Pixels to Prognosis: ResNet50 Analysis for Predicting Benign vs. Malignant Breast Cancer from Biopsy Scans

Naman Dhariwal
Department of Statistics
University of Michigan
Ann Arbor, MI, USA
namand@umich.edu

Dr. Xian Zhang
Department of Statistics
University of Michigan
Ann Arbor, MI, USA
xianz@umich.edu

***Abstract*—Breast cancer diagnosis is a critical task that can benefit from advancements in deep learning and computer vision. This study focuses on the analysis and application of the ResNet50 architecture to classify benign and malignant breast cancer types using RCL biopsy images. Three versions of the ResNet50 model were trained and evaluated using diverse hyperparameter configurations, optimization strategies, and learning rate schedulers to determine the most accurate and robust classifier. The performance of each model was assessed using metrics such as precision, recall, F1-score, and accuracy, with additional validation through k-fold cross-validation. ResNet50 (Classifier 3), utilizing the AdamW optimizer and ReduceLROnPlateau scheduler, emerged as the top performer with an average accuracy of 90% and balanced performance across all metrics. These findings underscore the potential of ResNet50, when optimized effectively, for reliable breast cancer classification. The research highlights the importance of model tuning and evaluation in developing robust AI-driven diagnostic tools.**

***Index Terms*—Breast Cancer Classification, ResNet50, Deep Learning, Medical Imaging, Hyperparameter Optimization, Computer Vision, Biopsy Image Analysis, Diagnostic AI.**

I. Introduction

Cancer is a class of disorders characterised by uncontrollable cell growth and the ability to metastasize to other regions of the body. It is caused by genetic abnormalities or disturbances in normal cellular processes, which may be induced by environmental, genetic, or lifestyle factors. These alterations lead to the dysregulation of mechanisms governing cell division, differentiation, and death.[1, 2] Breast cancer is a disease in which abnormal breast cells develop uncontrollably forming tumours. If left untreated, tumours can spread throughout the body and be fatal. Breast cancer originates in the milk ducts or lobules and may initially remain localized, posing no immediate threat. If invasive, it spreads to surrounding breast tissue, forming lumps or thickened areas, and can metastasize to lymph nodes or distant organs, which can become life-threatening [3].

A biopsy is a diagnostic procedure in which a small sample of tissue or cells is extracted from the body and examined under a microscope. It is often used to detect the presence, kind, and severity of diseases such as cancer by analysing cellular abnormalities. A Radiologically Controlled Localisation (RCL) breast cancer biopsy is a technique used to collect tissue samples from areas of concern in the breast that are difficult to identify by touch but apparent on imaging modalities such as mammography, ultrasound, or MRI [4].

Machine Learning (ML) has been increasingly applied to analyze RCL biopsy images, enhancing the diagnostic process by automating and improving the accuracy of cancer detection. ML models, particularly those based on deep learning, can identify patterns and anomalies in biopsy images that may not be readily apparent to human observers. These models aid in distinguishing between benign and malignant tissues, classifying cancer subtypes, and predicting tumor aggressiveness. Additionally, ML can streamline workflows by reducing the time required for manual analysis and minimizing diagnostic variability, contributing to more consistent and reliable patient outcomes [5, 6].

Deep Neural Networks (DNNs), like Convolutional Neural Networks (CNNs), process complex data by learning hierarchical features. CNNs are used in image tasks, using convolutional layers for feature extraction and pooling for dimensionality reduction, enabling accurate classification and detection. Residual networks (ResNets) address the challenges of training deep neural networks by introducing a residual learning framework, reformulating layers to learn residual functions relative to their inputs. This innovation enables the training of networks with significantly greater depth, improving accuracy while maintaining computational efficiency. ResNets with up to 152 layers demonstrated superior performance on benchmarks such as ImageNet, achieving a 3.57% error rate. Additionally, ResNets achieved a 28% improvement on COCO object detection, excelling in detection, localization, and segmentation tasks [7].

Chen et al. proposes an automatic breast cancer classification method for mammography using a fine-tuned ResNet with transfer learning and data augmentation. The approach improves feature extraction, reduces training time, and mitigates overfitting with limited data. Using the CBIS-DDSM dataset, the method achieved precision of 93. 15%, specificity of 92. 17%, sensitivity of 93. 83%, AUC of 0.95 and loss of

0.15, demonstrating robustness and generalization. However, the method relies on mammography images of breast cancer datasets and may face challenges in generalizing to diverse clinical settings [8].

Ferreira et al. propose a deep neural network approach using transfer learning to classify breast cancer histology images into four categories. The model, based on Inception ResNet V2, applies data augmentation and fine-tuning to improve accuracy. Tested on the ICIAR 2018 BACH-Challenge, it achieved 0.76 accuracy on the blind test set. However, this model was developed around a small size dataset proposed in a challenge. This does not indicate the model's performance on real-world data [9].

