Few-shot Image Classification for Breast Cancer Detection

Qirun Dai **Jingzhi Sun** **Pengyu Chen** **Xiaowei Zeng**

1 Introduction

Breast cancer is a pressing concern worldwide, ranking as the second leading cause of cancer death in women [Miller et al., 2022]. Fortunately, early detection and identification of breast cancer can lead to timely treatment, effectively reducing the risk of further deterioration or death. Breast ultrasound image classification serves as a primary method for such detection. However, traditional medical image classification often demands considerable human expertise and time, making it impractical for many underdeveloped countries and regions. Hence, there has been a shift towards pattern recognition and machine learning approaches to automate and streamline medical image classification and breast cancer detection. Numerous learning-based methods, ranging from traditional machine learning to modern deep learning, have been proposed.

However, two major challenges remain relatively unaddressed. Firstly, due to the inherent data sparsity of real-life medical image data [Varoquaux and Cheplygina, 2022], current machine learning methods that thrive on extensive datasets struggle to generalize and prevent overfitting in a few-shot training setting. Secondly, the swift advancements in the deep learning community mean a constant influx of new vision models, complicating comprehensive comparisons and evaluations between traditional machine learning models and cutting-edge deep learning models. Our project aims to address these challenges with the following guidelines:

1. Our research seeks to mimic the data-scarce real-world environment of the medical image classification domain. To achieve this, we utilize a special breast ultrasound image dataset comprising only 780 training and testing images combined. By doing so, our project aligns with a genuine few-shot training scenario, adding significant real-world application value.
2. We also aim to perform a comprehensive evaluation involving both traditional machine learning models and state-of-the-art deep learning models. Specifically, we experiment on four traditional machine learning models representing three distinct learning paradigms: for **parametric supervised learning**, we use Logistic Regression and Naive Bayes; for **non-parametric supervised learning**, we use Support Vector Machine; for **ensembling methods**, we use Random Forest. For the deep learning models, we employ two leading vision architectures: **Convolutional Neural Networks (CNN)** and **Vision Transformers (ViT)**. Furthermore, to tackle the problem of data scarcity posed by the few-shot setting, we employ a prevalent training paradigm called **Transfer Learning**, which takes CNN and ViT as the base architecture for pretraining and then finetuning. Through this approach, our project endeavors to create a comprehensive evaluation framework for few-shot medical image classification, encompassing both statistical learning and deep learning paradigms indicative of prior research.

2 Background Knowledge and Related Works

Naïve Bayes Naïve Bayes proves significantly effective due to the randomness in medical images. [Zaw et al., 2019] extracts features from the brain tumor images and trains them in the Naïve Bayes classifier so that it can predict the test image whether it is normal or tumor. Compared with other methods that wrongly detect the eyes and bones as tumors, Naïve Bayes can reliably distinguish between the tumors and other non-tumor objects (eyes, bones). In scenarios with high-dimension

Preprint. Under review.

features, [Ramesh Kumar and Vijaya, 2022] introduces feature ranking techniques and Principal Component Analysis (PCA) to enhance the Naïve Bayes Classifier, reducing feature dimensionality and computational modeling costs while increasing interpretability at minimized information loss. The results show that Naïve Bayes consistently delivered high predictive accuracy.

Logistic Regression Logistic regression stands as a pivotal supervised learning technique grounded in probability functions and predominantly employed for classification tasks. Logistic regression can perform well even with relatively small datasets, making it advantageous in scenarios where limited labeled medical imaging data is available. This is especially relevant in medical fields where obtaining large labeled datasets can be challenging due to issues like data privacy and the cost of obtaining labeled medical images. [Dinesh and Kalyanasundaram, 2022] finds that Logistic Regression appears to be better than the SVM, KNN, Decision Tree, Random Forest in breast cancer detection using the Wisconsin dataset.

Support vector machine SVM, Support Vector Machine, has been a groundbreaking development in the realm of machine learning. Developed by Vapnik and his team in the 1990s, SVM theory is influenced by neural networks and can be seen as a mathematical extension of them. The concept that only the training samples lying on the class boundaries are crucial for classification is the fundamental principle behind SVM's approach. In medical imaging, where the acquisition of training data can be costly and limited due to factors like expert annotation or specific imaging techniques, SVM's ability to classify based on these critical support vectors becomes highly valuable. [Chi et al., 2008] emphasizes SVM's efficiency in classifying small-sized training datasets, showcasing its ability to generalize well in scenarios with high-dimensional input spaces. [Alrais and Elfadil, 2020] proposes SVM model for classifying the medical image into two types of tumors. Author before classifying the images uses DWT (Discrete Wavelet Transform) to remove noise from the images and PCA (Principal Component Analysis) for reducing the dimensional feature to get the only required storage and computational space. Similarly, our utilizing mask data also effectively eliminates noise from the dataset, resulting in improved classification performance.

