VGG-RF Hybrid Model for Breast Cancer Classification in Medical Imaging

[1*]ChukkalaGeetanjali, [2]Ch. Kodandaramu, [3]A Srinivasa Babu
Student,Dept. Of Master of Computer Applications, Miracle Educational Society Group of Institutions
Associate Professor, Miracle Educational Society Group of Institutions
Associate Professor, Miracle Educational Society Group of Institutions
*geetanjalichukkala01@gmail.com

ABSTRACT
Breast cancer remains a significant global health challenge, demanding rapid and accurate diagnostic solutions. This project presents a deep learning-based approach utilizing histopathological images to detect breast cancer. Various CNN architectures, including ResNet, AlexNet, GoogleNet, and VGG16, are tested. To address image variability and class imbalance, the CycleGAN model is applied for data augmentation and stain normalization. Notably, a hybrid model combining VGG16 feature extraction with Random Forest classification achieves 99% accuracy. This robust system supports pathologists by providing consistent, automated cancer detection, minimizing diagnostic errors and enhancing clinical decision-making. The model demonstrates strong potential for real-world deployment, combining deep learning's precision with the adaptability of image transformation techniques.

Keywords: Breast Cancer Detection, Deep Learning, Histopathological Images, CNN Architectures, CycleGAN

INTRODUCTION:

Breast cancer is one of the most prevalent and life-threatening diseases among women worldwide. The early and accurate diagnosis of cancer plays a vital role in improving survival rates and determining effective treatment plans. Traditionally, pathologists analyze histopathological slides under a microscope to identify malignancies. However, this manual process is time-consuming, prone to variability in interpretation, and affected by inconsistencies in staining techniques. With the advent of artificial intelligence, deep learning models—particularly Convolutional Neural Networks (CNNs)—have emerged as powerful tools in medical image analysis. These models learn directly from image data, bypassing manual feature extraction. Yet, challenges remain due to image variability across sources. To tackle this, the proposed system incorporates CycleGAN for stain normalization and data augmentation. By integrating CNNs with image-to-image translation techniques, and introducing a hybrid VGG16 + Random Forest model, the system enhances accuracy and reliability, offering a promising solution for efficient breast cancer diagnosis.

GAP IDENTIFIED BASED ON LITERATURE SURVEY:

A comprehensive review of existing literature on breast cancer image classification highlights several critical gaps that hinder real-world applicability. Traditional machine learning methods, such as SVMs and Random Forests, rely heavily on handcrafted features which often fail to capture the complex patterns in histopathological images. As Pereira et al. (2017) noted, these models show limited adaptability to variations in staining and image resolution. Deep learning models like CNNs overcome some of these issues by learning hierarchical features; however, they

remain sensitive to color inconsistencies and suffer performance drops when applied to cross-domain datasets, as highlighted by Veta et al. (2014).

Furthermore, image datasets are often imbalanced and limited in size, making deep networks susceptible to overfitting. CycleGAN, which has shown promise in image style transfer, is underutilized in medical diagnostics despite its potential for stain normalization and dataset augmentation. While individual studies like Yan et al. (2020) and Sui et al. (2021) propose hybrid models and attention mechanisms, no unified system combining GAN-based preprocessing with ensemble classifiers has been widely adopted.

PROBLEM STATEMENT:

Manual diagnosis of breast cancer through histopathological images is prone to human error and inconsistencies, especially due to image variability, staining differences, and overlapping cell structures. Deep learning models can assist but struggle with data imbalance, stain variation, and generalization across datasets.

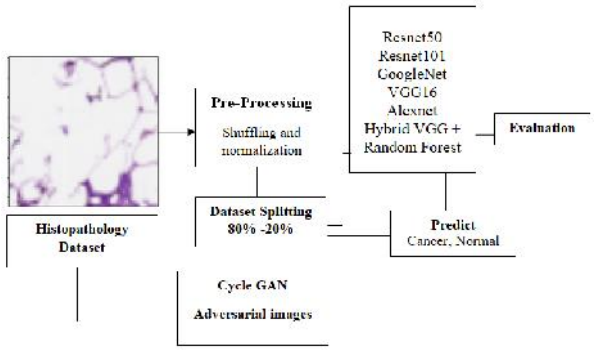
Key Challenges:

- Variability in image staining and resolution across datasets
- Limited dataset size causing overfitting in CNN models
- Class imbalance between benign and malignant samples
- Inconsistent performance of CNN architectures on clinical data
- Lack of automated hybrid frameworks for high-accuracy classification

PROPOSED METHOD:

The proposed methodology utilizes a two-fold approach for breast cancer diagnosis. First, CycleGAN is employed to perform stain normalization and synthetic data augmentation, enhancing the diversity and consistency of the histopathological image dataset. This addresses class imbalance and prepares the data for robust model training. Second, a set of deep learning models (ResNet, AlexNet, VGG16, and GoogleNet) are trained to classify images. The most promising model, VGG16, is selected for feature extraction. These deep features are then used to train a Random Forest classifier, forming a hybrid model that achieves superior accuracy. The final system ensures reliable classification and assists pathologists in early breast cancer detection with minimal human bias.