Shahidi et al. in their study evaluates deep learning models for classifying breast cancer histopathology images, focusing on binary, four, and eight-class classifications. Models such as ResNeXt, Dual Path Net, SENet, and NASNet were tested, with an emphasis on the effects of pre-processing, data augmentation, and transfer learning. Experiments on the BreakHis and BACH datasets revealed that Inception-ResNet-V2 performed best for binary and eight-class classifications [10].

From the study of literature, it is evident that although research has been conducted on predicting cancer extent in terms of benign and malignant, Resnet has proven to be one of the best performing models. Researchers have tried comparing Resnet50 with other computer vision based models and have found better results from Resnet50, However, not much analysis have been done in term of hyper-parameter tuning the Resnet50 itself. Their are multiple activation functions and parameters that, if experimented with, may produce a model that is more reliable and robust with the large datasets.

This paper aims with the analysis of Resnet50, a computer vision model by Microsoft. The aim is to experiment with various versions of Resnet50 classifier. Finally, this study aims to present a model that is accurate and robust in predicting benign vs. malignant breast cancer types from the analysis of RCL biopsy images.The results will be substantiated by evaluation metrics like accuracy, area under the ROC curves, learning curves, and more, to analyze the Resnet50 architecture and parameters.

II. Methodology

A. Data Collection and Preprocessing

The Breast Cancer Histopathological Image Classification (BreakHis) dataset comprises 9,109 microscopic images of breast tumor tissues obtained from 82 patients. The images are captured at four distinct magnification levels: 40X, 100X, 200X, and 400X. The dataset includes 2,480 benign and 5,429 malignant samples, with each image having a resolution of 700x460 pixels, stored in 3-channel RGB format with 8-bit depth per channel in PNG format. Developed in collaboration with the P&D Laboratory in Paraná, Brazil, this dataset provides a standardized resource for benchmarking and evaluating classification models in breast cancer histopathology. For the entirity of this paper, we consider the label classes benign as 0 and malignant as 1, respectively.

The BreaKHis dataset is categorized into benign and malignant tumors. Benign tumors are non-cancerous, grow slowly, and remain localized, lacking traits like cellular atypia or metastasis. Malignant tumors, synonymous with cancer, are invasive, capable of destroying nearby tissues, and can metastasize to distant sites, often resulting in fatal outcomes. Each image file represents a tissue sample, with separate folders categorizing the images based on their class. Figures 1 and 2 visualize the two classes from the dataset, benign and malignant respectively. These images also support the need of the research on image processing based diagnosis as it is very difficult to practically differentiate between the two cancer classes.

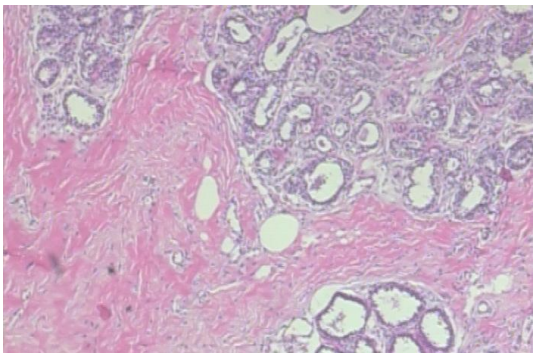

Fig. 1. Biopsy image of a benign breast cancer section.

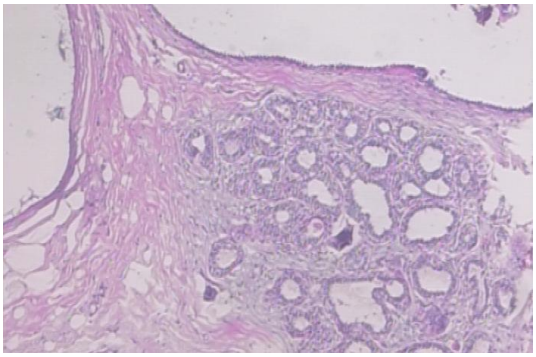

Fig. 2. Biopsy image of a malignant breast cancer section.

The preprocessing of data is a critical step in ensuring the effective training and evaluation of deep learning models, particularly when working with a complex architecture such as ResNet50. In this study, several preprocessing techniques were applied to the input images to standardize the dataset and enhance model performance. These transformations were implemented using the torchvision.transforms library, ensuring compatibility with PyTorch-based frameworks.

- Image Resizing and Normalization: The input images were first resized to a uniform dimension of 224×224 pixels to align with the input size requirement of the ResNet50 model. Resizing ensures consistency across all

TABLE I
HYPERPARAMETERS VALUES FOR CLASSIFIERS 1, 2, AND 3

Hyperparameter	Classifier 1	Classifier 2	Classifier 3
Batch Size	32	64	64
Learning Rate	0.005	0.01	0.0075
Epochs	5	3	3
Weight Decay	0.01	0.01	0.01
Momentum	0.9	0.9	0.92
Number of Workers	2	2	2
Image Size	224	224	224
Optimizer Choice	SGD	SGD	AdamW
Scheduler Choice	StepLR	StepLR	ReduceLROnPlateau
Step Size	3	3	3
Gamma	0.1	0.1	0.1
K-Folds (Cross-Validation)	N/A	3	3

images, preventing shape mismatches during batch processing. Furthermore, pixel values were normalized using the mean and standard deviation values of the ImageNet dataset ([0.485,0.456,0.406] and [0.229,0.224,0.225], respectively). This normalization standardizes the range of pixel values, therefore accelerating the convergence of the training process and mitigating the risk of vanishing or exploding gradients.

