Details on our MedFM Challenge Results

Dr. Adrian Krenzer, Amar Hekalo, Marcel Roth, Micha Nowak,
Prof. Dr. Frank Puppe

Julius-Maximilians-Universität Würzburg

Department of Artificial Intelligence and Knowledge Systems

15.12.2023

1. **Introduction and Challenge Overview**:
    1. Emphasis on foundation models in medical imaging, addressing challenges with limited high-quality annotated datasets.
    2. Three tasks: Thoracic Disease Screening, Pathological Tumor Tissue Classification, Lesion Detection in Colonoscopy Images.
2. **Methodology**:
    1. Split into exploratory data analysis, preprocessing, data augmentation, model training, and ensemble strategies.
    2. Discussion on various models used, including Vision Transformers and ResNets.
    3. In-depth look at data augmentation techniques for different tasks.
3. **Results**:
    1. Performance of various models across tasks.
    2. Analysis of results using metrics like AUC, Accuracy, and Aggregate Score.
    3. Demonstrated the benefits of ensemble models and fine-tuning strategies.
4. **Conclusion**:
    1. Emphasizes learning experience and the effectiveness of ensemble models.
    2. Highlights the importance of grid search in optimizing models and strategies.

MedFM Challenge:

Focus: Participation in the prestigious MedFM Challenge 2023.

Challenge Relevance: Addresses the scarcity of high-quality, annotated medical imaging datasets.

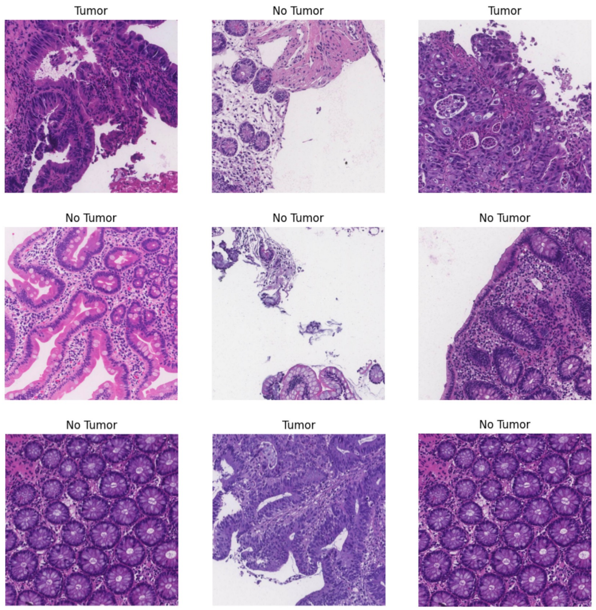
Main Objectives: Enhancing foundation models for medical image analysis.

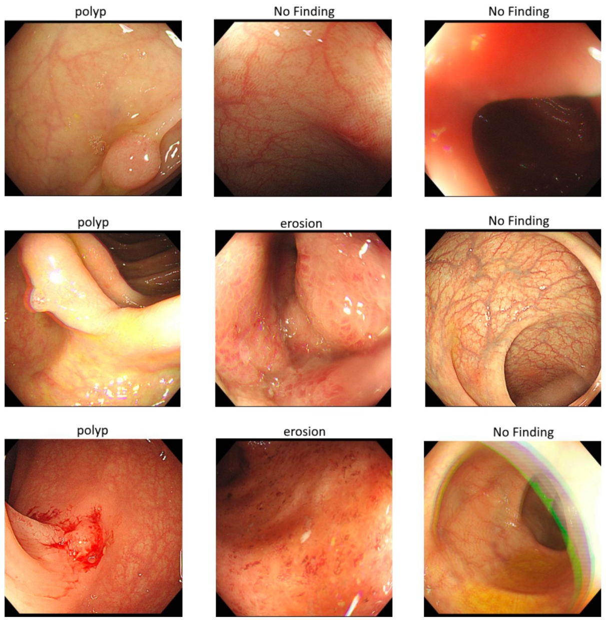
Key Tasks: Involves Thoracic Disease Screening, Pathological Tumor Tissue Classification, and Lesion Detection in Colonoscopy Images.