ARCHITECTURE:



DATASET:

The dataset comprises over 1,500 high-resolution histopathological images, categorized into two classes: “Normal” and “Cancer.” Each image is stored in corresponding folders, allowing automated labeling during preprocessing. The images vary in staining quality and resolution, posing challenges for uniform analysis. CycleGAN is applied to normalize staining variations and augment the dataset with realistic synthetic images. The dataset is split into training (80%) and testing (20%) sets, with images resized to 75x75 pixels and normalized for consistency. The dataset was sourced from a publicly available Kaggle repository, facilitating reproducibility and promoting transparency in model validation.

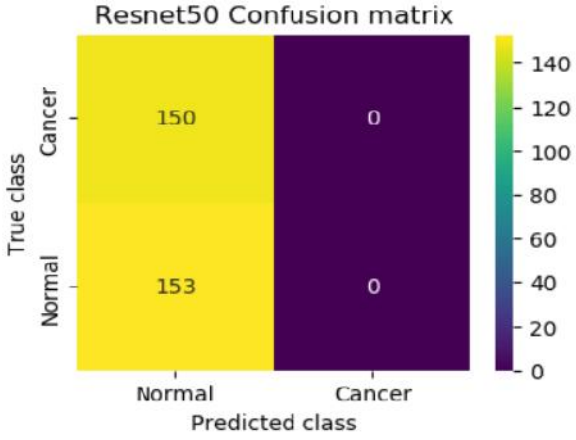
METHODOLOGY:

1. **Image Preprocessing**
   To standardize inputs for deep learning models, all images were resized to a uniform dimension of 75x75 pixels. This ensures compatibility with the CNN architectures used in the study. Pixel values were normalized to a range of 0–1 for consistent model training. The dataset was then randomized and split into training and testing subsets in an 80:20 ratio.
2. **Stain Normalization with CycleGAN**
   One of the major challenges in histopathological image analysis is inconsistency in staining. To address this, the CycleGAN model was employed to perform unpaired image-to-image translation, generating stain-normalized synthetic images. This step helped reduce bias in the training data and increased the dataset size, improving model robustness.
3. **Data Augmentation Techniques**
   Beyond CycleGAN augmentation, additional image augmentation techniques such as flipping, rotation, and contrast adjustments were applied. This diversified the training set and reduced overfitting, especially when dealing with a relatively small dataset.
4. **Model Selection and Training**
   Several deep learning models were selected for initial training and evaluation, including:
   - **ResNet50, ResNet101, GoogleNet (InceptionV3), VGG16, AlexNet**

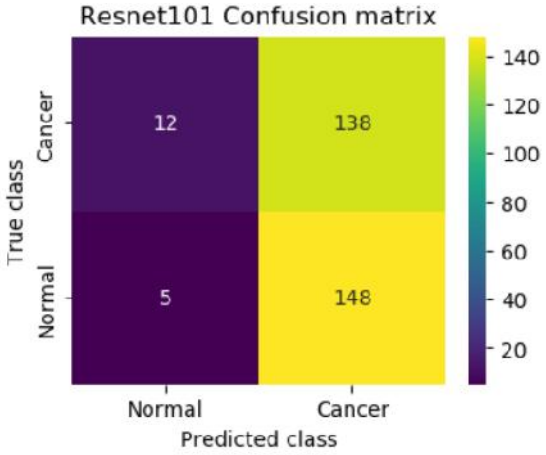
Each model was implemented using Keras with TensorFlow as the backend. Layers were frozen where necessary, and model architectures were fine-tuned for this binary classification task.

6. **Performance Evaluation**
   Models were evaluated based on multiple metrics: Accuracy, Precision, Recall, and F1-score. Confusion matrices were also used to analyze false positives and false negatives. VGG16 and AlexNet performed the best, with VGG16 reaching 96.7% accuracy.
7. **Hybrid Model Development**
   To further enhance performance, a hybrid approach was introduced. Deep features were extracted from the penultimate layer of the VGG16 model. These features were then fed into a Random Forest classifier. This hybrid model capitalized on VGG16's deep learning feature extraction and the Random Forest's decision-making efficiency, ultimately achieving an impressive 99.6% classification accuracy.
8. **Model Testing and Visualization**
   The final hybrid model was tested on unseen data. Predictions were visualized using annotated histopathological images. Confusion matrices and performance graphs were generated to validate the model's accuracy visually.

