A Comparative Evaluation of Accuracy-Robustness Trade-offs in Modern Image Data Augmentation Methods

Bachelor Thesis

Bachelor of Science in Computer Science

Juncheng Gong

April 13, 2026

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Francesco Di Salvo, M.Sc.

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

This thesis studies how different data augmentation methods affect clean performance, corruption robustness, and calibration in image classification. Augmentation is still often judged mainly by clean test accuracy, but that perspective becomes too narrow once the data distribution changes. Methods that perform similarly on clean data can behave quite differently under corruption. To study this more systematically, the thesis uses a three-stage evaluation pipeline. Stage 1 selects stable learning rates for each dataset–architecture pair. Stage 2 compares augmentation methods on clean in-distribution data. Stage 3 evaluates robustness under common corruptions using CIFAR-C and MedMNIST-C.

The experiments cover two natural-image datasets, CIFAR-10 and CIFAR-100, and two medical-image datasets, DermaMNIST and PathMNIST. ResNet-18 serves as the main benchmark architecture, while ViT-B is included as an additional architecture to test whether the main patterns remain similar across model families. The results show that clean performance alone is not sufficient to judge robustness. They also show that robustness and calibration do not always improve together, so confidence reliability must be considered alongside accuracy. More broadly, augmentation effects vary across domains and architectures. Overall, the findings suggest that augmentation should be evaluated with the target setting in mind, rather than chosen based on clean-data performance alone.

Contents

List of Figures

List of Tables

List of Acronyms

CNN	Convolutional Neural Network
ECE	Expected Calibration Error
mCA	Mean Corruption Accuracy
MCE	Maximum Calibration Error
mECE	Mean Expected Calibration Error
SGD	Stochastic Gradient Descent

Notation

This section lists only the notation that appears repeatedly in the thesis. It is not intended to be a full deep-learning notation catalogue.

Data and Labels

$\mathcal{D}$	Dataset used for training or evaluation
$x^{(i)}$	Input image of the i-th sample
$y^{(i)}$	Ground-truth label of the i-th sample
p_{train}	Training data distribution
p_{corr}	Corrupted test distribution used in robustness evaluation

Model and Training

$f_\theta(x)$	Model with parameters θ applied to input x
$\hat{y}$	Predicted class label
$p(y \mid x)$	Predicted class probabilities for input x
$\mathcal{L}$	Training loss
θ	Trainable model parameters
η	Learning rate

Evaluation

Acc	Classification accuracy on clean in-distribution data
mCA	Mean corruption accuracy across corruption types and severities
ECE	Expected calibration error
MCE	Maximum calibration error
mECE	Mean expected calibration error under corruption

1 Introduction

1.1 Motivation

Deep learning models have achieved strong performance on standard image classification benchmarks when they are evaluated on clean and carefully curated datasets [1, 2]. In real applications, however, the input data is usually less controlled than in benchmark settings. Images may be affected by noise, blur, changes in illumination, compression artifacts, or sensor-related distortions. These effects are commonly discussed under the term *distribution shift* and can lead to clear drops in model performance, even when the semantic content of the image is still easy for a human observer to recognize [3].

Data augmentation is one of the standard tools used to improve generalization in deep learning [4]. By modifying training samples during learning, augmentation aims to reduce overfitting and expose the model to a wider range of plausible input variations. Over time, many augmentation strategies have been proposed, ranging from simple geometric or photometric transformations to policy-based methods such as AutoAugment and RandAugment [5, 6], mixing-based methods such as Mixup and CutMix [7, 8], and more recent approaches that target robustness more directly through style variation or synthetic sample generation [9–11].

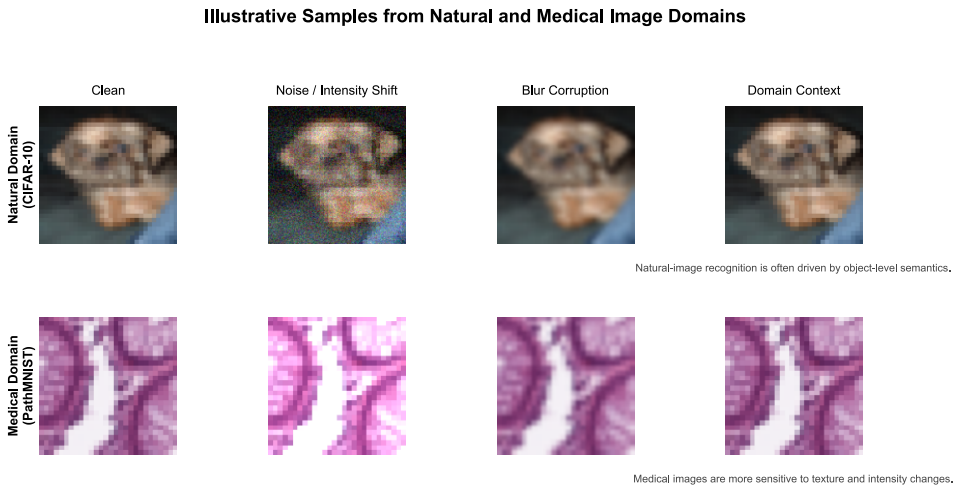


Figure 1: Illustration of the distribution-shift problem. A model trained on clean data may perform confidently on standard test samples, yet fail on corrupted inputs that remain semantically recognizable to humans. This highlights the need to evaluate robustness beyond conventional clean-test accuracy.

Although augmentation is widely used, it is still often judged mainly through clean test accuracy. This leaves an important gap. A method that improves performance on clean data does not necessarily remain strong once the input is corrupted or shifted. In addition, robustness is often discussed only in terms of accuracy, while calibration receives less attention. This matters in settings such as medical image

analysis, where it is important not only that a model predicts correctly, but also that its confidence scores remain trustworthy [12].

A further difficulty is that conclusions drawn from natural-image benchmarks do not necessarily transfer to other domains. Medical image datasets differ from natural-image datasets in several ways, including visual statistics, class imbalance, semantic variability, and task-specific decision boundaries [13, 14]. Because of these differences, augmentation methods that work well on natural images may behave differently on medical imaging tasks [15]. The same is true for model architecture. Convolutional neural networks and transformer-based vision models rely on different inductive biases and training dynamics, so the same augmentation strategy may not help both architectures in the same way [16–19].

For these reasons, this thesis studies traditional and modern data augmentation methods with a joint focus on clean performance, calibration, and robustness under distribution shift.

1.2 Problem Statement

Modern image classification models can achieve strong results on clean, in-distribution test data [16, 17]. However, good performance in this setting does not automatically imply robustness in more realistic deployment scenarios. In practice, inputs are often affected by common corruptions or domain shifts. Even relatively small perturbations can therefore lead to noticeable drops in performance.

Data augmentation is widely used to improve generalization, but its effect is still often evaluated mainly through clean accuracy. As a result, several questions remain open. First, it is not always clear whether methods that improve clean-data performance also improve robustness under corrupted conditions. Second, robustness should not be measured only by accuracy. A model may remain reasonably accurate under corruption while still producing poorly calibrated confidence estimates, which reduces its usefulness in settings where reliability matters. Third, augmentation effects may vary across domains and model architectures, so conclusions drawn from a single benchmark or a single model family can easily be too narrow.

The main problem addressed in this thesis is therefore the lack of a unified evaluation setting for comparing augmentation methods beyond clean accuracy alone. In this work, augmentation methods are assessed jointly with respect to clean performance, calibration, corruption robustness, domain dependence, and architectural stability. To do so, the thesis applies a single three-stage evaluation pipeline to both natural-image and medical-image benchmarks, using ResNet-18 as the main benchmark backbone and ViT-B as an additional architecture.

1.3 Research Questions

Based on the motivation and problem statement outlined above, this thesis addresses the following research questions:

- **RQ1:** How do different data augmentation methods affect classification performance on clean test data?
- **RQ2:** How do different data augmentation methods affect calibration on clean test data?
- **RQ3:** To what extent do different data augmentation methods improve robustness under common corruptions on natural-image and medical-image benchmarks?
- **RQ4:** Are the robustness effects of data augmentation domain-dependent, that is, do the observed trends differ between natural-image and medical-image datasets?
- **RQ5:** Are the observed effects of data augmentation stable across model architectures, or do they depend on whether the underlying model is a convolutional neural network or a transformer-based vision model?

These research questions are examined through one unified three-stage evaluation pipeline. The main experiments are conducted with ResNet-18 in order to establish one consistent benchmark setting. Additional experiments with ViT-B are included to check which trends remain similar across architectures under a fixed training budget.

1.4 Contributions

The main contributions of this thesis are as follows:

1. **A unified three-stage evaluation pipeline.** The study uses a three-stage pipeline that separates learning-rate selection, clean-data comparison, and robustness evaluation under corruption. This keeps the evaluation systematic and makes the later comparisons easier to interpret.
2. **A comparative study of traditional and modern augmentation methods.** The study compares several augmentation families within one experimental framework, including traditional transformations, policy-based methods, mixing-based methods, and more recent robustness-oriented approaches.
3. **The inclusion of calibration in the evaluation.** In addition to clean performance and corruption robustness, the thesis also evaluates calibration through metrics such as Expected Calibration Error (ECE), mean Expected Calibration Error (mECE), and Maximum Calibration Error (MCE). This makes it possible to check whether stronger robustness is actually accompanied by more reliable confidence estimates and whether average calibration and worst-case calibration tell the same story.

4. **A cross-domain evaluation on natural and medical image datasets.** By considering both natural-image and medical-image benchmarks, the thesis examines whether augmentation effects remain similar across domains or shift with the properties of the data.

5. **An additional cross-architecture comparison.** While ResNet-18 serves as the main benchmark architecture, ViT-B is included as an additional architecture in order to examine which augmentation trends remain similar across different model families and which do not.

6. **Practical observations for robustness-oriented evaluation.** The results show that augmentation choice should be interpreted in relation to the target domain, the evaluation goal, and the model architecture instead of being treated as a one-time benchmark ranking.

1.5 Thesis Structure

The remainder of this thesis is organized as follows. Chapter 2 reviews the relevant background and related work. Chapter 3 describes the methodology and experimental setup. Chapter 4 presents the main experimental results. Chapter 5 discusses these findings and relates them to the broader literature. Finally, Chapter 6 concludes the thesis and outlines possible directions for future work.

2 Background and Related Work

This chapter introduces the background needed for the rest of the thesis. It reviews the augmentation families considered in the experiments, outlines corruption robustness under distribution shift, summarizes the benchmark suites used in this work, and explains the calibration metrics used later in the evaluation. It also discusses why augmentation effects may vary across model architectures, which becomes important in the later comparison between ResNet-18 and ViT-B.

2.1 Data Augmentation for Image Classification

Data augmentation is a common regularization technique in image classification. It increases the diversity of the training data by applying label-preserving transformations to input images [4]. By exposing the model to a broader range of visual patterns during training, augmentation aims to reduce overfitting and improve generalization to unseen data. In practice, augmentation is often one of the first tools used when the available training data is limited or when the model is expected to handle more variation at test time.

In this thesis, the evaluated augmentation methods are grouped into four families: traditional augmentations, policy-based augmentations, mixing methods, and style-

or generative methods. This grouping is useful because the methods differ not only in how they transform training data, but also in the kind of invariance they encourage. Some mainly introduce small local variations, while others change the training distribution more strongly and may therefore affect calibration and robustness in different ways.

2.1.1 Traditional and Geometric Augmentations

Traditional augmentation methods are among the earliest and most widely used approaches in image classification. They are usually hand-crafted, computationally inexpensive, and intended to introduce simple invariances into the model. In this thesis, the traditional augmentation settings are represented by the standard baseline pipeline and by a setting that combines random rotation with *Random Erasing*.

- **Standard Baseline (Random Crop + Random Horizontal Flip):** A common baseline in image classification is to apply random cropping and horizontal flipping during training [1]. For small natural-image datasets such as CIFAR, this helps the model become less sensitive to small translations and mirror variations. In this thesis, this setting serves as the main reference point for comparing the other augmentation methods.
- **Rotation + Random Erasing:** One of the evaluated traditional settings combines small random image rotations with *Random Erasing*. Random Erasing, introduced by [20], randomly selects a rectangular image region and replaces it with random values or a constant fill. This encourages the model to learn from partially occluded inputs and can improve robustness to local information loss. The additional rotation transform is meant to reduce sensitivity to moderate orientation changes. Compared with more complex augmentation families, this remains a fairly classical hand-crafted strategy.

Traditional augmentations are simple and efficient, and they still provide a strong baseline in many image-classification settings. At the same time, the transformation space they cover is relatively limited. In this thesis, they are therefore useful not only as practical methods, but also as a reference point for judging whether more advanced augmentation families actually add something beyond standard regularization.

2.1.2 Policy-based Augmentations

Policy-based augmentation methods aim to automate the selection of augmentation operations instead of relying only on manual design. The main idea is that a useful augmentation strategy does not need to be chosen entirely by hand, but can be searched for or simplified through a more systematic design process.

- **AutoAugment:** [5] formulates augmentation-policy selection as a search problem. A controller searches for sequences of image transformations and associated magnitudes that improve validation performance. AutoAugment can produce strong results, but its search procedure is computationally expensive.
- **RandAugment:** [6] simplifies the AutoAugment idea by removing the search controller and reducing the policy space to a small number of global parameters. In practice, RandAugment often remains competitive while being much easier to apply.

Policy-based methods are relevant to this thesis because they represent a practical alternative to hand-crafted augmentation design and often serve as strong baselines. They are also useful for comparison because they sit between simple classical augmentations and more aggressive modern methods.

2.1.3 Image Mixing Methods

Mixing methods generate new training examples by combining two samples and their labels. Compared with ordinary geometric transforms, they modify the training distribution more strongly and often encourage smoother decision boundaries. Because of this, their effect is not limited to standard generalization, but may also influence robustness and calibration.

- **Mixup:** [7] creates a virtual example $(\tilde{x}, \tilde{y})$ by linearly interpolating two samples (x_i, y_i) and (x_j, y_j):

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda)y_j \tag{1}$$

  where $\lambda \in [0, 1]$ is sampled from a Beta distribution. Mixup is often associated with smoother decision boundaries and, in some settings, improved calibration.
- **CutMix:** [8] replaces a rectangular patch of one image with a patch from another image and mixes the labels according to the patch area. In contrast to Mixup, CutMix preserves more local image structure while still encouraging the model to rely on broader evidence.

Mixing methods are particularly relevant here because they often improve clean performance, but their effect on corruption robustness and calibration is not always consistent across datasets and architectures. This makes them especially interesting for the joint evaluation carried out in this thesis.

2.1.4 Style and Generative Augmentations

More recent augmentation methods try to create broader appearance variation by using augmentation chains, style transfer, or generative models. Compared with traditional methods, these approaches aim to expose the model to richer distributional changes during training.

- **AugMix:** [9] combines several stochastic augmentation chains and mixes them with the original image. It is designed particularly with corruption robustness in mind and is often paired with a consistency-based training objective.
- **StyleAug:** [10] uses style transfer to vary image texture while preserving semantic structure. The goal is to reduce texture bias and encourage the model to rely more on shape-related cues, which is especially relevant in light of earlier work on texture-biased image recognition models [21].
- **DiffuseMix:** [11] uses diffusion-based generation together with mixing ideas to create more diverse but semantically related training samples. It represents a more recent generative direction in augmentation research.

These methods are especially relevant to this thesis because they go beyond simple invariance-based augmentation and expose the model to broader appearance variation during training. At the same time, this broader variation may also change how the model behaves under corruption or how well calibrated its predictions remain. This makes them particularly important for the later empirical analysis.

Table 1: Overview of the data augmentation methods considered in this thesis.

Category	**Method**	**Key Mechanism**	**Typical Motivation**
Traditional	Random Crop/Flip	Geometric transformation	Invariance
Traditional	Rotation + Random Erasing	Rotation and local occlusion	Robustness
Policy-based	AutoAugment	Policy search	Accuracy
Policy-based	RandAugment	Randomized policy selection	Efficiency
Mixing	Mixup	Linear interpolation	Calibration
Mixing	CutMix	Patch replacement	Generalization
Generative	AugMix	Multi-chain mixing	Robustness
Generative	StyleAug	Style transfer	Domain variation
Generative	DiffuseMix	Diffusion-based generation	Sample diversity

2.2 Robustness under Distribution Shift

Standard supervised learning usually assumes that training and test data are sampled from the same underlying distribution. In practice, however, this assumption is often violated. When the test distribution differs from the training distribution, the model is evaluated under *distribution shift*.

In this thesis, the main focus is on corruption robustness, that is, robustness to non-adversarial perturbations such as noise, blur, compression artifacts, and intensity changes [3]. This setting differs from adversarial robustness, which studies worst-case perturbations designed specifically to fool the model. Corruption robustness instead focuses on more realistic average-case degradations that may occur in practical applications.

This distinction matters because data augmentation is often motivated as a way to improve generalization, but improvements on clean data do not automatically imply stronger robustness under shifted conditions. A method may improve standard test accuracy while still remaining sensitive to corruption, which is one of the main reasons why clean performance alone is not sufficient in this thesis.

2.3 Benchmark Suites

To evaluate robustness in a controlled and comparable way, standardized corruption benchmarks are needed. This thesis uses benchmark suites for both natural-image and medical-image domains.

2.3.1 CIFAR-C

CIFAR-10-C and CIFAR-100-C [3] extend the original CIFAR datasets [1] by applying common corruptions to the clean test images. In this thesis, robustness is evaluated over 19 corruption types, covering categories such as noise, blur, weather effects, and digital artifacts.

The benchmark is useful because it allows performance to be measured under controlled corruption strength through multiple severity levels. This makes it possible to analyze not only average robustness, but also how rapidly performance degrades as corruption severity increases. Since CIFAR-C has become one of the most common corruption benchmarks in the literature, it also makes the results of this thesis easier to interpret relative to earlier work.

