# Deep Learning for Medical Image Classification: Evaluating ResNet18, Vision Transformer, and DenseNet in Breast Cancer, Pneumonia, and Chest Pathology

**Bandi Jagadeesh (z2016628)**

**Mallipudi Gayathri (z1981153)**

## 1. Abstract

Medical image classification using deep learning has shown remarkable progress in recent years, yet the challenge of class imbalance remains a significant hurdle in practical applications. This project addresses the challenge of imbalanced medical image classification by developing and evaluating deep learning models optimized for Area Under the ROC Curve (AUC) performance. We implement and compare two primary architectures: ResNet18 and Vision Transformer (ViT), evaluating their performance across three distinct medical imaging tasks from the MedMNIST dataset - breast cancer detection, pneumonia identification, and chest pathology classification. Our approach incorporates various techniques to address class imbalance, including specialized data augmentation strategies, careful consideration of loss functions, and transfer learning from pretrained models. We explore the effectiveness of different optimization strategies and regularization techniques to enhance model generalization while maintaining high sensitivity to minority classes. Experimental results demonstrate the comparative advantages of different architectural choices and training strategies across the three medical domains, providing insights into building robust medical image classification systems for imbalanced datasets.

## 2. Introduction

The rapid advancement of deep learning technologies has revolutionized medical image analysis, offering new possibilities for automated diagnosis and screening. Medical image classification presents unique challenges, particularly in handling class imbalance and achieving reliable performance across different medical domains. This study investigates the effectiveness of three prominent deep learning architectures - ResNet18, Vision Transformer (ViT), and DenseNet - in

classifying medical images across three distinct domains: breast cancer detection, pneumonia identification, and chest pathology classification.

The significance of this research lies in its comprehensive evaluation of modern deep learning architectures applied to medical imaging tasks. While previous studies have often focused on single domains or architectures, our work provides a comparative analysis across multiple medical domains and model architectures, offering insights into their relative strengths and limitations.

## 3. Method

### a. Dataset Selection and Preprocessing

We utilized three datasets from the MedMNIST collection that includes Breast Cancer Dataset suitable for binary classification task for breast cancer detection, Pneumonia Dataset for binary classification for pneumonia identification, and Chest Dataset supporting Multi-label classification with 14 different chest pathologies.

Data preprocessing included:

- Resizing images to 224x224 pixels to maintain consistency across models
- Normalization using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
- Channel expansion from grayscale to RGB through replication
- Random sampling to manage computational resources while maintaining statistical significance

### b. Model Architectures

Our study implemented and evaluated three distinct deep learning architectures, each chosen for their proven capabilities in image classification tasks. The ResNet18 architecture, our first model, was pretrained on the ImageNet dataset to leverage transfer learning benefits. We modified its final fully connected layer to accommodate our specific classification tasks while maintaining the original architectural features, particularly the crucial residual connections that address the vanishing gradient problem in deep networks.

The Vision Transformer (ViT) represented our second architectural approach, implementing a base configuration with 16x16 patch size for image processing. Like ResNet18, we utilized pretrained weights from ImageNet as our starting point. The model's head was carefully adapted for our specific classification tasks, and we leveraged its attention mechanism for capturing global dependencies within medical images. This architecture's unique ability to process relationships between distant image regions made it particularly interesting for medical image analysis.

Our third architecture, DenseNet121, implemented a dense connectivity pattern that ensures maximum information flow between layers. We configured it with a growth rate of 32 and incorporated transition layers for effective dimensionality reduction. The classifier was modified to align with our specific tasks while maintaining the network's characteristic feature reuse capabilities.

**c. Training Strategy**

The implementation of our training strategy was carefully designed to optimize model performance while maintaining consistency across experiments. We employed the Adam optimizer with a learning rate of 0.001, chosen for its adaptive learning rate capabilities and robust performance in deep learning tasks. The loss function selection was task-specific: Binary Cross-Entropy with Logits was implemented for the chest dataset, while Cross-Entropy Loss was utilized for breast and pneumonia datasets.

Batch sizes were optimized according to task complexity, with 16 samples per batch for binary classification tasks and 32 for multi-label classification. Training proceeded for 5 epochs, with an early stopping mechanism based on validation performance to prevent overfitting. This approach ensured efficient training while maintaining model generalization capabilities.

**d. Evaluation Metrics**

Our primary evaluation metric was the Area Under the ROC Curve (AUC), chosen for its robustness in handling imbalanced datasets and its independence from decision thresholds. For binary classification tasks, we implemented the standard ROC AUC calculation, providing a comprehensive measure of model discrimination ability. Multi-label classification required a more sophisticated approach, utilizing macro-averaged ROC AUC to account for multiple classes.

The evaluation framework included careful handling of edge cases and missing classes, ensuring robust performance measurement across all scenarios. This comprehensive evaluation approach allowed us to make meaningful comparisons between different architectures and across various medical imaging tasks.

