Project 2

FYS-STK4155

Krithika Gunasegaran, Oskar Idland, Erik Røset & Arangan Subramaniam
University of Oslo, Department of Physics
(Dated: November 23, 2024)

Add abstract here

https://github.com/Oskar-Idland/FYS-STK4155-Projects

I. INTRODUCTION

II. THEORY

III. METHODS & IMPLEMENTATION

A. Data Preprocessing and Model Architecture

The dataset consisted of breast cancer histopathological images classified into three categories: normal, malignant and benign. Each malignant and benign image included an associated mask file denoting regions of interest, which were excluded from this analysis. All images were preprocessed to a standardized size of 224 × 224 pixels to maintain consistency with standard deep learning architectures while preserving sufficient detail for medical diagnosis.

TThe dataset was partitioned into training (563 images), validation (100 images), and test (117 images) sets, representing approximately 72%, 12% and 15% of the data respectively. Stratification was maintained across splits to preserve class distribution, particularly important given the significant class imbalance (benign: 437, malignant: 210, normal: 133). Data augmentation techniques were employed to address this imbalance, specifically targeting the minority classes (malignant and normal). The augmentation pipeline included random horizontal flips with a high probability (p=0.9) to maximize the diversity of cell orientations, constrained rotation (±15 degrees) to maintain anatomical plausibility, and color jittering to enhance robustness against staining variations common in histopathological samples.

Three distinct deep learning architectures were implemented and compared:

- A custom simple CNN was designed with specific consideration for the small dataset size. The architecture comprised two convolutional layers ($3 \rightarrow 16 \rightarrow 32$ channels) with 3×3 kernels, chosen to capture local cellular features while maintaining computational efficiency. ReLU activation was selected for its non-linearity and reduced likelihood of vanishing gradients. Max pooling operations follow each convolutional layer to achieve spatial dimensionality reduction while preserving important features. The resulting 56×56×32 feature maps are flattened into a 100,352-dimensional vector, followed by fully connected layers reducing to 512 and 128 neurons respectively. This progressive reduction in dimensionality ($100,352 \rightarrow 512 \rightarrow 128 \rightarrow 3$) creates an information bottleneck that forces the network to learn increasingly abstract representations of the input data, with the final layer outputting probabilities for our three classes.

- A MobileNet architecture[1], selected for its efficient depthwise seperable convolutions which significantly reduce computational complexity while maintaining performance. This architecture has demonstrated success in medical image classification tasks where computational resources may be limited in clinical settings.[2]

- A pre-trained ResNet101 model[3], chosen for its deep residual learning framework that effectively addresses the degradation problem in deep networks and general good performance in medical classification tasks[4]. The model was pre-trained on ImageNet, allowing it to leverage general feature detection capabilities, while its final layer was modified to accommodate our three-class classification task.

B. Training and Evaluation

The models were trained using the Adam optimizer, selected for its adaptive learning rate capabilities and robust performance across different neural architectures. Learning rates were empirically determined: 0.001 for both Simple CNN and MobileNet, and a lower rate of 0.00005 for ResNet101 to prevent catastrophic forgetting of pre-trained features. A step learning rate scheduler (step size=7, γ=0.1) was implemented to facilitate convergence to better optima. Training utilized cross-entropy loss, appropriate for our multi-class classification task, and incorporated early stopping (patience=2, minimum epochs=5) to prevent overfitting while ensuring sufficient learning time.

Models were evaluated using a comprehensive set of metrics: accuracy, precision, recall, and F1-score, with particular attention to per-class performance given the imbalanced nature of our dataset. Confusion matrices

were generated to provide detailed insight into class-wise performance and misclassification patterns.

C. Implementation Details

The implementation used the PyTorch framework and was executed on GPU hardware. Training was conducted with a batch size of 8, chosen to balance between computational efficiency and stable gradient updates. Data normalization used ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) to ensure compatibility with pre-trained models and standardize feature scales. All experiments were conducted with a fixed random seed for reproducibility, with final evaluation performed on the held-out test set to ensure unbiased performance assessment.

