POLITECNICO DI TORINO

Master Degree course in Biomedical Engineering

Master Degree Thesis

Integration of Uncertainty into Explainability Methods to Enhance AI Transparency in Brain MRI Classification

Supervisors
Prof. Massimo Salvi
PhD. Silvia Seoni

Candidate
Letizia Quattrocchio

Academic Year 2023-2024

“Almeno un punto”

Acknowledgements

Vorrei ringraziare il professore Massimo Salvi e l'ingegner Silvia Seoni per avermi permesso di partecipare a questo entusiasmante lavoro di Tesi Magistrale. Mi é stato offerto tutto il supporto necessario per affrontare questo progetto, e sono stata anche guidata nel vasto mondo dell'analisi delle immagini mediche utilizzando l'intelligenza artificiale. Massimo Salvi e Silvia Seoni sono stati sempre disponibili e pazienti durante questo percorso che é durato quasi un anno condividendo le loro conoscenze e dandomi la possibilitá di navigare da sola sviluppando nuovi metodi per l'analisi delle immagini mediche, apportando il mio personale contributo alla letteratura. Senza di loro questo lavoro non sarebbe stato possibile.

Un ringraziamento speciale va alla mia famiglia. Sono tanti i cugini, gli zii e anche i prozii che mi hanno supportato in tutti questi anni.

I ringraziamenti piú cari li faccio a mia mamma Antonella, a mio papá Fulvio, a mia nonna Carla e a mio nonno Francesco. Nessuno della mia famiglia ha mai studiato ingegneria, e ho sorpreso un po' tutti quando ho deciso di iniziare questo percorso, ma non é mai mancato il loro supporto, soprattutto quello morale che é tanto servito in questi anni. I miei nonni hanno seguito le mie avventure giornalmente raccontate alle 19:00, ormai questo é l'orario della storia della mia vita a Torino. Hanno seguito tutto, dalla vita universitaria alle avventure di tutti i giorni, anche se alcune volte si dimenticano dei particolari, ma non importa. Penso di essere la persona piú fortunata del mondo, i miei nonni mi vedranno laurearmi per la seconda volta. Avranno modo di vantarsi con tutti i loro amici.

Grazie alla mia mamma e al mio papá. Non mi hanno mai messo pressione in alcun modo standomi sempre vicino nel bello e nel cattivo tempo. Sono stati dei veri eroi venendo a trovarmi quando non tornavo a casa per tempi molto lunghi. Mi hanno permesso di trasferirmi a Torino dandomi la possibilitá di farmi una vita tutta mia. All'inizio mi hanno aiutato molto con i traslochi quando abitavo al campus, e si sono trasformati in tutto-fare quando mi sono trasferita in appartamento. La fine di un'era é stata segnata quando da sola ho completato il mio ultimo trasloco da studente. Ormai sono entrata ufficialmente nel mondo degli adulti.

Ringrazio anche gli amici del liceo, pochi ma buoni, con cui ho continuato a sentirmi e che ho sempre piacere di ritrovare quando torno a casa.

Vorrei ringraziare anche gli amici che ho incontrato durante tutto il mio percorso: dai "ragazzi del campus" con cui ho convissuto, ai "ragazzi del gruppo Skype" conosciuti

in periodo di lockdown, passando alle ultime amicizie strette durante questo percorso magistrale, tra vari laboratori, progetti, e lezioni impegnative. E' anche grazie a loro che sono diventata la persona che sono adesso.

Dicono che le amicizie per la vita si facciano a 16 anni, e se fosse davvero vero io sarei messa molto male. Non lo dico con cattiveria, ma ho molta difficoltá a stringere amicizie. Gli ultimi ringraziamenti vorrei farli a coloro che sono diventati la mia nuova famiglia. Siamo venti e piú persone che hanno stretto un'improbabile amicizia durante la sessione estiva 2023 diventando subito molto affiatati dopo un paio di merende con il gelato. Chiunque sia sia seduto attorno al tavolo alle coordinate 45° 4' 9.17" N
7° 39' 22.7" E, su cui sono state versate lacrime di gioia, ma anche di dolore, ha il mio ringraziamento finale. Siete le persone che mi hanno salvato la vita nel mio ultimo periodo. Sono infinitamente grata che siate diventati i miei amici piú cari, e che a distanza di un anno ancora mi sopportiate. Siete coloro che in solo una sessione estiva, scusate se utilizzo le sessioni d'esame come metro per indicare alcuni mesi dell'anno, siete riusciti a tirar fuori da me il lato che sembravo aver dimenticato da qualche parte nei miei ricordi piú lontani. Le coordinate del tavolo le ho tatuate assieme ad alcuni di voi, e mi ricorderanno sempre di questo incredibile e molto improbabile incontro. Vorrei lasciarvi con la frase simbolo che ha segnato le nostre ultime avventure:

a tutti voi auguro nuovi percorsi da affrontare " . . . almeno un punto" alla volta.

Abstract

The performance achieved by artificial intelligence tools is so remarkable that they can be applied in various fields of research and development. One of the most rapidly advancing areas is deep learning, which leverages neural networks - algorithms designed to mimic the behavior of the human brain. Neural networks' ability to analyze data in different forms, such as signals, numerical data, and images, has greatly expanded scientific research, particularly in the field of medicine.

Despite the wide range of applications and the impressive results they can achieve, deep learning networks are often referred to as "black boxes". The data processing and the specific features used to make decisions are largely hidden or not transparent to humans. This opacity has implications for the results produced by the network, leading to difficulties in interpretation, unpredictability, and reduced reliability. Over the years, explainability techniques, also known as Explainable Artificial Intelligence, have been developed to make neural networks more transparent by providing understandable explanations. In parallel, methods have been developed to assess the uncertainty in a network's predictions, quantifying the degree of trust one can place in the model. Explainability and uncertainty are two aspects that have not yet been combined in the study of neural networks.

In this thesis, a series of methods have been developed to investigate the variability and uncertainty of the most important features used by the neural network to make decisions. The neural network used is the Cross-Covariance Image Transformer, tasked with classifying a set of brain MRI images into four distinct classes: no tumor, pituitary tumor, meningioma tumor, and glioma tumor. The explainability methods applied are Grad-CAM and Score-CAM, two visual techniques that provide heatmaps highlighting which parts of an image were crucial for classification. Monte Carlo Dropout is the method used to estimate uncertainty by randomly disabling some neurons during inference, thus generating multiple predictions for each input image. This method was applied with two different dropout probabilities. The first objective of this thesis was to identify which of the devised methods performs best, while the second was to characterize the overall behavior of the neural network. By using the developed method, it was also possible to determine which of the two explainability techniques and which dropout probability generated more robust heatmaps and less uncertain predictions.

Although the proposed method has its limitations, the applications of this work can be extended to any neural network and any image classification task. It provides not only clarity in the heatmaps and a quantification of prediction reliability but also an indication of whether the implemented explainability methods are compatible with the neural network in use and which dropout probability should be associated with it.

Abstract

Le prestazioni che gli strumenti dell’intelligenza artificiale raggiungono sono cosí sorprendenti che questi possono essere applicati in vari settori sia di ricerca che di sviluppo. Uno dei campi che si sta sviluppando di piú é rappresentato dal deep learning che fa uso di reti neurali, algoritmi che tentano di emulare il comportamento del cervello umano. La capacitá di analizzare dati che si presentano in varie forme, come segnali, dati numerici, e anche immagini, ha permesso alle reti neurali di ampliare la ricerca scientifica soprattutto nel campo della medicina.

Nonostante la varietá di applicazioni e i risultati che riescono ad ottenere, le reti di deep learning sono anche definite "scatole nere". L’elaborazione dei dati e le caratteristiche specifiche utilizzate per prendere una decisione, infatti, sono in gran parte nascosti o non trasparenti agli esseri umani. Questo comportamento ha delle conseguenze sui risultati che la rete propone provocando difficoltá di interpretazione, imprevedibilitá e poca affidabilitá. Negli anni si sono sviluppate delle tecniche di spiegabilitá, anche denominate Explainable Artificial Intelligence, che cercano di rendere piú trasparente il funzionamento delle reti neurali fornendo spiegazioni comprensibili. In parallelo sono stati sviluppati anche metodi che indagano l'incertezza nelle predizioni che la rete fornisce, quantificando quanto é possibile fidarsi del modello. La spiegabilitá e l’incertezza sono due aspetti che non sono mai stati combinati nello studio delle reti neurali.

In questo lavoro di tesi sono stai ideati una serie di metodi che permettono di indagare la variabilitá e l’incertezza delle caratteristiche piú importanti utilizzate dalla rete neurale per prendere una decisione. La rete neurale utilizzata é la Cross-Covariance Image Transformer che ha il compito di classificare una serie di immagini di risonanza magnetica del cervello in quattro classi distinte: no tumore, tumore ipofisario, meningioma e glioma. I metodi di spiegabilitá applicati sono Grad-CAM e Score-CAM, due metodi visivi che forniscono delle mappe chiamate "heatmap" in cui si evidenziano quali parti di un'immagine sono state determinanti per la classificazione. Monte Carlo Dropout é invece il metodo impiegato per stimare l’incertezza che, nella fase di inferenza, spegne casualmente alcuni neuroni della rete ottenendo varie previsioni per ogni immagine fornita come input. Il metodo é stato applicato utilizzando due probabilitá di dropout diverse. Il primo obiettivo di questo lavoro é stato individuare quale tra i metodi ideati possiede le prestazioni migliori; il secondo obiettivo é stato caratterizzare il comportamento complessivo della rete neurale. Sfruttando il metodo ideato é stato anche possibile individuare quale tra i due metodi di spiegabilitá utilizzati e quale probabilitá di dropout forniscono heatmap piú robuste e predizioni meno incerte. Nonostante il metodo ideato possieda delle limitazioni, le applicazioni di questo lavoro si possono estendere a qualsiasi rete neurale e a qualsiasi immagine nell'ambito della classificazione ottenendo non solo chiarezza nelle heatmap e quantificazione dell'affidabilitá delle predizioni, ma anche un'indicazione se i metodi di spiegabilitá implementati sono compatibili con la rete neurale utilizzata e quale probabilitá di dropout associarci.

Contents

Chapter 1

Introduction

Artificial Intelligence (AI) is a broad field of computer science dedicated to building smart machines capable of performing tasks that typically require human intelligence. These tasks include, but are not limited to, learning, reasoning, problem-solving, perception, language understanding, and decision-making. AI aims to simulate human cognitive processes, enabling machines to process information, recognize patterns, and make decisions based on data. AI encompasses a wide range of subfields, including Machine Learning (ML), natural language processing, robotics, and computer vision. The ultimate goal of AI is to create systems that can function autonomously and intelligently in a variety of environments, offering potential applications across numerous domains such as healthcare, finance, autonomous vehicles, and more.

Deep Learning (DL) is a subset of machine learning, which itself is a branch of AI. DL is inspired by the structure and function of the human brain, specifically the neural networks. It involves training artificial neural networks, which are composed of layers of nodes, on large amounts of data to recognize patterns and make predictions. DL models, often referred to as deep neural networks, are designed to automatically extract and learn features from raw data. Unlike traditional ML, where features are manually crafted, DL models can identify and learn hierarchical representations of data through multiple layers of abstraction. This capability allows deep learning to excel in complex tasks such as image and speech recognition, natural language processing, and game playing. The term "deep" refers to the number of layers in the neural network. A deep neural network has many layers between the input and output layers, allowing it to learn more intricate patterns. These networks are typically trained using large datasets and significant computational power, often with specialized hardware such as GPUs.

DL has led to remarkable advancements in AI, powering applications like virtual assistants, facial recognition systems, autonomous vehicles, and more. Its ability to automatically discover the intricate structures in high-dimensional data makes it a crucial technology in the development of intelligent systems. DL has demonstrated exceptional performance in the realms of image processing and data analysis, surpassing traditional methods in many challenging tasks [1]. The ability of DL models to automatically learn and extract intricate features from large datasets has revolutionized how images and complex data are analyzed. This proficiency stems from the deep neural networks'ability to capture and model the complex relationships within the data, making them particularly effective in tasks like image classification, object detection, and segmentation. In image processing, DL models, particularly Convolutional Neural Networks (CNNs), have achieved state-of-the-art results. These models excel at identifying patterns, textures, and structures in images, making them ideal for applications such as facial recognition, autonomous driving, and, crucially, medical imaging.

The high accuracy and robustness of DL in image analysis have made it a transformative tool in the medical field. Medical imaging, which includes techniques such as MRI, CT scans, X-rays, and ultrasounds, is critical for diagnosing and monitoring a wide range of conditions. Traditionally, the interpretation of medical images relied heavily on the expertise of radiologists and clinicians. However, DL models have significantly enhanced this process by providing automated, accurate, and fast analysis of medical images. These models can detect subtle abnormalities, classify diseases, and even predict patient outcomes with remarkable precision. For example, DL algorithms have been developed to detect early signs of cancer in mammograms [1] [2], identify diabetic retinopathy in retinal images, determine the type of skin cancer from histopathological images [3], diagnose pneumonia from chest X-rays, and even diagnose Alzheimer's disease through the analysis of MRI scans [4]. The use of DL in medical imaging has led to improved diagnostic accuracy, reduced human error, and faster processing times, ultimately leading to better patient outcomes.

Moreover, DL is not limited to image analysis but also extends to other forms of data analysis in the medical field. For instance, DL models are used to analyze genomic data, predict disease progression, and assist in drug discovery by identifying potential compounds for treatment. The ability of these models to handle and make sense of vast and complex datasets is driving innovation in personalized medicine, where treatments and interventions are tailored to individual patients based on their unique data profiles.

In summary, the high performance of DL in image processing and data analysis has established it as a powerful tool in the medical field, enabling more accurate, efficient, and personalized healthcare.

Despite the impressive performance of DL models in medical applications, their trustworthiness remains a significant concern, particularly due to their black-box nature [5]. Before deploying these models in clinical settings, it's crucial to rigorously evaluate their reliability. Understanding the reasons behind a model's predictions is essential for building trust, especially when decisions with high stakes, such as medical diagnoses, are involved [6]. The opacity of AI decision-making processes has led to debates about the necessity of explainability in high-stakes scenarios like healthcare [7]. While explainable AI is often advocated as a solution to foster trust, transparency, and bias mitigation, there are doubts about whether current explainability methods can truly meet these expectations, particularly in providing clear insights or accounting for the uncertainty inherent in the model's predictions [8].

1.1 State of the art

This section reviews the literature by separately characterizing neural networks, explainability methods, and uncertainty identification techniques. These are the three main pillars on which this Master's Thesis is based. The literature analysis highlights what has already been discovered, developed, and applied, as well as the challenges that remain unresolved in the application of AI in the medical field.

1.1.1 Deep neural networks

The explainability of neural networks has become a prominent topic in recent literature. Many scholars are engaging with this issue, contributing their insights by proposing solutions and training networks that do not rely on data from the healthcare sector. As a result, the literature is replete with articles addressing this challenge. However, an intriguing observation emerges from the analysis of the literature: regardless of the specific domain discussed - whether it is human tumor recognition or the identification of specific animals within images - there does not appear to be a unified, global approach. Each study presents its own potential solution, claiming to achieve accuracy and precision superior to those previously reported in the literature, yet each study tends to stand alone. This has led to a proliferation of new networks in the literature, either created from scratch or derived from pre-existing networks that have been modified in their layers.

In the healthcare sector, AI has been extensively applied to a variety of tasks, including image classification, detection, segmentation, content-based image retrieval, image generation, and image enhancement, as well as the integration of image data with clinical reports. Image classification is employed not only for the categorization of medical exams but also for the identification of objects and lesions. Segmentation tasks encompass the delineation of organs, substructures, and lesions, as well as image registration. A wide range of anatomical regions have been explored, including the brain, eye, chest, digital pathology and microscopy, breast, cardiac, abdomen, musculoskeletal system, among others [9]. Only a selection of the neural networks presented in the literature will be discussed below: Entropy-based Elastic Ensemble [1], ResNet-18 [3], DenseNet201-Xception-SIE [2], DenseNet-121 [5], VGG19 [4] [10], Actionable Uncertainty Quantification Optimization [11], BayesNetCNN [12], BIRADS-Net [13], Collaborative Human-AI [13], AlexNet [14], VGG16 [14], ResNet [15], Multi-level Context and Uncertainty aware [16], DenseNet-169 [17] [10], MultiResUNet [10], Inception [10], Xception [10], ResNet-50 [10], sMRI-PatchNet [18], and ResNet-18 [19]. Some studies also incorporate attention gates into standard CNN models to solve localisations and classification tasks separately [20].

Given the variety of neural networks employed, the clear objective reported in the literature is to achieve accuracy rates surpassing those of all other studies, with values ranging around 86% [12], 90% [17], 94.47% [3], 94.64% [21], 97.12% [2], 98.11% [16], and ambitiously aiming towards the nearly unattainable absolute perfection that is so highly sought after in every field, specially the medicine one.