The input image I_{RGB} of size $H_{\text{raw}} \times W_{\text{raw}}$ is resized to a fixed size $H \times W$ (e.g., 224×224):

$$I_{\text{resized}}(x, y) = I_{\text{RGB}}\left(\frac{xH_{\text{raw}}}{H}, \frac{yW_{\text{raw}}}{W}\right),$$

where (x, y) are the pixel coordinates in the resized image, and bilinear interpolation is commonly used [7]. The pixel values $I_{\text{resized}}(x, y, c)$, where $c \in \{R, G, B\}$, are normalized to zero mean and unit variance based on predefined mean μ_c and standard deviation σ_c values [7]:

$$I_{\text{normalized}}(x, y, c) = \frac{I_{\text{resized}}(x, y, c) - \mu_c}{\sigma_c}.$$

- Dataset Transformation and Preparation: A custom function was defined to apply the transform operations on each image in the dataset, replacing the raw image data with its processed equivalent. This function ensures consistent preprocessing across all samples while maintaining the integrity of the dataset by retaining essential metadata like labels.
- Data Loading and Batching: After transformation, the datasets were prepared for input into the model by creating DataLoaders. The DataLoaders facilitate efficient data handling by dividing the dataset into manageable batches and shuffling the data for the training subset to enhance the model's robustness against overfitting. The batch size, set as a hyperparameter, was initialized to 32 (later increased to 64), with further customization options available to adjust for hardware limitations or specific experimental requirements. The number of workers for data loading was also specified, enabling parallel data preprocessing to optimize computational efficiency.

B. Model Configuration

The model is implemented using the Hugging Face transformers library with a pre-trained ResNet-50 model, modified for our binary classification task. In addition, K-Fold cross-validation is employed to evaluate the model's performance robustly across different subsets of the dataset, mitigating the risk of overfitting and ensuring more reliable results.

The experimental setup aimed to evaluate the performance of a ResNet-50-based image classification model under three distinct configurations, referred to as classifiers 1, 2, and 3. Variations in hyperparameters such as batch size, learning rate, optimizer, scheduler, and cross-validation were introduced to assess their impact on the model's accuracy and generalizability.

Batch size and learning rate were the most prominent factors varied across the classifiers. Classifier 1 employed a smaller batch size of 32 and a lower learning rate of 0.005, promoting gradual learning but requiring longer training time. In contrast, classifiers 2 and 3 utilized a batch size of 64, enabling smoother gradients and faster convergence, coupled with higher learning rates of 0.01 and 0.0075, respectively, to optimize performance. Classifier 3 fine-tuned the balance between faster convergence and stability by leveraging the AdamW optimizer's adaptive learning rate adjustments.

Weight decay (0.01) was consistently applied across classifiers to prevent overfitting, while momentum was increased to 0.92 in Experiment 3 to complement the AdamW optimizer. Classifiers 1 and 2 utilized SGD with a StepLR scheduler for gradual learning rate reduction, whereas Classifier 3 leveraged AdamW with a dynamic ReduceLROnPlateau scheduler, enhancing adaptability and convergence.

Cross-validation was incorporated in classifiers 2 and 3, employing three folds to evaluate robustness and generalizability. This method provided more reliable performance metrics compared to the single train-test split used in classifier 1. Other parameters, including a fixed image size of 224×224, two data-loading workers, and consistent StepLR parameters (step size of 3 and gamma of 0.1), ensured uniformity across classifiers. Table I summarizes the various hyperparameters that have been used to train each classifier in this study.

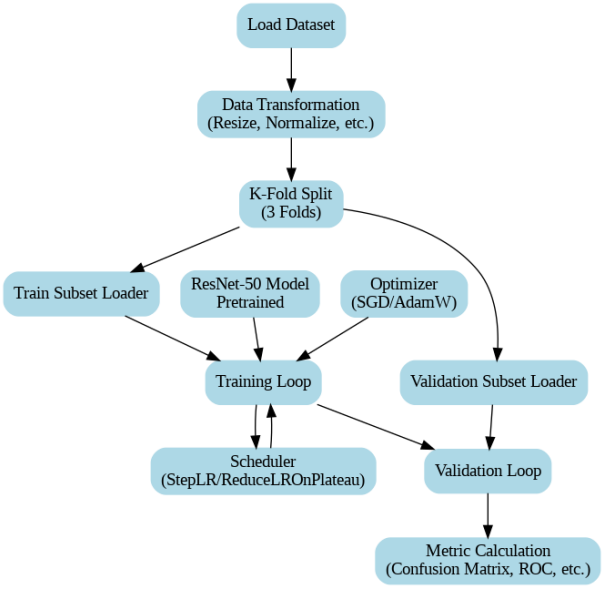

Fig. 3. Pipeline diagram of Classifier 2.