IV. RESULTS & DISCUSSION

We present a comprehensive analysis of three neural network architectures' performance on the breast cancer histopathological dataset, examining their effectiveness in handling a relatively small medical imaging dataset (780 images). Our analysis spans model accuracy, class-wise performance, and training dynamics to understand how architectural complexity and transfer learning influence classification performance under data constraints.

A. Model Performance Overview

We present our analysis through confusion matrices (fig. 1) and classification metrics (table I), which reveals distinct patterns in how each architecture handles the classification task. ResNet101 demonstrates superior and balanced performance across all classes, while both the Simple CNN and MobileNet show characteristic behaviors reflecting their architectural differences and the challenges of limited training data.

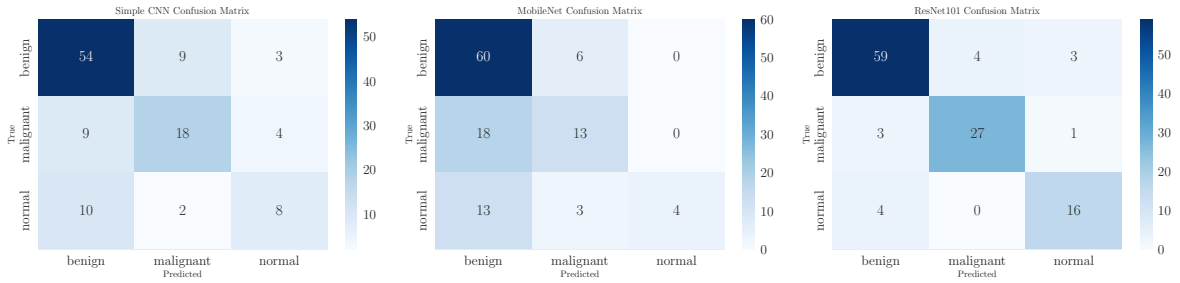

FIG. 1: Confusion matrices for the three neural network architectures (Simple CNN, MobileNet, and ResNet101) evaluated on the breast cancer histopathological test dataset. Values and color intensity indicate the number of images classified in each category.

The confusion matrices demonstrate distinct classification behaviors across architectures. ResNet101 exhibits balanced prediction distribution across classes, while both the simple CNN and MobileNet display systematic bias toward the majority class (benign), as evidenced by the disproportionate predictions in their respective primary columns.

ResNet101 shows markedly superior behavior compared to both other models, achieving not just higher but more consistent performance across all classes. The confusion matrix demonstrates balanced prediction patterns: correctly identifying 59/66 benign, 27/31 malignant, and 16/20 normal cases. This balanced performance suggests that the pre-trained features, developed on a large diverse dataset, provide a robust foundation that can be effectively fine-tuned even with limited medical imaging data.

The Simple CNN, despite its basic architecture, demonstrates more balanced performance across classes than MobileNet, though with lower overall accuracy. Its confusion matrix reveals consistent prediction patterns, correctly identifying 54 of 66 benign cases (0.82 recall) while maintaining modest but balanced performance on minority classes (malignant: 18/31, normal: 8/20 correct identifications). This suggests that simpler architectures might be more robust to class imbalance when training data is limited, possibly due to fewer parameters needing optimization and thus less opportunity for overfitting to the majority class.

	Simple CNN	MobileNet	ResNet101
Precision			
Benign	0.74	0.66	0.89
Malignant	0.62	0.59	0.87
Normal	0.53	1.00	0.80
Weighted Avg.	0.67	0.70	0.87
Recall			
Benign	0.82	0.91	0.89
Malignant	0.58	0.42	0.87
Normal	0.40	0.20	0.80
Weighted Avg.	0.68	0.66	0.87
F1-score			
Benign	0.78	0.76	0.89
Malignant	0.60	0.49	0.87
Normal	0.46	0.33	0.80
Weighted Avg.	0.68	0.62	0.87

TABLE I: Classification metrics grouped by metric type, comparing performance across models. The weighted averages account for class imbalance in the dataset.