2.3.2 MedMNIST-C

While CIFAR-C covers the natural-image setting, medical imaging involves different visual characteristics and task constraints. For this reason, this thesis also uses MedMNIST-C [22], which is built on the MedMNIST v2 benchmark [23].

In this thesis, the medical robustness evaluation focuses on a subset of eight corruption types: Defocus Blur, Motion Blur, JPEG Compression, Pixelate, Brightness Up, Brightness Down, Contrast Up, and Contrast Down. This subset is used because it provides a comparable corruption protocol across the selected medical datasets and reflects realistic image degradations that may appear in medical imaging pipelines.

Using both CIFAR-C and MedMNIST-C is important in this thesis because one of the main questions is whether augmentation effects transfer across domains or

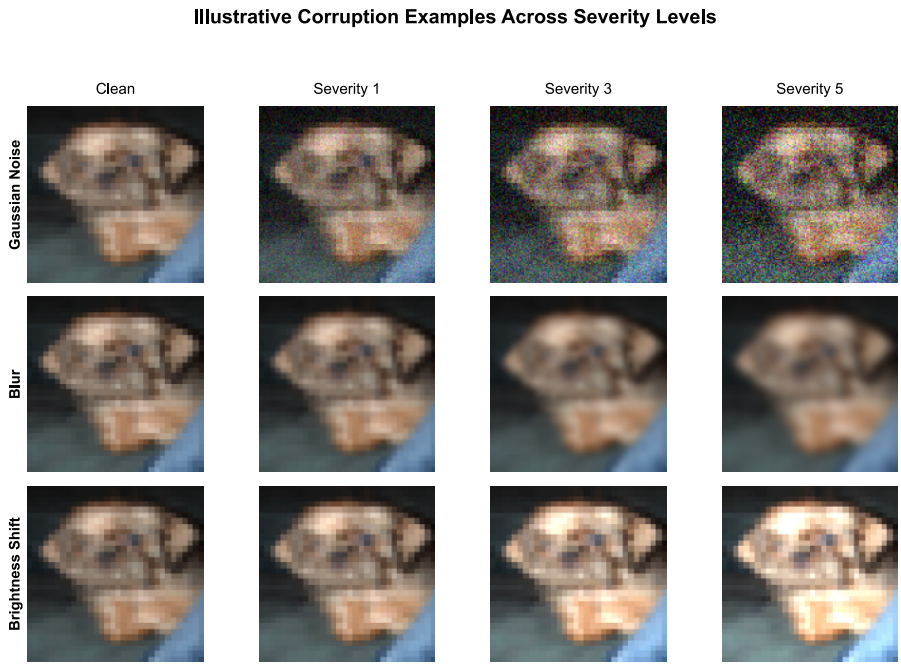


Figure 2: Illustrative examples of selected CIFAR-C corruption types across increasing severity levels.

remain domain-dependent. The medical benchmark is therefore not included only as an additional test set, but as an important part of the overall research design.

2.4 Calibration Metrics and Predictive Reliability

In many applications, especially safety-relevant ones, accuracy alone is not sufficient. A model should also provide confidence estimates that reflect how likely its predictions are to be correct. This property is usually referred to as calibration. A model is well calibrated if predictions made with confidence p are correct approximately p percent of the time [12, 24].

Calibration is particularly relevant in this thesis because one of the main goals is to move beyond the usual evaluation based only on clean accuracy. A model that predicts accurately but produces unreliable confidence scores may still be difficult to trust in practical settings.

2.4.1 Expected Calibration Error (ECE)

A common metric for calibration is Expected Calibration Error (ECE) [12]. ECE measures the weighted average gap between confidence and empirical accuracy across M confidence bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{2}$$

where N is the number of samples, B_m is the set of predictions assigned to bin m, $\text{acc}(B_m)$ is the average accuracy within that bin, and $\text{conf}(B_m)$ is the average predicted confidence.

ECE is used throughout this thesis as one of the main calibration metrics because it provides a compact summary of average miscalibration.

2.4.2 Maximum Calibration Error (MCE)

While ECE summarizes average calibration quality, Maximum Calibration Error (MCE) measures the worst discrepancy across the confidence bins [12]:

$$\text{MCE} = \max_{m \in \{1,\ldots,M\}} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{3}$$

MCE is useful because it highlights whether a model has particularly badly calibrated regions, even when the average calibration error appears moderate. This is relevant for the later experiments because some augmentation methods may look competitive on average while still producing unstable confidence behavior in specific settings.

Under distribution shift, modern neural networks often become overconfident on incorrect predictions [12, 24]. For this reason, this thesis evaluates not only clean accuracy and corruption robustness, but also calibration under both clean and corrupted conditions.

2.5 Architecture Biases and Robustness Transfer

Another question relevant to this thesis is whether augmentation effects remain stable across model architectures. Convolutional neural networks and transformer-based vision models differ in their inductive biases, feature extraction behavior, and training dynamics [16, 17]. As a result, an augmentation method that works well for one model family may not behave in the same way for another.

This issue is especially relevant here because the main benchmark experiments of the thesis are built around ResNet-18, while ViT-B is used as an additional architecture. Vision Transformers have shown strong performance in image recognition [17], but their behavior is also known to depend more strongly on training recipe, augmentation, and regularization choices than is often the case for standard CNN baselines [25, 26]. In addition, prior work suggests that robustness comparisons between transformers and CNNs are not entirely straightforward and may depend on the evaluation setting and on how comparable the training setup actually is [18, 19, 27].

For this reason, a later cross-architecture comparison is meaningful in this thesis, even though the goal is not to provide a fully exhaustive CNN-versus-ViT study. Instead, the aim is to examine which augmentation trends remain similar once the backbone changes and which ones do not.

2.6 Summary and Research Gap

This chapter reviewed the main augmentation families considered in the thesis and introduced the background on corruption robustness, benchmark suites, calibration, and architecture-related differences. Overall, prior work suggests that augmentation can help clean generalization and sometimes also corruption robustness, but the picture is still mixed once domains, metrics, and architectures are considered together.

In particular, prior work often focuses on clean accuracy, studies only natural-image benchmarks, or considers one architecture family in isolation. This leaves several open questions about how augmentation affects robustness, calibration, domain transfer, and cross-architecture stability under a unified evaluation protocol. These gaps motivate the methodology presented in the next chapter.

3 Methodology

This chapter describes the experimental framework used in the thesis. The aim is to compare data augmentation methods not only in terms of clean predictive performance, but also with respect to calibration and robustness under distribution shift. To keep the comparison fair, the experiments follow a three-stage pipeline and use a consistent protocol across datasets, architectures, and augmentation settings.

3.1 Overview of the Three-Stage Evaluation Pipeline

The experiments follow a three-stage pipeline that separates hyperparameter selection, clean training, and robustness evaluation. This design helps avoid tuning decisions based on corrupted test results and makes the later robustness analysis easier to interpret.

The three stages are summarized below. The same structure is used throughout the thesis and later linked directly to the result analysis in Chapter 4.

1. **Stage 1: Learning-Rate Selection.** A grid search is performed with the baseline augmentation in order to select a stable learning rate for each dataset and architecture. Performance is monitored only on a validation split. The purpose of this stage is not to optimize a preliminary run as strongly as possible, but to choose a learning rate that leads to stable training and can later be reused in Stage 2 and Stage 3.

2. **Stage 2: Clean Training and Checkpoint Preparation.** After the learning rate has been selected, models are trained on the clean training set with the chosen augmentation method. No corrupted test data is used at this stage. The final checkpoint is saved at the end of training so that all methods can later be evaluated under the same protocol.

3. **Stage 3: Evaluation under Distribution Shift.** The saved checkpoints from Stage 2 are evaluated on the clean test set and on the corrupted benchmarks. This stage is used to compute clean reference metrics, corruption robustness metrics, and calibration metrics under distribution shift.

Both ResNet-18 and ViT-B are evaluated under this three-stage pipeline. ResNet-18 serves as the main benchmark backbone throughout the study, while ViT-B is included as an additional architecture that still follows the same overall protocol. This makes it possible to compare augmentation trends across model families without changing the evaluation logic itself.

3.2 Datasets and Corruption Protocols

The evaluation covers two domains: natural images and medical images. This design makes it possible to check whether augmentation effects remain similar across domains or whether they depend strongly on the type of data.

3.2.1 Natural Image Datasets

For the natural-image setting, this thesis uses CIFAR-10 and CIFAR-100 [1]. Both datasets contain 60,000 RGB images of size 32×32.

- **CIFAR-10**: 10 classes, with 50,000 training images and 10,000 test images.
- **CIFAR-100**: 100 classes, with the same overall split, but with a finer-grained label structure.

For corruption robustness, this thesis uses CIFAR-10-C and CIFAR-100-C [3]. These benchmarks evaluate models on 19 corruption types grouped into broad categories such as noise, blur, weather, and digital corruptions. Each corruption is provided at five severity levels. This setup makes it possible to analyze both average robustness and failure patterns under increasing corruption strength.

3.2.2 Medical Image Datasets

For the medical-image setting, this thesis uses two datasets from the MedMNIST collection [23]. These datasets are standardized to 28×28 pixels.

- **DermaMNIST**: A dermatology dataset for classifying seven skin lesion categories [23]. It contains 10,015 samples split into 7,007 training, 1,003 validation, and 2,005 test samples.
- **PathMNIST**: A histology dataset of colorectal tissue patches classified into nine categories [23]. It contains 107,180 samples split into 89,996 training, 10,004 validation, and 7,180 test samples.

3.2.3 Corruption Benchmarks

For medical robustness evaluation, this thesis uses MedMNIST-C [22]. Although the full benchmark contains additional task-specific corruptions, this thesis focuses on a common subset of eight corruption types so that comparisons remain consistent across the selected medical datasets. These corruptions include:

- **Blur and artifact corruptions:** Defocus Blur, Motion Blur, JPEG Compression, and Pixelate.
- **Intensity-related corruptions:** Brightness Up, Brightness Down, Contrast Up, and Contrast Down.

This restricted subset is used because it provides a more comparable evaluation protocol across DermaMNIST and PathMNIST, while still covering several practically relevant image degradations.

Table 2: Overview of datasets and corruption protocols used in this thesis.

Domain	Dataset	Clean Setting	Corruption Benchmark
Natural	CIFAR-10	10 classes, 32×32 RGB	CIFAR-10-C with 19 corruption types and 5 severity levels
Natural	CIFAR-100	100 classes, 32×32 RGB	CIFAR-100-C with 19 corruption types and 5 severity levels
Medical	DermaMNIST	7 classes, 28×28	MedMNIST-C subset with 8 corruption types and 5 severity levels
Medical	PathMNIST	9 classes, 28×28	MedMNIST-C subset with 8 corruption types and 5 severity levels

The table above provides a compact overview of the benchmark design and the domain split used throughout the experiments.

3.3 Model Architectures and Training Setup

The main experiments in this thesis use ResNet-18 as the primary benchmark backbone. In addition, ViT-B is included as an additional architecture in order to examine whether the main augmentation trends remain similar across model families. The goal is not to construct a fully symmetric CNN-versus-ViT comparison, but to test whether the main observations from the primary benchmark also appear under a different architecture.

3.3.1 Primary Backbone: ResNet-18

To keep the main comparison focused on augmentation effects, this thesis uses ResNet-18 [16] as the default architecture for the core experiments.

ResNet-18 is a convolutional neural network that uses residual connections to make optimization easier in deeper models. A residual block can be written as

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{4}$$

where x is the input and y is the output. The term $\mathcal{F}(x, \{W_i\})$ denotes the residual mapping to be learned. The shortcut connection allows gradients to pass more directly through the network during backpropagation and helps reduce optimization difficulties in deeper architectures. ResNet-18 also uses Batch Normalization [28] and ReLU activations [29].

With roughly 11 million parameters, ResNet-18 offers a useful balance between model capacity and computational cost. For this reason, it is a practical main backbone for the benchmark setting used in this thesis.

3.3.2 Additional Backbone: ViT-B

In addition to ResNet-18, this thesis includes ViT-B as an additional architecture. Vision Transformers process images differently from convolutional networks and model interactions across image patches through self-attention [17]. This difference makes ViT-B useful for checking whether augmentation effects are architecture-dependent.

In this thesis, ViT-B is evaluated under the same three-stage framework as ResNet-18. At the same time, it is treated as an additional architecture rather than as a fully symmetric second benchmark track. The purpose is to examine whether the main augmentation trends observed under the CNN-based benchmark remain broadly similar under a transformer-based model family.

The inclusion of ViT-B is also motivated by prior work showing that transformer-based vision models can respond differently to augmentation, regularization, and data scale compared with standard CNNs [25–27]. Because of computational constraints, the Stage 2 training budget for ViT-B is fixed to 30 epochs across all datasets. This shorter budget keeps the extension study manageable while still allowing a meaningful cross-architecture comparison within a consistent evaluation protocol.

This compact comparison is sufficient for the methodology chapter, since the goal here is to clarify the experimental role of each backbone rather than to provide a full architectural tutorial.

3.3.3 Optimization and Hyperparameter Protocol

All models are trained with stochastic gradient descent (SGD) [30] and Nesterov momentum. To keep the comparison fair, the optimization setup is fixed as much as

Table 3: Roles of the two backbone families in this thesis.

Backbone	**Architectural tendency**	**Role in this thesis**
ResNet-18	CNN with strong local spatial bias	Primary benchmark for the full three-stage comparison
ViT-B	Transformer with patch-based global interaction	Additional backbone for cross-architecture trend checking

possible across experiments, while the learning rate is selected separately in Stage 1 for each dataset and architecture.

The main Stage 2 training setup used in this thesis is summarized in Table 4. For ResNet-18, the training budget is set to 100 epochs on CIFAR-10 and CIFAR-100, and to 50 epochs on DermaMNIST and PathMNIST. For ViT-B, the Stage 2 budget is fixed to 30 epochs across all four datasets. This choice is mainly motivated by computational cost and reflects the role of ViT-B as an additional architecture rather than the primary benchmark backbone.

Table 4: Training hyperparameters used in Stage 2.

Hyperparameter	**ResNet-18**	**ViT-B**
Optimizer	SGD [30]	SGD [30]
Momentum	0.9	0.9
Weight Decay	5×10^{-4}	5×10^{-4}
Batch Size	128	128
Training Epochs	100 / 50 (dataset-dependent)	30
LR Scheduler	Cosine Annealing	Cosine Annealing

The learning rate itself is not fixed globally. Instead, it is selected in Stage 1 for each dataset and architecture and then reused consistently in Stage 2 and Stage 3. This keeps the later comparison more interpretable, since performance differences are less likely to come from arbitrary optimizer settings.

3.3.4 Stage 1 Search Space and Selection Logic

The Stage 1 learning-rate search is performed separately for each dataset and architecture. This is necessary because the optimization behavior differs across domains and across model families.

For ResNet-18, the candidate learning rates are:

- CIFAR-10 / CIFAR-100: {0.1, 0.05, 0.01, 0.001}
- DermaMNIST / PathMNIST: {0.01, 0.005, 0.001, 0.0005}

For ViT-B, a lower learning-rate range is used throughout:

- CIFAR-10 / CIFAR-100 / DermaMNIST / PathMNIST: {0.001, 0.0005, 0.0001, 0.00005}

The Stage 1 epoch budget also differs across datasets and architectures. For ResNet-18, Stage 1 runs are trained for 20 epochs on CIFAR-10 and CIFAR-100, and for 15 epochs on DermaMNIST and PathMNIST. For ViT-B, Stage 1 runs are trained for 15 epochs on CIFAR-10, 20 epochs on CIFAR-100, and 10 epochs on both DermaMNIST and PathMNIST.

Based on this search protocol, the final learning rates selected for Stage 2 are:

- **ResNet-18:** 0.01 for CIFAR-10, 0.001 for CIFAR-100, 0.005 for DermaMNIST, and 0.001 for PathMNIST
- **ViT-B:** 0.0005 for CIFAR-10, 0.0005 for CIFAR-100, 0.001 for DermaMNIST, and 0.001 for PathMNIST

This stage is designed to provide stable and reusable optimization settings for the later comparison, rather than to maximize the performance of a single preliminary run.

3.3.5 Experimental Roles of the Two Architectures

The roles of the two architectures in this thesis are intentionally different.

- **ResNet-18:** It serves as the primary benchmark backbone. The full three-stage pipeline and the main comparisons are centered on this architecture.
- **ViT-B:** It serves as an additional architecture. It is also evaluated through the same three-stage framework, but with a reduced Stage 2 training budget in order to keep the overall study computationally manageable.

This distinction is important for the interpretation of the later results. The thesis does not aim to provide a fully exhaustive comparison between CNNs and Vision Transformers. Instead, the architecture extension is used to test whether the main conclusions from the ResNet-18 experiments remain broadly stable under a transformer-based model family.

3.4 Data Augmentation Methods

The evaluation includes nine augmentation strategies. For clarity, they are grouped into four categories. The implementations follow standard libraries or official repositories where available. Within the traditional category, this includes the standard

Table 5: Data augmentation methods evaluated in this thesis.

Category	**Methods**
Baseline	Standard (Crop+Flip), Rotation + Random Erasing
Policy-based	AutoAugment [5], RandAugment [6]
Mixing	Mixup [7], CutMix [8]
Generative	AugMix [9], StyleAug [10], DiffuseMix [11]

baseline pipeline and an augmentation setting that combines random rotation with *Random Erasing*.

In the experiments of this thesis, the traditional augmentation setting beyond the standard baseline is not a standalone Random Erasing transform. Instead, it combines small random rotations with *Random Erasing*. In later chapters, this setting is therefore referred to as Rotation + Random Erasing.

The baseline methods provide the reference point for the later comparison. Policy-based methods define transformation policies, mixing methods combine information from multiple samples, and the generative or style-related methods aim to expose the model to broader appearance changes during training.

3.5 Evaluation Metrics

Model behaviour is evaluated from three perspectives: classification performance, calibration, and robustness under corruption.

3.5.1 Classification Metrics

For balanced datasets such as CIFAR-10 and CIFAR-100, the main classification metric is accuracy (Acc), that is, the proportion of correctly classified samples.