1.1.2 Explainability

Literature shows that the pursuit of ever-higher accuracy and performance often mirrors the quest for the best neural network. However, this approach may not be the optimal solution for studies conducted in the healthcare sector. Since it became evident that DL is a more powerful and necessary tool than its counterpart, ML, research in these techniques has expanded significantly. DL is preferred for tasks that require the analysis of large and complex datasets. For instance, in image and signal analysis, without focusing too much on the purpose of the analysis, the datasets necessary for conducting a study must include at least over 1,000 elements. DL is advantageous because it is considerably faster than human analysis, saving valuable time.

The benefits of DL do not end here: unlike ML, DL does not require operator-dependent preparation of the network. In ML, data must be analyzed to identify the most important features that will enable classification - if the ultimate goal is to classify a series of elements into well-defined groups with quantifiable performance that remains consistent over time. This process, known as feature extraction and subsequent feature selection, is no longer present in DL. To be precise, the operator-dependent process is no longer necessary; instead, the network itself identifies patterns in the data that will allow it to provide the results for which it was trained and tested.

The astonishing results produced by neural networks initially led to the belief that they were tools highly similar to the human brain. The basic unit of a neural network is called a neuron - a mathematical model designed to mimic the biological behavior of the brain cell - and it was thought that neural networks could reason similarly to the human brain. This sparked a more in-depth analysis of neural networks. The remarkable aspect of a neural network is its ability to "learn" on its own, seemingly leaving the operator with a reduced workload. The operator is then tasked with the "mere" job of analyzing the data and attempting to understand how the network arrived at a particular result rather than another - a task far from easy. Although the process of training a network, which involves iterative optimization of weights by minimizing a loss function, is well understood, it is not always clear what factor led the network to favor a specific response.

This is where explainability (XAI) comes into play-understanding how a neural network "thinks." Early findings in this field were quite shocking: neural networks, it turns out, understand nothing at all! It is important to clarify this point. Even though a neural network is left alone to navigate the vast sea of data provided to it, and despite being made up of elements named after the cells that allow humans to perform incredible reasoning, it interprets data exactly like a machine. In the medical field, a radiologist examining chest X-rays identifies structures such as bone tissue, muscle tissue, and pneumonia. However, a neural network does not identify these structures in the same way. For a neural network, it is important whether a group of pixels has the same shade of gray, whether adjacent pixels have different shades, whether the gradient of an image shows clusters of pixels, and so on-making the task of identifying the network's "reasoning" much more challenging.

The first studies on XAI highlighted precisely this aspect. The most well-known case involves distinguishing between photos containing huskies and those containing wolves [6].

Despite high performance, XAI techniques revealed that the network was not differentiating between the two canines by analyzing the shape of the muzzle or the color of the fur, as a human would, but rather by analyzing the background. Any object placed against a forest background was identified as a wolf, while a domestic background indicated a dog. This was the most striking case during the study of neural networks and led to a more in-depth investigation of techniques that allow for explaining neural network results.

The field dedicated to analyzing the explainability of neural networks is known as visual analytics [22], which focuses on representing data and models in a way that makes them interpretable. Visual approaches primarily concentrate on how visualization techniques are employed to represent data and architectures, analyze performance, and provide both local and global explanations. XAI methods are categorized by explanation level, implementation level, and model dependency. The explanation level indicates whether an explainability technique focuses on the entire model or on a single instance. The subcategories of the explanation level are the global level, which focuses on the explainability of the entire model, and the local level, which explains the decisions of a model by analyzing individual instances or subpopulations.

The implementation level [22] has two main subcategories: intrinsic and post-hoc. Intrinsic explanations are generated by the model itself, indicating how a prediction was made using the model's parameters, decision trees, and rule-based methods. Post-hoc explanations, on the other hand, reveal the internal workings and decision-making mechanisms of black-box models. Post-hoc explanations can be applied both to pre-trained models and to models after training is completed.

Model dependency [22] encompasses both model-specific and model-agnostic explainers. Model-specific XAI techniques are tailored to explain only a particular type of algorithm. Intrinsic explanations, serving as model-specific methods, are not universally applicable and require modification of the explanation mechanism when applied to different models. In contrast, model-agnostic explanations can be applied to any type of model and are independent of the model's architecture. Given that many model-agnostic explainers also offer post-hoc explanations, these methods are frequently utilized for their versatility.

Given the classification described above, here some of the most popular XAI techniques are presented. ANCHORS [22], Shapley Additive Explanations [22] [23] [24] [25] [21] [18], Gradient-weighted Class Activation Mapping (Grad-CAM) [22] [26] [24] [27] [25] [28] [29] [30] [10] [31] [32] [7], Saliency Maps [22], Integrated Gradients [22], Deep Learning Important FeaTures [22] [23] [24] [30], Class Activation Mapping [22] [24] [25] [33] are model agnostic local and post-hoc level explanations. Local Interpretable Model-Agnostic Explanations [6] [3] [22] [23] [24] [25] [21] and Layer-wise Relevance Propagation [22] [8] [24] [25] [34] [35] [7] like before, they also fall on the global explanation level. Bayesian Rule Lists [22], Generalized Additive Models [22], and Mean Decrease Impurity [22] are model specific global intrinsic level explanations, the latter is also local explanation level. Distillation tecnhnique [22] is a model agnostic global post-hoc explanation level.

Among all the methods presented, the approach that predominates in computer vision is local explanation using saliency methods. Three types of approaches are employed:

feature map weighting, backpropagation, and input image perturbation. In the first approach, we find methods such as Grad-CAM++ [24] [36], Score-CAM [24] [37], Ablation-CAM [24], LIFT-CAM [24], and aXiom-based Grad-CAM [12] [24]. Some backpropagation approaches include gradient map [24], guided backpropagation [24] [27] [25] [38] [39] [10] [31] [35] [32], integrated gradients [24] [27] [39] [32], SmoothGrad [24] [27] [39] [40] [7], and VarGrad [24]. The final group includes the method randomized input sampling for explanation [24] [41].

XAI methods do not end here. Some techniques have been further refined to reduce computational cost, achieve better performance, and provide results that are more interpretable by the operator, such as: Ablation-CAM [26], Deep SHAP [23] [25], guided Grad-CAM [27] [42] [30] [10] [31] [7], guided integrated gradients [27], HR-CAM [29], Kernel SHAP [23], Linear SHAP [23], SP-LIME [6], Low-Order SHAP [23], Max SHAP [23], NeuroXAI [27], Uncertain-CAM [43], vanilla gradient [27].

Given the vast number of techniques presented, one might think that the field of XAI is nearing its conclusion, but the ongoing refinement of these methods suggests otherwise. In reality, the issue mentioned in Section 1.1.1 reemerges: the literature is abundant with numerous techniques, some even designed for specific studies, all striving to provide results that satisfy operators, yet without channeling the research in a unified direction. There are studies attempting to identify the most effective techniques, or to cluster them in order to determine if certain methods are more suitable for specific studies. One thing is certain: it is essential to select the appropriate XAI method based on the specific neural network being used. Not all methods fit every network perfectly; some yield better results than others. However, as the trend of creating new networks continues, it is only natural that new explainability techniques will be developed accordingly.

1.1.3 Uncertainty

During the research and development of models aimed at uncovering the "reasoning" behind the responses provided by neural networks and making these "black-box" systems transparent to human operators, the exploration of uncertainty (UQ) in neural networks has emerged alongside this endeavor in recent years.

UQ can be categorized into two types: epistemic uncertainty [17] [44], which reflects the uncertainty in the model's parameters due to insufficient data for training, and can be reduced by increasing the amount of data available for analysis; and aleatoric uncertainty [17] [44], which describes the inherent noise in the data, arising from hidden variables or measurement errors, and cannot be reduced by simply acquiring more data.

What some researchers have realized in their attempts to explain neural networks is that, despite the network providing an output, this result is still subject to UQ. Here, two aspects of the same problem converge: a human operator constructs an artificial algorithm based on a mathematical model in an effort to emulate human reasoning; upon building and testing the algorithm, it is used to perform the tasks for which it was designed, producing an output or a series of outputs. Even if the output is correct, the operator still seeks to understand the steps that led to that result, as the algorithm, though designed by the operator, has effectively learned to identify a set of features and rules to generate a response autonomously.

Explainability serves as the bridge that renders the model interpretable to a human being, utilizing visual tools that are ideally suited for human interpretation. However, while these tools are highly favored, they are not sufficient to make a model fully interpretable. It remains necessary for a human operator to evaluate and interpret why a visual tool indicates a particular pixel region as crucial for the provided response [45]. Thus, explainability methods still require human interpretation.

Explainability methods alone, therefore, are insufficient in determining whether a network has autonomously learned optimal rules during the training phase. For this reason, some researchers have begun to calculate the UQ of neural network predictions. The study of UQ is relatively new and has sparked significant interest in the field. By quantifying the UQ of a single prediction, it is possible to obtain a measurable value that indicates the robustness of the network, thereby bypassing the need for human interpretation that is still required when using XAI methods.

Neural network uncertainty can be modeled through Bayesian analysis, where uncertainties are formalized as probability distributions over the model's parameters, in the case of epistemic uncertainty, or over the model's inputs, in the case of aleatoric uncertainty.

In the literature, the most commonly used metrics to quantify uncertainty include: Bhattacharyya coefficient [44] between distributions, Deep Ensemble [46], entropy [44], Ensemble Monte Carlo dropout [46], Monte Carlo Dropout [11] [31] [46], Shannon's entropy [47], and variance [44]. Depending on the study, one or more metrics are used to quantify UQ. When multiple metrics are employed, they are combined with the network's output to gather more information about the model and assess whether the model is overconfident.

Because UQ is estimated numerically, thresholds are frequently used across studies. These thresholds are set according to the data obtained during model training and are particularly useful in uncertainty research due to the nature of uncertainty itself: it measures the degree of disorder, and thus the amount of information, in a system. In a random process, a common event carries less information than a rare one. Given a random event, higher probabilities indicate lower information content. This explains why thresholds are an effective and convenient tool: any uncertainty below a fixed threshold indicates useful information [1] [46].

Since the study of UQ is relatively young, most research utilizes the metrics mentioned above, with only a few opting to employ alternative metrics for its calculation.

1.2 Objectives

The objective of this Master's Thesis is to combine two aspects - XAI and UQ - in the context of medical image classification tasks. Although the number of studies integrating these two areas is not extensive, research is increasingly moving in this direction.

The novelty of this work lies in the concept of globality, aiming to automate XAI and UQ to ultimately derive one or two numerical metrics that comprehensively indicate how confident a neural network is in its predictions and the degree of uncertainty it possesses. The confidence of a prediction can be translated into the robustness of the

XAI method used. Specifically, by testing the localization phase, which is present in any neural network, we can evaluate certainty. Given an input, if it is perturbed and the XAI method consistently identifies the same spatial region used to obtain the output, the network can be considered confident in its localization.

The UQ of a prediction can be understood as the model's uncertainty regarding a specific example. This involves testing the second phase that characterizes neural networks, following the localization phase: the interpretation phase. This phase is the most delicate and is highly sensitive to the type of data provided, particularly the number of examples used during training and validation. During training, the network autonomously analyzes the features and "creates" a set of rules to arrive at a given prediction. Incorrect rules lead to incorrect predictions, which is why neural networks undergo a tuning phase to adjust parameters or even modify the network's architecture to improve performance. Care must be taken to avoid the network "learning" overly specific rules, which can lead to overfitting. Overfitting occurs when a neural network becomes too precise in its predictions, making no errors. A network that experiences overfitting is said to be too specialized for the cases it has been trained on, resulting in significantly poorer performance during the testing phase, where new, unseen data is used to evaluate final performance. A neural network that cannot generalize and analyze real-world data is not a successful network.

UQ is tied to the rules the network "learns" and, consequently, to the prediction it outputs. A neural network's prediction in a classification task is not merely a binary response (0 or 1), where 0 indicates non-membership in one or more classes and 1 indicates membership in a single class. Instead, it is a probabilistic distribution. The highest probability in the distribution indicates the final output class of the network. A high probability of belonging to a class indicates confidence in the prediction, while a high probability that is still comparable to others in the distribution indicates uncertainty in the prediction.

The focus of this study on achieving globality with XAI and UQ is not limited to a single classification network. The research will test whether the metric or metrics that identify variability and uncertainty in predictions hold true for other neural networks as well. This will involve verifying whether the method can be applied to other tasks in the healthcare domain, providing the literature with the much-needed element to make these "black boxes" increasingly transparent.

Chapter 2

Materials and methods

2.1 Dataset

For this study, a publicly available MRI image dataset is employed [11] [21]. The dataset comprises 3,264 images in JPG format, encoded in RGB, which includes three channels despite the images displaying only grayscale tones typical of MRI scans. The images represent various orientations, including axial, sagittal, and coronal views. The dataset is divided into two primary sets: 394 images are designated for testing, while 2,870 images are allocated for training.

Each set includes identical annotations for the images, categorizing them into four classes: no tumor, pituitary tumor, meningioma tumor, and glioma tumor.

2.1.1 Data pre-processing

To train a neural network, it is necessary to further divide the training set into two additional sets, referred to as the training set and validation set. To create these two sets, the 2,870 images were divided such that 85% of the images were used for the actual training set, while the remaining 15% were used as the validation set. Following this process, the resulting sets are as follows: a training set containing 2,437 images, a validation set with 433 images, and the test set with 394 images.

However, the images in these sets do not all share the same dimensions. The image dimensions were analyzed, and the results are presented in Table 2.1.

2.1.2 Resize

It is essential to standardize the image dimensions before feeding them into the neural network. Another insight from the dimension analysis is that only 4% of the entire dataset contains at least one dimension, either row or column, exceeding 512 pixels. A 512×512 resolution is one of the most commonly used dimensions in image analysis. Given that the dimension analysis suggests 512×512 as the optimal size, the images undergo a resizing process accordingly.

The decision was made to process images with dimensions smaller than 512×512 using zero padding, which involves adding a black border that carries no information.

dim	%	Train	Validation	Test
x<200	0.18	1	1	4
200≤ x<300	17	308	42	198
700≤ x<800	0.25	4	3	1
x>800	1.6	27	6	18

Table 2.1. Analysis of image dimensions in pixels. The analysis focused solely on the number of rows and columns of the images. The percentage column (%) refers to the entire dataset, while the dimension (x) in column (dim) refers to at least one of the dimensions, either rows or columns.

For images with significantly larger dimensions, background cropping was first applied to check whether the brain fits within the desired dimensions. If not, the cropped image underwent resizing to meet the target size.

After applying zero padding to all images smaller than 512×512, the output images were analyzed. It is well known that neural networks perform best when images are centered, with good resolution and a moderate black border framing the content. Visually, the images that align with this description and offer a good trade-off between brain visibility and black background are those with dimensions of 380×380. An example of what has just been said is provided by the Fig. 2.1. Images smaller than this tend to have a brain that is too small, with an excessive amount of black background.

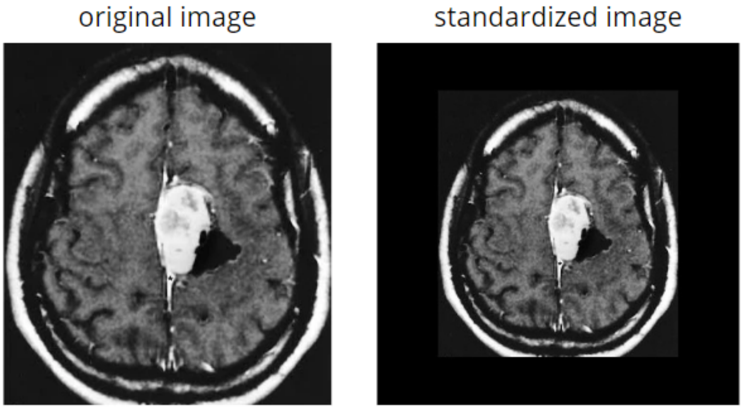

Figure 2.1. This is the image m1(117) labelled meningioma tumor. On the left there is the original image which has dimensions 341×377, on the right is the standardized image 512×512. This is the perfect example of how this orginal dimensions have a good trade-off between brain visibility and black background

Thus, it was decided to resize images with initial dimensions smaller than 380×380 to this optimal size, utilizing the zoom mode of the transformation. Zero padding was then applied to reach the desired dimensions of 512×512.

Care must be taken when applying resizing, as it alters the image resolution: whether zooming in or shrinking the image, each pixel's value is recalculated using different methods, resulting in a change in resolution. Therefore, it is crucial to carefully select which images undergo transformation and consider the size difference between the original and final images. By avoiding direct resizing from the original dimensions to 512×512, significant degradation of image resolution was prevented. The intermediate step of resizing to dimensions around 380×380, combined with the use of zero padding, is the most optimal solution.

Accordingly, the images subjected to this transformation are those with at least one dimension falling within the range indicated in Table 2.1, specifically x<200 and 200≤x<300. A more detailed analysis of image dimensions is provided in Table 2.2.

Set	min column	min row
Test	175	167
	174	195
	200	208
	201	202
Train	512	512
	200	207
	200	198
	201	202
Validation	512	512
	180	218
	201	217
	256	256

Table 2.2. Detailed analysis of image dimensions in pixels. For each set and for each image class, the minimum and maximum row and column dimensions are analyzed. The smallest dimensions within the range x<200 are highlighted in yellow, while those within the range 200≤ x<300 are shown in green.

