Bern University
of Applied Sciences

UNIVERSITÄT
BERN

Faculty of Medicine
Biomedical Engineering

Master of Science Thesis

Interpretability-based Robustness Analysis of Medical Image Classification Models

by

Pierre Treyer

of Switzerland

Supervisors
Prof. Dr. Mauricio Reyes and Zixin Shu

Institutions
ARTORG Center for Biomedical Engineering Research, University of Bern
Medical Image Analysis Group (MIA)

Examiners
Prof. Dr. Mauricio Reyes and Zixin Shu

Bern, August 2024

Abstract

The abstract should provide a concise (300-400 word) summary of the motivation, methodology, main results and conclusions. For example:

Acknowledgements

Before starting the development of this report, it seems appropriate to begin by expressing my gratitude to the people who made it possible for me to successfully complete this work and make it more enjoyable.

First and foremost, I would like to thank my supervising professor, Mauricio Reyes, for proposing this project and guiding me throughout the semester in this fascinating work. It has been a real pleasure to work on this project with him, thanks to his enthusiasm, availability, and reliability. His thoughtful management and genuine care for the team fostered a collaborative environment that greatly contributed to a positive and supportive atmosphere in the workplace.

I would also like to thank Ms. Zixin Shu for her support and advice on my work, which allowed me to better structure my work and find a clear direction. Her insightful guidance and warm encouragement were truly invaluable, and I deeply appreciate her dedication and kindness throughout the project.

Additionally, I would like to extend my gratitude to the entire Medical Image Analysis team for warmly welcoming me and providing an encouraging environment throughout the project. Their camaraderie and support played a significant role in making this experience both productive and enjoyable.

Finally, I would like to express my deepest gratitude to my family for their unwavering support throughout this project and, more broadly, during my entire Master's journey. Their constant encouragement and belief in my abilities have been a source of strength and motivation, enabling me to overcome challenges and achieve my goals.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Als Hilfsmittel habe ich Künstliche Intelligenz verwendet. Sämtliche Elemente, die ich von einer Künstlichen Intelligenz übernommen habe, werden als solche deklariert und es finden sich die genaue Bezeichnung der verwendeten Technologie sowie die Angabe der «Prompts», die ich dafür eingesetzt habe. Mir ist bekannt, dass andernfalls die Arbeit mit der Note 1 bewertet wird bzw. der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.
Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Bern, August 31$^{\text{th}}$ 2024

Pierre Treyer

Contents

List of Figures

List of Tables

List of Abbreviations

2D, 3D	...	Two- or Three-dimensional, referring to spatial dimensions of an image
ACC	...	Accuracy
AI	...	Artificial Intelligence
AUC	...	Area under the ROC Curve
CAM	...	Class Activation Map
CNN	...	Convolutional Neural Network
CT	...	Computed Tomography
DL	...	Deep Learning
FC	...	Fully Connected
FN	...	False Negative
FP	...	False Positive
FPR	...	False Positive Rate
GAP	...	Global Average Pooling
GradCAM	...	Gradient-weighted Class Activation Map
IDE	...	Integrated Development Environment
ML	...	Machine Learning
MedMNIST	...	Med for medical and MNIST which is a reference to the famous
	...	MNIST dataset used for handwritten digit classification.
MIA	...	Medical Image Analysis
MSE	...	Mean Square Error
MRI	...	Magnetic Resonance Imaging
OCT	...	Optical Coherence Tomography
ReLU	...	Rectified Linear Unit
RMSE	...	Root Mean Square Error
ROC	...	Receiver Operating Characteristic
TN	...	True Negative
TP	...	True Positive
TPR	...	True Positive Rate
XAI	...	Explainable AI

Chapter 1

Introduction

The application of Artificial Intelligence (AI) in medical image classification has gained significant momentum in recent years. Despite the rapid development and proliferation of novel AI methodologies, the integration of these systems into clinical practice still necessitates human oversight to manage errors and corrections. Many errors produced by AI systems are challenging to interpret, as these systems frequently function as opaque "black boxes" [8]. This lack of transparency impedes the ability to understand and rectify errors, thereby undermining clinical trust and decision-making processes.

To address these challenges, there has been a growing emphasis on the development of quality control and error-detection mechanisms that can effectively monitor and assess the outputs of AI systems. These mechanisms aim to ensure that AI-generated results are accurate and reliable [6], thereby increasing trust in AI applications within the medical community.

Building on previous findings, it is proposed that leveraging interpretability information during model training can enhance robustness analyses of medical image classification models [8]. By integrating interpretability methods, such as gradient-based saliency maps [21], into the training process, researchers can gain insights into the model's decision-making patterns. Typically, saliency maps are generated using the last convolutional layer of the model [21], as this layer captures high-level features that are crucial for the final classification. The hypothesis is that the gradient fingerprints of the saliency maps for correctly and incorrectly classified images are distinct and can serve as a proxy for quality control. This distinction can potentially reveal failure patterns, offering a pathway to understanding and mitigating errors.

This Master's thesis, therefore, aims to investigate the mapping between failure patterns or so-called failure modes and Magnetic Resonance Imaging (MRI) sequence-specific image perturbations [23]. By examining how different MRI sequences influence the AI model's performance, the study seeks to identify specific perturbations that correlate with classification errors. Given this mapping, the study will analyze the levels of separability between corresponding layers obtained from unperturbed and perturbed images. By assessing the differences in saliency maps, the research intends to develop a mechanism for flagging incorrect AI results, thereby providing a new approach to quality control.

Chapter 2

Materials and Methods

The following chapter, Material and Methods, outlines the experimental framework and methodologies employed to investigate the robustness of medical image classification models through interpretability techniques. This section details the datasets used, the AI models employed, the specific interpretability methods integrated, and the procedures for evaluating and analyzing model performance. By providing a comprehensive overview of the experimental setup, this chapter lays the foundation for understanding how the research objectives were pursued and how the results were derived.

2.1 Datasets

2.1.1 MedMNIST dataset

The MedMNIST version 2 (V2) [25] dataset was chosen for this work due to its extensive and diverse collection of standardized biomedical images, which is essential for developing and testing robust machine learning (ML) models in the medical domain. The dataset's broad coverage of imaging modalities, including X-ray, Optical Coherence Tomography (OCT), ultrasound, Computed Tomography (CT), and electron microscopy, provides a comprehensive foundation for tackling various biomedical classification tasks. The availability of images at multiple resolutions (28x28, 64x64, 128x128, and 224x224 pixels) allows for flexible experimentation and optimization of model performance across different scales. Additionally, the inclusion of detailed labels with each image obviates the need for specialized medical knowledge, making it accessible for a wide range of users. The official train-validation-test splits for each subset and the extensive number of images—ranging from approximately 800 to 236'000 ensure rigorous evaluation and validation of machine learning algorithms. These attributes make MedMNIST V2 [25] an ideal choice for advancing research in biomedical image analysis and enhancing the reliability and generalizability of predictive models.

This work focuses on 2D datasets and smaller datasets, such as BreastMNIST, DermaMNIST [1], RetinaMNIST, and PneumoniaMNIST, due to several compelling reasons.

Firstly, 2D datasets are particularly advantageous for their simplicity and reduced computational demands compared to 3D datasets. They allow for quicker prototyping and iteration, which is essential for developing and fine-tuning machine learning models. The 2D images in these datasets are representative of a wide range of clinical scenarios, making them suitable for a variety of diagnostic tasks. For example, BreastMNIST, DermaMNIST, RetinaMNIST, and PneumoniaMNIST cover key medical imaging domains such as mammography, dermatology, retinal imaging, and chest X-rays, respectively.

Secondly, smaller datasets are chosen to facilitate rapid experimentation and model development. They are often more manageable in terms of computational resources and training time, enabling researchers to explore different models and techniques more efficiently. Additionally, smaller datasets can serve as valuable benchmarks for evaluating initial model performance before scaling up to larger and more complex datasets.