Decision Tree and Random Forest Decision Trees (DTs) prove invaluable in various applications. DTs excel at organizing diverse image data, ensuring accurate categorization. Each DT assesses image features, including pixel brightness, texture changes, and breast tissue-related details. By iteratively partitioning data based on these features, DTs differentiate 'normal,' 'benign,' and 'malignant' images, ensuring precision [Riri et al., 2016].

Moreover, [Kaganov et al., 2018] find that a decision tree model based on MRI signal intensities aids in diagnosing uterine leiomyosarcomas, revealing a significant relationship between histopathological type and T1 and T2 intensity signals. However, it remains crucial to control the level of detail in DTs to avoid overfitting, particularly in medical imaging, as excessive focus on irrelevant details can lead to misclassification.

In Random Forests (RF), each tree independently evaluates images, scrutinizing various features such as texture, color patterns, and brightness to detect changes in breast tissue. Since each tree is trained with different data and features, they offer unique perspectives. This diversity yields a broader and more detailed understanding of breast cancer images.

[Criminisi et al., 2010] proposes a paper on multi-class random regression forests as an algorithm for the efficient, automatic detection, and localization of anatomical structures within three-dimensional CT scans. Researchers perform quantitative validation on a database of 100 highly variable CT scans, demonstrating that localization errors are lower (and more stable) than those from global affine registration approaches. The regressor's parallelism and the simplicity of its context-rich visual features yield typical runtimes of only 1 second. Applications include semantic visual navigation, image tagging for retrieval, and initializing organ-specific processing.

CNN and ViT. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are two most prevalent and state-of-the-art architectures in the current field of computer vision. At their core, CNNs [Krizhevsky et al., 2012] utilize convolution kernels to recognize patterns in images. These kernels operate with two primary insights: firstly, they detect features or patterns that are typically smaller than the complete input image, and secondly, they acknowledge that these features can emerge multiple times in various parts of the image. This design inherently understands the spatial hierarchies in visual data compared to traditional Multi-layer Perceptrons (MLPs), effectively recognizing intricate structures within images. In contrast, Vision Transformers (ViTs) dissect an

image into fixed-size patches and then transform them into a series of vectors using a transformer encoder [Dosovitskiy et al., 2020]. This approach is rooted in harnessing the potent long-range modelling capability of the attention mechanism present in transformer architectures. By doing so, ViTs evade the inductive bias inherently introduced by CNNs' convolution design [Lippe, 2023], and push back the frontiers of long-range dependency modelling in computer vision.

Transfer Learning in Medical Image Classification. Despite the proven capabilities of CNN and ViT in general vision tasks, it is not guaranteed that they will consistently show strong performance on medical image classification, mainly due to the severe training data scarcity of medical images. In order to tackle this problem, we resort to a new training paradigm called Transfer Learning (TL), which has been trending in few-shot learning tasks in recent years. The core idea of transfer learning is to pretrain a vision model on a very large dataset (e.g. ImageNet [Deng et al., 2009], which contains 1.2 million images with 1000 categories), and then use the pretrained model either as an initialization or a fixed feature extractor for the task of interest. [Kim et al., 2022] conducted a comprehensive survey of transfer learning in the field of medical image classification, demonstrating its efficacy with deep convolutional neural networks (e.g. ResNet [He et al., 2016] and GoogLeNet [Szegedy et al., 2015]) as the pretrained backbone model. Moreover, with the advent of Vision Transformers, a large proportion of recent research also concentrates on the application of ViT under the transfer learning framework. [Matsoukas et al., 2022] studied the working mechanism of two data-efficient ViT models (DeiT [Touvron et al., 2021] and Swin Transformers [Liu et al., 2021]) in transfer learning for medical image classification, and found that the benefits from transfer learning increase with reduced data size and models with fewer inductive biases. However, their work did not focus on the extreme data-insufficient scenario like ours, and mainly concentrated on the domain discrepancy between the natural image domain (i.e. ImageNet that the backbone model was trained on) and the medical image domain.

3 Methods

3.1 Statistical Learning Methods

Naïve Bayes Naïve Bayes calculates the posterior probability for each class and predicts the class with the highest probability. The term 'naïve' comes from a strong assumption that the features are conditionally independent of one another given the class label, which enhances the algorithm's computational efficiency. While this assumption may not always hold in real-world scenarios, it actually demonstrates competitive classification accuracy in most cases for the resilience against image noise [Webb et al., 2010]. It performs even better in the breast ultrasound image segmentation (BUSI) scenarios because the randomness of medical events preserves the independence assumption.