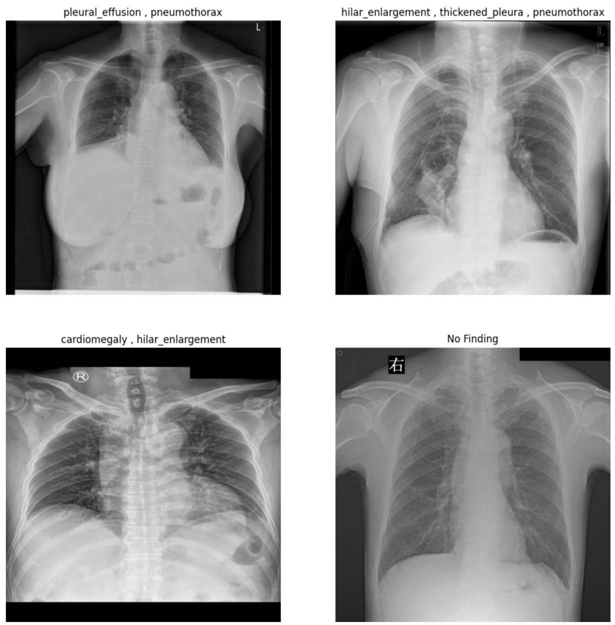
Methodology

1. **Exploratory Data Analysis**: Initial assessment of datasets to understand characteristics and challenges.
2. **Data Preprocessing**: Techniques to clean and standardize the data for model input.
3. **Data Augmentation**: Application of various augmentation strategies to enhance dataset diversity.
4. **Model Selection and Training**: Discussion of models like Vision Transformers and ResNets, and their training approaches.

Data







Colon: 5656 images (9.5 GB)

2 classes

Endo: 1811 images (2.7 GB)

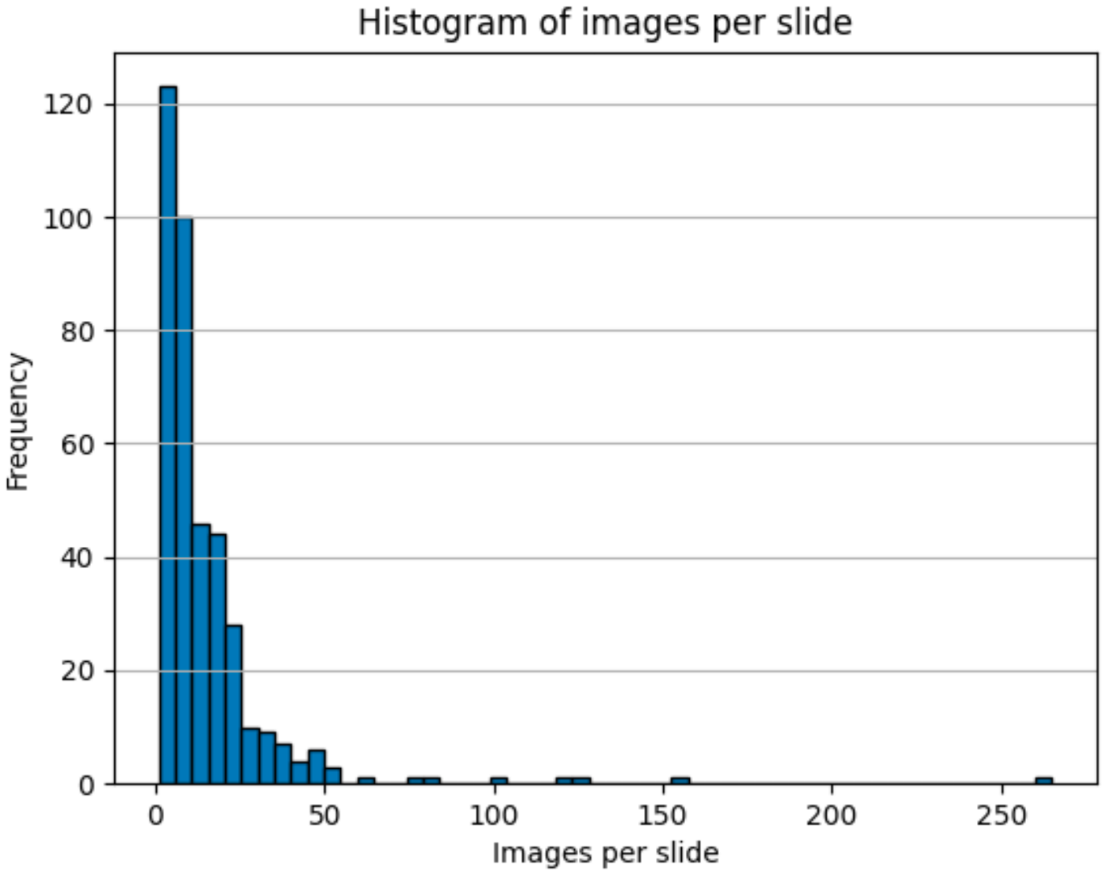
4 classes

Chest: 2141 images (2.6 GB)