For medical datasets such as DermaMNIST, class imbalance makes standard accuracy less informative. In such cases, this thesis also uses balanced accuracy (BalAcc), defined as the average recall across classes:

$$\mathrm{BalAcc} = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{5}$$

where K is the number of classes.

Using both metrics where appropriate is important for the later comparison, because the natural-image and medical-image settings differ in class structure and evaluation emphasis.

3.5.2 Calibration Metrics

To measure predictive reliability, this thesis uses Expected Calibration Error (ECE) [12]. Predictions are grouped into M confidence bins, and ECE computes the weighted average gap between empirical accuracy and confidence:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{6}$$

Maximum Calibration Error (MCE). In addition to ECE, this thesis also considers Maximum Calibration Error (MCE) as a supplementary calibration metric. While ECE summarizes the average gap between confidence and empirical accuracy across bins, MCE measures the largest such gap:

$$\mathrm{MCE} = \max_{m \in \{1,\ldots,M\}} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{7}$$

MCE is useful because it highlights worst-case miscalibration in particular confidence regions, even when the average calibration error remains moderate. In this thesis, ECE and mean ECE (mECE) are used as the primary calibration metrics, while MCE is used as a supplementary metric to provide additional detail on calibration risk.

In later analyses, this thesis also reports calibration behavior under corruption, because accurate but poorly calibrated predictions can still be problematic in practice.

3.5.3 Robustness Metrics under Corruption

Robustness is evaluated by aggregating performance over corruption types and severity levels [3]. For natural-image corruption benchmarks, this thesis uses mean corruption accuracy (mCA), defined as

$$\mathrm{mCA} = \frac{1}{|C|} \sum_{c \in C} \left(\frac{1}{5} \sum_{s=1}^{5} \mathrm{Acc}_{c,s} \right) \tag{8}$$

where C is the set of corruption types and $s \in \{1, \ldots, 5\}$ denotes the severity level.

Calibration under corruption is summarized mainly through mean ECE (mECE). Where useful, the thesis also reports a corruption-aggregated MCE view as a supplementary indicator of worst-case calibration error under corruption.

3.6 Implementation and Reproducibility

All experiments are implemented in PyTorch [31]. To keep the experiments reproducible, the following decisions are used throughout the thesis:

- Random seeds are fixed for data splitting and model initialization.
- Augmentation parameters follow literature defaults where applicable, for example Mixup with $\alpha = 0.2$ and CutMix with $\alpha = 1.0$.
- Configuration files are organized through a YAML-based setup so that experimental settings can be tracked consistently.
- Results used in Chapter 4 are exported from the unified analysis pipeline and, where applicable, aggregated over multiple runs to reduce stochastic variance.

These implementation choices help keep the experimental pipeline consistent and make the later analysis easier to reproduce and interpret.

4 Experimental Results

4.1 Experimental Setup and Reporting Protocol

This chapter presents the main empirical results of the thesis. The order of presentation follows the evaluation pipeline from Chapter 3, so that the later robustness comparisons can be read against the earlier optimization and clean-data results.

The main benchmark track is based on ResNet-18. ViT-B is included where it adds something substantive to the interpretation, especially when a result looks stable or unstable across architectures.

The natural-image experiments are based on CIFAR-10 and CIFAR-100, while the medical-image experiments are based on DermaMNIST and PathMNIST. Robustness is evaluated on 19 corruption types for CIFAR-C and on a common subset of 8 corruption types for MedMNIST-C.

This chapter reports both predictive performance and predictive reliability. On clean data, the main reference metric is classification accuracy, or balanced accuracy where class imbalance makes ordinary accuracy less informative. Under corruption, robustness is summarized through mean corruption accuracy (mCA), where higher values indicate stronger robustness. Calibration is discussed mainly through Expected Calibration Error (ECE) on clean data and mean ECE (mECE) under corruption. Maximum Calibration Error (MCE) and its corruption-aggregated form are used as supplementary indicators.

The chapter starts with the Stage 1 learning-rate search, then turns to clean performance and clean calibration, and finally analyses corruption robustness in the natural-image and medical-image settings before closing with a short synthesis across domains and architectures.

The reported results should also be read in light of the training budget used in each stage. In Stage 1, the learning-rate search is performed with a short budget in order to keep the search practical while still making learning-rate differences visible. For

ResNet-18, the Stage 1 budget is 20 epochs on CIFAR-10 and CIFAR-100 and 15 epochs on DermaMNIST and PathMNIST. For ViT-B, the Stage 1 budget is 15 epochs on CIFAR-10, 20 epochs on CIFAR-100, and 10 epochs on DermaMNIST and PathMNIST. In Stage 2, the main benchmark with ResNet-18 is trained for 100 epochs on CIFAR-10 and CIFAR-100 and 50 epochs on DermaMNIST and PathMNIST, whereas the ViT-B extension is trained for 30 epochs across all four datasets. This difference reflects practical computational cost constraints and should be understood as part of the study design: ResNet-18 serves as the main benchmark track, while ViT-B is included as a supplementary cross-architecture check rather than as a fully symmetric full-budget benchmark.

4.2 Stage 1: Learning Rate Selection

The purpose of Stage 1 is to fix learning rates before the augmentation comparison begins. This stage is not meant to tune every augmentation separately. Its role is simply to find settings that produce stable and comparable training behaviour for each dataset and backbone, so that later differences are more likely to reflect augmentation effects rather than optimization noise.

A grid search is run with the baseline augmentation setting. The candidate values differ across datasets and architectures so that the range stays realistic for each model family. For ResNet-18, the search is centred on larger values for the CIFAR datasets and smaller values for the medical-image datasets. For ViT-B, the entire range is shifted downward. The final choice is based on validation behaviour rather than test performance, which keeps the later robustness evaluation separate from hyperparameter selection.

Even this short search already shows that optimization behaviour changes with both domain and architecture. On the CIFAR datasets, ResNet-18 still trains stably at comparatively large learning rates, while the medical-image datasets require more conservative choices. ViT-B uses smaller selected values throughout, which fits the generally more sensitive optimization behaviour of the transformer backbone.

For the main stages, the selected learning rates are 0.01, 0.001, 0.005, and 0.001 for ResNet-18 on CIFAR-10, CIFAR-100, DermaMNIST, and PathMNIST, respectively. For ViT-B, the selected values are 0.0005, 0.0005, 0.001, and 0.001. Figure 3 is included mainly as a compact methodological overview, while the full backbone-specific curves are moved to Appendix A.

For the medical-image datasets, validation balanced accuracy is considered alongside ordinary validation accuracy when interpreting the curves, because class imbalance makes balanced metrics more informative in these settings. This is especially relevant for DermaMNIST and PathMNIST, where learning-rate stability should not be judged by overall accuracy alone.

Stage 1 is therefore mainly preparatory, but it is still informative: before any augmentation comparison is made, it already becomes clear that the two backbones should not be expected to behave identically.

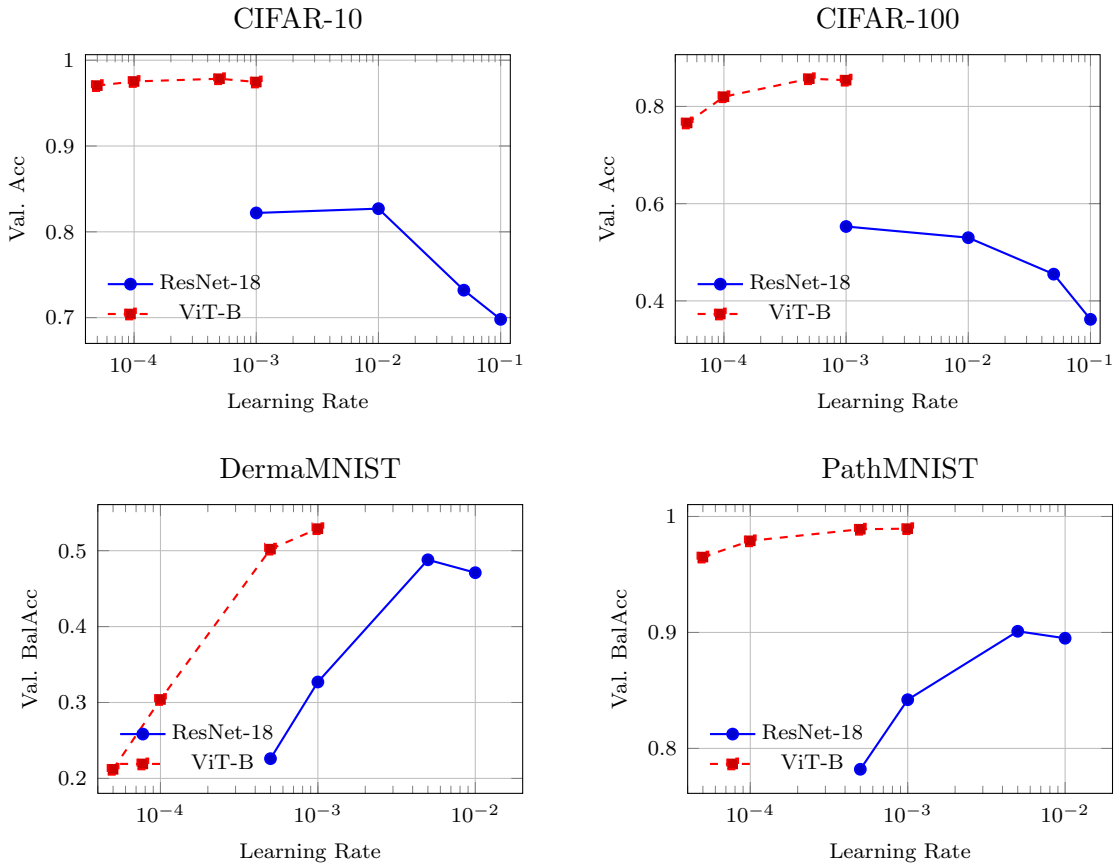


Figure 3: Compact Stage 1 learning-rate overview across the four datasets. Each panel overlays ResNet-18 and ViT-B so that dataset-level selection trends can be compared more directly in the main text. For CIFAR-10 and CIFAR-100, validation accuracy is reported. For DermaMNIST and PathMNIST, validation balanced accuracy is used because class imbalance makes it more informative than ordinary validation accuracy. The full backbone-specific Stage 1 curves are provided in Appendix A.

4.3 Stage 2: Clean Performance and Calibration

After fixing the learning rates, Stage 2 compares the augmentation methods on clean, in-distribution data. These results matter because they provide the reference point for the corruption analysis in Stage 3. Without that reference, it would be hard to tell whether later robustness differences simply mirror cleaner training performance or whether they reflect something more specific.

The clean results should also be read together with the training budget. ResNet-18 is the main benchmark setting, whereas ViT-B is included as a shorter but fixed extension. The cross-architecture comparison is therefore useful for checking trend stability, but it is not intended as a fully compute-matched backbone study.

4.3.1 Clean Predictive Performance

On clean data, many methods stay fairly close to each other, but the ranking still shifts with the dataset and the backbone. For CIFAR-10 and CIFAR-100, the natural reference is clean accuracy. For DermaMNIST and PathMNIST, balanced accuracy is more informative because of the class structure of the medical tasks.

Clean rankings are relatively compressed in the natural-image setting. On CIFAR-10 / ResNet-18, AutoAugment is highest at 0.8540, but margins are small; under ViT-B, CutMix reaches 0.9829. On CIFAR-100, Mixup leads under ResNet-18 (0.6074), while CutMix leads under ViT-B (0.8853). So the clean ranking is already architecture-dependent before corruption is introduced.

The medical setting shows the same point through balanced accuracy: on DermaMNIST, CutMix is strongest under ResNet-18 (0.7627), while ViT-B shares the top value between CutMix and Rotation + Random Erasing (0.7936). PathMNIST is more stable, with CutMix leading on both backbones (0.9207 and 0.9968). Overall, clean results are informative but not strongly separating; full Stage 2 tables are kept in Appendix C.

4.3.2 Clean Calibration Behavior

Although clean predictive performance is often close across methods, calibration separates them more clearly. Stage 2 therefore provides not only a performance comparison, but also an early view of predictive reliability. Some augmentations produce smoother confidence behaviour on clean data, while others remain comparatively overconfident even when their predictive performance looks competitive.

The clean-stage results show that predictive performance and calibration should be read jointly. Figure 4 gives one representative view, while additional clean trade-off plots and the multi-setting dual-axis clean overview are reported in Appendix C. On CIFAR-10 / ResNet-18, Mixup has the lowest clean ECE (0.0336) without the highest accuracy, whereas DiffuseMix is weaker on both dimensions. On CIFAR-100 / ResNet-18, Mixup is strongest overall, combining the best clean accuracy (0.6074) with the lowest clean ECE (0.0612).

The same pattern appears under ViT-B at a higher absolute level: on CIFAR-10, CutMix combines highest clean accuracy (0.9829) and lowest clean ECE (0.0051). In the medical setting, calibration also remains method- and backbone-dependent; for example, CutMix reaches low ECE on DermaMNIST (0.0566 under ResNet-18, 0.0270 under ViT-B), while PathMNIST shows a large backbone gap for CutMix ECE (0.1294 vs 0.0040).

The supplementary MCE results support this interpretation but add a worst-case view: low average calibration error does not always imply low worst-bin error. Overall, Stage 2 already shows why Stage 3 cannot be interpreted from accuracy alone.

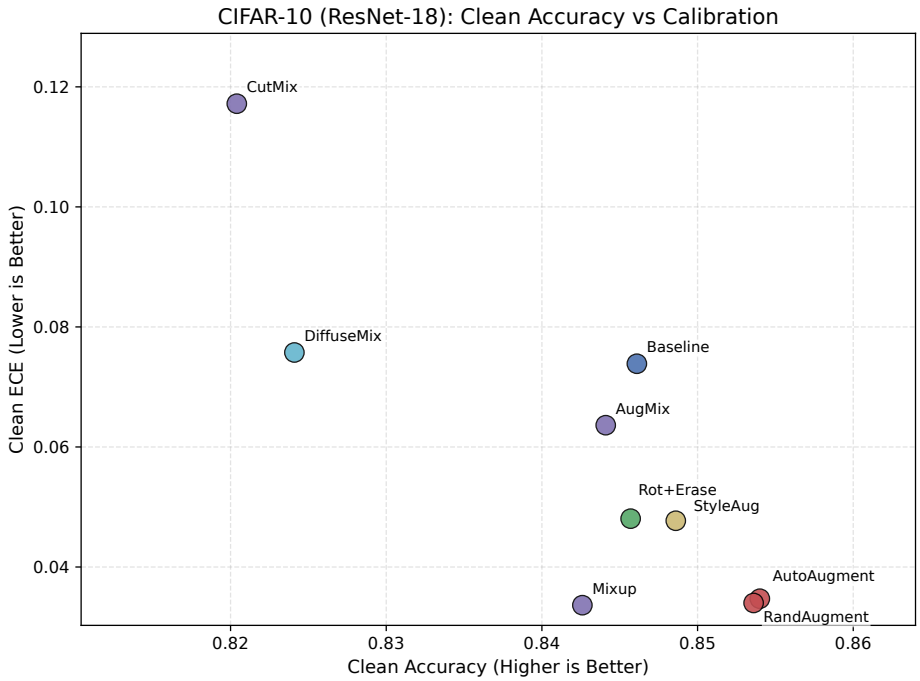


Figure 4: Representative Stage 2 clean trade-off example on CIFAR-10 with ResNet-18.

4.3.3 Cross-Architecture Observations on Clean Data

The cross-architecture clean comparison helps separate more stable from less stable trends. Some methods stay broadly in a similar part of the ranking across ResNet-18 and ViT-B, while others move much more once the backbone changes.

On CIFAR-10, several methods remain reasonably competitive across both backbones, but CutMix changes the most: it is the weakest method in clean accuracy under ResNet-18 and the strongest under ViT-B. This kind of reversal is important because it shows how misleading a single-backbone reading could be. On CIFAR-100, Mixup looks more stable. It stays near the top on both architectures and remains attractive in both clean performance and calibration. On DermaMNIST, the clean balanced-accuracy ranking changes more noticeably, which suggests that the medical-image clean comparison is already more architecture-sensitive. PathMNIST is the clearest exception, since CutMix remains the top clean balanced-accuracy method under both architectures, even though its calibration pattern still differs.

These early shifts matter because they prepare the ground for the later synthesis in Section 4.6. Architecture effects are already visible on clean data, so it would be misleading to discuss corruption robustness as if the backbone only mattered at Stage 3.

4.3.4 Interpretation of Stage 2

Stage 2 gives a useful reference point, but it is still only a first step. On the natural-image datasets, clean accuracy remains the main reference; on DermaMNIST and

PathMNIST, balanced accuracy is the more meaningful clean metric. Across both domains, however, clean performance alone still leaves too much hidden.

Calibration adds an extra layer that is already informative before any corruption is applied. The supplementary MCE results support the same reading by showing that average calibration and worst-case calibration are related, but not identical. In that way, Stage 2 works as a bridge: it provides the clean reference for Stage 3 and already hints that part of the augmentation effect only becomes visible once the evaluation moves beyond in-distribution performance.

4.4 Stage 3: Robustness Analysis on Natural Images

This section turns to robustness under distribution shift in the natural-image setting. Using CIFAR-C, it asks how the augmentation methods behave once the clean test images are replaced by corrupted variants. Compared with Stage 2, the focus is no longer on in-distribution performance, but on how well the models hold up under a broad set of non-adversarial corruptions.

4.4.1 Quantitative Summary

At the aggregate level, robustness is summarized through corruption-averaged metrics such as mCA, together with calibration-oriented quantities such as mECE. These metrics help separate methods that merely stay accurate under corruption from methods that also keep their confidence estimates under better control. Supplementary MCE results add a worst-case view, but the main discussion remains centred on mCA and mECE.

To keep the main text readable while still reporting concrete values, Table 6 provides a condensed ResNet-18 snapshot (best mCA, best mECE, and baseline), whereas the full per-method tables remain in Appendix C.