Logistic Regression Unlike traditional regression, logistic regression assesses the likelihood of an event happening by applying the sigmoid function to transform the output into a range between 0 and 1. This transformation helps in mapping the continuous input space into a restricted output range, effectively assigning probabilities to different classes in a more intricate and nuanced manner. Logistic regression assumes that the relationship between the log-odds of the variables should be approximately linear. While this assumption might be reasonable in certain cases, the complex nature of medical images introduces challenges to this linearity assumption.

Support Vector Machine The core principle behind SVM is to find an optimal hyperplane for classification. By mapping data into a higher-dimensional space, SVM creates a decision plane that optimally separates different classes. Kernel function is fundamental for nonlinear mapping in higher-dimensional spaces, which allows SVM to handle the classification of data in higher dimensions based on the originally input data space. The kernel function computes the value of the dot product of mapped data points in the feature space. Its advantage lies in ensuring that the complexity of the problem is primarily dependent on the dimensionality of the input space rather than the feature space.

Random Forest In our study, we also investigate the efficacy of **Random Forest** (RF) models for the few-shot learning task, particularly focusing on the BUSI task. We leverage the robust and interpretable nature of RF models, which are well-suited for medical image classification, especially in scenarios with limited training data.

Model Fitting Outlines The approach for implementing the aforementioned statistical methods in our study is detailed as follows:

1. **Feature Extraction:** Initially, we extract relevant features from the BUSI dataset. This step involves preprocessing the images to enhance their characteristics, crucial for capturing distinct patterns in normal, benign, and malignant cases.
2. **Model Training with Non-masked Images**: The models are initially trained using features extracted from non-masked images. This phase involves tuning standard hyperparameters and results in a baseline model for the classification task.
3. **Model Training with Masked Images and Grid Search**: Subsequently, all the classifiers are trained using features from masked images. The training process includes tuning hyperparameters (Table 1) to optimize the model's performance. In this stage, we employ a grid search strategy to fine-tune hyperparameters, optimizing the model's performance specifically for the masked image dataset. This methodical approach is aimed at achieving the best possible accuracy and generalization for the BUSI task.
4. **Model Evaluation**: The performance of our models, trained separately on non-masked and masked images, is evaluated on a distinct test set. We use metrics such as accuracy, precision, recall, and F1-score to assess the model's ability in classifying ultrasound images under different preprocessing conditions.

Model	**Hyperparameters**
Naïve Bayes	`var_smoothing`
Logistic Regression	`C, penalty`
Support Vector Machine	`C, gamma, kernel`
Random Forest	`n_estimators, max_depth, minsamples_split`

Table 1: Hyperparameters tuned in statistical models.

Consistent with studies emphasizing the importance of hyperparameter optimization in machine learning models Criminisi et al. [2010], our approach involves a meticulous tuning process, ensuring that the models are well-adapted to the specific characteristics of the BUSI task. Furthermore, the interpretability of our models remains a significant advantage, providing insights into feature importance and decision-making processes.

3.2 Deep Learning Methods

Since another focus of our research is on evaluating the few-shot learning performance of **state-of-the-art** deep learning models, we employ the prevalent framework of transfer learning, with CNN and ViT as the base models for pretraining and finetuning. Specifically, we follow previous studies [Matsoukas et al., 2022] by using two SOTA vision architectures, EfficientNet [Tan and Le, 2019] and Swin Transformer [Liu et al., 2021], as the base models pretrained on ImageNet-1k, and employing weight transfer as the initialization strategy for finetuning. The complete transfer learning pipeline is shown below:

1. Pretrain the base model on ImageNet-1k task, resulting in parameters W and b.
2. Initialize a second network of the same architecture as the base model with all of W and b, except for the final linear classification layer whose output dimension is reset to match the number of classes in BUSI task.
3. Train (Finetune) the second network on BUSI task, resulting in final parameters W' and b' for testing.

When finetuning the second network on BUSI task, we employ an end-to-end finetuning paradigm by not freezing any of the hidden layers in the base model after weight transfer, which is also consistent with previous studies in transfer learning for medical image classification [Matsoukas et al., 2022]. Moreover, prior study of representation learning [Huang et al., 2023] shows that end-to-end finetuning leads to consistently better performance regardless of the dataset size in medical image classification.

4 Experiments and Main Results

4.1 Experimental Setup

Dataset Description. The dataset we use for our classification task includes Breast Ultrasound Images (BUSI) collected from 600 female patients aged between 25 and 75 years old [Al-Dhabyani et al., 2020]. BUSI consists of 780 grayscale images, each of which containing 500*500 pixels, and is categorized into three classes: normal, benign, and malignant. One of the most intriguing features of BUSI is the exceptionally small size. With a 80%/20% training/testing split, only about 600 images can be used to train a model for three-class classification. Such a characteristic not only vividly simulates the real-world scenario of medical data sparsity but also poses a great challenge to the efficient training of machine learning models.

