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 application of Artificial Intelligence (AI) in medical image classification has gained significant momentum in recent years. Despite the rapid advancement and spread of novel AI methodologies, many errors produced by AI systems remain difficult to interpret, as these systems often operate as opaque "black boxes." While previous research has made significant strides in enhancing model transparency, the focus has predominantly been on the final layers of the model, which may limit the understanding of errors occurring earlier in the decision-making process. In contrast, this thesis seeks to explore saliency maps across different layers of the network and investigate the mapping between failure patterns, or so-called failure modes, and MRI sequence-specific image perturbations. It is intended to develop a mechanism for flagging incorrect AI results, thereby providing a new approach to quality control.

Five specific perturbations, at five levels of severity, were applied to the biomedical images of the MedMNIST collection: BreastMNIST, DermaMNIST, PneumoniaMNIST, and RetinaMNIST. To classify these images, a ResNet18 model was employed. Following this, saliency maps were generated across eight layers for both the original and perturbed images using the GradCAM method. The differences between these saliency maps were then quantified using the RMSE metric. Subsequently, the relationship between model interpretability and performance was analyzed by plotting changes in AUC against RMSE values. This approach allowed for the exploration of how variations in saliency maps correlate with model performance. Ultimately, by examining these differences, the research aimed to develop a mechanism for flagging incorrect AI results, thus offering a novel approach to quality control in AI systems.

The study's preliminary results revealed that there is a challenge in explaining complex correlations between features, particularly when visualizing data with a large number of modifiable inputs. Nevertheless, a few observable trends emerged. The linear relationships between delta AUC and RMSE values, especially under contrast and brightened perturbations, suggest that these patterns can indeed serve as proxies for identifying failure modes within the model. However, the findings also indicate that these correlations are often specific to particular datasets and perturbation types. This specificity suggests that while it is possible to flag incorrect AI results using saliency map discrepancies, this approach may be most effective when tailored to specific types of data and perturbations.

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
ID	...	Intrinsic Dimension
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.

One of the key solutions to improving the interpretability of AI models in medical image classification is the use of saliency maps, which are a type of heatmap visualization. Saliency maps highlight the regions of an image that the AI model considers most important for making its classification decisions. By assigning different colors or intensities to these regions, saliency maps provide a visual representation of the features that influence the model's output. This approach provides clinicians and researchers with insights into the AI system's decision-making, facilitating the identification of potential errors or biases. Saliency maps are also essential for quality control, helping detect inconsistencies between the model's focus and clinical expectations. Integrating saliency maps into AI workflows is a crucial step toward demystifying deep learning models and building greater trust in AI-driven diagnostics.

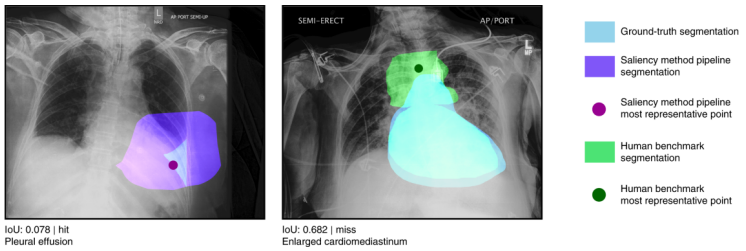

Figure 1.1: Comparison of saliency method pipeline and human benchmark segmentations on chest X-ray images. Source: `https://www.nature.com/articles/s42256-022-00536-x`

A review of existing literature reveals numerous studies that have leveraged saliency maps to enhance the explainability of deep learning models in image classification. For instance, the work by Simonyan, Vedaldi, and Zisserman (2014) [22] pioneered the use of saliency maps to visualize the features that influence a model's final decision, primarily focusing on the last convolutional layer. Similarly, Chattopadhyay et al. (2018) [3] extended this concept through the development of Grad-CAM++, which generalizes gradient-based visual explanations for deep convolutional networks, also emphasizing the last layer. While these studies have made significant contributions to improving model transparency, they largely concentrate on the final layers of the network, which may limit the understanding of errors occurring earlier in the decision-making process. In contrast, this thesis aims to explore saliency maps across various layers of the network, to identify and analyze error patterns throughout the model. By doing so, the research seeks to develop more comprehensive quality control mechanisms that go beyond the final classification layer, thereby offering deeper insights into the sources of errors and enhancing the reliability of AI in clinical settings.

Furthermore, a recent study by Konz and Mazurowski titled "Pre-processing and Compression: Understanding Hidden Representation Refinement Across Imaging Domains via Intrinsic Dimension" [12] provides crucial insights into how the intrinsic dimension (ID) of neural network representations evolves across different layers, particularly in medical imaging domains. Their findings reveal a strong correlation between the peak representation ID within the network and the ID of the dataset in its input space, highlighting a deep connection between the information content of a model's learned representations and the intrinsic properties of the training data. While this research does not directly engage with intrinsic dimension analysis, it offers a foundational perspective on how the inherent characteristics of medical data might influence model behavior and, consequently, the saliency maps generated across different layers. This understanding could enrich the interpretation of saliency maps, particularly in identifying and analyzing the layers where the model's focus may align with or diverge from the expected medical features.

Building on these insights, the primary hypothesis of this thesis 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. By distinguishing these gradient patterns, the study aims to uncover failure patterns within the model, offering a systematic approach to understanding and mitigating errors. Consequently, this Master's thesis aims to investigate the mapping between failure patterns or so-called failure modes and Magnetic Resonance Imaging (MRI) sequence-specific image perturbations [24]. 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 and Saliency maps. 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) [26] 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 [26] 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.

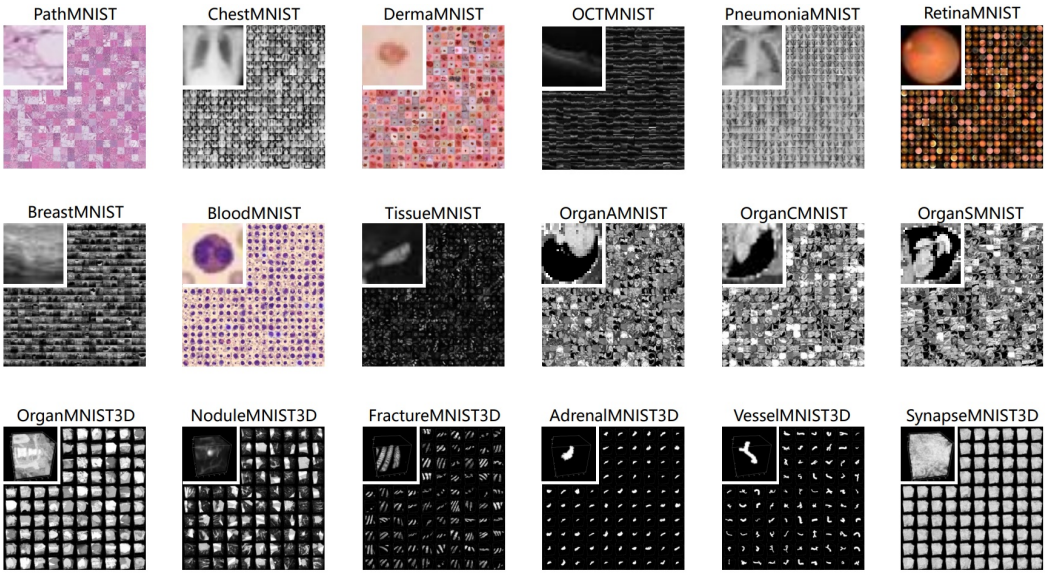

Figure 2.1: Overview of MedMNIST Datasets. Source: `https://medmnist.com/`

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 (2)	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 [15] 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 [21] 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) [27] architecture 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.

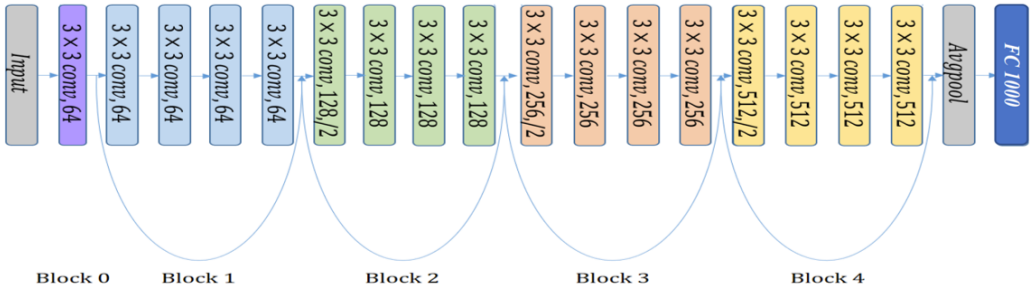

Figure 2.2: ResNet18 model architecture, showing the input layer, 18 convolutional layers, average pooling, and the fully connected layer. Source: `https://www.semanticscholar.org/reader/57dc772091632ea97adfb9a7f8c45dafd40a1e1a`

ResNet18 has demonstrated competitive performance on prominent benchmark datasets [26], 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. 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 [13] 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 [18] 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.