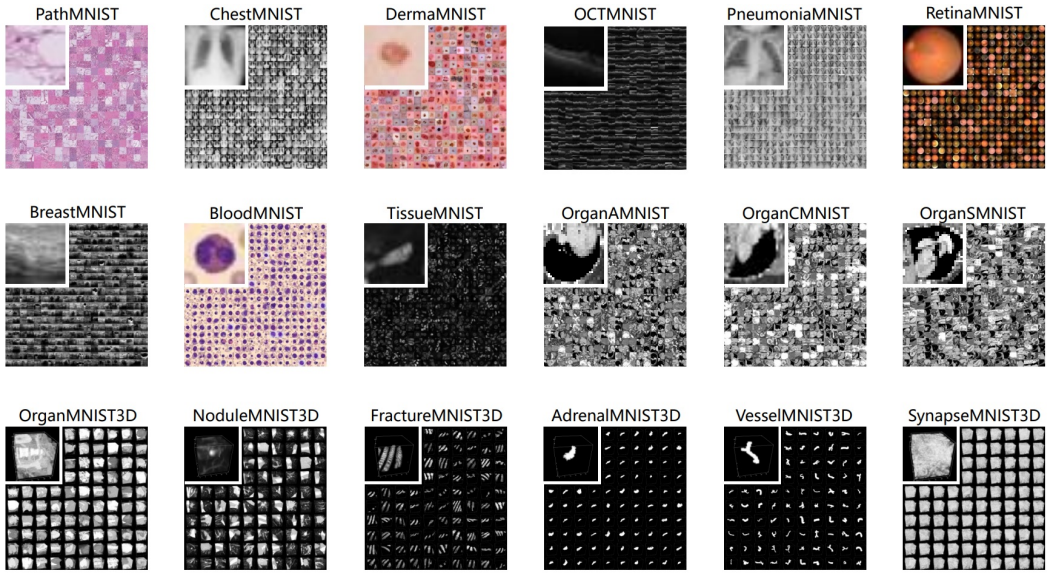

Figure 2.1: Overview of MedMNIST Datasets. Source: `https://medmnist.com/`

To better understand the MedMNIST datasets, the following sections detail each dataset. The BreastMNIST dataset simulates breast cancer diagnostic imaging, offering a benchmark for classifying benign versus malignant tumors. DermaMNIST provides images of various skin conditions, such as melanoma and basal cell carcinoma, to support dermatological diagnosis. PneumoniaMNIST includes chest X-ray images categorized into healthy, bacterial pneumonia, and viral pneumonia, aiding in pneumonia detection research. RetinaMNIST contains retinal fundus images for the diagnosis of various retinal diseases, covering a broad spectrum of retinal conditions. Collectively, these datasets are designed to enhance medical image classification and diagnostic research across different medical specialties.

The table below summarizes the MedMNIST datasets, detailing data modality, classification tasks, and sample distribution across training, validation, and test sets. This overview aids in understanding each dataset's scope and application:

Dataset	**Data Modality**	**Tasks (N° of classes)**	**# Samples (Training/Validation/Test)**
BreastMNIST	Breast Ultrasound	Binary-Class (2)	780 (546 / 78 / 156)
DermaMNIST	Dermatoscope	Multi-Class (7)	10,015 (7,007 / 1,003 / 2,005)
RetinaMNIST	Fundus Camera	Ordinal Regression (4)	1,600 (1,080 / 120 / 400)
PneumoniaMNIST	Chest X-Ray	Binary-Class (2)	5,856 (4,708 / 524 / 624)

Table 2.1: MedMNIST selected Datasets Overview and Distribution. Source: `https://medmnist.com/`

2.1.2 Data Pre-Processing

In this study, data pre-processing and augmentation were critical steps in preparing the MedMNIST dataset for training and evaluation. The preprocessing pipeline began with normalizing the input images [14] to a standardized format that the model could efficiently process. Specifically, each image was first converted to a tensor, which scaled the pixel values to a range of [0,1]. Subsequently, the images were normalized by centering the pixel values around zero and scaling them to have a standard deviation of one. This normalization step ensures that the data fed into the model has consistent statistical properties, which can lead to more stable and efficient training.

2.1.3 Data Augmentation

In addition to basic pre-processing, data augmentation [20] was employed to improve the model's robustness and generalization capability. The augmentation process involved applying a variety of perturbations to the training images. Details of these perturbations and their implementation will be discussed further in the paper. This approach ensures that the model is exposed to a wide spectrum of possible input variations, thereby enhancing its ability to generalize to diverse and potentially noisy data. However, data augmentation was not always utilized; in some experiments, the model was trained on the original, unaltered dataset to serve as a baseline for comparison with the augmented data.

This pre-processing and augmentation strategy aimed to create a more diverse training dataset, thereby enhancing the model's ability to generalize well to unseen data during testing.

2.2 AI Model

2.2.1 ResNet18 Architecture

ResNet18 [9] is a convolutional neural network (CNN) architecture [26] developed by Microsoft Research, distinguished by its incorporation of residual connections. The architecture comprises 18 layers, leveraging these residual connections to facilitate the training of deep networks by mitigating the vanishing gradient problem and enabling the learning of residual mappings. This design allows ResNet18 to achieve high accuracy across a range of image classification tasks.

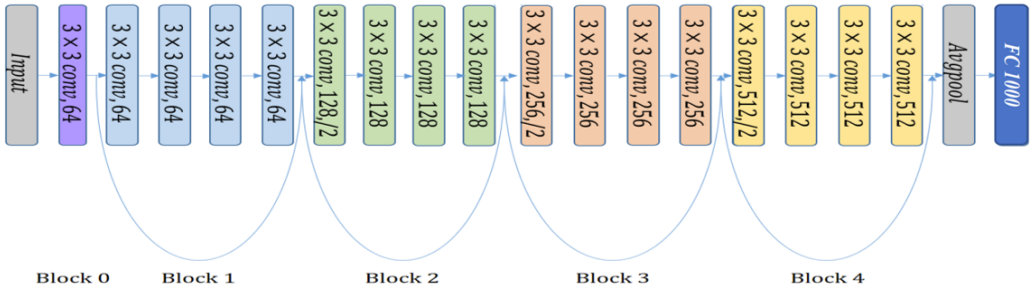

Figure 2.2: ResNet18 model architecture, showing the input layer, 18 convolutional layers, average pooling, and the fully connected layer. Source: `https://www.semanticscholar.org/reader/57dc772091632ea97adfb9a7f8c45dafd40a1e1a`

ResNet18 has demonstrated competitive performance on prominent benchmark datasets [25], including ImageNet, where it has consistently achieved high accuracy in image classification challenges. Its effectiveness extends to various other benchmarks and datasets, including MedMNIST , where it has excelled as a benchmarking method, showcasing its robustness in medical image analysis.

	PathMNIST		ChestMNIST		DermaMNIST		OCTMNIST		PneumoniaMNIST		RetinaMNIST	
Methods (Image size)	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18 (28)	0.983	0.907	0.768	0.947	0.917	0.735	0.943	0.743	0.944	0.854	0.717	0.524
ResNet-18 (224)	0.989	0.909	0.773	0.947	0.920	0.754	0.958	0.763	0.956	0.864	0.710	0.493

Table 2.2: Benchmark performance Metrics for ResNet-18 with different image sizes on the MedMNIST datasets. Source: `https://medmnist.com/`

The choice of ResNet18 for this study is based on several key factors. Firstly, its performance is notable for its accuracy and efficiency. The residual connections not only enhance training stability but also contribute to improved generalization across diverse tasks. Additionally, ResNet18's relatively compact architecture makes it more computationally efficient than deeper variants, which is advantageous for both research and industrial applications. These characteristics make ResNet18 a versatile and practical choice for a wide range of computer vision tasks, ensuring high performance while maintaining manageable computational demands.

2.2.2 Hyperparameters

In this study, the selected hyperparameters for training the AI model on the MedMNIST dataset were carefully chosen to balance performance and computational efficiency. The input image size was set to the maximum resolution of 224 x 224 pixels, a standard size that preserves sufficient detail for accurate medical image analysis while being compatible with widely used pre-trained models. The maximum number of epochs was fixed at 200 to ensure adequate training time, allowing the model to converge while minimizing the risk of overfitting. A batch size of 16 was selected to accommodate memory constraints, given the resolution of the images, while still enabling effective gradient updates.

The learning rate was set to 0.001 [7], a commonly used value that ensures steady convergence without overshooting minima in the loss function. A patience of 10 was introduced as part of an early stopping strategy, halting training if no improvement in validation loss was observed for 10 consecutive epochs. This helps prevent unnecessary training beyond the point of convergence, reducing overfitting and saving computational resources. Finally, the dropout rate was set to 0, as initial experiments indicated that regularization was not necessary for this dataset and model configuration.

These hyperparameters were chosen based on a combination of best practices and empirical testing, ensuring a robust and efficient training process tailored to the MedMNIST dataset's specific characteristics.

2.3 Generation of Perturbations

Perturbations [12] such as Gaussian noise, speckle noise, motion blur, contrast adjustments, and brightness variations were applied using the imgaug library [11]. These perturbations were designed to simulate real-world variations [17] in MRI scans, such as motion blur caused by patient movement, contrast changes due to differences in imaging protocols or settings, and brightness variations resulting from inconsistent scanner calibrations or signal intensity. Each perturbation was implemented across five levels of severity, with level 1 being the least severe and level 5 being the most severe, to comprehensively assess the model's robustness to different types and intensities of noise and distortions.