Masking for Medical Images. Despite the significant classification improvement brought by image masking, it is not always possible to apply simple and effective masking to general medical images with more complicated and fine-grained details. The inherent noisy nature of medical images is always a critical problem that machine learning models have to face and solve. Therefore, for experiments with deep learning models, we exclusively focus on the classification scenario without masking, and endeavor to utilize the robust vision modelling capability of state-of-the-art deep learning models to overcome the intrinsic noise in medical images.

4.2 Main Results

Overall Results. Table 2 compares the best results across various statistical learning and deep learning methods. Specifically, for the original task without masking, the deep learning paradigm - Transfer Learning - significantly outperforms all the other statistical learning methods, demonstrating prominently higher robustness and capability in noisy visual modelling.

The Effect of Masking. Table 2 also shows that after image masking is applied, all of the statistical learning methods show significant improvement except for Logistic Regression, demonstrating the efficacy of masking in reducing background noise for grayscale medical images like BUSI. In later analysis, we also investigate the reason why masking fails to improve the classification performance of Logistic Regression, and further develop a two-stage Logistic Regression method that utilizes both the original images and masked images and achieves significantly improved performance.

Accuracy(%)	**Naïve Bayes**	**Logistic Regression**	**SVM**	**Random Forest**	**Transfer Learning**
Original	52	69	67	70	**83.4**
With Masking	79	67	**84**	80	/

Table 2: Comparison of best accuracy results across statistical learning methods and deep learning methods. The best result of Transfer Learning uses `swin_base` as the base model, along with data augmentation techniques and the image resolution set to 224.

Base Models of Various Parameter Sizes. Table 3 compares the results of Transfer Learning using 4 base models of different parameter sizes and FLOPs, and presents several interesting observations. First, ViTs (Swin Transformer) consistently outperform CNNs (EfficientNet) by a large margin regardless of the model size. For instance, `Efficient_b5` and `swin_tiny` have similar parameter size of about $30M$, but the performance of the latter outperforms the former by 11.5%. Second, the CNNs show a tendency of overfitting and drop in test accuracy when model size increases, while the test accuracy of ViTs continues to increase with the scaling of the base model. In later analysis we show that these two critical observations can both be explained by the inherent and fundamental architectural differences between CNNs and ViTs.

5 Ablation Studies and Analysis

According to the main results in Table 2, training statistical models on masked data effectively improves the accuracy of prediction for all the statistical learning methods except for Logistic

	Image Resolution	Data Augmentation	#Parameters	FLOPs	Accuracy(%)
Efficient_b5	224	✓	30M	9.9G	**69.4**
Efficient_b7			66M	37.0G	67.5
Swin_tiny			29M	4.5G	80.9
Swin_base			88M	15.4G	**83.4**

Table 3: Results of Transfer Learning with 4 base models of different parameter sizes and floating-point operation counts (FLOPs). `swin_base`, ViT with the largest parameter size, achieves the best performance. All the results displayed here are obtained with data augmentation techniques and image resolution set to 224.

Regression. Therefore, we conduct an error analysis and extend the Logistic Regression into a two-stage model. Then, ablation studies for deep learning methods are introduced.

5.1 Error Analysis

Using grid search strategy, we optimizes the classifiers with tuned hyperparameters on masked images (Table 4). These hyperparameters are instrumental in enhancing the model's performance.

Model	Hyperparameters After Tuning
Naïve Bayes	`var_smoothing = 0.0032`
Logistic Regression	`C = 0.025, penalty = "l2"`
Support Vector Machine	`C = 22.5, gamma = 0.003, kernel = "rbf"`
Random Forest	`n_estimators = 70, max_depth = 9, minsamples_split = 4`

Table 4: Values of the hyperparameters tuned in statistical models.

Based on the provided results from the parameter tuning using grid search, the following analysis can be made.

Class	Precision	Recall	F1-score	Support
Benign	0.76	0.90	0.83	87
Malignant	0.71	0.48	0.57	46
Normal	1.00	1.00	1.00	27
Accuracy: 0.79		Total Support: 160		

Table 5: **Naïve Bayes:** Classification Report (Masked Images)

Class	Precision	Recall	F1-score	Support
Benign	0.72	0.85	0.78	84
Malignant	0.56	0.35	0.43	43
Normal	0.97	1.00	0.98	27
Accuracy: 0.66		Total Support: 160		

Table 6: **One-stage Logistic Regression:** Classification Report (Masked Images)

The confusion matrices of different models indicate a strong classification performance and show the same pattern: All 27 normal cases (without tumor) are accurately classified; The prediction performance on normal cases is the best, then the benign cases, and the worst is the performance on