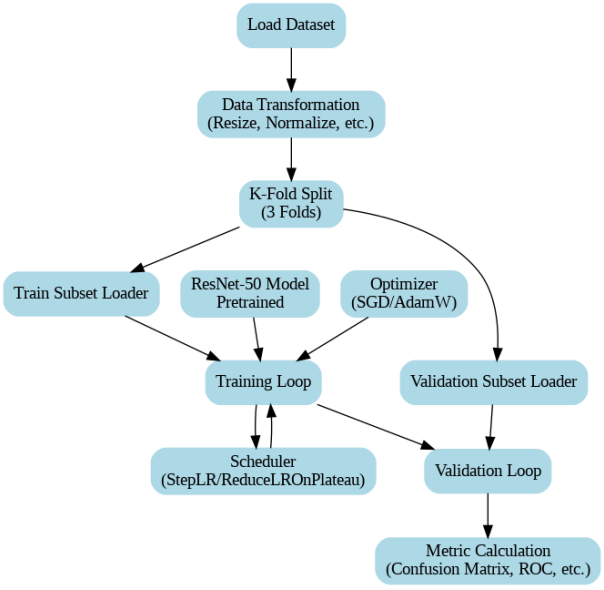

Fig. 4. Pipeline diagram of Classifier 3.

C. Training and Evaluation Process

The training loop follows a standard deep learning approach, where the model is first trained on the training data for a specified number of epochs. During each epoch, the model's parameters are updated using the gradients computed from the loss function. The loss function used is the cross-entropy loss, which is appropriate for binary classification tasks. After each training epoch, the model is evaluated on the validation data to track its performance.

For evaluation, metrics such as the confusion matrix, classification report, and Receiver Operating Characteristic (ROC) curve are computed. The confusion matrix helps to visualize the classification performance by showing the number of true positives, false positives, true negatives, and false negatives. The ROC curve, along with the area under the curve (AUC), provides insights into the model's ability to distinguish between the two classes. The classification report provides additional details on precision, recall, and F1-score for each class.

To visualize the model's performance across different folds, the confusion matrix and ROC curve are plotted for each fold. After completing all folds, aggregate metrics such as the average confusion matrix and average ROC AUC are computed, providing a final measure of model performance. These metrics help ensure that the model generalizes well across different subsets of the data.

The pipelines of the classifiers 2 and 3 are duly visualized in Figures 3 and 4 respectively.

III. Results and Discussions

A. Classifier 1

The evaluation of Classifier 1 demonstrates its efficacy in binary classification tasks, achieving robust performance metrics on the test set. Table II summarizes the detailed classification report, including precision, recall, F1-score, and support for each class. The overall accuracy of the model was 90%, showcasing its capability to generalize effectively across the dataset.

For Class 0, the precision was 0.85, and the recall was 0.97, resulting in an F1-score of 0.91. This indicates that the classifier performed exceptionally well in identifying true negatives, with only a minimal proportion of false positives. Similarly, Class 1 achieved a precision of 0.96 and a recall of 0.82, leading to an F1-score of 0.88. These metrics highlight the model's strong ability to minimize false negatives, though there remains a slight trade-off in precision for recall when predicting Class 1.

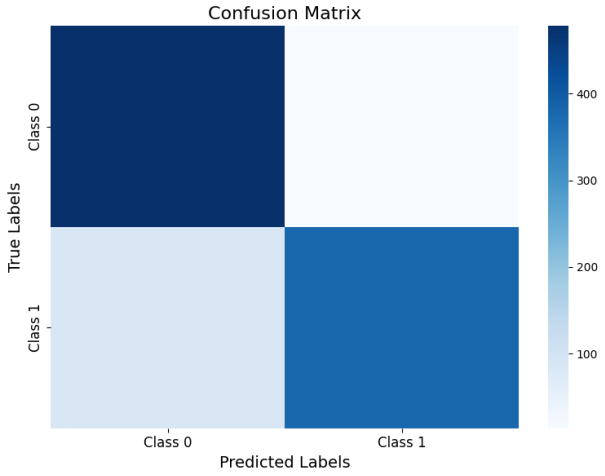

Fig. 5. Confusion Matrix for Classifier 1.

The macro-averaged precision, recall, and F1-score were 0.91, 0.89, and 0.89, respectively, reflecting balanced performance across both classes. The weighted averages, considering

class support, were consistent with the macro averages, further reinforcing the classifier's reliable performance. Figures 5 and 6 visualize the confusion matrix and receiver operating characteristic curve for classifier 1.