Of particular note is MobileNet's handling of the normal class, where it shows perfect precision (1.00) but poor recall (0.20). This seemingly contradictory result stems from the model's extreme conservatism in predicting the normal class: out of 117 test cases, it only predicted "nor-

mal" 4 times. While all of these predictions were correct (hence the perfect precision), the model failed to identify 16 out of 20 normal cases, instead classifying them primarily as benign (13 cases) or malignant (3 cases). This behavior suggests that despite our data augmentation efforts, MobileNet struggled with the class imbalance in the training data, adopting an overly conservative strategy for the minority class.

B. Architectural Performance Analysis

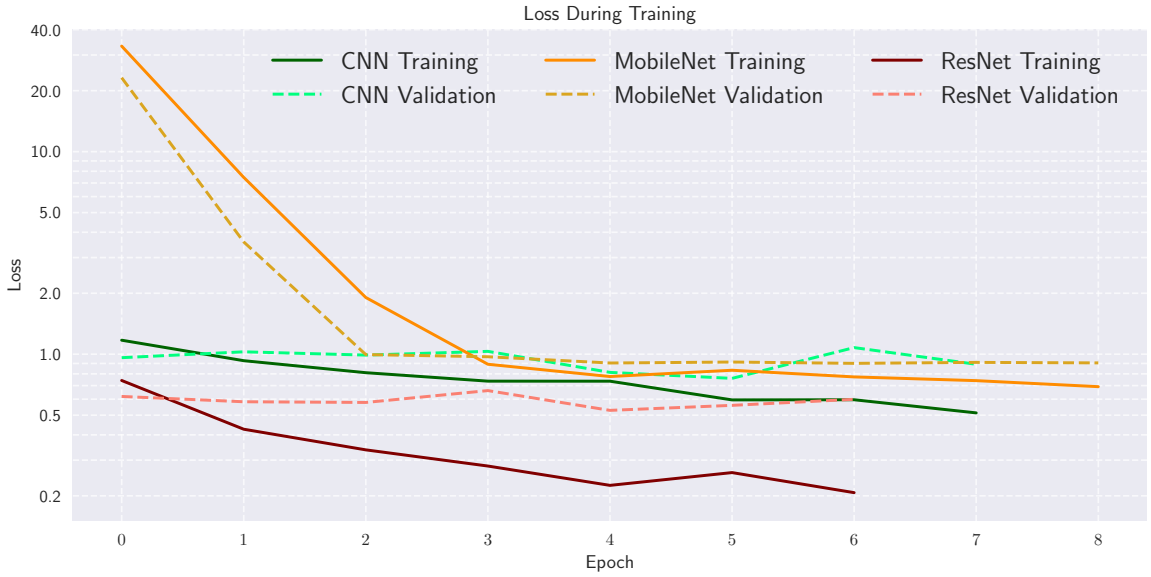

FIG. 2: Loss values during model training plotted on a logarithmic scale. Solid lines represent training loss and dashed lines represent validation loss for each architecture over training epochs.