Class	Precision	Recall	F1-score	Support
Benign	0.85	0.89	0.87	87
Malignant	0.76	0.70	0.73	46
Normal	1.00	1.00	1.00	27
Accuracy: 0.85		Total Support: 160		

Table 7: **Support Vector Machine:** Classification Report (Masked Images)

Class	Precision	Recall	F1-score	Support
Benign	0.83	0.93	0.88	87
Malignant	0.91	0.63	0.74	46
Normal	0.90	1.00	0.95	27
Accuracy: 0.86		Total Support: 160		

Table 8: **Random Forest:** Classification Report (Masked Images)

malignant cases. Therefore, there exists a noticeable challenge in classifying malignant cases, where fewer than three-fourth of the total 46 were correctly identified, indicating some misclassifications.

In terms of malignant cases, the precision is the higher than the recall, indicating that when the model predicts a tumor as malignant, it is more likely to be correct (higher precision), but the model may not be capturing all the malignant tumors present in the BUSI dataset (lower recall). The high F1-scores for benign and normal classes reflect the balanced nature of precision and recall for these categories.

The random forest model has the highest overall accuracy of the models, standing at 0.86, which is a strong indicator of its effectiveness in classifying the cases. The total support of 160 shows a substantial dataset size, giving credence to the reliability of these metrics.

In conclusion, the model, with its optimized hyperparameters, performed well in distinguishing between benign and normal cases but showed some limitations in accurately identifying all malignant cases, as reflected in the lower recall rate for this category. Further refinement in the model could aim at improving its sensitivity towards malignant cases without compromising the high accuracy achieved in other classes.

Our study further delves into the impact of feature selection on the classification accuracy of the models when applied to the BUSI dataset. The rationale behind this analysis is twofold. Firstly, the intrinsic nature of medical images, characterized by their grayscale properties and relatively simpler structures, demands a thoughtful selection of features. Secondly, considering the task's limitation to three classes, there's a need to determine whether a reduced set of features would suffice for effective discrimination, unlike more complex image classification tasks that require a larger feature set.

In our analysis, presented in Table 9 and Table 10, we observe an intriguing trend: Reducing the number of features does not necessarily lead to a decrease in classification accuracy. Contrarily, a carefully curated subset of features often results in comparable, if not better, performance. This phenomenon can be attributed to the reduction of noise and irrelevant information, allowing the classifiers to focus on the most informative features. While the training accuracy is relatively stable across different feature sets, the testing accuracy for a reduced feature set (masked pictures) is consistently higher or at par with that using the full feature set (raw pictures). This indicates a reduction in overfitting, a common challenge in machine learning models applied to medical imaging.

Confusion Matrix	Benign	Malignant	Normal
Benign	81	3	3
Malignant	17	29	0
Normal	0	0	27

Table 9: **Random Forest:** Confusion Matrix (Masked Images)

Confusion Matrix	Benign	Malignant	Normal
Benign	82	2	0
Malignant	24	19	0
Normal	20	1	8

Table 10: **Random Forest:** Confusion Matrix (Raw Images)

In conclusion, our investigation into feature selection for the BUSI dataset suggests that a carefully chosen subset of features can lead to efficient model performance. This is especially significant given the simpler nature of the medical images and the limited classification categories involved. Overcomplicating the model with an excessive number of features may lead to overfitting and diminished performance, underscoring the importance of precise feature selection in machine learning applications in medical imaging.

5.2 Two-stage Logistic Regression

In contrast to other statistical learning models, the incorporation of masked images has consistently enhanced the accuracy of image classification. However, logistic regression exhibits an opposite trend, experiencing a decline in accuracy from 0.79 to 0.76 upon the introduction of masked images. The key to explaining this anomaly lies in the two tables below, which present the confusion matrices for the use of original images and masked images, respectively. It is evident that the introduction of a mask image results in a rapid increase in the model's accuracy in distinguishing between tumor(benign and malignant) and non-tumor(normal) cases, with the accuracy for the 'normal' category reaching 0.97, and a recall rate of 1.00. Nevertheless, models trained with mask images are more prone to confusion between benign and malignant tumors, ultimately leading to an overall decrease in accuracy.

Confusion Matrix	Benign	Malignant	Normal
Benign	70	11	3
Malignant	12	30	1
Normal	14	7	8

Table 11: **One-stage Logistic Regression:** Confusion Matrix (Raw Images)

Confusion Matrix	Benign	Malignant	Normal
Benign	71	12	1
Malignant	28	15	0
Normal	0	0	29

Table 12: **One-stage Logistic Regression:** Confusion Matrix (Masked Images)

Consequently, we decide to train a two-stage logistic regression model to maximize recognition accuracy. Feature extraction and model preprocessing are performed similarly to other statistical learning models. In the first stage, we fit a binary logistic regression model (tumor vs. non-tumor) using all data from the training set. Then we fit another binary logistic regression model (benign vs. malignant) using only the tumor data from the training set as the second stage model. To obtain the final predictions, we initially use the model from the first stage to predict whether a sample in the test set has a tumor. For samples predicted to have a tumor, we then use the model from the second stage to predict whether the tumor is benign or malignant. The confusion matrix and classification report for the final predictions from both stages are presented below.