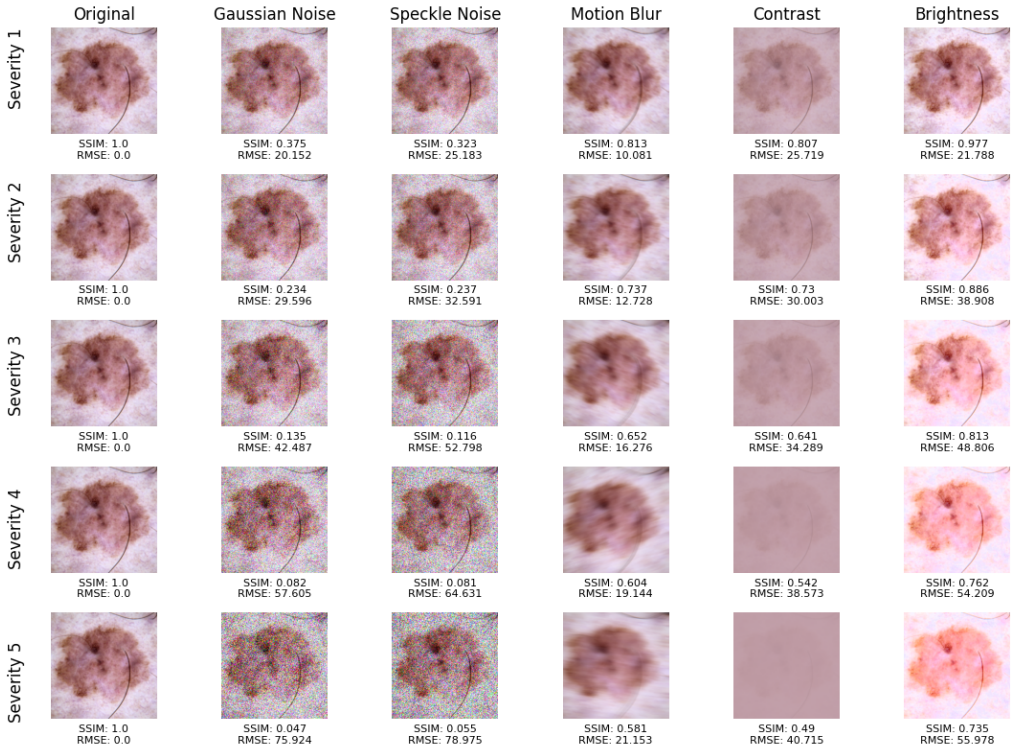

Figure 2.3: Generation of 5 types of perturbations at severity levels ranging from 1 to 5 applied to the original image. The dataset used is Dermamnist, a 2D dataset. The image shown is a 'melanoma,' with dimensions of 224x224 pixels.

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 [25] 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 [22], 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) [16] 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 [22]. 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.

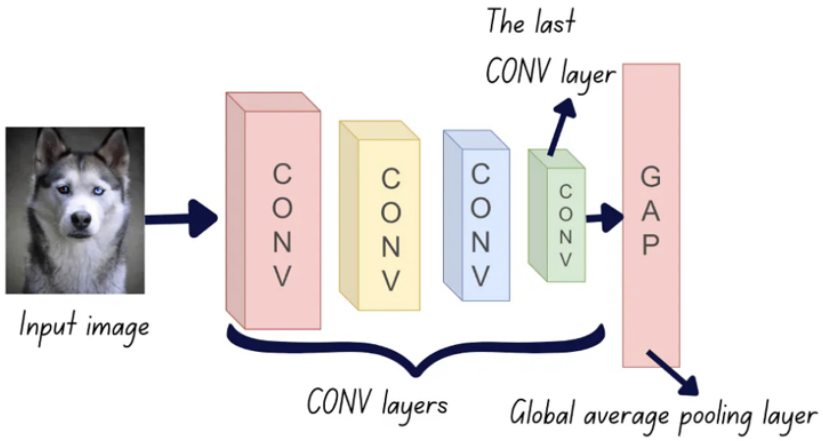

Figure 2.4: CNN processing an image of a dog, with convolutional layers and a GAP layer, used for generating the Class Activation Map for feature visualization. Source: `https://www.pinecone.io/learn/class-activation-maps/`

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 [22], 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.

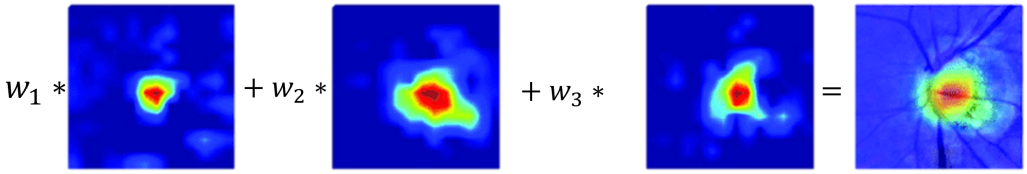

Figure 2.5: Generation of a class activation map with a weighted sum of the feature maps.

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 [20] 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.

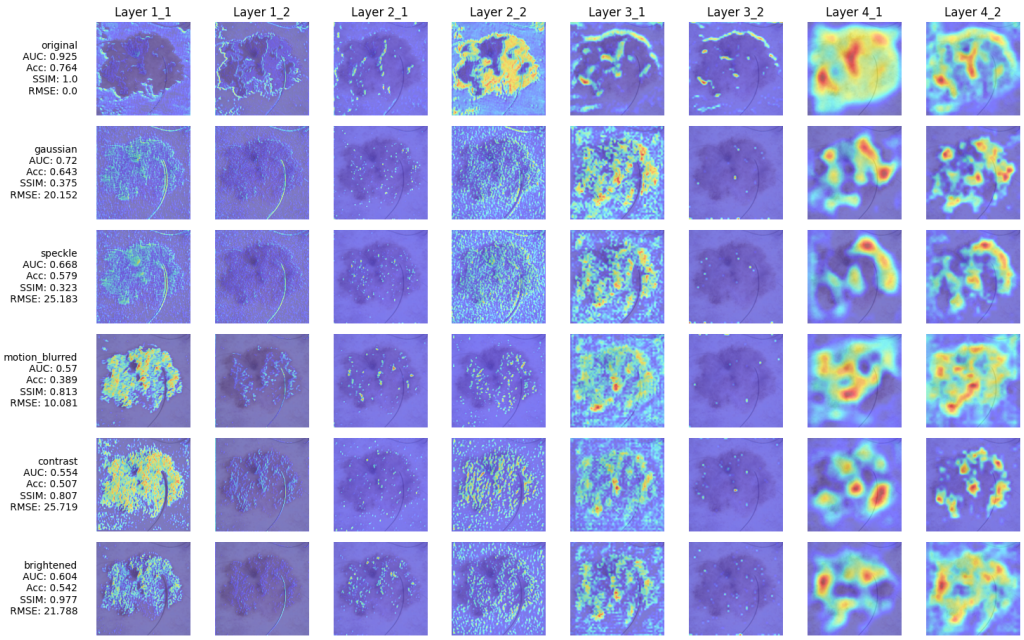

Figure 2.6: Saliency map generation across various layers and different perturbations at severity level 1. The x-axis represents the model's layers, while the y-axis corresponds to the different perturbations. The dataset used is Dermamnist, which is a 2D dataset with images of size 224x224. The model 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 [17].

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

$\ell(x, y)$	:	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.

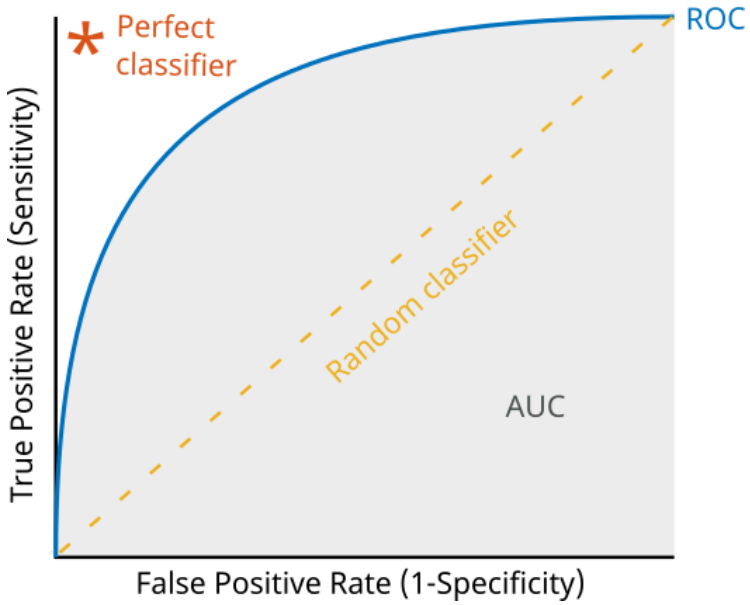

Figure 2.7: The ROC curve illustrates the performance of a classifier. The True Positive Rate (TPR), also known as sensitivity or recall, is plotted on the y-axis, while the False Positive Rate (FPR) is shown on the x-axis. The curve demonstrates the trade-off between sensitivity and specificity as the 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.

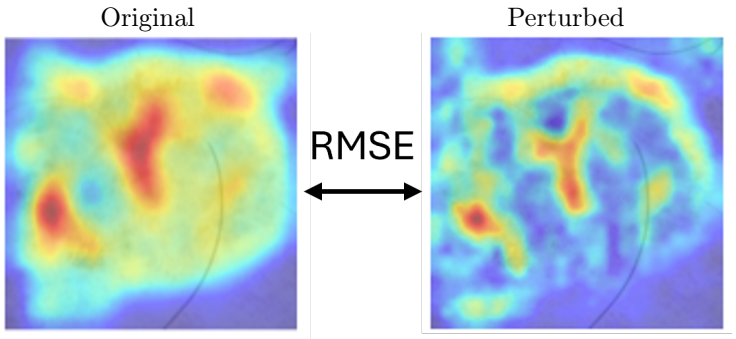

Figure 2.8: RMSE between two saliency maps of the 2D Dermamnist dataset. The left side shows the saliency map for the original image, and the right side displays the saliency map for the perturbed image.

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.