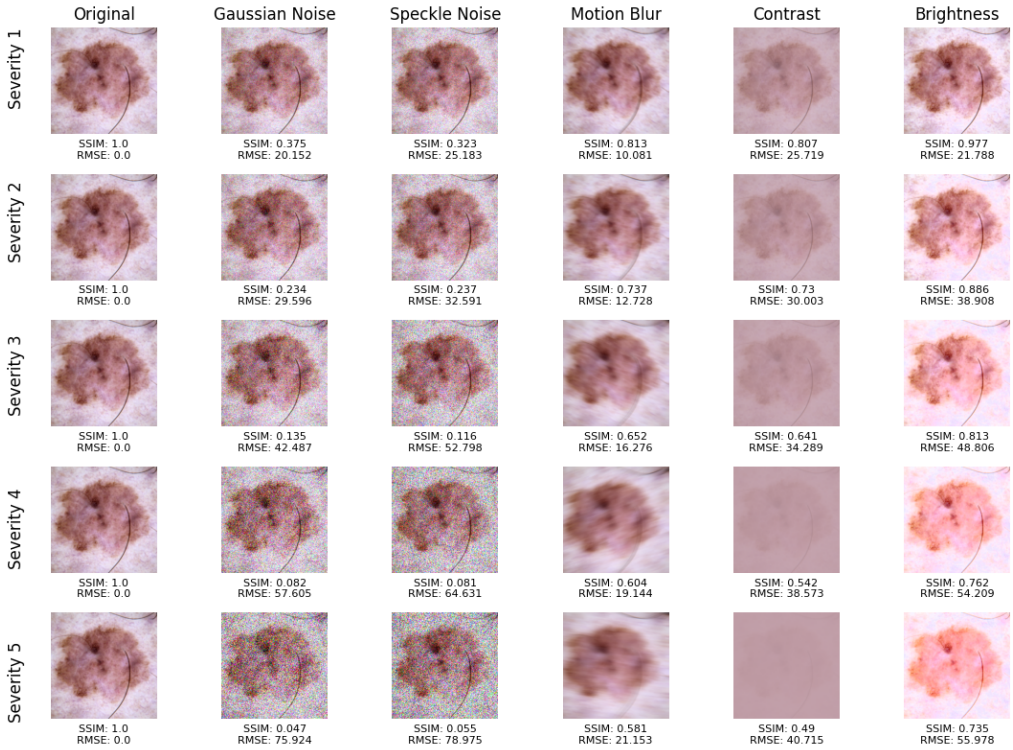

Figure 2.3: Generation of 5 types of perturbations at severity levels ranging from 1 to 5 applied to the original image. The dataset used is Dermamnist, a 2D dataset. The image shown is a 'melanoma,' with dimensions of 224x224 pixels.

These perturbations were utilized not only for augmenting the training dataset but also for generating new testing datasets. Specifically, the original testing dataset was subjected to these perturbations to create additional test datasets. This process resulted in a total of 26 distinct datasets comprising the original testing dataset and 25 perturbed variations. The model was then evaluated on these datasets to assess its performance under various conditions of noise and distortion.

2.4 Explainable AI (XAI)

XAI [24] aims to provide insights into how models arrive at specific conclusions, ensuring that they are not only accurate but also reliable and trustworthy. In this project, saliency maps [21], a common XAI technique, will be leveraged to visually demonstrate the areas of input data that a model considers important when making predictions.

2.4.1 Class activation map (CAM)

Class Activation Maps (CAMs) [15] are a pivotal technique in the field of interpretability within deep learning (DL), particularly for visualizing the regions of an image that are most influential in the model's class predictions. By providing spatial information about which parts of the input image the model is focusing on, CAMs offer valuable insights into the decision-making process of CNNs [21]. These heatmaps can be superimposed upon the original image to identify and emphasize critical areas within medical images, potentially aiding in the diagnosis of diseases and enhancing the overall trust in AI-driven medical tools. This overlay allows to directly visualize the correspondence between the highlighted regions and the anatomical structures, thereby facilitating more informed and confident decision-making in clinical practice. However, CAMs are inherently dependent on the model architecture, requiring both a Global Average Pooling (GAP) layer and a Fully Connected (FC) layer.

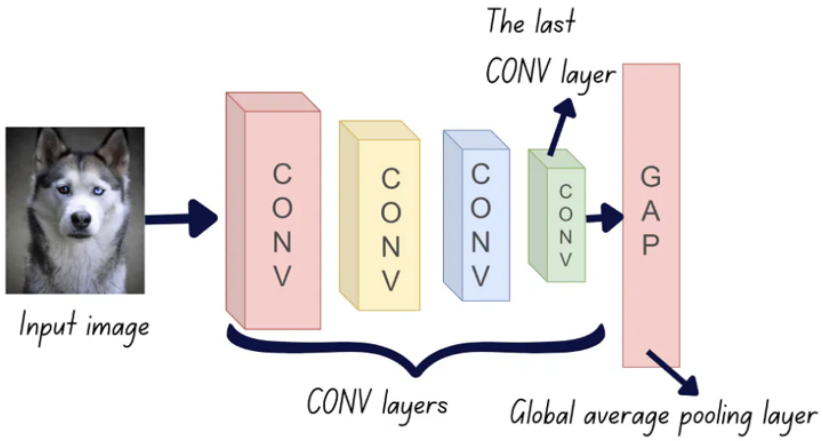

Figure 2.4: CNN processing an image of a dog, with convolutional layers and a GAP layer, used for generating the Class Activation Map for feature visualization. Source: `https://www.pinecone.io/learn/class-activation-maps/`

CAMs work by aggregating the information from the feature maps produced by the convolutional layers of a neural network. Specifically, It involves a weighted sum of the feature maps from the final convolutional layer, where each feature map is associated with a particular class. Typically, they are generated from the last convolutional layer [21], allowing for a detailed spatial understanding of the model's decision-making process. In this thesis, however, saliency maps is analyzed across all layers of the network, encompassing the entire model, to provide a comprehensive view of the model's decision-making process and to visualize the sensitivity of the model throughout its architecture.

CAM can be mathematically described as follows:

$$y^c = \sum_k w_k^c \cdot \frac{1}{Z} \sum_i \sum_j A_{ij}^k \tag{2.1}$$

- A_{ij}^k: Pixel at location (i, j) in the k-th feature map.
- Z: Total number of pixels in the feature map.
- w_k^c: Weight of the k-th feature map for class c.
- y^c: Output score for class c.

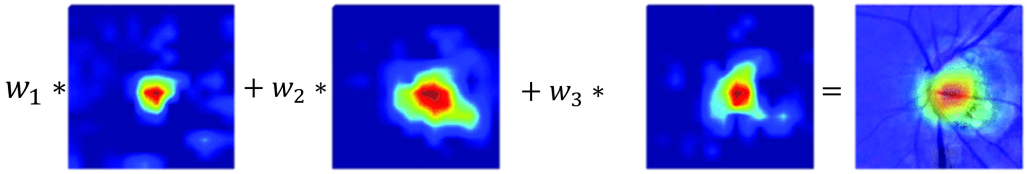

Figure 2.5: Generation of a class activation map with a weighted sum of the feature maps.

2.4.2 Gradient-weighted Class Activation Map ++ (Grad CAM ++)

Grad-CAM++ [3] is an advanced variant of CAM that provides more precise localization of object boundaries within an image. Unlike CAM, which requires specific architectural components like a GAP layer and a FC layer, Grad-CAM++ can be applied to a broader range of model architectures. It achieves this flexibility through a gradient-based approach. Grad-CAM++ can be mathematically described as follows:

$$L_{\text{Grad-CAM++}}^c = \text{ReLU}\left(\sum_k w_k^c A^k\right) \tag{2.2}$$

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \text{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right) \tag{2.3}$$

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{\partial (A_{ij}^k)^2}}{2 \cdot \frac{\partial^2 y^c}{\partial (A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \frac{\partial^3 y^c}{\partial (A_{ij}^k)^3}} \tag{2.4}$$

- A_{ij}^k: Activation at location (i, j) in the k-th feature map.
- w_k^c: Weight for the k-th feature map for class c.
- α_{ij}^{kc}: Coefficient for the pixel (i, j) in the k-th feature map for class c.
- y^c: Output score for class c.
- ReLU: Rectified Linear Unit activation function.

Numerous methods derived from CAM have been developed, each offering various enhancements to improve interpretability and localization accuracy. Among these, Grad-CAM++ was chosen primarily for its computational efficiency, in addition to its ability to provide more detailed and accurate visual explanations. Unlike standard CAM and its predecessors, Grad-CAM++ incorporates higher-order gradients and pixel-level contributions, enabling it to capture finer details and more nuanced areas of importance within the feature maps, all while maintaining faster calculation times.

As seen in the previous equations, Grad-CAM++ computes the gradients of the output score for a particular class with respect to the feature maps in the final convolutional layer of the CNN. These gradients flow back through the network, indicating the importance of each neuron in the feature maps for the target class. Grad-CAM++ improves upon standard Grad-CAM [19] by considering not only the first-order gradients but also higher-order gradients and the pixel-level contributions, allowing it to generate more finely detailed heatmaps.

By aggregating these gradients, Grad-CAM++ assigns importance weights to each feature map, which are then used to produce a weighted sum of the feature maps. This results in a class-discriminative localization map that highlights the regions in the input image most relevant to the class prediction. This precise localization is particularly beneficial for tasks involving complex images, where understanding exactly which regions contribute to the model's decision is critical.