19 classes

Data augmentation colon

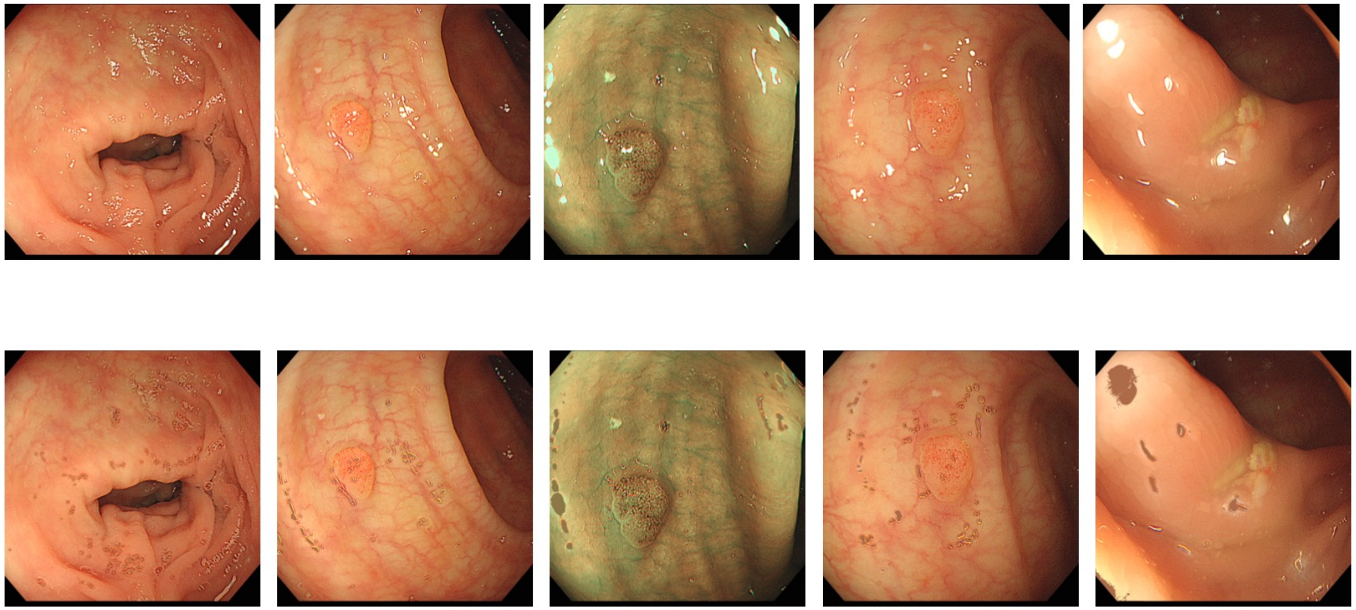
1. **Dataset Overview**: Uses the Colon dataset, divided into 2 categories for tumor tissue classification.
2. **Augmentation Process**: Implements random resized crops and horizontal/vertical flips to enhance robustness.
3. **ColorJitter Augmentation**: Adjusts image brightness, contrast, color saturation, and hue to improve classification accuracy.
4. **Data Loading Configurations**: Varies in batch size and pipeline across training, validation, and testing phases for optimal performance.


Histogram of images per slide
Frequency
Images per slide
0
20
40
60
80
100
120
0
50
100
150
200
250

Data augmentation Endo

1. **Dataset and Challenge**: Focuses on the Endoscopy dataset with four classes, addressing challenges from specular reflections.
2. **Adaptive-RPCA Implementation**: Utilizes Adaptive-RPCA for identifying and removing reflections, enhancing image clarity for classification.
3. **Training Data Augmentation**: Applies random resized cropping and horizontal/vertical flips to enhance model generalization.
4. **Data Processing and Loading**: Simplified testing data processing with consistent resizing, and tailored data loading configurations for training and validation.

Application of Adaptive-RPCA on a selection of endoscopic images



Data augmentation Chest

1. **Dataset and Preprocessing**: Utilizes the Chest dataset with 19 classes; RGB images normalized using specific means and standard deviations.
2. **Training Augmentation**: Incorporates random affine transformations, resizing, and horizontal flips, with crop scales tailored to the model.
3. **Testing Process**: Simplifies augmentation, resizing images to specific resolutions, with a focus on model evaluation.
4. **Optimization**: Employs a hyperparameter grid-search to determine optimal training batch sizes and randomization seeds for each model.

Models

1. **Vision Transformer (ViT)**: Transforms images into sequences of patches for analysis, excelling at capturing long-range dependencies through self-attention.
2. **Visual Prompt Tuning (VPT)**: Fine-tunes pre-trained transformer models using visual prompts, optimizing efficiency by learning only the prompt's embedding.
3. **ResNet**: Employs skip connections to overcome gradient issues in deep learning, allowing for efficient learning and rapid hyperparameter iteration.
4. **Swin Transformer**: Enhances ViT with a hierarchical approach and sliding window self-attention, offering better efficiency and scalability, especially in Swin V2.