## 4. Experiments

### a. Experimental Setup

We conducted three sets of experiments corresponding to each dataset:

1. Breast Cancer Classification:

    - Sample size of 100 training samples was used
    - Balanced validation and test sets
    - All three architectures evaluated under identical conditions

2. Pneumonia Detection:

    - Sample size of 100 training samples was used
    - Focus on binary classification performance
    - Evaluation of model generalization

3. Chest Pathology Classification:

    - Larger sample size (1000 samples) was used
    - Multi-label classification scenario
    - Evaluation of model performance across 14 classes

### b. Results and Analysis

#### 1. Binary Classification Tasks:

Breast Cancer Detection:

- DenseNet achieved highest performance (AUC: 0.8686)

- ResNet18 showed moderate performance (AUC: 0.4883)
- ViT demonstrated intermediate results (AUC: 0.5439)

Pneumonia Detection:

- ResNet18 achieved superior performance (AUC: 0.9618)
- DenseNet showed strong results (AUC: 0.9409)
- ViT performed notably lower (AUC: 0.5401)

1. **Multi-label Classification Task:**

   Chest Pathology:

   - DenseNet showed best performance (AUC: 0.6360)
   - ResNet18 achieved moderate results (AUC: 0.5681)
   - ViT performed similarly to ResNet18 (AUC: 0.5509)

2. **Comparative Analysis:**

   Architecture Performance:

   - DenseNet showed consistent performance across tasks
   - ResNet18 excelled in pneumonia detection
   - ViT demonstrated lower performance, possibly due to limited training data

   Task-Specific Observations:

   - Binary classification tasks generally achieved higher AUC scores
   - Multi-label classification proved more challenging
   - Model performance varied significantly across medical domains

   Computational Considerations:

   - Training time and resource requirements varied across architectures

- Trade-offs between model complexity and performance observed
- Sample size impact on model performance noted

## 5. Discussion

The experimental results from our comprehensive study of deep learning architectures in medical image classification reveal several crucial insights that warrant detailed discussion. Our analysis encompasses both architecture-specific performance metrics and domain-specific considerations, providing a thorough understanding of the strengths and limitations of each approach.

### a. Model Performance Analysis

In examining architecture-specific performance, ResNet18 demonstrated exceptional capability in pneumonia detection, achieving an AUC of 0.9618. This outstanding performance suggests that ResNet18's architecture is particularly well-suited for identifying clear structural abnormalities in medical images. The success can be attributed to its residual learning framework, which enables effective feature extraction for distinct pathological patterns.

DenseNet exhibited remarkable consistency across all tasks, indicating robust performance in handling complex feature hierarchies. This consistency stems from its dense connectivity pattern, which facilitates feature reuse and promotes stronger gradient flow during training. The architecture's ability to maintain high performance across varied medical domains suggests its potential as a versatile solution for medical image classification tasks.

Vision Transformer (ViT), while innovative in its approach, showed relatively lower performance compared to its CNN counterparts. This limitation can be attributed to several factors: the restricted size of our training datasets, the absence of inductive biases inherent in CNNs, and the model's inherent complexity requiring larger datasets for optimal performance. These findings align with existing literature suggesting that transformer architectures may require substantial data resources to achieve their full potential.

### b. Domain-Specific Considerations

In breast cancer detection, we observed moderate performance variations across models, with AUC ranges from 0.4883 to 0.8686. This variance reflects the inherent complexity of breast cancer

classification, where subtle tissue variations and complex patterns must be identified. DenseNet's superior performance in this domain suggests enhanced capability in capturing fine-grained tissue characteristics, while ResNet18's lower performance indicates potential limitations in discriminating subtle pathological features.

Pneumonia detection presented a notably different scenario, with consistently high performance across models (AUC: 0.9409-0.9618). The clear structural changes characteristic of pneumonia cases likely contributed to this superior performance. ResNet18's architecture proved particularly effective, suggesting that its design is well-suited for identifying distinct pathological patterns in chest radiographs.

The chest pathology classification task, being multi-label in nature, presented unique challenges. Overall performance was lower compared to binary tasks, reflecting the increased complexity of simultaneous multiple pathology detection. DenseNet's architecture demonstrated superior capability in handling these multiple pathology patterns, likely due to its efficient feature reuse mechanism.

**c. Limitations and Challenges**

Our study encountered several notable limitations that warrant discussion.

- Reduced sample sizes may not fully reflect real-world performance characteristics: The challenge of class imbalance significantly affected model training dynamics, potentially biasing model performance toward majority classes.
- Memory limitations often necessitated smaller batch sizes than optimal, potentially affecting training stability: Computational resource constraints influenced model selection and training strategies, sometimes requiring compromises between model sophistication and practical feasibility. Training time considerations for larger models, particularly ViT, highlighted the need for efficient training strategies in practical applications.
- The gap between research performance metrics and clinical requirements remains substantial: The need for model interpretability - crucial for clinical adoption - often conflicts with the black-box nature of deep learning models. Real-world deployment

considerations, including integration with existing clinical workflows and regulatory compliance, pose additional challenges that extend beyond pure technical performance.

## 6. Future Directions

Based on our findings, several promising directions for future research emerge. In terms of model improvements, the investigation of hybrid architectures combining CNN and transformer elements could potentially leverage the strengths of both approaches. Advanced data augmentation techniques specific to medical imaging could help address the limited data availability. Domain-specific architectural modifications, informed by medical domain knowledge, may improve model performance for specific tasks.

Clinical integration represents another crucial area for future work. The integration of clinical metadata with image features could provide additional context for more accurate diagnoses. Development of explainable AI components would enhance trust and adoption in clinical settings. Validation on larger, more diverse datasets remains essential for establishing robust clinical utility.

Technical advancements should focus on implementing more efficient training strategies to reduce computational requirements while maintaining performance. The development of lightweight model variants could facilitate deployment in resource-constrained environments. Investigation of semi-supervised learning approaches might help address the limited availability of labeled medical data.

## 7. Conclusion

The study reveals that deep learning architectures are not consistently effective in medical image classification. Instead, task-specific characteristics should guide architecture selection, considering trade-offs between model complexity and performance. The findings have clinical implications, as deep learning models' performance varies across medical domains. Future research should focus on larger, diverse datasets, task-specific architectural optimization, and finding the right balance between model performance and computational efficiency. Collaboration between technical experts and medical professionals is crucial for achieving the full potential of deep learning in medical image classification.