The analysis of dimensions focuses exclusively on the minimum number of rows and columns for each set of each class. From Table 2.2, it can also be seen that most images do not have square dimensions. The goal is to increase both dimensions of the images by a certain percentage so that the final dimensions are as close as possible to 380×380. The dimensions calculated through this percentage are then used in the resizing function. The resized image subsequently undergoes zero padding to achieve a final output size of 512×512.

To determine the percentages to be applied to images within the x<200 range and those in the 200≤ x<300 range, the following approach is used: for both ranges, the smaller dimension between rows and columns is identified. From Table 2.2, the minimum dimension is highlighted in yellow for the x<200 range, and in green for the 200≤ x<300 range. This dimension is then used to calculate the percentage by which the image dimensions should be increased. The relationship used is shown by the equation (2.1).

$$
\begin{gathered}
100 : 380 = x : 167 \\
100 - x = 56\% \\
100 : 380 = x : 200 \\
100 - x = 47\%
\end{gathered}
\tag{2.1}
$$

Considering equation (2.1), the images will undergo a resize that will increase their dimensions by 56% and 47%, respectively, followed by zero padding to achieve the desired 512×512 dimensions. The Fig. 2.2 shows the output of an image that has undergone first reshape and then zero padding.

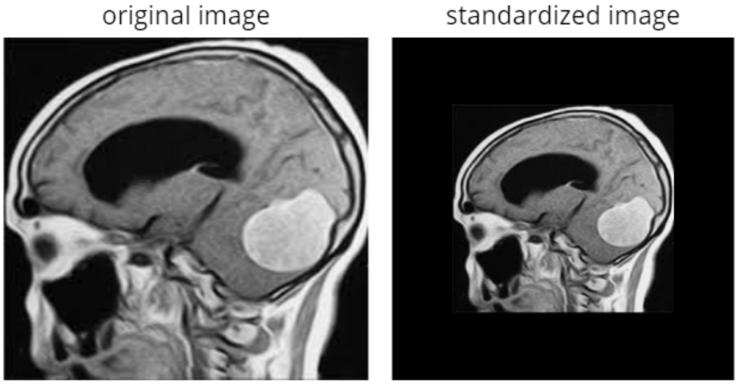

Figure 2.2. This is the image m1(11) labelled meningioma tumor. On the left there is the original image which has dimensions 226×212, on the right is the standardized image 512×512. This images has undergone first an encrease of its dimensions by a factor of 0.47, then zero padding was applied to get the standardized image.

Images with dimensions larger than 512×512 will be treated differently: the brain region will be identified in each image, and the background will be cropped to try to achieve the desired dimensions while avoiding excessive use of the resize function. If, after cropping the background, the image is still too large, a resize will be applied to produce a 512×512 image.

2.2 Neural network

The neural network most commonly used in the field of medicine, particularly for image analysis, is CNN. CNNs do not treat images as a flat vector of pixels but rather analyze their spatial structure by leveraging the fact that nearby pixels contain visually correlated information. The architecture of CNNs consists of:

- Convolutional layer: Filters, called kernels, are applied to the image to capture local patterns such as edges, corners, or textures. Each filter produces a feature map;
- Activation function: Each convolutional layer has a nonlinear activation function, which introduces nonlinearity into the model, enabling the network to learn complex patterns;
- Pooling: A process applied after convolution to reduce the size of the feature map by retaining only the most relevant information, thus reducing computational complexity;
- Fully connected layer: After a series of convolutions and pooling, the feature maps are transformed into a vector and passed through a series of fully connected layers similar to those in traditional neural networks. These layers are used to perform the final classification.

The final layer of the network is called the softmax layer, where the probabilities for each pixel belonging to a specific class are accumulated. This is also referred to as a probability map, and it is from this that the final softmax probability is obtained. The softmax probability is a vector that reports the likelihood of the image belonging to each possible class in the classification task. The sum of all probabilities is equal to 1, and the highest probability determines the final class chosen as the network's output. A schematic example of the architecture of a CNN is shown in Fig. 2.3.

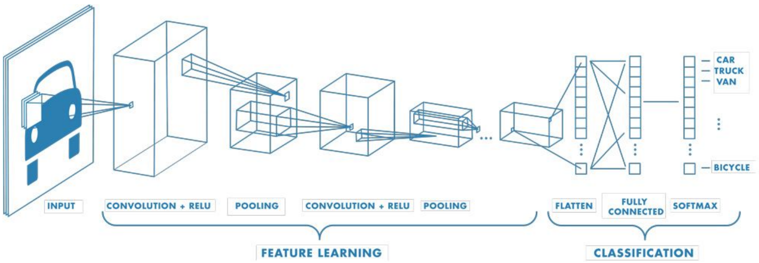

Figure 2.3. This is the most used representation of the architecture of a CNN for image classification. This scheme highlights each layer and their objective, and shows how the softmax is presented.

Deep Convolutional Neural Networks (DCNNs), some of which are mentioned in Section 1.1.1, are an extension of CNNs. They have many more layers compared to traditional CNNs, but they are based on the same principles and can extract increasingly complex features as data is processed. A DCNN contains the same layers as a CNN, but in a much larger number.

Some DCNNs and CNNs may have an architecture divided into two parts called encoder and decoder. The encoder is the first part of the network, with the objective of reducing the size and complexity of the input by extracting the most important features. The aforementioned layers - convolutional layers, activation function, and pooling - are used in sequence, and at the end of the encoder, a highly condensed and compact representation of the data is obtained. The decoder is the second part of the network and aims to reconstruct the original image starting from the representation provided by the encoder. The layers used in this part of the network include:

- Upsampling: Operations are applied to increase the dimensions of the compressed representation;
- Deconvolution: The resolution of the original image is restored by applying a series of filters;
- Decoder output: A reconstructed image or a mask representing the desired output is obtained.

The final layer of a DCNN, whether with or without a decoder, remains the softmax layer, from which the softmax probability and the final classification are obtained. Fig. 2.4 presents a schematic overview of a DCNN that includes both an encoder and a decoder in its structure.

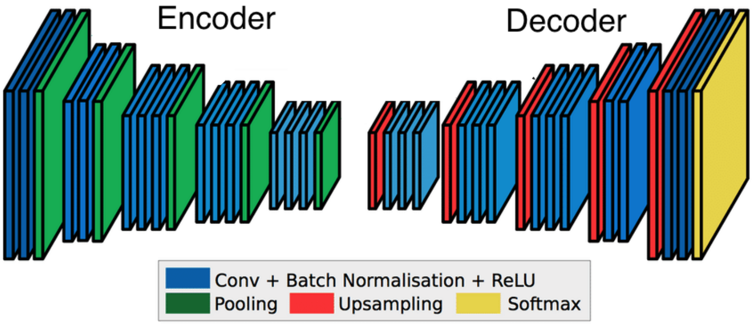

Figure 2.4. This is the most used representation of the architecture of a DCNN for image classification. This scheme highlights how the encoder and the decoder are structured, each layer and their objective. In the legend of this scheme the Batch Normalization is presented, some DCNN use this technique to improve the training speed and stability of deep networks. ReLu is one of the non-linearity activation functions that can be used to introduce non-linearity in the model. This scheme also represents very well the difference between a CNN and a DCNN. Despite they share the convolutional layer, the non-linearity function, and the pooling layer, a DCNN has a very high number of these layers. That is why these networks are called "deep".

The network used for this project is the Cross-Covariance Image Transformer (XCiT) [48]. XCiT consists of an encoder and a decoder:

- The encoder includes a series of layers based on the multi-head attention mechanism and a position-wise fully connected feed-forward network. Each layer is accompanied by a residual connection and normalization;

- The decoder has the same structure as the encoder and includes the same layers but with one additional layer. This layer, called multi-head attention, takes the output of the encoder as input. A residual connection and a normalization layer are also applied to this layer.

XCiT shares the same architecture as the Transformer network [49]. The difference between the two architectures lies in the multi-head attention layer. XCiT uses its own specific XCiT layer, which consists of three blocks, each preceded by a LayerNorm. The three blocks are: Cross-Covariance Attention, Local Patch Interaction, and Feed-Forward Network. The introduction of this block in place of the traditional multi-head attention layer reduces computational cost while achieving performance comparable to networks in the literature across various tasks, such as image classification, object detection, instance and semantic segmentation. This network can also be used as a backbone for self-supervised learning.

The parameters used to train XCiT are as follows:

- Input size: 512;
- Number of epochs: 50;
- Batch size: 8;
- Learning rate: 1e-05.

The network was pre-trained using ImageNet.

2.3 Monte Carlo Dropout

Monte Carlo Dropout (MCD) is a variant of Variational Inference [50]. Both approaches are used for estimating uncertainty in models. Since Variational Inference has a computational cost that is directly proportional to the size of the initial dataset, the MCD method was developed. To use MCD, dropout layers must be present in the neural network after each layer that contains network weights [51]. These dropout layers help prevent the network from overfitting. The method defines a function that approximates the distribution of weights for each layer in the neural network. This function uses a Bernoulli random variable to decide, based on a probability, whether a particular input should be dropped or not [50].

Given a dataset D, where X represents the records of all variable predictions and Y represents the predictions provided by the model, the objective of uncertainty estimation is to predict a new set y^* by providing new data x^*. The calculation of uncertainty is performed by knowing a model described by a series of layer weights or parameters W. The formulation of the uncertainty estimation problem is an optimization problem that seeks to obtain an optimal set W, and consequently, a single posterior prediction y^*. This problem is expressed in equation (2.2).

The function defined by MCD attempts to solve the first integral in equation (2.2) and minimize it [11].

$$
\begin{aligned}
P(D) &= \int P(D|W)P(W)dW \\
P(W|D) &= \frac{P(D)P(W)}{P(D)} \\
P(y^*|x^*, D) &= \int P(y^*|x^*, W)P(W|D)dW
\end{aligned}
\tag{2.2}
$$

MCD is tasked with capturing the model's uncertainty, specifically epistemic uncertainty, but its formulation does not allow for the calculation or elimination of aleatoric uncertainty.

Operationally, MCD randomly deactivates neurons in the neural network using dropout during the inference phase. This process is repeated multiple times, and with each deactivation of a set of neurons, the network produces different predictions. By observing how these predictions vary, it is possible to determine whether the model is robust or affected by uncertainty. In this thesis work, two dropout probabilities were used: 0.001 and 0.005. The MCD method was applied 10 times, resulting in 10 heatmaps and 10 softmax outputs for each classified image.

2.4 Grad-CAM and Score-CAM

The XAI methods that have been implemented are Grad-CAM and Score-CAM, two model-agnostic methods that provide post-hoc explanations and generate saliency maps. These saliency maps, also known as heatmaps, are represented using a color scale ranging from blue to red. The blue color indicates low importance of the spatial region, while the red color indicates high importance.

Grad-CAM [42] [28] allows visualization of which regions of an image are responsible for the prediction output by the neural network. The method is outlined as follows:

- Perform a forward pass through the network with the input image;
- Identify the last convolutional activation map before the classification operation;
- Compute the gradient of the prediction with respect to the activation map, obtaining information on how each pixel of the map contributes to the predicted class;
- Spatially average the obtained gradients to get a weight for each channel of the activation map;
- Combine the weights with the convolutional activation map to create a heatmap that shows the importance of each area of the image relative to a target.

The advantage of Grad-CAM is that it can be applied to any neural network. However, the method depends on gradients, which can introduce noise or vanish in deeper networks, reducing the quality of the explanation. An example of a Grad-CAM visualization is shown in Fig. 2.5.

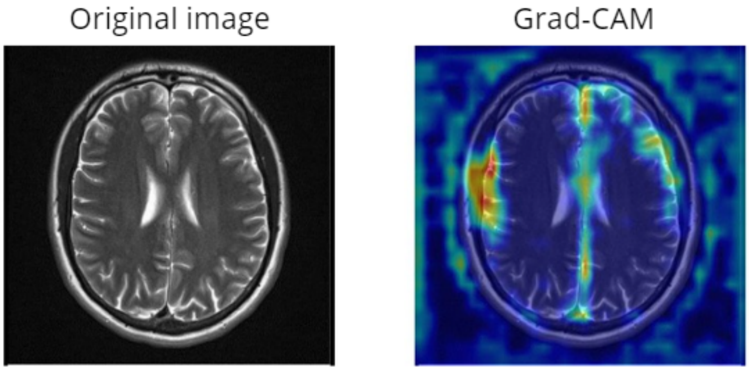

Figure 2.5. This is the image(276) labelled no tumor. On the left there is the original image, on the right there is the heatmap provided by the Grad-CAM method placed on the original image. It is possible note the characterized colors of the heatmap.

Score-CAM [37] was developed as an improvement over Grad-CAM because it does not use gradients, thus eliminating the issue of noise and gradient vanishing in deep neural networks. Like Grad-CAM, Score-CAM can be applied to any neural network, and the method is outlined as follows:

- Perform a forward pass to obtain the activation map of the last convolutional layer;
- For each activation channel, the map is normalized and overlaid onto the original image;
- Perform a forward pass with the modified image to evaluate how each specific channel affects the target class score;
- The obtained scores are used as weights to combine the channels of the activation map, producing the final heatmap.

Score-CAM avoids the gradient-related issues that affect Grad-CAM, resulting in a more stable and accurate visualization. However, this method is computationally more expensive, as it requires more forward passes compared to Grad-CAM. An example of a Score-CAM visualization is shown in Fig. 2.6.

2.5 Re-evaluating Pearson Correlation Coefficient method

The primary objective of this thesis is to quantify the variability and uncertainty of the most important features computed by the neural network. DL neural networks automatically extract features from the images provided to them: the more complex the neural network, the more complex the features extracted. It is not possible to know which features are calculated or utilized by the network for image classification. However, through visual XAI methods, it is possible to visualize the spatial regions where the features are

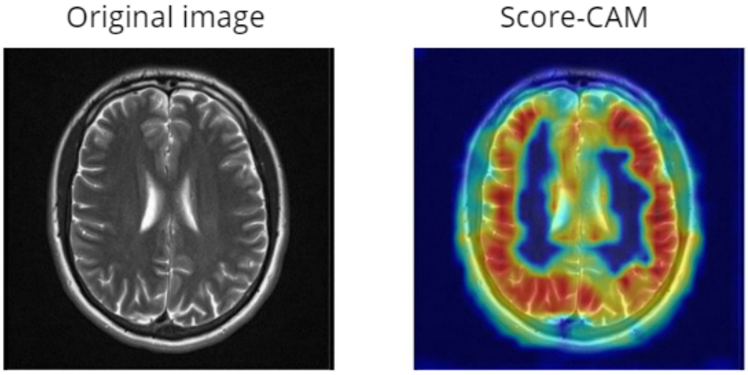

Figure 2.6. This is the image(276) labelled no tumor. On the left there is the original image, on the right there is the heatmap provided by the Score-CAM method placed on the original image. It is possible note the characterized colors of the heatmap.

so critical to the neural network that interpretation and subsequent classification rely solely on them.

The XAI method generates an image known as a heatmap. In this heatmap, each pixel is assigned a color that ranges from blue to bright red. The closer a pixel's color is to red, the more important the feature calculated at that pixel is. The pixel color, however, is not directly proportional to the feature's value.

By utilizing the MCD method, various neurons in the network are deactivated, leaving classification to the remaining active neurons. As a result, the features used for the new classification may differ from those employed when all neurons in the neural network are active. After each round of neuron deactivation, a new heatmap and corresponding classification are generated. The MCD method is used to determine whether the heatmaps change and how the classification is affected when certain neurons are turned off, thus assessing the robustness of the network. Additionally, the MCD method is applied during an initial network evaluation phase: if the most important features are extracted from the background of the images or do not entirely fall within the brain region, this indicates that the network is suboptimal, necessitating adjustments to certain parameters or modifications to its architecture.

At the conclusion of the network training phase, the following data are obtained:

- The heatmaps of the images, along with their associated softmax values and classifications;
- The heatmaps, softmax values, and classifications for all instances where neurons were deactivated using MCD.

In this specific case, the network's neurons were deactivated 10 times, resulting in 11 heatmaps, 11 softmax values, and 11 classifications for each analyzed image. This yields a total of 4,763 images, 4,763 softmax values, and 4,763 classifications just for the validation set.

According to the MCD method, if all the heatmaps are identical and the classification remains unchanged, the neural network is robust in its predictions. This reasoning is straightforward: invariant localizations correspond to consistently accurate interpretations.

If the results deviate from this expected outcome, what can be inferred about the reliability of the neural network? The neural network can exhibit behaviors that are not fully accounted for by the MCD method:

- The network may produce similar heatmaps but different classifications;
- The network may generate differing heatmaps, yet the classification remains the same;
- The heatmap produced by the neural network with all neurons active may result in an incorrect classification, but applying the MCD method consistently yields correct classifications;
- Conversely, the heatmap produced by the neural network with all neurons active may yield a correct classification, but this classification may not always be maintained when the MCD method is applied.

The use of the MCD method opens up a range of hypotheses that must be examined and analyzed to understand the variability and uncertainty of the most important features extracted by the neural network.