TABLE II
EVALUATION METRICES OF CLASSIFIER 1

Class	Precision	Recall	F1-Score	Support
0	0.85	0.97	0.91	492
1	0.96	0.82	0.88	465
Accuracy		0.90		957
Macro Avg	0.91	0.89	0.89	957
Weighted Avg	0.90	0.90	0.89	957

Classifier 1 was trained using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.005 and momentum of 0.9. Weight decay was set to 0.01 to regularize the model and prevent overfitting. The StepLR scheduler was utilized to reduce the learning rate by a factor of 0.1 after every three epochs (step size), facilitating fine-tuning as training progressed. Training was conducted for three epochs with a batch size of 32 and an image input size of 224×224 pixels. The model employed two workers for efficient data loading during training and evaluation.

The performance of Classifier 1 indicates a strong ability to discriminate between the two classes in the dataset, with a balanced trade-off between precision and recall. The chosen hyperparameter configuration, including the combination of SGD and StepLR, played a pivotal role in achieving stable and accurate predictions. The results validate the effectiveness of this setup for image classification tasks, particularly in scenarios where the dataset may exhibit class imbalance or complex feature distributions.

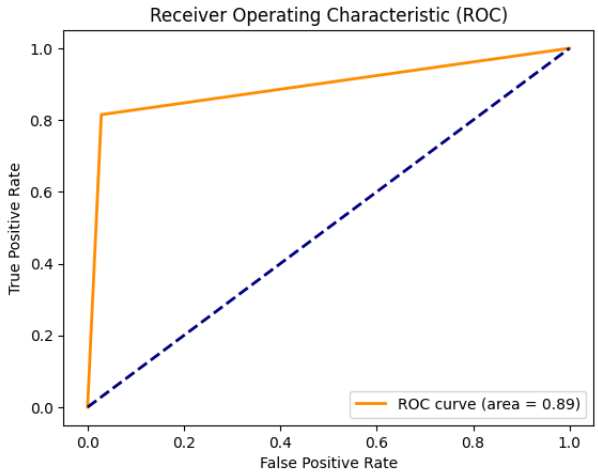

Fig. 6. Receiver operating characteristic for Classifier 1.

Future improvements may involve fine-tuning the hyperparameters further or exploring alternative optimizers such as AdamW, which could adaptively adjust learning rates for faster convergence. Additionally, increasing the number of training epochs may provide incremental performance gains, given the observed trends in accuracy and loss curves during training.

B. Classifier 2

Classifier 2 was evaluated using a 3-fold cross-validation approach to ensure robustness and generalizability of the results. The metrics for each fold, as well as the average performance across all folds, are presented in Tables III and IV. The classifier achieved a consistent level of accuracy across folds, with an average accuracy of 89%, demonstrating reliable performance across different subsets of the data.

In Fold 1, the classifier attained an accuracy of 79%. The precision, recall, and F1-score for Class 0 were 0.93, 0.64, and 0.75, respectively, while for Class 1, these metrics were 0.73, 0.95, and 0.82. The relatively lower recall for Class 0 indicates a need for further optimization to improve the model's sensitivity toward this class.

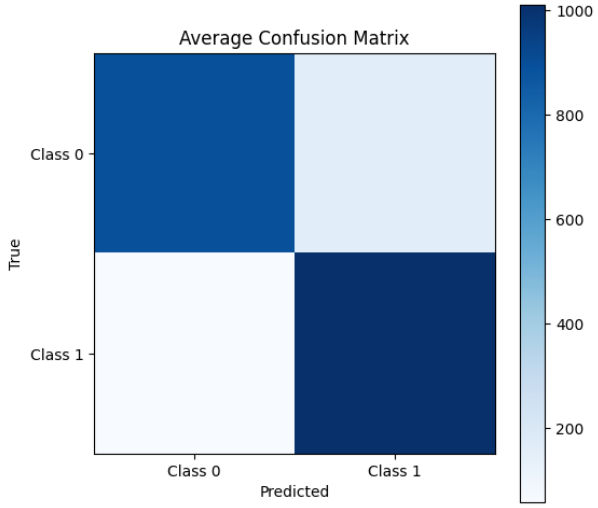

Fig. 7. Confusion Matrix for Classifier 2.

In Folds 2 and 3, the classifier's performance significantly improved, achieving accuracies of 93% in both cases. For Class 0, the F1-scores were 0.94 and 0.93, respectively, indicating effective identification of true negatives with minimal false positives. Similarly, Class 1 exhibited high F1-scores of 0.93 for both folds, demonstrating strong performance in identifying true positives with reduced false negatives.

TABLE III
EVALUATION METRICES OF K-FOLDS FOR CLASSIFIER 2; WITH K=3

Fold	Class	Precision	Recall	F1-Score	Support
Fold 1	0	0.93	0.64	0.75	1055
	1	0.73	0.95	0.82	1070
	Accuracy		0.79		2125
Fold 2	0	0.91	0.97	0.94	1053
	1	0.97	0.90	0.93	1072
	Accuracy		0.93		2125
Fold 3	0	0.90	0.97	0.93	1060
	1	0.96	0.90	0.93	1065
	Accuracy		0.93		2125

The classifier's macro-averaged precision, recall, and F1-score across the three folds were 0.90, 0.89, and 0.88, respectively. The high precision reflects the model's ability to minimize false positives, while the balanced recall values indicate its capacity to identify true positives effectively. The overall weighted averages were consistent with the macro averages, reinforcing the robustness of the classifier across different class distributions. Figures 7 and 8 visualize the confusion matrix and receiver operating characteristic curve for classifier 2.