Table 6: Condensed natural-image robustness summary (ResNet-18). Full per-method tables are provided in Appendix C.

Dataset	Best mCA Method	mCA	Best mECE Method	mECE	Baseline (mCA/mECE)
CIFAR-10	StyleAug	0.7815	RandAugment	0.0745	0.7128 / 0.1629
CIFAR-100	AugMix	0.4846	CutMix	0.0602	0.4266 / 0.3155

The quantitative summary gives the ranking, but not the mechanism; corruption-wise and severity-wise details are therefore discussed next, while full tables are moved to Appendix C.

On ResNet-18, CIFAR-10 and CIFAR-100 already diverge: CIFAR-10 is led by StyleAug in mCA (0.7815), whereas CIFAR-100 is led by AugMix (0.4846). Calibration changes the reading in both cases: on CIFAR-10, RandAugment/StyleAug are more balanced than DiffuseMix (mECE 0.0745/0.0802 vs 0.1088), and on CIFAR-100, CutMix has the lowest mECE (0.0602) despite not leading mCA.

Under ViT-B, the trade-off remains: on CIFAR-10, RandAugment has the highest mCA (0.9092), while CutMix has the lowest mECE (0.0315); on CIFAR-100, AugMix stays strongest in mCA, but CutMix/Mixup are calibration-friendlier. The MCE results support the same direction and show that average and worst-case calibration are not identical.

4.4.2 Corruption-wise and Severity-wise Analysis

Aggregate robustness metrics are informative, but they can hide important differences between augmentation methods. Two methods may achieve similar average mCA while failing on different corruption families. To make these differences visible, this section also considers corruption-wise and severity-wise patterns.

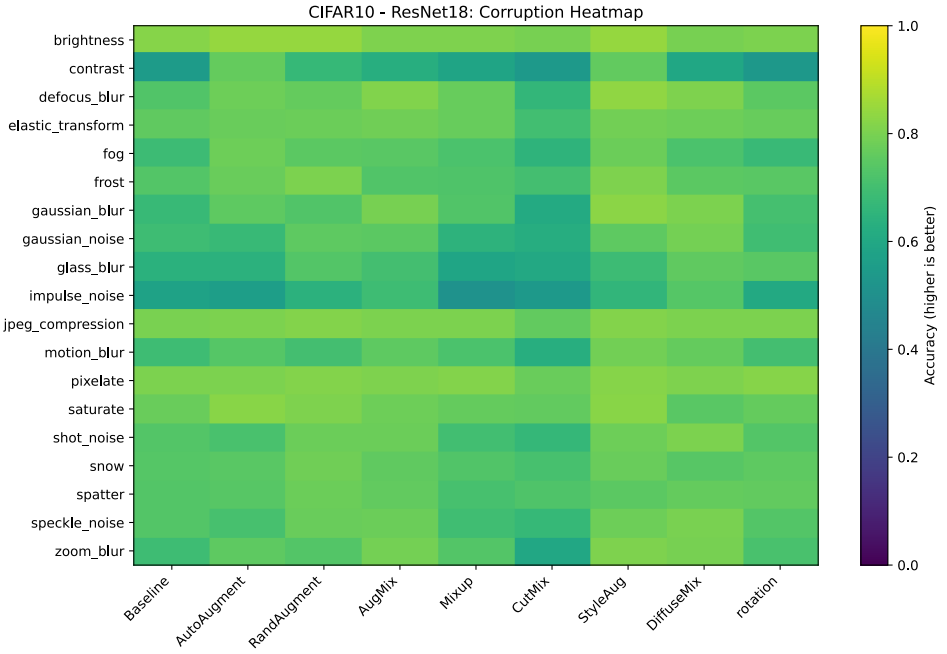


Figure 5: CIFAR-10 corruption-wise robustness heatmap for ResNet-18. The corresponding ViT-B view is reported in Appendix C.

The CIFAR-10 heatmap in Figure 5 already shows the key pattern: architecture changes affect both average robustness and the structure of corruption sensitivity. Under ResNet-18, variation across corruption types is stronger, while ViT-B results are generally stronger. Severity curves for CIFAR-10 and the corresponding CIFAR-100 details are moved to Appendix C.

For readability, the main text keeps CIFAR-10 as the representative detailed example and moves the fuller CIFAR-100 visual material to the Appendix.

4.4.3 Robustness–Calibration Trade-off

In addition to raw robustness, this thesis also asks whether robustness gains are accompanied by reliable confidence estimates. A method may preserve relatively

high mCA under corruption, but still become poorly calibrated when it fails. For this reason, robustness and calibration are examined jointly rather than treated as independent observations.

Figure 6 highlights why natural-image robustness should not be read from mCA alone. ViT-B occupies a stronger robustness region than ResNet-18, but method-level trade-offs remain visible: under ResNet-18, DiffuseMix is robust but less calibrated; under ViT-B, CutMix combines strong robustness with low mECE. On CIFAR-100, the same pattern holds in a harder setting: AugMix leads in mCA, while CutMix/Mixup are more calibration-friendly.

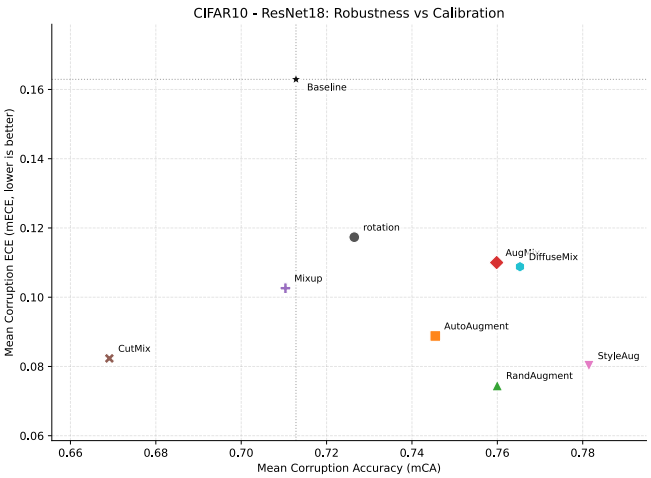


Figure 6: Representative natural-domain trade-off on CIFAR-10 with ResNet-18.

The supplementary MCE results reinforce this point: average and worst-case calibration do not always rank methods identically.

4.4.4 Interpretation of the Natural-Image Results

The CIFAR-C results give the first full view of how the augmentation methods behave once distribution shift is introduced. The aggregate metrics show which methods are strongest on average, while the detailed plots make it easier to see where those rankings come from. Reading mCA together with mECE and MCE also makes an accuracy-only interpretation of robustness hard to justify.

The natural-image results also show that architecture matters here. On both CIFAR-10 and CIFAR-100, ViT-B reaches a noticeably higher robustness level than ResNet-18. But the change is not just a simple upward shift. Corruption sensitivity and the robustness–calibration trade-off still depend on the augmentation choice.

4.5 Stage 3: Robustness Analysis on Medical Images

This section extends the robustness analysis to the medical-image setting. Using MedMNIST-C, it asks whether methods that look strong on CIFAR-C remain strong once the data domain, corruption types, and visual structure change more substantially.

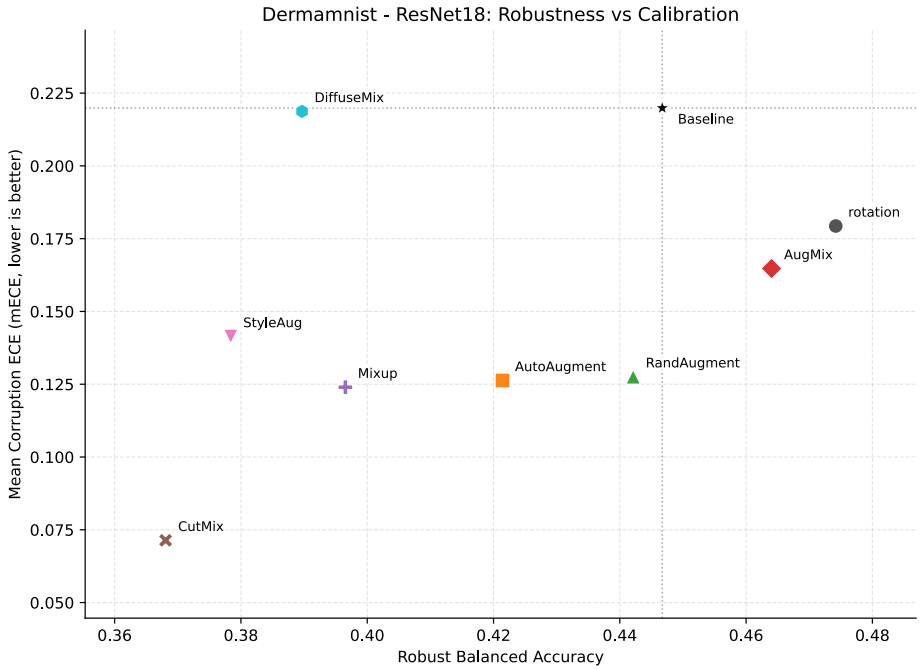


Figure 7: Representative medical-domain robustness–calibration trade-off on DermaMNIST with ResNet-18.

4.5.1 Quantitative Summary

As in the natural-image setting, the medical-image results start with aggregate metrics for corruption-averaged performance and corruption-aggregated calibration. These values give a first ranking, but they do not explain by themselves why some methods transfer better to the medical domain than others. Again, ECE-based metrics remain the main calibration reference, while MCE is used as a supplementary worst-case indicator.

Table 7 reports the corresponding condensed ResNet-18 view for the medical domain (best mCA, best mECE, and baseline), with complete per-method tables kept in Appendix C.

Table 7: Condensed medical-image robustness summary (ResNet-18). Full per-method tables are provided in Appendix C.

Dataset	Best mCA Method	mCA	Best mECE Method	mECE	Baseline (mCA/mECE)
DermaMNIST	RandAugment	0.7334	CutMix	0.0714	0.7124 / 0.2199
PathMNIST	RandAugment	0.8581	RandAugment	0.0730	0.5983 / 0.3505

Figure 7 shows stronger setting dependence than the natural-image case. Under ResNet-18, RandAugment is strongest on PathMNIST (mCA 0.8581), but DermaMNIST has no single winner across mCA and mECE. Under ViT-B, rankings shift again, with AugMix becoming especially strong on PathMNIST. The supplementary MCE view supports the same conclusion; full medical tables and split plots are reported in Appendix C.

4.5.2 Corruption-wise Failure Analysis

To better understand the medical-image results, this section also considers corruption-specific behaviour. Aggregate performance can hide the fact that a method is stable under one type of corruption but fails clearly under another. A more detailed view is therefore needed to identify which medical corruption patterns remain difficult for different augmentation families.

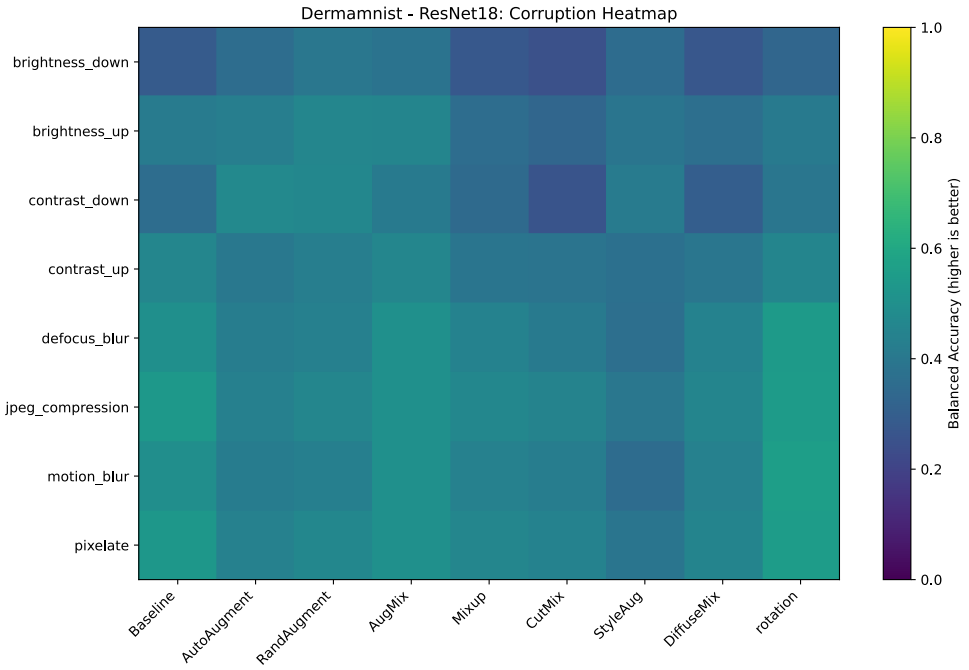


Figure 8: DermaMNIST corruption-wise robustness heatmap for ResNet-18. The corresponding ViT-B view is reported in Appendix C.

The heatmap in Figure 8 highlights corruption-specific variability on DermaMNIST more directly than the aggregate mCA values. Under ResNet-18, several corruption types show visibly sharper performance drops, which supports the earlier point that medical-image robustness remains strongly setting-dependent.

Together with the supplementary severity curves in Appendix C, Figure 8 indicates that architecture changes affect both overall robustness level and the structure of corruption sensitivity. ViT-B is generally stronger, but improvements are not uniform across methods.

For PathMNIST, the heatmaps, severity curves, and trade-off plots still add useful evidence, but they work better as supporting material than as a second main-text overview figure. The visual pattern also differs from the natural-image setting because intensity-related corruption behaviour becomes more central.

In the medical domain, this detailed analysis is especially informative because the relevant corruption types differ from those emphasized in natural-image benchmarks. Intensity-related corruptions such as brightness and contrast changes may play a larger role, while some augmentation methods that improve robustness to texture-

like changes on CIFAR-C may be less useful when the discriminative medical patterns are more subtle.

4.5.3 Domain-Specific Interpretation

The medical-image results should be interpreted more cautiously because augmentation may interact with this domain in a more task-specific way. A method that increases diversity in natural images is not automatically a good fit for medical data if it distorts structure that is actually relevant for the prediction.

The medical analysis is therefore not treated as a simple repeat of the CIFAR-C benchmark. It is meant to test whether the augmentation effects seen in natural images still hold in a second domain with different visual and task-specific properties.

The plots and calibration results support that cautious reading. In DermaMNIST, for example, some methods appear to benefit more clearly from the ViT-B extension than others, which suggests that medical-image robustness remains architecture-sensitive even before the later cross-domain synthesis. The supplementary MCE results support the same point by showing that average calibration and worst-case calibration do not always move in exactly the same way.

4.5.4 Interpretation of the Medical-Image Results

The MedMNIST-C results provide the second main robustness perspective of the study. The aggregate metrics show how the methods rank in the medical domain, while the detailed plots make the ranking shifts easier to interpret. This allows the discussion to compare not only performance levels, but also the structure of failure patterns across domains.

The medical-image setting should not be read as a simple extension of CIFAR-C. The dominant corruption sensitivities change, the most useful augmentation families can change, and the robustness–calibration trade-off remains strongly setting-dependent.

4.6 Cross-Domain and Cross-Architecture Synthesis

This final section of the chapter brings the earlier results together. The point is not to repeat every dataset-level result, but to identify the broader patterns that become visible once the methods are compared across domains and, where possible, across architectures.

4.6.1 Domain Dependency of Augmentation Effects

The comparison between CIFAR-C and MedMNIST-C shows clearly that augmentation effects are not domain-invariant. Methods that perform strongly on natural-image corruptions do not necessarily keep the same advantage once the evaluation

shifts to medical images. Robustness gains therefore need to be read in relation to the target domain rather than treated as universally transferable.

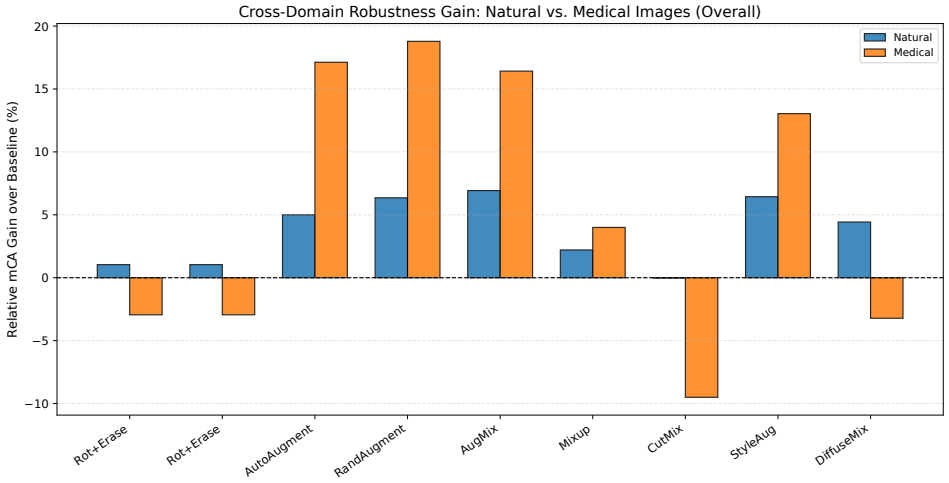


Figure 9: Cross-domain overview of corruption-robustness gains relative to the baseline. The figure summarizes how strongly each augmentation improves mean corruption accuracy across the natural-image and medical-image settings.

A central reason is that relevant corruption families differ by domain: texture/noise variation is more dominant in natural images, while intensity-related shifts are more central in medical images. Figure 9 reflects this directly: gains are broad on CIFAR-C but more selective on MedMNIST-C, and the top methods are not identical across domains.

At dataset level, natural-image top performers are often StyleAug, RandAugment, or AugMix, whereas medical rankings shift: RandAugment is strongest on DermaMNIST and PathMNIST under ResNet-18, but on PathMNIST / ViT-B AugMix leads in mCA. Calibration trends (ECE and supplementary MCE) follow the same message: robustness gains do not transfer uniformly across domains.