One certainty should be acknowledged: if the heatmaps generated by the network, both with all neurons active and with a subset of neurons deactivated in turn, are similar to one another, then the localization phase is robust. This is affirmed by the MCD method and is not questioned in this thesis project. If all heatmaps were identical and the variability of each pixel was calculated, this variability would be zero; thus, the features identified as most important by the network would have zero variability. However, it is important to note that the actual value of the feature at each pixel is not known; only whether the feature at each pixel is important or not for the neural network can be determined. Therefore, when attempting to calculate the variability of the features, one is actually evaluating whether the heatmaps provided by the network exhibit variability.

Suppose the heatmaps generated by the neural network differ from one another, but the classification remains accurate. In this case, the network's interpretation would be robust: deactivating neurons in a neural network is analogous to removing a tree from a mountain landscape photo, and if the response continues to be "mountain landscape", the network is robust in its predictions. If one were to calculate the pixel-level variability in this scenario, a variability map would be obtained, corresponding to the variability of each feature computed by the network, not just the features used for classification. Given varying localizations, how could one calculate the variability of the most important features? In this case, it is reasonable to assume that the heatmaps provided could simultaneously be both correct and incorrect. This consideration led to the development of ReP method which stands for re-evaluating Pearson Correlation Coefficient.

ReP method was designed to manipulate heatmaps and their associated softmax values to identify the spatial region that appears most consistently across the heatmaps. If a spatial region is always present, it indicates that different features calculated in the same location continue to be decisive for classification. And what about the spatial regions that are not consistent? It cannot be definitively stated that these regions are not critical for classification; therefore, they cannot simply be discarded from the image. Any information that contributed to classification is "hidden" in the heatmaps and their corresponding predictions.

By averaging all the heatmaps pixel by pixel, a heatmap is obtained that contains all the most important spatial regions, including the common area. This first step is schemed in Fig. 2.7

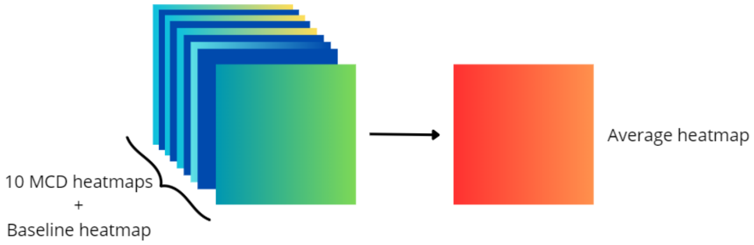

Figure 2.7. This is a scheme of the first step of the ReP method. All the 10 MCD heatmaps and the baseline heatmap, which is the one provided when the network has all its neuron activated, are averaged.

To highlight this region, cumulative averaging is used. In cumulative averaging, the earlier elements carry less weight than the later ones. To extract the common spatial region, the heatmaps that are least similar to the average are averaged first, followed by those that are more similar. The Pearson Correlation Coefficient (PCC) is used as the indicator to determine the similarity between two images. The PCC is a correlation coefficient that measures the linear relationship between two sets of data. The closer the PCC value is to 1, the more similar the two images are; the closer the PCC value is to zero, the more different the images are. In this method, the PCC is calculated between the average heatmap and all the original heatmaps; the heatmap with the lowest value is averaged first. The PCC is continuously recalculated between the cumulative average being constructed and the remaining original heatmaps, with the heatmap possessing the lowest PCC value being averaged first.

The method also manipulates the softmax values, which are treated in the same manner as the heatmaps, following the averaging order of the corresponding images. The initial step is the same as the heatmaps one and is schemed in Fig. 2.8 At the end of this process, a heatmap is obtained that represents the spatial region consistently decisive for classification, along with its associated softmax.

The method developed seeks to identify the smallest common area across all the heatmaps. This approach is inspired by feature selection techniques used in ML: such methods aim to identify the smallest number of salient features that allow for accurate classification while maintaining the same informational content. At the conclusion of the ReP method process, a heatmap with the smallest common activation region is obtained.

An additional hypothesis was used during the development of the method: a network

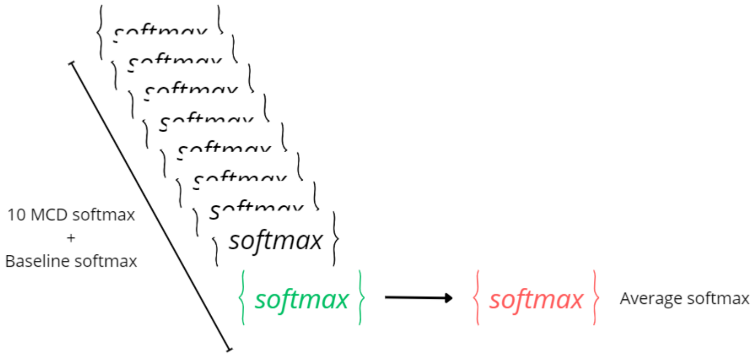

Figure 2.8. This is a scheme of the first step of the ReP method. All the 10 MCD softmax and the baseline softmax, which is the one provided when the network has all its neuron activated, are averaged.

with all neurons active that correctly classifies an image will likely continue to classify it correctly even when the MCD method deactivates some neurons. This hypothesis was incorporated into ReP method following an analysis of the available data. The same hypothesis applies to misclassified images: if the neural network with all neurons active misclassifies an image, it will likely continue to do so when the MCD method is applied.

ReP method treats correctly classified images differently from misclassified ones: for the former, the PCC is sorted in ascending order, so the heatmaps most different from the average are averaged first; for the latter, the PCC is sorted in descending order, so the heatmaps most different from the average are averaged last.

Is the PCC an optimal indicator capable of correlating correct/incorrect image predictions? Can a PCC threshold be identified? The answer is negative, and the reason is straightforward: if the PCC alone could discriminate between a heatmap associated with the correct class and one associated with the wrong class, the heatmaps would have different localizations, and the network would misclassify due to a localization issue. It cannot be assumed with certainty that the network consistently errs in the localization phase; it may also have issues during the interpretation phase, specifically with the rules learned during the training and validation stages.

This brings us to the third condition utilized by the method: the necessity of having ground truth (GT), the correct class for each image. GT is incorporated into the PCC calculation during heatmap manipulation as follows: always using the PCC order as previously described, all heatmaps classified as GT are always averaged last. This approach "cheats" by constructing heatmaps that more closely resemble those of correctly classified images, ensuring that the information leading to correct classification is preserved. Heatmaps associated with incorrect classifications are also averaged, as the network's performance depends on these images as well, and it is essential to determine in which phase the network makes errors. This third condition is based on the fact that the neural network may produce similar heatmaps, but only some of them may correspond to correct classifications. Fig. 2.9 presents the schematic of the ReP method, illustrating how the cumulative average heatmap is constructed using the PCC and the GT class.

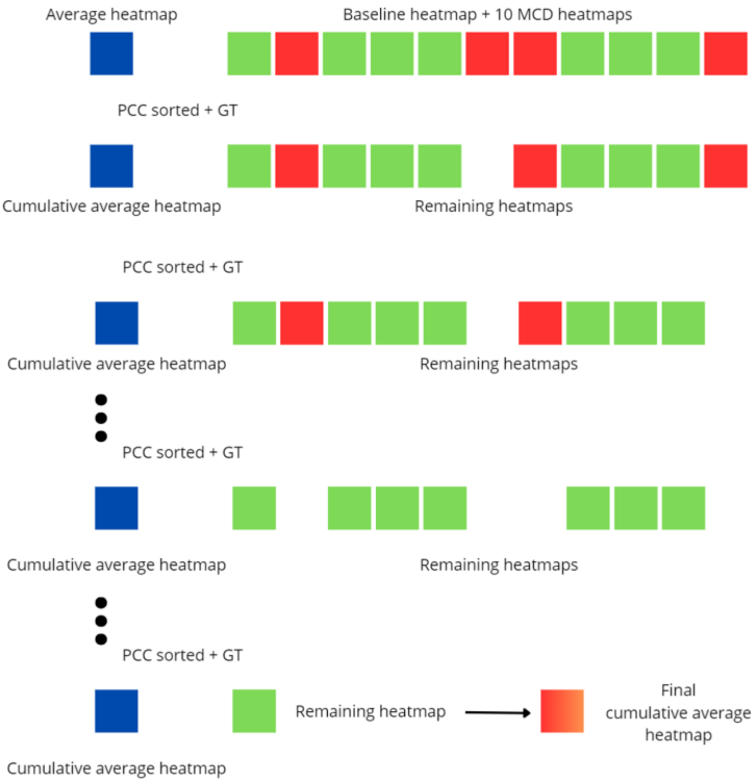
Average heatmap
Baseline heatmap + 10 MCD heatmaps
PCC sorted + GT
Cumulative average heatmap
Remaining heatmaps
PCC sorted + GT
Cumulative average heatmap
Remaining heatmaps
PCC sorted + GT
Cumulative average heatmap
Remaining heatmaps
PCC sorted + GT
Remaining heatmap
Final cumulative average heatmap
Cumulative average heatmap

Figure 2.9. This is a scheme of the construction of the cumulative average heatmap by the ReP method. In this scheme the green heatmaps are the ones labelled with the GT class, the red heatmaps are the ones with a label different from the GT class. At first the PCC is evaluated between the average heatmap and the 11 original heatmaps. The PCC is sorted given the order described in this paragraph. Whatever the order, the first heatmaps that are averaged are the one whitouth the GT label, so the red ones. At a certian point only the green heatmaps remain, but the process is not finished yet. After the last heatmap is averaged the cumulative average heatmaps is provided. The same reasoning is applied to softmax values.

Fig. 2.10 refers to image p(721), which belongs to the pituitary tumor class. The image is correctly classified by the neural network, even when the MCD method is applied. The two heatmaps in the first row are generated by the neural network. Although the image is consistently classified correctly, the heatmaps provided are not identical. The heatmap provided by the ReP method is the one on the bottom right, while the original p(721) image is shown on the bottom left.

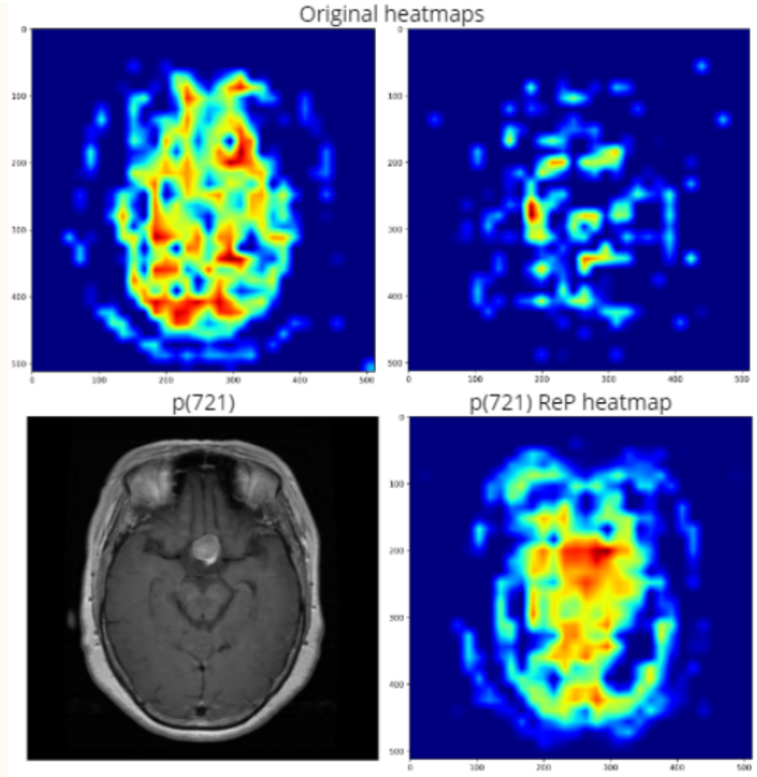

Figure 2.10. This is the image p(721) and belongs to the pituitary tumor class. In the first row two of the 11 generated heatamps are presented. Even though the image is consistently correctly classified, the heatmaps privided are not the same. The most representative heatmaps provided by the ReP method is the one at the bottom right, on the bottom left there is the original p(721) image.

2.5.1 Variants of ReP method

As described in Section 2.5, the ReP method is developed based on three hypotheses:

- The method provides a heatmap representing the smallest area common to all the original heatmaps;
- The classification obtained when all the neurons in the neural network are active is likely to remain consistent when the MCD method is applied, regardless of whether the image is correctly or incorrectly classified;
- It is necessary to have the GT for each image to associate it during the calculation of the PCC, in order to optimally aggregate the heatmaps in the cumulative average.

Based on these hypotheses, correctly classified and misclassified images are processed differently to obtain the final heatmap and softmax. However, the ReP method is not the only possible approach.

The OtP method, an acronym for One-time PCC, is the second method developed. In this method, the PCC is calculated only once between the average of the heatmaps and the original heatmaps. The PCC values are subsequently ordered, following the considerations outlined in Section 2.5, treating correctly classified images differently from misclassified ones, and always averaging last the heatmaps where the prediction matches the GT. This method also manipulates both heatmaps and softmax outputs. The OtP and ReP methods differ in how the heatmaps are aggregated and can be applied to both correctly classified and misclassified images.

Fig. 2.11 illustrates the PCC calculation performed by the OtP method. Fig. 2.12, on the other hand, shows how the PCC values are ordered when the original image is correctly classified and how the ordering appears when the GT class is also used. Both figures represent the manipulation of heatmaps, but it should be noted that the same manipulation is applied to the softmax outputs as well.

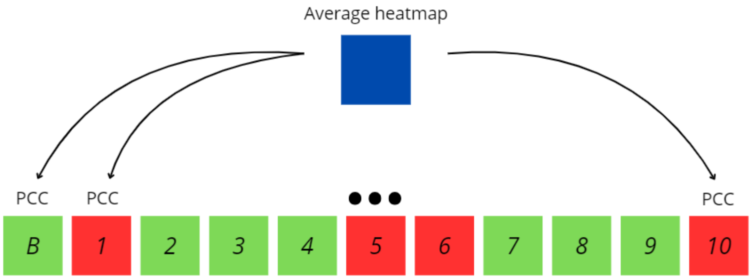

Figure 2.11. This is a scheme of how the PCC is evaluated just one time between the average heatmap and the 11 original heatmaps. The green heatmaps are the labelled with the GT class, the red ones have a different classification. In this scheme *B* stands for baseline heatmap, while the numbers refer to the MCD heatmaps.

Fig. 2.13 provides an example of what happens when the original image is misclassified. The figure illustrates the PCC calculation performed by the OtP method. In Fig. 2.14, the change in the ordering of the PCC values is shown when it is known that the original heatmap is misclassified, and how the ordering appears when the GT class is used. These two figures only represent the manipulation of the heatmaps, but the same manipulation, following the same ordering, is applied to the softmax outputs as well.

In Section 2.5, it was also discussed how all the heatmaps provided by a neural network can simultaneously be considered both correct and incorrect. A new hypothesis could be introduced: an image that is correctly classified provides a heatmap that is sufficiently representative of the spatial region encompassing the most important features. The heatmaps generated by the MCD method serve to adjust the initial heatmap, always aiming to obtain a heatmap that represents the smallest common area. Based on this new hypothesis, the OtP and ReP methods can be modified as follows:

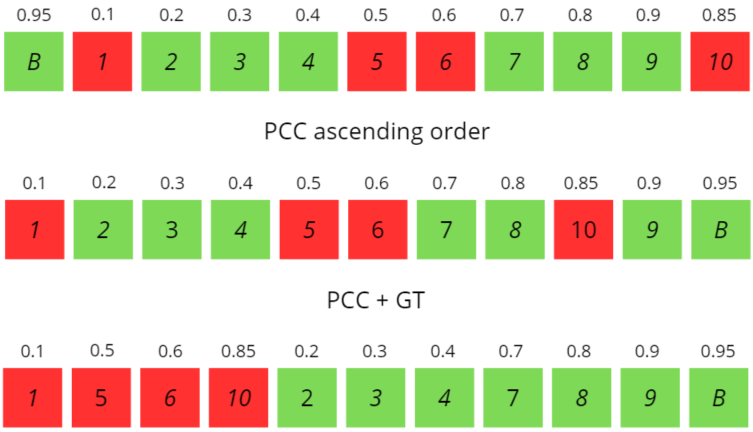

Figure 2.12. In this scheme the baseline heatmap is correclty classified, *B* has green colour. Above each heatmap there a hypothetic PCC value. Because the original image is correclty classified the PCC values are sorted in ascending order. Then all the heatmaps that are classified as GT are putted as last in the averaging order list.

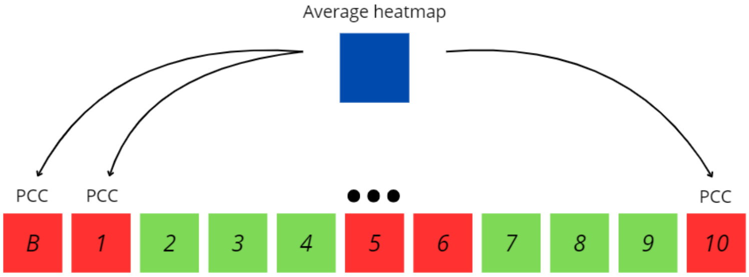

Figure 2.13. This is a scheme of how the PCC is evaluated just one time between the average heatmap and the 11 original heatmaps. The green heatmaps are the labelled with the GT class, the red ones have a different classification. In this scheme *B* stands for baseline heatmap, while the numbers refer to the MCD heatmaps.