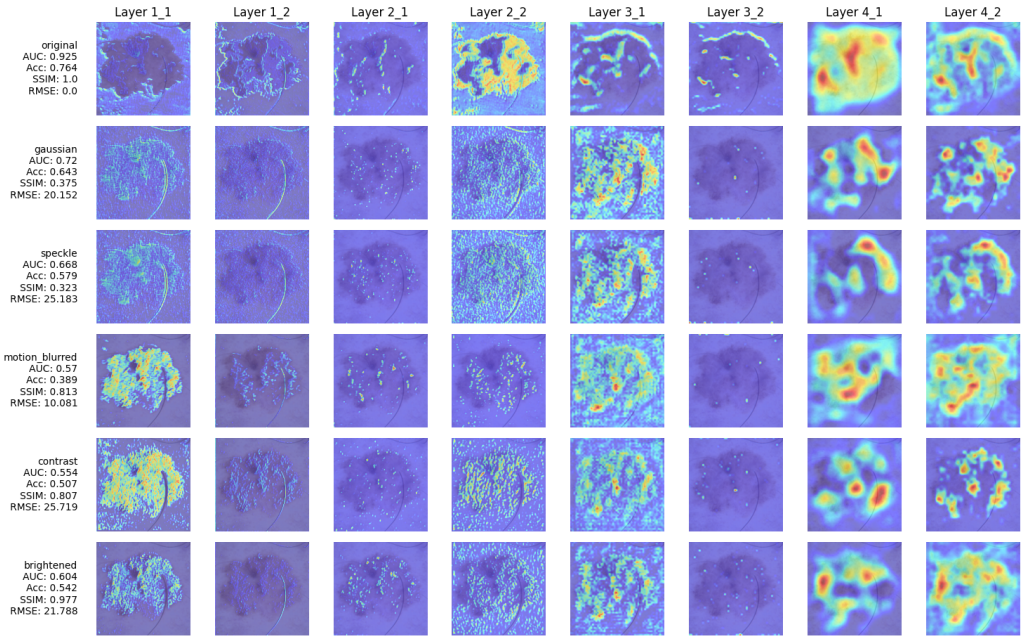

Figure 2.6: Saliency map generation across various layers and different perturbations at severity level 1. The x-axis represents the model's layers, while the y-axis corresponds to the different perturbations. The dataset used is Dermamnist, which is a 2D dataset with images of size 224x224. The model achieved correct predictions for the label "melanoma." It was trained for 78 epochs without data augmentation.

In this study, heatmaps were generated using the torchcam.methods. The GradCAM++ extractor was employed to produce heatmaps corresponding to the class with the highest predicted score. These heatmaps were generated across 5 different perturbations and 1 to 5 severity levels, with a total of 8 layers. This setup results in the generation of 200 heatmaps for each original image. Subsequently, the heatmaps were normalized to an 8-bit format and resized to align with the original dimensions of the input images.

2.5 Metrics

2.5.1 Loss function

A loss function, also known as a cost function, quantifies how far off the model's predictions are from the actual target values during training. In this project, BCEWithLogitsLoss [4] is used as a loss function specifically designed for binary- and multi-class tasks. It is a combination of a sigmoid activation function and binary cross-entropy loss. The sigmoid function converts raw model outputs (logits) into probabilities, and the binary cross-entropy then measures the difference between these predicted probabilities and the actual binary labels. This loss function is found in the torch.nn module of PyTorch [16].

$$\ell(x, y) = L = \begin{pmatrix} l_1 \\ \vdots \\ l_N \end{pmatrix}, \quad l_n = -w_n \left[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n)) \right] \tag{2.5}$$

$\ell(x, y)$	:	The overall loss for a given input x and target y.
L	:	The loss vector, where each component l_n corresponds to the loss for a specific output.
l_n	:	The individual loss for the n-th output.
w_n	:	A weight associated with the n-th output, which could be used to give different importance t
y_n	:	The actual target label for the n-th output (either 0 or 1 in binary classification).
x_n	:	The raw model output (logit) for the n-th output.
$\sigma(x_n)$	:	The sigmoid function applied to the logit x_n, converting it into a probability.
$\log \sigma(x_n)$	:	The log probability of the model predicting the correct class.
$\log(1 - \sigma(x_n))$	:	The log probability of the model predicting the incorrect class.

2.5.2 Accuracy (ACC)

Accuracy is a metric that measures the proportion of correctly predicted instances out of the total instances in a dataset. It is calculated using the equation:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

TP	:	True Positive	FP	:	False Positive
TN	:	True Negative	FN	:	False Negative

In this project, accuracy was used to evaluate the model's performance, indicating how well the model is classifying the data overall. The evaluation was conducted using the scikit-learn library. However, accuracy can be misleading in imbalanced datasets where one class dominates, which is why other metrics need to be implemented.

2.5.3 Area under the ROC Curve (AUC)

The Area Under the Curve (AUC) metric is a widely used evaluation measure in classification problems. It quantifies the overall ability of a classifier to discriminate between positive and negative examples, representing the area under the Receiver Operating Characteristic (ROC) curve.

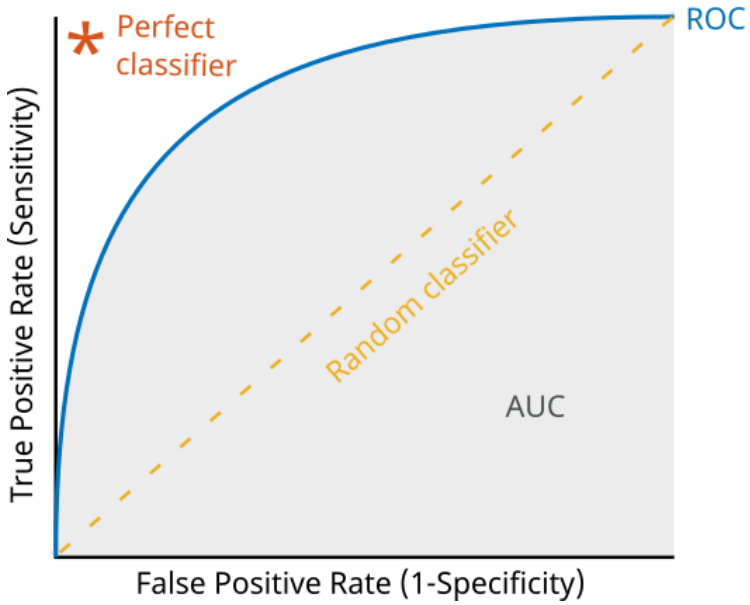

Figure 2.7: The ROC curve illustrates the performance of a classifier. The True Positive Rate (TPR), also known as sensitivity or recall, is plotted on the y-axis, while the False Positive Rate (FPR) is shown on the x-axis. The curve demonstrates the trade-off between sensitivity and specificity as the decision threshold varies. The diagonal dashed line represents a random classifier, serving as a baseline with an AUC of 0.5, indicating no discriminative power. Source: `https://ch.mathworks.com/help/deeplearning/ug/compare-deep-learning-models-using-ROC-curves.html`

In binary classification, AUC is computed as the integral of the ROC curve, which plots the true positive rate against the false positive rate at various threshold settings. This calculation is performed using the roc-auc-score function from the scikit-learn library.

For multi-class classification, AUC is extended through a one-vs-rest approach, where the ROC curve is computed for each class against all others. The final AUC is then averaged across all classes, providing a comprehensive evaluation of the classifier's performance across multiple classes.

Scikit-learn's roc-auc-score function efficiently handles both binary and multi-class scenarios, offering robust AUC calculations that facilitate model comparison and performance analysis.

In this project, the AUC metric was utilized to evaluate the performance of each dataset. Additionally, the AUC deltas were computed to quantify the differences in performance between the original dataset and its various perturbed versions.

$$\Delta\text{AUC} = \text{AUC}_{\text{original}} - \text{AUC}_{\text{perturbed}} \tag{2.7}$$

2.5.4 Root mean square error (RMSE)

RMSE is a widely used metric for quantifying the difference between predicted and observed values. It measures the square root of the average squared differences between these values, providing a measure of the average magnitude of errors in predictions.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2.8}$$

In this work, RMSE is applied using the skimage library to quantify the discrepancy between saliency maps derived from the original image and its perturbed version. By calculating RMSE for each layer of the saliency maps, the effect of perturbations on the saliency representation can be systematically evaluated, facilitating a comprehensive comparison of how these perturbations alter the saliency maps.

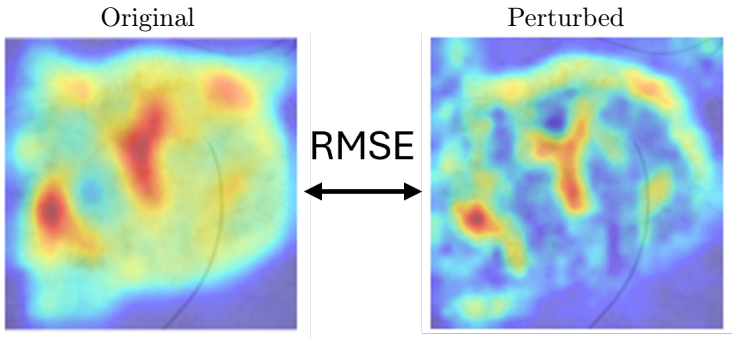

Figure 2.8: RMSE between two saliency maps of the 2D Dermamnist dataset. The left side shows the saliency map for the original image, and the right side displays the saliency map for the perturbed image.

2.6 Visualization and Axis Selection

2.6.1 First part - Saliency map variations across layers

In the initial phase of experimentation, the RMSEs between the saliency maps derived from the original and perturbed images for each class within a specific dataset were plotted as a function of the targeted layers. Each data point in these plots represents the average RMSE calculated from a set of up to 100 original images, with some classes having fewer images.