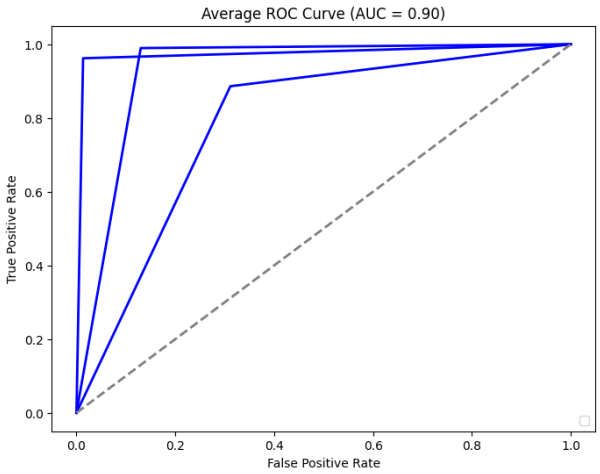

Fig. 8. Receiver operating characteristic for Classifier 2.

Classifier 2 was trained using a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01 and momentum set to 0.9. A weight decay of 0.01 was applied to regularize the model and prevent overfitting. The learning rate was adjusted dynamically using a StepLR scheduler, with the learning rate reduced by a factor of 0.1 after every three epochs. The training process was carried out over three epochs with a batch size of 64, ensuring efficient updates while maintaining computational feasibility.

TABLE IV
EVALUATION METRICES OF CLASSIFIER 2

Metric	Value
Average Precision	0.90
Average Recall	0.89
Average F1-Score	0.88
Average Accuracy	0.89

The results obtained for Classifier 2 demonstrate its reliability and adaptability in binary classification tasks across varied data subsets. The hyperparameter configuration, particularly the use of the SGD optimizer with StepLR scheduling, contributed to consistent improvements in accuracy and F1-scores during training. The model's strong performance across all three folds underscores its ability to generalize effectively, even when exposed to diverse data distributions.

The higher recall for Class 1 compared to Class 0 in Fold 1 suggests a potential class imbalance or feature complexity that may require further exploration. This could be addressed by implementing techniques such as data augmentation, oversampling, or alternative loss functions that penalize misclassifications unevenly.

C. Classifier 3

Classifier 3 was evaluated using 3-fold cross-validation to ensure reliable and generalizable performance metrics. The evaluation results across folds, as well as the average metrics, are summarized in Tables V and VI. The classifier consistently demonstrated strong performance, achieving an average accuracy of 90% across the folds.

In Fold 1, Classifier 3 achieved an accuracy of 79%, with Class 0 obtaining a precision of 0.86, recall of 0.69, and F1-score of 0.76. Class 1 achieved a precision of 0.74, recall of 0.89, and F1-score of 0.81. These results suggest a moderate level of performance for this fold, likely influenced by class-specific variations or the inherent data distribution.

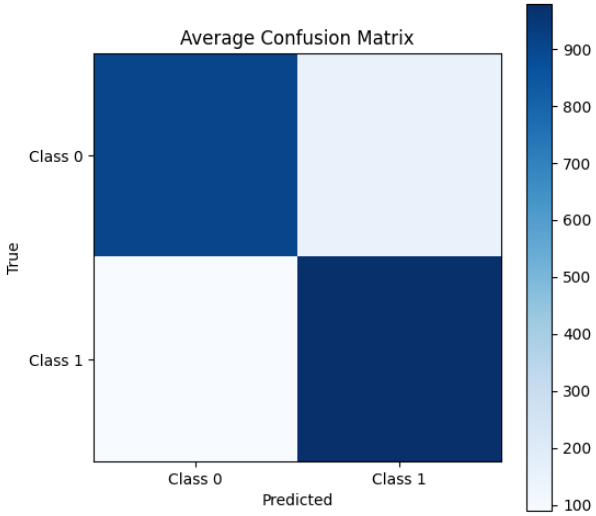

Fig. 9. Confusion Matrix for Classifier 3.

The classifier's performance improved significantly in Fold 2, achieving an accuracy of 93%. Class 0 attained an F1-score of 0.93, reflecting improved sensitivity and precision in identifying true negatives. Similarly, Class 1 achieved an F1-score of 0.93, indicating balanced performance across both classes. Figures 9 and 10 visualize the confusion matrix and receiver operating characteristic curve for classifier 3.

Fold 3 demonstrated the highest performance, with an accuracy of 97%. Both classes exhibited identical F1-scores of 0.97, highlighting the classifier's exceptional ability to accurately predict both positive and negative classes in this subset of data.

Across all three folds, Classifier 3 achieved average precision, recall, and F1-scores of 0.90, demonstrating its balanced performance across classes. The consistent macro and weighted averages indicate that the classifier performed robustly regardless of the class distribution in the dataset.