- The first step of pixel-by-pixel averaging of all the provided heatmaps is not performed, as a base heatmap is already available from which to start building the cumulative mean. This is the heatmap created when all the neurons in the neural network are activated;
- The PCC is calculated between the original heatmap and those provided by the MCD method, and ordered either in ascending or descending order, depending on whether the image is correctly or incorrectly classified, continuing to average last those heatmaps that present the GT class;
- The softmax outputs are manipulated exactly like their corresponding heatmaps, following the aggregation order dictated by the PCC and the GT class.

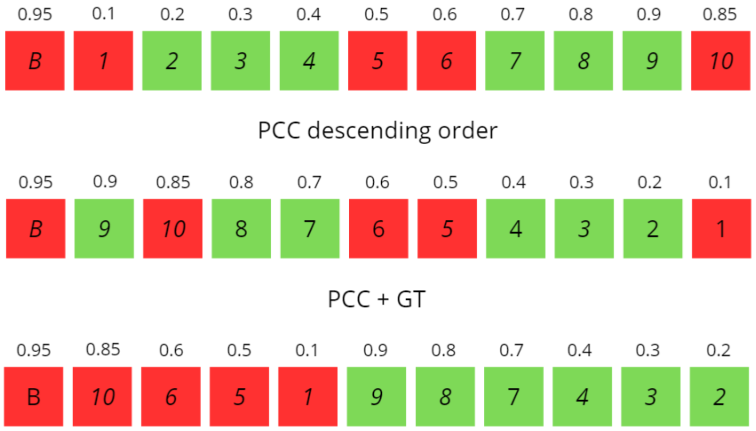

Figure 2.14. In this scheme the baseline heatmap is incorreclty classified, *B* has red colour. Above each heatmap there a hypothetic PCC value. Because the original image is incorreclty classified the PCC values are sorted in descending order. Then all the heatmaps that are classified as GT are putted as last in the averaging order list.

Both the OtP and ReP methods can be modified to adhere to this new hypothesis, resulting in the OtP-A and ReP-A methods. The last letter of the acronym indicates that the first step of pixel-by-pixel averaging of all the heatmaps is unnecessary. These two new methods retain the difference in image aggregation: in the OtP-A method, the PCC is calculated only once between the heatmap obtained when all the neurons in the neural network are activated and the heatmaps provided when the MCD method is applied; in the ReP-A method, the PCC continues to be recalculated between the cumulative heatmap under construction and the remaining original heatmaps.

It is important to note that not all methods can be applied to all images:

- Given the three hypotheses outlined at the beginning of this paragraph, the OtP and ReP methods can be applied to all images;
- Given the fourth new hypothesis, the OtP-A and ReP-A methods can only be applied to images that are correctly classified by the neural network.

All the methods presented provide a heatmap that is more representative of the spatial region used to classify the image, with the respective softmax associated. From the heatmaps obtained by the various methods, a series of metrics can be calculated to identify the variability of the most important features and the uncertainty. From the softmax, it is possible to extract the uncertainty of the prediction.

Fig. 2.15 refers to image p(721), which belongs to the pituitary tumor class. The image is correctly classified by the neural network, even when the MCD method is applied. Because is a correclty classified image, ReP, OtP, ReP-A, and OtP-A methods can be applied to the image. The final heatmaps are presented for each method. The heatmaps are different from one another, especially in the shading and in the extent of the most important areas.

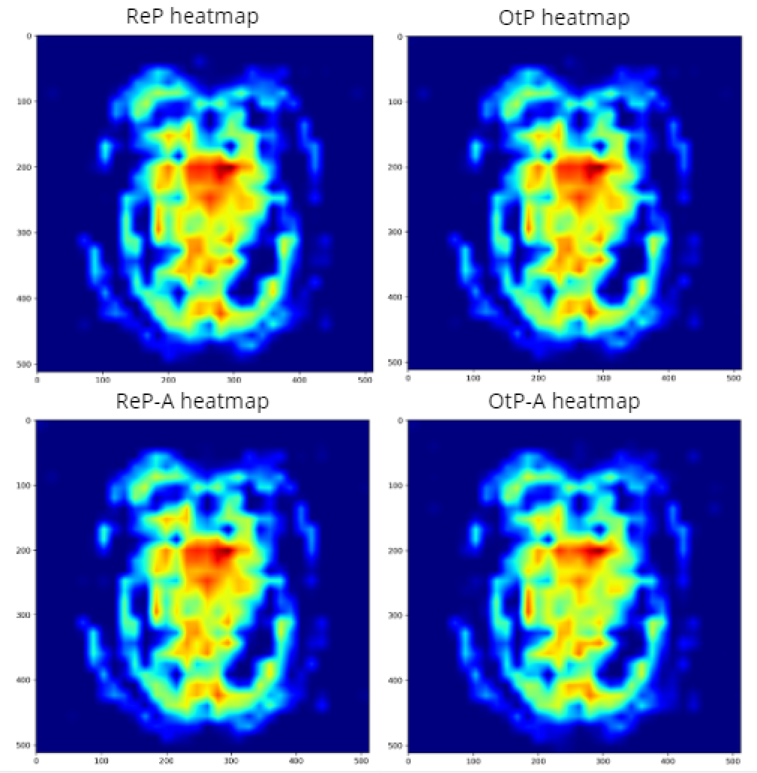

Figure 2.15. These are the ReP, OtP, ReP-A, and OtP-A heatmaps of the image p(721) which belongs to the pituitary tumor class. All the four methods can be applied because the image is correclty classified by the network. These heatmaps difference in shading and in the extent of the most important areas. On the top left the heatmap provided by the method ReP, on the top right the heatmap provided by the method OtP, on the bottom left the heatmap provided by the method ReP-A, on the bottom right the heatmap provided by the method OtP-A.

Fig. 2.16 refers to image gg(704), which belongs to the glioma tumor class. The image is misclassified by the neural network, but is correclty classified seven times when the MCD is applied. Because is a misclassified image, the only methods that can be applied to analyse this image are ReP, and OtP. The final heatmaps are presented for each method and over each heatmap there is the method that generates it. The heatmaps are different from one another, especially in the shading and in the extent of the red and green areas.

2.5.2 Variability and uncertainty on the heatmap

The heatmap obtained from each method displays the spatial region containing the most important features used by the network to classify the images. This heatmap retains its characteristic colors, ranging from blue to red.

From this heatmap, it is now possible to calculate the variability of the most important

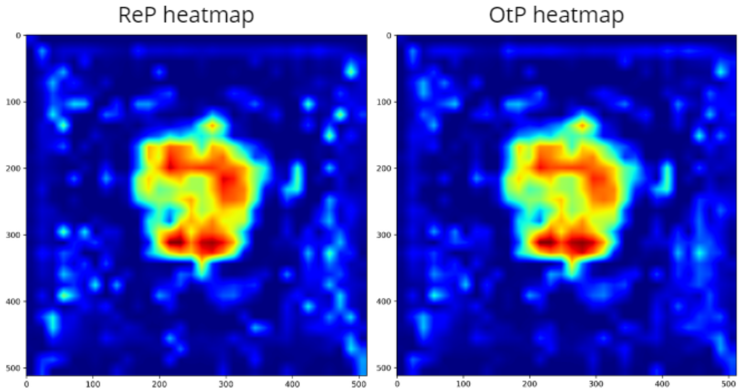

Figure 2.16. This is the image gg(704) which belongs to the glioma tumor class. Because it is a misclassified image only ReP and OtP methods can be applied. On the left there is the heatmap provided by the ReP method, on the right there is the heatmap provided by the OtP method. These heatmaps difference in shading and in the extent of the red and green areas

features and the associated uncertainty. This calculation can now be performed because the heatmap provided by the various methods accurately reflects the spatial region that is consistently identified and utilized for image classification. As described in Section 2.5, the most relevant area, depicted in red, represents the smallest common area across all original heatmaps.

To calculate the variability of the most important features, the original heatmaps are first stacked, and then pixel-by-pixel variability is computed, resulting in a variability map. This variability map represents the variability of all the features calculated by the network. The map is then multiplied by the heatmap provided by the various methods. Since the heatmap is normalized, with values ranging from 0 to 1, the variability of pixels that fall within a significant area - where the red area contains pixels with values approaching 1 - is accentuated, while the variability of pixels in an insignificant area, which possess values approaching 0, is diminished. At the end of the multiplication, a map is obtained that allows one to observe how variable the features present in the heatmap generated by the method are. Furthermore, the map also includes the variability of the most important features.

To isolate solely the variability of the most important features, it is necessary to find a method that identifies this variable region. To isolate this region, the 95% confidence interval was calculated, with its mathematical formulation presented in Equation (2.3):

$$\begin{gathered} \mu - 1.96\sigma \\ \mu + 1.96\sigma \end{gathered} \tag{2.3}$$

where μ and σ represent the mean and standard deviation of the values that the pixels belonging to the map, obtained by multiplying the heatmap from the methods with the variability map of all features, assume. The variable pixels that belong to the most significant region are those that do not fall within the upper limit of the 95% confidence interval, specifically, those pixels with values exceeding $\mu + 1.96\sigma$.

At the end of this operation, the variable region of the most important features is obtained. As mentioned in Section 2.5, the calculation of the variability of only the most important features translates into a problem of variability regarding the region encompassing those features used by the network for classification.

The GT, if the network classifies an image as such, is also utilized in this calculation, allowing for the identification of two regions of variability: one dependent on the GT class and one dependent on all classes that are not GT. If the GT is never present, or if the network never misclassifies, only one area of variability is identified instead of two.

All original heatmaps classified as GT are separated from all other heatmaps. Once these two distinct groups are obtained, and the original heatmaps are stacked, the procedure continues as outlined above:

- The pixel-by-pixel variability of each stack of heatmaps is calculated, resulting in a variability map for the GT and a variability map for the misclassified images;
- Each variability map is multiplied by the heatmap provided by the method, identifying the variable regions;
- Given the resulting maps from the multiplication, the 95% confidence intervals are calculated;
- After identifying the upper limits, the variable regions of the most important features are obtained, yielding a contribution derived from the GT and a contribution derived from misclassification.

Since the heatmap provided by the methods is constructed using both those resulting from correct classifications and those from misclassifications, it is reasonable to identify which areas of the most significant region derive from the GT and which stem from the misclassifications.

Fig. 2.17 refers to image gg(706), which belongs to the glioma tumor class. The image is correctly classified by the neural network, but misclassified two times when the MCD method is applied. The heatmap used to evaluate the variability areas is the one provided by the ReP method. On the left column the two variability maps are evaluated, one for the misclassified images (fc), and one for the correctly classified images (tc). On the right column the multiplication between the variability heatmaps and the ReP heatmap.

Uncertainty is calculated in the following manner. Given the heatmap provided by the methods, the pixel-by-pixel squared difference with the original heatmaps is computed. This results in as many difference maps as there are original heatmaps. The Root Mean Square Error (RMSE) is calculated for each pixel of the difference map using Equation (2.4):

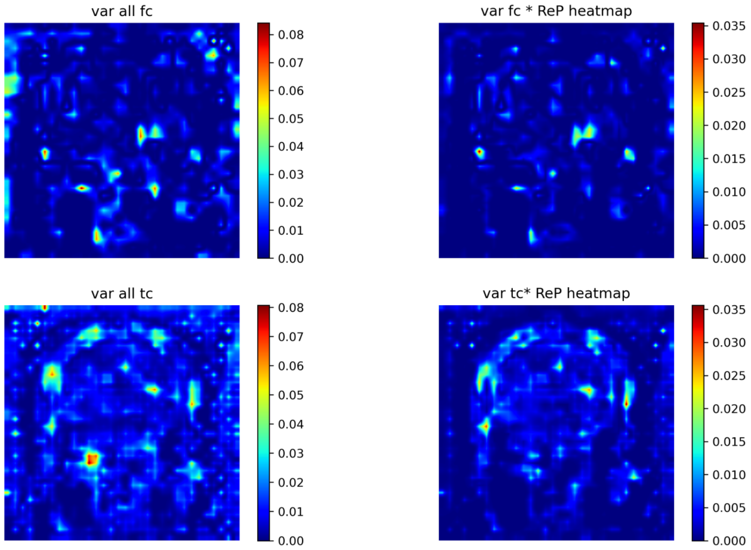

Figure 2.17. This is the image gg(706) which belongs to the glioma tumor class. The image is correctly classified by the network, but misclassified two times when the MCD method is applied. On the left columns there are the two variability maps, on the top left corner there is the variability map for the misclassified images (fc), on the bottom left corner there is the variability map for the correctly classified images (tc). Each variability map is the multiplied with the heatmap provided by the method, here the ReP heatmap is used.

$$RMSE = \sqrt{\frac{\sum_S b[m,n,s]}{S}} \tag{2.4}$$

where $b[m,n,s]$ represents the various difference maps positioned at different levels s. All values that the pixels assume at the various coordinates $[m,n]$ are summed, and then divided by the number of heatmaps present S.

Once an RMSE map is obtained, it is multiplied by the heatmap provided by the methods, thereby better highlighting which areas of the features are affected by error. To determine which areas impact the most important features, a 95% confidence interval is calculated. All pixels exceeding the value identified by the upper bound of this interval are those that delineate the spatial regions of the most important features affected by error.

Since the cumulative heatmap provided by the various methods is constructed using both correctly classified and misclassified heatmaps, it is reasonable to indicate which areas of this heatmap do not correspond to the original areas. The calculation of the RMSE allows for the identification of these erroneous areas, treating them as areas of uncertainty. It is important to emphasize that in this case, uncertainty pertains to the process of constructing the heatmap. Without knowing the features utilized for classification or the values they assume, it is not possible to extrapolate the uncertainty of the features from the heatmap. One might think that the green area present in the heatmaps could represent some form of feature uncertainty or a region of uncertainty; however, the green region should only be treated as an important region for the network, albeit not sufficiently significant to determine the classification of the image.

Once again, the GT is utilized to identify two contributions of uncertainty: one dependent on the GT class and one dependent on misclassification. If the GT is always present, or if the network consistently misclassifies, only one area is identified instead of two. When the network correctly classifies and misclassifies an image, all heatmaps reporting the GT classification are separated from the others and stacked. Once these two groups are obtained, the procedure continues as follows:

- The squared difference between the method's heatmap and the original heatmaps is calculated, resulting in difference maps for the GT class and difference maps for the misclassified images;
- The pixel-by-pixel RMSE is calculated for each stack of maps, yielding two RMSE maps;
- Each RMSE map is multiplied by the heatmap from the method to highlight the areas affected by error;
- Given the resulting maps from the multiplication, the 95% confidence intervals are calculated;
- After identifying the upper limits, the regions affected by error in the most important features are obtained, yielding contributions derived from the GT and from misclassification.

Fig. 2.18 refers to image gg(706), which belongs to the glioma tumor class. The image is correctly classified by the neural network, but misclassified two times when the MCD method is applied. The heatmap used to evaluate the uncertainty areas is the one provided by the ReP method. On the left column the two RMSE maps are evaluated, one for the misclassified images (fc), and one for the correctly classified images (tc). On the right column the multiplication between the RMSE heatmaps and the ReP heatmap.

Fig. 2.19 refers to image gg(706), which belongs to the glioma tumor class. The image is correctly classified by the neural network, but misclassified two times when the MCD method is applied. The heatmap used to evaluate the uncertainty areas is the one provided by the ReP method. On the left there is the final uncertainty map. The blue areas represent the contributions that the misclassified heatmaps bring to the ReP heatmap, the green areas represent the contributions from the correctly classified heatmaps, and the yellow areas show the overlap between these two contributions. These areas indicate how different the ReP heatmap is from the original heatmaps. On the right, the final variance map is shown. The blue areas identify the contributions from the misclassified heatmaps, the green areas show the contributions from the correctly classified heatmaps, and the yellow areas indicate the overlap between the two contributions. These areas highlight where the most important features in the ReP heatmap are more variable.