Hyperparmeter optimization

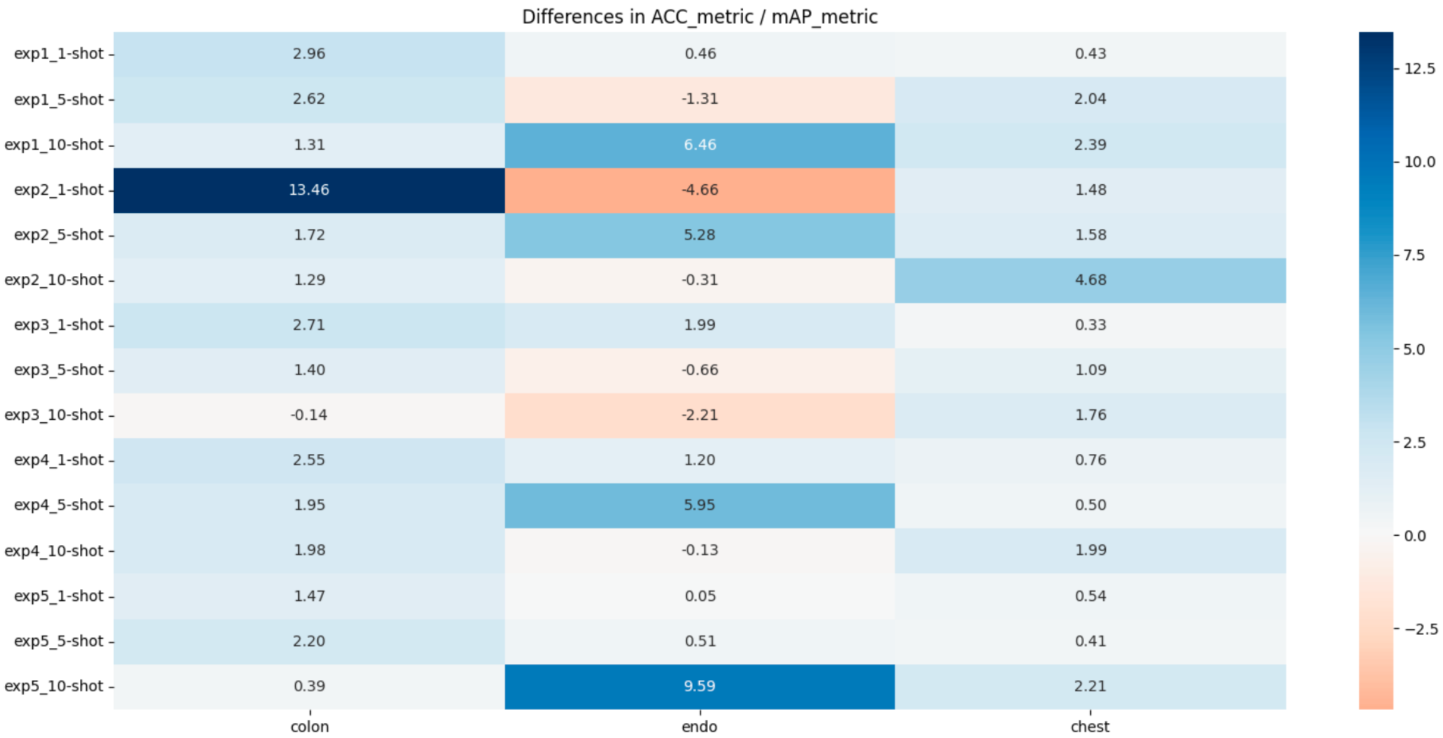
```
BASE_PARAMS_CONFIG = {
'model': ["clip-b_vpt", "dinov2-b_vpt", "eva-b_vpt", "swin-b_vpt",
"swinv2-b", "vit-b_vpt"],
'dataset': ["chest", "colon", "endo"],
'shot': [1, 5, 10],
'exp_num': [1, 2, 3, 4, 5],
'lr': [1e-6, 1e-7, 1e-8],
'train_bs': [2, 4, 6, 8]
}
```