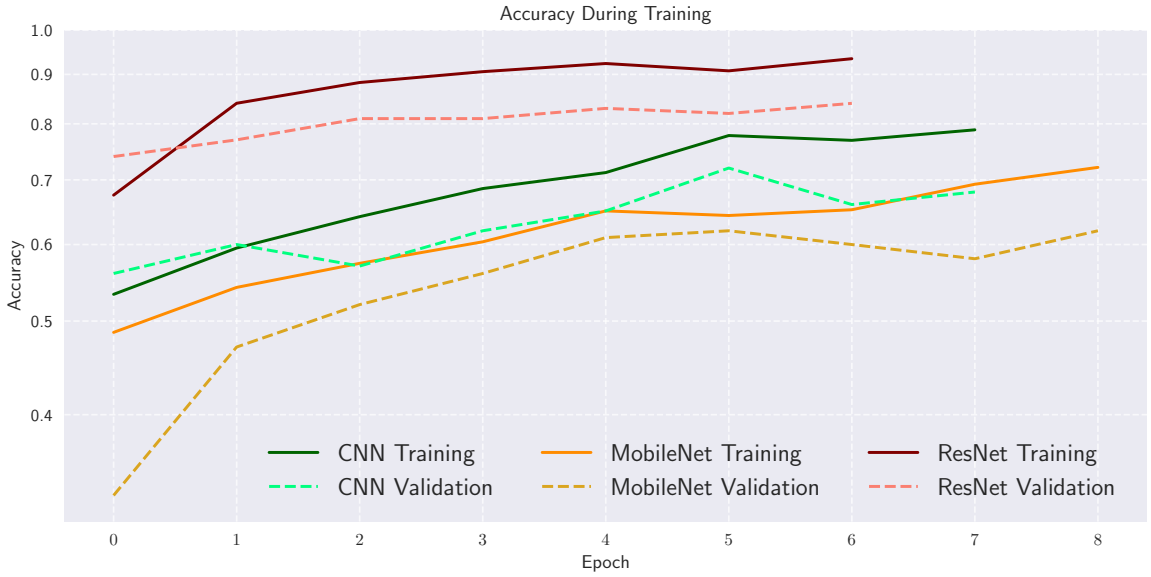

FIG. 3: Model accuracy during training. Solid lines show training accuracy and dashed lines show validation accuracy for each architecture over training epochs.

The training dynamics of each model architecture (fig. 2 and fig. 3) provide additional insight into model

convergence and learning efficiency. ResNet101 initiated training with lower loss values (0.74) compared to the Simple CNN (1.17) and MobileNet (33.19), attributable to its pre-trained weights. The convergence trajectories also differ significantly: ResNet101 exhibited steady descent to a final loss of 0.23, while MobileNet required several epochs to stabilize from its initially high loss values. The Simple CNN maintained intermediate loss values throughout training, converging to 0.59.

Accuracy measurements correlate with these observations. ResNet101 achieved initial validation accuracy of 0.74, improving to 0.83 through training. In contrast, the Simple CNN and MobileNet started at 0.56 and 0.33 respectively, with final validation accuracies of 0.72 and 0.60. Early stopping activated at 7-8 epochs across all architectures, and the proximity between validation and training curves suggests minimal overfitting in all cases. This rapid convergence indicates that performance limitations stem from architectural and data constraints rather than insufficient training duration.

V. CONCLUSION

These experimental results yield several crucial insights for developing practical medical image classification systems:

- **Transfer Learning Effectiveness:** Pre-trained models can significantly outperform custom architectures on small datasets, with ResNet101 demonstrating nearly 20% higher F1-scores (0.87) than models trained from scratch (CNN: 0.68, MobileNet: 0.62).
- **Architectural Complexity Trade-offs:** The superior performance of the Simple CNN compared to MobileNet contradicts the general assumption that more complex architectures yield better results. This finding suggests that architectural complexity must be carefully balanced against dataset size, particularly in medical imaging applications where large, annotated datasets are often unavailable.
- **Class Imbalance Challenges:** The degraded performance of both Simple CNN and MobileNet on minority classes, despite data augmentation efforts, highlights persistent challenges in handling class imbalance in medical datasets. ResNet101's more balanced performance suggests that transfer learning can help mitigate these issues.
- **Resource Considerations:** While transfer learning from large-scale pre-trained models offers superior performance, the requirements for computational resources and initial training data may limit accessibility. The adequate performance of simpler architectures suggests that targeted, dataset-appropriate model design might offer a more practical solution in resource-constrained settings.

These findings demonstrate that successful medical image classification with limited data requires careful consideration of model complexity, transfer learning opportunities, and computational constraints rather than defaulting to more complex architectures. Future work should investigate methods to improve minority class performance and explore architectural modifications that better utilize limited training data.

REFERENCES

[1] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, CoRR **abs/1704.04861** (2017), 1704.04861.

[2] P. Sharma, D. R. Nayak, B. K. Balabantaray, M. Tanveer, and R. Nayak, Neural Networks **169**, 637 (2024).

[3] The PyTorch Foundation, "ResNet101," https://pytorch.org/vision/main/models/generated/torchvision.models.resnet101.html (Accessed: November 2024).

[4] Y. Wang, Z. Feng, L. Song, X. Liu, and S. Liu, Computational and Mathematical Methods in Medicine **2021**, 2485934 (2021), https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/2485934.

Appendix A: Code

Link to our GitHub repository: https://github.com/Oskar-Idland/FYS-STK4155-Projects