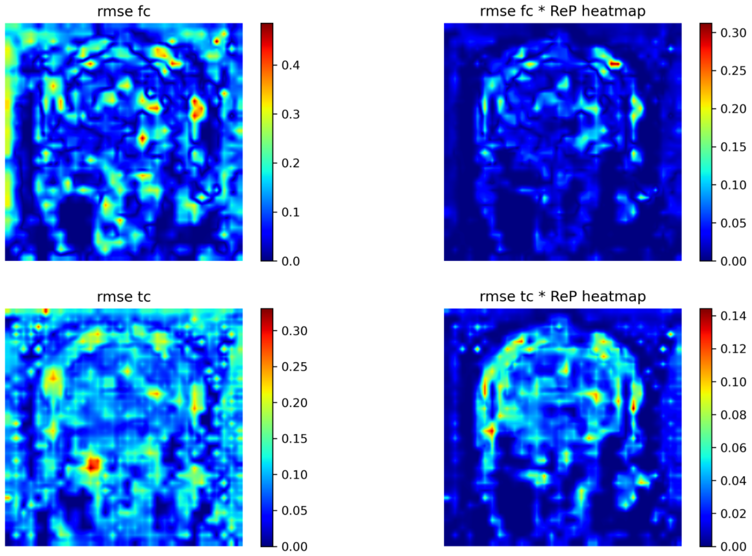

Figure 2.18. This is the image gg(706) which belongs to the glioma tumor class. The image is correctly classified by the network, but misclassified two times when the MCD method is applied. On the left columns there are the two RMSE maps, on the top left corner there is the RMSE map for the misclassified images (fc), on the bottom left corner there is the RMSE map for the correctly classified images (tc). Each RMSE map is the multiplied with the heatmap provided by the method, here the ReP heatmap is used.

Thus, from the heatmap generated by the methods, it is possible to identify the variability of the most important features, pinpointing which areas of the region presented in red are more variable. If classifications different from the GT are present, two contributions of variability can be identified: one stemming from correctly classified heatmaps and one from misclassified heatmaps. Another contribution that can be identified is that of uncertainty. Uncertainty indicates, within the red area of greater importance, spatial regions that differ from the original heatmaps. These areas signify uncertainty regarding the construction of the method's heatmap, rather than the uncertainty of the most relevant features.

2.5.3 Prediction Uncertainty

The uncertainty of the most important features is not identifiable on the heatmaps; however, it is reflected in the classification of images. The softmax is a vector that indicates the probabilities of the image belonging to various classes. In this study, the classes are four: no tumor, pituitary tumor, meningioma tumor, and glioma tumor. Therefore, the softmax contains four probabilities that sum to 1. From the softmax, it is possible to identify the uncertainty of the prediction by calculating the margin.

The margin is the difference between the probability assigned to a certain class and the probability assigned to another class. A wide margin indicates that the network is very

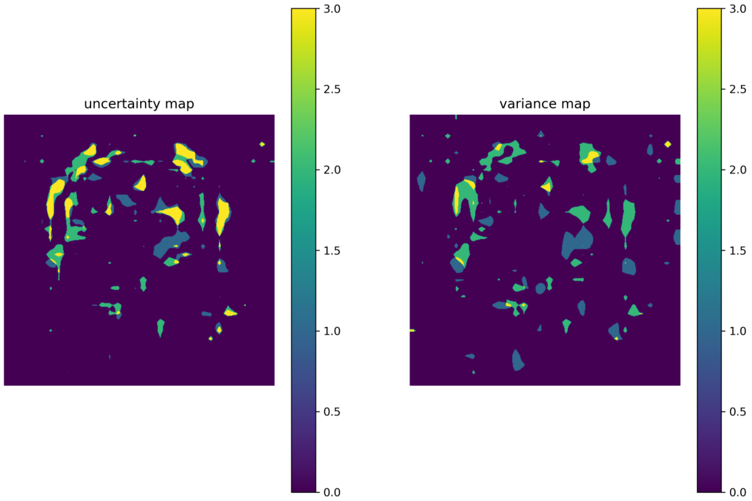

Figure 2.19. This is the image gg(706) which belongs to the glioma tumor class. The image is correctly classified by the network, but misclassified two times when the MCD method is applied. The heatmap used to evaluate the uncertainty areas is the one provided by the ReP method. On the left there is the final uncertainty map. The blue areas represent the contributions that the misclassified heatmaps bring to the ReP heatmap, the green areas represent the contributions from the correctly classified heatmaps, and the yellow areas show the overlap between these two contributions. On the right, the final variance map is shown. The blue areas identify the contributions from the misclassified heatmaps, the green areas show the contributions from the correctly classified heatmaps, and the yellow areas indicate the overlap between the two contributions.

confident in the provided response, while a narrow margin indicates that the network is not very certain about the classification made. Depending on whether the image is correctly classified or not by the network, the margin is calculated as follows:

- For a correctly classified image, the margin is calculated between the highest probability, corresponding to the GT class, and the second most probable class, which corresponds to an incorrect class;
- For a misclassified image, the margin is calculated between the highest probability, which does not correspond to the GT class, and the probability associated with the GT class.

The original margin is calculated using the softmax provided by the network when all neurons are activated, following the aforementioned distinction between correct and incorrect classifications. Subsequently, the margin is calculated using the softmax computed by the various methods, which is associated with the produced heatmap. Since the heatmap provided by the methods offers the most accurate localization of the most important features, it is appropriate to use the associated softmax to quantify the uncertainty of the prediction.

Fig. 2.20 provides an example of margin calculation. The softmax output is presented in the form of a graph and shows the four classes into which an image can be classified. The example focuses on the margin calculation for a correctly classified image. The margin is always calculated between the probability of GT class and the second most probable class within the softmax.

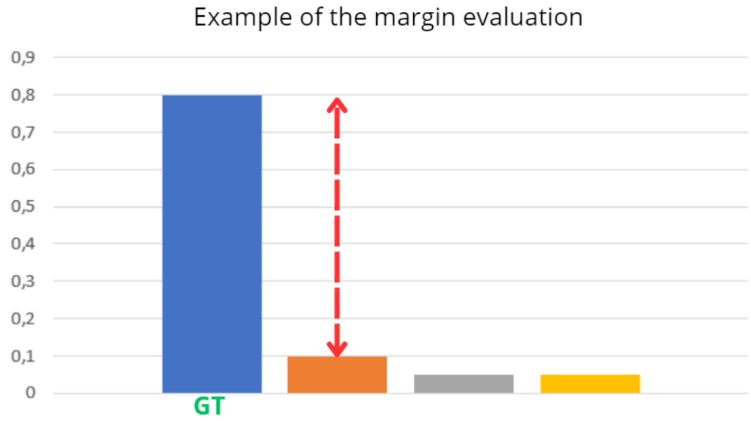

Figure 2.20. This is an example of margin calculation. The softmax output is presented in the form of a graph and shows the four classes into which an image can be classified. The example focuses on the margin calculation for a correctly classified image. The margin is always calculated between the probability of the GT class and the second most probable class within the softmax. The sum of all probabilities is always 1.

Fig. 2.21 provides an example of margin calculation. The softmax output is presented in the form of a graph and shows the four classes into which an image can be classified. The example focuses on the margin calculation for a misclassified classified image. The original margin is calculated between the probability of the wrong predicted class and the probability of the GT class.

The heatmap created may exhibit a more robust prediction, indicating a wider margin than the original, or a more uncertain prediction, indicating a narrower margin than the original.

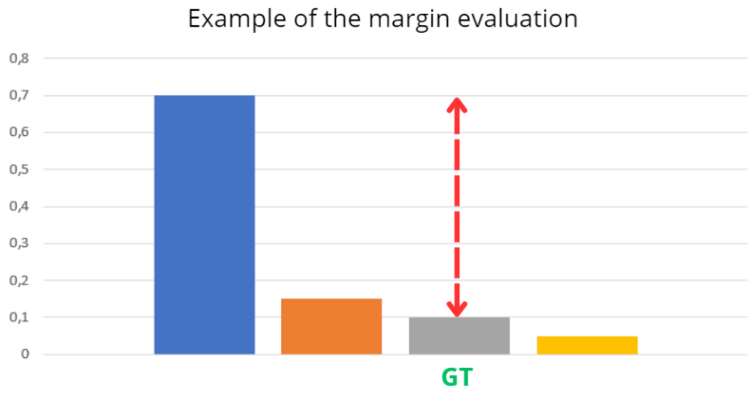

Figure 2.21. This is an example of margin calculation. The softmax output is presented in the form of a graph and shows the four classes into which an image can be classified. The example focuses on the margin calculation for a misclassified image. The original margin is calculated between the probability of the probability of the wrong predicted class and the probability of the GT class. The sum of all probabilities is always 1.

Chapter 3

Results

By applying the ReP, OtP, ReP-A, and OtP-A methods described in Section 2.5 and Section 2.5.1, identifying variability and uncertainty areas on the heatmaps as illustrated in Section 2.5.2, and calculating prediction uncertainty as reported in Section 2.5.3, the images in the validation set are analyzed.

As presented in Section 2.1.1, the validation set consists of 433 images divided into four classes: no tumor, pituitary tumor, meningioma tumor, and glioma tumor. For each image, the heatmap generated by the network with all neurons activated and the heatmaps derived from the MCD method are saved in NPZ format; in total, the validation set contains 4,763 heatmaps.

The images in the validation set were classified by the XCiT neural network four times: two different XAI methods were applied to the neural network, and each XAI method was tested by adjusting the dropout probability from 0.001 to 0.005. In total, Score-CAM and Grad-CAM produced 9,526 datasets each, including heatmaps and softmax outputs for investigation.

The first objective of this Master's Thesis is to determine which of the four proposed methods demonstrates the best performance by specifying:

- which method produces the most optimal heatmap construction, resulting in reduced variability and uncertainty areas;
- which method results in the lowest prediction uncertainty on the softmax outputs.

Once the best method is identified, the second objective of the thesis is to determine which XAI method is most compatible with the XCiT neural network and which dropout rate between 0.001 and 0.005 is optimal.

3.1 Flags

In this analysis, the term flag is introduced to differentiate the behavior of the images. Flags are assigned to both correctly classified and misclassified images according to the following criteria:

- A correctly classified image that remains correctly classified after applying the MCD method is assigned flag 0;
- A correctly classified image that is misclassified at least once by the MCD method is assigned flag 1;
- A misclassified image that remains misclassified after applying the MCD method is assigned flag 0;
- A misclassified image that is correctly classified at least once by the MCD method is assigned flag 1.

Images assigned flag 0 indicate a single contribution of variability and a single contribution of uncertainty on the heatmap. Images assigned flag 1 are characterized by two contributions, one from the GT and one from misclassification, both for variability and uncertainty. Additionally, it is possible to quantify the overlap between the two contributions of variability and between the two contributions of uncertainty, as well as to determine the overlap that may occur between areas of variability and areas of uncertainty.

Flags are initially assigned by the methods to each image in the validation set prior to the cumulative heatmap and softmax construction phases. It is not possible for an image, whether correctly classified or not, to be without a flag. The flags are subsequently reviewed and used to verify the final classification of the applied methods. Regardless of the flag, a correctly classified image must remain so, meaning that the GT class must have the highest probability in the methods' softmax. A misclassified image with flag 1 may experience a correction in the classification performed by the neural network, with the GT class attaining the highest probability in the final softmax. Given the design of the heatmap and softmax manipulation, it is reasonable to expect this case to occur, depending on the weight that correctly classified MCD contributions have in the final heatmap and softmax.

Based on the margin definition reported in Section 2.5.3, the following hypotheses are made:

- A correctly classified image with flag 0 should exhibit a larger margin than the original, indicating lower prediction uncertainty;
- A misclassified image with flag 1 should exhibit a larger margin than the original, meaning lower prediction uncertainty.

These hypotheses are consistent with the flag definition: if the MCD method classifies the image into the GT class, it is reasonable to assume that the final heatmap constructed by the methods contains more information than the one provided by the network when all neurons are activated, and that the softmax produced by the method will have a larger margin and lower prediction uncertainty.

A similar reasoning can be applied to correctly classified images with flag 1: in this case, it can be assumed that the heatmap provided by the methods is more influenced by the contribution from misclassified examples, and that the final prediction will have a smaller margin than the original, leading to higher prediction uncertainty.

Flag 0 for a misclassified image identifies a particular set of images. The network and the MCD method consistently misclassify the image, so the softmax provided by the methods predicts an incorrect class as the final image classification. However, it is possible that each original softmax has high prediction uncertainty, with the second most probable prediction always being the GT class. The classification verified after applying the methods is as follows: a misclassified image with flag 0 is an image that remains misclassified even when the MCD method is applied, with the softmax from the methods positioning the GT class as the second most probable class. Given the description of this image set's classification, it is necessary to verify whether the final heatmap, which remains misclassified, has high prediction uncertainty, meaning that the margin between the most probable incorrect class and the GT class, the second most probable, decreases.

Based on the flag description, the following schematic of potential classification and prediction uncertainty scenarios is presented:

- A correctly classified image with flag 0 continues to have the GT class in the methods' final prediction, while the margins may increase, reducing prediction uncertainty;
- A correctly classified image with flag 1 continues to have the GT class in the methods' final prediction, while the margins may decrease, increasing prediction uncertainty;
- A misclassified image with flag 1 may be classified by the methods into the GT class, with margins exceeding the original, resulting in lower prediction uncertainty;
- A misclassified image with flag 0 remains misclassified, but the second most probable class may be the GT class, while the margins may decrease, increasing prediction uncertainty.

After applying each method, the flags are re-evaluated to verify whether they correspond to the respective classification as described above. It is possible for the methods to make mistakes. When the new classification does not fall into the identified scenarios, the method is considered to have made an error, and the analysis of heatmaps and softmax for these images is not performed.

3.2 First objective

With the flags introduced in Section 3.1, the images from the validation set are analyzed to identify the most optimal method for manipulating heatmaps and softmax.

All the images in the validation set are divided into correctly classified and incorrectly classified categories by reviewing the predictions of the XCiT network for each image. The ReP, OtP, ReP-A, and OtP-A methods manipulate the heatmaps and softmax, distinguishing between correctly classified and misclassified images, as outlined in

Section 2.5 and Section 2.5.1. Once the heatmaps and their corresponding softmax outputs are obtained, the methods' new classifications are generated by reapplying the flags for each image according to the rules established in Section 3.1. Only the images that match the flag-based classification are analyzed, with variability and uncertainty areas on the heatmaps quantified, and the final prediction uncertainty compared to the original uncertainty of the neural network.

The number of errors made by the methods, the behavior of the margins, and the quantification of the areas are all data used to determine which method delivers the best performance. Once the optimal method is identified, it is also possible to observe preliminary trends in the performance of the XAI methods and determine which dropout probability is most suitable.

The results are presented by indicating the XCiT network's prediction, the applied XAI method, and the corresponding dropout probability.

3.2.1 Misclassified Grad-CAM 0.005

Table 3.1 and Table 3.2 display in their headers the number of misclassified images identified by the neural network and the method under analysis. The analysis is divided into images that carry a flag 0 and those that carry a flag 1. As described in Section 3.1, a misclassified image with flag 1 may have the GT class as the new prediction after applying the method, while a misclassified image with flag 0 may have the GT class as the second most probable class within the method's softmax. All images that do not align with the classification of their respective flags are considered errors, and their data regarding heatmaps and softmax are not included in the table.

Subsequently, the number of images that conform to the margin behavior is presented: a misclassified image with flag 1 should have a larger margin than the original, while a misclassified image with flag 0 should have a smaller margin than the original.

The heatmaps are analyzed, and the results are presented under the following categories:

- unFC: Uncertainty Area FC, the uncertainty area for heatmaps not classified in the GT class. This area is identified for both flag 1 and flag 0 images;
- unTC: Uncertainty Area TC, the uncertainty area for heatmaps classified in the GT class. This area is identified only for flag 1 images;
- varFC: Variance Area FC, the variance area for heatmaps not classified in the GT class. This area is identified for both flag 1 and flag 0 images;
- varTC: Variance Area TC, the variance area for heatmaps classified in the GT class. This area is identified only for flag 1 images;
- un ov: Uncertainty Overlap, the overlap area between the unFC and unTC areas. This overlap is identified only for flag 1 images;
- var ov: Variance Overlap, the overlap area between the varFC and varTC areas. This overlap is identified only for flag 1 images;

- t ov: Total Overlap, the overlap area between variance and uncertainty areas. This area is identified for both flag 1 and flag 0 images.

In cases where a category in the table cannot be filled, the fields "unTC", "varTC", "un ov", and "var ov" cannot be completed for flag 0 images, and the symbol / appears in the table. The area sizes are reported as the number of pixels.

25 misclassified images			
ReP			
flag 1		flag 0	
number	7	number	9
margin gain	3	margin decrease	8
unFC	13386	unFC	11157
unTC	12567	/	/
varFC	9901	varFC	9753
varTC	9767	/	/
un ov	7300	/	/
var ov	3571	/	/
t ov	14406	t ov	8270

Table 3.1. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.1, it can be observed that the ReP method makes 9 errors, but is able to correct the prediction for 7 images, while 9 images have the GT class as the second most probable class in the method's softmax. Only 3 flag 1 images out of 7 achieve a larger margin than the original, while 8 flag 0 images out of 9 obtain a smaller margin than the original. Given that the total number of pixels per image is 262,144, as the dimensions are 512×512, the variability, uncertainty, and overlap areas are very small. The largest area reported in Table 3.1 is "t ov = 14,406", which corresponds to 5% of the total image area.

In Table 3.2, it can be observed that the OtP method makes 9 errors, but is able to correct the prediction for 7 images, while 9 images have the GT class as the second most probable class in the method's softmax. Only 3 flag 1 images out of 7 achieve a larger margin than the original, while 8 flag 0 images out of 9 obtain a smaller margin than the original. Given the total number of pixels per image, the variability, uncertainty, and overlap areas are very small. The largest area reported in Table 3.2 is "t ov = 14,450", which corresponds to 6% of the total image area.