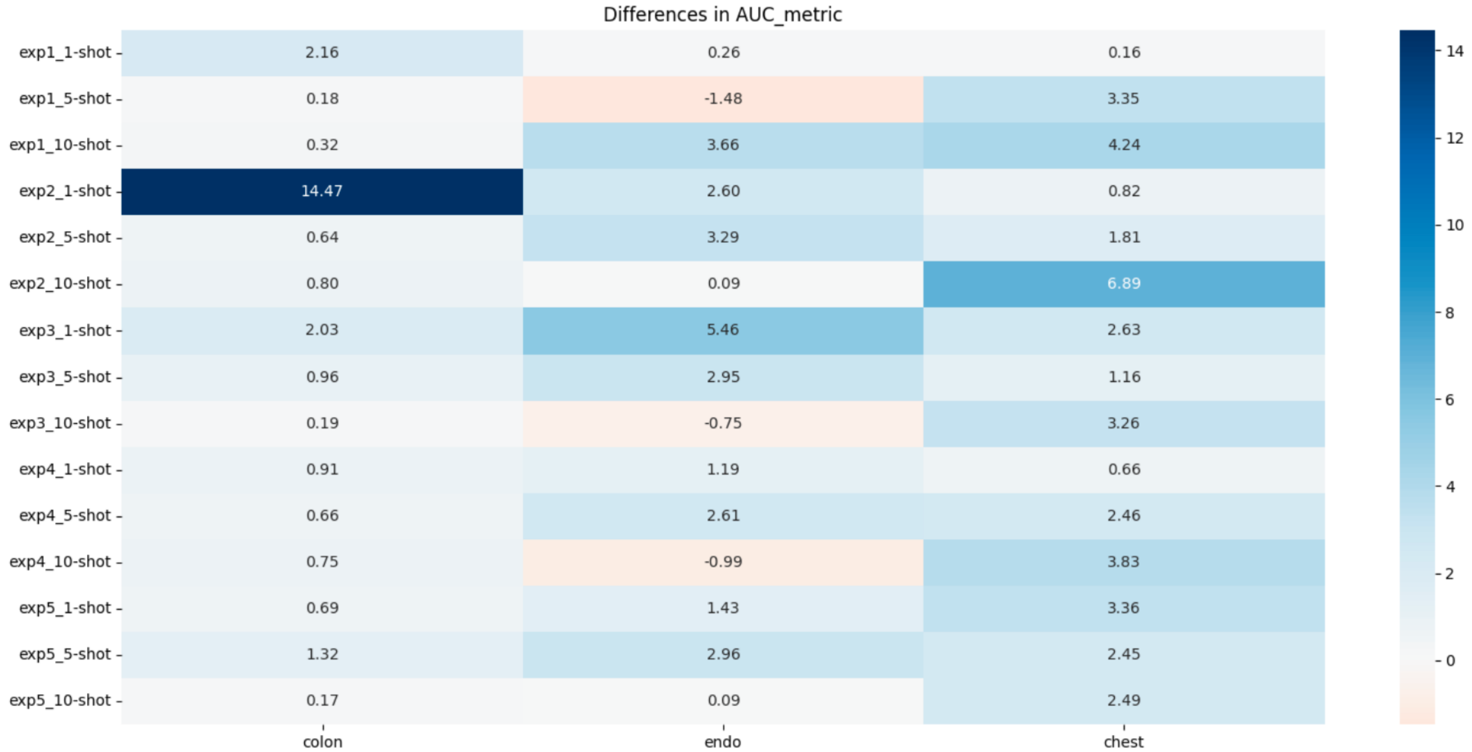
Ensemble technics

1. **Expert-per-Task**: Selects the top model for each task/setting.
2. **Expert-per-Class**: Chooses top model for each class in a task/setting.
3. **Weighted**: Assigns weights to top-k models per class, normalized before combining predictions.
4. **Performance-Difference-Weighted**: Weights models based on performance difference from the lowest performer.
5. **Performance-Difference-Log-Weighted**: Similar to above but with log scaling.
6. **Weighted-Expert-per-Class**: Builds on weighted strategy with optional scaling; normalization on final predictions.
7. **Log-Weighted-Expert-per-Class**: Uses log scaling on the weighted-expert-per-class strategy.
8. **Expo-Weighted-Expert-per-Class**: Applies exponential scaling in the weighted-expert-per-class method.
9. **SM-Weighted-Expert-per-Class**: Incorporates softmax scaling in the weighted-expert-per-class strategy.
10. **Rank-Based-Weighted**: Ranks models, setting weights inversely proportional to their ranks.

Ensemble Gridsearch

1. **Model Assessment**: Evaluated 1,200 models with an average of 18.2 models per setting, leading to a top-k parameter of 20.
2. **Unique Approaches**: Resulted in 8,190 unique approaches (45 settings x average 18.2 top-k x 10 strategies).
3. **Structured Approach**: Implemented a timestamped directory for systematic segregation of model outputs based on task, shot, and experiment.
4. **Selection and Application**: Adopted a grid-search-like strategy for model selection, applying the best-performing ensemble strategies from the validation set to the final submission.


Differences in ACC_metric / mAP_metric
colon
endo
chest
exp1_1-shot 2.96 0.46 0.43
exp1_5-shot 2.62 -1.31 2.04
exp1_10-shot 1.31 6.46 2.39
exp2_1-shot 13.46 -4.66 1.48
exp2_5-shot 1.72 5.28 1.58
exp2_10-shot 1.29 -0.31 4.68
exp3_1-shot 2.71 1.99 0.33
exp3_5-shot 1.40 -0.66 1.09
exp3_10-shot -0.14 -2.21 1.76
exp4_1-shot 2.55 1.20 0.76
exp4_5-shot 1.95 5.95 0.50
exp4_10-shot 1.98 -0.13 1.99
exp5_1-shot 1.47 0.05 0.54
exp5_5-shot 2.20 0.51 0.41
exp5_10-shot 0.39 9.59 2.21
12.5
10.0
7.5
5.0
2.5
0.0
−2.5