The experiments were conducted on the four selected MedMNIST datasets and various types of perturbations with five different severities. Only the breastmnist dataset was augmented with perturbed images due to its small size. Augmenting other datasets was not feasible as it required too much GPU memory. Additionally, the experiments were conducted on both correctly predicted and incorrectly predicted images.

This methodology facilitated the visualization of variations in saliency maps across the model. It allowed for an in-depth analysis of how these different factors influence variations in saliency maps across different layers of the model.

The primary goal of this analysis was to identify tendencies and patterns in the differences observed in saliency maps across various datasets and classes. By examining these variations, the experiment sought to uncover which components of the model demonstrated greater robustness and to identify the specific factors contributing to this robustness. This information is crucial for understanding how different MRI sequence perturbations, with varying severities, impact the model saliency maps and can provide a basis for targeted improvements in model interpretability.

To further contextualize these findings within the framework of quality control, it is important to correlate the observed patterns with the broader objectives of enhancing model robustness and interpretability. The analysis of saliency map differences aligns with the hypothesis that distinct gradient fingerprints for correctly and incorrectly classified images can serve as a proxy for identifying failure modes. By identifying layers and datasets where significant discrepancies occur, this experiment contributes to the development of mechanisms for flagging incorrect AI results and improving quality control processes.

2.6.2 Second part - Correlation between AUC and RMSE

In the second phase of experimentation, scatterplots and boxplots were created to examine the relationship between the deltas AUC performance (between original and perturbed datasets) and the RMSEs (between saliency maps). For each class, up to a hundred points were plotted, with some classes having fewer due to a smaller number of images.

The initial visualizations included labels representing the five distinct perturbation types, with the AUC plotted as a function of the corresponding RMSEs. This approach aimed to elucidate how variations in RMSEs relate to changes in AUC performance for each perturbation type.

Additionally, alternative visualizations were produced where the labels represented different levels of perturbation severity rather than the perturbation types. This modification allows for a more nuanced analysis of how varying severities impact the relationship between RMSEs and AUC performance. By including both types of labels, the visualizations provide comprehensive insights into how both perturbation types and their severities influence model performance

To further investigate the effects of perturbation severity and type, different plots were generated for various perturbation severities and types, depending on the visualization focus. These plots were analyzed across all model layers to comprehensively assess how both the nature of the perturbations and their severities impact the relationship between RMSEs and AUC performance. Similar to the initial analysis, certain datasets were enhanced with perturbed images, while others were left unaltered. The experiments also included evaluations of both correctly and incorrectly predicted images.

The selection of deltas AUC and RMSE as axes for visualization was motivated by their direct relevance to assessing the robustness and reliability of AI models in medical image classification. Plotting deltas AUC against RMSEs of saliency maps aims to uncover how changes in model interpretability, as indicated by saliency maps, correlate with changes in classification performance under various perturbations. This correlation is crucial for investigating whether the model's decision-making patterns are reliable and consistent, especially under different conditions of data perturbation. Understanding this relationship addresses the challenges identified in the introduction, where the opacity of AI systems often hinders error correction and quality control. Utilizing scatterplots and boxplots enhances this analysis by providing clear, visual insights into the distribution and trends of these relationships across different classes and perturbation severities.

2.7 Experimental Setup and Training Configuration

The experimental work for this project was conducted using Python version 3.11.8. The training of models was performed with PyTorch [5], specifically version 2.2.2, which includes CUDA 11.8 support for accelerated computation on compatible GPUs. The GPU resources utilized for these experiments were provided by the Nerve server of the MIA Group, which is part of the ARTORG Center for Biomedical Engineering Research. It offers high-performance computing capabilities essential for efficient model training.

For development and code management, the Integrated Development Environment (IDE) utilized was PyCharm, provided by JetBrains. The JetBrains Gateway was employed to facilitate remote development through a secure SSH connection. This setup allows for efficient code editing and debugging directly on the server where the experiments are executed, enhancing the workflow by providing a seamless and integrated development experience.

Version control was managed using Git, with code repositories hosted on GitHub. The experimental setup was documented to ensure reproducibility.

Chapter 3

Results

This chapter presents the results obtained in this thesis. It begins by illustrating the variation in saliency maps across different layers for different parameters. Subsequently, it highlights the correlation between AUC and RMSE under varying conditions. The results shown here are selected to provide the most relevant insights, and not every figure generated during the analysis will be presented.

3.1 Saliency map variations across layers

3.1.1 Test on correctly classified images

Figure 3.1: RMSEs of saliency maps across different perturbations and target layers for correctly classified images of severity 1 without augmentation. There are four distinct datasets: breastmnist, dermamnist, pneumoniamnist, and retinalmnist. Each row corresponds to a different dataset, while each column represents a specific image perturbation type: gaussian noise, speckle noise, motion blur, contrast adjustment, and brightness adjustment. For each dataset and perturbation, the RMSEs are plotted across different target layers of the neural network, with separate lines representing different classes within each dataset. Every point plotted represents an average of up to 100 data points depending on the size of the dataset, ensuring robust statistical representation. The plot also includes a "Deltas AUC" value, indicating the AUC of the differences between RMSEs for various classes, providing a measure of separability. A larger "Deltas AUC" value suggests a more significant drop in model accuracy when compared to the original dataset, thus highlighting the impact of the perturbation.

Dataset-Specific Observations:

BreastMNIST

The RMSE values show fluctuation across different layers, but the overall differences in RMSE between malignant and benign classes under various perturbations, such as Gaussian noise, Speckle noise, Motion Blur, Contrast adjustment, and Brightening, remain moderate. Among the perturbations, Speckle noise results in the highest AUC delta of 0.523, indicating the most significant divergence between saliency maps of the original and perturbed dataset. In contrast, Gaussian noise produces the lowest AUC delta of 0.401. When examining the layers, the first few layers exhibit relatively lower RMSE values, with a trend of increasing RMSE in the deeper layers.

DermaMNIST

The DermaMNIST dataset exhibits more erratic behavior, with RMSE values scattered across the target layers. The observed trends are inconsistent, making it challenging to identify a clear pattern. When examining the impact of different perturbations, the lowest AUC delta is associated with Gaussian noise (0.246), while contrast adjustment results in the highest AUC delta (0.320). This indicates that the separability between classes, in terms of RMSE, is more pronounced under contrast adjustments. Across the layers, there is no consistent trend, suggesting that the model's robustness is more influenced by the type of perturbation rather than by any specific layer.

PneumoniaMNIST

Similar to the BreastMNIST dataset, the RMSE values in this case are more stable, displaying a gradual increase across the layers. However, certain layers show significant drops in RMSE, indicating that these layers might be less sensitive to perturbations. Among the perturbations applied, Gaussian noise results in the highest Delta AUC (0.506), signaling a substantial divergence in saliency maps between the classes. Other perturbations show lower Delta AUCs, with the Brightened perturbation having the lowest value at 0.243. The middle layers exhibit more noticeable differences in RMSE, suggesting that these layers may capture more class-distinctive features when subjected to perturbations.

RetinaMNIST

The RMSE values in this dataset are relatively flat, with only minor fluctuations across the layers. The overlapping RMSE values between the different classes suggest low separability based on the saliency maps. Regarding perturbations, all of them exhibit low AUC deltas, ranging from 0.162 to 0.169, indicating minimal divergence in saliency maps between correctly and incorrectly classified images. Across the layers, the trend remains relatively stable, demonstrating consistent robustness to perturbations, though there is no significant separation in RMSE values.

Layer-Specific Observations:

Across most datasets, there is a noticeable trend where the RMSE values tend to increase as we move from the shallow layers (closer to the input) to the deeper layers (closer to the output). This is expected because deeper layers generally capture more abstract and class-specific features, which are more sensitive to perturbations and contribute directly to the final decision.

The effect of perturbations like Gaussian noise and Speckle noise tends to be more pronounced in deeper layers across multiple datasets. This suggests that these types of noise may cause more significant distortions in high-level features, which are critical for final classification. This observation could be particularly important when designing defenses against such perturbations, as focusing on deeper layers might yield better robustness improvements.

The DermaMNIST dataset stands out because its RMSE trends are less stable across layers and perturbations, indicating a more complex interaction between the image characteristics and model layers. This could be due to the inherent variability in skin lesions, which are less homogeneous compared to other datasets like PneumoniaMNIST.

3.1.2 Test on incorrectly classified images

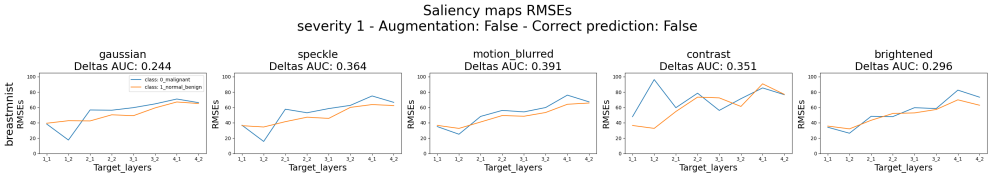

Figure 3.2: Results of Saliency map RMSEs across different datasets and perturbations for incorrectly predicted images of severity 1 without augmentation. The layout and axes are consistent with the previous figure; however, the key differ- ence is that this figure represents incorrectly classified images only. Similar to the earlier figure, RMSEs are plotted across various target layers of the neural network, with distinct lines representing different classes within each dataset. The "Deltas AUC" value is again provided, reflecting the AUC of the differences between RMSEs for the classes.