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. Here is the GitHub repository: `https://github.com/cailloumagic/MedMNIST`

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 non-augmented datasets

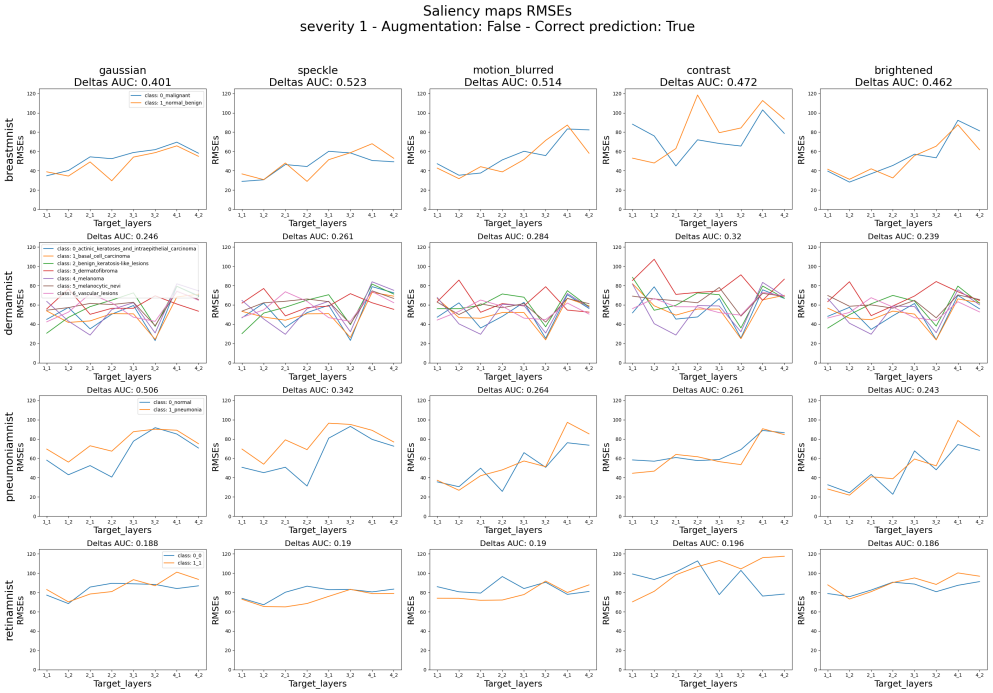

Figure 3.1: RMSEs of saliency maps across different perturbations and target layers for correctly classified images of severity 1 without augmentation. There are four distinct datasets: breastmnist, dermamnist, pneumoniamnist, and retinalmnist. Each row corresponds to a different dataset, while each 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.

General Observations:

Across all the datasets examined, several general observations can be made regarding the model's behavior under perturbations. First, the BreastMNIST dataset demonstrates the most pronounced degradation in performance with consistently high Delta AUC values. Datasets like BreastMNIST, DermaMNIST and RetinaMNIST show specific vulnerabilities, particularly in deeper layers. Commonly, there is an observable trend where RMSE values increase as the layers deepen, with significant peaks at certain layers, such as 4.1, across multiple datasets. Additionally, contrast perturbations tend to induce greater fluctuations in RMSE across multiple datasets, highlighting a potential weakness in handling variations in contrast. Despite these commonalities, there are dataset-specific sensitivities; for instance, DermaMNIST shows erratic RMSE behavior, particularly in the dermatofibroma class, and RetinaMNIST exhibits low deltas AUC, reflecting a low difference in performance between the original and perturbed datasets. Overall, these observations suggest that while the model demonstrates some robustness, certain perturbations, particularly contrast changes, and specific layers, present significant variations in RMSE and performance.

Dataset-Specific Observations:

BreastMNIST

The model's performance significantly deteriorates under even small perturbations, as indicated by the high Delta AUC values across all types of perturbations. This suggests that the model is particularly sensitive to variations in input data. The RMSE trends show a consistent increase towards deeper layers, with a peak at layer 4.1 across all perturbation types, indicating that the deeper layers of the model are more prone to error accumulation under perturbed conditions. Additionally, while the RMSE lines for different classes remain relatively close to each other for most perturbations, the contrast perturbation stands out due to its substantial fluctuation. This fluctuation may indicate that the BreastMNIST dataset is particularly sensitive to changes in contrast, which could be a critical factor affecting model robustness when dealing with varying image qualities or conditions.

DermaMNIST

When examining the impact of different perturbations on the DermaMNIST dataset, the Delta AUC values reflect smaller effects compared to those observed in the BreastMNIST dataset. The analysis of RMSE, reveals more erratic behavior for the DermaMNIST dataset, with RMSE values scattered across the target layers. Despite this variability, there is a consistent low peak observed at layer 3.2 and a pronounced high peak at layer 4.1, indicating layers where the saliency map differences are more or less pronounced. The RMSE values for different classes remain fairly consistent, except for the dermatofibroma class, which exhibits a notable high peak at layer 3.2. This behavior contrasts with the other layers and classes, suggesting that the dermatofibroma class is particularly sensitive to certain perturbations, which could lead to significant changes in the saliency maps and impact the interpretability of the model's decision-making process for this class.

PneumoniaMNIST

Similar to the BreastMNIST dataset, the RMSE values for the RetinaMNIST dataset exhibit a more stable trend, with a gradual increase across the layers. However, a significant drop in RMSE is observed at layer 2.2 for the normal class and a high peak at layer 4.1, suggesting that these layers may have varying levels of sensitivity to perturbations. Among the various perturbations applied, Gaussian noise results in the highest Delta AUC (0.506), indicating a substantial impact on the saliency maps and a notable difference in the model's sensitivity to this perturbation. In contrast, other perturbations produce lower Delta AUC values.

RetinaMNIST

The RMSE values in the RetinaMNIST dataset are relatively flat, displaying only minor fluctuations across the layers, although consistently high values are observed. Notably, peaks are observed at layer 2.2 for every perturbation, indicating a specific sensitivity at this layer across different conditions. The overlapping RMSE values between the different classes suggest low separability based on the saliency maps, implying that the model's response is similar across classes despite the perturbations. Regarding the perturbations themselves, all exhibit low Delta AUC values, ranging from 0.162 to 0.169, which indicates minimal divergence in the saliency maps between the original and perturbed images. However, once again, the RMSE curve for the contrast perturbation stands out, showing higher values and more pronounced peaks compared to other perturbations, suggesting that the RetinaMNIST dataset may be particularly sensitive to contrast changes.

3.1.2 Test on augmented dataset

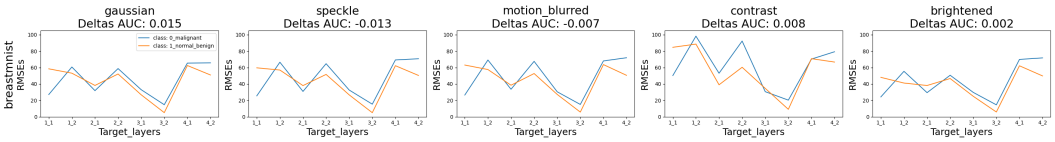

Figure 3.2: RMSEs for the augmented BreastMNIST dataset across different perturbations for incorrectly predicted images of severity 1. It focuses exclusively on the BreastMNIST dataset due to the high memory usage required by the other datasets. The layout and axes remain consistent with the previous 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.

The RMSE values for the augmented dataset exhibit more erratic and pronounced fluctuations across different layers compared to the non-augmented dataset. Similar to the non-augmented datasets, the model demonstrates increased sensitivity in RMSE to contrast variations. Despite the application of data augmentation, these perturbations continue to introduce greater instability in the model's delta saliency maps, indicating that augmentation has not fully mitigated the model's sensitivity to such factors. The layers consistently show alternating high and low RMSE peaks between two layers across all perturbations, suggesting that even with augmentation, the saliency maps remain sensitive to perturbations. However, the overall impact on performance is lower compared to the non-augmented dataset, indicating some level of robustness has been gained through augmentation, but not enough to fully stabilize the model against all perturbations.

3.1.3 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 without augmentation

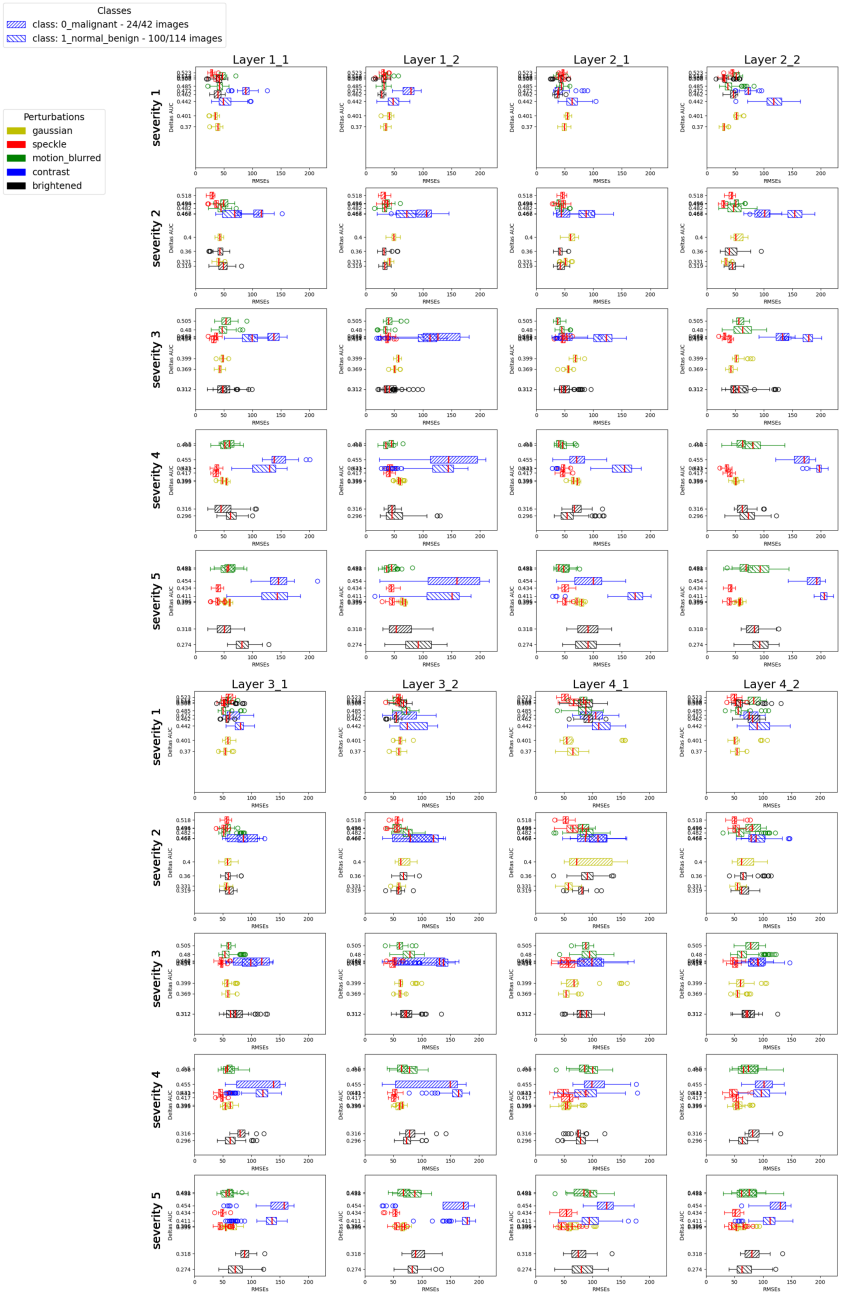

Figure 3.3: Correlation analysis of delta AUC and delta Saliency for the BreastMNIST dataset without augmentation.