Differences in AUC_metric
exp1_1-shot
2.16
0.26
0.16
exp1_5-shot
0.18
-1.48
3.35
exp1_10-shot
0.32
3.66
4.24
exp2_1-shot
14.47
2.60
0.82
exp2_5-shot
0.64
3.29
1.81
exp2_10-shot
0.80
0.09
6.89
exp3_1-shot
2.03
5.46
2.63
exp3_5-shot
0.96
2.95
1.16
exp3_10-shot
0.19
-0.75
3.26
exp4_1-shot
0.91
1.19
0.66
exp4_5-shot
0.66
2.61
2.46
exp4_10-shot
0.75
-0.99
3.83
exp5_1-shot
0.69
1.43
3.36
exp5_5-shot
1.32
2.96
2.45
exp5_10-shot
0.17
0.09
2.49
colon
endo
chest
14
12
10
8
6
4
2
0

Results and Performance Analysis

1. **Model Performance Metrics**: Evaluation using AUC, Accuracy, and Aggregate Score.
2. **Comparison Across Tasks**: Analysis of model effectiveness in different tasks.
3. **Ensemble Model Benefits**: Demonstrating improved results with ensemble strategies.
4. **Fine-Tuning Impact**: The significance of model fine-tuning in performance enhancement.

Results

model	shot	AUC-Acc	AUC	Acc
convnext-v2-b	1-shot	76.340726	80.486287	72.195166
convnext-v2-b	10-shot	94.368138	97.469795	91.266467
convnext-v2-b	5-shot	89.386615	93.324950	85.448270
resnet101-CSRA	1-shot	82.332940	85.415590	79.250287
resnet101-CSRA	10-shot	94.494678	97.241228	91.748126
resnet101-CSRA	5-shot	90.814232	93.917209	87.711257
swin-b_vpt	1-shot	78.045265	83.389229	72.701312
swin-b_vpt	10-shot	96.832200	98.822100	94.842200
swin-b_vpt	5-shot	94.013033	98.007867	90.018200
swinv2-b	1-shot	84.573043	90.568157	78.577936
swinv2-b	10-shot	95.782956	98.972222	92.593711
swinv2-b	5-shot	92.548450	96.931593	88.165314

Table 1: Performance results of a selection of our best-peforming models on the pathological tumor tissue classification task.

Results

model	shot	AUC-Acc	AUC	Acc
convnext-v2-b	1-shot	76.340726	80.486287	72.195166
convnext-v2-b	10-shot	94.368138	97.469795	91.266467
convnext-v2-b	5-shot	89.386615	93.324950	85.448270
resnet101-CSRA	1-shot	82.332940	85.415590	79.250287
resnet101-CSRA	10-shot	94.494678	97.241228	91.748126
resnet101-CSRA	5-shot	90.814232	93.917209	87.711257
swin-b_vpt	1-shot	78.045265	83.389229	72.701312
swin-b_vpt	10-shot	96.832200	98.822100	94.842200
swin-b_vpt	5-shot	94.013033	98.007867	90.018200
swinv2-b	1-shot	84.573043	90.568157	78.577936
swinv2-b	10-shot	95.782956	98.972222	92.593711
swinv2-b	5-shot	92.548450	96.931593	88.165314

Table 1: Performance results of a selection of our best-peforming models on the pathological tumor tissue classification task.

Results

model	shot	AUC-Acc	AUC	Acc
convnext-v2-b	1-shot	76.340726	80.486287	72.195166
convnext-v2-b	10-shot	94.368138	97.469795	91.266467
convnext-v2-b	5-shot	89.386615	93.324950	85.448270
resnet101-CSRA	1-shot	82.332940	85.415590	79.250287
resnet101-CSRA	10-shot	94.494678	97.241228	91.748126
resnet101-CSRA	5-shot	90.814232	93.917209	87.711257
swin-b_vpt	1-shot	78.045265	83.389229	72.701312
swin-b_vpt	10-shot	96.832200	98.822100	94.842200
swin-b_vpt	5-shot	94.013033	98.007867	90.018200
swinv2-b	1-shot	84.573043	90.568157	78.577936
swinv2-b	10-shot	95.782956	98.972222	92.593711
swinv2-b	5-shot	92.548450	96.931593	88.165314

Table 1: Performance results of a selection of our best-peforming models on the pathological tumor tissue classification task.