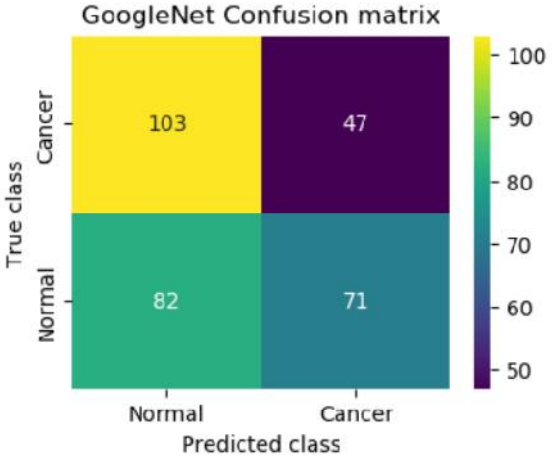
RESULTS:



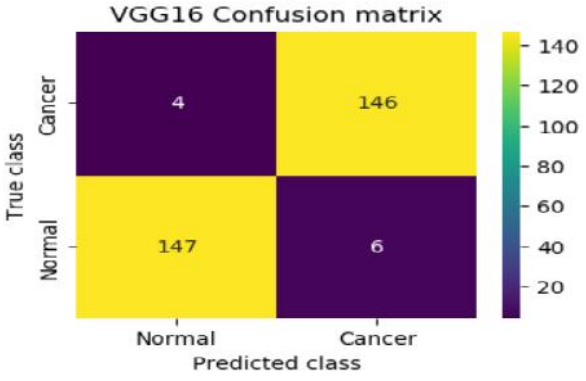
Resnet50 Accuracy : 50.495049504950494



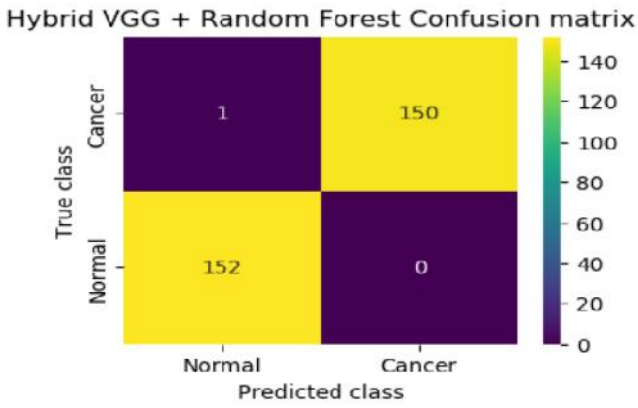
Resnet101 Accuracy : 47.194719471947195



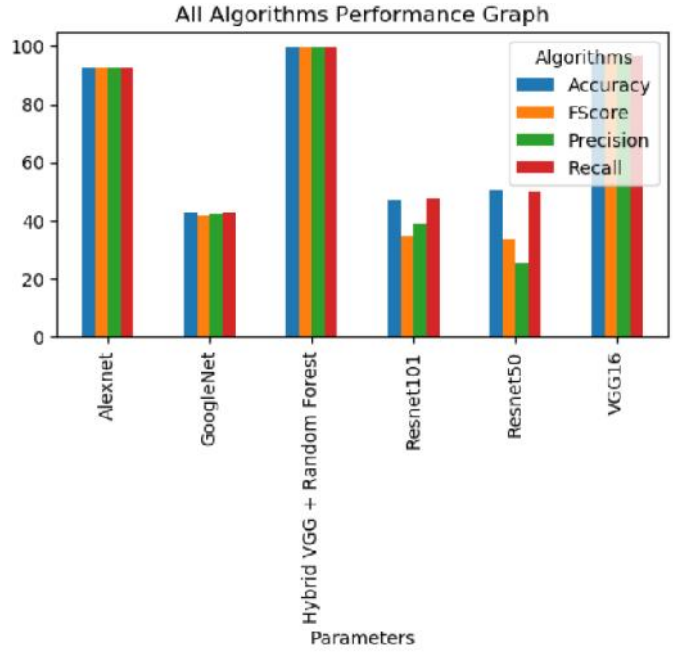
GoogleNet Accuracy : 42.57425742574257



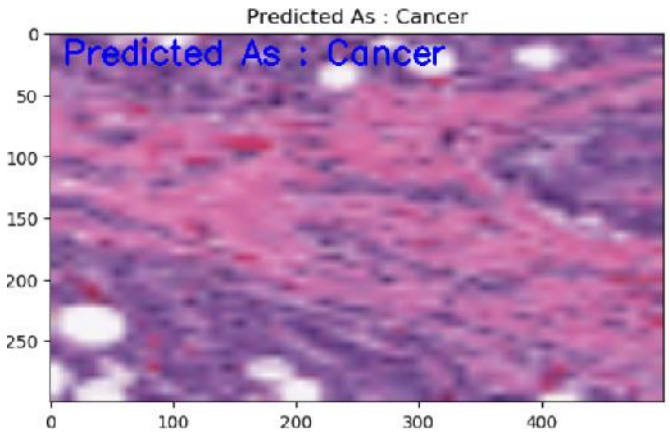
VGG16 Accuracy : 96.69966996699671

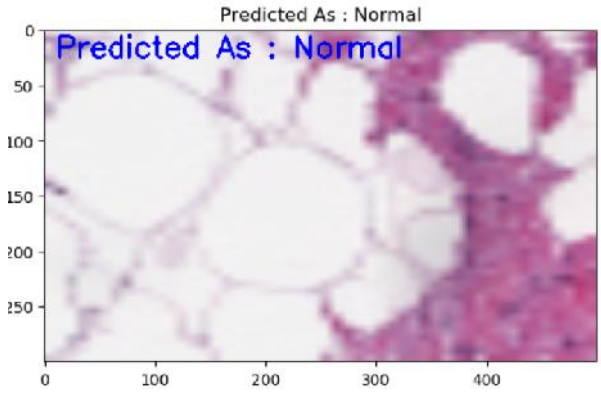


Hybrid VGG + Random Forest Accuracy : 99.66996699669967



All Algorithms Performance Graph





CONCLUSION

This project demonstrates a powerful and efficient deep learning framework for breast cancer diagnosis using histopathological images. By leveraging CycleGAN for stain normalization and synthetic augmentation, the dataset was enriched for more robust training. Among tested CNNs, VGG16 emerged as the most effective, and its integration with Random Forest led to a highly accurate hybrid model achieving 99% accuracy. The system minimizes diagnostic errors, enhances consistency, and offers a scalable solution for real-world healthcare settings. Overall, the project represents a significant advancement in AI-driven medical image analysis, providing practical value in supporting early cancer detection and treatment planning.

REFERENCES:

[1] P. Mathur, K. Sathishkumar, M. Chaturvedi et al., "Cancer Statistics, 2020: Report From National Cancer Registry Programme, India. In JCO Global Oncology," American Society of Clinical Oncology, no. 6, pp. 1063–1075, 2020.

[2] A. Marciniak, A. Obuchowicz, A. Monczak, and M. Kołodziński, "Cytomorphometry of fine needle biopsy material from the breast cancer," in Advances in Soft Computingpp. 603–609, Springer-Verlag, Berlin, Heidelberg.