4.6.2 Architecture Dependency of Augmentation Effects

Alongside the domain comparison, the thesis also asks whether the main findings remain stable across architectures. The main benchmark is still ResNet-18, while ViT-B serves as an additional architecture for selected analyses.

Figure 10 shows this more directly than the earlier table. Some trends remain fairly stable when the backbone changes, but others are clearly reshaped. In the natural-image setting, moving from ResNet-18 to ViT-B often raises the overall robustness level, yet the augmentation ranking is not preserved exactly. On CIFAR-10, for example, StyleAug is strongest under ResNet-18, whereas RandAugment leads under ViT-B and CutMix becomes more attractive once calibration is considered jointly. On CIFAR-100, AugMix stays very strong in both architectures, but the surrounding trade-off structure still changes.

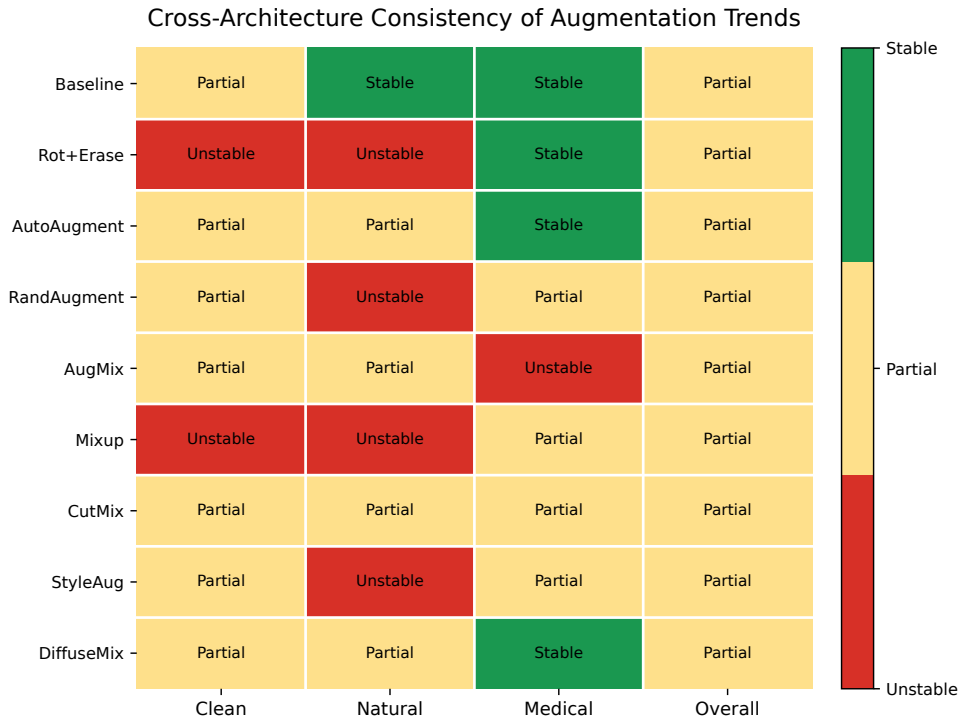


Figure 10: Cross-architecture consistency of augmentation trends. Cell colours summarize how stable each augmentation remains between ResNet-18 and ViT-B in clean, natural-corruption, medical-corruption, and overall settings.

The medical-image comparison shows the same general pattern in a different form. On DermaMNIST, RandAugment remains highly robust in both architectures, but the best calibration is not always achieved by the same method. On PathMNIST, the strongest method changes more clearly between architectures, with RandAugment leading under ResNet-18 and AugMix leading under ViT-B. This means that architecture changes do not simply strengthen all augmentation methods equally; instead, they reshape which trade-offs become most favourable.

The supplementary MCE analysis also supports the idea that architecture changes influence not only average calibration quality but sometimes also worst-case calibration behaviour. This makes the cross-architecture comparison more informative than a pure accuracy-based reading.

4.6.3 Stable and Unstable Findings Across Settings

The synthesis separates stable findings from setting-dependent ones. Stable findings are: policy-based methods (especially RandAugment) are frequently competitive, robustness should be read jointly with calibration (highest mCA is not always lowest mECE), and ViT-B usually raises the absolute robustness level.

Setting-dependent findings remain substantial: the best method changes with dataset, corruption family, and backbone (e.g., StyleAug strongest on CIFAR-10 / ResNet-18, AugMix strong on CIFAR-100 and PathMNIST / ViT-B, CutMix often strongest

from calibration). Supplementary MCE results support the same interpretation. Overall, augmentation effectiveness is interaction-driven across domain, corruption structure, and architecture.

5 Discussion

5.1 Clean Performance Is Informative but Not Sufficient

One important finding is that clean performance is useful, but not enough on its own. The Stage 2 results show that many augmentation methods remain fairly close on clean evaluation, especially on the natural-image benchmarks. Once the analysis moves to Stage 3, the gaps become easier to see. Strong clean performance should therefore not be read as a sufficient indicator of strong corruption robustness.

This pattern appears in both architectures, even though the absolute performance levels are different. Under ResNet-18, the clean-data gaps are often modest, while the differences under CIFAR-C or MedMNIST-C are clearer. Under ViT-B, both clean and robust performance are usually higher, but the same general conclusion still holds: stronger clean performance does not by itself determine the later robustness ranking. A stronger backbone does not remove the need for explicit robustness evaluation.

The results therefore argue against relying on clean metrics alone when selecting augmentation methods. Clean accuracy, or clean balanced accuracy in the medical-image setting, is still a necessary reference, but it should not be the main target of evaluation. A method can look competitive on clean data and still become much less reliable once realistic corruption is introduced.

This also helps explain why the Stage 2 and Stage 3 rankings are not identical. On clean data, part of the augmentation effect is hidden because many methods already fit the in-distribution benchmark reasonably well. Under corruption, the same methods have to show whether the invariances they encourage still help once the input changes. Seen this way, the value of augmentation lies less in improving clean fit alone and more in shaping representations that remain useful under shift.

A practical implication is that augmentation selection should not be driven by clean validation performance alone. When robustness matters, clean metrics should be complemented by explicit corruption evaluation. Otherwise, the evaluation protocol may overestimate methods that appear strong only because the clean benchmark is not difficult enough to expose their weaknesses.

5.2 Robustness and Calibration Are Related but Not Identical

Another central finding is that corruption robustness and calibration should not be treated as interchangeable objectives. Across both the natural-image and medical-

image benchmarks, the augmentation with the highest corruption robustness is not always the one with the best corruption calibration. In several settings, the ranking by mCA differs from the ranking by mECE. A model can therefore remain comparatively accurate under corruption while still becoming less reliable in its confidence estimates.

This distinction is already visible in the natural-image results. On CIFAR-10 with ResNet-18, DiffuseMix achieves strong corruption robustness, but its corruption calibration is weaker than that of RandAugment or StyleAug. It is therefore a good example of a method that remains strong in terms of accuracy while being less convincing from the confidence perspective. RandAugment and StyleAug look more attractive in the joint robustness–calibration view because they combine strong mCA with more controlled calibration error.

The same issue remains visible under ViT-B. On CIFAR-10, RandAugment reaches the highest mCA, but CutMix becomes especially competitive once calibration is considered jointly because it combines almost the same robustness with the lowest corruption calibration error. On CIFAR-100, AugMix reaches the highest mCA, yet CutMix and Mixup remain especially attractive from the calibration perspective. Even when the overall robustness level rises, the distinction between predictive robustness and predictive reliability remains.

The medical-image results reinforce the same conclusion. On DermaMNIST with ResNet-18, RandAugment has the highest mCA, but CutMix yields the lowest mECE. On PathMNIST with ResNet-18, RandAugment is unusually strong because it combines the highest mCA with the best corruption calibration. But that alignment does not remain universal once the backbone changes. On DermaMNIST with ViT-B, RandAugment still has the highest mCA, while StyleAug gives the lowest mECE. On PathMNIST with ViT-B, AugMix becomes strongest in both mCA and mECE. So the relationship between robustness and calibration is not fixed; it depends on the dataset and the backbone.

The additional MCE analysis makes this point more precise. ECE and mECE describe average calibration quality, which is useful for understanding overall confidence behaviour. MCE, in contrast, highlights the worst calibration gap across confidence bins. In the experiments of this thesis, these measures often move in a similar direction, but they are not fully identical. A method may look relatively well calibrated on average while still showing an unfavourable worst-bin error.

This distinction becomes more relevant in safety-related settings. In practice, a model that is accurate on average but still produces isolated high-confidence mistakes may remain problematic. This concern is especially clear in the medical-image setting, where confidence scores may affect whether a prediction is trusted directly or reviewed by a human. From this perspective, MCE does not replace ECE or mECE, but it adds another layer of interpretation by showing whether a method is only well calibrated on average or whether its confidence behaviour is also more stable in the worst case.

The results therefore support a broader evaluation principle: augmentation methods should be judged from a joint robustness–calibration perspective. A method with the highest mCA should not automatically be preferred if its calibration worsens strongly under corruption. Conversely, a method with slightly lower mCA may still be the better choice when it produces more reliable confidence estimates. This is one of the clearest lessons that appears repeatedly across the experiments.

5.3 Augmentation Effects Are Domain-Dependent

One of the central questions of the thesis was whether robustness gains observed on natural-image benchmarks transfer to the medical-image setting. The results suggest that the answer is only partial. Some augmentation families remain competitive across domains, but the strongest methods, the most relevant corruption sensitivities, and the robustness–calibration trade-offs all change noticeably between the two settings.

The contrast already appears in the overall ranking patterns. In the natural-image setting, the leading methods often include StyleAug, RandAugment, and AugMix, depending on the dataset and architecture. On CIFAR-10 under ResNet-18, for example, StyleAug reaches the highest corruption robustness, while under ViT-B the top position shifts to RandAugment and CutMix becomes especially attractive once calibration is considered jointly. On CIFAR-100, AugMix becomes particularly strong, especially under ViT-B. In the medical-image setting, however, the ranking changes. On DermaMNIST and PathMNIST under ResNet-18, RandAugment is especially competitive, while under ViT-B the strongest method can change again, most clearly on PathMNIST where AugMix reaches the highest mCA.

This already suggests that augmentation should not be treated as a universally portable robustness solution. A method that performs very well in the natural-image setting cannot automatically be expected to transfer with the same advantage to the medical-image setting. In broad terms, the reason is that the two domains reward different kinds of invariance.

In the natural-image benchmarks, corruption sensitivity appears to be closely related to texture variation, noise, blur, and broad appearance change. Augmentation methods such as StyleAug and AugMix appear to be well aligned with this setting because they encourage broader invariance to changes in texture and appearance. This alignment helps explain why they are often especially effective in CIFAR-C-style benchmarks.

In the medical-image setting, the situation is different. Here, intensity-related shifts and domain-specific visual structure become more important. On datasets such as PathMNIST, the image content is not defined by large-scale object shape in the same way as in natural-image classification. Instead, the task depends more strongly on subtle local patterns, tissue structure, and medically relevant visual statistics. As a result, an augmentation strategy that is highly beneficial in natural images may become less effective, or less appropriate, once the target domain changes.

This interpretation also fits the detailed plots in Chapter 4. In the natural-image setting, the heatmaps and severity plots emphasize broad sensitivity to noise-like and texture-related corruptions. In the medical-image setting, the more relevant failure modes are tied more closely to intensity shifts and medically meaningful visual variation. The difference is therefore not only quantitative, but also qualitative.

A useful way to read this contrast is to think of augmentation as introducing a prior about which variations the model should learn to ignore. In the natural-image setting, it is often helpful to reduce sensitivity to texture variation, corruption noise, or other low-level appearance changes. In the medical-image setting, however, some of these low-level patterns may themselves remain task-relevant. If an augmentation encourages invariance to exactly those features, its transfer to the medical domain may naturally be limited.

This helps explain why the results do not support a one-size-fits-all view of augmentation. Augmentation should therefore be selected with the target domain in mind. Rather than transferring the strongest method from a natural-image benchmark directly into a medical setting, it is more reasonable to ask whether the invariances induced by that method are compatible with the information structure of the downstream task.

The addition of ViT-B strengthens this conclusion rather than weakening it. Although ViT-B often raises the overall robustness level, it does not remove the domain gap. Instead, the ranking changes observed under ViT-B show that domain dependency remains visible even under a stronger architecture. This suggests that the domain gap is not just a property of one backbone, but may reflect a broader pattern in how augmentation interacts with the task and the data distribution.

5.4 Architecture Matters for Robustness Conclusions

The architecture comparison shows that the conclusions about augmentation are not fully architecture-invariant. In general, the ViT-B extension reaches a higher overall robustness level than the ResNet-18 benchmark, especially in the natural-image setting. But that does not mean the same augmentation ranking is preserved exactly when moving from one architecture to the other.

This point is already visible in the natural-image results. On CIFAR-10, StyleAug is strongest under ResNet-18, whereas RandAugment reaches the highest mCA under ViT-B. At the same time, CutMix becomes especially competitive under ViT-B once robustness and calibration are evaluated jointly. On CIFAR-100, AugMix remains very strong in both architectures, but the trade-off structure still changes because CutMix and Mixup become more attractive from the robustness–calibration perspective. These examples show that moving to a stronger backbone does not simply scale all augmentation methods upward by the same amount.

The medical-image results reveal the same phenomenon in a different form. On DermaMNIST, RandAugment remains highly robust in both architectures, but the best corruption calibration is not always achieved by the same method. Under

ResNet-18, CutMix provides the lowest mECE, whereas under ViT-B, StyleAug becomes the strongest calibration-oriented method. On PathMNIST, the strongest method changes more clearly between architectures, with RandAugment leading under ResNet-18 and AugMix leading under ViT-B. This means that architecture changes do not merely increase robustness; they can also change which augmentation family becomes most favourable.

A useful interpretation is that the backbone influences how augmentation-induced invariances are absorbed by the model. ResNet-18 and ViT-B do not process visual structure in exactly the same way, so the same augmentation policy can interact differently with the learned representation. In this sense, augmentation is not an isolated training trick. Its effect depends partly on the representational biases of the backbone to which it is applied.

The additional MCE analysis supports this point in a more subtle way. The architecture change does not only influence accuracy and average calibration, but in some settings also affects worst-case calibration behaviour. This suggests that architecture matters not only for how robust a model becomes, but also for how evenly its confidence quality is distributed across different confidence regions.

The addition of ViT-B therefore changes the discussion in an important way. In a ResNet-18-only study, the main question would simply be which augmentation performs best for a fixed backbone. Once the transformer extension is included, a second question becomes unavoidable: does the same augmentation remain strong when the model family changes? The results suggest that the answer is often only partial.

This has an important methodological implication. Conclusions about augmentation quality should not be drawn from a single backbone alone, especially when the goal is to make broader claims about robustness. A method that appears strongest under a CNN baseline may lose that advantage under a transformer-based model, while another method may become more attractive once calibration is taken into account. For this reason, architecture-aware evaluation should be viewed as an important part of robustness-oriented augmentation research rather than as an optional extension.

5.5 Practical Guidelines for Robust Augmentation Selection

The results of this thesis suggest that augmentation selection should not be treated as a search for one universally best method. Instead, the choice should be guided by the interaction between the target domain, the expected corruption structure, the calibration requirements, and the model architecture.

One practical guideline is that clean performance alone is not a sufficient selection criterion. Several methods remain close on clean accuracy or clean balanced accuracy, while their robustness under corruption differs much more strongly. Augmentation should therefore not be chosen only on the basis of clean validation results when corruption robustness is an important requirement.

Another practical guideline is that robustness and calibration should be optimized jointly. Across multiple settings in this thesis, the method with the highest mCA is not always the one with the lowest mECE. This means that a method should not automatically be preferred only because it achieves the strongest corruption robustness. If reliable confidence estimates are important, then calibration needs to be included explicitly in the evaluation protocol. The supplementary MCE results strengthen this point further by showing that even methods with reasonable average calibration may still differ in their worst-case confidence behaviour.

A third practical guideline is that augmentation choice should be domain-aware. In the natural-image setting, methods such as StyleAug, RandAugment, and AugMix often perform strongly because they align well with corruption structures that emphasize texture, blur, noise, and broad appearance variation. In the medical-image setting, however, the strongest methods shift, and intensity-related variation becomes more central. Augmentation should therefore be selected in relation to the information that the task actually depends on, rather than transferred directly from a benchmark in another domain.

A further practical guideline is that targeted invariance is often more useful than generic augmentation strength. Augmentation methods implicitly encode assumptions about which variations the model should learn to ignore. If these assumptions match the dominant corruption mechanisms of the target domain, the augmentation is more likely to help. If they suppress information that remains relevant for the task, the transfer of augmentation benefits may be limited. The key question is therefore not only whether an augmentation is strong, but whether it is strong in the right direction.

Finally, architecture-aware evaluation is necessary. The addition of ViT-B shows that augmentation rankings can change once the backbone changes. A method that is strongest under ResNet-18 is not guaranteed to remain strongest under ViT-B, and the robustness–calibration balance can change as well. When augmentation is selected for a real deployment setting, it should therefore ideally be evaluated on the model family that is actually intended for use.

These observations point to a more careful evaluation strategy for robust model design. Rather than asking which augmentation is best in general, it is more useful to ask which augmentation provides the best balance of clean performance, corruption robustness, calibration quality, domain fit, and architectural compatibility for the specific application.

From this perspective, the contribution of the thesis is not just the comparison itself, but also a clearer way to read that comparison. Robust augmentation selection is better understood as a context-dependent design decision than as a one-time benchmark ranking problem.

5.6 Limitations

Several limitations should be kept in mind when interpreting the results.

A first limitation concerns the architecture scope. The comparison between ResNet-18 and ViT-B provides a useful contrast between a convolutional backbone and a transformer-based backbone, but it does not cover the full range of architectures that are relevant in modern image classification. Deeper CNNs, alternative transformer variants, hybrid architectures, or models with substantially different capacity may interact with augmentation in different ways. The current architecture comparison therefore strengthens the analysis, but it should still be understood as a focused extension rather than a complete backbone study.