Results

model	shot	AUC-Acc	AUC	Acc
convnext-v2-b	1-shot	76.340726	80.486287	72.195166
convnext-v2-b	10-shot	94.368138	97.469795	91.266467
convnext-v2-b	5-shot	89.386615	93.324950	85.448270
resnet101-CSRA	1-shot	82.332940	85.415590	79.250287
resnet101-CSRA	10-shot	94.494678	97.241228	91.748126
resnet101-CSRA	5-shot	90.814232	93.917209	87.711257
swin-b_vpt	1-shot	78.045265	83.389229	72.701312
swin-b_vpt	10-shot	96.832200	98.822100	94.842200
swin-b_vpt	5-shot	94.013033	98.007867	90.018200
swinv2-b	1-shot	84.573043	90.568157	78.577936
swinv2-b	10-shot	95.782956	98.972222	92.593711
swinv2-b	5-shot	92.548450	96.931593	88.165314

Table 1: Performance results of a selection of our best-peforming models on the pathological tumor tissue classification task.

model	shot	AUC-mAP	AUC	mAP
resnet101-CSRA	1-shot	36.201033	55.211860	17.190204
resnet101-CSRA	10-shot	41.102515	60.542692	21.662354
resnet101-CSRA	5-shot	40.259908	59.399269	21.120550
swin-b_vpt	1-shot	37.887900	58.086800	17.688900
swin-b_vpt	10-shot	40.895092	61.337317	20.452883
swin-b_vpt	5-shot	40.476100	60.760660	20.191500
swinv2-b	1-shot	38.649044	58.041118	19.256985
swinv2-b	10-shot	49.042674	68.695778	29.389556
swinv2-b	5-shot	42.768421	62.754743	22.782114

Table 2: Performance results of a selection of our best-peforming models on the endoscopic image analysis task.

model	shot	AUC-mAP	AUC	mAP
resnet101-CSRA	1-shot	36.201033	55.211860	17.190204
resnet101-CSRA	10-shot	41.102515	60.542692	21.662354
resnet101-CSRA	5-shot	40.259908	59.399269	21.120550
swin-b_vpt	1-shot	37.887900	58.086800	17.688900
swin-b_vpt	10-shot	40.895092	61.337317	20.452883
swin-b_vpt	5-shot	40.476100	60.760660	20.191500
swinv2-b	1-shot	38.649044	58.041118	19.256985
swinv2-b	10-shot	49.042674	68.695778	29.389556
swinv2-b	5-shot	42.768421	62.754743	22.782114

Table 2: Performance results of a selection of our best-peforming models on the endoscopic image analysis task.

model	shot	AUC-mAP	AUC	mAP
resnet101-CSRA	1-shot	36.201033	55.211860	17.190204
resnet101-CSRA	10-shot	41.102515	60.542692	21.662354
resnet101-CSRA	5-shot	40.259908	59.399269	21.120550
swin-b_vpt	1-shot	37.887900	58.086800	17.688900
swin-b_vpt	10-shot	40.895092	61.337317	20.452883
swin-b_vpt	5-shot	40.476100	60.760660	20.191500
swinv2-b	1-shot	38.649044	58.041118	19.256985
swinv2-b	10-shot	49.042674	68.695778	29.389556
swinv2-b	5-shot	42.768421	62.754743	22.782114

Table 2: Performance results of a selection of our best-peforming models on the endoscopic image analysis task.

Results

model	shot	AUC-mAP	AUC	mAP
densenet121	1-shot	42.5952	63.0547	22.1356
densenet121	10-shot	42.3039	63.5197	21.0881
densenet121	5-shot	40.6961	62.5318	18.8604
swinv2-b	1-shot	38.0050	58.8893	17.1206
swinv2-b	10-shot	47.6696	68.8499	26.4893
swinv2-b	5-shot	46.8300	68.2585	25.4014
resnet101	1-shot	39.8241	60.3921	19.2562
resnet101	10-shot	49.6390	72.0935	27.1844
resnet101	5-shot	46.0372	68.7443	23.3303
efficientnetv2-s	1-shot	39.8216	60.4742	19.1689
efficientnetv2-s	10-shot	47.5157	68.4753	26.5562
efficientnetv2-s	5-shot	46.4347	68.7042	24.1653
clip-b_vpt	1-shot	39.080433	58.467100	19.693767
clip-b_vpt	10-shot	43.204700	64.367350	22.042150
clip-b_vpt	5-shot	47.774067	66.540733	29.007383
dinov2-b_vpt	1-shot	37.973650	57.284150	18.663150
dinov2-b_vpt	10-shot	42.399700	63.591150	21.208250
dinov2-b_vpt	5-shot	41.358143	62.555186	20.161129
eva02-b_vpt	1-shot	36.286150	56.385100	16.187250
eva02-b_vpt	10-shot	40.685300	61.088350	20.282250
eva02-b_vpt	5-shot	39.248550	60.477550	18.019500
swin-b_vpt	1-shot	38.177986	57.733214	18.622743
swin-b_vpt	10-shot	45.386900	66.872100	23.901700
swin-b_vpt	5-shot	42.426650	63.928450	20.924800
vit-b_vpt	1-shot	39.294200	59.060600	19.527700
vit-b_vpt	10-shot	44.048400	65.015900	23.080800
vit-b_vpt	5-shot	41.416600	62.775200	20.058100

Table 3: Performance results of a selection of our best-peforming models on the thoracic disease screening task.