The accuracy increases from 0.66 (0.69) in the one-step model to 0.83, indicating a significant improvement in accuracy with our two-step logistic regression.

Class	Benign	Malignant	Normal
Benign	73	11	0
Malignant	16	27	0
Normal	0	0	29

Table 13: **Two-stage Logistic Regression:** Confusion Matrix

Class	Precision	Recall	F1-score	Support
Benign	0.82	0.87	0.84	84
Malignant	0.71	0.63	0.67	43
Normal	1.00	1.00	0.82	29
Accuracy: 0.83		Total Support: 156		

Table 14: **Two-stage Logistic Regression:** Classification Report

5.3 Ablation Studies of Transfer Learning Results

5.3.1 Ablation Study 1: Image Resolution

The first ablation study looks into the effect of image resolution on classification accuracy. The motivation of this study is straightforward: Original images from BUSI dataset are of 500*500 pixel size, which does not match the input dimension of most computer vision architectures. Therefore, in transfer learning, the size of input images must be rescaled to match the input dimension of the base model, and the effect of rescaling on final classification performance is worth investigating. Ideally, for the same base model, the higher the resolution of the rescaled image has, the better classification results will be produced, as more fine-grained features are included in the image and the representation quality of the model also increases. However, as can be seen in Table 15, for both CNN and ViT, larger input resolution uniformly leads to lower classification accuracy. Larger input resolution actually contributes to more severe overfitting, as is indicated by Figure 1 in which the training accuracies of the two input resolutions are close, while the testing accuracy of the larger resolution (yellow) is consistently and significantly lower than that of the lower resolution (blue).

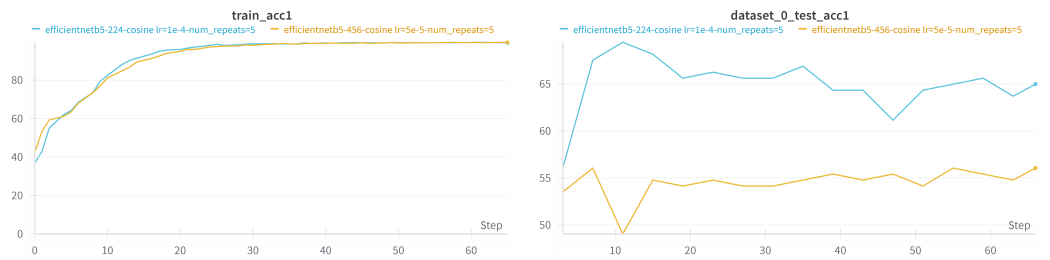
An investigation into the properties of BUSI images can generally explain the surge in overfitting. Firstly, the medical images are grayscale and naturally simpler than RGB images on which the base models are pretrained. Secondly, this image classification task contains only three classes, so it does not need a large number of latent features for effective discrimination, unlike typical image classification tasks where the model needs sufficient features in order to classify an image into usually hundreds of different classes. In summary, classification of the BUSI dataset does not need so much latent features or excessively complicated representations due to its relative simplicity, and an inappropriately high input resolution can thus easily lead to overfitting and performance drop.

	Image Resolution	Accuracy(%)		Image Resolution	Accuracy(%)
Efficient_b5	**224**	**69.4**	**Swin_base**	**224**	**83.4**
Efficient_b5	456	56	**Swin_base**	384	81.5

Table 15: Ablation Study 1: For both CNN and ViT, larger resolution leads to lower classification accuracy.

5.3.2 Ablation Study 2: Data Augmentation

Data augmentation is a technique of artificially increasing the training set by creating modified copies of a dataset based on existing data. The Augmented data is derived from the original data with some minor changes such as Geometric Transformation (e.g. Cropping and Rotation) and Photometric Transformation (e.g. Color Jittering and Gaussian blurring). Due to its efficacy in augmenting the



(a) Training Accuracy

(b) Test Accuracy

Figure 1: Ablation Study 1: The training and test accuracy of EfficientNet-b5 with different input image resolutions. The resolution of blue line is 224, and that of yellow line is 456. Larger input resolution leads to more severe overfitting.

training set and enriching the supervised information, data augmentation is widely employed in few-shot learning and other data-scarce scenarios [Wang et al., 2020]. In this ablation study, we employ Repeated Augmentation [Hoffer et al., 2020] which replicates instances of samples within the same batch with different data augmentations, and experiment to see how the number of augmentation repetition affects the final classification accuracy.