A second limitation concerns the robustness scope. The thesis focuses on common corruption benchmarks, namely CIFAR-C and MedMNIST-C, because they provide a structured and reproducible way to study distribution shift. However, corruption robustness is only one aspect of model reliability. The present study does not cover adversarial robustness, open-set recognition, label shift, or broader out-of-distribution scenarios beyond the selected corruption benchmarks. The conclusions should therefore be interpreted as statements about robustness under common corruptions rather than as a universal account of robustness in all deployment conditions.

A third limitation concerns the dataset scope, especially in the medical-image setting. The medical experiments are limited to DermaMNIST and PathMNIST. These two datasets provide a useful contrast within the medical domain, but they cannot represent all medical imaging tasks. Other settings such as radiology, volumetric imaging, multi-modal diagnosis, or higher-resolution pathology tasks may involve different corruption mechanisms and different relationships between robustness and calibration. As a result, the medical findings of this thesis should be interpreted as informative but not exhaustive.

A fourth limitation concerns the augmentation scope. Although the thesis evaluates a broad set of traditional and modern augmentation methods, the conclusions remain tied to the methods that were actually included in the benchmark. Other recent augmentation strategies, different parameterizations of the same methods, or combinations with post-hoc calibration techniques may show different behaviour. This is especially relevant for methods such as DiffuseMix, where calibration-aware post-processing could potentially change the practical interpretation of the results.

A fifth limitation concerns the metric level of analysis. The thesis emphasizes aggregated metrics such as clean accuracy, clean balanced accuracy, mCA, and mECE because these are useful for structured comparison across many settings. However, aggregated metrics cannot fully capture every practically relevant error pattern. For example, a method may look competitive on average while still failing badly on a small subset of corruption types or severity levels. The additional heatmaps, severity plots, trade-off figures, and the supplementary MCE view reduce this limitation to some extent, but they do not remove it completely.

A final limitation concerns the interpretive scope of the discussion. Several explanations proposed in this thesis, such as the role of domain-specific invariance or the interaction between augmentation and backbone bias, are supported by the empir-

ical results but remain interpretive rather than mechanistically proven. In other words, the thesis provides empirical evidence for certain patterns, but it does not fully identify the internal causal mechanisms by which all augmentation methods produce those patterns.

These limitations do not undermine the main contribution of the thesis, but they do define its scope more clearly. The results are best understood as a systematic benchmark-based comparison of augmentation effects on clean performance, corruption robustness, and calibration across selected datasets and architectures. That scope is broad enough to support meaningful conclusions while still leaving clear room for future work.

6 Conclusion and Future Work

6.1 Conclusion

This thesis examined how different data augmentation methods affect clean performance, robustness under corruption, and calibration in image classification. To study this in a structured way, a unified three-stage evaluation pipeline was applied across two natural-image datasets (CIFAR-10 and CIFAR-100) and two medical-image datasets (DermaMNIST and PathMNIST). ResNet-18 served as the main benchmark architecture, while ViT-B was included as an additional architecture to check whether the main findings remain similar across model families.

The results lead to four main conclusions. First, clean performance is informative, but it is not sufficient on its own. Many augmentation methods remain relatively close on clean evaluation, especially in the natural-image setting, while the differences become clearer once corruption is introduced. Second, robustness and calibration should be evaluated together. Across both domains, the method with the highest mCA is not always the one with the lowest mECE, and the supplementary MCE view shows that average calibration and worst-case calibration are related but not identical. Third, augmentation effects are domain-dependent. Methods that perform strongly on natural-image benchmarks do not transfer uniformly to the medical-image setting, where the relevant corruption sensitivities and the most useful augmentation families change. Fourth, the backbone architecture affects the overall conclusion as well. The addition of ViT-B shows that augmentation rankings are not fully preserved when the model family changes.

The main lesson is that augmentation should not be evaluated only by asking which method gives the best clean result. A more useful evaluation has to consider clean performance, corruption robustness, calibration quality, domain characteristics, and architecture choice together. In that sense, the contribution of the thesis lies not only in comparing augmentation methods, but also in showing how such comparisons can be interpreted in a more careful and practically meaningful way.

Taken together, these conclusions provide direct answers to the research questions introduced in Chapter 1. RQ1 and RQ2 are addressed through the Stage 2 com-

parison, which shows that clean predictive performance and clean calibration do not always favour the same augmentation methods. RQ3 is addressed through the Stage 3 robustness evaluation on CIFAR-C and MedMNIST-C, where the methods separate more clearly under corruption. RQ4 is answered by the cross-domain comparison, which shows that augmentation effects are not fully preserved between natural-image and medical-image settings. RQ5 is addressed through the ResNet-18 and ViT-B comparison, which shows that augmentation rankings are not fully architecture-invariant.

6.2 Future Work

The findings of this thesis also point to several clear directions for future work. One direction is to extend the architecture comparison. The present study already shows that augmentation conclusions are not fully architecture-invariant, so it would be valuable to include additional CNNs, transformer variants, or hybrid architectures in order to see how broadly these ranking shifts generalize. A second direction is to develop more domain-aware augmentation strategies. The results suggest that augmentation tends to be more effective when the induced invariances match the target domain, which is especially relevant for medical-image settings where intensity changes and domain-specific structure matter more strongly.

A third direction is to investigate joint robustness–calibration optimization more directly. Since the experiments show that robustness and calibration are related but not identical, future work could look at training strategies that address both objectives at once, for example through calibration-aware losses, post-hoc calibration methods, or augmentation selection procedures that explicitly include confidence reliability under corruption. A fourth direction is to move beyond the current corruption benchmarks. CIFAR-C and MedMNIST-C are structured and useful test beds, but they do not cover every practically relevant type of distribution shift. It would therefore be worthwhile to check whether the findings remain stable under broader out-of-distribution settings, more realistic real-world shifts, or additional medical-imaging tasks.

Finally, more fine-grained failure analysis would be valuable. Although this thesis already includes corruption-wise heatmaps, severity plots, and robustness–calibration trade-off analysis, there is still room for a closer examination of which corruption families and which visual patterns are most responsible for the observed ranking changes. That kind of analysis could help move the discussion from empirical comparison toward a more mechanistic understanding of why certain augmentation methods remain strong in some settings but not in others.

Taken together, these directions point to a broader shift in how augmentation research should be done. Rather than continuing to search for one method that wins everywhere, future work will likely be more useful if it explains how augmentation interacts with domain characteristics, model architecture, and evaluation objectives.

Data and Code Availability

The datasets and corruption benchmarks used in this thesis are publicly available through their respective official sources and are cited in the main text, including CIFAR-10, CIFAR-100, CIFAR-C, MedMNIST, and MedMNIST-C. Their use remains subject to the original licenses and access conditions of the corresponding benchmark providers.

To support transparency and reproducibility, the experimental code and analysis scripts used in this thesis are available in the following public GitHub repository:

`https://github.com/Junsche/xai_project`

At submission, a corresponding resource containing the developed code base and associated digital attachments is provided alongside the digital thesis.

A Experimental Configuration Details

This appendix reports implementation details that are necessary for reproducibility but would interrupt the flow of the main argument in Chapter 4.

A.1 Software and Hardware Environment

Table 8: Experimental environment specifications.

Component	**Specification**
Hardware	
GPU	NVIDIA RTX A5000 (24GB VRAM)
CPU	Intel Xeon Scalable Processors
RAM	64 GB System Memory
Software (Core Libraries)	
Operating System	Linux (Ubuntu 22.04 LTS)
Python	3.11.14
PyTorch	2.5.1
CUDA / cuDNN	12.1 / 9.1.0
Torchvision	0.20.1
MedMNIST	3.0.1
MedMNIST-C	0.1.0
Software (Data & Analysis)	
NumPy	2.2.6
Pandas	2.3.3
Scikit-Learn	1.7.2
Matplotlib	3.10.7
WandB	0.22.3

A.2 Learning-Rate Search Space and Final Selections

The Stage 1 learning-rate search is performed separately for each dataset and architecture. The candidate sets and final selections are listed below. The selected values are then reused consistently in Stage 2 and Stage 3.

A.3 Stage-Specific Training Budgets

To keep the comparison practical while still informative, the epoch budgets differ by stage, dataset, and backbone role.

The ViT-B Stage 1 schedule intentionally uses a longer search budget on CIFAR-100 than on CIFAR-10. This setting follows the experiment protocol used to generate the reported results.

Table 9: Stage 1 learning-rate search space and final selected learning rates.

Backbone	Dataset	Candidate LRs	Selected LR
ResNet-18	CIFAR-10	{0.1, 0.05, 0.01, 0.001}	0.01
ResNet-18	CIFAR-100	{0.1, 0.05, 0.01, 0.001}	0.001
ResNet-18	DermaMNIST	{0.01, 0.005, 0.001, 0.0005}	0.005
ResNet-18	PathMNIST	{0.01, 0.005, 0.001, 0.0005}	0.001
ViT-B	CIFAR-10	{0.001, 0.0005, 0.0001, 0.00005}	0.0005
ViT-B	CIFAR-100	{0.001, 0.0005, 0.0001, 0.00005}	0.0005
ViT-B	DermaMNIST	{0.001, 0.0005, 0.0001, 0.00005}	0.001
ViT-B	PathMNIST	{0.001, 0.0005, 0.0001, 0.00005}	0.001

Table 10: Stage-specific epoch budgets used in this thesis.

Stage	Backbone	Dataset	Epochs
Stage 1	ResNet-18	CIFAR-10 / CIFAR-100	20
Stage 1	ResNet-18	DermaMNIST / PathMNIST	15
Stage 1	ViT-B	CIFAR-10	15
Stage 1	ViT-B	CIFAR-100	20
Stage 1	ViT-B	DermaMNIST / PathMNIST	10
Stage 2	ResNet-18	CIFAR-10 / CIFAR-100	100
Stage 2	ResNet-18	DermaMNIST / PathMNIST	50
Stage 2	ViT-B	All datasets	30

A.4 Augmentation Hyperparameters

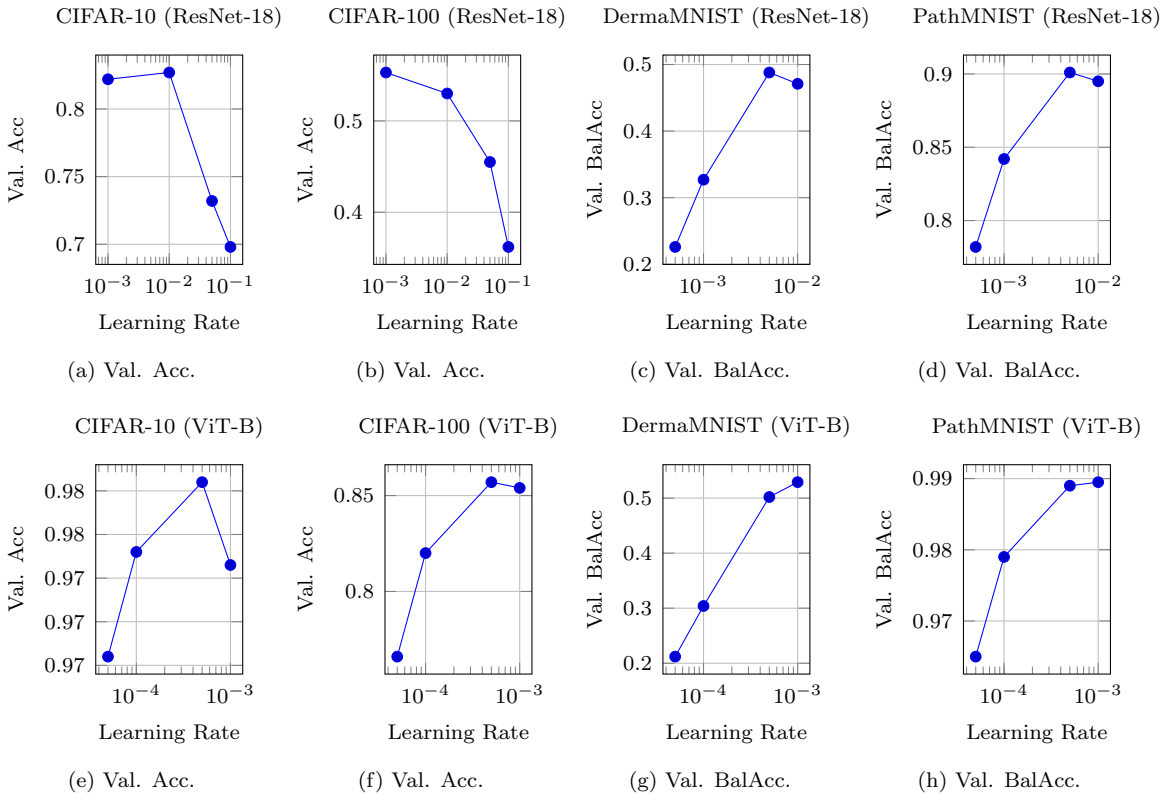
Table 11: Hyperparameter configurations for data augmentation methods.

Method	Parameter	Value
Baseline	Crop Size	32×32 (CIFAR), 28×28 (MedMNIST)
	Padding	4 pixels
	Horizontal Flip	$p = 0.5$
Rotation + Random Erasing	Probability (p)	0.5
	Scale Range	$(0.02, 0.33)$
	Aspect Ratio	$(0.3, 3.3)$
Mixup	Alpha (α)	0.2
CutMix	Alpha (α)	1.0
	Probability	0.5
AutoAugment	Policy	CIFAR-10 policy
RandAugment	Number of Ops (N)	2
	Magnitude (M)	9
AugMix	Severity / Width / Depth / Alpha	3 / 3 / random / 1.0
StyleAug	Style Model	Pre-trained style-transfer network
DiffuseMix	Diffusion Setting	Default implementation setting

For methods with external implementations (for example, StyleAug and DiffuseMix), the project uses one fixed implementation configuration across all datasets and backbones in order to keep the comparison controlled.

A.5 Full Backbone-Specific Stage 1 Curves

The main text reports only a compact dataset-level overview of the Stage 1 search behaviour. For completeness, the full backbone-specific curves are collected here in one place.



(a) Val. Acc. (b) Val. Acc. (c) Val. BalAcc. (d) Val. BalAcc.

(e) Val. Acc. (f) Val. Acc. (g) Val. BalAcc. (h) Val. BalAcc.

Figure 11: Full Stage 1 learning-rate search curves for each dataset–backbone combination.

B Corruption Taxonomy and Evaluation Scope

This appendix documents the corruption protocol used in Stage 3. The goal is to make the evaluation scope explicit without repeating the interpretation from the main chapters.

B.1 Natural-Image Corruption Benchmark

Table 12: Corruption taxonomy used for CIFAR-C evaluation (19 corruption types).

Family	**Corruption Types**
Noise	Gaussian Noise, Shot Noise, Impulse Noise, Speckle Noise
Blur	Defocus Blur, Glass Blur, Motion Blur, Zoom Blur, Gaussian Blur
Weather	Snow, Frost, Fog, Brightness, Spatter
Digital	Contrast, Elastic Transform, Pixelate, JPEG Compression, Saturate

B.2 Medical-Image Corruption Benchmark

Table 13: Shared corruption subset used for MedMNIST-C evaluation (8 corruption types).

Family	**Corruption Types**
Blur & Artifacts	Defocus Blur, Motion Blur, JPEG Compression, Pixelate
Intensity-related	Brightness Up, Brightness Down, Contrast Up, Contrast Down

C Additional Results

This appendix provides supplementary tables and selected figures that support Chapter 4. The main text keeps only the central summary displays, while dataset-specific details and secondary visualizations are collected here.

C.1 Additional Stage 2 Clean Results

This subsection reports the full Stage 2 clean summaries that are referenced but not shown in full in the main text.

Table 14: Clean-data performance and calibration on CIFAR-10 in Stage 2.

Backbone	**Augmentation**	**Acc**	**ECE**	**Loss**
resnet18	augmix	0.8441	0.0636	0.5172
resnet18	autoaugment	0.8540	0.0347	0.4362
resnet18	baseline	0.8461	0.0739	0.5380
resnet18	cutmix	0.8204	0.1172	0.6178
resnet18	diffusemix	0.8241	0.0757	0.5946
resnet18	mixup	0.8426	0.0336	0.4985
resnet18	randaugment	0.8536	0.0340	0.4415
resnet18	rotation_erasing	0.8457	0.0481	0.4891
resnet18	styleaug	0.8486	0.0477	0.4782
vit_b	augmix	0.9692	0.0161	0.1021
vit_b	autoaugment	0.9801	0.0071	0.0591
vit_b	baseline	0.9784	0.0132	0.0835
vit_b	cutmix	0.9829	0.0051	0.0577
vit_b	diffusemix	0.9737	0.0148	0.0968
vit_b	mixup	0.9805	0.0151	0.0804
vit_b	randaugment	0.9809	0.0093	0.0676
vit_b	rotation_erasing	0.9813	0.0071	0.0567
vit_b	styleaug	0.9805	0.0098	0.0640

Table 15: Clean-data performance and calibration on CIFAR-100 in Stage 2.

Backbone	Augmentation	Acc	ECE	Loss
resnet18	augmix	0.6002	0.1862	1.8916
resnet18	autoaugment	0.6017	0.1509	1.6708
resnet18	baseline	0.5875	0.2225	2.1212
resnet18	cutmix	0.5997	0.0869	1.5403
resnet18	diffusemix	0.5456	0.2306	2.3498
resnet18	mixup	0.6074	0.0612	1.6046
resnet18	randaugment	0.6006	0.1640	1.7400
resnet18	rotation_erasing	0.5597	0.1910	2.0308
resnet18	styleaug	0.5821	0.1785	1.8837
vit_b	augmix	0.8786	0.0317	0.4039
vit_b	autoaugment	0.8791	0.0176	0.3992
vit_b	baseline	0.8730	0.0478	0.4712
vit_b	cutmix	0.8853	0.0136	0.3725
vit_b	diffusemix	0.8587	0.0447	0.5030
vit_b	mixup	0.8839	0.0133	0.4100
vit_b	randaugment	0.8770	0.0315	0.4125
vit_b	rotation_erasing	0.8817	0.0233	0.3933
vit_b	styleaug	0.8808	0.0265	0.4024

Table 16: Clean-data performance and calibration on DermaMNIST in Stage 2.