TABLE V
EVALUATION METRICES OF K-FOLDS FOR CLASSIFIER 3; WITH K=3

Fold	Class	Precision	Recall	F1-Score	Support
Fold 1	0	0.86	0.69	0.76	1063
	1	0.74	0.89	0.81	1062
	Accuracy		0.79		2125
Fold 2	0	0.99	0.87	0.93	1061
	1	0.88	0.99	0.93	1064
	Accuracy		0.93		2125
Fold 3	0	0.96	0.99	0.97	1044
	1	0.99	0.96	0.97	1081
	Accuracy		0.97		2125

Classifier 3 was optimized using the AdamW optimizer, which effectively balances convergence speed and stability through weight decay and adaptive learning rates. The learning rate was set to 0.0075, with a weight decay of 0.01 to minimize overfitting. Momentum was configured at 0.92, and training was conducted over three epochs using a batch size of 64 to ensure computational efficiency.

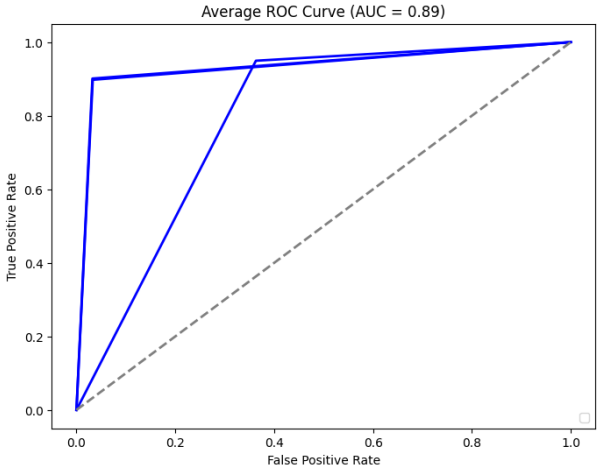

Fig. 10. Receiver operating characteristic for Classifier 3.

A ReduceLROnPlateau learning rate scheduler was employed to dynamically adjust the learning rate based on validation loss, enhancing the model's ability to converge. The lower recall for Class 0 in Fold 1 suggests that further optimization or rebalancing techniques, such as oversampling or synthetic data augmentation, could improve sensitivity for this class. Despite this, the classifier achieved high precision, indicating its ability to minimize false positives effectively.

TABLE VI
EVALUATION METRICES OF CLASSIFIER 3

Metric	Value
Average Precision	0.90
Average Recall	0.90
Average F1-Score	0.90
Average Accuracy	0.90

D. Comparative Study of Classifiers

Each classifier was trained with distinct configurations and assessed using cross-validation to ensure robustness and generalizability. The comparative analysis focuses on the performance metrics, learning configurations, and trade-offs between precision, recall, and accuracy.

1) Performance Across Folds: Classifier 1 demonstrated a balanced performance with an overall accuracy of 90%. While it achieved high precision for Class 1 (0.96), its recall for this class was relatively lower (0.82), indicating a tendency to underpredict positive cases. Conversely, Classifier 2 outperformed Classifier 1 in terms of cross-validation accuracy, achieving an average accuracy of 89%. Classifier 2 showed stable performance across folds, with its best accuracy recorded in Fold 3 (93%). However, its recall for Class 0 in Fold 1 (0.64) suggests sensitivity issues for negative cases in certain scenarios. Classifier 3 emerged as the top-performing model, achieving the highest average accuracy of 90% across folds, with Fold 3 recording a peak accuracy of 97%. Its performance was consistent across both classes, achieving precision, recall, and F1-scores of 0.90 on average.

2) Class-Level Observations: For Class 0, Classifier 3 demonstrated superior precision and recall across most folds, particularly excelling in Fold 3 with a precision of 0.96 and recall of 0.99. Classifier 2 exhibited competitive precision for this class but struggled in recall for Fold 1. Classifier 1, while achieving good precision, had comparatively lower recall, indicating a trade-off where it was less effective at identifying all true negatives. For Class 1, Classifier 3 again showed balanced metrics, achieving the highest F1-score (0.97) in Fold 3. Classifier 2 demonstrated a slight edge in precision for this class during Fold 2, but its performance varied across folds. Classifier 1 displayed a strong F1-score for Class 1 but with a lower recall, suggesting room for improvement in identifying true positives.

3) Impact of Hyperparameters: Classifier 1 and Classifier 2 were both trained using the SGD optimizer, with Classifier 1 employing a learning rate of 0.005 and Classifier 2 using a higher learning rate of 0.01. This difference in learning rate, combined with the batch size of 32 for Classifier 1 and 64 for Classifier 2, influenced their convergence behaviors and abilities. In contrast, Classifier 3 utilized the AdamW optimizer and a learning rate of 0.0075, coupled with a ReduceLROnPlateau scheduler. This configuration enabled it to adapt effectively to the data distribution, minimizing overfitting and achieving consistently high performance across folds.