The ReP and OtP methods are compared, and the results are shown in Table 3.3. The table header continues to indicate the number of images misclassified by the neural network. It is verified whether the methods identify the same flag 0 and flag 1 images. Only the images common to both methods are considered. The margins are analyzed

25 misclassified images			
OtP			
flag 1		flag 0	
number	7	number	9
margin gain	3	margin decrease	8
unFC	13457	unFC	11209
unTC	12523	/	/
varFC	9941	varFR	9791
varTC	9792	/	/
un ov	7321	/	/
var ov	3592	/	/
t ov	14450	t ov	8257

Table 3.2. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

as follows: the method with the largest number of flag 0 images with a smaller margin and the largest number of flag 1 images with a larger margin is identified. The areas are compared, and the method with the smallest variability, uncertainty, and overlap areas on the heatmaps is determined.

Table 3.3 shows how the ReP and OtP methods identify the same flag 0 and flag 1 images. Both methods have the same number of errors, which is 9. The ReP method exhibits the largest margins for flag 1 images and the smallest margins for flag 0 images. Only the cells highlighted in green, which represent the number of images meeting the condition "ReP < OtP", are in favor of the ReP method. Only 4 out of 10 areas favor the ReP method.

The Grad-CAM 0.005 images show low prediction uncertainty when analyzed by the ReP method; however, this method generates heatmaps with larger variability and uncertainty areas compared to the heatmaps generated by the OtP method.

3.2.2 Misclassified Grad-CAM 0.001

Table 3.4 and Table 3.5 present in their headers the number of misclassified images identified by the neural network and the method analyzed. The analysis is divided into flag 0 and flag 1 images, reporting only the images that comply with the classification of each flag. For each flag, the number of images that reflect the behavior of the margins is indicated, along with the size of the various identifiable areas on the heatmaps, measured in pixels.

In Table 3.4, it can be observed that the ReP method makes 5 errors but is able to correct the prediction for 4 images, while 16 images have the GT class as the second most probable class in the method's softmax. Only 1 flag 1 image out of 4 manages to achieve a

25 misclassified images	
share flag 0	9
share flag 1	7
margin decrease ReP < OtP	6
margin gain ReP > OtP	4
flag 1 - ReP < OtP	
unFC	1
unTC	1
varFC	3
varTC	6
un ov	2
var ov	5
t ov	4
flag 0 - ReP < OtP	
unFC	5
varFC	4
t ov	2

Table 3.3. Comparison between OtP and ReP method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP" is reported. Just the green entries are those in favor of the ReP method.

25 misclassified images			
ReP			
flag 1		flag 0	
number	4	number	16
margin gain	1	margin decrease	12
unFC	13763	unFC	11406
unTC	15027	/	/
varFC	9484	varFR	9348
varTC	9513	/	/
un ov	7182	/	/
var ov	2307	/	/
t ov	14551	t ov	8068

Table 3.4. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

wider margin than the original, while 12 flag 0 images out of 16 obtain a narrower margin than the original. Knowing the total number of pixels, the variability, uncertainty, and overlap areas are found to be very small. The largest area shown in Table 3.4 is "unTC = 15,027", which corresponds to 6% of the image's total area.

25 misclassified images			
OtP			
flag 1		flag 0	
number	4	number	16
margin gain	1	margin decrease	10
unFC	13419	unFC	11292
unTC	14738	/	/
varFC	9481	varFR	9249
varTC	9579	/	/
un ov	6989	/	/
var ov	2334	/	/
t ov	14361	t ov	7833

Table 3.5. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.5, it is observed that the OtP method makes 5 errors but is able to correct the predictions for 4 images, while 16 images have the GT class as the second most probable class in the method's softmax. Only 1 flag 1 image out of 4 manages to achieve a wider margin than the original, whereas 10 flag 0 images out of 16 obtain a narrower margin than the original. Given the total number of pixels in the image, the variability, uncertainty, and overlap areas are found to be very small. The largest area presented in Table 3.5 is "unTC = 14,738", which corresponds to 6% of the image's total area.

The ReP and OtP methods are compared, and the results are reported in Table 3.6. The table header continues to indicate the number of images misclassified by the neural network. It is verified whether the methods identify the same flag 0 and flag 1 images. Only the images common to both methods are taken into consideration. The margins are analyzed to identify which method has the highest number of flag 0 images with a narrower margin and the highest number of flag 1 images with a wider margin. The areas are compared to identify the method that has smaller areas of variability, uncertainty, and overlap on the heatmaps.

Table 3.6 illustrates how the ReP and OtP methods identify the same flag 0 and flag 1 images. Both methods exhibit the same number of errors, which is 5. The OtP method possesses wider margins for flag 1 images; however, no method demonstrates narrower margins for flag 0 images. Only the cells highlighted in green, which correspond to the number of images that satisfy the condition "ReP $<$ OtP", are in favor of the ReP method. In this case, the two methods identify an equal number of flag 0 and flag 1 images. All

25 misclassified images	
share flag 0	16
share flag 1	4
margin decrease OtP = ReP	8
margin gain OtP > ReP	3
flag 1 - ReP < OtP	
unFC	2
unTC	1
varFC	1
varTC	3
un ov	3
var ov	4
t ov	2
flag 0 - ReP < OtP	
unFC	7
varFC	8
t ov	3

Table 3.6. Comparison between OtP and ReP method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP" is reported. Just the green entries are those in favor of the ReP method, while the yellow entries are not decisive.

entries that report a number of images corresponding to "number of images flag/2" are highlighted in yellow and are not considered for the final comparison. Of the remaining 7 areas, 3 areas favor the ReP method.

The Grad-CAM 0.001 flag 1 images exhibit lower prediction uncertainty when analyzed using the OtP method. Neither method appears optimal for analyzing the margins of flag 0 images. The ReP method produces heatmaps that are more variable and uncertain compared to those generated by the OtP method.

3.2.3 Misclassified Score-CAM 0.005

Tables 3.7 and 3.8 include in their headers the number of misclassified images identified by the neural network and the method analyzed. The analysis is divided into flag 0 and flag 1 images, reporting only those images that adhere to the classification of each flag. For each flag, the number of images that reflect the behavior of the margins is presented, along with the dimensions of the various areas identifiable on the heatmaps.

In Table 3.7, it can be observed that the ReP method incurs 7 errors but is able to correct the predictions of 7 images, while 11 images possess the GT class as the second most probable class in the softmax output of the method. Only 4 flag 1 images out of 7 manage to achieve a margin wider than the original, whereas 8 flag 0 images out of 11 obtain a margin smaller than the original. Considering the total number of pixels, the

25 misclassified images			
ReP			
flag 1		flag 0	
number	7	number	11
margin gain	4	margin decrease	8
unFC	15230	unFC	12793
unTC	16212	/	/
varFC	14249	varFR	13688
varTC	14140	/	/
un ov	7417	/	/
var ov	4017	/	/
t ov	18718	t ov	9118

Table 3.7. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.7 is "t ov = 18,718", which corresponds to 7% of the total image area.

25 misclassified images			
OtP			
flag 1		flag 0	
number	7	number	11
margin gain	4	margin decrease	10
unFC	15006	unFC	11963
unTC	16167	/	/
varFC	14320	varFR	13418
varTC	14091	/	/
un ov	7677	/	/
var ov	3996	/	/
t ov	18111	t ov	8584

Table 3.8. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.8, it can be observed that the OtP method incurs 7 errors but is able to correct the predictions of 7 images, while 11 images possess the GT class as the second most probable class in the softmax output of the method. Only 4 flag 1 images out of 7 manage to achieve a margin wider than the original, whereas 10 flag 0 images out of

11 obtain a margin smaller than the original. Considering the total number of pixels in the image, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.8 is "t ov = 18,111", which corresponds to 7% of the total image area.

The ReP and OtP methods are compared, and the results are presented in Table 3.9. The header of the table continues to indicate the number of images misclassified by the neural network. It is also verified whether the methods identify the same flag 0 and flag 1 images. Only the images common to both methods are taken into consideration. The margins are analyzed by identifying which method has the greater number of flag 0 images with smaller margins and the greater number of flag 1 images with larger margins. The areas are compared, and the method that has smaller areas of variability, uncertainty, and overlap on the heatmaps is identified.

25 misclassified images	
share flag 0	11
share flag 1	7
margin decrease ReP $<$ OtP	6
margin gain ReP $>$ OtP	5
flag 1 - ReP $<$ OtP	
unFC	3
unTC	3
varFC	5
varTC	1
un ov	5
var ov	2
t ov	1
flag 0 - ReP $<$ OtP	
unFC	2
varFC	3
t ov	0

Table 3.9. Comparison between OtP and ReP method. The number of shared images, the method thas has the bes margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP $<$ OtP" is reported. Just the gree entries are those in favor of the ReP method.

Table 3.9 presents how the ReP and OtP methods identify the same flag 0 and flag 1 images. Both methods have the same number of errors, which amounts to 7. The ReP method exhibits the smallest margins for flag 0 images and the largest margins for flag 1 images. Only the cells highlighted in green, which correspond to the number of images satisfying the condition "ReP $<$ OtP", favor the ReP method. Only 2 out of 10 areas favor the ReP method.

The Score-CAM images at 0.005 exhibit low prediction uncertainty when analyzed by the ReP method; however, this method generates highly variable and uncertain heatmaps.

3.2.4 Misclassified Score-CAM 0.001 and discussion

Table 3.10 and Table 3.11 include in their headers the number of misclassifications identified by the neural network and the method analyzed. The analysis is divided into flag 0 and flag 1 images, presenting only the images that comply with the classification of each flag. For each flag, the number of images that reflect the behavior of the margins is reported, along with the size of the various areas identifiable on the heatmaps.

25 misclassified images			
ReP			
flag 1		flag 0	
number	6	number	15
margin gain	2	margin decrease	13
unFC	15032	unFC	12728
unTC	14690	/	/
varFC	12020	varFR	12585
varTC	10057	/	/
un ov	7048	/	/
var ov	2855	/	/
t ov	16290	t ov	8445

Table 3.10. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.10, it can be observed that the ReP method makes 4 errors but is able to correct the predictions of 6 images, while 15 images have the GT class as the second most probable class in the softmax of the method. Only 2 images with flag 1 out of 6 manage to achieve a margin wider than the original, whereas 13 images with flag 0 out of 15 obtain a margin smaller than the original. Given the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.10 is "t ov = 16,290", which corresponds to 6% of the total image area.

In Table 3.11, it can be observed that the OtP method makes 4 errors but is able to correct the predictions of 6 images, while 15 images have the GT class as the second most probable class in the softmax of the method. Only 2 images with flag 1 out of 6 manage to achieve a margin wider than the original, whereas 12 images with flag 0 out of 15 obtain a margin smaller than the original. Given the total number of pixels in the image, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.11 is "t ov = 16,399", which corresponds to 6% of the total image area.

The ReP and OtP methods are compared, and the results are presented in Table 3.12.

25 misclassified images			
OtP			
flag 1		flag 0	
number	6	number	15
margin gain	2	margin decrease	12
unFC	15228	unFC	12991
unTC	15045	/	/
varFC	12013	varFR	12420
varTC	10198	/	/
un ov	7672	/	/
var ov	2876	/	/
t ov	16399	t ov	8508

Table 3.11. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

The table header continues to indicate the number of misclassified images identified by the neural network. It is also verified whether the methods identify the same images with flags 0 and 1. Only the images common to both methods are taken into consideration. The margins are analyzed by identifying which method has the greater number of images with flag 0 and smaller margins, as well as the greater number of images with flag 1 and larger margins. The areas are compared, and the method that possesses the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.12, it can be observed that the ReP and OtP methods identify the same images with flags 0 and 1. Both methods have the same number of errors, which amounts to 4. The OtP method has smaller margins for images with flag 0; however, no method shows larger margins for images with flag 1. Only the cells highlighted in green, which correspond to the number of images that meet the condition "ReP < OtP", are in favor of the ReP method. Since both methods identify an equal number of images with flag 1, all entries in the table that report a number of images corresponding to "number of images with flag/2" are highlighted in yellow and are not considered for the final comparison. Of the remaining 7 areas, 4 areas favor the ReP method.

The Score-CAM 0.001 images with flag 0, when analyzed using the OtP method, exhibit smaller margins than the originals. However, no method achieves low prediction uncertainties for the images with flag 1. The ReP method is the one that provides more robust and less uncertain heatmaps.

The ReP method appears to be the most optimal for analyzing the misclassified images of the neural network. The image analysis method is capable of achieving better margin behavior for both flags, although the Grad-CAM 0.001 data show that the OtP method is more optimal. The Score-CAM 0.001 data are inconclusive regarding the margins: the OtP method obtains a greater number of images with flag 0 and smaller margins, but

25 misclassified images	
share flag 0	15
share flag 1	6
margin decrease OtP < ReP	8
margin gain OtP = ReP	3
flag 1 - ReP < OtP	
unFC	🟨 3
unTC	🟩 4
varFC	2
varTC	🟨 3
un ov	🟩 6
var ov	🟨 3
t ov	🟩 5
flag 0 - ReP < OtP	
unFC	🟩 8
varFC	7
t ov	6

Table 3.12. Comparison between OtP and ReP method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP" is reported. Just the green entries are those in favor of the ReP method, while the yellow entries are not decisive.

no method achieves larger margins for the images with flag 1. The heatmaps generated by this method are also robust and exhibit low uncertainty, as demonstrated by the Score-CAM 0.001 data. The Score-CAM 0.005 data are the only ones that favor the OtP method, while the Grad-CAM 0.005 and Grad-CAM 0.001 data indicate that the OtP method is optimal for only one additional entry compared to the ReP method.

By analyzing the correctly classified images, it is assessed whether this method remains the most optimal for manipulating the heatmaps and the softmax.

3.2.5 Correctly classified Grad-CAM 0.005

Tables 3.13, 3.14, 3.16, and 3.17 include in their headers the number of correctly classified images identified by the neural network and the analyzed method. The analysis is divided into images with flag 0 and flag 1, reporting only the images that conform to the classification of each flag. As stated in Section 3.1, the correctly classified images with flag 1 and flag 0 continue to maintain their original predictions.

Below the header, the number of images reflecting the behavior of the margins for each flag is reported: a correctly classified image with flag 0 may obtain a margin larger than the original margin, while a correctly classified image with flag 1 may obtain a margin smaller than the original. In the table, for flag 1, only cases where the final margin is greater than the initial margin are reported. The analysis continues by displaying the sizes of the various identifiable areas on the heatmaps, measured in pixels.

Regarding correctly classified images, the fields referring to the areas "unFC", "varFC", "un ov", and "var ov" cannot be filled for images that possess flag 0; in these cases, the symbol / will be reported in the table.

408 correctly classified images			
ReP			
flag 1		flag 0	
number	14	number	394
margin gain	3	margin gain	53
unFC	11996	/	/
unTC	11992	unTC	12813
varFC	9667	/	/
varTC	8985	varTC	10997
un ov	7350	/	/
var ov	4495	/	/
t ov	12836	t ov	9639

Table 3.13. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.13, it can be observed that the ReP method does not commit any errors; 394 images have flag 0, while 14 images have flag 1. Only 3 images with flag 1 out of 14 and 53 images with flag 0 out of 394 are able to achieve a margin greater than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.13 is "t ov = 12,836", which corresponds to 5% of the total area of the image.

In Table 3.14, it can be observed that the OtP method does not commit any errors; 394 images have flag 0, while 14 images have flag 1. Only 3 images with flag 1 out of 14 and 50 images with flag 0 out of 394 are able to achieve a better margin than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.14 is "t ov = 12,998", which corresponds to 5% of the total area of the image.

The ReP and OtP methods are compared, and the results are presented in Table 3.15. The table header continues to report the number of correctly classified images identified by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only images common to both methods are taken into consideration. The margins are analyzed by identifying which method has the greatest number of images with flag 0 and flag 1 that have wider margins. The areas are compared, and the method that possesses the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.15, it can be observed that the ReP and OtP methods identify the same

408 correctly classified images			
OtP			
flag 1		flag 0	
number	14	number	394
margin gain	3	margin gain	50
unFC	12397	/	/
unTC	12160	unTC	12961
varFC	9743	/	/
varTC	8997	varTC	11101
un ov	7586	/	/
var ov	4478	/	/
t ov	12998	t ov	9690

Table 3.14. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

408 correctly classified images	
share flag 0	394
share flag 1	14
margin gain OtP > ReP	215
margin gain OtP > ReP	11
flag 1 - ReP < OtP	
unFC	🟩 10
unTC	🟨 7
varFC	🟩 9
varTC	🟩 8
un ov	🟩 9
var ov	6
t ov	🟩 9
flag 0 - ReP < OtP	
unFC	🟩 211
varFC	🟩 223
t ov	🟩 201

Table 3.15. Comparison between OtP and ReP method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP" is reported. Just the green entries are those in favor of the ReP method, while the yellow entries are not decisive.

images with flag 0 and flag 1. Both methods do not commit any errors. The OtP method has larger margins compared to the ReP method for both flags, resulting in lower prediction uncertainty. Only the cells highlighted in green, corresponding to the number of images that meet the condition "ReP < OtP", are in favor of the ReP method. Since the two methods identify an equal number of images with flag 0 and flag 1, all entries in the table that report a number of images corresponding to "number of images flag/2" are highlighted in yellow and are not considered for the final comparison. Of the 9 remaining areas, 8 lean toward the ReP method.