As can be seen in Table 16, a moderate number of repeated augmentation (e.g. 3) is helpful for improving few-shot image classification accuracy, but the improvement saturates as the number of augmentation repetition increases further (e.g. to 10). We explain these two observations step by step. Firstly, the improvement itself is easy to explain, as data augmentation not only increases the total size of the training set but also promotes generalization to a broader range of data by reducing the degree of overfitting. Secondly, when it comes to improvement saturation, we can consider the effect of data augmentation on transforming the training set distribution. Since data augmentation generates new data based on the original training set, the biases in the original dataset persist in the augmented data. Popular data augmentation methods like cropping, flipping, rotation and color-jitting cannot generate completely new data that perfectly bridge the gap between the training distribution and test distribution, and will somehow amplify the biases of the original training distribution. Such amplified biases eventually prevent the model from further generalizing to the test distribution after the number of repeated augmentation exceeds a threshold, resulting in the saturation of improvement brought by repeated augmentation.

Moreover, the quality of augmented data is usually not guaranteed. Data of poor quality can be generated due to inappropriate perturbation such as overly cropping or excessive color jitting, and they are not guaranteed to be relevant enough to the original training distribution. Such augmented data can instead introduce noise, bias, or inconsistency to the model, leading to inaccurate or misleading predictions. Specifically, for the grayscale medical images in BUSI dataset, data augmentation is especially prone to introduce extra harmful noise, as the original images already have a relatively high amount of inherent noise.

	#Repeated Augmentation	**Accuracy(%)**
Swin_tiny	0	78.9
Swin_tiny	**3**	**80.9**
Swin_tiny	10	80.3

Table 16: Ablation Study 2: The classification accuracy of transfer learning (using `swin_tiny` as the base model) first increases and then saturates as the number of repeated augmentation increases.

5.3.3 Ablation Study 3: Architectural Differences

As is shown in Table 17, for CNNs and ViTs of similar model size, ViTs consistently and significantly outperform CNNs in classification accuracy. Such an observation actually reveals an inherent and fundamental architectural difference between CNN and ViT - the quantity of Inductive Biases (IB).

Convolutional Neural Networks are built on the premise that images exhibit translation invariance. As a result, convolutions using identical filters are applied throughout the image, aiming to detect similar local patterns in different regions. Additionally, the architecture of a CNN incorporates the idea of spatial proximity: pixels in close proximity bear more relation than those far apart. These local patterns are progressively downsampled into more extensive patterns leading up to the final classification prediction. These characteristics form the inductive biases inherent in CNNs [Cohen and Shashua, 2016].

In Contrast, a Vision Transformer lacks inherent knowledge about the proximity or distance between two pixels, relying entirely on the limited learning signals from the classification task to grasp this information. This poses a significant challenge when working with small datasets, as this spatial understanding is vital for generalization to new, unseen test datasets. However, with **sufficiently large datasets or effective pre-training**, a Transformer is capable of acquiring this spatial knowledge without the need for built-in inductive biases, thus offering more flexibility compared to a CNN. Particularly, handling long-distance relations between local patterns, which can be problematic in CNNs as the prior spatial knowledge from the inductive biases obstructs the model from learning such inevident relations between local patterns in an image, is more feasible in Transformers where all patches have the same distance of one [Lippe, 2023].

	#Parameters	Accuracy		#Parameters	Accuracy
Efficient_b5	30M	69.4	**Efficient_b7**	66M	67.5
Swin_tiny	29M	**80.9**	**Swin_base**	88M	**83.4**

Table 17: Ablation Study 3: For CNNs and ViTs of similar model size, ViTs consistently and significantly outperform CNNs in classification accuracy, demonstrating inherent differences between these two model architectures.

6 Conclusion and Future Work

6.1 Conclusion

Our research presents a comprehensive evaluation on the task of few-shot medical image classification, involving both traditional statistical learning methods and state-of-the-art deep learning paradigms. Specifically, we conclude our research with the following statements derived from holistic experiments and in-depth analyses.

1. Deep Learning Methods (especially Transfer Learning using ViT as the base model) demonstrate significantly higher capability and robustness in few-shot noisy medical image classification, compared with traditional statistical learning methods.

2. Among various deep learning architectures, ViTs are consistently and significantly more capable than CNNs, probably due to fewer inductive biases and correspondingly stronger long-range modelling capability. They show a prominently lower degree of overfitting than CNNs.

3. Image masking proves to be a crucial data-preprocessing technique for grayscale medical images such as BUSI. It effectively removes the inherent noise of medical images and thus promotes the feature extraction and image modelling capabilities of traditional statistical learning methods. However, it is not always possible to apply simple masking techniques to more general medical images which include more complicated and fine-grained features, so enhancing the image modelling robustness of statistical learning methods in a noisy background remains a pivotal research direction.