4) Discussion and Trade-Offs: Classifier 1 exhibited a balanced trade-off between precision and recall but fell short of the overall accuracy achieved by the other models. Its smaller batch size and lower learning rate contributed to more conservative updates, which may have limited its performance. Classifier 2, while achieving high accuracy in some folds, demonstrated variability in class-level recall, suggesting it may require additional fine-tuning to address sensitivity issues. Classifier 3, with its advanced optimizer and adaptive learning rate scheduler, achieved the most consistent and robust

performance. Its high recall and precision across classes make it particularly suitable for tasks where misclassification costs are high.

Classifier 3 emerged as the most reliable model, achieving superior accuracy and balanced class-level performance. While Classifier 1 and Classifier 2 excelled in precision and accuracy, respectively, they exhibited inconsistent recall. Future research could explore extended training, ensemble techniques, and advanced data augmentation to enhance performance further.

A significant challenge in breast cancer diagnostics is inter-observer variability among pathologists when interpreting complex histopathological images. The ResNet50-based classifiers provided consistent and reproducible results across multiple folds, reducing subjective bias and serving as a complementary tool for pathologists. Automated systems like these can prioritize high-risk cases, alleviate the workload of medical professionals, and ensure timely diagnosis, particularly in resource-limited settings.

The balanced performance for benign and malignant cases supports effective risk stratification in clinical workflows. High-risk patients can be directed to intensive diagnostic pathways, while low-risk patients avoid unnecessary procedures, aligning with personalized medicine approaches to optimize patient care.

IV. Conclusion

This study performs a comprehensive analysis of the ResNet50 architecture, a state-of-the-art computer vision model, to classify benign and malignant breast cancer types using RCL biopsy images. By experimenting with three versions of ResNet50, each tuned with distinct hyperparameters and optimization strategies, the research aimed to identify the most accurate and robust model for this critical medical imaging task. Evaluation metrics, including accuracy, precision, recall, and F1-scores, were utilized alongside cross-validation techniques to ensure reliable and generalizable results.

Classifier 1 provided a balanced trade-off between precision and recall but lagged in overall accuracy and recall consistency compared to the other models. Classifier 2 showed strong accuracy in specific folds but demonstrated variability in class-level recall, highlighting the need for further tuning to improve sensitivity to true negative cases. Classifier 3 emerged as the top-performing model, achieving the highest average accuracy (90%) across folds and demonstrating balanced performance across all evaluation metrics. Its use of the AdamW optimizer and an adaptive learning rate scheduler was critical in ensuring stability and consistency across the dataset.

This study highlights the importance of hyperparameter optimization and cross-validation in developing reliable deep learning models for medical imaging. Classifier 3, with its balanced and precise performance, shows strong potential for clinical use in breast cancer diagnostics, where accurate predictions are critical.

Future work will focus on extending training, testing generalization with larger datasets, and leveraging ensemble methods to enhance performance. Additionally, integrating explainable AI could provide valuable insights into the model's decisions, fostering trust and adoption in clinical settings. This research advances AI in healthcare, aiming to improve diagnostic accuracy and patient outcomes.

Code and GitHub Repository

The code and documentation for the methods discussed in this paper can be accessed through the following GitHub repository:

Naman Dhariwal GitHub Repository

The 3 classifiers discussed in this paper can be accessed through the following Google Drive link:

Naman Dhariwal Google Drive

References

[1] Siegel, Rebecca L., Angela N. Giaquinto, and Ahmedin Jemal. "Cancer statistics, 2024." CA: a cancer journal for clinicians 74.1 (2024): 12-49.

[2] Brown, Joel S., et al. "Updating the definition of cancer." Molecular Cancer Research 21.11 (2023): 1142-1147.

[3] Wang, Jun, and San-Gang Wu. "Breast cancer: an overview of current therapeutic strategies, challenge, and perspectives." Breast Cancer: Targets and Therapy (2023): 721-730.

[4] Nadeem, R., et al. "Occult breast lesions: a comparison between radioguided occult lesion localisation (ROLL) vs. wire-guided lumpectomy (WGL)." The Breast 14.4 (2005): 283-289.

[5] Zhang, Bo, Huiping Shi, and Hongtao Wang. "Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach." Journal of multidisciplinary healthcare (2023): 1779-1791.

[6] Sebastian, Anu Maria, and David Peter. "Artificial intelligence in cancer research: trends, challenges and future directions." Life 12.12 (2022): 1991.

[7] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[8] Chen, Yuanqin, et al. "Fine-tuning ResNet for breast cancer classification from mammography." Proceedings of the 2nd International Conference on Healthcare Science and Engineering 2nd. Springer Singapore, 2019.

[9] Ferreira, Carlos A., et al. "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2." International conference image analysis and recognition. Cham: Springer International Publishing, 2018.

[10] Shahidi, Faezehsadat, et al. "Breast cancer classification using deep learning approaches and histopathology image: A comparison study." Ieee Access 8 (2020): 187531-187552.

[11] Spanhol, F. A., Oliveira, L. S., Petitjean, C., Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification. IEEE Transactions on Biomedical Engineering (TBME), 63(7):1455-1462.