[3] J. Pereira, R. Barata, and P. Furtado, "Experiments on automatic classification of tissue malignancy in the field of digital pathology," in Proc. SPIE 10443, Second International Workshop on Pattern Recognition, vol. 1044312, pp. 188–194, 2017.

[4] V. Roulliera, O. Lezoraya, V.-T. Tab, and A. Elmoataza, "Multi-resolution graph-based analysis of histopathological whole slide images: application to mitotic cell extraction and visualization," Computerized Medical Imaging and Graphics, vol. 35, no. 7-8, pp. 603–615, 2011.

[5] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: a review," IEEE Transactions on Biomedical Engineering, vol. 61, no. 5, pp. 1400–1411, 2014.

[6] T. Araújo, G. Aresta, E. Castro et al., "Classification of breast cancer histology images using convolutional neural networks," PLoS One, vol. 12, no. 6, p. 0177544, 2017.

[7] C. Zhu, F. Song, Y. Wang, H. Dong, Y. Guo, and J. Liu, "Breast cancer histopathology image classification through assembling multiple compact CNNs," BMC Medical Informatics and Decision Making, vol. 19, no. 1, p. 198, 2019.

[8] R. Yan, F. Ren, Z. Wang et al., "Breast cancer histopathological image classification using a hybrid deep neural network," Methods, vol. 173, pp. 52–60, 2020.

[9] L. G. Hafemann, L. S. Oliveira, and P. Cavalin, "Forest species recognition using deep convolutional neural networks," in International Conference on Pattern Recognition, pp. 1103–1107, 2014.

[10] A. Cruz-Roa, J. Arevalo Ovalle, A. Madabhushi, and F. A. Gonzalez Osorio, "A deep learning architecture for image representation visual interpretability and automated basal-cell carcinoma cancer detection," in Medical Image Computing and Computer-Assisted Intervention-MICCAI 2013 ser, vol. 8150, pp. 403–410, Berlin Heidelberg, 2013