Missing observations and correct plot, waiting for the entire plot to load on nerve

3.1.3 Test on augmented dataset

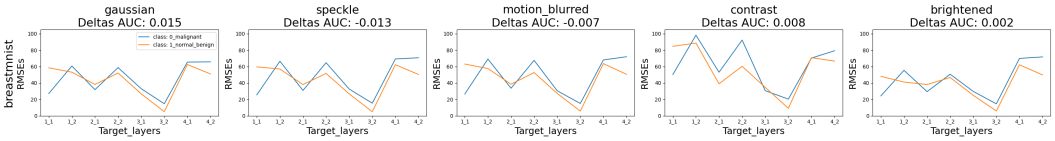

Figure 3.3: RMSEs for the augmented BreastMNIST dataset across different perturbations for incorrectly predicted images of severity 1. It focuses exclusively on the BreastMNIST dataset due to the high memory usage required by the other datasets. The layout and axes remain consistent with the previous figures, allowing for direct comparison of the effects of augmentation across the same perturbation types.

General Observations:

All the perturbations exhibit Delta AUC values close to zero, indicating minimal divergence between the model's performance on the original and perturbed datasets. This suggests that the augmentation process has effectively improved the model's robustness, as the performance remains relatively stable even when perturbations are applied.

Layer-Specific Observations:

The RMSE values for the augmented dataset exhibit fluctuations across different layers, and these fluctuations are more erratic and pronounced compared to the non-augmented dataset. Similar to the BreastMNIST dataset, the model demonstrates increased sensitivity to contrast variations. Despite the use of augmentation, these perturbations introduce greater instability in the model's performance, indicating that the augmentation has not fully mitigated the sensitivity to such factors.

The deep layers (3.2 and 4.1) consistently show a peak in RMSE for all perturbations, suggesting that even with augmentation, these deeper layers are more sensitive to perturbations; however, the overall impact on performance remains lower compared to the non-augmented dataset.

3.1.4 Additional tests and observations

Tests were also conducted on images with different severity levels of perturbations beyond severity 1. While these results are not included here, they follow similar trends and further support the findings presented. The decision to omit these additional figures was made to ensure clarity and conciseness in the presentation of the results.

3.2 Correlation between AUC and RMSE

3.2.1 BreastMNIST

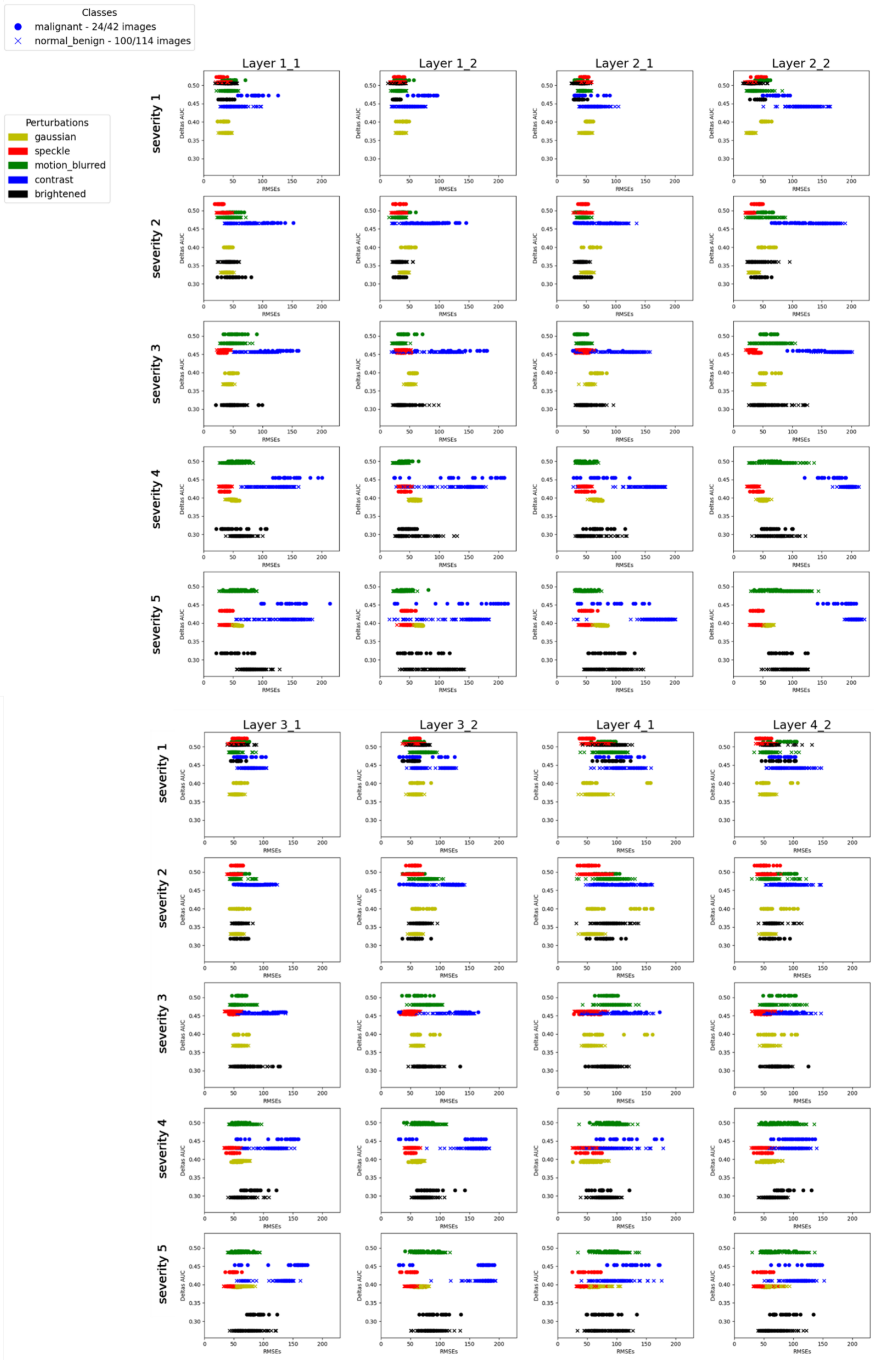

Figure 3.4: Correlation analysis of delta AUC and delta Saliency for the BreastMNIST dataset without augmentation.

This figure presents a series of scatter plots showing the relationship between Delta AUC and RMSE values across various perturbations, target layers, and perturbation severities. The analysis was conducted on the BreastMNIST dataset, specifically on correctly classified images. The x-axis represents the RMSE values of the saliency maps across different target layers, while the y-axis represents the Delta AUC, which measures the divergence between the original model performance and its performance under perturbation. The plots are organized in a grid where the columns correspond to different target layers of the model, and the rows correspond to different perturbation severities from 1 to 5. The legend indicates two classes: malignant and benign. For the malignant class, 24 out of 42 images were correctly predicted, while for the benign class, 100 out of 114 images were correctly predicted.

General Observations:

The analysis of the correlation between Delta AUC and RMSEs for the BreastMNIST dataset across all layers and perturbation severities reveals hardly any consistent linear or non-linear relationship between these two metrics. The scatter plots further illustrate that high RMSE values do not necessarily correspond to either high or low Delta AUCs. This indicates that significant changes in saliency maps do not always coincide with notable performance drops or improvements, highlighting a complex relationship between perturbations and model behavior within the BreastMNIST dataset.

The Delta AUC seems to decrease with increasing severity for most perturbations. This trend is consistent across layers and is particularly noticeable in higher severity levels (severity 4 and 5).

Layer-Specific Observations:

For the first layer and low severities, the Delta AUC values for most perturbations are relatively close together, indicating that the perturbations have similar effects on the model's performance. However, as the severity increases, the spread in Delta AUC values widens significantly, particularly for the "motion blurred" perturbation, which exhibits a substantial reduction in performance.

In the second layers, the impact of perturbations becomes more apparent compared to the first layers, especially as severity increases. The variance in Delta AUC values is broader, reflecting that the model's performance is more sensitive to perturbations at this stage. Among the various perturbations, "motion blurred" continues to have the most significant adverse effect, whereas "contrast" and "brightened" perturbations show a relatively smaller impact.

Moving to the third layers, the Delta AUC values display a moderate spread for different perturbations. "Motion blurred" remains the most impactful perturbation, causing a noticeable reduction in performance. As the severity of the perturbations increases, the spread of Delta AUC values becomes even more significant, indicating that higher layers are increasingly sensitive to these disruptions.

In the highest layers, the impact of perturbations is generally more pronounced. The Delta AUC values exhibit a wider spread, particularly for "motion blurred" and "speckle" perturbations. This suggests that the higher layers are more vulnerable to the effects of perturbations, especially as severity increases.