This figure presents a series of box 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 at least were correctly predicted.

General observations:

The analysis of RMSE and Delta AUC values across different layers and perturbation severities reveals several key trends. One consistent observation is that deeper layers, particularly layers like 4.1 and 4.2, tend to exhibit broader and more erratic RMSE ranges across most perturbation types. This indicates that these deeper layers are generally more sensitive to perturbations, resulting in greater variability in the saliency maps.

A common trend observed across perturbations, especially contrast and brightened conditions, is a linear relationship where increasing severity leads to higher RMSE values and a decrease in Delta AUC. This correlation suggests that as the perturbations become more intense, the model's ability to maintain consistent performance deteriorates, as reflected by larger discrepancies in the saliency maps and reduced classification accuracy (lower Delta AUC).

In contrast, Gaussian noise and motion blur demonstrate relatively stable RMSE values across layers and severities, though with broader ranges in the deeper layers. The consistent patterns in Delta AUC and RMSE correlations across different perturbations and severities highlight that the model's robustness is significantly challenged by certain types of perturbations, particularly those that introduce variations in contrast or brightness.

Perturbations-specific observations:

Gaussian Noise

The Gaussian noise values remain relatively stable across different severity levels, showing minimal fluctuation. However, a broader range of RMSE values is observed in the deepest layers (4.1 and 4.2), indicating that these layers may be more susceptible to variability introduced by the noise at higher severities.

Speckle Noise

Similar to the Gaussian noise perturbation, an increase in the range of RMSE values is observed in the deeper layers. Additionally, as the severity increases, there is a corresponding decrease in Delta AUC, indicating an enhancement in model performance under more severe perturbations.

Motion Blurred

The performance of the model on the motion blurred perturbation dataset remains stable across layers and severities. However, the range of RMSE values differs significantly from layer to layer. The deeper layers, specifically 2.2, 3.2, 4.1, and 4.2, appear more prone to exhibiting wider and more inconsistent RMSE distributions. Additionally, as the severity of the perturbation increases, the range of data broadens, with a slight increase in RMSE values accompanying it. This suggests that while the model maintains overall stability, the deeper layers are more vulnerable to variability under stronger motion blur conditions.

Contrast

First, what stands out is the behavior of the contrast perturbation (blue). The RMSE values for contrast are notably higher for both classes compared to the other perturbations, with a much wider range of data points. This wide range indicates that contrast perturbations introduce a significant amount of variability in the saliency maps, making the model's output more inconsistent. This variability is not only observed across different layers but also varies substantially with the severity of the perturbations. These observations reinforce previous findings that the BreastMNIST dataset is particularly sensitive to contrast variations, making it a critical factor in the model's robustness assessment. Additionally, across all layers, a noticeable linear trend emerges: as the severity of the contrast perturbation increases, the Delta AUC consistently decreases while the RMSE increases. This trend suggests that the model becomes progressively less reliable in maintaining consistent performance under greater contrast perturbation, leading to larger discrepancies in the saliency maps and a more pronounced decline in overall model accuracy.

Brightened

It is observed that as the severity of the perturbation increases, the range of RMSE values becomes wider. This effect is particularly pronounced at severity level 5, where RMSE values are notably higher. As with the contrast perturbation, a linear relationship is observed: as the severity increases, the Delta AUC decreases, and the RMSE values rise. However, there is no significant variation in this behavior across different layers, indicating that the impact of increased severity is consistent throughout the network.

3.2.2 BreastMNIST with augmentation

Figure 3.4: Correlation analysis of delta AUC and delta Saliency for the BreastMNIST dataset with augmentation.

This figure presents a series of box 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, legends and layout are the same as in the previous figure.

General Observations:

First, it is observed that the overall Delta AUC values for all perturbations are very low and concentrated after the data augmentation. This indicates that the performance between the original and perturbed datasets remains quite similar across the model and different severity levels, which aligns with the intended outcome of the augmentation process. This consistency suggests that the augmentation successfully mitigates the variability introduced by the perturbations, leading to a stable model performance regardless of the applied perturbation. Regarding the correlation between Delta AUC and RMSE values, no significant correlation with performance was found. This is primarily because the performance remains relatively consistent across different perturbations and severities after data augmentation.

From the observations across various perturbations, a consistent pattern emerges: the RMSE values generally remain stable across most layers, with only minor fluctuations observed in the middle layers (3.1 and 3.2), where a decrease in RMSE values is common. This suggests that the model exhibits reduced sensitivity to perturbations in these specific layers, leading to more stable saliency maps. Notably, the contrast perturbation continues to show a broader range of RMSE values, indicating greater variability in saliency map responses compared to other perturbations, though this variability is less pronounced in the deeper layers and at lower severity levels. Additionally, the RMSE values for different perturbations tend to converge in the deeper layers, where they are almost all the same, suggesting that the model's response becomes more uniform as the data progresses through the network.

3.2.3 PneumoniaMNIST

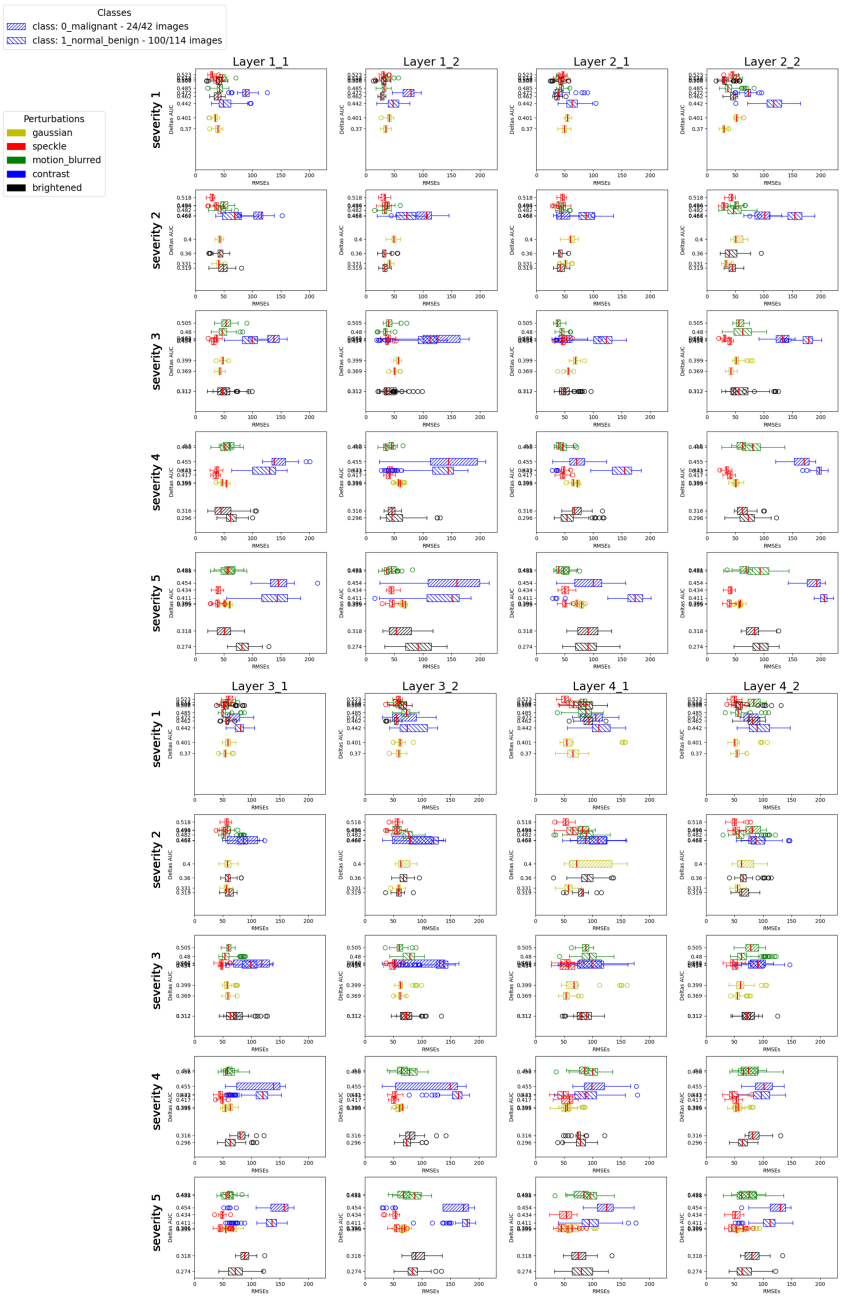

Figure 3.5: Correlation analysis of delta AUC and delta Saliency for the PneumoniaMNIST dataset without augmentation.

This figure presents a series of box 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 at least correctly predicted, while for the pneumonia class, 100 out of 390 images were plotted.

General Observations:

A consistent observation is that deeper layers, particularly layers 4.1 and 4.2, are more prone to variability, as evidenced by the broader range of RMSE values under perturbations like Speckle Noise, Motion Blurred, Contrast, and Brightened. Conversely, shallower layers (1.1 and 1.2) show more stable RMSE values, except under contrast and brightened perturbations, where these layers also experience increased fluctuation as severity rises.