Backbone	Augmentation	BalAcc	ECE	Loss
resnet18	augmix	0.7328	0.1329	0.9208
resnet18	autoaugment	0.7328	0.0943	0.7692
resnet18	baseline	0.7478	0.1785	1.2222
resnet18	cutmix	0.7627	0.0566	0.6923
resnet18	diffusemix	0.7368	0.1628	1.0899
resnet18	mixup	0.7507	0.0647	0.7112
resnet18	randaugment	0.7398	0.0731	0.7646
resnet18	rotation_erasing	0.7398	0.1199	0.8451
resnet18	styleaug	0.7029	0.0950	0.8457
vit_b	augmix	0.7856	0.0771	0.6237
vit_b	autoaugment	0.7787	0.0354	0.6060
vit_b	baseline	0.7866	0.1272	0.8016
vit_b	cutmix	0.7936	0.0270	0.5728
vit_b	diffusemix	0.7767	0.0922	0.6578
vit_b	mixup	0.7886	0.0402	0.6059
vit_b	randaugment	0.7866	0.0335	0.5932
vit_b	rotation_erasing	0.7936	0.0622	0.5940
vit_b	styleaug	0.7817	0.0382	0.6349

Table 17: Clean-data performance and calibration on PathMNIST in Stage 2.

Backbone	Augmentation	BalAcc	ECE	Loss
resnet18	augmix	0.8548	0.0443	0.4380
resnet18	autoaugment	0.8266	0.0470	0.5179
resnet18	baseline	0.8931	0.0564	0.3978
resnet18	cutmix	0.9207	0.1294	0.3426
resnet18	diffusemix	0.8134	0.1022	0.6512
resnet18	mixup	0.9144	0.0358	0.2834
resnet18	randaugment	0.8893	0.0201	0.3057
resnet18	rotation_erasing	0.8743	0.0519	0.3904
resnet18	styleaug	0.7551	0.0939	0.7572
vit_b	augmix	0.9927	0.0014	0.0221
vit_b	autoaugment	0.9920	0.0022	0.0290
vit_b	baseline	0.9927	0.0037	0.0265
vit_b	cutmix	0.9968	0.0040	0.0142
vit_b	diffusemix	0.9788	0.0078	0.0673
vit_b	mixup	0.9962	0.0199	0.0331
vit_b	randaugment	0.9906	0.0030	0.0292
vit_b	rotation_erasing	0.9945	0.0021	0.0154
vit_b	styleaug	0.9893	0.0029	0.0366

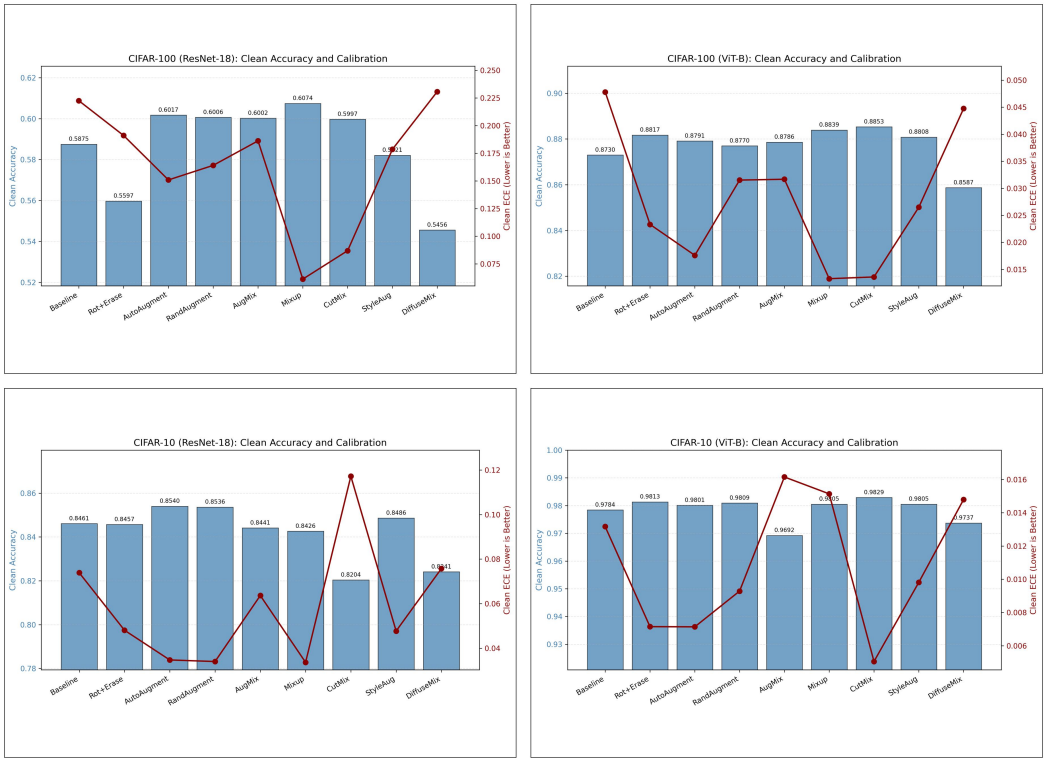


Figure 12: Supplementary Stage 2 dual-axis clean overview across datasets and backbones.

Table 18: Cross-architecture clean comparison in Stage 2.

Dataset	Augmentation	R18 Acc	ViT-B Acc	R18 ECE	ViT-B ECE
cifar10	augmix	0.8441	0.9692	0.0636	0.0161
cifar10	autoaugment	0.8540	0.9801	0.0347	0.0071
cifar10	baseline	0.8461	0.9784	0.0739	0.0132
cifar10	cutmix	0.8204	0.9829	0.1172	0.0051
cifar10	diffusemix	0.8241	0.9737	0.0757	0.0148
cifar10	mixup	0.8426	0.9805	0.0336	0.0151
cifar10	randaugment	0.8536	0.9809	0.0340	0.0093
cifar10	rotation_erasing	0.8457	0.9813	0.0481	0.0071
cifar10	styleaug	0.8486	0.9805	0.0477	0.0098
cifar100	augmix	0.6002	0.8786	0.1862	0.0317
cifar100	autoaugment	0.6017	0.8791	0.1509	0.0176
cifar100	baseline	0.5875	0.8730	0.2225	0.0478
cifar100	cutmix	0.5997	0.8853	0.0869	0.0136
cifar100	diffusemix	0.5456	0.8587	0.2306	0.0447
cifar100	mixup	0.6074	0.8839	0.0612	0.0133
cifar100	randaugment	0.6006	0.8770	0.1640	0.0315
cifar100	rotation_erasing	0.5597	0.8817	0.1910	0.0233
cifar100	styleaug	0.5821	0.8808	0.1785	0.0265
dermamnist	augmix	0.7328	0.7856	0.1329	0.0771
dermamnist	autoaugment	0.7328	0.7787	0.0943	0.0354
dermamnist	baseline	0.7478	0.7866	0.1785	0.1272
dermamnist	cutmix	0.7627	0.7936	0.0566	0.0270
dermamnist	diffusemix	0.7368	0.7767	0.1628	0.0922
dermamnist	mixup	0.7507	0.7886	0.0647	0.0402
dermamnist	randaugment	0.7398	0.7866	0.0731	0.0335
dermamnist	rotation_erasing	0.7398	0.7936	0.1199	0.0622
dermamnist	styleaug	0.7029	0.7817	0.0950	0.0382
pathmnist	augmix	0.8548	0.9927	0.0443	0.0014
pathmnist	autoaugment	0.8266	0.9920	0.0470	0.0022
pathmnist	baseline	0.8931	0.9927	0.0564	0.0037
pathmnist	cutmix	0.9207	0.9968	0.1294	0.0040
pathmnist	diffusemix	0.8134	0.9788	0.1022	0.0078
pathmnist	mixup	0.9144	0.9962	0.0358	0.0199
pathmnist	randaugment	0.8893	0.9906	0.0201	0.0030
pathmnist	rotation_erasing	0.8743	0.9945	0.0519	0.0021
pathmnist	styleaug	0.7551	0.9893	0.0939	0.0029

C.2 Additional Stage 3 Natural-Image Results

This subsection provides the complete natural-image robustness tables together with selected split CIFAR-100 and trade-off supplements.

Table 19: Robustness summary on CIFAR-10 using ResNet-18. Clean performance and corruption-averaged metrics are reported jointly.

Augmentation	N Corr	Clean Acc.	Clean ECE	Clean MCE	mCA	mECE	mMCE	mLoss
Baseline	95	0.8461	0.0739	0.2205	0.7128	0.1629	0.3247	1.1492
AutoAugment	95	0.8540	0.0347	0.1493	0.7455	0.0888	0.2354	0.8183
RandAugment	95	0.8536	0.0340	0.1833	0.7600	0.0745	0.2040	0.7713
AugMix	95	0.8441	0.0636	0.2426	0.7598	0.1100	0.2487	0.8245
Mixup	95	0.8426	0.0336	0.1105	0.7104	0.1026	0.2352	0.9474
CutMix	95	0.8204	0.1172	0.2149	0.6691	0.0823	0.1769	1.0211
StyleAug	95	0.8486	0.0477	0.1384	0.7815	0.0802	0.2103	0.7150
DiffuseMix	95	0.8241	0.0757	0.2406	0.7653	0.1088	0.2708	0.8265
Rotation + Random Erasing	95	0.8457	0.0481	0.1499	0.7265	0.1173	0.2532	0.9641

Table 20: Robustness summary on CIFAR-10 using ViT-B. Clean performance and corruption-averaged metrics are reported jointly.

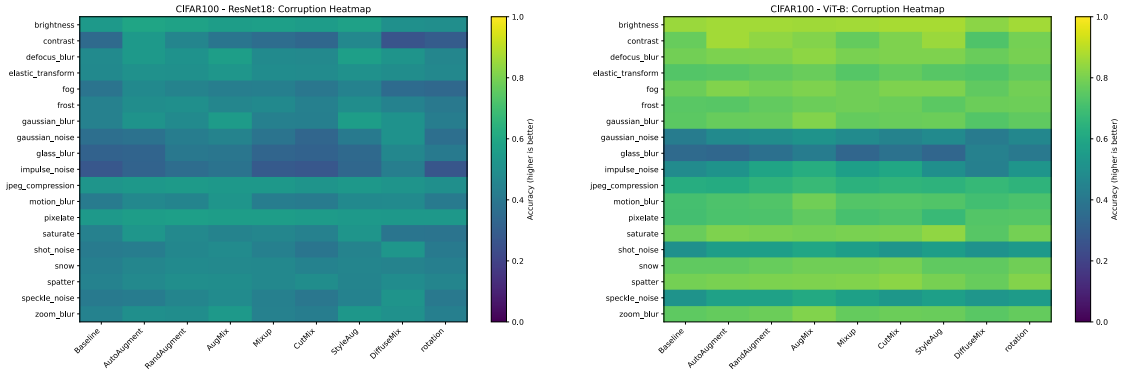
Augmentation	N Corr	Clean Acc.	Clean ECE	Clean MCE	mCA	mECE	mMCE	mLoss
Baseline	95	0.9784	0.0132	0.4483	0.8914	0.0718	0.3799	0.4770
AutoAugment	95	0.9801	0.0071	0.7173	0.8971	0.0502	0.3993	0.3633
RandAugment	95	0.9809	0.0093	0.6986	0.9092	0.0522	0.3891	0.3508
AugMix	95	0.9692	0.0161	0.3665	0.8842	0.0646	0.3293	0.4346
Mixup	95	0.9805	0.0151	0.6853	0.9028	0.0445	0.3390	0.3459
CutMix	95	0.9829	0.0051	0.6307	0.9038	0.0315	0.2995	0.3161
StyleAug	95	0.9805	0.0098	0.8295	0.8955	0.0584	0.3335	0.3966
DiffuseMix	95	0.9737	0.0148	0.3357	0.9005	0.0608	0.3628	0.4023
Rotation + Random Erasing	95	0.9813	0.0071	0.6947	0.9019	0.0479	0.3185	0.3428

Table 21: Robustness summary on CIFAR-100 using ResNet-18. Clean performance and corruption-averaged metrics are reported jointly.

Augmentation	N Corr	Clean Acc.	Clean ECE	Clean MCE	mCA	mECE	mMCE	mLoss
Baseline	95	0.5875	0.2225	0.4195	0.4266	0.3155	0.4904	3.3764
AutoAugment	95	0.6017	0.1509	0.2678	0.4746	0.1997	0.3362	2.4200
RandAugment	95	0.6006	0.1640	0.3071	0.4770	0.2175	0.3560	2.4928
AugMix	95	0.6002	0.1862	0.3372	0.4846	0.2410	0.3900	2.5988
Mixup	95	0.6074	0.0612	0.1225	0.4443	0.1247	0.2215	2.4932
CutMix	95	0.5997	0.0869	0.1698	0.4259	0.0602	0.1446	2.3723
StyleAug	95	0.5821	0.1785	0.3134	0.4778	0.2305	0.3728	2.5646
DiffuseMix	95	0.5456	0.2306	0.3988	0.4673	0.2789	0.4406	2.9874
Rotation + Random Erasing	95	0.5597	0.1910	0.3260	0.4145	0.2731	0.4212	3.0858

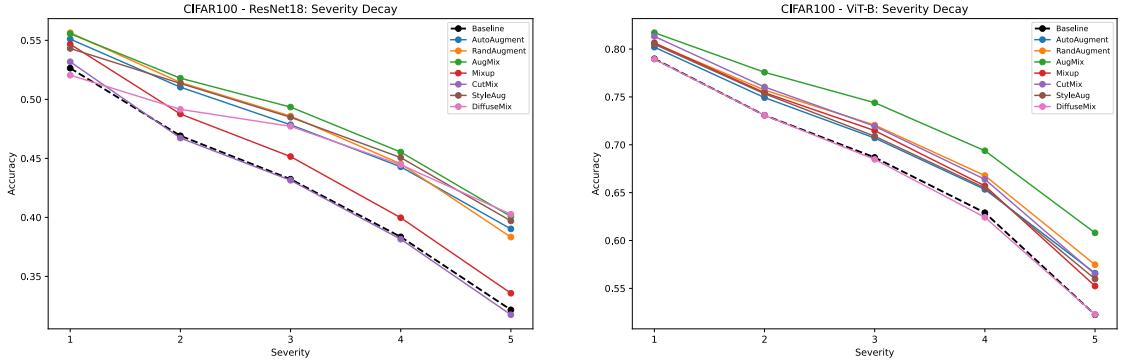
Table 22: Robustness summary on CIFAR-100 using ViT-B. Clean performance and corruption-averaged metrics are reported jointly.

Augmentation	N Corr	Clean Acc.	Clean ECE	Clean MCE	mCA	mECE	mMCE	mLoss
Baseline	95	0.8730	0.0478	0.2050	0.6718	0.1370	0.3041	1.5135
AutoAugment	95	0.8791	0.0176	0.1078	0.6955	0.0892	0.2006	1.2672
RandAugment	95	0.8770	0.0315	0.1356	0.7052	0.1030	0.2369	1.2502
AugMix	95	0.8786	0.0317	0.1446	0.7277	0.0964	0.2324	1.1399
Mixup	95	0.8839	0.0133	0.2251	0.6970	0.0608	0.1809	1.2257
CutMix	95	0.8853	0.0136	0.0983	0.7044	0.0590	0.1624	1.1569
StyleAug	95	0.8808	0.0265	0.1186	0.6964	0.0979	0.2199	1.2866
DiffuseMix	95	0.8587	0.0447	0.2844	0.6704	0.1297	0.2931	1.4744
Rotation + Random Erasing	95	0.8817	0.0233	0.2097	0.6979	0.0879	0.1984	1.2585



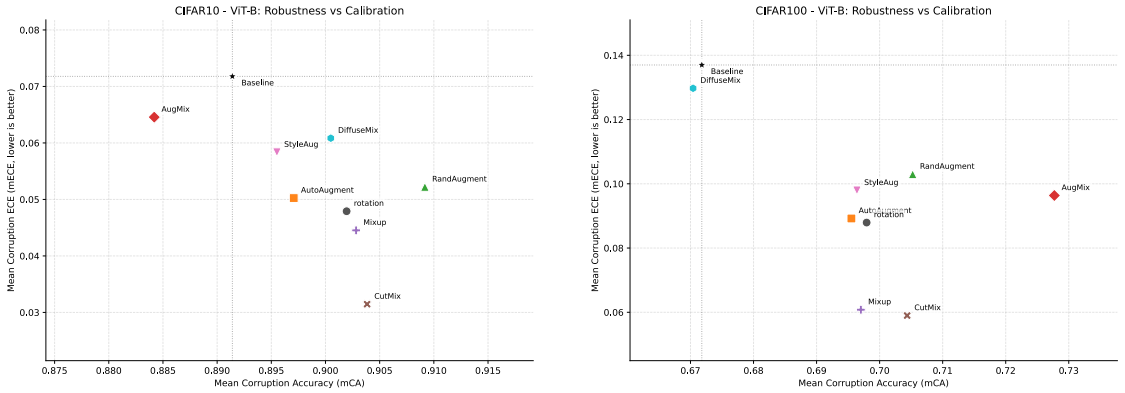
(a) ResNet-18 (b) ViT-B

Figure 13: Supplementary CIFAR-100 corruption-wise heatmaps for ResNet-18 and ViT-B.



(a) ResNet-18 (b) ViT-B

Figure 14: Supplementary CIFAR-100 severity-wise robustness curves for ResNet-18 and ViT-B.



(a) CIFAR-10 / ViT-B (b) CIFAR-100 / ViT-B

Figure 15: Supplementary natural-domain trade-off examples (ViT-B).