Perturbation-Specific Observations:

Some perturbations have a more significant impact on the Delta AUC than others. For example, the "motion blurred" perturbation tends to have a broader range of Delta AUC values across all layers, suggesting it may have a more substantial impact on the model's performance.

"Speckle" and "Gaussian" perturbations also show significant impacts, though they are generally less severe than "motion blurred."

3.2.2 PneumoniaMNIST

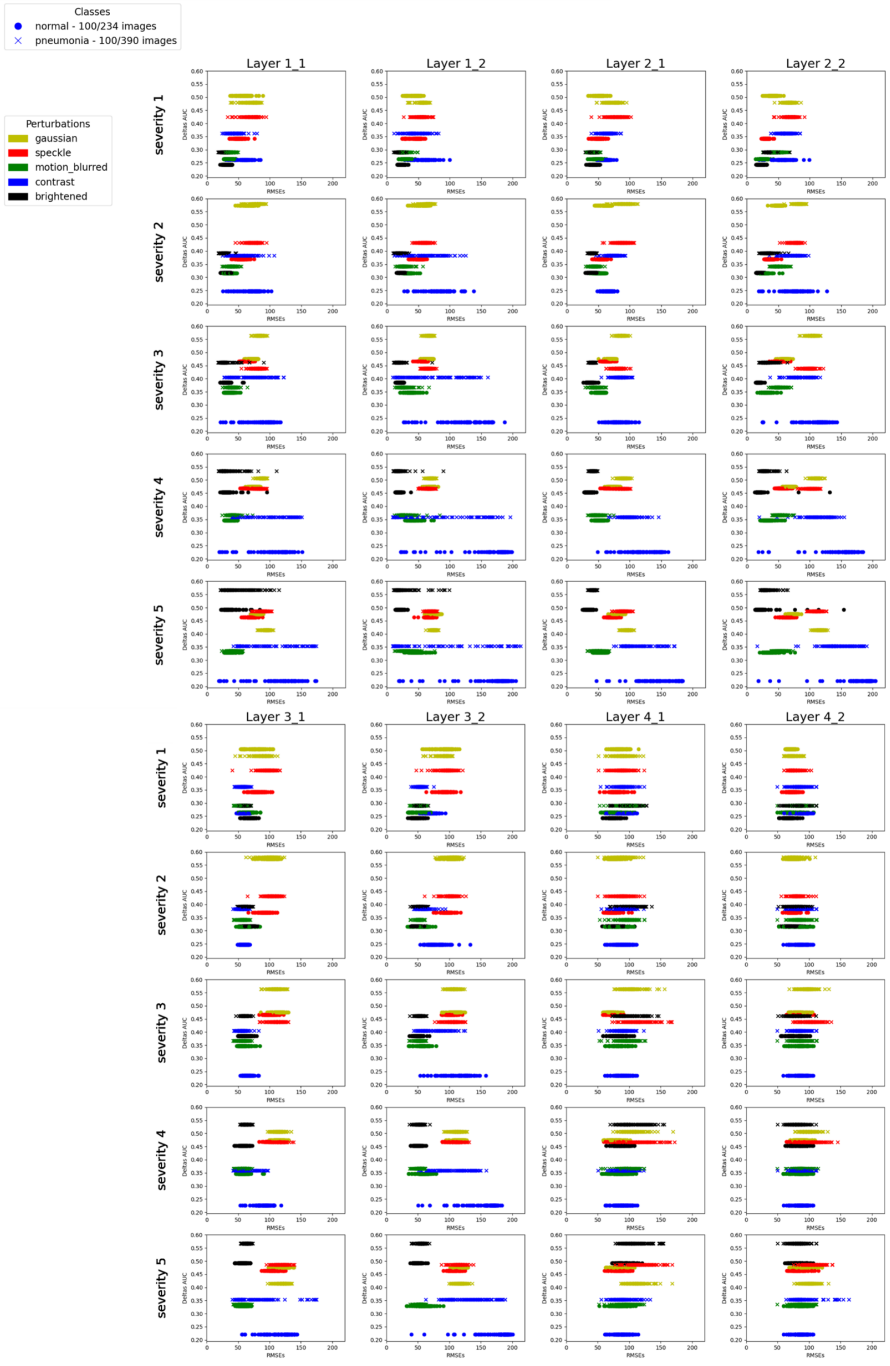

Figure 3.5: Correlation analysis of delta AUC and delta Saliency for the PneumoniaMNIST dataset without augmentation.

This figure presents a series of scatter plots showing the relationship between Delta AUC and RMSE values across various perturbations, target layers, and perturbation severities. The analysis was conducted on the PneumoniaMNIST dataset, specifically on correctly classified images. The axes and layout are the same as in the previous figure. The legend indicates two classes: normal and pneumonia. For the normal class, 100 out of 234 images were correctly predicted, while for the pneumonia class, 100 out of 390 images were correctly predicted.

General Observations:

Similar to previous analyses on the BreastMNIST dataset, there is no clear linear or non-linear correlation between Delta AUC and RMSEs across layers and severities. This analysis further reinforces the complexity of the relationship between perturbations and model performance, as the overall pattern remains inconsistent across different perturbation types and layers. The brightness perturbation exhibits the most pronounced variability in Delta AUC and RMSE values, indicating a broader range of effects on model performance compared to other perturbation types.

Layer-Specific Observations:

Layer 1, specifically 1.1 and 1.2, exhibits relatively inconsistent RMSEs and AUC values across all severities. Specifically the brightness performance across all severity and the contrast with a really wide RMSE range of values. This variability could reflect the model's difficulty in generalizing under different perturbation conditions, particularly in the initial layers where fundamental feature extraction occurs.

Moving to Layer 2 and Layer 3, there is still no clear correlation between RMSE and Delta AUC. Moving to Layers 2 and 3, there is still no clear correlation between RMSE and Delta AUC, with no significant changes observed compared to Layer 1. Additionally, the saliency maps for these layers appear slightly less sensitive to contrast perturbations, indicating a marginal reduction in the impact of contrast variations on model performance as the data passes through deeper layers.

In Layer 4, particularly in sublayers 4.1 and 4.2, the deepest layers demonstrate the least variability in RMSE, especially under higher severity perturbations. The RMSE values for contrast are more concentrated, indicating greater robustness to contrast changes, whereas the model exhibits increased sensitivity to Gaussian and speckle noise, suggesting these noise types have a more pronounced impact on performance in the deeper layers.

Perturbation-Specific Observations:

Motion blur exhibits relatively stable saliency RMSE values across all severities and layers, indicating that the model is robust to this type of perturbation, with consistent differences in saliency maps observed. Conversely, contrast perturbations introduce greater variability in Delta AUC, particularly in the lower layers and at higher severities, reinforcing the observation that contrast is a more disruptive factor. Additionally, Gaussian noise, speckle noise, and brightness perturbations display a smaller range of RMSE as severity increases, with the model showing heightened sensitivity to these perturbations in the deeper layers.