4. Traditional statistical learning methods are more interpretable than deep learning ones, especially in such tasks as there are only 3 classes, making it possible to conduct detailed statistical analyses. In contrast, though extensive ablation studies have been conducted on the deep learning method, much of the observation can only be intuitively interpreted based on empirical experiences of past practitioners.

6.2 Future Work

1. **Explore Hybrid Policies**: Future research can explore the potential synergy between a variety of traditional machine learning techniques, such as Random Forest, Naive Bayes, Support Vector Machines (SVM), and Logistic Regression, in conjunction with deep learning models. By integrating these diverse statistical learning methods with advanced deep learning architectures, a hybrid approach could be developed. Such an approach would not only capitalize on the unique strengths of both traditional and contemporary methodologies but also potentially enhance overall classification accuracy and robustness in complex tasks like medical image analysis.
2. **Advanced Feature Engineering**: There is still room for more sophisticated feature engineering, particularly in extracting and utilizing more complex characteristics of medical images that may be overlooked by current statistical learning methods.
3. **Improve Interpretability for Safer Medical Use**: Given the critical nature of medical diagnostics, future work can also focus on improving the interpretability and explainability of the models, ensuring that medical professionals can fully trust the AI-assisted diagnostic process.
4. **Task Expansion**: There are two dimensions for task expansion. Firstly, the research on few-shot medical image classification can extend to datasets with a larger number of classes and correspondingly fewer training images for each class. Reducing the size of the finetuning dataset to 50-shot, 10-shot or even 5-shot may further elicit the few-shot learning ability of state-of-the-art machine learning models. Secondly, it is also beneficial to conduct few-shot learning research on more complicated medical image tasks such as RGB images with fine-grained features, so that the few-shot learning recipes developed in our research might be generalized further.
5. **Investigate the Effect of Domain Discrepancy in Transfer Learning.** Further research can also experiment with base models that are pretrained on large medical image datasets instead of daily image datasets like ImageNet, to see how the domain discrepancy between the pretraining data and the transfer target affects final classification performance.

References

Kimberly D Miller, Leticia Nogueira, Theresa Devasia, Angela B Mariotto, K Robin Yabroff, Ahmedin Jemal, Joan Kramer, and Rebecca L Siegel. Cancer treatment and survivorship statistics, 2022. *CA: a cancer journal for clinicians*, 72(5):409–436, 2022.

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022.

Hein Tun Zaw, Noppadol Maneerat, and Khin Yadanar Win. Brain tumor detection based on naïve bayes classification. In *2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*, pages 1–4, 2019. doi: 10.1109/ICEAST.2019.8802562.

P Ramesh Kumar and A Vijaya. Naïve bayes machine learning model for image classification to assess the level of deformation of thin components. *Materials Today: Proceedings*, 68:2265–2274, 2022. ISSN 2214-7853. doi: https://doi.org/10.1016/j.matpr.2022.08.489. 4th International Conference on Advances in Mechanical Engineering.

Paidipati Dinesh and P Kalyanasundaram. Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: Svm, knn, logistic regression, random forest, and decision tree to measure accuracy. *ECS Transactions*, 107(1):12681, 2022.

Mingmin Chi, Rui Feng, and Lorenzo Bruzzone. Classification of hyperspectral remote-sensing data with primal svm for small-sized training dataset problem. *Advances in space research*, 41(11): 1793–1799, 2008.

Reem Alrais and Nazar Elfadil. Support vector machine (svm) for medical image classification of tumorous. *Int. J. Comput. Sci. Mob. Comput*, 9(6):37–45, 2020.

Hicham Riri, A. Elmoutaouakkil, A. Beni-hssane, and Farid Bourezgui. Classification and recognition of dental images using a decisional tree. *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*, pages 390–393, 2016. doi: 10.1109/CGIV.2016.82.

Helen Kaganov, A. Ades, and David S. Fraser. Preoperative magnetic resonance imaging diagnostic features of uterine leiomyosarcomas: A systematic review. *International Journal of Technology Assessment in Health Care*, 34:172 – 179, 2018. doi: 10.1017/S0266462318000168.

A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. Regression forests for efficient anatomy detection and localization in ct studies. pages 106–117, 2010. doi: 10.1007/978-3-642-18421-5_11.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Phillip Lippe. UvA Deep Learning Tutorials. `https://uvadlc-notebooks.readthedocs.io/en/latest/`, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. URL `http://arxiv.org/abs/1409.4842`.

Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: Feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9225–9234, June 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714, 2010.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020.

Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. *arXiv preprint arXiv:1605.06743*, 2016.