Results

model	shot	AUC-mAP	AUC	mAP
densenet121	1-shot	42.5952	63.0547	22.1356
densenet121	10-shot	42.3039	63.5197	21.0881
densenet121	5-shot	40.6961	62.5318	18.8604
swinv2-b	1-shot	38.0050	58.8893	17.1206
swinv2-b	10-shot	47.6696	68.8499	26.4893
swinv2-b	5-shot	46.8300	68.2585	25.4014
resnet101	1-shot	39.8241	60.3921	19.2562
resnet101	10-shot	49.6390	72.0935	27.1844
resnet101	5-shot	46.0372	68.7443	23.3303
efficientnetv2-s	1-shot	39.8216	60.4742	19.1689
efficientnetv2-s	10-shot	47.5157	68.4753	26.5562
efficientnetv2-s	5-shot	46.4347	68.7042	24.1653
clip-b_vpt	1-shot	39.080433	58.467100	19.693767
clip-b_vpt	10-shot	43.204700	64.367350	22.042150
clip-b_vpt	5-shot	47.774067	66.540733	29.007383
dinov2-b_vpt	1-shot	37.973650	57.284150	18.663150
dinov2-b_vpt	10-shot	42.399700	63.591150	21.208250
dinov2-b_vpt	5-shot	41.358143	62.555186	20.161129
eva02-b_vpt	1-shot	36.286150	56.385100	16.187250
eva02-b_vpt	10-shot	40.685300	61.088350	20.282250
eva02-b_vpt	5-shot	39.248550	60.477550	18.019500
swin-b_vpt	1-shot	38.177986	57.733214	18.622743
swin-b_vpt	10-shot	45.386900	66.872100	23.901700
swin-b_vpt	5-shot	42.426650	63.928450	20.924800
vit-b_vpt	1-shot	39.294200	59.060600	19.527700
vit-b_vpt	10-shot	44.048400	65.015900	23.080800
vit-b_vpt	5-shot	41.416600	62.775200	20.058100

Table 3: Performance results of a selection of our best-peforming models on the thoracic disease screening task.

Results

model	shot	AUC-mAP	AUC	mAP
densenet121	1-shot	42.5952	63.0547	22.1356
densenet121	10-shot	42.3039	63.5197	21.0881
densenet121	5-shot	40.6961	62.5318	18.8604
swinv2-b	1-shot	38.0050	58.8893	17.1206
swinv2-b	10-shot	47.6696	68.8499	26.4893
swinv2-b	5-shot	46.8300	68.2585	25.4014
resnet101	1-shot	39.8241	60.3921	19.2562
resnet101	10-shot	49.6390	72.0935	27.1844
resnet101	5-shot	46.0372	68.7443	23.3303
efficientnetv2-s	1-shot	39.8216	60.4742	19.1689
efficientnetv2-s	10-shot	47.5157	68.4753	26.5562
efficientnetv2-s	5-shot	46.4347	68.7042	24.1653
clip-b_vpt	1-shot	39.080433	58.467100	19.693767
clip-b_vpt	10-shot	43.204700	64.367350	22.042150
clip-b_vpt	5-shot	47.774067	66.540733	29.007383
dinov2-b_vpt	1-shot	37.973650	57.284150	18.663150
dinov2-b_vpt	10-shot	42.399700	63.591150	21.208250
dinov2-b_vpt	5-shot	41.358143	62.555186	20.161129
eva02-b_vpt	1-shot	36.286150	56.385100	16.187250
eva02-b_vpt	10-shot	40.685300	61.088350	20.282250
eva02-b_vpt	5-shot	39.248550	60.477550	18.019500
swin-b_vpt	1-shot	38.177986	57.733214	18.622743
swin-b_vpt	10-shot	45.386900	66.872100	23.901700
swin-b_vpt	5-shot	42.426650	63.928450	20.924800
vit-b_vpt	1-shot	39.294200	59.060600	19.527700
vit-b_vpt	10-shot	44.048400	65.015900	23.080800
vit-b_vpt	5-shot	41.416600	62.775200	20.058100

Table 3: Performance results of a selection of our best-peforming models on the thoracic disease screening task.

Conclusion

1. **Swin Transformer V2 Efficiency**: Swin Transformer V2 demonstrated remarkable performance, highlighting its effectiveness in medical image classification tasks.
2. **Ensemble Model Strength**: The use of ensemble strategies significantly enhanced prediction accuracy, proving their value in complex challenges.
3. **Crucial Role of Data Preparation**: Effective data preparation played a key role in model performance, underscoring its importance.
4. **Impact of Data Augmentation**: Data augmentation techniques were crucial in improving model robustness and generalization capabilities.