3.2.3 RetinaMNIST

Missing correct plot, waiting for the entire plot to load on nerve

3.2.4 DermaMNIST

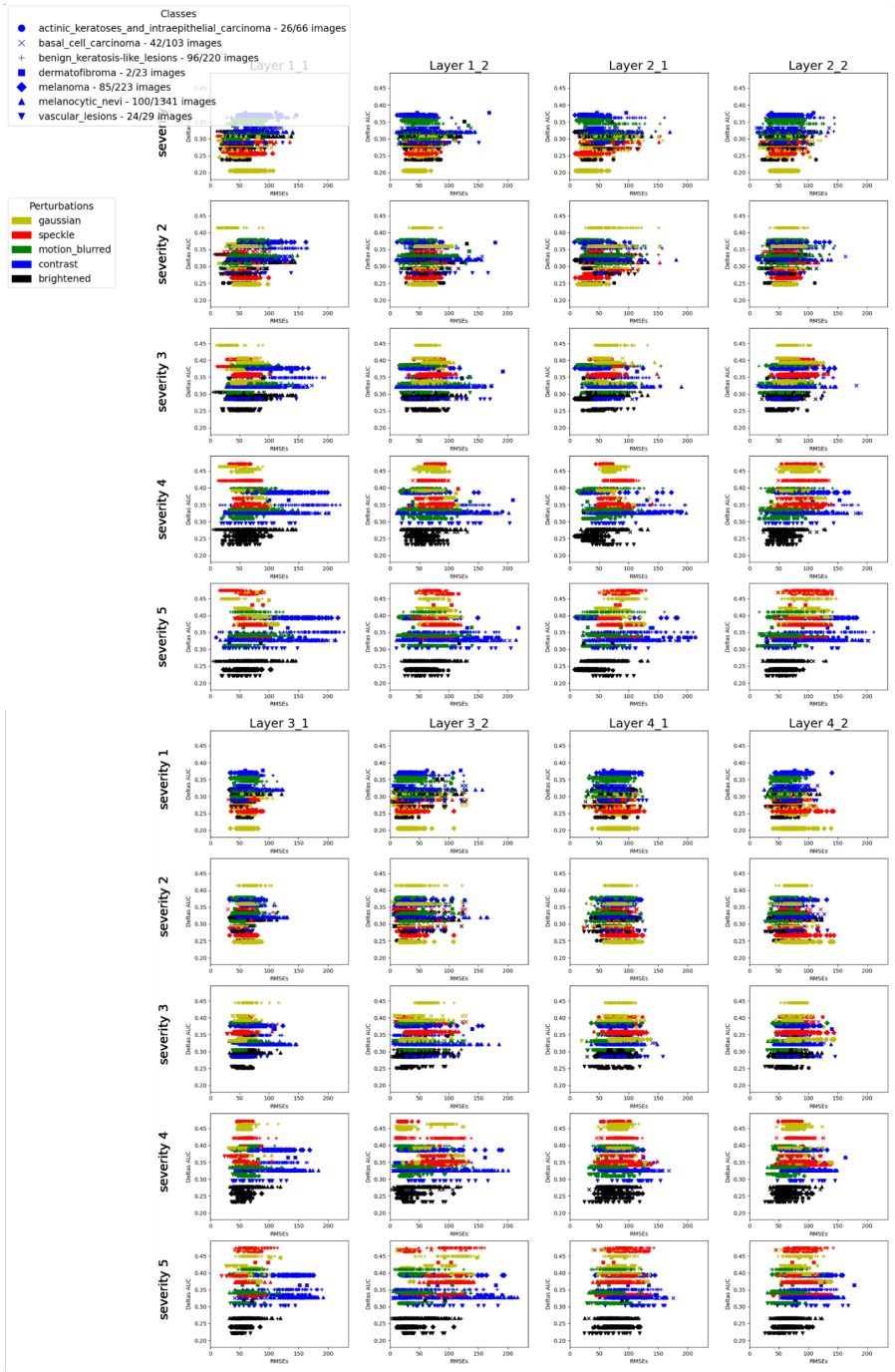

Figure 3.6: Correlation analysis of delta AUC and delta Saliency for the DermaMNIST dataset without augmentation.

This figure presents a series of scatter plots showing the relationship between Delta AUC and RMSE values across various perturbations, target layers, and perturbation severities. The analysis was conducted on the DermaMNIST dataset, specifically on correctly classified images. The axes and layout are the same as in the previous figures. The legend categorizes different classes of skin lesions. Actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions are the classes represented.

Challenges in analysis due to the complexity of the figure:

The figure contains seven different classes, each represented by a different symbol and color. The figure's effectiveness is significantly compromised due to the excessive number of classes and the resulting visual clutter. The dense overlapping of data points makes it difficult to analyze the correlation between Delta AUC and RMSEs across layers, severities, and perturbations. While the intention of the figure is to reveal insights into how perturbations affect the model's performance and saliency maps for the DermMNIST dataset, the overcrowded visual presentation prevents a clear and meaningful analysis. To address these challenges, it would be advisable to separate the classes into individual plots or reduce the number of classes visualized in a single figure, allowing for a more focused and interpretable analysis.

Chapter 4

Discussion and Conclusions

4.1 Discussion

In process

4.2 Conclusions

In process

4.3 Limitations

Despite the advancements presented, several limitations affect the current work. Specifically, the performance of the model, as reflected by the AUC, remains suboptimal for the RetinaMNIST dataset. This indicates that the model's ability to distinguish between classes may not be as effective as desired, potentially affecting its overall utility and accuracy in real-world scenarios. Moreover, the application of Grad-CAM++ for generating heatmaps, while useful, may not consistently yield high-quality visual explanations. This inconsistency can impact the interpretability of the model, making it difficult to understand and trust the model's decision-making process.

Another significant limitation lies in the challenge of explaining complex correlations between features, particularly when visualizing data with a large number of classes. The presence of numerous modifiable inputs, coupled with the high number of classes, adds to the difficulty of discerning and articulating how these features interact and influence the model's outputs. This complexity is further compounded by the potential for nonlinear relationships and interactions among features, which make it challenging to isolate and understand the contributions of individual variables. Additionally, the abundance of parameters and data points on the plots can overwhelm the visual representation, making it difficult to draw clear insights. Consequently, this limitation affects the overall effectiveness of feature analysis and hinders the ability to provide robust explanations that could guide further model development and refinement.

On top of that, the generation of Grad-CAM explanations and the augmentation of datasets for large-scale or real-time applications present substantial computational and temporal challenges. Specifically, the process of augmenting each dataset by a factor of 26 demands significant GPU memory and processing time. This high resource consumption,

particularly when working with large datasets such as those in the MedMNIST collection, may limit the practicality of these methods in real-world applications, where efficiency and scalability are crucial. The large size of some MedMNIST datasets exacerbates these issues, potentially restricting the applicability and effectiveness of the proposed techniques in scenarios requiring rapid or extensive data processing.

Chapter 5

Outlook

Future work could focus on optimizing the computational efficiency of the algorithms developed in this thesis. This can be achieved by exploring advanced techniques for real-time and lightweight dataset augmentation or implementing lightweight augmentation, where only a subset of images is augmented. It could significantly lower resource demands. This optimization would enable training on larger and more complex datasets, thereby enhancing the scalability and applicability of the AI models.

Moreover, expanding the research scope to include more diverse and complex datasets, as well as integrating sophisticated model architectures and interpretability methods, could provide valuable insights into better-identifying patterns associated with classification errors. This approach may lead to the development of more effective mechanisms for flagging incorrect AI results, ultimately offering a better approach to quality control.

It would also be important to seek advice and engage in discussions with clinicians to better understand the practical usefulness and reliability of these indicators. Ensuring they align with clinical needs and enhance the overall quality of the AI systems. Incorporating clinician feedback into the design and evaluation of these mechanisms could also help in fine-tuning the system to address real-world challenges and improve its integration into clinical workflows, thereby fostering greater acceptance and utility in medical practice.

Ultimately, these efforts would aim to build more reliable, efficient, and clinically relevant AI solutions that can significantly enhance the quality of healthcare.

Bibliography

[1] K. Abhishek, A. Jain, and G. Hamarneh. Investigating the quality of dermamnist and fitzpatrick17k dermatological image datasets. arXiv preprint arXiv:2401.14497, 2024.

[2] P. T. M. Anh. Overview of class activation maps for visualization explainability. Unpublished manuscript.

[3] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 839–847. IEEE, 2018.

[4] P. Contributors. Bcewithlogitsloss — pytorch 2.4 documentation, 2024.

[5] P. Contributors. Tensors and dynamic neural networks in python with strong gpu acceleration, 2024.

[6] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? arXiv preprint arXiv:2112.00639, 2021.

[7] A. J. Fetterman, E. Kitanidis, J. Albrecht, Z. Polizzi, B. Fogelman, M. Knutins, B. Wróbĺewski, J. B. Simon, and K. Qiu. Tune as you scale: Hyperparameter optimization for compute efficient training. arXiv preprint arXiv:2306.08055, 2023.

[8] B. Fresz, V. P. Göbels, S. Omri, D. Brajovic, A. Aichele, J. Kutz, J. Neuhüttler, and M. F. Huber. The contribution of xai for the safe development and certification of ai: An expert-based analysis. arXiv, 2024.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.

[10] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261, 2019.

[11] imgaug contributors. imgaug.augmenters.imgcorruptlike — imgaug 0.4.0 documentation. `https://imgaug.readthedocs.io/en/latest/source/api_augmenters_imgcorruptlike.html#imgaug.augmenters.imgcorruptlike.GaussianNoise`.

[12] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2017.

[13] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. arXiv preprint arXiv:1908.01224, 2019.

[14] S. G. K. Patro and K. K. Sahu. Normalization: A preprocessing stage. IARJSET, pages 20–22, March 2015.

[15] S. Poppi, L. Furini, J. Cavazza, F. Galasso, and S. Calderara. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14453–14462, 2021.

[16] pytorch contributors. torch.nn — pytorch 2.4 documentation.

[17] F. D. Salvo, S. Doerrich, and C. Ledig. Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. arXiv preprint arXiv:2406.17536, 2024.

[18] scikit-learn contributors. scikit-learn: A set of python modules for machine learning and data mining. Software.

[19] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 618–626, 2017.

[20] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. Journal of Big Data, 6:60, 2019.

[21] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2014.

[22] S. Suara, A. Jha, P. Sinha, and A. A. Sekh. Is grad-cam explainable in medical images? arXiv, 2023. arXiv:2307.10506.

[23] N. C. Wang, D. C. Noll, A. Srinivasan, J. Gagnon-Bartsch, M. M. Kim, and A. Rao. Simulated mri artifacts: Testing machine learning failure modes. BME Frontiers, 2022:9807590, 2022.

[24] H. Xiong, X. Li, X. Zhang, J. Chen, X. Sun, Y. Li, Z. Sun, and M. Du. Towards explainable artificial intelligence (xai): A data mining perspective. arXiv preprint arXiv:2401.04374, 2024.

[25] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data, 10(1):41, January 2023.

[26] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, 2014.

Appendices

Appendix A

Saliency map variations across layers

A.1 Introduction

...

A.2 Variable Types

...

Appendix B

Correlation between AUC and RMSE

B.1 Section 1

...

B.2 Section 2

...