C.3 Additional Stage 3 Medical-Image Results

This subsection provides the complete medical-image robustness tables together with selected split PathMNIST and trade-off supplements.

Table 23: Robustness summary on DermaMNIST using ResNet-18. Clean performance and corruption-averaged metrics are reported jointly for the medical-image domain.

Augmentation	N Corr	Clean Acc.	Clean BalAcc.	Clean ECE	Clean MCE	mCA	mBalAcc.	mECE	mMCE	mLoss
Baseline	40	0.7471	0.5406	0.1761	0.4114	0.7124	0.4467	0.2199	0.4482	1.7868
AutoAugment	40	0.7297	0.4206	0.1006	0.2661	0.7307	0.4214	0.1263	0.3113	0.8872
RandAugment	40	0.7426	0.4571	0.0630	0.2627	0.7334	0.4421	0.1276	0.3392	0.9213
AugMix	40	0.7137	0.4642	0.1504	0.7476	0.7154	0.4640	0.1648	0.3632	1.1203
Mixup	40	0.7676	0.4874	0.0597	0.2321	0.7251	0.3965	0.1240	0.3234	0.9889
CutMix	40	0.7571	0.3940	0.0484	0.0871	0.7087	0.3681	0.0714	0.2322	0.8365
StyleAug	40	0.7147	0.4198	0.0889	0.2123	0.7237	0.3784	0.1414	0.3578	1.0298
DiffuseMix	40	0.7212	0.4335	0.1779	0.3540	0.7114	0.3897	0.2187	0.4097	1.7867
Rotation + Random Erasing	40	0.7506	0.5362	0.1104	0.3239	0.7093	0.4742	0.1793	0.3861	1.2932

Table 24: Robustness summary on DermaMNIST using ViT-B. Clean performance and corruption-averaged metrics are reported jointly for the medical-image domain.

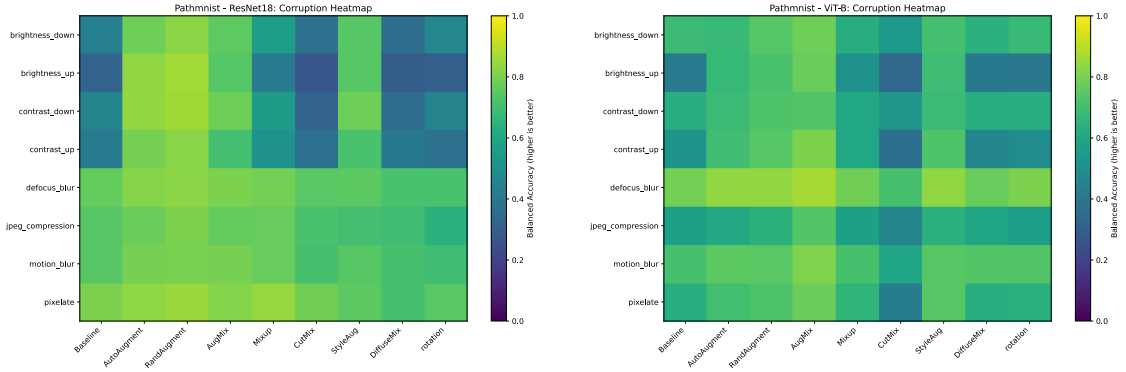
Augmentation	N Corr	Clean Acc.	Clean BalAcc.	Clean ECE	Clean MCE	mCA	mBalAcc.	mECE	mMCE	mLoss
Baseline	40	0.7900	0.6159	0.1172	0.6800	0.7122	0.4444	0.1423	0.3138	1.0302
AutoAugment	40	0.7786	0.5540	0.0315	0.2629	0.7334	0.4871	0.0422	0.1779	0.7239
RandAugment	40	0.7960	0.6157	0.0328	0.1014	0.7342	0.4949	0.0387	0.2087	0.7544
AugMix	40	0.7915	0.5968	0.0672	0.2599	0.7110	0.5010	0.0630	0.2189	0.8361
Mixup	40	0.7855	0.5758	0.0379	0.2518	0.7124	0.4021	0.0548	0.1809	0.8281
CutMix	40	0.7855	0.5597	0.0237	0.2401	0.7008	0.4055	0.0659	0.1676	0.8693
StyleAug	40	0.7776	0.5506	0.0334	0.2638	0.7224	0.4845	0.0372	0.2020	0.7464
DiffuseMix	40	0.7830	0.5995	0.0868	0.2496	0.7068	0.4589	0.1130	0.2750	0.9260
Rotation + Random Erasing	40	0.8000	0.5970	0.0520	0.2563	0.6995	0.3500	0.0855	0.2489	0.9005

Table 25: Robustness summary on PathMNIST using ResNet-18. Clean performance and corruption-averaged metrics are reported jointly for the medical-image domain.

Augmentation	N Corr	Clean Acc.	Clean BalAcc.	Clean ECE	Clean MCE	mCA	mBalAcc.	mECE	mMCE	mLoss
Baseline	40	0.8453	0.7997	0.0981	0.3251	0.5983	0.5829	0.3505	0.5446	5.3706
AutoAugment	40	0.7967	0.7592	0.0793	0.1766	0.8344	0.8083	0.0788	0.2892	0.5732
RandAugment	40	0.8604	0.8350	0.0522	0.2500	0.8581	0.8302	0.0730	0.3198	0.4882
AugMix	40	0.7804	0.7718	0.1145	0.2553	0.7848	0.7676	0.1302	0.2886	0.8963
Mixup	40	0.8571	0.8286	0.0287	0.1511	0.6655	0.6505	0.1362	0.3145	1.1981
CutMix	40	0.8125	0.7913	0.0630	0.2517	0.5383	0.5358	0.2187	0.4169	2.0653
StyleAug	40	0.7358	0.7034	0.1197	0.2312	0.7560	0.7397	0.1380	0.3567	0.8999
DiffuseMix	40	0.7560	0.7317	0.1559	0.3424	0.5381	0.5292	0.3966	0.6019	6.4197
Rotation + Random Erasing	40	0.7624	0.7264	0.1394	0.2747	0.5503	0.5477	0.3707	0.5139	4.1519

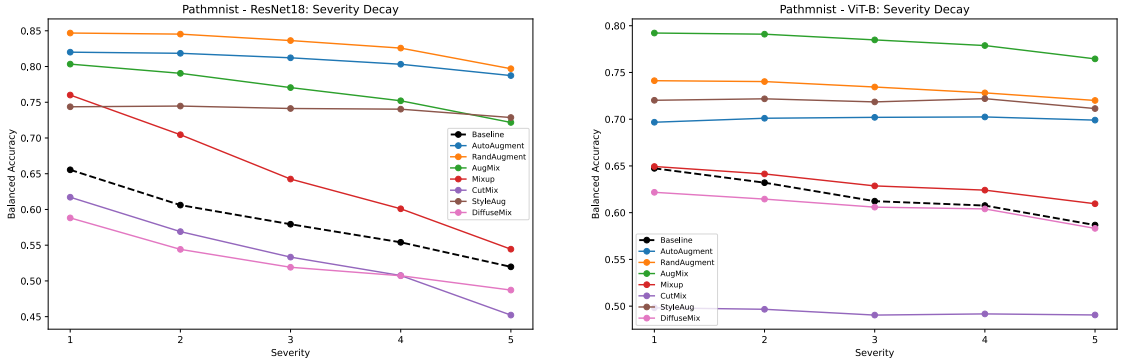
Table 26: Robustness summary on PathMNIST using ViT-B. Clean performance and corruption-averaged metrics are reported jointly for the medical-image domain.

Augmentation	N Corr	Clean Acc.	Clean BalAcc.	Clean ECE	Clean MCE	mCA	mBalAcc.	mECE	mMCE	mLoss
Baseline	40	0.9209	0.8994	0.0527	0.3073	0.5781	0.6172	0.2502	0.3771	1.7324
AutoAugment	40	0.9369	0.9120	0.0326	0.2276	0.7139	0.7002	0.1850	0.3664	1.2864
RandAugment	40	0.9465	0.9255	0.0256	0.1682	0.7266	0.7328	0.1765	0.3253	1.2688
AugMix	40	0.9361	0.9151	0.0362	0.3257	0.7762	0.7823	0.1219	0.2756	0.9180
Mixup	40	0.9343	0.9153	0.0058	0.2586	0.5951	0.6306	0.1456	0.2953	1.3029
CutMix	40	0.9130	0.8983	0.0513	0.2834	0.4286	0.4935	0.4375	0.6108	2.7848
StyleAug	40	0.9515	0.9327	0.0247	0.3012	0.7099	0.7188	0.1933	0.3678	1.3821
DiffuseMix	40	0.8897	0.8662	0.0668	0.7467	0.5670	0.6059	0.2610	0.4483	1.8288
Rotation + Random Erasing	40	0.9407	0.9225	0.0317	0.3011	0.5691	0.6133	0.2338	0.4091	1.7430



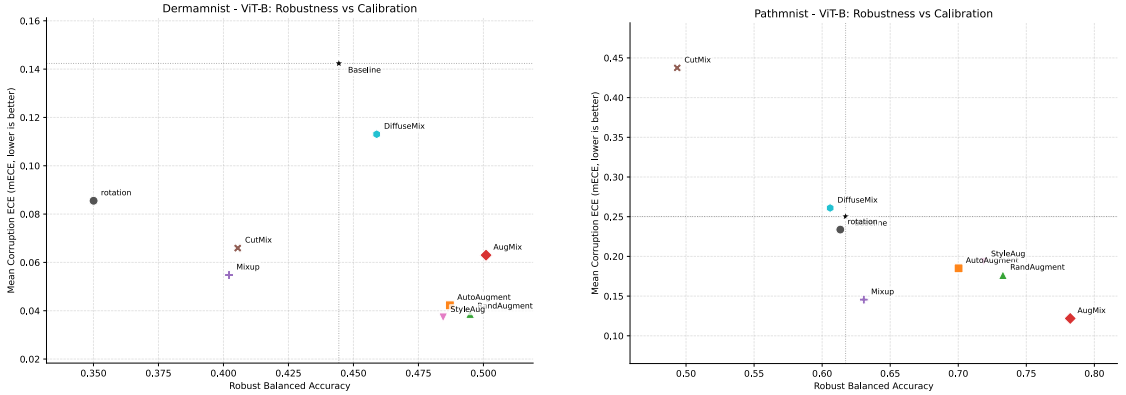
(a) ResNet-18 (b) ViT-B

Figure 16: Supplementary PathMNIST corruption-wise heatmaps for ResNet-18 and ViT-B.



(a) ResNet-18 (b) ViT-B

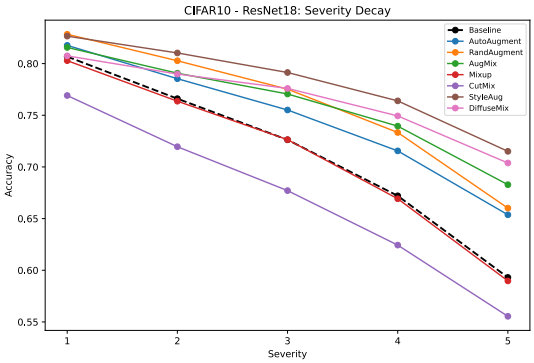
Figure 17: Supplementary PathMNIST severity-wise robustness curves for ResNet-18 and ViT-B.



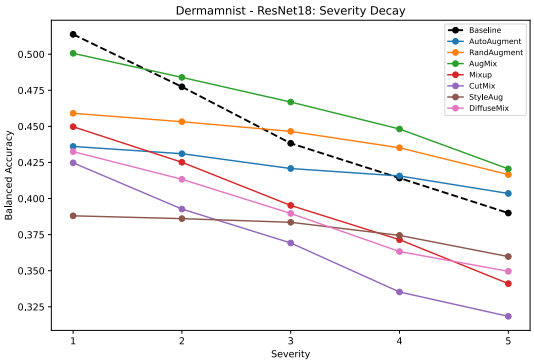
(a) DermaMNIST / ViT-B (b) PathMNIST / ViT-B

Figure 18: Supplementary medical-domain trade-off examples (ViT-B).

C.4 Additional Split Figure Examples

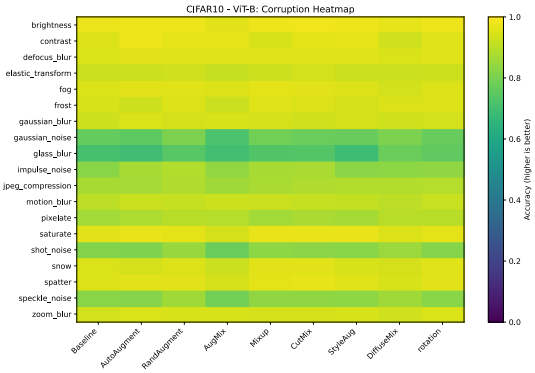


(a) CIFAR-10 / ResNet-18 severity

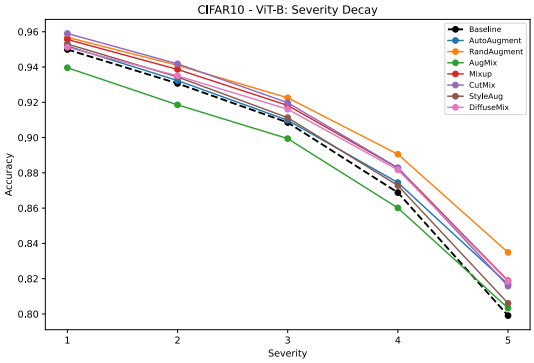


(b) DermaMNIST / ResNet-18 severity

Figure 19: Supplementary ResNet-18 severity-wise robustness curves referenced from the main text.

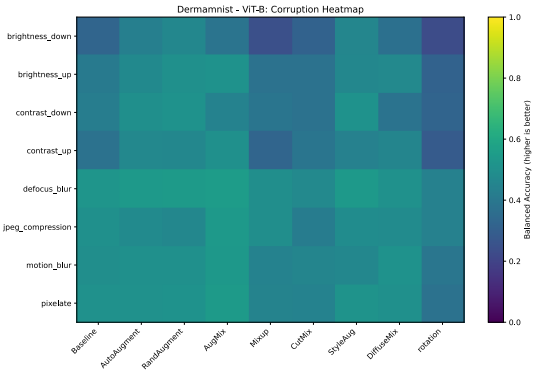


(a) CIFAR-10 / ViT-B heatmap

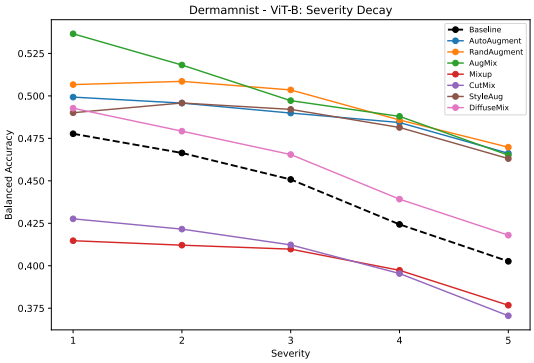


(b) CIFAR-10 / ViT-B severity

Figure 20: Supplementary CIFAR-10 ViT-B detailed plots referenced from the main text.

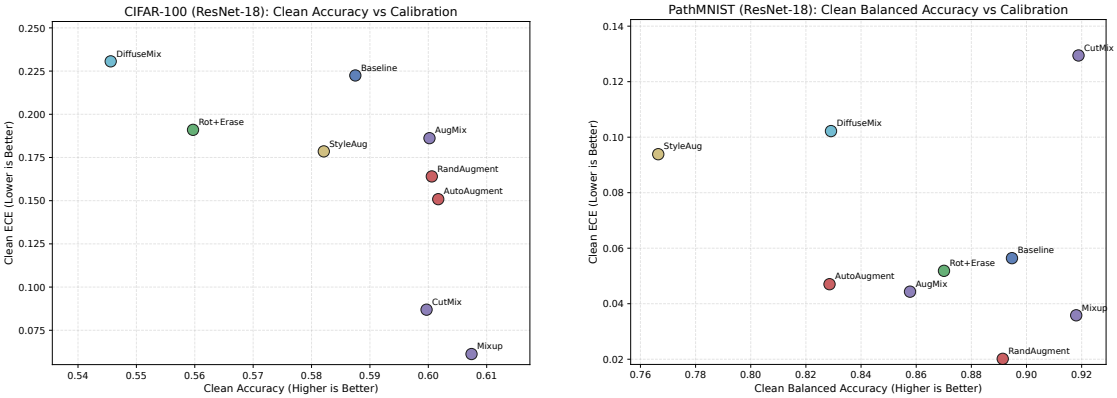


(a) DermaMNIST / ViT-B heatmap



(b) DermaMNIST / ViT-B severity

Figure 21: Supplementary DermaMNIST ViT-B detailed plots referenced from the main text.



(a) CIFAR-100 / ResNet-18 (b) PathMNIST / ResNet-18

Figure 22: Supplementary Stage 2 clean trade-off examples referenced from the main text.

Bibliography

[1] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[3] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[4] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019.

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 702–703, 2020.

[7] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[8] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

[9] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020.

[10] Philip T Jackson, Amir Atapour-Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: Data augmentation via style randomization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 83–92, 2019.

[11] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27621–27630, 2024.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[13] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A W M van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[14] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[15] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[16] 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)*, pages 770–778, 2016.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[18] Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 26831–26843, 2021.

[19] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 27378–27394, 2022.

[20] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

[21] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

[22] Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig. Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. *arXiv preprint arXiv:2406.17536*, 2024.

[23] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10 (1):41, 2023.

[24] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 15682–15694, 2021.

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021.

[26] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research (TMLR)*, 2022.

[27] Zhiying Lu, Hongtao Xie, Chuanbin Liu, and Yongdong Zhang. Bridging the gap between vision transformers and convolutional neural networks on small datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 14698–14711, 2022.

[28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[29] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[30] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Bamberg 17.04.2026

Place, Date

Signature