The correlation between Delta AUC and RMSE is evident in several cases, particularly with contrast and brightened perturbations. A clear trend emerges: for the contrast perturbation, as severity increases, RMSE values rise and Delta AUC decreases, indicating an increase in model performance. In contrast, for the brightened perturbation, both RMSE values and Delta AUC increase with higher severity, reflecting a decline in model performance under these conditions.

However, some perturbations, like Gaussian Noise, exhibit little to no correlation between Delta AUC and RMSE, suggesting that the model's performance remains relatively stable despite changes in severity, although RMSE still slightly increases in deeper layers. Overall, these observations suggest that while the model handles certain perturbations with more stability, deeper layers are consistently more vulnerable to errors.

Perturbations-specific observations:

Gaussian Noise

The Gaussian perturbation exhibits a relatively stable and concentrated range of RMSE values across layers and severities, with only slight increases in RMSE as the severity level rises. Despite these slight changes, there is no apparent trend or correlation between the model's performance and the RMSE differences observed in the saliency maps.

Speckle Noise

For the Speckle noise perturbation, a widening of the RMSE range is observed, particularly in the deeper layers (4.1 and 4.2), reflecting some inconsistency in the model's Saliency maps at these layers. In contrast, the RMSE values in the shallower layers remain relatively concentrated, indicating more stable behavior in the earlier stages of the network. This suggests that while the model generally handles speckle noise well, the deeper layers are more vulnerable to variability introduced by this perturbation.

Motion Blurred

The Motion blurred perturbation shows minimal fluctuations across layers and severity levels, suggesting that this type of perturbation has a uniform impact on the model's performance and saliency maps throughout the network. It indicates that the severity level does not significantly influence the results, further implying that the model maintains stable behavior in the presence of motion blur, regardless of its intensity.

Contrast

Similar to the BreastMNIST dataset, the PneumoniaMNIST dataset shows larger RMSE value ranges for the contrast perturbation compared to other perturbations. This effect is particularly pronounced in the shallow layers (1.1 and 1.2), where the range of RMSE values increases with each successive severity level. Furthermore, there is a significant difference in performance between the normal class and the pneumonia class, with the normal class showing more variability in RMSE values. Additionally, there is a clear linear correlation in the normal class: as the severity of the contrast perturbation increases, the Delta AUC decreases while the RMSE values increase. However, this trend is accompanied by a significant number of outliers in the normal class, indicating variability and potential instability in the model's response to contrast perturbations.

Brightened

This type of perturbation shows more fluctuation across layers and severity levels. Specifically, the first layers (1.1 and 1.2) and the last layers (4.1 and 4.2) exhibit a broader range of RMSE values as severity increases, while the middle layers tend to show more concentrated values. Furthermore, a clear trend emerges with increasing severity, where each layer experiences higher RMSE values and higher Delta AUC, reflecting a decrease in overall model performance. This indicates that the brightened perturbation not only introduces greater variability in the model's saliency maps but also adversely impacts its accuracy, particularly in the shallower and deeper layers.

3.2.4 RetinaMNIST

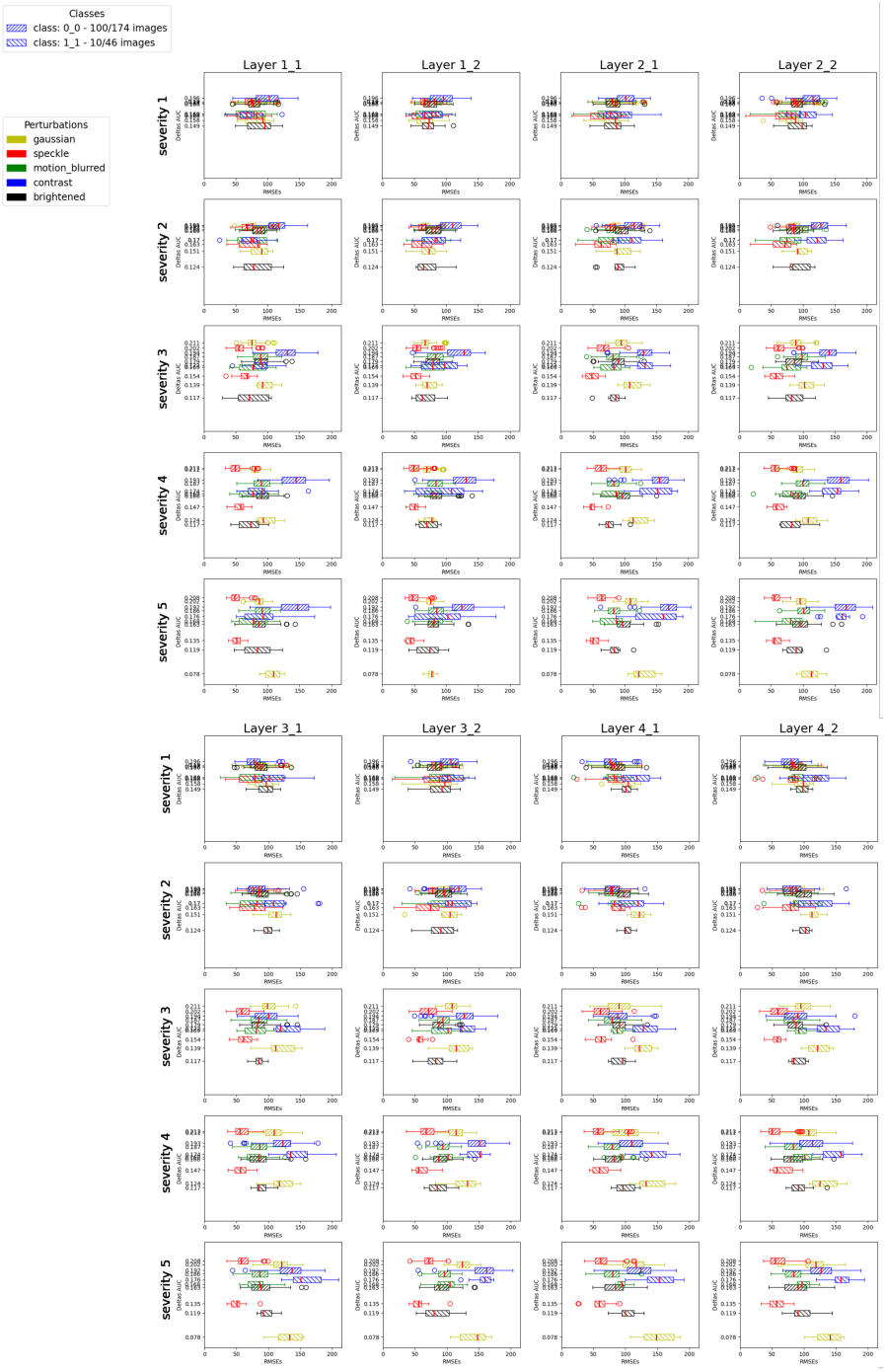

Figure 3.6: Correlation analysis of delta AUC and delta Saliency for the RetinaMNIST dataset without augmentation.

This figure presents a series of box plots showing the relationship between Delta AUC and RMSE values across various perturbations, target layers, and perturbation severities. The analysis was conducted on the RetinaMNIST dataset, specifically on correctly classified images. The axes and layout are the same as in the previous figure. The legend indicates four classes.

General observations:

What is striking about the RMSE values is that they are relatively more spread out for each perturbation across all subplots than in other datasets. This broader range suggests greater variability in the saliency maps under different conditions. Additionally, the Delta AUC values are consistently quite small, indicating only a minor difference in performance between the original and perturbed datasets.

Across the various perturbations and severities, several key trends emerge in the model's behavior. A consistent pattern is the correlation between increasing severity and rising RMSE values, especially in deeper layers. This indicates that the deeper layers are more sensitive to perturbations, leading to greater variability in saliency maps. For perturbations like Gaussian Noise and Contrast, a linear relationship between severity, Delta AUC, and RMSE values is evident, suggesting a clear connection between model performance and the discrepancies in saliency maps as the perturbations intensify.

In contrast, perturbations like Speckle Noise and Motion Blurred exhibit stable model performance with minimal fluctuations across layers and severities. Although these perturbations affect the saliency maps, the impact on Delta AUC and RMSE is less pronounced, indicating that the model handles these perturbations more consistently.

Perturbations-specific observations:

Gaussian Noise

What is first notable is the very low Delta AUC value, indicating that the model's performance on the original dataset and the perturbed dataset is quite similar. Additionally, there is a clear increase in RMSE values as the model processes deeper layers and as severity levels rise, particularly for class 1. A linear relationship is also observed: as the severity increases, the Delta AUC decreases, and the RMSE values increase, suggesting a direct correlation between model performance and saliency maps for various severities.

Speckle Noise

For the Speckle Noise perturbation, it is observable that as severity levels increase, the range of RMSE values becomes thinner and more concentrated. However, there is no significant fluctuation in either model performance or RMSE values across different layers and severity levels. This suggests that the model maintains a stable response to Speckle Noise, with little variation in saliency map differences and performance as the perturbation severity changes.

Motion Blurred

The Motion blurred perturbation has a consistent impact and does not fluctuate significantly across different layers and severity levels. The effect of the perturbation is present, but it remains steady throughout the model, indicating that while the model's performance and saliency maps are affected, they do not exhibit notable variability as the perturbation severity increases or as the data passes through different layers.

Contrast

As with the previous datasets, the Contrast perturbation exhibits a wide range of RMSE values across all subplots, with noticeable variations across different layers. Additionally, these RMSE values tend to be higher with increasing severity levels and as the data progresses deeper into the layers. This pattern suggests that the model's sensitivity to contrast variations increases both with severity and layer depth, leading to greater discrepancies in the saliency maps.

Brightened

As with the last perturbation, similar to the contrast, the RMSE values increase with both rising severity levels and deeper layers. However, the inconsistency in RMSE values varies depending on the layer, without reflecting any specific or consistent trend. Despite these fluctuations, the model's performance remains relatively stable across the different levels of perturbed datasets, indicating that the overall impact of the brightened perturbation does not vary significantly with severity.

3.2.5 DermaMNIST

For this last dataset, due to the high number of classes (7), the data could not be displayed in its entirety without compromising readability. To maintain clarity and ensure the data remains interpretable, each perturbation was plotted on a separate graph, allowing for a more focused and detailed analysis of the RMSE values and Delta AUC across different classes and severity levels.

Gaussian Noise

Figure 3.7: Correlation between delta AUC and delta Saliency for the DermaMNIST dataset under Gaussian noise perturbation without augmentation.

This figure presents a series of box plots showing the relationship between Delta AUC and RMSE values across the Gaussian perturbation, 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.

Gaussian Noise observations:

It is observed that there are minimal variations in Delta AUC and RMSE values between classes within each subplot. However, a broader range of values is evident in layers 2.2 and 3.2, indicating inconsistencies in the saliency maps at these levels. Additionally, a noticeable shift in performance is observed from severity 1 to 5, with Delta AUC values increasing as severities rise, signaling a decline in performance. This trend is accompanied by a linear correlation between Delta AUC and RMSE values, where both metrics increase with higher severity levels, further highlighting the model's growing instability under more severe perturbations.

Speckle Noise

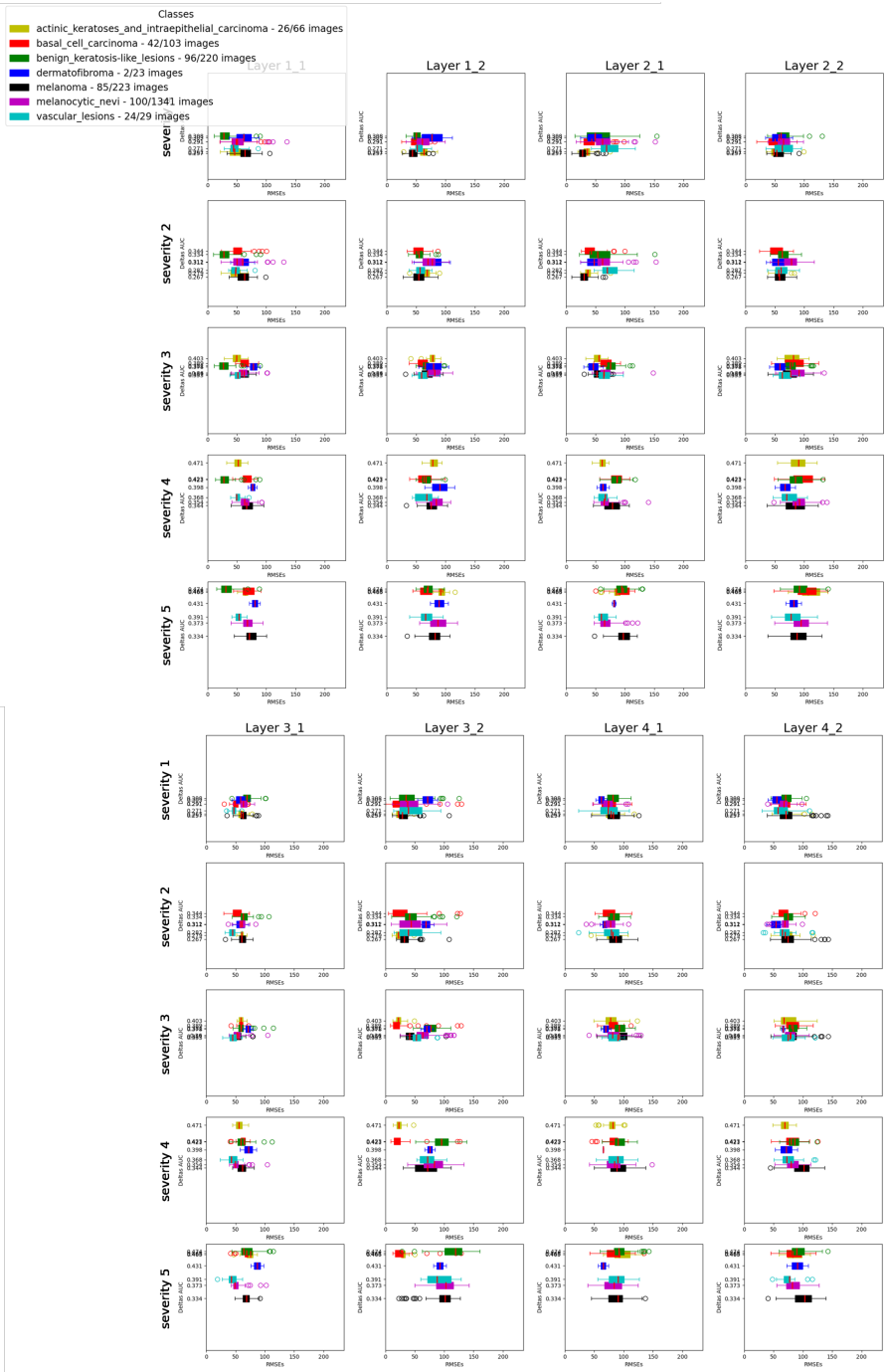

Figure 3.8: Correlation analysis of delta AUC and delta Saliency for the DermaMNIST dataset without augmentation.

This figure presents a series of box plots showing the relationship between Delta AUC and RMSE values across the Speckle perturbation, target layers, and perturbation severities. The analysis was conducted on the DermaMNIST dataset, specifically on correctly classified images. The axes, legend and layout are the same as in the previous figures.

Speckle Noise observations:

The performance between classes remains quite similar across severities, except for the two highest severity levels, where an increase in Delta AUC is observed for various classes, indicating a worsening in performance. Additionally, the differences in RMSE metrics between the different classes are minimal, suggesting consistent saliency map behavior across classes. Overall, similar to the pattern observed with Gaussian noise, a subtle trend emerges for numerous classes where both Delta AUC and RMSE metrics increase with rising severity levels, indicating a linear correlation between these metrics and the increasing perturbation severity.

Motion blurred

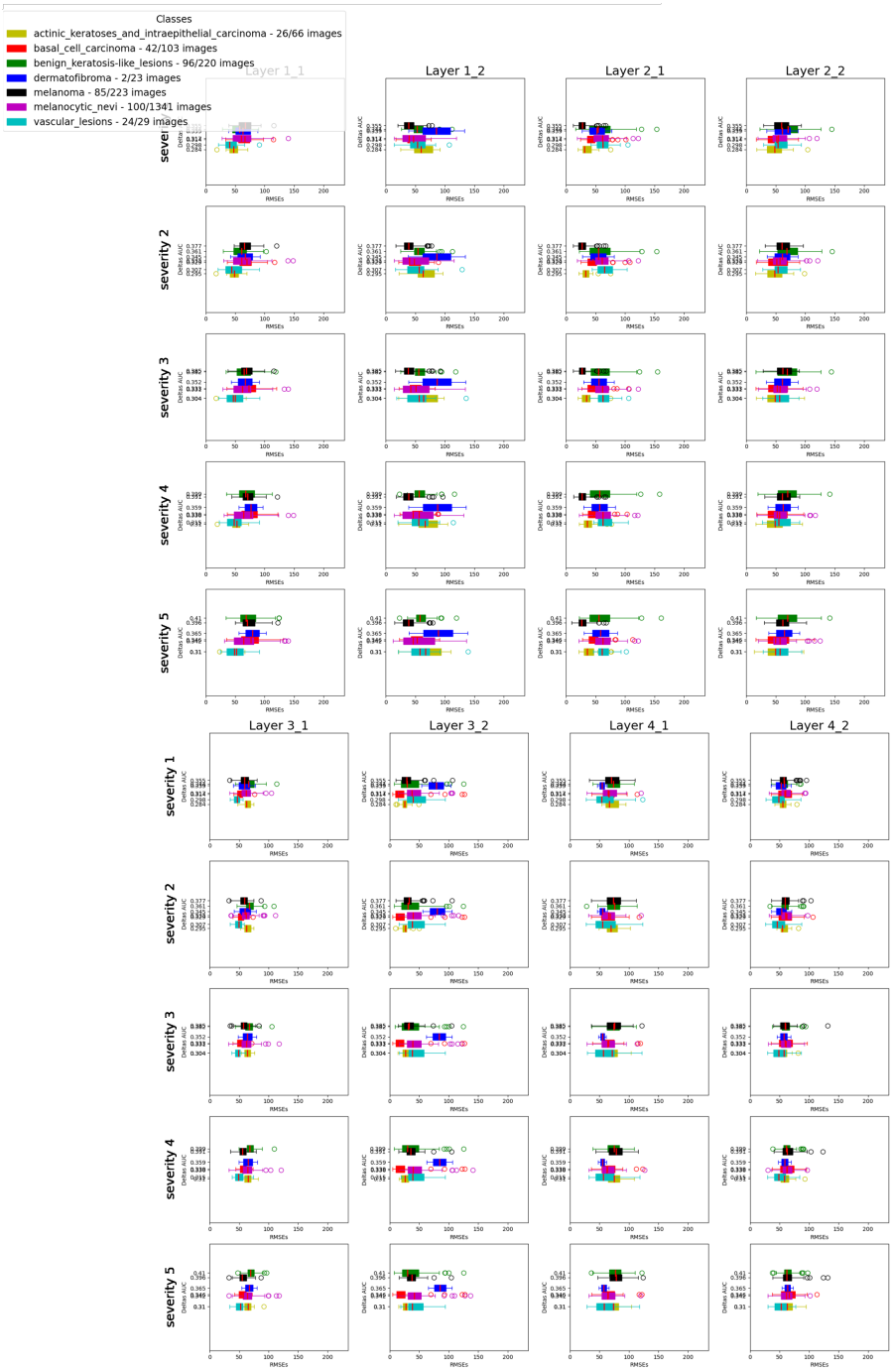

Figure 3.9: Correlation between delta AUC and delta Saliency for the DermaMNIST dataset under Speckle noise perturbation without augmentation.

This figure presents a series of box plots showing the relationship between Delta AUC and RMSE values across the Motion blurred perturbation, target layers, and perturbation severities. The analysis was conducted on the DermaMNIST dataset, specifically on correctly classified images. The axes, legend and layout are the same as in the previous figures.

Motion blurred observations:

To begin with, the RMSE values remain similar across different classes for most severities and layers, except for layers 1.2 and 3.2, where more variation is observed. Additionally, in the deeper layers, these RMSE values become more concentrated and stable, indicating less variability in the saliency maps as the data progresses through the network. Despite this, there appears to be no clear or obvious pattern in these subplots. The overall performance remains stable across subplots, while the RMSE values exhibit rather erratic behavior, lacking a consistent trend or correlation with the severity levels and layers.

Contrast

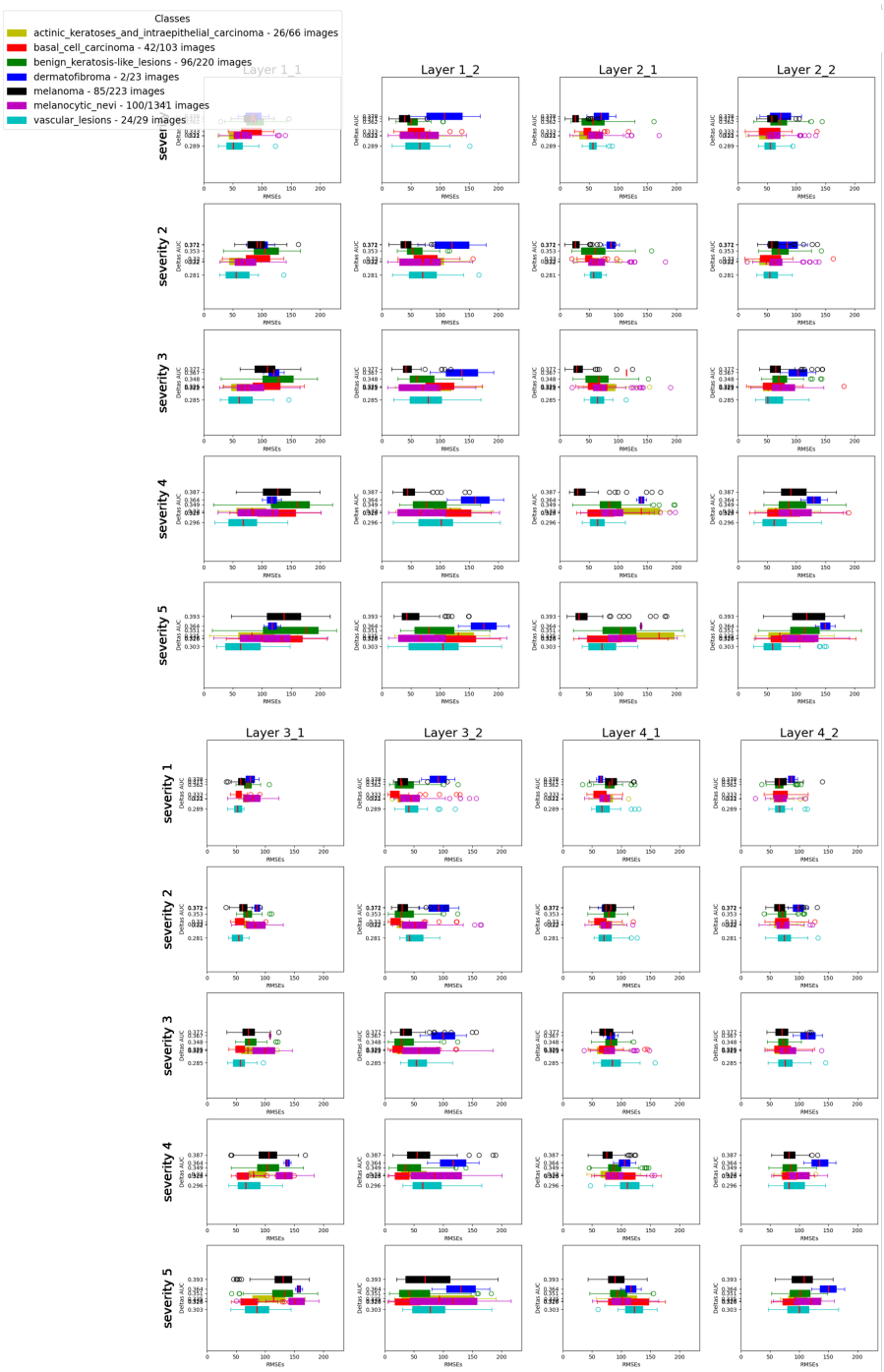

Figure 3.10: Correlation between delta AUC and delta Saliency for the DermaMNIST dataset under Speckle noise perturbation without augmentation.

This figure presents a series of box plots showing the relationship between Delta AUC and RMSE values across the Contrast perturbation, target layers, and perturbation severities. The analysis was conducted on the DermaMNIST dataset, specifically on correctly classified images. The axes, legend, and layout are the same as in the previous figures.

Contrast observations:

As always, perturbing the dataset with contrast results in a broader range of data compared to other perturbations. Layer-wise, the RMSE values are more concentrated and consistent, particularly in layer 3.1 and the deeper layers 4.1 and 4.2. Additionally, the delta Saliency maps show an increase as the layers become deeper, indicating that the saliency map discrepancies grow with depth. Moreover, a slight trend is observed across all classes: as the severity level increases, both the RMSE and Delta AUC metrics rise, creating a linear trend across the perturbed datasets. This suggests that contrast perturbations consistently lead to greater variability and reduced model performance as the severity increases.

Brightened

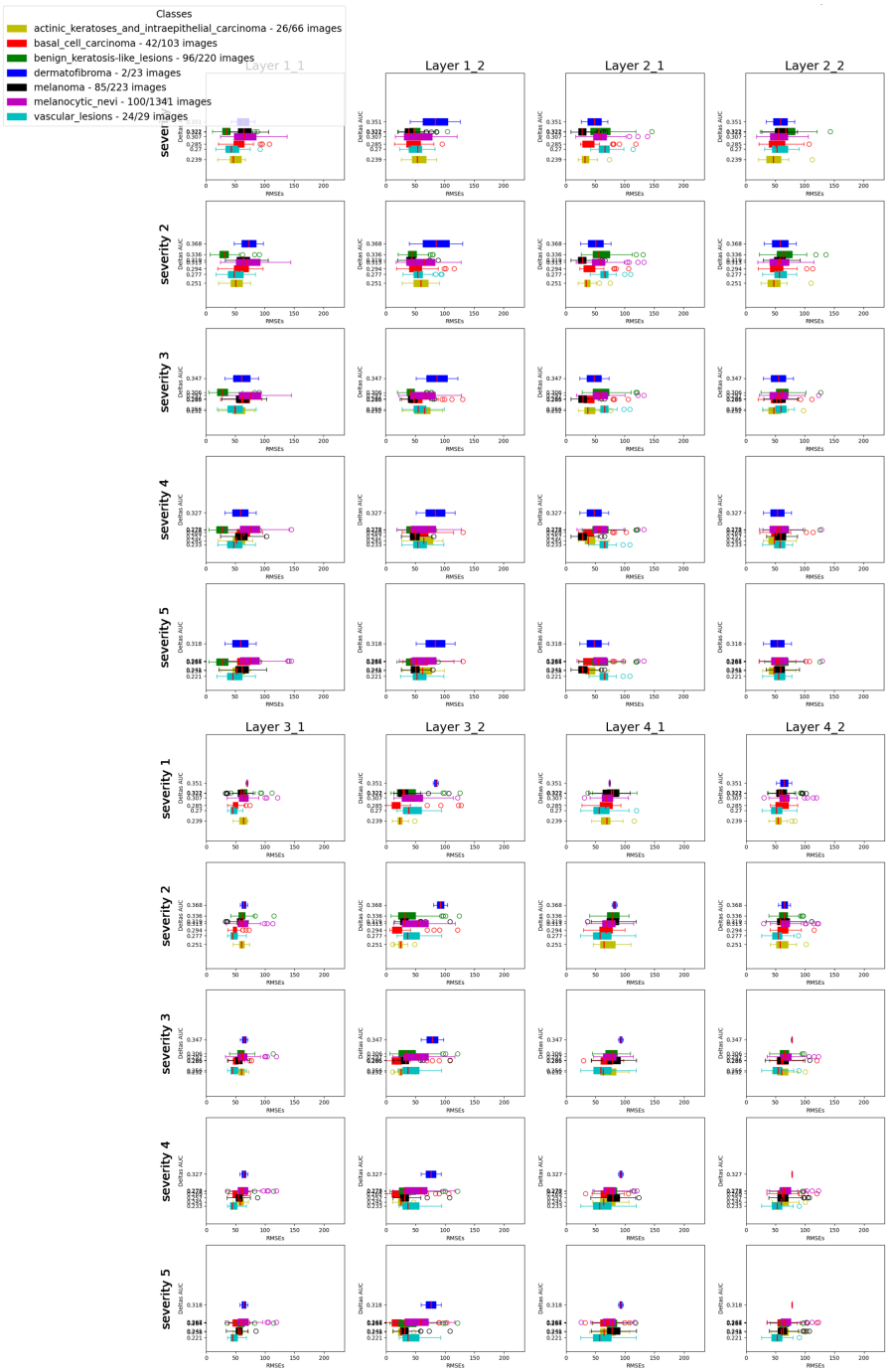

Figure 3.11: Correlation between delta AUC and delta Saliency for the DermaMNIST dataset under Speckle noise perturbation without augmentation.

This figure presents a series of box plots showing the relationship between Delta AUC and RMSE values across the Brightened perturbation, target layers, and perturbation severities. The analysis was conducted on the DermaMNIST dataset, specifically on correctly classified images. The axes, legend and layout are the same as in the previous figures.

Brightened observations:

What is striking in this graphic is that, regardless of the severity level, the RMSE values for each layer remain similar. This indicates that the severity level has no significant influence on the saliency maps. Additionally, the range of these values is quite spread out across most layers, except in the deeper layers, where the boxplot shows a more concentrated distribution of RMSE values. This suggests that the deeper layers exhibit more consistent behavior, with less variability in saliency map discrepancies, even as the severity of perturbations increases. What is also notable in this graphic is a small decrease in Delta AUC with increasing severities across all classes, suggesting an improvement in performance with higher levels of brightness. This trend indicates that the model might be better adapted to handling brightness variations, leading to more consistent and accurate predictions as the brightness level increases.

General observations:

Across all perturbations and severity levels, several consistent trends emerge. RMSE values generally remain stable across layers, with variations mainly occurring in specific layers such as 1.2, 2.2, and 3.2, where broader ranges or more significant shifts are observed. In deeper layers (such as 4.1 and 4.2), RMSE values tend to be more concentrated and stable, suggesting that the model's behavior becomes more consistent as the data progresses through the network, regardless of the type or severity of perturbation applied.

The correlation between Delta AUC and RMSE values varies depending on the perturbation. For Gaussian Noise and Speckle Noise, there is a noticeable trend where both Delta AUC and RMSE values increase with rising severity, indicating a linear relationship and suggesting that the model's performance degrades as perturbations become more severe. In contrast, for Motion Blurred perturbations, the model shows stable performance with minimal fluctuations in RMSE values, indicating that the model is less affected by severity in this case.

For Contrast and Brightened perturbations, the RMSE values show greater variability, particularly in certain layers, reflecting the model's increased sensitivity to these types of perturbations. Interestingly, a small decrease in Delta AUC with increasing severities is observed for Brightened perturbations, indicating a slight improvement in performance under these conditions, likely due to the model's adaptation to handling brightness variations.

Overall, these observations highlight that while the model demonstrates varying degrees of robustness depending on the perturbation and class, the correlation between Delta AUC and RMSE values is most evident in scenarios where the perturbation severity directly impacts model stability and performance.

Chapter 4

Discussion and Conclusions

4.1 Discussion

In this discussion, the results were interpreted in the context of the research objectives, specifically identifying and analyzing failure modes through gradient fingerprints in Saliency maps. The study examined how different MRI sequence-specific perturbations impacted the model's performance and the consistency of its saliency maps across various layers and perturbation severity levels. This was accomplished by first plotting the RMSEs between the saliency maps derived from the original and perturbed images for each class within a specific dataset as a function of the targeted layers. Following this, key findings regarding the correlation between Delta AUC and RMSE values were highlighted, accompanied by an exploration of the observed stability or variability within specific layers and perturbation types. These results were connected to the broader implications for quality control and error detection in AI-driven medical imaging, emphasizing the significance of understanding layer-wise model behavior in response to diverse perturbations.

4.1.1 Layer-wise discrepancies in Saliency maps

A common trend observed across all datasets was the increase in RMSE values as layers deepened, with significant peaks in the final layers. Additionally, RMSE values across different perturbations tended to converge in deeper layers, indicating a more uniform response as data progressed through the network. It highlighted a general discrepancy in saliency maps within the model's deeper layers. This could imply that as features become more abstract in the later layers, the saliency maps become consistently less reliable or less representative of the model's decision-making process. This trend suggests that errors in interpretability are more likely to occur in the deeper layers of the network, potentially compromising the clarity and reliability of the generated saliency maps. While previous research has often relied on the last convolutional layer for interpretability, favoring its alignment with human visual perception, these findings indicate that it may also be more susceptible to errors when dealing with perturbed images.

Additionally, contrast perturbations were particularly disruptive, causing greater fluctuations in RMSE values across multiple datasets. This points to a weakness in the model's ability to generate consistent and explainable saliency maps in the presence of contrast variations, as well as in maintaining similarity to the original saliency maps. The model's interpretability may therefore be more vulnerable to such perturbations, potentially posing a limitation, especially in real-world applications where image quality and technical setup can vary.

Despite these overarching trends, dataset-specific sensitivities were also identified. For example, the DermaMNIST dataset showed erratic RMSE behavior, especially in the dermatofibroma class, indicating a lack of consistency in model explainability for this particular class through the model. In contrast, the RetinaMNIST dataset exhibited flat and consistent RMSE values across the layers, suggesting that the model's explainability is more stable for this type of data. This variation indicates that the model's ability to provide consistent explanations is not uniform across different datasets, highlighting potential challenges in generalizing interpretability across diverse data types.

For the augmented BreastMNIST dataset, the delta AUC values across all perturbations remain close to zero, indicating that the model's predictive performance is largely unaffected by the perturbations. Despite this improved stability, the RMSE values for the augmented datasets show more erratic and pronounced fluctuations across different layers compared to the non-augmented datasets, especially regarding contrast variations. These inconsistencies suggest that more targeted interventions, such as adaptive augmentation techniques, layer-specific fine-tuning, or the incorporation of more advanced interpretability techniques may be necessary to fully stabilize the saliency maps. Moreover, the model's continued sensitivity to contrast perturbations implies that its interpretability could still be at risk in clinical applications.

4.1.2 Relationship between model performance and explainability

Across the datasets, a few subtle trends and patterns emerged regarding the model's response to perturbations. A consistent observation was the linear relationship between delta AUC and RMSE values. Specifically, as severity levels increased, particularly for contrast and brightened perturbations, delta AUC values tended to rise while RMSE values declined, indicating a deterioration in model performance. Other notable patterns were identified, such as the correlation between RMSE and delta AUC in the RetinaMNIST and DermaMNIST datasets for Gaussian noise, and a similar correlation for speckle noise in the DermaMNIST dataset.

Although the augmentation of the data successfully enhanced the model's performance, it did not produce any consistent trends across the datasets and make it challenging to effectively flag incorrect results based on saliency map discrepancies.

It was found in the first part of this study that saliency maps appeared more sensitive to perturbations in the last layer. However, these patterns seemed to be consistently present across all layers of the network. While previous research has emphasized the significance of the final layer for model interpretation, this focus might have missed important insights from earlier layers. The results of this study suggest that discrepancies in saliency maps, and the corresponding error patterns, may not be limited to the final layer but could also be occurring throughout the network. This possibility highlights the need for a more comprehensive layer-wise analysis to fully understand how perturbations affect model behavior across entire networks.

Despite the identification of some trends, these patterns are often highly specific to certain datasets and perturbation types. This specificity makes it challenging to generalize the findings across all cases. The complexity is further compounded by the large volume of data, numerous plots, and multiple axes, which make it difficult to discern clear trends with the human eye alone. Additionally, some RMSE data ranges were notably broad, particularly for the Contrast perturbation, leading to highly inconsistent saliency maps. This inconsistency complicates the interpretation of the results, as the variability within these broad RMSE ranges makes it harder to extract reliable conclusions. Consequently, the data's variability further limits the ability to draw definitive and consistent trends across all datasets, emphasizing the need for more refined analysis methods and balanced datasets to achieve more robust findings.

The few trends observed in this study align closely with the primary aim of investigating the distinct gradient fingerprints of saliency maps as indicators of model errors. The linear relationships between delta AUC and RMSE values, especially under contrast and brightened perturbations, provide evidence that these patterns can indeed serve as proxies for identifying failure modes within the model. However, the findings also indicate that these correlations are often specific to particular datasets and perturbation types, such as the heightened sensitivity to contrast and brightness variations. This specificity suggests that while it is possible to flag incorrect AI results using saliency map discrepancies, this approach may be most effective when tailored to specific types of data and perturbations. The study underscores the importance of focusing on these critical areas, particularly contrast and brightness, which play a significant role in influencing model performance and could be key to developing a more robust and targeted quality control mechanism.

4.1.3 Limitations

Despite the advancements presented, several limitations affect the current work. Specifically, 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.

Moreover, another significant limitation of this study is the imbalance in class data across several datasets, particularly in the DermaMNIST dataset. Some classes within this dataset contain only a few samples, making it challenging to draw reliable conclusions for those specific categories. This imbalance affects the robustness of the analysis and can lead to biased model interpretations, as the results may be disproportionately influenced by the more prevalent classes. The scarcity of data in underrepresented classes reduces the overall reliability of the findings and suggests that the model's performance on these classes might not be fully indicative of its general capability.

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.

4.2 Conclusions

In conclusion, this study demonstrates the potential of using saliency map discrepancies as indicators for identifying model errors. The consistent presence of error patterns across all network layers, rather than just the final layer, highlights the importance of a comprehensive layer-wise analysis in understanding and improving model behavior. Although the correlations between Delta AUC and RMSE values were limited, they do suggest a possible link, particularly under specific perturbations like contrast and brightness. However, these findings also underscore the challenges posed by the specificity of these trends to particular datasets and perturbation types, suggesting that a one-size-fits-all approach may not be effective. Additionally, the challenge of explaining complex correlations between features hinders the ability to provide robust explanations that could guide further model development and refinement. Moving forward, more work should be dedicated to developing a reliable mechanism for flagging incorrect AI results, ensuring that these insights can be effectively applied across diverse datasets and perturbation scenarios.

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, to address the limitations posed by the complexity of explaining correlations between features, future research could focus on developing techniques that can better handle nonlinear relationships and interactions among features. This would simplify the process of isolating and understanding the contributions of individual features and enhance the ability to detect and analyze these patterns effectively.

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 a more reliable, efficient, and clinically relevant approach to quality control.

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] N. Konz and M. A. Mazurowski. Pre-processing and compression: Understanding hidden representation refinement across imaging domains via intrinsic dimension. `http://arxiv.org/abs/2408.08381`, August 2024.

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

[14] 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.

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

[16] 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.

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

[18] 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.

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

[20] 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.

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

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

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

[24] 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.

[25] 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.

[26] 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.

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