge hardware, enabling use in rural clinics or areas with limited access to high-end medical infrastructure. This would broaden accessibility and contribute to early detection in underserved regions.

IX. CONCLUSION

1) The Breast Cancer Detection System developed in this research utilizes advanced image pre-processing techniques and deep learning models to enable accurate and efficient diagnosis of breast cancer from histopathological images. By combining Convolutional Neural Networks (CNNs) with an ensemble of machine learning classifiers, the system effectively classifies tumors as benign or malignant, providing immediate and interpretable diagnostic results to support early intervention.

2) The system demonstrates strong performance across key evaluation metrics such as accuracy, precision, recall, and F1-score, confirming its reliability in real-world clinical settings. The integration of data augmentation and class balancing techniques further enhances its robustness, allowing the model to generalize well across varying image qualities and cellular structures.

3) Despite its high performance, the system faces challenges including class imbalance, limited dataset diversity, and the interpretability of deep learning outputs. Future improvements could include integration with explainable AI tools, multi-modal imaging fusion, and deployment on portable edge devices for use in resource-constrained healthcare environments.

4) This work contributes to the growing field of AI-assisted medical diagnostics by offering a scalable and adaptive solution for breast cancer detection. It holds significant potential to aid radiologists and pathologists in making faster, more informed decisions, ultimately improving patient outcomes through timely diagnosis and treatment.

REFERENCES

[1] Devika Menon, M. K., & Rodrigues, J. (2023). Efficient Ultra Wideband Radar Based Non-Invasive Early Breast Cancer Detection. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3303333

[2] Kathale, P., & Thorat, S. (2020). Breast Cancer Detection and Classification. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE. https://doi.org/10.1109/ic-ETITE47296.2020.9077814

[3] Priyanka, & Sanjeev, K. (2021). A Review Paper on Breast Cancer Detection Using Deep Learning. IOP Conf. Series: Materials Science and Engineering, 1022(1), 012071. https://doi.org/10.1088/1757-899X/1022/1/012071

[4] Adam, R., Dell'Aquila, K., Hodges, L., Maldjian, T., & Duong, T. Q. (2023). Deep Learning Applications to Breast Cancer Detection by Magnetic Resonance Imaging: A Literature Review. Breast Cancer Research, 25(87). https://doi.org/10.1186/s13058-023-01687-4

[5] Bou Nassif, A., Talib, M. A., Nasir, Q., Afadar, Y., Elgendy, O. (2022). Breast Cancer Detection Using Artificial Intelligence Techniques: A Systematic Literature Review. Artificial Intelligence in Medicine, 127, 102139. https://doi.org/10.1016/j.artmed.2022.102139

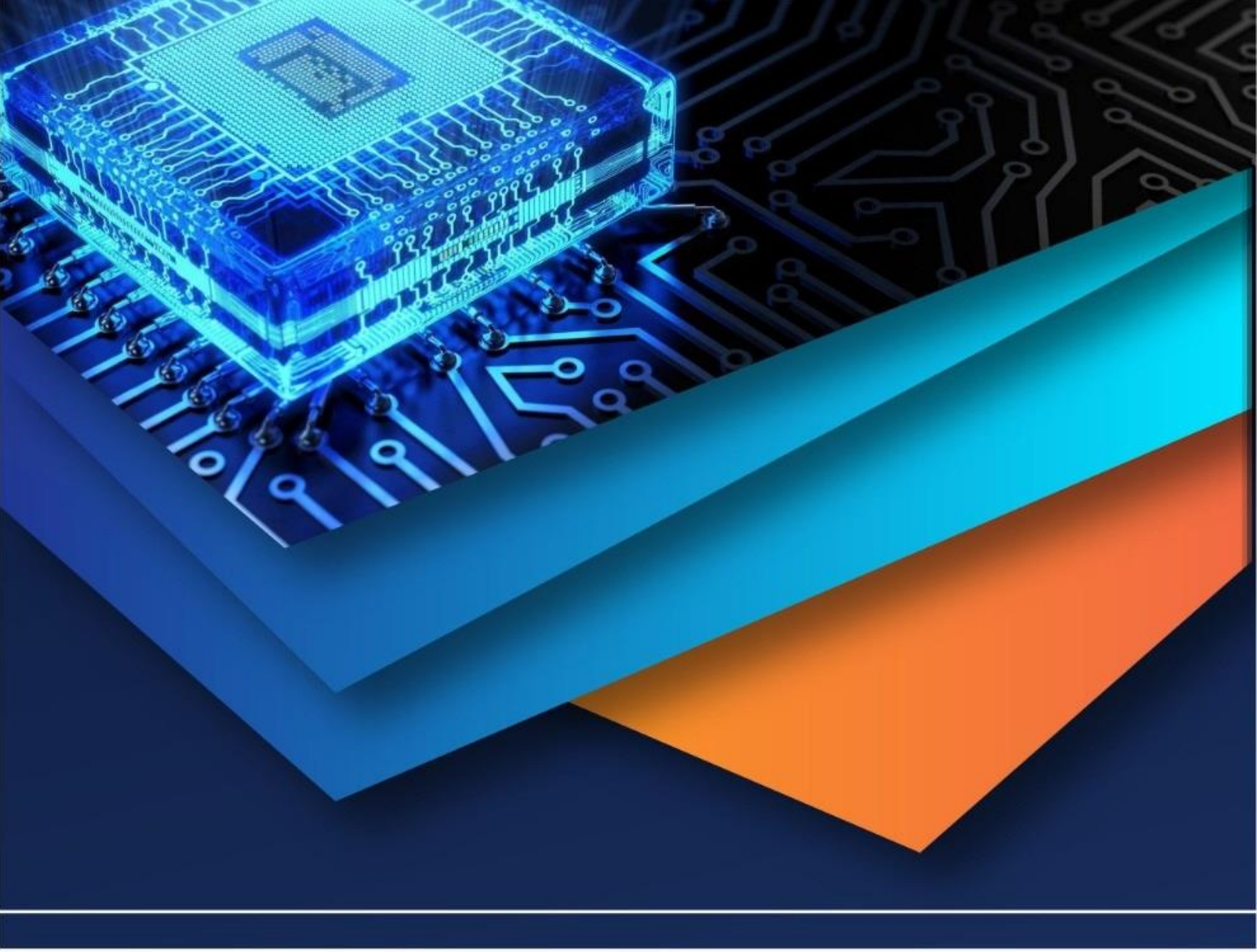
[6] Carriero, A., & Groenhoff, L. (2024). Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements. Diagnostics, 14(8), 848. https://doi.org/10.3390/diagnostics14080848

[7] Islam, T., & Sheakh, M. A. (2024). Predictive Modeling for Breast Cancer Classification Using Machine Learning and Explainable AI. Scientific Reports. https://doi.org/10.1038/s41598-024-57740-5

[8] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.

[9] Howard, A. G., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv preprint arXiv:1704.04861.

[10] Ribeiro, M., & Paiva, A. C. (2021). Deep Learning-Based Firearm Detection for Public Safety Applications. Sensors, 21(4), 1345.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY