UCLouvain

epl

École polytechnique de Louvain

Sensitivity of Explanations to Training Randomness in Vision Transformers

Author: **Pierre LEBOUTTE**
Supervisor: **François-Xavier STANDAERT**
Readers: **Jérémie BOGAERT, Benoît MACQ**
Academic year 2025–2026
Master [120] in Mathematical Engineering

Acknowledgments

First, I would like to thank François-Xavier Standaert and Jérémie Bogaert for their mentorship and guidance throughout the development of this master's thesis. Your consistent supervision, invaluable support, and our regular meetings has been paramount in the success of my work.

I would also like to thank my family for their continuous encouragement, especially my mom and her endless patience when listening to me explaining concepts totally new for her.

Last but certainly not least, I want to thank my girlfriend, Amélie, for her constant support and her help in reviewing law-related sections.

Abstract

Deep Neural Networks like Vision Transformers (ViT) achieve nowadays remarkable performances in image classification, but their opacity raises a fundamental question for their deployment in high-stakes domains: *can we trust the explanations they produce?* This master thesis studies the Layer-wise Relevance Propagation method applied to Vision Transformers, that aims to explain the model's decision by highlighting which areas of an image contributed the most to the prediction. By training a large number of models functionally equivalent differing only by the training seed, on two image classification tasks (facial emotion recognition and glaucoma diagnosis), this master thesis focuses on how these explanations are sensitive to the randomness inherent to the training phase, and shows how the explanations vary between models using the Mean Correlation With a Mean Explanation metric. The results reveal that models showing no significant difference in terms of accuracy can produce different explanations for an identical image. This phenomenon varies with the training regime: models fully fine-tuned show a higher instability of explanations but a better focus on discriminating areas, showing a trade-off between stability and quality of explanations. These results raise significant concerns about the use of artificial intelligence in critical contexts.

Table of Notations

Abbreviations

AUC	Area Under the Curve
CNN	Convolutional Neural Network
FEM	Functionally Equivalent Models
FER	Facial Expression Recognition
FT	Fine-Tuning / Fine-Tuned
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
MCWME	Mean Correlation With a Mean Explanation
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
SHAP	SHapley Additive exPlanations
ViT	Vision Transformer
xAI	Explainable Artificial Intelligence

Contents

1 Introduction **1**

2 Background **5**

2.1 Vision Transformers (ViT) 5

2.1.1 Architecture 5

2.1.2 ViT Training 7

2.2 Datasets 9

2.2.1 FER+ Dataset 9

2.2.2 PAPILA Dataset 10

2.3 Explainability in Deep Learning 11

2.3.1 Overview of Existing Methods 11

2.3.2 Adaptation of LRP to Vision Transformers 14

2.4 Sensitivity and Robustness of Explanations 16

2.4.1 Rashomon Effect & Equivalent Models 16

2.4.2 Sensitivity to Randomness 17

2.5 Metrics and Relevance Maps 17

2.5.1 Relevance Maps 17

2.5.2 MCWME 18

2.5.3 Top-k MCWME 19

2.5.4 Variance of the relevance 19

3 Methodology **21**

3.1 Overview 21

3.2 Main Pipeline 23

3.2.1 Dataset Preprocessing 23

3.2.2 Models Generation 25

3.2.3 Equivalent Models 28

3.2.4 Input compatibility 30

3.3 Research Questions 31

3.3.1 Core Analysis 31

3.3.2 Exploratory Analysis 33

4 Results **35**
4.1 Fine-tuning Regime . . . 35
4.1.1 Quantitative Analysis . . . 36
4.1.2 Qualitative Analysis . . . 37
4.1.3 Summary . . . 38
4.2 Contrastive Analysis . . . 38
4.2.1 The `happiness` class . . . 39
4.2.2 The `sadness` class . . . 41
4.2.3 The `neutral` class . . . 42
4.2.4 Summary . . . 43
4.3 Patch-level Analysis . . . 43
4.3.1 Top-k MCWME Analysis . . . 44
4.3.2 Variance Analysis . . . 46
4.3.3 Summary . . . 49

5 Exploratory Analysis **51**
5.1 Corpus Subjectivity . . . 51
5.1.1 FER+ . . . 51
5.1.2 PAPILA . . . 52
5.2 Task Difficulty . . . 54

6 Conclusion **57**

Chapter 1

Introduction

Deep Neural Networks like Vision Transformers (ViT) have revolutionized image classification, reaching impressive performance on various tasks, from Natural Language Processing (NLP) to computer vision. However, despite this spectacular soaring, the black box nature of these models raises an important challenge in order to understand their internal reasoning. Murdoch et al. (2019) define interpretable machine learning as the extraction of relevant knowledge concerning a model; this knowledge can be located in the data it uses, or in what the model learned. If this information is not extracted, the model remains essentially a black box. In high-stakes domains like health, law, and finance, the need for transparency and trustworthy explanations is therefore becoming paramount, and the current opacity of machine learning models raises a major ethical and scientific issue, described by Ribeiro, Singh, and Guestrin (2016):

How do we trust a machine learning model when we cannot understand its reasoning?

To address this question, Doshi-Velez and Kim (2017) insist on the need for a scientific framework in interpretability for machine learning models. This necessity for transparency gave birth to a new field called *Explainable AI* (xAI), seeking to make model decisions understandable for humans. A first major distinction is between *intrinsic* (or model-based) interpretability, concerning models from which the decision process is inherently readable, like decision trees, and *post-hoc* methods that seek to extract the explanation on an already trained model (Murdoch et al. 2019). Among existing (post-hoc) methods, *Layer-wise Relevance Propagation* (LRP) allows to re-dispatch the final decision of the model to each part of the input image, allowing the visualization of the areas that contributed the most to the prediction through relevance maps. This approach was adapted to Vision Transformers and gave promising results to explain models decisions on image classification tasks (Chefer, Gur, and Wolf 2021).

Jacovi and Goldberg (2020) evaluate explanations through the lens of two concepts: *faithfulness*, i.e. to what extent the explanation reflects the true internal reasoning of the model, and *plausibility*, i.e. at which point the explanation is convincing for a human user. Often, obtaining high levels of both is difficult; and this goes with several consequences. Indeed, Yeh et al. (2019) and Bansal, Agarwal, and Nguyen (2020) show that explanations can be highly sensitive to small changes in hyperparameters of the explanation method itself, which raises important concerns about how reliable these tools are.

Yet, a major question linked to this topic, but largely unexplored in the current literature, is whether these explanations remain stable despite the randomness inherent to model training. Indeed, recent work on language models, conducted by Bogaert (2026) and Loncour (2025), showed that two models trained exactly in the same way, except for the random seed used during training, can produce very different explanations for the exact same input, and exact same correct prediction. This phenomenon, called *sensitivity to training randomness*, raises concerns about the trustworthiness of explanation methods. If a model can justify its decision in multiple ways depending on the training randomness, how can we know among all the explanations which one(s) we should show to the user?

While this phenomenon of sensitivity of explanations to training randomness has been studied in the Natural Language Processing (NLP) context, no equivalent exists for now for Vision Transformers and image classification tasks, to the best of my knowledge; this master thesis proposes therefore to focus on computer vision and to study the stability of explanations produced by the LRP method on Vision Transformers. This work is based on two distinct domains: facial emotion recognition with the FER+ dataset and glaucoma diagnosis with the PAPILA dataset. These two tasks show different characteristics: one is rather subjective (emotions detection), and the other is more objective (medical diagnosis), allowing me to study the stability of explanations through diverse angles.

This work is articulated around several major axes seeking to dissect factors influencing this stability through point-by-point comparisons, mostly quantified by the Mean Correlation With a Mean Explanation metric (MCWME), quantifying stability of explanations across models. At the end, the main contribution of this master thesis is to show how training regime, contrastive class choice, and spatial variance of the relevance affect the stability of LRP explanations on Vision Transformers.

To achieve this, the remainder of the document is divided as follows:

- **Chapter 2** describes in detail the theoretical Background on Vision Transformers, the datasets used in this master thesis, as well as Explainability methods, quantitative metrics and qualitative representations.

- **Chapter 3** details in its first part the common methodology (the stem) used for all experiments explored in this master thesis. In the second part, I detail the leaves, i.e. each one of the five research questions I asked myself during this semester and how I planned to answer them.
- **Chapter 4** describes both qualitative and quantitative results of the experiments related to three of these research questions, and confronts them to my initial assumptions.
- **Chapter 5** presents exploratory results centered on the two last research questions detailed in Chapter 3.
- **Chapter 6** discusses limitations of this master thesis, summarizes the conclusions and details their implications for the use of xAI in real-world applications.

The results presented in this master thesis reveal that, in computer vision tasks as well, the explanation behind a correct prediction is not unique. Beyond the experimental scope of this thesis, the stability of explanations is a complex challenge with multiple ways to tackle it. At the end, finding answers will require contributions from law, philosophy, medicine, among other domains. This master thesis constitutes one step towards a better understanding of this issue.

Chapter 2

Background

2.1 Vision Transformers (ViT)

2.1.1 Architecture

The Attention Mechanism

The Transformer architecture (Vaswani et al. 2017) was originally introduced for Natural Language Processing (NLP) tasks. This architecture quickly became a standard first in NLP (Devlin et al. 2019), and the Vision Transformer (ViT) architecture then outperformed Convolutional Neural Networks (CNN's) on image classification tasks (Dosovitskiy et al. 2021).

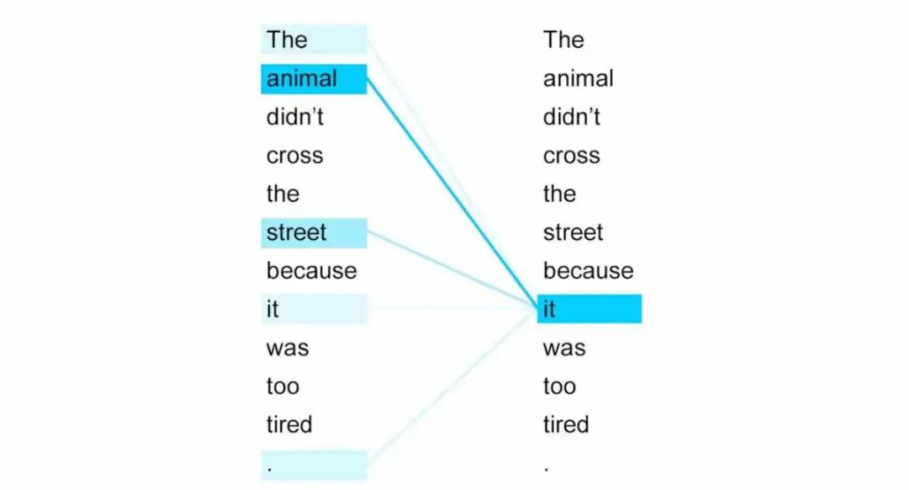


Figure 2.1: Attention mechanism. The high attention from "it" to "animal" allows the model to use directly this link in its embedding.[1]

This architecture uses the attention mechanism, allowing it to model global dependencies. In NLP, attention allows the model to give weights to each input token, as depicted in Figure 2.1. Self-attention is used to link words with the ones

[1]Image from `https://spotintelligence.com/2024/02/29/bert-nlp/`.

that are related to it in a sentence.

Formally, the self-attention mechanism projects an input sequence into three different matrices Queries (Q), Keys (K), and Values (V). For a certain input token, the information it is looking for in a sequence is the Query, while the Key contains the information it holds. Then, the dot product between Queries and all Keys is computed, to obtain scores which are then divided by the square root of the key vector dimension ($\sqrt{d_k}$) because, for large values of d_k, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients. The softmax function is then applied to these scores in order to obtain a probability distribution, called the attention weights, finally used to compute a weighted sum of the Value vectors.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

This process is often done in parallel through a mechanism called Multi-Head Attention, allowing the model in order to capture multiple types of dependencies in the input.

The Vision Transformer

Instead of CNN's that use local convolutions on the pixels' neighborhood, the ViT treats the entire image as a sequence. The image is first divided into fixed-size patches (e.g.14x14 patches), before being embedded linearly into a vector and adding positional embedding, and this sequence of vectors is then sent to a standard Transformer Encoder. A special vector called the `[CLS]` token is added to this sequence of embeddings. This token gathers the global representation of the image and at the output of the encoder, it is the only vector given to the classification head.

The Encoder itself has L=12 blocks. In each block, as described in Figure 2.2, several operations are applied successively: layer normalization, multi-head self-attention, residual connection, layer normalization, and a two-layer MLP with GELU (Gaussian Error Linear Unit) activation function, followed by a second residual connection.

The Vision Transformers get rid of CNN's inductive biases [2], giving the model the ability to model interactions between very distant zones of the image, from the first layer, with the cost of an increased data requirement.

This transition from the text to the images allowed the emergence of multimodal systems (combining text and images). Some other models like DeiT (Touvron et al.

[2]The main inductive bias of CNN's is that they assume that pixels close from each other are linked, and consider therefore small neighborhood when training.

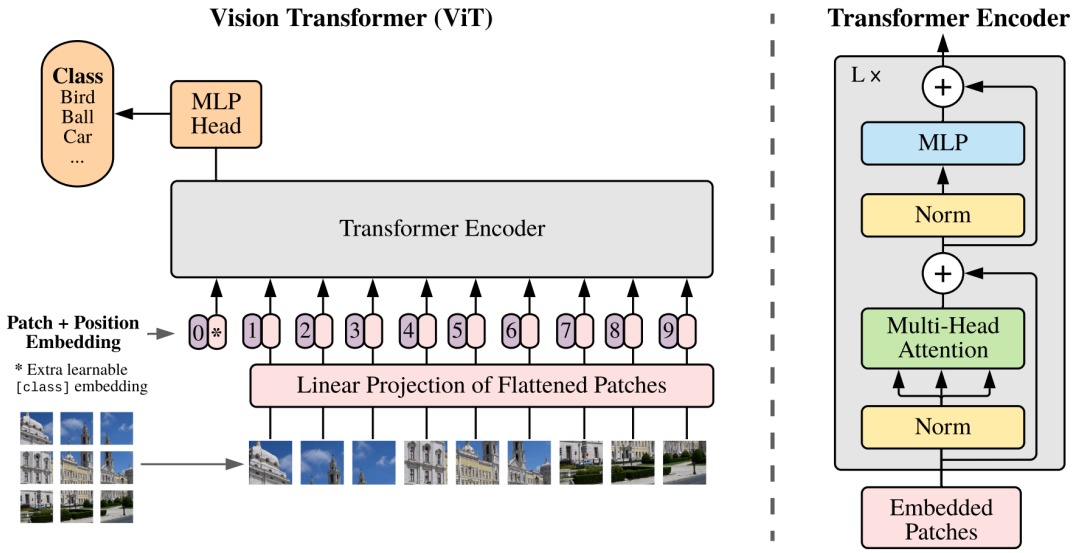


Figure 2.2: ViT Architecture as described by Dosovitskiy et al. (2021)

2021) and Swin (Z. Liu et al. 2021) emerged later, aiming to reduce the flaws of the original ViT.

2.1.2 ViT Training

Unlike the CNN, that has strong inductive bias on image structure, the ViT has to learn by itself how to interpret the image, its geometry, its shapes, using a large amount of data. The training is divided into two main phases:

Pre-training

In order for the ViT to have competitive performances compared to CNN's (Dosovitskiy et al. 2021), it is almost always pre-trained on massive databases, like, in my case, ImageNet-21k (14 millions of images, 21k classes). Indeed, training a ViT without this pre-training phase would most of the time not generalize well to smaller datasets. During this phase, the weights obtained describe very generic structures. In this thesis, these weights are shared and fixed for all trained models, ensuring that any future divergence between models is not due to this pre-training phase.

Fine-tuning

Once the model has a global visual understanding, I use fine-tuning to adapt it to specific tasks, often coming along with limited-size datasets. Practically speaking, the pre-trained classification head is discarded and replaced by a randomly initialized linear layer $\mathbf{W}_{\text{head}} \in \mathbb{R}^{D \times K}$ adapted for a specific downstream task (D is the embedding dimension and K is the number of downstream classes; $K = 2$ for all binary tasks in this thesis). Two main fine-tuning methods are proposed in this thesis, inspired by Devlin et al. (2019) and used by Bogaert (2026) in its work on

Natural Language Processing.

The first one, called *Full fine-tuned method* [3], allows all the parameters of both the encoder and the classification head to be modified. This regime is expected to produce a much larger diversity in the models.

In contrast, the *Frozen* regime keeps the weight of the Encoder fixed, allowing only the weights of the classification head $\mathbf{W}_{\text{head}}$ to be modified, and reducing therefore the amount of degrees of freedom during the training.

While the former is often preferred for NLP (Bogaert 2026) or computer vision (Dosovitskiy et al. 2021) and may reach better performances, both are used in the literature and the choice of one training regime or the other is particularly interesting with the angle of stability of explanations against training randomness.

Randomness in model training

Let us now introduce a fundamental distinction between model inference and model training. While the inference is strictly deterministic (meaning that, during evaluation, a given trained model will always produce the same output when given the same input), the model training is, however, stochastic. Indeed, even when using the exact same architecture and images from the same training set, training a model would multiple times never lead to the exact same weights.

This non-deterministic property of the training is due to the presence of several sources of randomness in the training : the initialization of $\mathbf{W}_{\text{head}}$ (at random values), the shuffling of images batches[4] (which will change the sequence of the gradient updates), random data augmentation (introducing some variability in the input[5]) and the drop-out[6] of specific neurons (to avoid overfitting, as described in Srivastava et al. (2014). All these factors ensure that each training run follows its unique trajectory, even when all hyperparameters are kept constant.

In order to study the impact of these factors, I control all this randomness by fixing a random seed produced by a Pseudo Random Number Generator. The study of how this seed will influence the diversity of explanations is the center of this thesis.

[3] In this document I use *Full FT* as an abbreviation for *Full fine-tuned*.
[4] A batch is the subset of images considered simultaneously during a forward-backward pass.
[5] This variability is described in Chapter 3.
[6] The default value of 10% was not changed.

2.2 Datasets

Two datasets for image classification were used in this thesis, with the goal to produce complementary results, the first one being **FER+**, a facial expression analysis dataset, and the second one, **PAPILA**, a medical imaging dataset for ophthalmology. This pair allows me to analyze whether the explanation stability varies depending on the dataset.

2.2.1 FER+ Dataset

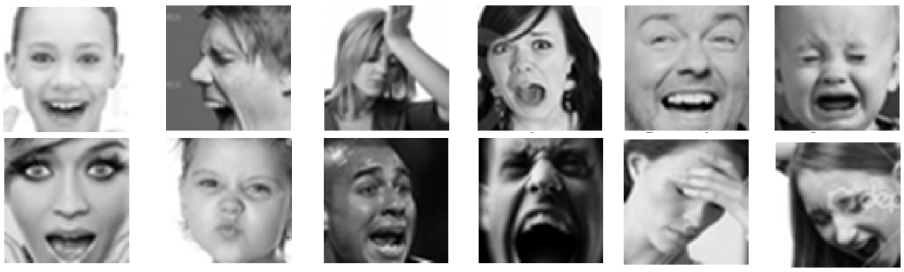


Figure 2.3: Representative examples from the FER+ dataset

The FER+ dataset (Barsoum et al. 2016) extends the original FER2013 dataset (Goodfellow et al. 2013). FER+ is composed of 35,887 images of faces in gray scale, with a 64×64 resolution. Examples of images from this dataset are depicted at Figure 2.3. This dataset features ten different labels. Eight are classical emotions: `neutral`, `happiness`, `sadness`, `anger`, `surprise`, `disgust`, `fear`, and `contempt` and the last two labels are `unknown` (corresponding to an image of a face, but the annotator is not able to identify the emotion), and `NF` (short for "Not Face", i.e. image clearly not depicting a face). This dataset shows a clear unbalance, mostly in favor of two classes: `neutral` and `happiness` : 62, 67 % of the images from the FER+ dataset belong to one of these two classes.

Emotion	**Number of images**
Neutral	13 013
Happiness	9 367
Surprise	4 493
Sadness	4 415
Anger	3 124
Fear	825
Disgust	253
Contempt	221
Total	**35 711**

Table 2.1: Repartition of the classes in the FER+ dataset (after filtering `NF` and `unknown` entries). For each image, the label corresponds to the majority vote among the annotators

The main contribution of FER+ compared to its predecessor, on top of the relabeling of some images, is how this dataset is annotated: each image was annotated by a panel of 10 labelers, yielding no single label but rather a probability distribution on the emotions classes for each image. This characteristic is used during the training when computing the loss function, and better reflects the ambiguity in labeling emotions as well as it reduces the errors due to a labeling made by a single person.

2.2.2 PAPILA Dataset

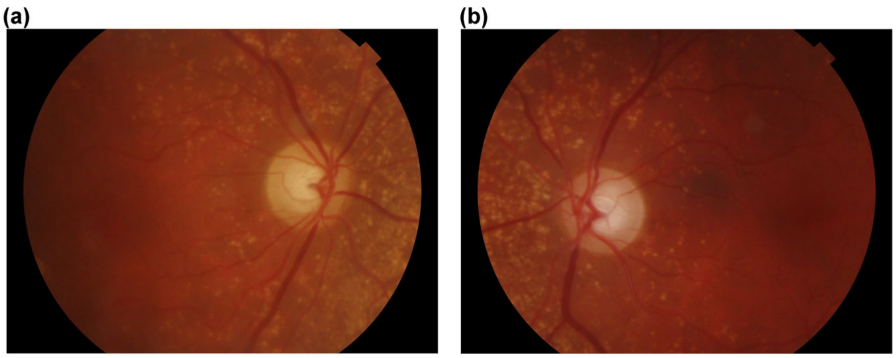


Figure 2.4: Retina images from the PAPILA dataset, showing both eyes from the same patient. a) RET047OD.jpg and b) RET047OS.jpg

The PAPILA dataset (Kovalyk et al. 2022) is a public dataset for glaucoma diagnosis, collected from the Hospital General Universitario Reina Sofía (Spain). It is composed of 488 color fundus images of size 2576×1934 pixels, annotated after the analysis of clinical data into three medical classes: `Healthy`, `Suspect` and `Glaucoma`. Two examples of images from this dataset are depicted at Figure 2.4. As shown in Table 2.2, the dataset is small and unbalanced (68.24 % of the images belong to the `Healthy` class), explaining why drawing general conclusions about results coming from this dataset requires carefulness. This dataset is therefore rather used as a complementary dataset for my analyses.

Class	**Number of images**
Healthy	333
Suspect	87
Glaucoma	68
Total	**488**

Table 2.2: Class Repartition for the PAPILA dataset

The `Suspect` class is gathering cases where the comprehensive evaluation of the clinical data of the subject does not allow the authors to assign the `Glaucoma` or

`Healthy` label (Kovalyk et al. 2022). This class does not gather images of retina at an intermediate level of the disease but represent rather a gray area; as described by Kovalyk et al. (2022) through t-SNE and PCA plots, the `Suspect` class is mixed with both `Glaucoma` and `Healthy` classes. A characteristic of PAPILA is the presence of manual segmentations of the optic disc and optic cup of the retina [7], given by two independent experts[8]. These optic disc and optic cup segmentations are used by ophthalmologists to identify glaucoma in patients retina (Almazroa et al. 2015).

2.3 Explainability in Deep Learning

2.3.1 Overview of Existing Methods

Motivation

With the ever-growing development of deep neural networks in high-stakes domains, such as medical diagnosis or elections, the opacity of these models has raised major concerns. A model able to reach high level of performance, but unable to produce meaningful explanations for its decision, raises serious ethical and scientific concerns. This need for models able to produce significant explanations understandable by humans gave birth to a brand new field: Explainable AI (xAI).

Faithfulness vs Plausibility

Two criteria are in general cited to evaluate the quality of an explanation: **faithfulness** and **plausibility**.

Faithfulness (Lyu, Apidianaki, and Callison-Burch 2022) is defined as the degree to which an explanation reflect with precision the true internal process of the model leading to its prediction. It is a critical criterion, as an explanation that does not reflect the internal functioning of the model is misleading, and therefore unusable. It is indeed critical that the explanation reveals the elements truly used by the models to take its decision, even if this internal logic seems counter-intuitive or far from the human reasoning.

Plausibility (Jacovi and Goldberg 2020) is the capacity for an explanation to be understandable and convincing for a human user. Unlike Faithfulness, Plausibility has a high subjective dimension: the convincing nature of an explanation would inevitably depend on the knowledge and expertise domain of those evaluating it. The first challenge is to show an explanation in a form directly interpretable by humans, like a heatmap showing areas considered important by the model, in order for them

[7] see Appendix: Additional Figures.

[8] the true label for each image has well a unique assignation, but the visual segmentations are given by two experts.

to validate or reject the reasoning of the model.

However, simultaneously optimizing these two criteria would inevitably lead to a trade-off. A completely faithful explanation would require to give all parameters and mathematical reasoning of the neural network, that would not be understandable for a human and therefore not plausible for him. Conversely, a method could show what seems to be a plausible explanation, illustrating what a human would expect from the model, but this explanation might not be reflecting the true computation leading to the prediction. The goal of xAI is to navigate between these two concepts, using methods able to give representations interpretable visually, while making sure that they stick to the internal reasoning of the model.

Taxonomy

Explainability methods (also called interpretability methods) can be divided into two main categories: **intrinsic** and **post-hoc** interpretability (Du, N. Liu, and Hu 2019). The intrinsic-based models are designed in order for their decision process to be immediately readable from their structure. This family includes decisions tree, rule-based models, linear models, among other types of models. In opposition, post-hoc methods are used on an already trained model to dissect its reasoning and behavior. In my context focusing on Vision Transformers, as this architecture is not intrinsically explainable, its decision process is not directly readable from its design. Explainability methods used in this thesis will then fall to the post-hoc methods category.

Another distinction point is whether the method is **global** or **local** (Du, N. Liu, and Hu 2019). Whereas global methods aim to describe the overall behavior of the system (e.g., using the parameters and the structure of a model to explain its reasoning), local methods locally examine an individual prediction of a model, trying to figure out why the model makes the decision it makes.

Within the **local post-hoc** explainability methods, the literature has explored four methods of interest:

1. Gradient-based methods (Simonyan, Vedaldi, and Zisserman 2014) (Sundararajan, Taly, and Yan 2017)
2. Attention-based methods (Abnar and Zuidema 2020)
3. Perturbation-based methods (Ribeiro, Singh, and Guestrin 2016)
4. Propagation-based methods (Bach et al. 2015)

Gradient-based methods

Gradient-based methods describe the relationship between a model's input and output, through the computation of its partial derivatives. Simonyan, Vedaldi, and Zisserman (2014) proposed the **Saliency map**, defining the importance of an input k for a class c as the sensitivity of the output to small perturbation of each input dimension. Saliency maps are cheap to compute, as a simple backward pass is sufficient, but are known to be very noisy (Smilkov et al. 2017).

Some other methods have been proposed, as **SmoothGrad** (Smilkov et al. 2017), and **Integrated Gradients** (Sundararajan, Taly, and Yan 2017) that integrates gradients along a straight path from a baseline x' to the input x. Finally, Grad-CAM (Selvaraju et al. 2017) is commonly used to highlight which areas influenced the decision of the model, by combining the activation and the gradients of the layers.

Attention-based Methods

With the emergence of attention-based models, it is natural to express the explanation through the computation of attention weights α_{ij} at head h and layer l, describing how much token i attends to token j. It is natural to think that high attention tokens are the most important in the explanations to lead to the prediction.

Abnar and Zuidema (2020) described **Attention Rollout**, a method computing all the attention flow from the output layer. However, this method makes no difference between the attention given to a token, and this token being relevant for a prediction. Indeed, whether the raw attention weights can truly and only be interpreted as faithful explanations is still debated (Wiegreffe and Pinter 2019); this is the reason why these methods are not the main focus of this thesis.

Perturbation-based Methods

These methods compute how specific areas of the image are important by perturbing these areas (for example, by masking specific pixels) and calculating how much the output changed because of these perturbations.

LIME (Local Interpretable Model-agnostic Explanations), described by (Ribeiro, Singh, and Guestrin 2016) fits a locally interpretable model by perturbing the input and observing the outputs. **SHAP** (SHapley Additive exPlanations) uses the Shapley Value (used in Game Theory), exposed by (Lundberg and Lee 2017) for a model f and an input $\mathbf{x} \in \mathbb{R}^d$, for a feature k:

$$\phi_k(f, \mathbf{x}) = \sum_{S \subseteq [d] \setminus \{k\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \Big[f_{S \cup \{k\}}(\mathbf{x}) - f_S(\mathbf{x}) \Big] \tag{2.2}$$

where $f_S(\mathbf{x})$ denotes the expected model output when only the features in subset S

are known. While Shapley values are the unique attribution method satisfying four desirable properties (Molnar 2022), both methods are unusable for my thesis because of their **randomness**: as they use random perturbations of the input, repeated measure on the same image and the same model would lead to different results. I would thus need a deterministic method for my experimental pipeline.

Propagation-based Methods

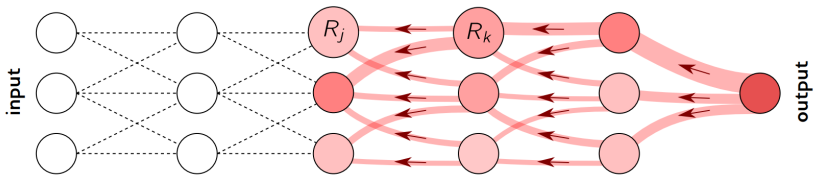


Figure 2.5: Illustration of the Layer-wise Relevance Propagation method.[9]

In the propagation-based methods family, **Layer-wise relevance Propagation** (LRP), is the most common method. Introduced by Bach et al. (2015) for deep learning in classification, it is based on backpropagation (c.f. Figure 2.5) of the output signal. The center of this method is the **conservation principle**, i.e. a principle stating that the total relevance of an output of the network (its prediction score), is conserved when redistributing it backwards through the layers of the network, until the individual pixel. This is a desirable property for an explainability method, because it ensures that the whole decision of the model is explained, without any part of the signal being created or lost during the redistributing process.

The conservation rule is, for all layer l:

$$\sum_j R_j^{(l)} = \sum_k R_k^{(l-1)} = f_c(\mathbf{x}) \tag{2.3}$$

where $R_j^{(l)}$ is the relevance score assigned to each layer l and each neuron j. $f_c(\mathbf{x}) \in \mathbb{R}$ is the output score for class c on input $\mathbf{x}$, and therefore, the output layer score is of course $R_c^{(L)} = f_c(\mathbf{x})$ for the target class, and $R_j^{(L)} = 0$ for all other output neurons. This makes sure that the total relevance is equal to the model's output score.

2.3.2 Adaptation of LRP to Vision Transformers

As Layer-wise Relevance Propagation was initially created for networks with linear or pointwise non-linear layers, adaptations are required for the Vision Transformer architecture because of the non-linear property of the self-attention mechanism (e.g. with the scalar product between Queries and Keys).

[9] Image from Montavon et al. (2019).

To address this issue, Chefer, Gur, and Wolf (2021) show that operations in Transformers (like the matrix multiplications of the attention mechanism) do not respect the conservation principle. To solve this, for a specific Transformer block b, the weighted attention relevance $\bar{\mathbf{A}}^{(b)}$ is computed by combining the attention map $\mathbf{A}^{(b)}$, its gradients $\nabla\mathbf{A}^{(b)}$, and the layer's relevance $R^{(n_b)}$:

$$\bar{\mathbf{A}}^{(b)} = \mathbf{I} + \mathbb{E}_h\left(\left(\nabla\mathbf{A}^{(b)} \odot R^{(n_b)}\right)^+\right) \tag{2.4}$$

where n_b is the layer that corresponds to the softmax operation[10] of block b, and $R^(n_b)$ is the layer's relevance. $\odot$ is the Hadamard product, $\mathbb{E}_h$ is the mean across all attention heads, and $(\cdot)^+ = \max(0, \cdot)$ keeps only the positive part. The identity matrix $\mathbf{I}$ represents the skip connections.

The final explanation matrix $\mathbf{C}$ is obtained by performing the matrix multiplication of these weighted relevances across all B blocks, in order, from block 1 to block B:

$$\mathbf{C} = \bar{\mathbf{A}}^{(1)} \cdot \bar{\mathbf{A}}^{(2)} \cdot \ldots \cdot \bar{\mathbf{A}}^{(B)} \tag{2.5}$$

For the visualization, I extract the row of $\mathbf{C}$ corresponding to the [CLS] token. The content tokens are then reshaped into a 14×14 grid of patches (for a ViT-B/16 architecture) and interpolated to the original image dimensions using bilinear interpolation to produce the final map. The process used by Chefer, Gur, and Wolf (2021) is described at Figure 2.6.

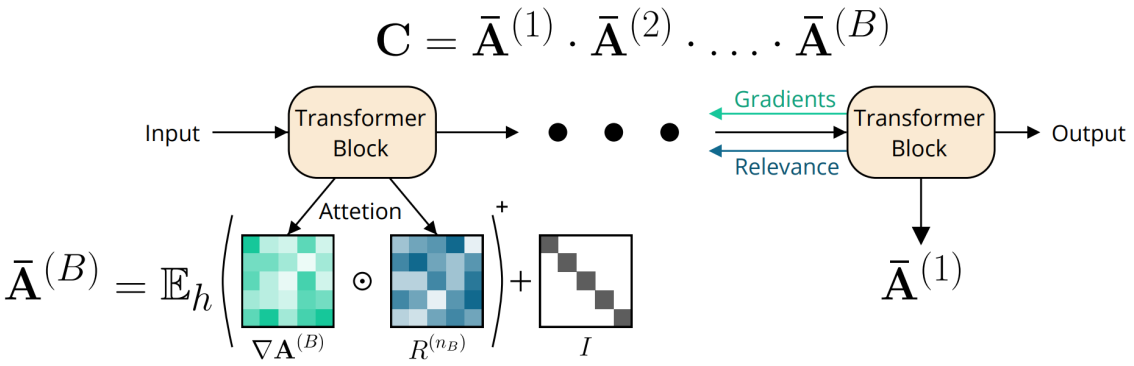


Figure 2.6: LRP for ViT architectures as described in Chefer, Gur, and Wolf (2021). The innovation from the paper is the normalization formula on the bottom left, making sure that the conservation principle is kept across all layers.

Suitability for this thesis

The use of this method is motivated by several fundamental properties. Firstly, LRP is entirely deterministic, guaranteeing that each variation observed between

[10] see Equation (2.1).

two explanation only comes from a difference between models training [11], and is not due to randomness in the explanation process itself, as it might be the case for other methods presented above. Moreover, this approach is mathematically consistent thanks to the conservation axioms. By integrating residual connections, the method of Chefer, Gur, and Wolf (2021) produces indeed explanations that appear to be more faithful than GradCAM.

2.4 Sensitivity and Robustness of Explanations

2.4.1 Rashomon Effect & Equivalent Models

Rashomon Effect

The concept of **Rashomon Effect** introduced by Breiman (2001) and named after Kurosawa's movie in which an event is described differently by several witnesses, states that there exists a large number of models able to solve the same task with equally good performance, but with different solution strategies (Müller et al. 2023). The set of these models is called the **Rashomon Set** (or statistically equivalent models set).

This phenomenon raises an important issue linked with explainable artificial intelligence: even though several models output the exact same prediction, and have similar performance (e.g. similar accuracy), there is absolutely no guarantee that they are reasoning identically.

Equivalent Models

To analyze the stability of explanations, let us define the notion of **functionally equivalent models**. In this thesis, two models are considered functionally equivalent when they are satisfying two criteria. Firstly, there should be no significant difference in their accuracy (this is validated by a statistical test described in § 3.2.3). Secondly, they have to produce an identical and correct prediction for the same input x. This restriction is important as it makes sure that every distinction observed in explanations is not due to different predictions, or mistakes in the prediction, but reflects truly the instability against the training randomness. The set obtained is the functionally equivalent models set (FEM set), and is a *Rashomon set* of models that output the same correct prediction for a specific input set. I refer to these inputs that are given the same correct prediction by all models in a FEM set as **compatible inputs**.

[11] This is guaranteed thanks to deterministic model inference (see § 2.1.2).

2.4.2 Sensitivity to Randomness

The main theme explored by Loncour (2025) and Bogaert (2026) in NLP tasks is the vulnerability of explanation methods against the randomness of the training. Both studies show that modifying the seed generate functionally equivalent models, yet able to produce different explanations for the same test instance. This phenomenon shows that the explanation given by a model is not an absolute representation of its reasoning but is rather governed by the random seed used in the optimization process. This sensitivity to randomness has deep ethical and practical implications, especially when developing deep learning models in high-stakes domains (Agarwal, Tanneru, and Lakkaraju 2024).

2.5 Metrics and Relevance Maps

2.5.1 Relevance Maps

Let us recall that the LRP method outputs relevance score $R_i(x) \in \mathbb{R}^{196}$, the dimension of R coming from the grid size of the ViT, composed of $14 \times 14 = 196$ patches. To obtain visual results, the relevance computed for each patch is displayed through relevance maps. These maps show visually the relevance score attributed to each patch. Patches colored in red were assigned a high relevance score, while patches in blue show relevance scores close to zero. After computed the relevance on each patch, the relevance map is upsampled back to the size of the original image using bilinear interpolation (Chefer, Gur, and Wolf 2021). We use a *mean relevance map* when displaying the mean relevance across models, and a *masked mean relevance map* to display the mean relevance of specific pixels. An example of a relevance map is shown in Figure 2.7.



(a) Original Image (b) Relevance map

Figure 2.7: Illustration of a LRP relevance map

2.5.2 MCWME

The quantitative comparison between two explanations requires a measure able to catch the spatial similarity between two visual explanations, i.e. two maps generated thanks to the LRP method. The first one that comes to mind is the spatial Pearson correlation coefficient (ρ_P), a standard measure in the literature to compare two distributions. The Pearson correlation measures the linear relationship between the scores given to pixels (or patches). Even if using a linear measure might seem relatively simple to analyze complex neural networks, searching for higher order relations would risk to produce more noisy and less interpretable values.

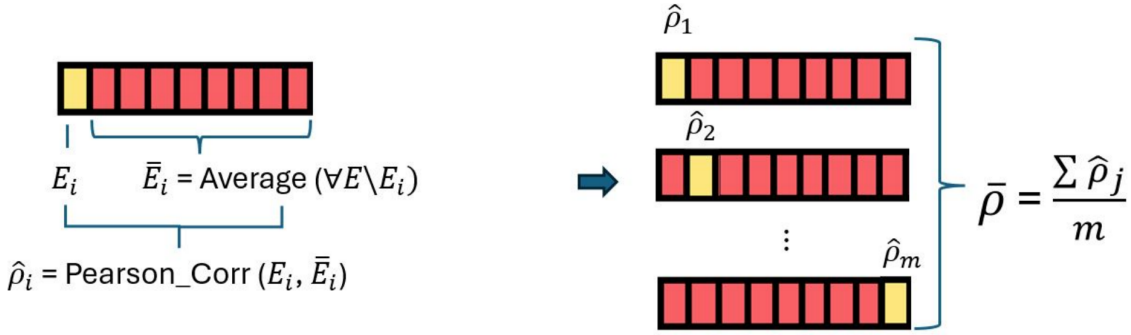


Figure 2.8: MCWME computation

To aggregate these model-to-model comparisons in the functionally models set (FEM set), this thesis uses the **Mean Correlation With a Mean Explanation (MCWME)** metric (J. Bogaert, Descampe, and Standaert 2026). The computation is done in two parts, as described in Figure 2.8.

Firstly, for each model in the set, let us define a mean Leave-one-out explanation $\bar{R}_{-i}(x)$, obtained by computing the patch-wise arithmetic mean of the relevance produced by all the $N-1$ other models of the set.

$$\bar{R}_{-i}(x) = \frac{1}{N-1} \sum_{j \neq i}^{N} R_j(x) \tag{2.6}$$

Secondly, I evaluate the stability of each model i by computing the Pearson correlation coefficient, denoted $\hat{\rho}_i$, between its individual explanation $R_i(x)$ and the corresponding leave-one-out mean $\bar{R}_{-i}(x)$:

$$\hat{\rho}_i = \rho_P \left(R_i(x), \bar{R}_{-i}(x) \right) \tag{2.7}$$

Since the distribution of these correlation coefficients is usually not normally distributed, the **Fisher Z-transform** is applied (Silver and Dunlap 1987) on each sample $\hat{\rho}_i$:

$$Z_i = \text{arctanh}(\hat{\rho}_i) \tag{2.8}$$

In this new space, I compute the mean $\bar{Z}$ and the empirical variance $\hat{\sigma}_Z^2$, before obtaining then the final value of the MCWME and its 95% confidence interval by applying the inverse transformation. The 95% confidence interval is:

$$\text{MCWME}(x) \in \left[\tanh\left(\bar{Z} - 1.96\frac{\hat{\sigma}_Z}{\sqrt{N}}\right); \tanh\left(\bar{Z} + 1.96\frac{\hat{\sigma}_Z}{\sqrt{N}}\right)\right] \tag{2.9}$$

The MCWME is bounded in the $[-1, 1]$ interval, even if in practice it is located between 0 and 1. A MCWME close to 1 means that each model agrees completely with the mean, acting as a proxy for a high consensus between models.

2.5.3 Top-k MCWME

To answer specific research questions, I also use a **Top-k explanation** measure. Instead of evaluating the relevance on the whole image, this method fixes to 0 the relevance scores, except the p patches having the most important relevance scores averaged on the N models of the functionally equivalent models set (FEM set), before computing the MCWME when considering only the patches selected. This masking filter has three advantages. Firstly, it is more interpretable, as it allows the MCWME to focus the most on critical areas. Secondly, it reduces the noise created by areas considered as less relevant. Finally, it allows to compare more equally the variability between images and tasks. To make the measure more interpretable, I rescale the value of the number of patches p (ranging from 1 to 196)[12] to the number of pixels k (ranging from 1 to 50176, as each patch contains 16×16 pixels).

By plotting the values of MCWME with its confidence interval for values of k (ranging from 1 to 50176), I obtain plots describing the evolution of MCWME with the number of pixels k, as the one given in Figure 2.9.

2.5.4 Variance of the relevance

The MCWME described above allows me to describe *how much* models disagree but not *where* this disagreement is located in the image. To answer this, I introduce the spatial variance of relevance scores.

Let us recall that for a compatible input x, the LRP method gives for each model M_i a relevance score $R_i(x) \in \mathbb{R}^{196}$. For each patch p, the variance of its relevance

[12] 196 is the number of total patches used in the ViT-B/16 architecture.

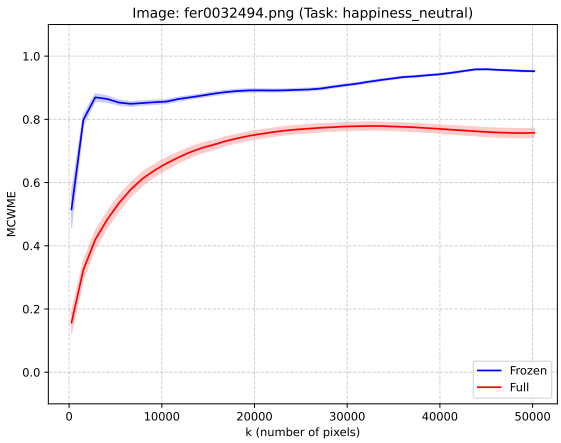


Figure 2.9: Example of a plot depicting the evolution of top-k MCWME with k

score across the N models of the FEM set is

$$\sigma^2_{(p)}(x) = \frac{1}{N}\sum_{i=1}^{N}\left(R_i^{(p)}(x) - \bar{R}^{(p)}(x)\right)^2 \tag{2.10}$$

where $\bar{R}^{(p)}(x)$ is the mean relevance score of patch p averaged for all the models. Repeating this computation for all the 196 patches, I obtain a **variance map** $\sigma^2(x) \in \mathbb{R}^{14\times14}$ which is then interpolated using bilinear interpolation to the original image's dimensions to produce a map, like the relevance map. For short, the variance map is similar to the relevance map, but the quantity displayed visually for each patch is the variance of the relevance across all models, not the relevance itself. Therefore, a high value for a patch shows that models disagree on this patch, and a low value shows consensus.

Chapter 3

Methodology

3.1 Overview

The goal of this Chapter is to study the experimental process I used to describe the Rashomon Effect applied to Vision Transformers (ViT). Concretely, I want to prove *to what extent models having similar performances can still radically differ in terms of explanations*. This is done by studying building bricks that I call *settings*.

> A *setting* is a combination of a binary task (on a specific dataset), and a fine-tuning regime.[1].

An example of a setting being:

`happiness` vs `sadness` (FER+ dataset) in the *Frozen* regime

In this master thesis, I will therefore use the word *task* to describe a classification task without mentioning the regime, like `happiness` vs `sadness`. When specifying the regime is needed, I am using the word *setting*.

To give an overview of the general pipeline used in this section, Figure 3.1 show the flowchart of my methodology, showing the 2 key parts of my core structure.

The first part is the *Main Pipeline*, in common for all experiments studied in this thesis. This section is composed of five steps. I begin with the pre-processing of the datasets (FER+ and Papila) (1). Then, I generate multiple models for each different setting (2). The third phase applies a statistical test to isolate models having similar accuracy (3). The fourth step filters the test set, by keeping images for which the prediction by the models are unanimous (4). I then proceed by generating the relevance vectors through the LRP method (5).

[1] Fine-tuning regimes are described in § 2.1.2.

The second key part answers the *Research Questions* that I describe at the end of this Chapter. This analytical part is divided into two parts: the Core Analysis of my work, based on a rigorous framework and the Exploratory Analysis part where I explore open questions that invite further investigations.

The rest of this Chapter will describe in details this pipeline. I will detail, for each of these steps, the implementation and the choices I made about it. This analysis begins naturally with the description of the main pipeline.

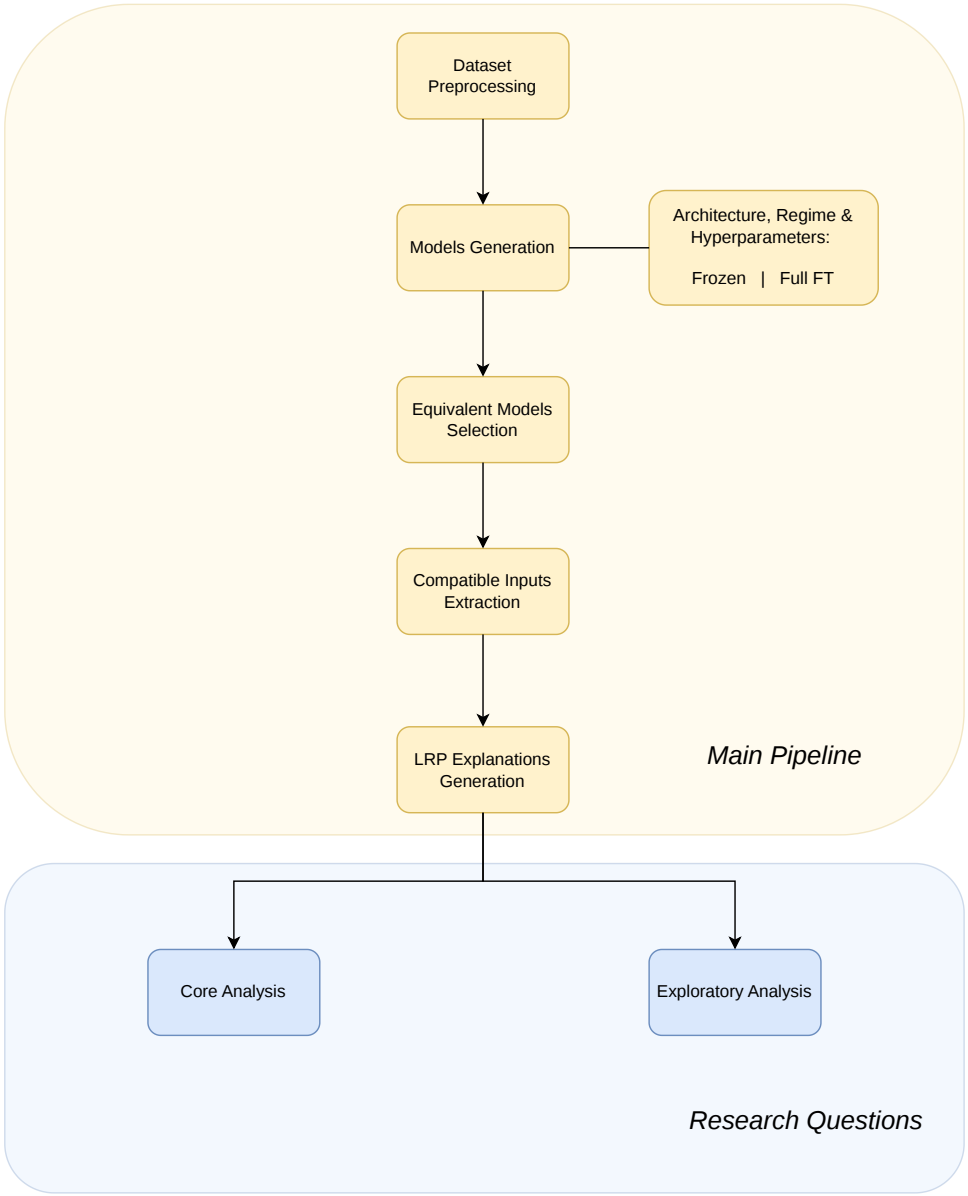


Figure 3.1: Overview of my experimental pipeline. This flowchart is described in details in Chapter 3.

3.2 Main Pipeline

This section details the main pipeline, applied for all experiments designed in this thesis. This main pipeline describes how I process data, generate my models, select functionally equivalent models and identify the compatible inputs; in short: how I build the suitable environment to address my research questions. To ensure that my methodology is valid, it is directly build on previous work from Loncour (2025) and Bogaert (2026).

3.2.1 Dataset Preprocessing

This pipeline naturally begins with the preprocessing of my data from the two datasets described in Chapter 2, to be able to use it to perform my analysis. The choice for these two datasets is motivated by the willingness to cover two applications with a different level of subjectivity. FER+ was chosen because it is a large-sized and well-studied dataset commonly used with Vision Transformers. The choice for PAPILA as a complementary dataset is motivated by the fact that this dataset is a three classes (like FER+) recent (2022) medical imaging dataset composed of colored high resolution images, annotated by experts.

FER+

For the preprocessing of this dataset, recall that this dataset contains ten labels: `neutral`, `happiness`, `sadness`, `anger`, `surprise`, `disgust`, `fear`, `contempt`, `unknown`, and `NF` (not face). The first operation consists in discarding images annotated as `NF` by a majority of annotators. These kinds of images are often either irrelevant images or NaN values. The split between training, validation and test sets is already done in FER+, and directly inherited from the FER2013 dataset (Barsoum et al. 2016). The next step is the extraction of a unique label for each image; this is done by selecting the majority vote.

Once every image is associated with a specific label, I proceed with the selection of relevant classes. For the purpose of this thesis, I use FER+ with 3 binary classification tasks, meaning that only three pairwise classification tasks are considered: `happiness` vs `sadness`, `happiness` vs `neutral` and `sadness` vs `neutral`. Therefore, I discard all images not belonging to one of these 3 classes: `neutral`, `happiness`, and `sadness`. These three classes were chosen because they are among the four most represented classes of the dataset and correspond to common and easily identifiable facial expressions. The number of images from each remaining class is listed in Table 3.1. It is noteworthy that two classes appears to be intuitively quite close to each other: `neutral` and `sadness`. This is an element of interest for the rest of this study as it allows me to explore the influence of this proximity on the stability of explanations.

Table 3.1: Class distribution in the training set of the FER+ dataset after filtering the three relevant classes.

Emotion	**Number of images**
Neutral	10 363
Happiness	7 532
Sadness	3 543
Total	**21 438**

The next step involves adjusting the structuring properties of the images, such as their size or color scale. The goal here is to make my images easily usable by a machine learning model, more specifically for the architecture used by Chefer, Gur, and Wolf (2021). I first resize the images to 224x224 pixels (standard size needed for the ViT-B/16 architecture I am using) with `transforms.Resize(224)`. I also convert these images in RGB to be compatible with the pre-trained weights of the model (we recall these weights are based on the training on the ImageNet-21k dataset, using a RGB scale). Finally, to stick to the methodology used by Chefer, Gur, and Wolf (2021), I normalize using a mean and standard deviation of 0.5 to convert the color values of the pixels (i.e, the RGB color levels) from the interval [0 ; 255] to an interval [-1; 1]. Indeed, normalizing the inputs is known to accelerate the convergence when using the gradient method during the optimization process (LeCun et al. 1998).

Finally, to avoid overfitting, I apply a random horizontal flip to the images of the dataset, directly inspired by Xue, Wang, and G. Guo (2021), who describes TransFER, a ViT model that reached state-of-the-art performance on FER+ when the paper was published. It is important to note that these random transformations are strictly used in the training, as no random transformations of this kind (except the resize) are applied to both test and validation set, to ensure deterministic inference.

PAPILA

The preprocessing of the PAPILA dataset follows a similar process to the one of FER+, beginning with the resizing to a 224× 224 resolution. Using all the images of this dataset, I extract three binary classification tasks for the sake of my study: `Healthy` vs `Suspect`, `Healthy` vs `Glaucoma`, and `Suspect` vs `Glaucoma`. Unlike the FER+ dataset, each image has its own unique label. The dataset was split into training and test set using a 80-20 split. The stratification ensures that the original distribution of classes is maintained in the split; this is particularly essential here because the dataset is unbalanced in favor of the `Healthy` class. Due to the small size of PAPILA, the validation set used during training was the same as the training set. While I acknowledge that this may be introducing some biases in the performances, this should not affect the comparison between models as they are all evaluated on

the same test set.

A similar image preparation is done with this dataset, by allowing transformations during the training: horizontal and vertical flips, slight rotation of 15°, light cropping of the image, and color jittering to randomly change the brightness, contrast, saturation, and hue of an image[2]. I also do the same RGB conversion and normalization as for the FER+ dataset. Once again, no transformations among those cited above, except the resizing, were used for both validation and test sets.

3.2.2 Models Generation

This section details the generation of models I will be using to study the *Rashomon Effect*. The goal of this section is to produce a population whose diversity only comes from training randomness.

Software and Hardware Infrastructures

The models were trained on the GPU of `CéCi` infrastructures (Consortium des Équipements de Calcul Intensif), more precisely the **lyra** cluster stored at Université libre de Bruxelles (ULB). All experiments were implemented in Python 3.11.3, using PyTorch 2.5.1 (compiled with CUDA 12.1). The cluster was running Rocky Linux 9.5. This HPC (High Performance Computing) cluster is specifically designed for machine learning tasks. The computations are done in parallel on virtual nodes equipped with AMD EPYC Genoa (32 cores, 128GB of RAM) and accelerated with graphics cards NVIDIA RTX 6000 Ada Generation with 48 GB of RAM. The use of this cluster was justified by its suitability for deep learning and particularly the computational cost of repeating the fine-tuning of many models.

Hyperparameters and Loss functions

All models used in this thesis use the ViT-B/16 architecture[3] described in Chapter 2. However, both datasets needed different hyperparameters choices due to their intrinsic fundamental difference, but I wanted to have a certain degree of similarity between the two experiences. Therefore, I initialized the parameters with values commonly used in the literature, before adjusting them empirically to obtain slightly more homogeneity between the two experiences. Overall, the global performance in terms of AUC or accuracy is not the main point of this study, as long as it aligns with the standards in the literature. The hyperparameters are detailed in Table 3.2.

Learning rates are inspired, for FER+, of those from Xue, Wang, and G. Guo (2021). As the dataset is quite recent, PAPILA lacks cited papers using Vision

[2]The documentation is available at `https://docs.pytorch.org/vision/main/generated/torchvision.transforms.ColorJitter.html`.

[3]This architecture is the one used by Chefer, Gur, and Wolf (2021) and can be downloaded here: `https://github.com/hila-chefer/Transformer-Explainability`.

Characteristics	**FER+**	**PAPILA**
Labels nature	*Soft Labels* (Probability distribution)	*Class unbalance*
Loss	Cross-Entropy	Weighted Cross-Entropy
Epochs	60	60
Patience	10	15
Learning rate	5×10^{-5} (found empirically)	1×10^{-5} (*Full FT*) 5×10^{-5} (*Frozen*)
Batch size	64	16
Weight decay	0.01	0.1
Optimizer	AdamW	AdamW
Optimization	`autocast` + `GradScaler`	`autocast` + `GradScaler`

Table 3.2: Summary and comparison of the training hyperparameters

Transformers, as most papers use the CNN architecture. I nonetheless inspired myself of Kovalyk et al. (2022) but decided to be more cautious with the learning rate to not destroy totally the weights learned from ImageNet-21k. The number of maximum epochs was chosen small (60 epochs for both) to be able to repeat the training of models efficiently, and a Patience parameter P was added to stop the learning phase if the model does not improve its performances after P epochs.

For the training, I use the (Weighted) Cross-entropy loss. This function measures the difference between the probabilities predicted by the model, and the true label.

$$\mathcal{L} = -\sum_{k=1}^{C} w_k \, y_k \log(\hat{y}_k)$$

where y_k is the true probability, $\hat{y}_k$ the predicted probability, and w_k is the weight associated with the class.

This loss is tailored to fit the characteristics of each dataset. For FER+, y_k is the probability distribution (*soft labels*) of the votes of the annotators. Conversely, for the PAPILA dataset, y_k is standard and single class (meaning that it is equal to 1 for the correct class and 0 for the others), but a weight[4] w_k is introduced to give more importance to `Suspect` and `Glaucoma` classes, under-represented in the dataset as described in Kovalyk et al. (2022). This means that misclassifying images from `Glaucoma` and `Suspect` classes leads to a higher penalty.

Finally, the batch size is slightly smaller for the PAPILA task in order to reach

[4]This weight for each class is computed as described in: `https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html`.

better accuracy, and the weight decay is also quite lower for tasks like FER+, that are more related to the ImageNet-21k dataset; whereas the images of PAPILA significantly differ from the pretraining dataset, requiring the model to prioritize learning new medical elements rather than relying on generic ImageNet-21k details.

For both dataset, I use the same optimizer and optimization. The AdamW[5] optimizer (Loshchilov and Hutter 2019), updates model's weights in order to reduce the error. This optimizer quickly became a standard in the literature because it deals better with weight penalty than the classical Adam optimizer, avoiding overfitting. Automatic Mixed Precision is used in all settings through Autocast+Gradscaler[6]. This technique allows Pytorch to do its computations with 16 bits instead of 32 whenever it's possible, reducing the memory consumption of the GPU and accelerating the training. The Gradscaler is there to avoid that small gradients transform into zero because of the use of 16 bits.

Fine-tuning level

In this study, I investigate two levels of fine-tuning: *Full fine-tuned models*, and *Frozen regimes* . As explained in Chapter 2, the first one allows all the weights of the encoder and classification head to be modified whereas the latter only train the classification head, by keeping all the weights of the Encoder unchanged (using `param.requires_grad = False`. In both regimes, the classification head is always discarded and replaced by a new one, randomly initialized, with output dimension of $K = 2$:

```
model.head = nn.Linear(model.head.in_features, 2)
```

A natural hypothesis is that fully fine-tuned models offer more freedom in weight optimization and should lead to more diverse explanations.

Training Loop

The Algorithm 1 in the Appendix section details the training loop used for both datasets.

The training uses a loop including an early stopping in order do prevent overfitting if the performance does not improve after P consecutive epochs, where P is the Patience hyperparameter (see Table 3.2). The selection metrics to save the best model at each epoch is the accuracy for FER+, and AUC for PAPILA. This corresponds to standards in the literature for these two datasets (Xue, Wang, and G. Guo (2021) and Kovalyk et al. (2022), respectively): FER+ uses the standard metric for evaluation and the use of AUC for PAPILA is justified by its nature: a highly unbalanced and

[5]The "W" stands for decoupled **w**eight decay.
[6]Documentation is available at `https://docs.pytorch.org/docs/2.12/amp.html`.

medical dataset. Results in terms of averaged accuracy in % (FER+) and averaged AUC (PAPILA) are displayed respectively in Tables 3.3 and 3.4 and appears to align with the state-of-the-art (Xue, Wang, and G. Guo (2021) and Kovalyk et al. (2022)).

Task (FER+)	**Frozen**	**Full**
`happiness` vs `sadness`	93.76	98.48
`happiness` vs `neutral`	94.79	96.91
`neutral` vs `sadness`	84.39	89.89

Table 3.3: Average accuracy for FER+ across all settings

Task (PAPILA)	**Frozen**	**Full**
`Healthy` vs `Glaucoma`	0.77	0.81
`Healthy` vs `Suspect`	0.92	0.94
`Suspect` vs `Glaucoma`	0.64	0.84

Table 3.4: Average AUC for PAPILA across all settings

Seeding

For each *setting* (binary task associated with a fine-tuning level), I generate a large number of models. This number of models can go from 200 to 700 in some settings. The choice of the number of models was driven by the necessity to obtain at least 100 equivalent models (see next section). For each dataset, the hyperparameters for all the generated models stay the same, with the exception of the random seed which is modified from one model to another. The training randomness can come from various elements, like the classification head initialization, the dropout, or the image shuffling. This is why I fix multiple seeds corresponding to the libraries used: Python, NumPy, and PyTorch across both GPU and CPU. The goal is to be able to isolate the randomness as the only source of diversity of the explanations, for each setting.

3.2.3 Equivalent Models

The next step is to obtain a sufficiently large (to have statistically significant results) and uniform (to guarantee homogeneous performances) pool of models. For each setting, I extract the top 100 models ranked by accuracy, while making sure that these models are equivalent in the statistical sense. This is done in 4 steps:

1. Compute the accuracy of the models generated in the previous step, in a deterministic way, on the same test set.

2. Sort them in decreasing order of accuracy

3. Perform a statistical test to extract a pool of models equivalent to the best model in terms of accuracy

4. From this pool, keep only the best 100.

For each setting, for all models in this setting, I do a forward pass on each image of the test set, and compute the mean accuracy on this set, before sorting the list of my models in the setting by decreasing order of accuracy. The statistical test mentioned in step 3 to measure if the performance difference between two models is significant is the Z-statistic test (Lehmann and Romano 2005). In my situation, I compare the accuracy a of the best model in the setting against the accuracy b of a candidate model, both computed on the same test set. The null hypothesis H_0 states that the two models have equivalent performances ($a = b$). Z is given by the following formula:

$$Z = \left| \frac{a - b}{\sqrt{\frac{p \times (1-p)}{N}}} \right| \tag{3.1}$$

where $p = \frac{a+b}{2}$ gives the mean of the accuracy of both models. When $Z < 1.96$ (for a confidence level of $\alpha = 0.05$), I cannot reject the null hypothesis H_0. Therefore, the candidate model can be considered functionally equivalent to my reference model, and is kept in my models pool used for my future experiments. The experience is repeated for all trained models in this setting until one model is not considered functionally equivalent. This pool is a functionally equivalent models set, as described in § 2.4.1.

However, one noteworthy element has to be considered here: indeed, as written by Dietterich (1998) in his paper, the z-statistic is supposed to be evaluated on two strictly independent groups. However, in my case, the two models evaluate the same images; their results are then correlated, which can lead the z-test to be inexact (and too much optimistic to reject the null hypothesis). The standard z-test was still used in this thesis given the need to compute easily equivalence between models, but other statistical tests like McNemar's test can be used to account for this dependency.

The exact process used to select my top 100 equivalent models is described in Algorithm 2 (see Appendix). I evaluate iteratively on increasing subset of my models (from $i = 1$ to N_{models}, where N_{models} is the number of models generated from the previous step). For each subset `list_of_models[:i]`, I compare the first model a (i.e. the best model in this setting) and the last model added, b. While $Z < 1.96$, the subset still contains only models considered as equivalent, and the loop continues. When the condition is violated, the loop breaks.

Although this process will typically yield more than 100 models per setting, I restrict myself to the top 100, to standardize the size of my model pool for all future computations and be able to compare settings together to extract meaningful analysis. Retaining exactly 100 models aligns with the methodology used by Loncour (2025) and Bogaert (2026).

3.2.4 Input compatibility

In order for the comparison between models to be valid, it is mandatory to analyze only the situations where all the models selected from the previous section are unanimous on the prediction given for an image, and that this prediction is correct. Indeed, it would not be surprising to obtain different explanations from two models giving different predictions.

This is exactly what motivates the selection of **compatible inputs**, i.e, a subset of the original test set composed of images for which the prediction $\hat{y}_i$ given by model i is equivalent to the prediction of all other models from the pool extracted in § 3.2.3, and corresponds to the true label y.

$$\hat{y}_1 = \hat{y}_2 = \cdots = \hat{y}_{100} = y$$

Therefore, an image of the test set is called **compatible** if and only if all the models predict the same label for it, and this label is correct. The selection process of these so-called compatible inputs, given as input the models, the prediction for each image, the test set and the ground truth for each image belonging to it, is shown in Algorithm 3.

It is important to note that for visual convenience, from these compatible inputs in the FER+ dataset, I only select a subset of **115 images** for each setting and each class when plotting comparative plots. To select this restricted set of compatible inputs, I intersect the set of compatible inputs linked to the exact same task, but with different regimes (e.g., I intersect the set of compatible inputs for the setting `happiness` vs `sadness` in the **Full FT** regime with the set of compatible inputs for the setting `happiness` vs `sadness` in the **Frozen** regime), before restricting to a subset of 115 compatible inputs for each class in setting. This specific number is determined by the most constrained setting: in the `neutral` vs `sadness` task, I only obtained precisely 115 compatible inputs belonging to the `sadness` class in common to both regimes.

However, for all other analyses, relevance vectors and relevance maps were well generated for the whole compatible input set. Finally, no restriction of this kind to a limited number of compatible inputs was needed for the PAPILA dataset, given its

smaller size.

3.3 Research Questions

Now that my pipeline for training and extraction is clearly established, I detail which use I will have of it. Inspired by the work of Loncour (2025) and Bogaert (2026), this section presents several point-by-point systematic comparisons.

3.3.1 Core Analysis

In this section, I present three axes that will structure Chapter 4: Results. Rather than listing the different experiments, I structure these axes as a progression following my thinking process during the elaboration of this master thesis, trying to investigate explanation stability.

As the training strategy naturally describes the degree of freedom given to the model, it was therefore natural to analyze this aspect at the beginning of my study, following the work done by Bogaert (2026).

> **Q1**
>
> *How will the regime used, linked to the degree of freedom given to a model, influence the stability of its explanations? More precisely, does the Frozen regime lead the network to converge to more unanimous explanations than the Full FT regime and do I see a stability-quality trade-off?*

I will compare each *regime* (i.e. *Frozen* vs *Full FT*) on the FER+ and PAPILA datasets, leading to 2 (datasets) × 2 (regimes) × 3 (binary tasks) = 12 settings. For each setting[7], I generate one *FEM set* and select for each setting after models generation the compatible inputs, following the main pipeline described in § 3.2. The comparison between regimes is then done both quantitatively and qualitatively, to assess whether restricting the number of weights the model can tune leads to a greater consensus.

Zooming in further, within a specific dataset and training regime, it seems reasonable to raise the hypothesis that all tasks are not equally challenging. Indeed, among all the binary classification tasks associated with a specific dataset, some seem easier. As an example, it is reasonable to expect that the `happiness` vs `sadness`

[7]As an example, I will obtain 100 models for the setting "`happiness` vs `sadness` in the *Frozen* regime" **and** 100 models for the setting "`happiness` vs `sadness` in the *Full FT* regime".

should be easier to discriminate than `sadness` vs `neutral`. This concept of relative task difficulty is a question in itself (discussed in § 3.3.2), but as it is a real challenge to assess the relative difficulty of a task, I will thus analyze firstly how the choice of the opposing class during training alters the stability of the explanations associated with a specific class.

Q2

Does the stability of explanations for a specific class depend on the complementary class used in the classification task?

Concretely, I will compare the explanations given for the `happiness` images given two distincts settings: `happiness` vs `sadness` (Task A) and `happiness` vs `neutral` (Task B), restricting myself to compatible inputs common to both settings. In both cases, the images used for training are identical, the only difference is the complementary class used during the training. If the Mean Correlation With Mean Explanation[8] (MCWME) for the `happiness` class differs significantly between the two settings, this would suggest that the model does not use the same markers of the face for `happiness` depending on if it has to compare it against `sadness` or `neutral`. I repeat the comparison for the other settings:

- `sadness` in (`happiness` vs `sadness`) against `sadness` in (`sadness` vs `neutral`) ;
- `neutral` in (`happiness` vs `neutral`) against `neutral` in (`sadness` vs `neutral`).

For each pair of settings sharing a common class, the MCWME is computed on their respective compatible inputs sets, and I then compare their relevance maps.

Finally, I wanted to study the stability through a complementary point of view, and study how the models agree or not on specific areas of the image. Indeed, the previous steps measure the stability using the MCWME, which summarizes in a single scalar the similarity between one explanation and the mean explanation, but it is highly desirable to obtain more grained results, a preciser information on how this stability is spread in the image, at the level of a patch.

Q3

How does the MCWME vary when I consider only the top-k relevant pixels? Furthermore, where is the variance of the relevance located in the image? Are some specific areas systematically more variable than others, and is this focus coherent with the nature of the task or centered on artifacts?

[8]This metric is described in § 2.5.2 .

To answer the first part of this section, I will focus on graphs depicting the evolution of the MCWME when I consider the top-k relevant pixels to compute it. On the same graph, I will depict for a specific task (e.g. `happiness` vs `sadness`) the evolution of this top-k MCWME with the number of pixels considered[9]. For each compatible input, I will rank all patches according to their mean relevance on the 100 models of the FEM set. Then, I fully mask the image before unmasking progressively, patch after patch, in decreasing order of relevance. At each step (i.e. for each k), the MCWME value is computed. The curve obtained show how this top-k MCWME is evolving with increasing areas of the images considered, with k on the x-axis (ranging from 1 to 50176) and the MCWME on the y-axis, with the lines and confidence intervals in blue for the *Frozen* regime, and in red for the *Full FT* regime.

To complete this analysis, I use a complementary approach by focusing instead on the spatial distribution of the *variance*. Instead of merely assessing whether the explanations are stable from one model to another, I am focusing on *where* this difference is located. To achieve this, I will compute the patch-level spatial variance (detailed in § 2.5.4), and then generate maps to observe exactly where this variance is located on the image.

3.3.2 Exploratory Analysis

Beyond the three main research axes investigated in this master's thesis, two additional questions emerged. However, their inherent complexity along with the need of a rigorous statistical comparison make these questions hard to address and will therefore be studied from a more exploratory point of view.

At the genesis of this work, a fundamental question emerged: which dataset should I use? I decided to work with two different datasets: FER+ and its facial expression, inherently subjective, and PAPILA, a medical imaging dataset for Glaucoma and is more objective, as images are labeled as `Healthy` or `Glaucoma` and `Suspect` cases have their own class. This choice mimicks the choice of Loncour (2025) to use datasets radically different in terms of subjectivity: AlloCiné and InfOpinions. This relation between corpus subjectivity and sensitivity to training randomness is therefore a natural topic to analyze.

QA

Does the stability of explanations vary depending if the task uses a subjective dataset (FER+) or a more objective dataset (PAPILA)?

However, addressing this question requires high caution. It is highly difficult

[9]This metric is described in § 2.5.3 .

to decide since I only have one representative dataset for each level of subjectivity, and given that these two datasets differ significantly in terms of size and nature, drawing general conclusions is a perilous task. However, I will still show some analysis about the PAPILA dataset I found noteworthy in Chapter 5: Exploratory Analysis.

Another aspect that comes into mind when analyzing the binary tasks used in this master thesis is their relative difficulty. As mentioned above, quantifying relative difficulty of a task is a challenge in itself. For the explanations to be meaningful, I need to restrict to compatible inputs (cf § 3.2.4), leading therefore to an accuracy of 100% for all my models on a FEM set. Therefore, the accuracy cannot be used as a proxy for task difficulty. Moreover, while some other metrics (like the confidence in the prediction[10]) might appear suitable, it would require advanced statistical test to validate its significance. In a similar way, using global measures like an "aggregated MCWME" across all models and compatible inputs belonging to a setting is perilous, and requires a rigorous statistical study. Without this highly rigorous statistical framework, I will therefore restrict my analysis to a more informal exploration in Chapter 5.

QB

Is the stability of explanations influenced by the difficulty of the classification task?

To explore this potential correlation, I will use the model's average confidence in the prediction as a proxy for the difficulty of the task. By analyzing these results along with aggregated MCWME values, I am aiming to observe qualitative trends that will require a more rigorous statistical framework as further exploration to be meaningful. Moreover, I will analyze the plots depicting the MCWME, along with its confidence interval, for compatible inputs in each setting.

Axis	Research Question	Dataset(s)	Chapter
1	**Fine-tuning Regime**	FER+, PAPILA	4
2	**Contrastive Analysis**	FER+	4
3	**Patch-level analysis**	FER+	4
a	**Corpus Subjectivity**	FER+, PAPILA	5
b	**Task Difficulty**	FER+	5

Table 3.5: Summary of Research Questions

[10]The confidence in the prediction is the probability value assigned by a model to its final prediction.

Chapter 4

Results

In this Chapter, I will answer the three core experimental questions designed in the last Chapter, both quantitatively and qualitatively. For each research axis, I recall briefly the central question, before confronting the results to my initial hypotheses. It is crucial to note that all results presented in this section are heavily dependent on the considered tasks and datasets used. Therefore, the results cannot be generalized blindly.

4.1 Fine-tuning Regime

Q1

How will the regime used, linked to the degree of freedom given to a model, influence the stability of its explanations? More precisely, does the Frozen regime lead the network to converge to more unanimous explanations than the Full FT regime and do I see a stability-quality trade-off?

Before the analysis of the results, I formulate my expectations based on the prior work of Bogaert (2026). Indeed, the *Frozen* regime should lead the model to learn only to use the embeddings without modifying them. Consequently, I am expecting that Layer-wise Relevance propagation (LRP) will distribute the relevance in a more uniform way across the *Frozen* models, compared to their *Full FT* counterparts. This would lead to a higher Mean Correlation With Mean Explanation[1] (MCWME) score for the *Frozen* models, showing a better stability across explanations. However, my hypothesis is that this higher stability is artificial: the *Frozen* regime, less expressive, will produce less discriminating relevance maps, with less focus on crucial areas[2] than the *Full FT* models. This raises the following question:

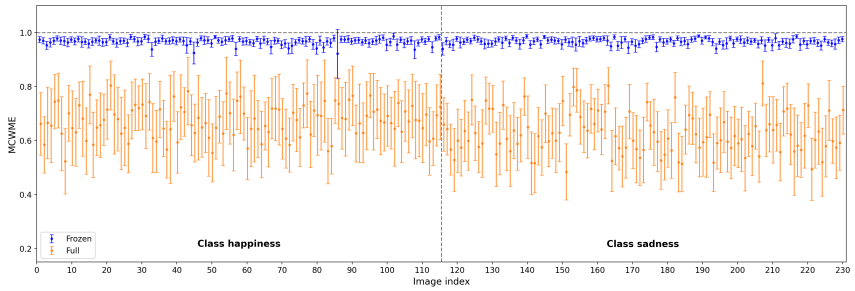
[1]This metric is described in § 2.5.2 .

[2]For emotion detection, these areas would be emotion markers like eyes, mouth, or lines of the face.

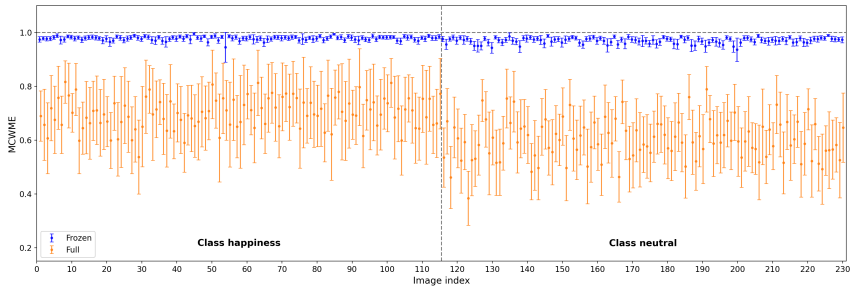
How do I characterize the trade-off between stability and quality of explanations across these two training regimes?

4.1.1 Quantitative Analysis

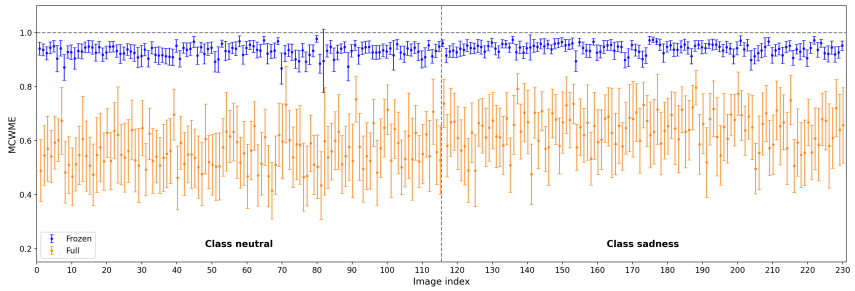
The three subplots in Figure 4.1 show the distribution of the mean MCWME (y-axis) and the confidence intervals across the 115 compatible inputs [3] (x-axis) for each setting [4] of FER+. Class A is on the left part of the graph, while class B is on the right part. Blue results are produced by explanations based on *Frozen* models, whereas orange intervals are those produced by *Full FT* models.



(a) `happiness` vs `sadness`.



(b) `happiness` vs `neutral`.



(c) `neutral` vs `sadness`.

Figure 4.1: Comparison of the mean MCWME on compatible inputs for FER+. No overlap in the confidence intervals shows a statistically significant difference.

The first conclusion extracted from these plots strongly confirms my initial hy-

[3] The selection of these 115 inputs is detailed in § 3.2.4 .

[4] A *setting* is a combination of a binary task (on a specific dataset) and a fine-tuning regime.

pothesis: the mean MCWME is higher in the *Frozen* regime than in the *Full FT* regime. Furthermore, the wider confidence intervals observed for the *Full FT* regime compared to those for the *Frozen* regime confirm my assumption: the LRP method used on *Frozen* models tend to give more consistent results across models. The complete absence of overlap between these blue and orange intervals for every compatible input is a visual confirmation of statistical significance. This narrower spread in *Frozen* models shows that LRP gives more stable and consistent explanations across different models from the same functionally equivalent models set (FEM set).

These two remarks are robust and particularly strong conclusions that I can draw for every setting and compatible input tested. Notably, similar experiments were repeated for the PAPILA dataset and led to similar trends.

4.1.2 Qualitative Analysis

When examining the relevance maps qualitatively , I observe that, overall, explanations from the *Frozen* are more homogeneous from one model to another, but they show a more dispersed relevance, without any focus on specific areas. In contrast, maps from the *Full FT* regime are more variable, but the relevance is more focused on specific areas of the face. For instance, in facial recognition tasks, the *Frozen* models relevance maps fail in general to accurately identify specific and accurate regions traditionally associated with emotion markers. Looking at Figure 4.2, I indeed observe this contrast through representative examples coming from both FER+ and PAPILA dataset.

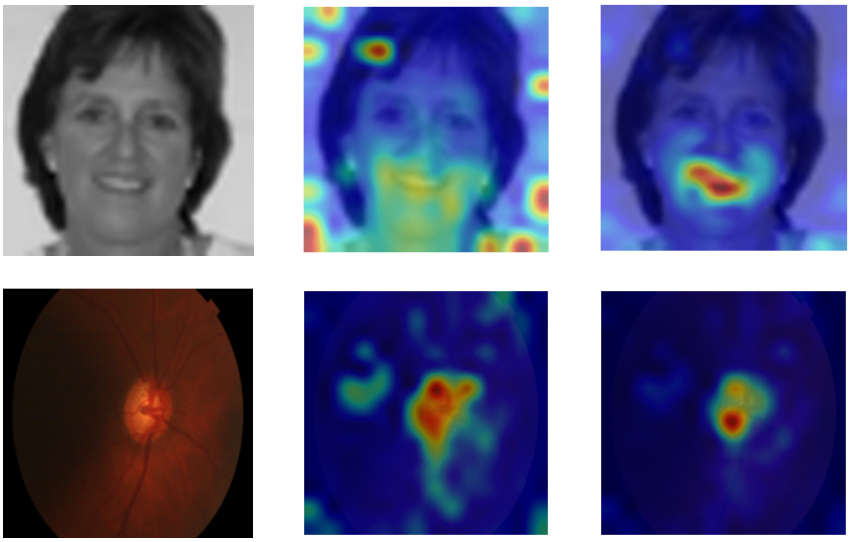


Figure 4.2: Mean relevance maps comparison. Top: An image belonging to the `happiness` class on FER+ (Original, Frozen, Full FT). Bottom: An image belonging to the `Healthy` class on PAPILA (Original, Frozen, Full).

Although drawing general conclusions for PAPILA is more challenging, due to

the smaller size of the dataset, the same trends can be observed across all its settings.

4.1.3 Summary

At the end, this contrast shows a clear trade-off between stability and quality of explanations: the *Frozen* regime being more stable but less informative, whereas the *Full FT* regime gives more meaningful, but also more variable explanations. In the second case, even if the explanations are less stable, they might be more informative and trustworthy for a human user. From a practical deployment perspective, this is a real dilemma: should we prioritize consistent explanations that are less informative (*Frozen*) or more plausible explanations that might vary across models (*Full FT*)? The answer likely depends on the specific application context, and is not black or white: some applications might require an in-between training regime where only a part of the weights can be modified.

If the stability appears to be constrained by the regime used, does it also vary depending on the nature of the task? I explore this question in the next section by examining if the choice of the contrastive class influence how the LRP method explains models decisions.

4.2 Contrastive Analysis

Q2

Does the stability of explanations for a specific class depend on the complementary class used in the classification task?

I investigate here an intriguing question related to explanation stability: when a model learns about a specific class (e.g, the `happiness` class), does the choice of the contrasting class (`sadness` or `neutral`) influence which facial features the model will focus on, and therefore the stability of the explanations?

It is hypothesized that the complementary class used will significantly affect the stability of explanations. With `sadness` being the class of interest, I expect that in the `happiness` vs `sadness` context, the MCWME of `sadness` will be higher compared to the other comparative context (`neutral` vs `sadness`), showing a better stability. Indeed, in the former task, the facial features for the `happiness` class are clearly opposed to those of `sadness` class (e.g, the shape of mouth, eye, eyebrows). However, I expect the `neutral` vs `sadness` task to lead to less stable explanations for

the `sadness` class as the difference between images belonging to these two classes is more subtle and might depend on more diverse face markers (not limited to obvious elements, like the shape of the mouth or the eyes).

Moreover, I might observe similar trends when `happiness` or `neutral` are my classes of interest. Nonetheless, I expect a weaker impact than when I consider `sadness` as my class of interest.

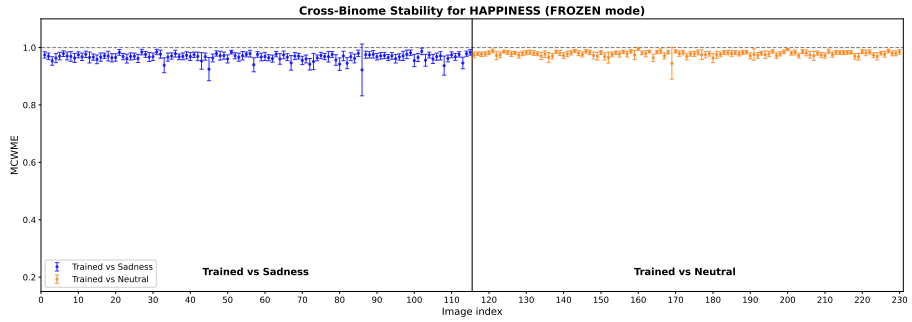
I divide the answer to this question in 3 main sections, covering each class of interest used in my restrained FER+ dataset. Following the same structure as above, Figures 4.3, 4.5 and 4.7 show the MCWME for each single compatible input, aggregated across the 100 models in the FEM set. The x-axis represents compatible inputs, while the y-axis shows the MCWME with confidence intervals, in both settings involving the two different comparative classes. Alongside this quantitative analysis, a qualitative analysis of relevance maps, depicted at Figures 4.4 and 4.6, is conducted.

However, it is critical to mention that the sets of compatible inputs is specific to each task. This means that the indices on the x-axis representing the compatible inputs for one task (1-115) and another task (116-230) are not identical from one task to another. As an example, input 15 in the `happiness` vs `sadness` task does not correspond to input 130 (in the left part of the graph depicting `happiness` vs `neutral`), since both settings do not share especially the same set of compatible inputs: an input compatible in one setting may not be compatible in another one.

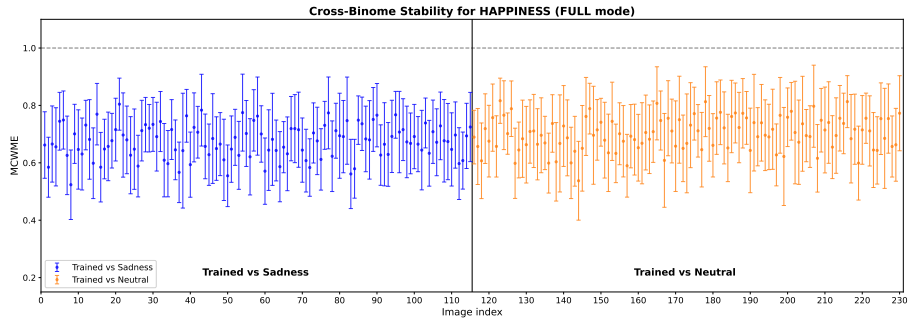
4.2.1 The `happiness` class

Figure 4.3 shows the MCWME distribution for `happiness` with two different contrastive classes: `sadness` and `neutral`. Inspecting closely this figure, no substantial differences depending on the setting image of this class are trained on can be observed. For both regimes, it appears that the mean MCWME is high and similar in both settings, with what seems to be a narrower confidence interval for MCWME when `happiness` is trained against `neutral` in the *Frozen* regime. The MCWME seems slightly higher when `happiness` is trained against `neutral`. One hypothesis, not formally tested, is that this difference, if significative, might be due to the larger number of images belonging to the `neutral` class in the training set, leading the models to converge more in their explanations. For what concerns the *Full FT* regime, results appear to be similar for both settings.

Overall, the differences observed in this section do not appear to be significative. This similarity in the MCWME results do not reveal any dependence to the contrastive class used during training, and seem to validate what I hypothesized at the beginning of this section, face markers of `happiness` are highly visible elements (like a smiling



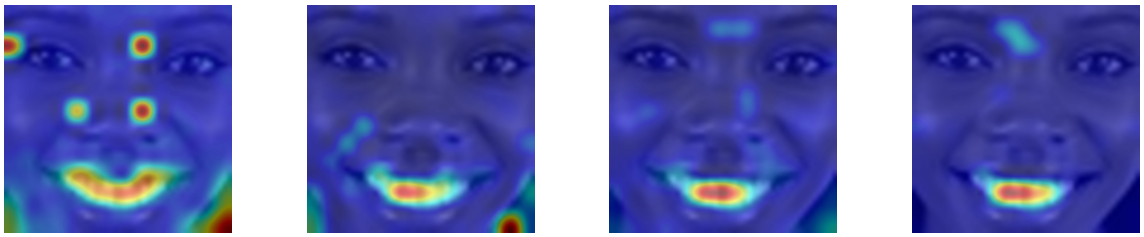
(a) Frozen



(b) Full

Figure 4.3: Contrastive Analysis of the `happiness` class

mouth, cheeks raised, or laughing eyes) that define this emotion, independently of the contrastive class used against it.



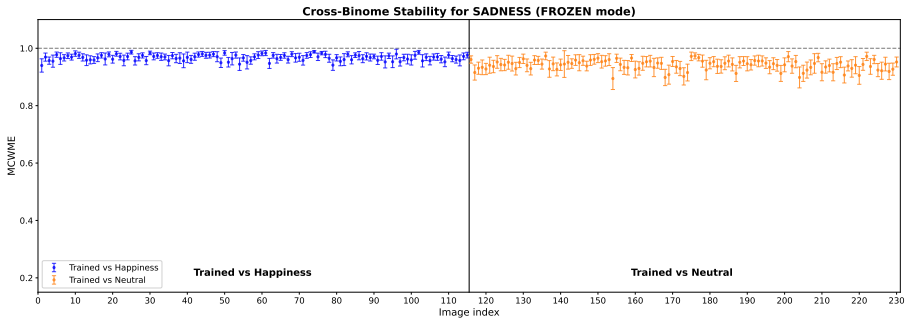
(a) (b) (c) (d)

Figure 4.4: Relevance maps of the image `fer0032419`. (a) and (b) depicts explanations from models in the `happiness` vs `sadness` context. (c) and (d) depicts explanations from models in the `happiness` vs `neutral` context. All LRP explanations are based on models in the *Full FT* regime. Maps selected are representative of relevance maps for this task.

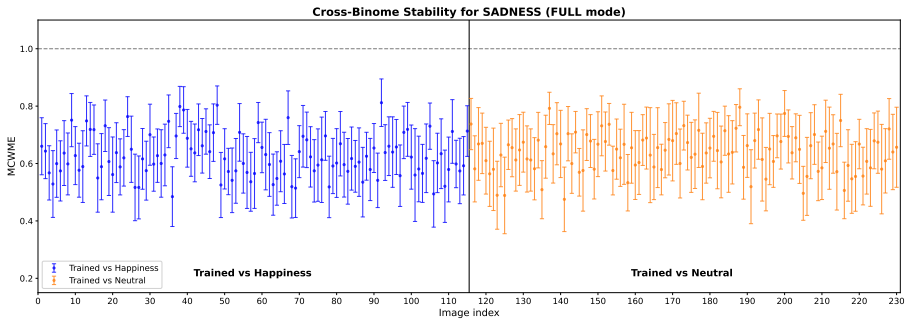
Inspection of Figure 4.4 representing an LRP explanations on an image from the class `happiness` shows that the first two maps appear to have very similar explanations when compared to the two last (they all target mostly the mouth), even if they are associated with models trained in two different contexts. Nonetheless, it is noteworthy that in my qualitative analysis, the main objective is not to describe the stability across models (this is the goal of Figures depicting quantitative results, like Figure 4.3), but rather visually confirm or infirm the conclusions I draw from the

quantitative analysis, by inspecting which areas are highlighted by the Layer-wise relevance propagation method.

4.2.2 The `sadness` class



(a) Frozen



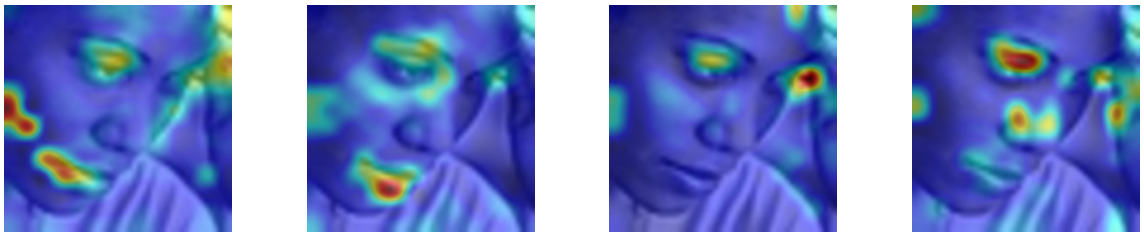
(b) Full

Figure 4.5: Contrastive Analysis of the `sadness` class

Inspecting the `sadness` class, shown in Figure 4.5, I observe a clear difference, at least in subfigure 4.5a depicting the *Frozen* regime. The MCWME when `sadness` is trained against `neutral` appears to be lower, and the confidence intervals appears to be wider in this setting, showing more variability in explanations between models.

In the *Full FT* regime, it is less clear to determine if there is a similar pattern, because of the higher confidence intervals. In this case, when distinguishing the `sadness` class from the `neutral` class, models have to rely on more subtle elements that may differ in relevance from one model to another.

Indeed, I obtain a visual confirmation of this analysis when looking at Figure 4.6. If I look closely at the first two maps, generated within the `happiness` vs `sadness` setting, I observe that the mouth is a clear face marker that the model assigns high relevance to, in order to classify the image as belonging to the `sadness` class. However, in the last two pictures, I observe that this is clearly less the case in the `neutral` vs `sadness` setting, where some elements like the eyes seem more important

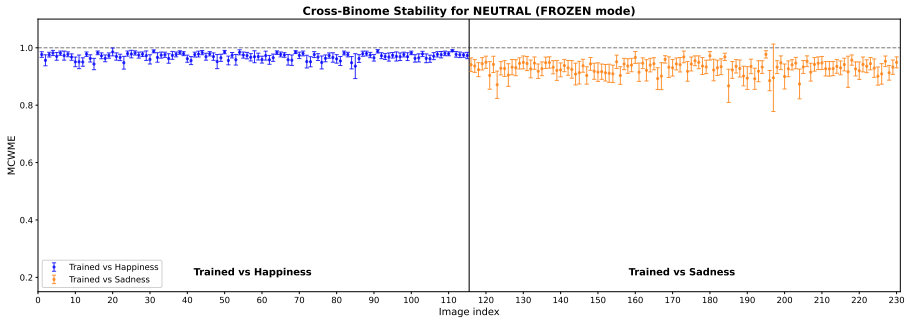


(a) (b) (c) (d)

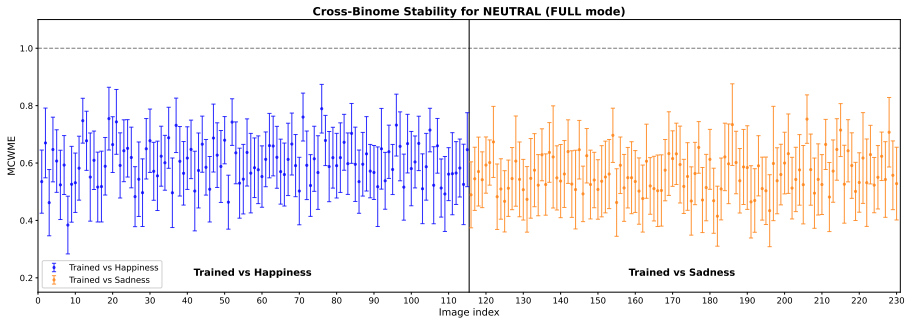
Figure 4.6: Relevance maps of the image `fer0032419` labeled as `sadness`. (a) and (b) depicts explanations from models in the `happiness` vs `sadness` context. (c) and (d) depicts explanations from models in the `neutral` vs `sadness` context. All LRP explanations are based on models in the *Full FT* regime.

for these models. This result is, however, not generalizable for each compatible input belonging to the `sadness` class, and facial features highlighted remains highly dependent of the specific compatible input considered.

4.2.3 The `neutral` class



(a) Frozen



(b) Full

Figure 4.7: Contrastive Analysis of the `neutral` class

Finally, focusing on the `neutral` class, I observe a similar pattern to the one observed for the `sadness` class, and results for this class are therefore discussed more concisely in this subsection. Indeed, I observe a slightly lower confidence interval the models when the class is trained against `happiness` in the *Frozen* regime, suggesting

more stable explanations; however, this difference in variability remains small.

4.2.4 Summary

The difference in context shown in Figure 4.6 reveals that, depending on the context of the training, maps fundamentally different yet equally plausible can be obtained. This is one central finding of this section: **a LRP explanation may not be specific to an image, but rather dependent of the task on which the model has been trained**. This is an expected result when comparing to the behavior of a human annotator. It is plausible (although no formal study of this kind has been conducted and this hypothesis remains speculative) that a person would prioritize different areas to circle if I ask her/him two different questions:

- Which areas of the face make you think that this face depicts `sadness` **rather than** `happiness`?
- Which areas of the face make you think that this face depicts `sadness` **rather than a** `neutral` **expression**?

However, the results I presented in this section remain specific to the FER+ dataset and the three emotions studied, as the lack of another large dataset in the dataset does not allow me to draw general conclusions. Studying a broader range of emotions may be a future research path. Moreover, a statistical test on specific inputs would be required to see if we observe statistically significant differences.

While my experiments showed that the change in contrastive class may lead to an increase or decrease of stability of explanations depending on the proximity between the class of interest and the adversarial class, the MCWME remains a global metric aggregating the variability and masking how the relevance is spatially distributed on the image. This is why I now focus on a patch-level analysis.

4.3 Patch-level Analysis

Q3

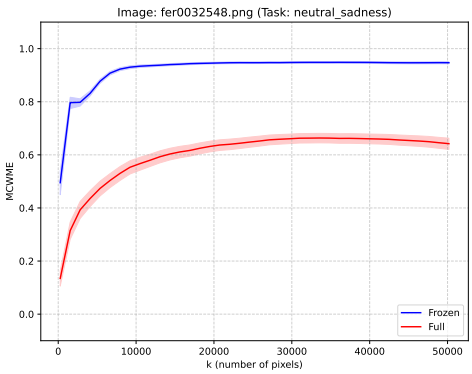
How does the MCWME vary when I consider only the top-k relevant pixels? Furthermore, where is the variance of the relevance located in the image? Are some specific areas systematically more variable than others, and is this focus coherent with the nature of the task or centered on artifacts?

As seen in the previous section, the MCWME is a useful indicator because it aggregates in one single scalar the stability of explanations across models for one

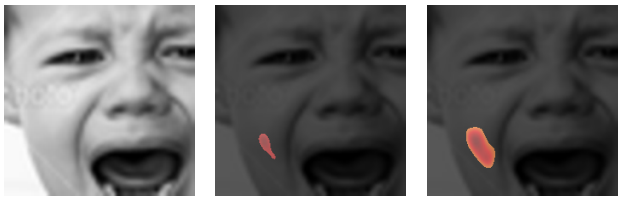
compatible input. However, it does not allow me to understand on which specific areas the models agree or disagree, and it is exactly what I will explore in this section. I am addressing this question through two complementary analyses. Firstly, I will study how the MCWME evolves when I consider only the top-k most relevant pixels, and in a second step I will be computing the spatial variance of the relevance to identify on which precise regions the models differ the most.

4.3.1 Top-k MCWME Analysis

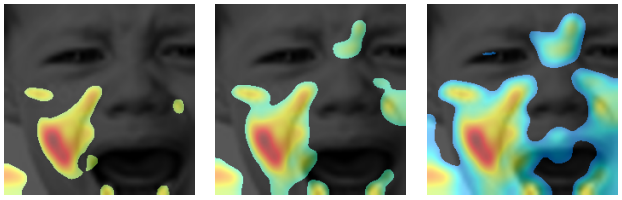
For what concerns the top-k MCWME behavior, I am expecting to see lower MCWME scores and a wider confidence interval when considering only the first pixels (small k). It is indeed reasonable to observe a pattern consistent with the work of Loncour (2025) and Bogaert (2026) in the NLP domain. More precisely, Bogaert (2026) discovered that the top-k MCWME always follows a specific pattern: it starts at a very low value of k=1, growing fast for small values of k, before stabilizing to a plateau at the approximate value of the global MCWME, showing that consensus is reached when considering more words. Adapting this observation to computer vision, my assumption is that models might disagree on which single specific patches of pixels are the most important, but agree on broader areas of interest. For instance, I can hypothesize that some models might focus certain parts of the mouth, others the eyes, or some contour lines of the cheeks. However, I expect all the models to agree on all these faces features when k (number of pixels) increases, showing a plateau when k reaches a sufficiently large value, showing that consensus is reached when considering enough pixels. Furthermore, as observed in § 4.1, I am expecting to see a higher top-k MCWME for the *Frozen* regime.



(a) Top-k MCWME curve



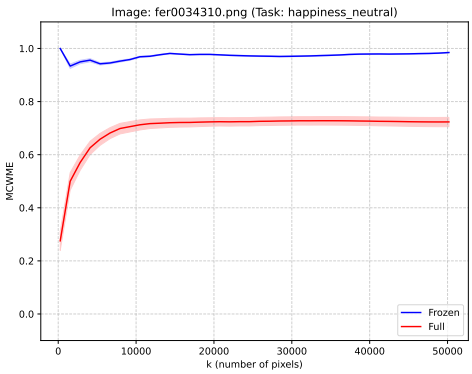
(b) Original (c) $k = 256$ (d) $k = 1024$



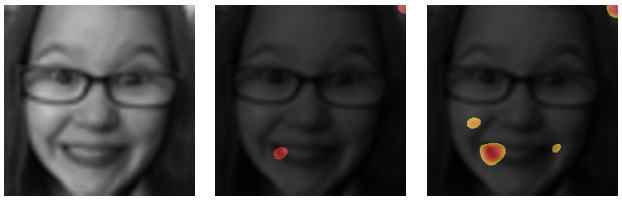
(e) $k = 6272$ (f) $k = 12544$ (g) $k = 25088$

Figure 4.8: Evolution of the top-k MCWME (a), original image `fer0032548.png` (label: `sadness`) (b) and masked mean relevance maps for different values of k (c-g). Setting: `neutral` vs `sadness` in the *Full FT* regime.

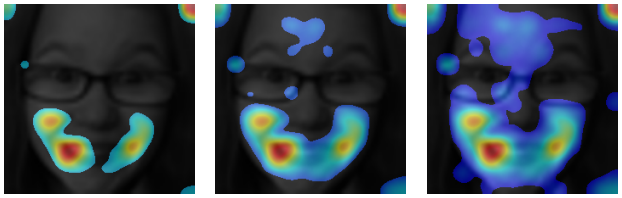
As depicted on Figure 4.8a, the top-k MCWME curve show a standard pattern representative of the images from the FER+ dataset. Indeed, I observe that the MCWME starts at a value dramatically low (especially for the *Full FT* regime), before rising when k increases and reaching a plateau around $k = 25000$, close to the (global) MCWME observed for this image (at $k = 50176$, when all the pixels are considered). For the *Frozen* regime, a similar behavior can be observed, but the plateau is reached at lower values of k ($k = 8000$). The confidence interval displayed is larger for the *Full FT* regime. This Figure brings me confirmation of what has already been observed in § 4.1, but it also allows me to confirm my hypothesis: the top-k MCWME curve show that a stable consensus is reached when considering higher values of k (wider regions). Visually, subfigure 4.8b displays the original image, while 4.8c to 4.8g show the masked mean relevance maps, i.e. the average relevance map of all models in the FEM set, masked to only consider the k most relevant pixels. These illustrations show that the mean relevance when k is low is concentrated on small specific area (like on specific line on the cheek), and increasing k leads to broader areas (face lines, shape of the mouth), leading the models to agree on those broader facial features. I observe that the eyes are considered non-relevant by the model: this is a behavior observed in a high number of examples, showing that, in general, a single facial feature (like the mouth) is considered enough for the model to make its prediction and aggregates the biggest part of the relevance (here, the line on the cheek).



(a) Top-k MCWME curve



(b) Original (c) $k = 256$ (d) $k = 1024$



(e) $k = 6272$ (f) $k = 12544$ (g) $k = 25088$

Figure 4.9: Evolution of the top-k MCWME (a), original image `fer0034310.png` (label: `happiness`) (b) and masked relevance maps for different values of k (c-g). Setting: `happiness` vs `sadness` in the *Full FT* regime.

Looking at Figure 4.9, I observe a similar curve trend compared to Figure 4.8: a *Full FT* curve rising slowly with a wide confidence interval, and a *Frozen* curve reaching the stable level of MCWME faster (even though, in this case, even when considering k small, the MCWME is already high: this is a behavior I presume more

frequent for the *Frozen* regime since, as we showed already, this regime shows more stability in explanations: this pattern is thus not surprising even when considering k small). The higher MCWME overall is due to a very small area of the face (essentially, the mouth) being targeted, even when considering k high ($k = 12544$). I observe that models reach a stable plateau around this value of k, showing that models reach consensus when considering the mouth. Again, the eyes are considered non-relevant by the model, the sole lines around the mouth and smiling teeth being considered to predict the `happiness` class for this image. This behavior of focusing mostly the mouth and ignoring other face markers, like the eyes, is a consistent pattern across images from the `happiness` class, and should be studied in further analysis.

4.3.2 Variance Analysis

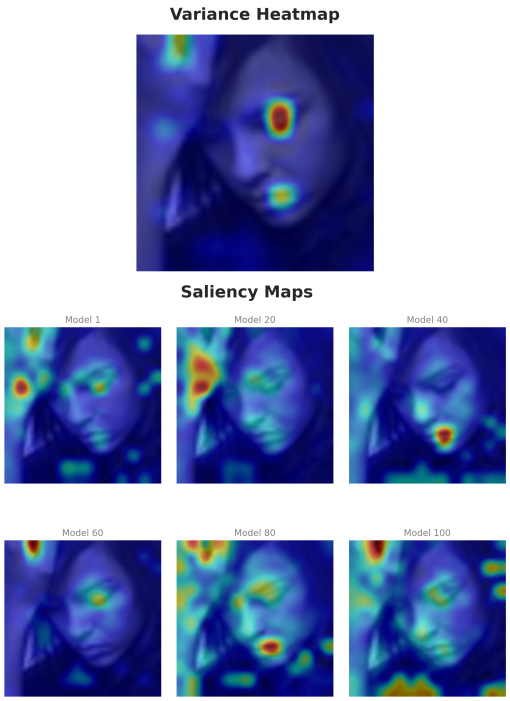
When considering the variance of the relevance analysis, I am expecting to see a non-uniform variance, concentrated on specific areas (facial features) because models generally agree on these areas being important, but disagree on more precise zones (e.g: should the focus be on the left, or right part of the eyebrow?) and also disagree on the relative importance to give to one feature compared to another (e.g: disagreement on the relative importance of the mouth compared to the relevance of the eyes). Less relevant regions like the background should display a lower variance regardless of which regime I use, as I expected either al the models to:

- not consider these areas as relevant;
- or be biased by a specific artifact in the background (but in this case, I am expecting all the models to agree on this artifact being relevant, leading to a low variance anyway)

This notion of models being biased by the background is an interesting topic in itself, but variance maps would probably not help to detect these bias.

I am expecting *Frozen* models to show a lower variance than *Full FT* models. The variance maps, described in § 2.5.4, will show with a blue to red color scale the variance of the patch-wise relevance across the 100 models of the FEM set, for a given compatible input.

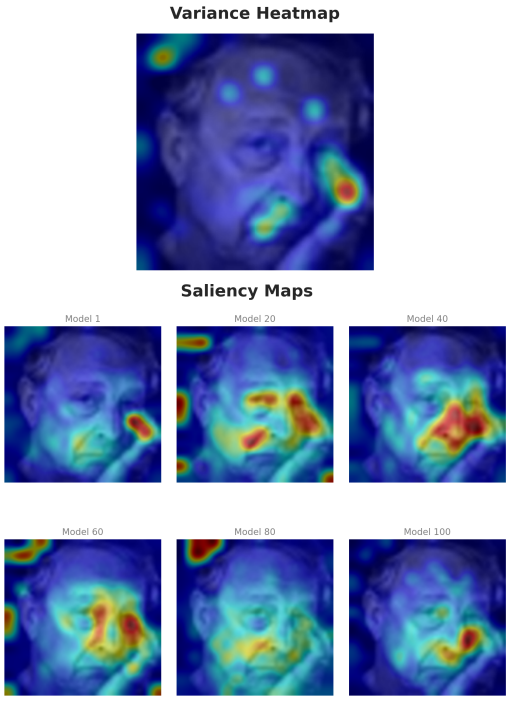
Looking at Figure 4.10, I observe four subfigures depicting representative cases of variance maps among the three classes, within different settings (but always in the *Full FT* regime, as we established in § 4.1 that relevance coming from the *Frozen* regime seem more spread-out and less informative). Of course, I had to make a choice among the amount of available compatible inputs and associated variance maps. The choice of these four subfigures is motivated by the interesting behavior of the Layer-wise Relevance Propagation method that are informative, in



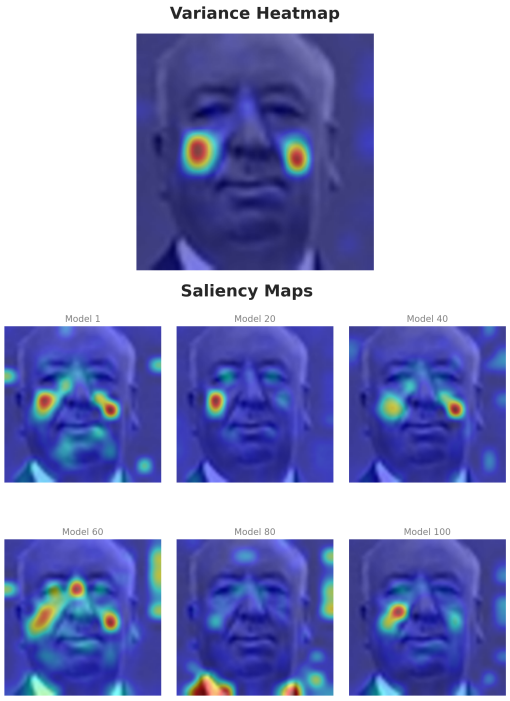
(a) fer0032583. label: `sadness`. Regime: Full, `neutral` vs sadness



(b) fer0033127. label: `happiness`. Regime: Full, `happiness` vs neutral



(c) fer0032487. label: `sadness`. Regime: Full, `happiness` vs sadness



(d) fer0032670. label: `neutral`. Regime: Full, `happiness` vs neutral

Figure 4.10: Representation of 4 variance maps, along with some examples of relevance maps. For each subfigure, the image at the top is the variance map (variance of relevance averaged for all models), and the six smaller figures are relevance maps.

particular cases that appear to be very *plausible*[5] for a human (and therefore showing hints of the strengths of this method for a potential use in practical contexts), or cases appearing conversely non-plausible for a human (these kinds of results being equally noteworthy). The methodology to analyze each of these variance maps will always follow the same order: analysis of both relevance maps and variance map and confrontation with what I expected initially.

Looking at subfigure 4.10a, the relevance appears to be distributed across areas that appear *plausible*: the eyes, the mouth, maybe the position of the nose, or the hand close to the face (this area being considered relevant is not very surprising, given that many images from the `sadness` class show this kind of posture, with a hand close to the face (see subfigure 4.10c)). However, some other elements seemingly less plausible for a human like some patches in the hair or in the neck also receive relevance. Overall, the variance map highlights distinctly the three main elements associated the sadness: the left eye, the mouth and the hand. This is the first interesting pattern I observed when analyzing the variance of relevance among models: *the variance of relevance appears to act as an interesting way to highlight regions of disagreements, that often show relevant parts of the images.* Indeed, if some regions of the face are judged irrelevant by the majority models, the variance associated with this area will be low. For the variance to be high, there are two extreme cases: it should be either a part of the image that is targeted by most models but with intensity varying considerably; or a part of the image targeted only by a subset of models, and ignored by the others. In this case, I observe as well that some elements less interpretable for a human user, like the wrist, the neck, or the nose, show moderate variance even if these areas are targeted by some models.

In subfigure 4.10b, I observe in the relevance maps as well what seems to be a high consensus on the mouth region (even though not all models target it with the same intensity). Some models also highlight the eyes, or facial lines, but models do not reach a consensus on these areas. Some artifacts in the background or in the hair/forehead are also highlighted. In the variance map, I observe a similar behavior as for subfigure 4.10a, with the features generally admitted as relevant to classify `happiness` faces (eyes and mouth) being highlighted.

Subfigure 4.10c shows that the bottom-right part of the face is assigned the majority relevance. However, for the variance map, some specific areas like the fingers or the left part of mouth are highlighted. It seemed noteworthy for me that mostly these markers are highlighted for the variance maps, especially because I expected the eyes to show greater variance. This appears to be attributable to the low relevance attributed by the models to this area in general, which lowers the variance itself.

[5]A plausible explanation for a human would be an explanation focusing on specific features generally considered as relevant emotion markers: mouth, eyes, face lines, eyebrows.

Finally, subfigure (4.10d) is a particularly relevant example depicting a `neutral` face which, despite the high number of training samples from this class in the training set, can be challenging for a model to classify as it lacks the distinct facial features specific to the two other classes, and because `neutral` are close in their facial expression to the `sadness` class. For this specific example, surprisingly, instead of targeting the mouth, the LRP method mostly highlights in the relevance maps the facial lines and creases, but with different levels of intensity (alongside other artifacts such as the background, or the suit). This translates to higher variance in these areas.

Across these four cases, I observed a common pattern: variance of relevance tends to be the highest in meaningful areas of the face.

4.3.3 Summary

What are the main conclusions and answers to extract from this analysis? Firstly, the top-k MCWME analysis confirms the main hypothesis: models do not always agree on very specific patches of pixels, but rather on slightly larger areas: these figures shows that models agree on which areas are essential, but disagree on their relative importance. Spatial variance acts slightly like **a relevance map of the relevance map**. It does not simply show where models look, but rather on where they disagree *productively*: the disagreement on areas judged as relevant by the models, but interpreted differently. The *Frozen* regime, as expected, shows variance maps less informative as the variance is almost totally uniform, showing that its apparent stability hides an absence of capability to accurately focus on genuinely relevant areas. This shows that models are using different representations (that can even be artifacts, like the outfit or the background) to predict the correct class, without any of these explanations being objectively correct.

Chapter 5

Exploratory Analysis

This section details the two additional questions mentioned in Chapter 3: (QA) and (QB). While these questions lack a rigorously enough framework, their results offer nonetheless qualitative information on two other axes of sensitivity to randomness of explanations.

5.1 Corpus Subjectivity

QA

Does the stability of explanations vary depending if the task uses a subjective dataset (FER+) or a more objective dataset (PAPILA)?

The choice to study two domains so different (emotion detection and glaucoma diagnosis) is motivated by an intuition: the subjectivity of the task should influence the stability of explanations. My initial hypothesis is that medical imaging is based on specific markers. However, emotion detection is a more subjective task where human annotators could legitimately diverge. More precisely, in the PAPILA dataset, I am expecting relevance to be high on the area of the optic disc (plausible area for an opthalmologist (Almazroa et al. 2015)), located at the center of the image. For what I will observe in the following section, this hypothesis was refuted partially.

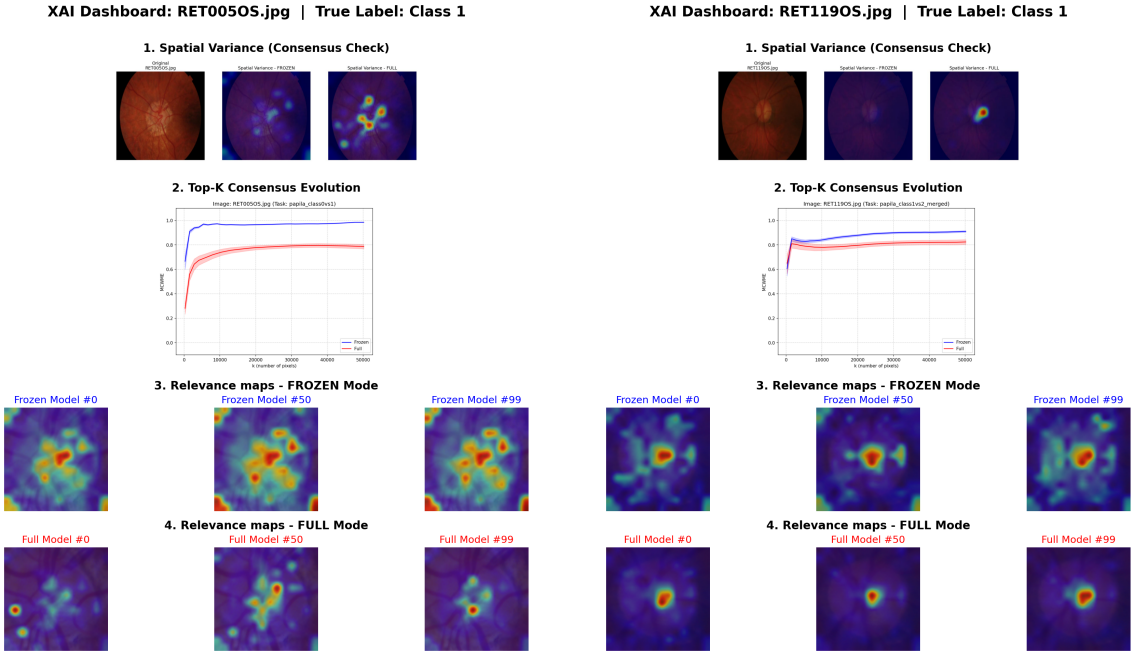
5.1.1 FER+

This section will be relatively short, as Chapter 4 already details the observations based on this dataset. These results show that the LRP method applied to FER+, my subjective dataset, gives explanations whose stability varies, but also show patterns observed across the whole dataset. The *Frozen* regime gives a higher MCWME than the *Full FT* regime, but this does translate into more spread-out relevance. The

choice of the contrastive class appears to affect the variance of the MCWME. Results shown by both relevance and variance map appear to be partially plausible and the disagreement often comes from face features plausible for a human.

5.1.2 PAPILA

A first difficulty appears when selecting compatible inputs. The vast majority of these compatible inputs come from the `Healthy` class. Since I only have a small number of compatible inputs belonging to other classes, it is hard to draw conclusions that are not class-dependent; and it is even more problematic that the `Glaucoma` class is the one showing specific markers of the disease that I expect the LRP method to reveal.

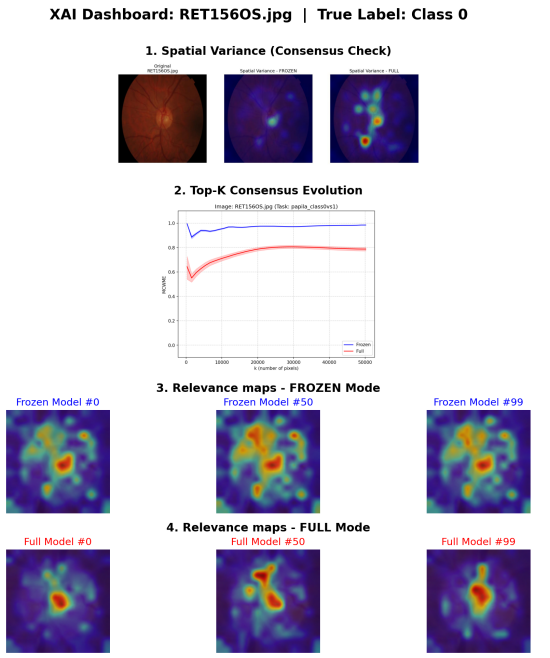


(a) RET005OS: `Suspect`. Task: `Healthy` vs `Suspect`

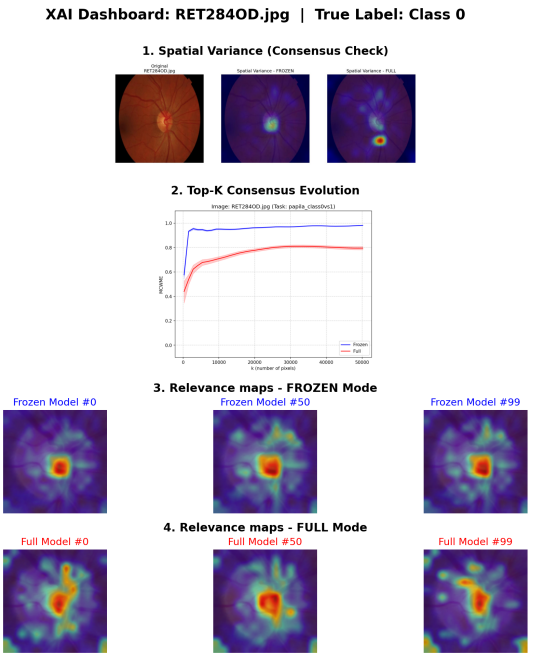
(b) RET119OS: `Suspect`. Task: `Suspect` vs `Glaucoma`

Figure 5.1: Summary of results from the PAPILA dataset. Each subfigure depicts the original image, the variance map, the top-k MCWME curve and six relevances maps (the top three based on *Frozen* models, and the bottom three based on *Full FT* models). For each subfigure, the class of the image, along with the task in which it has been extracted as compatible input, is indicated. (Part 1/2)

Figure 5.1 synthesizes four representative cases for the PAPILA dataset. For each of these images, the original image is displayed, along with the variance map, the top-k MCWME curves, and relevance maps of several models. Some models focus on the optic disc (see Figure 5.1b), but in general, explanations appear to be less plausible for a human as the relevance is really spread-out. Looking at subfigures 5.1c and 5.1d, I observe that even though these two images are classified as `Healthy`,



(c) RET156OS: `Healthy` Task: `Healthy` vs `Suspect`



(d) RET284OD: `Healthy`. Task: `Healthy` vs `Suspect`

Figure 5.1: Summary of results from the PAPILA dataset (continued).

their relevance maps are substantially different. Indeed, as there is no disease, it is not surprising that no specific area is highlighted. The fundamental limitation of these relevance map is that even if the explanation is very concentrated on the optic disc, it would be of very limited utility for the doctor as it just brings an already known information (you should look at the optic disc) but relevance maps do not focus on specific enough area of this optic disc.

Despite these difficulties, spatial variance remain an informative element as it is almost always concentrated in the center of the image, near the optic disc. Similarly to what I observed with the emotions, the variance of the relevance tends to be mainly concentrated on the important areas of the image. But, again, explanations appears to be very sensitive to the training randomness.

Some hypotheses can explain the limited results observed for PAPILA. Firstly, LRP may not be suitable for medical imaging, as this method appears to obtain convincing results on ImageNet-21k (Chefer, Gur, and Wolf 2021) but its generalization capabilities remain unclear. Secondly, the task may be too subtle as ViT-B/16 using 14×14 patches might miss important markers, coupled with the high resolution of the original medical images (2576×1934 pixels). While the resizing to a 224x224 resolution is mandatory to use these images as inputs for the ViT-B/16, this is clearly limiting the ability for the model to learn details of the structure of the retina. Model prediction might be driven by artifacts like luminosity, global texture of the eye

rather than features related to the pathology. Finally, the size of the dataset may be too small (N=488), preventing the models to learn robust enough representations.

Overall, what I learned is that relevance and variability of explanations are first and foremost highly dependent on the content and characteristics of the dataset used, rather than its subjective nature.

5.2 Task Difficulty

QB

Is the stability of explanations influenced by the difficulty of the classification task?

With a given dataset and training regime, my intuition suggests that not all binary classification tasks are equally difficult. For FER+, distinguishing `happiness` against `sadness` should be easier than `neutral` against `sadness`.

I formulate the hypothesis for this section that a task considered as easy by a quantitative criterion would lead to better stability, as having very distinguishable face markers (like a smile to distinguish `happiness` from sadness) to differentiate two classes would naturally lead to similar explanations across models.
Therefore, for FER+, `happiness` vs `sadness` and `happiness` vs `neutral` should show the highest MCWME.

However, several challenges can appear when trying to quantifying this intuition. Indeed, as explained in Chapter 3, even if I decide to use the accuracy as my objective criterion to measure the difficulty of a task (which is a choice debated in Chapter 3), restricting myself to the compatible inputs leads to an accuracy of 100 %. Another choice available is the confidence in the prediction. However, this metric carries limitations. Indeed, this metric do not always reflect the true correctness likelihood on modern architectures (C. Guo et al. 2017). Even when assuming that this metrics indeed reflects the difficulty of a task, and assuming that I observe a difference between settings for these indicators, it would require statistical tests to observe whether the aforementioned differences are indeed statistically significant.

Moreover, the **Aggregated MCWME** (AMCWME) for a specific class in a specific setting, aggregating the MCWME on all compatible inputs belonging to this class, appears to be relevant since it summarizes in one scalar the stability of explanations of a given class in a given setting, and therefore directly gives me the

material to answer my research question. However, this metric carries issues because it lacks clear interpretation and requires formal statistical tests to evaluate if any potential difference observed is statistically significant.

Given these constraints, answering this question would require a more rigorous statistical study. My approach is therefore to compute, for each setting, the confidence in the prediction for one class, and the aggregated MCWME, without drawing any conclusion about the significance of the results obtained, and confronting them carefully with my expectations. I acknowledge the limitations of this approach, and this section *suggests* rather than it *concludes*.

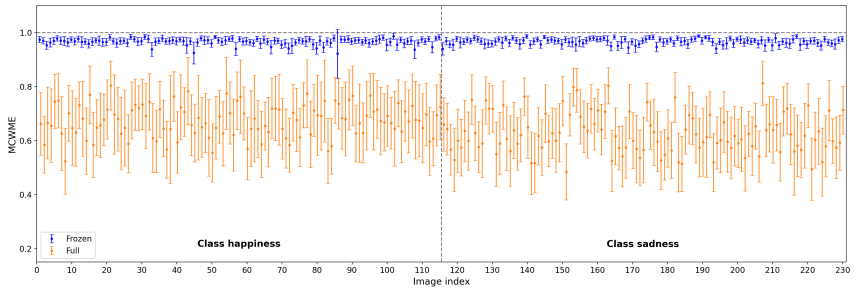
To answer this question, I am referring to Table 5.1 and Figure 5.2, a plot already used in § 4.1. The AMCWME for one class belonging to one setting is the mean of all 115 MCWME produced for all compatible inputs belonging to this class, for this setting.

Task (FER+)	Conf (C1)	Conf (C2)	AMCWME (C1)	AMCWME (C2)
Happiness Sadness (Frozen)	0.9597	0.9360	0.9669	0.9676
Happiness Sadness (Full)	0.9943	0.9909	0.6740	0.6255
Happiness Neutral (Frozen)	0.9420	0.9572	0.9786	0.9701
Happiness Neutral (Full)	0.9737	0.9866	0.7007	0.6000
Neutral Sadness (Frozen)	0.8556	0.7991	0.9295	0.9416
Neutral Sadness (Full)	0.9258	0.9087	0.5587	0.6381

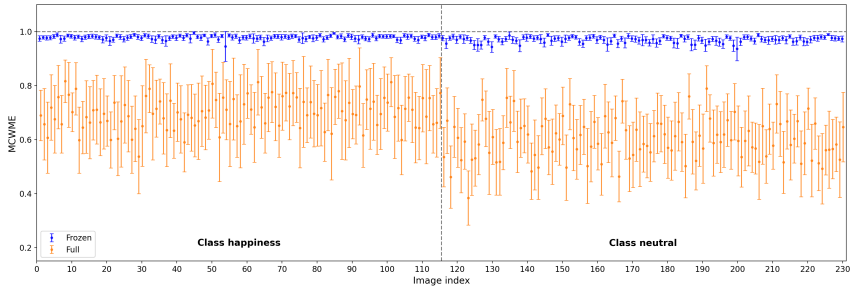
Table 5.1: Analysis of the confidence and AMCWME in each setting.

Following my intuition, `happiness` vs `sadness` appears to show high confidence score for the *Full FT* regime, `happiness` vs `neutral` showing a very high score as well. The `happiness` class show a higher AMCWME than the `sadness`[1], which is not surprising as we discovered in § 4.3 that explanations associated with `happiness` often focus the mouth only, whereas explanations of the `sadness` class appears to be more diverse. The `happiness` vs `neutral` task show similar values of confidence and AMCWME for both regimes to those observed for `happiness` vs `sadness`. Finally `neutral` vs `sadness` show a lower score of confidence, even if, in the *Frozen* regime, this score remains high. This setting also leads to the lowest AMCWME score (in the *Full FT* regime) for the `neutral` class.

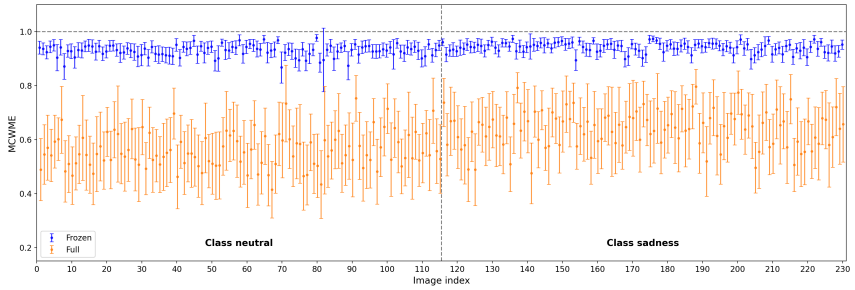
[1]This is also visible when looking at subfigure 5.2a.



(a) `happiness` vs `sadness`.



(b) `happiness` vs `neutral`.



(c) `neutral` vs `sadness`.

Figure 5.2: Comparison of the mean MCWME on compatible inputs for FER+. No overlap in the confidence intervals shows a statistically significant difference.

Overall, the results presented in this section are consistent with my hypotheses. The tasks featuring `happiness` as one of the two classes appear to show higher AMCWME, suggesting more stable explanation across this class. Moreover, in the `neutral` vs `sadness` setting, expected to be the most challenging (this seems to be confirmed with the confidence values), I observe the lowest AMCWME score in the *Full FT* regime, showing seemingly that when more subtle differences are at stake, explanations tend to vary more. However, these observations remain suggestive and would require a statistical framework along with a higher number of binary tasks for a future work.

Chapter 6

Conclusion

In this master thesis, I studied the stability of explanations for images through the prism of Vision Transformers and the LRP method.

In Chapter 2: Background, I defined the concepts I am using in my work: the architecture, datasets, as well as explainability methods and metrics, along with some taxonomy. Optimizing both *faithfulness* and *plausibility* is not always possible, but the aim of explainability is to develop methods that can be from different natures: intrinsic or post-hoc, local or global. In this work, the main focus is put on local post-hoc explainability methods, and among the family of propagation-based methods, the Layer-wise relevance Propagation, adapted for Vision Transformers, is the chosen method, as it is entirely deterministic. I used this method to study the Rashomon Effect (several models producing similar results, but different explanations), and to be able to quantify it, used several quantitative measures like the (top-k) MCWME and the spatial variance. This analysis is completed with qualitative analysis of relevance maps displaying the relevance.

In Chapter 3: Methodology, I first presented the main common pipeline, including a dataset preprocessing to adapt it adequately for my architecture, and my work, models generation and the selection of equivalent models, a central milestone to restrict myself to 100 models "equally good" (this notion being defined mathematically in Chapter 2): this is the functionally equivalent models set. Moreover, I complete this selection of a subset of models with a selection of a subset of images: the compatible inputs. This subset created from the test data is restricted to images correctly and unanimously predicted by the hundred models. After presenting this main pipeline, I guide the reader through my thoughts from the very beginning of my work, and the research questions I decided to explore in this master thesis.

In Chapter 4: Results, the hypotheses and results linked to each of these research questions are developed. Among the main results observed: there exists, similarly to what has been observed for the text, a significant difference in MCWME between *Full*

FT and *Frozen* regimes, showing a difference in stability between the explanations. Explanations obtained from the *Frozen* seem more stable, but less meaningful for a human. The contrastive class seems to have a real impact on which areas of the image are targeted preferentially, therefore impacting the stability of explanations. Finally, the patch-level analysis shows that models tends to agree on wide areas of the image, but disagree on the importance of smaller areas; the spatial variance acts like a meaningful measure of where the models disagree (productively).

Finally, in Chapter 5: Exploratory Analysis, I explored some research questions through a more informal point of view. The first one addresses whether the stability of explanations are influenced by the subjectivity level of the dataset. The second explores whether the difficulty of the task (using confidence in the prediction as a proxy) influences the stability of explanations.

The answers I obtained, or did not obtain, raises fundamental questions about the deployment of xAI in real-world applications. Two topics related to artificial intelligence that always interested me came immediately into my mind: artificial intelligence applied in a medical context, and the links between AI and law.

If we delve deeper into the fascinating topic of medical use of artificial intelligences, it is easy to figure out that the promising future of machine learning models replacing the doctors should be mitigated. In one of his courses at University of Berkeley (Recht 2023), Benjamin Recht uses cancer screening to illustrate many issues with the use of artificial intelligence in medicine.

The first main problem with machine learning models is that predicting is not sufficient; just being able to predict some outcome does not mean that you are able to solve it: prediction is not an intervention. Moreover, labels, particularly in medicine, are murky. This is something I experienced in the PAPILA dataset with the `Suspect` class: this class gathers retina images for which the doctors are not able to identify if these images show a retina belonging to the `Healthy` or `Glaucoma` class, illustrating an example of this gray area in the label attribution. Therefore, even though tasks related to the PAPILA datasets are probably more objective than those linked to FER+, it is important to keep in mind that correct and objective labeling of images is never guaranteed, especially for medical datasets. Furthermore, attributing the cause of the death to a disease is in many case rather an interpretation than an verifiable fact; deaths by cancer being a classic example of a causal attribution, in the majority of the cases. If the data by itself is subjective, the AI will reproduce and amplify this subjectivity. Another issue that can arise is that models are trained at at certain time T, with data available at this time T, meaning that a machine learning can become non usable due to a change in population ten or twenty years later.

More precisely, about the question of explainability in machine learing , I discovered that while explanations for artificial intelligence might look interesting, it also comes with major threats. Indeed, Babic et al. (2021) explored several of these risks: the overconfidence in the explanations is a major danger, exacerbated by the fact that the users are not always used to probabilistic predictions. A second issue is linked to the notion of *plausibility* that I explored in Chapter 2; this concept is crucial and can lead to biases for a human user. Indeed, doctors, like every user, like plausible stories that can confirm their diagnosis. Asking systematically for explanations can lead to a false sense of understanding. Furthermore, explainability lacks credibility because of the existence of the Rashomon Effect, that I explored in this thesis, as algorithms can produce very similar and possibly competing explanations not being necessarily justifiable or understood by experts.

If nonetheless we proceed to the implementation of xAI in a medical context, this should not be done without caution. As described by Tjoa and Guan (2021), failure is not an option in this context with human lives being at stake: there is therefore no error margin. The question of accountability is also crucial: along the ethical question of who between the doctor and the AI is responsible if things go wrong, there is also a major issue if someone takes advantage with malicious intents of this gray zone about accountability. Axes explored in this master thesis show multiple signs indicating that sensibility to randomness is a serious issue that can affect how trustworthy are the artificial intelligences, and therefore considerably affecting their accountability in these contexts.

In the legal context, two acts have a major importance in AI regulation in Europe: the General Data Protection Regulation (GDPR) (European Union 2016) and the Artificial intelligence Act (AI Act) (European Union 2024).

The first one is fundamental because, before it, Europe regulated under the 95/46/EC directive (European Union 1995), long before social networks and the rise of GAFAM companies. With the emergence of big data and scandals, like Cambridge Analytica, the European Union needed to modernize its laws; the objectives were to give citizens control on their personal data and create a unique regulation for Europe. In its articles 13, 14 and 15, the GDPR explains that the data controller is required to *provide meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.* This obviously resonates with one main topic of this thesis, the Rashomon Effect: it might be hard for an expert to provide these meaningful informations: what does a *meaningful* explanation really mean, and are these explanations accessible ? This master thesis suggests that for an explanation to be meaningful, it would require *a minima* for this explanation to be stable across functionally equivalent models: a criterion that

is not fulfilled, for now, by the LRP method applied to Vision Transformers.

If the GDPR regulates the data, the AI Act aims to handle the questions related to algorithms. With the fast deployment of artificial intelligence (e.g. ChatGPT), the European Union realized that these new algorithms had major risks for information and democracy. Instead of banning AI, the European Union chose an approach based on risk to regulate more or less artificial intelligences depending on the context and threats it creates. The link with explainability is clear: Article 13 mentions that *high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately.* This indeed excludes black-box models for application with high risks, and therefore promotes the use of intrinsically interpretable architectures.

The low number of datasets used in this master thesis is a clear limitation that do not allow me to draw general conclusions. Moreover, the small size of the PAPILA dataset is also an issue. The examination of other corpora could therefore be a noteworthy perspective for future work. It is also true that results observed in this master thesis remain highly dependent on the image considered. For certain images, relevance maps show results that can be surprising for a human (low plausibility) and, overall, explanations indeed appears to be very sensitive of the choice of the random seed, making therefore the drawing of general conclusions complicated.

Other ViT architectures (Swin, DeiT) might be another noteworthy axe to explore and may bring meaningful results as well. I only explored in this master thesis the LRP method. This choice was motivated by the need to rely on a deterministic explainability method to study the stability; however, when considering a larger scope than the stability of explanations, comparing both quantitative metrics and qualitative results with other explainability methods (SHAP, Grad-CAM, LIME,...) could be a research topic by itself, and still not deeply explored. Another extension of this work would be to compare this work with multi-class classification, in order to analyze the stability of explanations across these two contexts. Finally, the question of the high relevance given to several patches in the background of the image remains an interesting topic to explore, especially to better understand if this is a marker of unfaithfulness of the LRP method, or if the model simply learns unwanted artifacts during its training.

What I learned in this master thesis is that explanation stability is not guaranteed, even when using models that perform well on classification tasks simple and well explored in the literature. Implementing Explainable AI (xAI) in real-world application, especially in high-stakes domains like medicine, remains a major challenge with a high number of risks. But being able to explain a model's reasoning is a necessity to be able to comply with European laws aiming to protect our private life and data.

Investigating the limits of xAI in image classification tasks is my contribution to the daunting task of better knowing the reasoning of Artificial Intelligence.

References

Abnar, Samira and Willem Zuidema (May 31, 2020). *Quantifying Attention Flow in Transformers*. DOI: 10.48550/arXiv.2005.00928. arXiv: 2005.00928 [cs]. URL: http://arxiv.org/abs/2005.00928 (visited on 04/27/2026). Pre-published.

Agarwal, Chirag, Sree Harsha Tanneru, and Himabindu Lakkaraju (2024). "Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models". In.

Almazroa, Ahmed et al. (2015). "Optic Disc and Optic Cup Segmentation Methodologies for Glaucoma Image Detection: A Survey". In: *Journal of Ophthalmology* 2015, p. 180972. ISSN: 2090-004X. DOI: 10.1155/2015/180972. PMID: 26688751. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4673359/ (visited on 05/30/2026).

Babic, Boris et al. (July 16, 2021). "Beware Explanations from AI in Health Care". In: *Science* 373.6552, pp. 284–286. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abg1834. URL: https://www.science.org/doi/10.1126/science.abg1834 (visited on 05/21/2026).

Bach, Sebastian et al. (July 10, 2015). "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* 10.7. Ed. by Oscar Deniz Suarez, e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140. URL: https://dx.plos.org/10.1371/journal.pone.0130140 (visited on 04/27/2026).

Bansal, Naman, Chirag Agarwal, and Anh Nguyen (Apr. 13, 2020). *SAM: The Sensitivity of Attribution Methods to Hyperparameters*. DOI: 10.48550/arXiv.2003.08754. arXiv: 2003.08754 [cs.CV]. URL: http://arxiv.org/abs/2003.08754 (visited on 05/25/2026). Pre-published.

Barsoum, Emad et al. (Sept. 24, 2016). *Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution*. DOI: 10.48550/arXiv.1608.01041. arXiv: 1608.01041 [cs]. URL: http://arxiv.org/abs/1608.01041 (visited on 04/30/2026). Pre-published.

Bogaert (2026). "Explanation Sensitivity to the Training Randomness in Text Classification Transformers". UCLouvain. URL: https://perso.uclouvain.be/fstandae/PUBLIS4536/theses/JB2026.pdf (visited on 05/03/2026).

Bogaert, Jeremie, Antonin Descampe, and François-Xavier Standaert (2026). "Consolidating Explanation Stability Metrics". In: *Explainable Artificial Intelligence*. Ed. by Riccardo Guidotti, Ute Schmid, and Luca Longo. Vol. 2578. Cham: Springer Nature Switzerland, pp. 310–323. ISBN: 978-3-032-08326-5 978-3-032-08327-2. DOI: 10.1007/978-3-032-08327-2_15. URL: https://link.springer.com/10.1007/978-3-032-08327-2_15 (visited on 05/08/2026).

Breiman, Leo (2001). "Statistical Modeling: The Two Cultures". In: *Statistical Science* 16.3, pp. 199–215. JSTOR: 2676681. URL: http://www.jstor.org/stable/2676681.

Chefer, Hila, Shir Gur, and Lior Wolf (June 2021). "Transformer Interpretability Beyond Attention Visualization". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, pp. 782–791. ISBN: 978-1-6654-4509-2. DOI: 10.1109/CVPR46437.2021.00084. URL: https://ieeexplore.ieee.org/document/9577970/ (visited on 04/27/2026).

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT 2019. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/ (visited on 04/27/2026).

Dietterich, Thomas G. (Oct. 1, 1998). "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". In: *Neural Computation* 10.7, pp. 1895–1923. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976698300017197. URL: https://direct.mit.edu/neco/article/10/7/1895-1923/6224 (visited on 05/25/2026).

Doshi-Velez, Finale and Been Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. Version 2. DOI: 10.48550/ARXIV.1702.08608. URL: https://arxiv.org/abs/1702.08608 (visited on 05/30/2026). Pre-published.

Dosovitskiy, Alexey et al. (June 3, 2021). *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. DOI: 10.48550/arXiv.2010.11929. arXiv: 2010.11929 [cs]. URL: http://arxiv.org/abs/2010.11929 (visited on 04/27/2026). Pre-published.

Du, Mengnan, Ninghao Liu, and Xia Hu (May 19, 2019). *Techniques for Interpretable Machine Learning*. DOI: 10.48550/arXiv.1808.00033. arXiv: 1808.00033 [cs]. URL: http://arxiv.org/abs/1808.00033 (visited on 04/27/2026). Pre-published.

European Union (Oct. 24, 1995). *DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. URL: https://www.edps.europa.eu/sites/default/files/publication/dir_1995_46_en.pdf.

— (Apr. 27, 2016). *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (General Data Protection Regulation)*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

— (June 13, 2024). *REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (Artificial Intelligence Act)*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689.

Goodfellow, Ian J. et al. (July 1, 2013). *Challenges in Representation Learning: A Report on Three Machine Learning Contests*. DOI: 10.48550/arXiv.1307.0414. arXiv: 1307.0414 [stat]. URL: http://arxiv.org/abs/1307.0414 (visited on 04/30/2026). Pre-published.

Guo, Chuan et al. (Aug. 3, 2017). *On Calibration of Modern Neural Networks*. DOI: 10.48550/arXiv.1706.04599. arXiv: 1706.04599 [cs.LG]. URL: http://arxiv.org/abs/1706.04599 (visited on 05/30/2026). Pre-published.

Jacovi, Alon and Yoav Goldberg (July 2020). "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386/ (visited on 05/08/2026).

Kovalyk, Oleksandr et al. (June 9, 2022). "PAPILA: Dataset with Fundus Images and Clinical Data of Both Eyes of the Same Patient for Glaucoma Assessment". In: *Scientific Data* 9.1, p. 291. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01388-1. URL: https://www.nature.com/articles/s41597-022-01388-1 (visited on 05/01/2026).

LeCun, Yann et al. (1998). "Efficient BackProp". In: *Neural Networks: Tricks of the Trade*. Ed. by Genevieve B. Orr and Klaus-Robert Müller. Red. by Gerhard Goos, Juris Hartmanis, and Jan Van Leeuwen. Vol. 1524. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–50. ISBN: 978-3-540-65311-0 978-3-540-49430-0. DOI: 10.1007/3-540-49430-8_2. URL: http://link.springer.com/10.1007/3-540-49430-8_2 (visited on 05/25/2026).

Lehmann, E. L. and Joseph P. Romano (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer Texts in Statistics. New York: Springer. 784 pp. ISBN: 978-0-387-98864-1.

Liu, Ze et al. (Aug. 17, 2021). *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*. DOI: 10.48550/arXiv.2103.14030. arXiv: 2103.14030 [cs]. URL: http://arxiv.org/abs/2103.14030 (visited on 04/27/2026). Pre-published.

Loncour (2025). "Explainability and Machine Learning (for Text)". Mémoire de fin d'études. Louvain-la-Neuve: Université Catholique de Louvain.

Loshchilov, Ilya and Frank Hutter (Jan. 4, 2019). *Decoupled Weight Decay Regularization*. DOI: 10.48550/arXiv.1711.05101. arXiv: 1711.05101 [cs.LG]. URL: http://arxiv.org/abs/1711.05101 (visited on 05/25/2026). Pre-published.

Lundberg, Scott and Su-In Lee (Nov. 25, 2017). *A Unified Approach to Interpreting Model Predictions*. DOI: 10.48550/arXiv.1705.07874. arXiv: 1705.07874 [cs]. URL: http://arxiv.org/abs/1705.07874 (visited on 04/27/2026). Pre-published.

Lyu, Qing, Marianna Apidianaki, and Chris Callison-Burch (2022). "Towards Faithful Model Explanation in NLP: A Survey". In: *Computational Linguistics* 50.2, pp. 657–723. ISSN: 0891-2017, 1530-9312. DOI: 10.1162/coli_a_00511. URL: https://direct.mit.edu/coli/article/50/2/657/119158/Towards-Faithful-Model-Explanation-in-NLP-A-Survey (visited on 05/08/2026).

Molnar, Christoph (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Second edition. Munich, Germany: Christoph Molnar. 1 p. ISBN: 979-8-4114-6333-0.

Montavon, Grégoire et al. (2019). "Layer-Wise Relevance Propagation: An Overview". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Vol. 11700. Cham: Springer International Publishing, pp. 193–209. ISBN: 978-3-030-28953-9 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6_10. URL: http://link.springer.com/10.1007/978-3-030-28954-6_10 (visited on 05/30/2026).

Müller, Sebastian et al. (2023). "An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning". In: *Machine Learning and Knowledge Discovery in Databases: Research Track*. Ed. by Danai Koutra et al. Vol. 14171. Cham: Springer Nature Switzerland, pp. 462–478. ISBN: 978-3-031-43417-4 978-3-031-43418-1. DOI: 10.1007/978-3-031-43418-1_28. URL: https://link.springer.com/10.1007/978-3-031-43418-1_28 (visited on 05/08/2026).

Murdoch, W. James et al. (Oct. 29, 2019). "Definitions, Methods, and Applications in Interpretable Machine Learning". In: *Proceedings of the National Academy of Sciences* 116.44, pp. 22071–22080. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1900654116. URL: https://pnas.org/doi/full/10.1073/pnas.1900654116 (visited on 05/22/2026).

Recht, Ben (May 9, 2023). *Cancer Screening: A Parable of Prediction*. University of California, Berkeley.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 9, 2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. DOI: 10.48550/arXiv.1602.04938. arXiv: 1602.04938 [cs]. URL: http://arxiv.org/abs/1602.04938 (visited on 04/27/2026). Pre-published.

Selvaraju, Ramprasaath R. et al. (Oct. 2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, pp. 618–626. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.74. URL: http://ieeexplore.ieee.org/document/8237336/ (visited on 05/30/2026).

Silver, N. Clayton and William P. Dunlap (Feb. 1987). "Averaging Correlation Coefficients: Should Fisher's z Transformation Be Used?" In: *Journal of Applied Psychology* 72.1, pp. 146–148. ISSN: 1939-1854, 0021-9010. DOI: 10.1037/0021-9010.72.1.146. URL: https://doi.apa.org/doi/10.1037/0021-9010.72.1.146 (visited on 05/08/2026).

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (Apr. 19, 2014). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. DOI: 10.48550/arXiv.1312.6034. arXiv: 1312.6034 [cs]. URL: http://arxiv.org/abs/1312.6034 (visited on 04/27/2026). Pre-published.

Smilkov, Daniel et al. (June 12, 2017). *SmoothGrad: Removing Noise by Adding Noise*. DOI: 10.48550/arXiv.1706.03825. arXiv: 1706.03825 [cs]. URL: http://arxiv.org/abs/1706.03825 (visited on 04/27/2026). Pre-published.

Srivastava, Nitish et al. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research*.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (June 13, 2017). *Axiomatic Attribution for Deep Networks*. DOI: 10.48550/arXiv.1703.01365. arXiv: 1703.01365 [cs]. URL: http://arxiv.org/abs/1703.01365 (visited on 04/27/2026). Pre-published.

Tjoa, Erico and Cuntai Guan (Nov. 2021). "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11, pp. 4793–4813. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2020.3027314. URL: https://ieeexplore.ieee.org/document/9233366/ (visited on 05/21/2026).

Touvron, Hugo et al. (Jan. 15, 2021). *Training Data-Efficient Image Transformers & Distillation through Attention*. DOI: 10.48550/arXiv.2012.12877. arXiv: 2012.12877 [cs]. URL: http://arxiv.org/abs/2012.12877 (visited on 04/27/2026). Pre-published.

Vaswani, Ashish et al. (Aug. 2, 2017). *Attention Is All You Need*. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 04/27/2026). Pre-published.

Wiegreffe, Sarah and Yuval Pinter (Sept. 5, 2019). *Attention Is Not Not Explanation*. DOI: 10.48550/arXiv.1908.04626. arXiv: 1908.04626 [cs.CL]. URL: http://arxiv.org/abs/1908.04626 (visited on 05/25/2026). Pre-published.

Xue, Fanglei, Qiangchang Wang, and Guodong Guo (Aug. 25, 2021). *TransFER: Learning Relation-aware Facial Expression Representations with Transformers*. DOI: 10.48550/arXiv.2108.11116. arXiv: 2108.11116 [cs.CV]. URL: http://arxiv.org/abs/2108.11116 (visited on 05/25/2026). Pre-published.

Yeh, Chih-Kuan et al. (2019). "On the (In)Fidelity and Sensitivity of Explanations". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html (visited on 05/25/2026).

Appendix

Use of Artificial Intelligence

During coding, I used the Github Copilot tool for auto-completion.
During the writing process, Claude was used to check grammar, formatting tables, figures & algorithms, and some sentences reformulations.
All scientific content, analyses and conclusions are my own work.

Codes

Algorithm 1: Training Loop

Input: `train_loader` (training set), `val_loader` (validation set), `E` (max. number of epochs), `P` (patience)
Output: `weights.pth` (weights of the network)

```
best ← 0.0 ;
counter ← 0 ;
for epoch ← 1 to E do
    // Training
    model.train()
    foreach batch (x, y) ∈ train_loader do
        optimizer.zero_grad()
        ŷ ← model(x)                          // Forward pass
        ℒ ← criterion(ŷ, y)
        scaler.scale(ℒ).backward()            // Backward pass
        scaler.step(optimizer)
        scaler.update()
    // Validation
    model.eval()
    current ← evaluate(model, val_loader)
    // Early stop
    if current > best then
        best ← current
        Save (new) weights.pth
        counter ← 0
    else
        counter ← counter + 1
    if counter ≥ P then
        break                                 // Early stopping
```

Algorithm 2: Equivalent Models Extraction

Input: `list_of_models`, `n_test`, $Z_{thresh} = 1.96$
Output: `equivalent_subset`

```
// Sorting by accuracy
N_models ← len(list_of_models)
list_of_models.sort(key=lambda x:  x[acc], reverse=True)
equivalent_subset← [ ]
for  i = 1 to N_models do
    subset ← list_of_models[0:i]
    a ← subset[0][acc]
    b ← subset[−1][acc]
    // Z-statistic computation
    p ← (a + b)/2
    z ← |a − b|/√((p × (1 − p))/n_test)
    if z ≤ Z_thresh then
        equivalent_subset ←subset
    else
        break
return equivalent_subset[0:100]
```

Algorithm 3: Compatible Inputs

Input: `predictions`, `model_paths`, `true_labels`, `test_set`
Output: `compatible_images`

```
compatible_inputs ← [ ]
n_test ← len(test_set)
for i = 0 to n_test - 1 do
    image ← test_set[i]
    preds ← [predictions[p][i] for p in model_paths]
    label ← true_labels[i]
    // Check if the models are unanimous and predict correctly
    if len(set(preds)) == 1 and preds[0] == label then
        compatible_inputs.append(image)
return compatible_inputs
```

Algorithm 4: MCWME

Input: $\{R_i(x)\}_{i=1}^{N}$ (relevance scores for all models), x (compatible input)
Output: `MCWME`(x) with the 95% confidence interval

`// LOO mean explanations`
for $i \leftarrow 1$ **to** N **do**
  $\bar{R}_{-i}(x) \leftarrow \frac{1}{N-1} \sum_{j \neq i}^{N} R_j(x)$

`// Pearson correlation coefficient`
for $i \leftarrow 1$ **to** N **do**
  $\hat{\rho}_i \leftarrow \rho_P \left(R_i(x),\, \bar{R}_{-i}(x) \right)$

`// Fisher Z transform`
for $i \leftarrow 1$ **to** N **do**
  $Z_i \leftarrow \operatorname{arctanh}(\hat{\rho}_i)$

`// Mean and variance`
$\bar{Z} \leftarrow \frac{1}{N} \sum_{i=1}^{N} Z_i$
$\hat{\sigma}_Z^2 \leftarrow \frac{1}{N-1} \sum_{i=1}^{N} (Z_i - \bar{Z})^2$

`// Inverse transform`
`MCWME`$(x) \leftarrow \tanh(\bar{Z})$
`conf_interval`$_{95\%} \leftarrow \left[\tanh\left(\bar{Z} - 1.96 \frac{\hat{\sigma}_Z}{\sqrt{N}} \right),\ \tanh\left(\bar{Z} + 1.96 \frac{\hat{\sigma}_Z}{\sqrt{N}} \right) \right]$
return `MCWME` *(x),* `conf_interval`$_{95\%}$

Additional Figures

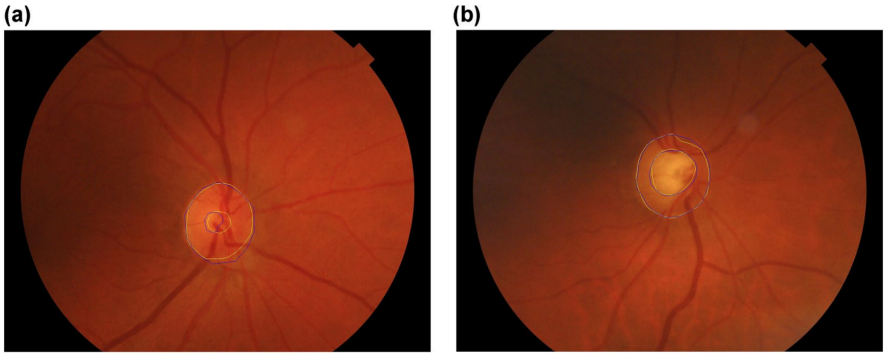


Figure 6.1: Segmentations of optic disc and optic cup performed by two ophthalmologists. (a) Right eye of Patient 73, (b) Right eye of Patient 20. Image from Kovalyk et al. (2022).

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**