The Grad-CAM 0.005 images exhibit low prediction uncertainty when analyzed using the OtP method. However, it is the ReP method that generates more robust and less variable heatmaps.

408 correctly classified images			
ReP-A			
flag 1		flag 0	
number	10	number	394
margin gain	1	margin gain	52
unFC	11476	/	/
unTC	12117	unTC	12744
varFC	9887	/	/
varTC	9361	varTC	10841
un ov	6484	/	/
var ov	4551	/	/
sovrap	12917	sovrap	9543

Table 3.16. Analysis of ReP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.16, it can be observed that the ReP-A method commits 4 errors, with 394 images having flag 0 and 10 images having flag 1. Only 1 image with flag 1 out of 10 and 52 images with flag 0 out of 394 achieve a better margin than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.16 is "t ov = 12,917", which corresponds to 5% of the total area of the image.

In Table 3.17, it can be observed that the OtP-A method commits 4 errors, with 394 images having flag 0 and 10 images having flag 1. Only 2 images with flag 1 out of 10 and 55 images with flag 0 out of 394 achieve a better margin than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.17 is "unTC = 12,647", which corresponds to 5% of the total area of the image.

The ReP-A and OtP-A methods are compared, and the results are presented in Table 3.18. The table header continues to indicate the number of images correctly classified

408 correctly classified images			
OtP-A			
flag 1		flag 0	
number	10	number	394
margin gain	2	margin gain	55
unFC	10566	/	/
unTC	11063	unTC	12647
varFC	9091	/	/
varTC	8672	varTC	10774
un ov	6069	/	/
var ov	4145	/	/
sovrap	11815	sovrap	9456

Table 3.17. Analysis of OtP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only the images common to both methods are taken into consideration. The margins are analyzed by identifying which method has the greatest number of images with flag 0 and flag 1 that have the widest margins. The areas are compared, and the method with the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.18, it can be observed that the ReP-A and OtP-A methods identify the same images with flag 0 but share only 9 images with flag 1 out of the 10 identified by both methods. The two methods commit the same number of errors, which is 4. The ReP-A method has larger margins compared to the OtP-A method, resulting in lower prediction uncertainty. Only the cells highlighted in green, corresponding to the number of images that meet the condition "ReP-A < OtP-A", support the ReP-A method. Only 1 area out of 10 favors the ReP-A method.

The Grad-CAM 0.005 images exhibit low prediction uncertainty when analyzed by the ReP-A method; however, the most stable and least uncertain heatmaps are provided by the OtP-A method.

3.2.6 Correctly classified Grad-CAM 0.001

Tables 3.19, 3.20, 3.22, and 3.23 present in the header the number of correctly classified images identified by the neural network and the method analyzed. The analysis is divided into images with flag 0 and flag 1, reporting only the images that conform to the classification of each flag.

Below the header, the number of images with flag 0 and flag 1 that reflect the behavior of the margins for each flag is reported. The analysis continues by showing the extent of

408 correctly classified images	
share flag 0	394
share flag 1	9
margin gain ReP-A > OtP-A	234
margin gain ReP-A > OtP-A	8
flag 1 - ReP-A < OtP-A	
unFC	4
unTC	3
varFC	3
varTC	2
un ov	4
var ov	3
t ov	4
flag 0 - ReP-A < OtP-A	
unFC	200
varFC	196
t ov	193

Table 3.18. Comparison between OtP-A and ReP-A method. The number of share images, the method that has the bes margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP-A < OtP-A" is reported. Just the green entries are those in favor of the ReP-A method.

the various identifiable areas on the heatmaps.

In Table 3.19, it can be observed that the ReP method makes no errors; 401 images possess flag 0, while 7 images possess flag 1. Only 5 images with flag 1 out of 7 and 95 images with flag 0 out of 401 manage to achieve a margin greater than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.19 is "unTC = 12,281", which corresponds to 5% of the total image area.

In Table 3.20, it can be observed that the OtP method makes no errors; 401 images possess flag 0, while 7 images possess flag 1. Only 4 images with flag 1 out of 7 and 97 images with flag 0 out of 401 are able to achieve a margin greater than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.20 is "unTC = 12,308", which corresponds to 5% of the total image area.

The ReP and OtP methods are compared, and the results are presented in Table 3.21. The table header continues to indicate the number of images correctly classified by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only the images common to both methods are taken into consideration. The margins are analyzed to identify which method possesses the greater number of images with flag 0 and flag 1 that have the widest margins. The areas are compared, and the

408 correctly classified images			
ReP			
flag 1		flag 0	
number	7	number	401
margin gain	5	margin gain	95
unFC	9709	/	/
unTC	10402	unTC	12281
varFC	7715	/	/
varTC	6206	varTC	9762
un ov	6244	/	/
var ov	3707	/	/
sovrap	10395	sovrap	8626

Table 3.19. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

408 correctly classified images			
OtP			
flag 1		flag 0	
number	7	number	401
margin gain	4	margin gain	97
unFC	9789	/	/
unTC	10299	unTC	12308
varFC	7727	/	/
varTC	6141	varTC	9787
un ov	6288	/	/
var ov	3714	/	/
t ov	10348	t ov	8640

Table 3.20. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

method that possesses the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.21, it can be observed that the ReP and OtP methods identify the same images with flag 0 and flag 1. Both methods make no errors. The OtP method possesses greater margins compared to the ReP method, resulting in lower prediction uncertainty. Only the cells highlighted in green, corresponding to the number of images that meet

408 correctly classified images	
share flag 0	401
share flag 1	7
margin gain OtP > ReP	204
margin gain OtP > ReP	4
flag 1 - ReP < OtP	
unFC	4
unTC	3
varFC	4
varTC	3
un ov	3
var ov	4
t ov	2
flag 0 - ReP < OtP	
unFC	208
varFC	211
t ov	211

Table 3.21. Comparison between OtP and ReP method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP" is reported. Just the green entries are those in favor of the ReP method.

the condition "ReP < OtP", favor the ReP method. In this case, 6 areas out of 10 lean towards the ReP method.

The Grad-CAM 0.001 images exhibit lower prediction uncertainty when analyzed using the OtP method; however, the method that generates more robust and less uncertain heatmaps is the ReP method.

In Table 3.22, it can be observed that the ReP-A method makes one error, with 401 images having flag 0 and 6 images having flag 1. Only 2 images with flag 1 out of 6 and 106 images with flag 0 out of 401 manage to achieve a margin greater than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.22 is "unTC = 12,285", which corresponds to 5% of the total area of the image.

In Table 3.23, it can be observed that the OtP-A method makes one error, with 401 images having flag 0 and 6 images having flag 1. Only 3 images with flag 1 out of 6 and 120 images with flag 0 out of 401 manage to achieve a margin greater than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.23 is "unTC = 12,334", which corresponds to 5% of the total area of the image.

The ReP-A and OtP-A methods are compared, and the results are presented in Table 3.24. The table header continues to indicate the number of images correctly classified

408 correctly classified images			
ReP-A			
flag 1		flag 0	
number	6	number	401
margin gain	2	margin gain	106
unFC	8917	/	/
unTC	11813	unTC	12285
varFC	6388	/	/
varTC	8826	varTC	9703
un ov	6068	/	/
var ov	3762	/	/
t ov	11235	t ov	8586

Table 3.22. Analysis of ReP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

408 correctly classified images			
OtP-A			
flag 1		flag 0	
number	6	number	401
margin gain	3	margin gain	120
unFC	9136	/	/
unTC	11962	unTC	12334
varFC	6387	/	/
varTC	8771	varTC	9750
un ov	6454	/	/
var ov	3622	/	/
t ov	11499	t ov	8541

Table 3.23. Analysis of OtP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only the images common to both methods are considered. The margins are analyzed to identify which method has the largest number of images with flag 0 and flag 1 that exhibit the widest margins. The areas are compared, and the method with the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

408 correctly classified images	
share flag 0	401
share flag 1	6
margin gain ReP-A > OtP-A	225
margin gain OtP-A > ReP-A	6
flag 1 - ReP-A < OtP-A	
unFC	🟩 4
unTC	🟩 5
varFC	🟨 3
varTC	🟩 4
un ov	🟩 4
var ov	🟨 3
t ov	🟩 4
flag 0 - ReP-A < OtP-A	
unFC	🟩 210
varFC	🟩 215
t ov	193

Table 3.24. Comparison between OtP-A and ReP-A method. The number of share images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP-A < OtP-A" is reported. Just the green entries are those in favour of the ReP-A method, while the yellow entries are not decisive.

In Table 3.24, it can be observed that the ReP-A and OtP-A methods identify the same images with flags 0 and 1. Both methods make only one error. Neither method possesses the best margins for both flag 0 and flag 1 images. Only the cells highlighted in green, corresponding to the number of images that satisfy the condition "ReP-A < OtP-A", favor the ReP-A method. Since the methods identify an equal number of images with flags 0 and 1, all entries that report a number of images corresponding to "number of images with flag/2" are highlighted in yellow and are not considered in the final comparison. Among the remaining 8 areas, 7 favor the ReP-A method.

The Grad-CAM 0.001 images exhibit more stable and less uncertain heatmaps when analyzed using the ReP-A method. However, the ReP-A method achieves low prediction uncertainty only for the images with flag 0, while the performance of the margins for flag 1 is better when using the OtP-A method.

3.2.7 Correctly classified Score-CAM 0.005

Tables 3.25, 3.26, 3.28, and 3.29 present in their headers the number of correctly classified images identified by the neural network and the method analyzed. The analysis is divided into flag 0 and flag 1 images, reporting only those images that conform to the classification of each flag.

Below the header, the number of flag 0 and flag 1 images that reflects the behavior of the margins is provided. The analysis continues by showing the extent of the various identifiable areas on the heatmaps.

408 correctly classified images			
ReP			
flag 1		flag 0	
number	16	number	392
margin gain	6	margin gain	35
unFC	14158	/	/
unTC	14061	unTC	13590
varFC	13421	/	/
varTC	13430	varTC	13121
un ov	6070	/	/
var ov	3718	/	/
t ov	15878	t ov	9262

Table 3.25. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.25, it can be observed that the ReP method makes no errors; 392 images have flag 0, while 16 images have flag 1. Only 6 images with flag 1 out of 16 and 35 images with flag 0 out of 392 manage to achieve a margin greater than the original. Given the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area presented in Table 3.25 is "t ov = 15,878", which corresponds to 6% of the total area of the image.

In Table 3.26, it can be observed that the OtP method makes no errors; 392 images have flag 0, while 16 images have flag 1. Only 8 images with flag 1 out of 16 and 36 images with flag 0 out of 392 manage to achieve a margin greater than the original. Given the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area presented in Table 3.26 is "t ov = 15,916", which corresponds to 6% of the total area of the image.

The ReP and OtP methods are compared, and the results are reported in Table 3.27. The table header continues to indicate the number of images correctly classified by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only the images common to both methods are considered. The margins are analyzed to identify the method that has the highest number of images with flag 0 and flag 1 with the widest margins. The areas are compared, and the method with the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.27, it can be observed that the ReP and OtP methods identify the same images with flag 0 and flag 1. Both methods make no errors. The OtP method has wider margins compared to the ReP method, resulting in lower prediction uncertainty for both

408 correctly classified images			
OtP			
flag 1		flag 0	
number	16	number	392
margin gain	8	margin gain	36
unFC	14353	/	/
unTC	14166	unTC	13674
varFC	13531	/	/
varTC	13544	varTC	13135
un ov	6182	/	/
var ov	3828	/	/
t ov	15916	t ov	9278

Table 3.26. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

408 correctly classified images	
share flag 0	392
share flag 1	16
margin gain OtP > ReP	214
margin gain OtP > ReP	9
flag 1 - ReP < OtP	
unFC	11
unTC	12
varFC	11
varTC	9
un ov	9
var ov	14
t ov	10
flag 0 - ReP < OtP	
unFC	215
varFC	203
t ov	190

Table 3.27. Comparison between OtP and ReP method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP" is reported. Just the green entries are those in favour of the ReP method

flags. Only the cells highlighted in green, which correspond to the number of images that meet the condition "ReP < OtP", favor the ReP method. In this case, 9 out of 10 areas lean toward the ReP method.

The Score-CAM images with a probability of 0.005 exhibit lower prediction uncertainty when analyzed using the OtP method; however, this method generates more variable and uncertain heatmaps than the ReP method.

408 correctly classified images			
ReP-A			
flag 1		flag 0	
number	13	number	392
margin gain	4	margin gain	37
unFC	14503	/	/
unTC	13896	unTC	13745
varFC	13189	/	/
varTC	13907	varTC	13091
un ov	5897	/	/
var ov	3548	/	/
t ov	15765	t ov	9311

Table 3.28. Analysis of ReP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.28, it can be observed that the ReP-A method makes 3 errors; 392 images have flag 0, while 13 have flag 1. Only 4 images with flag 1 out of 13 and 37 images with flag 0 out of 392 manage to achieve a wider margin than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area presented in Table 3.28 is "t ov = 15,765", which corresponds to 6% of the total area of the image.

In Table 3.29, it can be observed that the OtP-A method makes 2 errors; 392 images have flag 0, while 14 have flag 1. Only 3 images with flag 1 out of 14 and 30 images with flag 0 out of 392 are able to achieve a wider margin than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area presented in Table 3.29 is "t ov = 14,665", which corresponds to 6% of the total area of the image.

The ReP-A and OtP-A methods are compared, and the results are reported in Table 3.30. The table header continues to report the number of images correctly classified by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only the images common to both methods are taken into consideration. The margins are analyzed by identifying the method that has the highest number of images with flag 0 and flag 1 with the widest margins. The areas are compared, and the method that possesses the smallest areas of variability, uncertainty, and overlap on

408 correctly classified images			
OtP-A			
flag 1		flag 0	
number	14	number	392
margin gain	3	margin gain	30
unFC	14546	/	/
unTC	12838	unTC	13337
varFC	12778	/	/
varTC	13133	varTC	12955
un ov	6280	/	/
var ov	3023	/	/
sovrap	14665	sovrap	9090

Table 3.29. Analysis of OtP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

the heatmaps is identified.

In Table 3.30, it can be observed that the ReP-A and OtP-A methods do not identify the same images with flag 0 and flag 1; the OtP-A method identifies one additional image with flag 1 compared to the ReP-A method. The OtP-A method makes 2 errors, while the ReP-A method makes 3 errors. The ReP-A method has a lower prediction uncertainty solely for the images with flag 0, whereas the OtP-A method has lower prediction uncertainties for the images with flag 1. The cells highlighted in green, which correspond to the number of images that meet the condition "ReP-A < OtP-A", favor the ReP-A method. In this case, only 2 out of 10 areas lean towards the ReP-A method.

The Score-CAM images with 0.005 dropout probability exhibit more stable and less uncertain heatmaps when analyzed using the OtP-A method. This method also has low prediction uncertainty solely for the images with flag 1, but not for those with flag 0.

3.2.8 Correctly classified Score-CAM 0.001 and discussion

Tables 3.31, 3.32, 3.34, and 3.35 present in their headers the number of correctly classified images identified by the neural network and the analyzed method. The analysis is divided into images with flag 0 and flag 1, reporting the number of images that meet the classification for each flag.

Below the header, the number of images with flag 0 and flag 1 is provided, reflecting the behavior of the margins. The analysis continues by showing the extent of the various identifiable areas on the heatmaps.

In Table 3.31, it can be observed that the ReP method makes no errors; 401 images have flag 0, while 7 images have flag 1. Only 5 out of 7 images with flag 1 and 113 out of 401 images with flag 0 are able to achieve a margin wider than the original. Knowing the

408 correctly classified images	
share flag 0	392
share flag 1	13
margin gain ReP-A > OtP-A	265
margin gain OtP-A > ReP-A	7

flag 1 - ReP-A < OtP-A	
unFC	8
unTC	3
varFC	6
varTC	3
un ov	8
var ov	2
t ov	2

flag 0 - ReP-A < OtP-A	
unFC	153
varFC	178
t ov	166

Table 3.30. Comparison between OtP-A and ReP-A method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP-A < OtP-A" is reported. Just the green entries are those in favour of the ReP-A method.

408 correctly classified images			
ReP			
flag 1		flag 0	
number	7	number	401
margin gain	5	margin gain	113
unFC	14046	/	/
unTC	13043	unTC	13272
varFC	10728	/	/
varTC	13990	varTC	11877
un ov	4918	/	/
var ov	2909	/	/
t ov	14699	t ov	8883

Table 3.31. Analysis of ReP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.31 is "t ov = 14,699", which corresponds to 6% of the total area of the image.

408 correctly classified images			
OtP			
flag 1		flag 0	
number	7	number	401
margin gain	5	margin gain	116
unFC	13980	/	/
unTC	13153	unTC	13278
varFC	10713	/	/
varTC	14076	varTC	11873
un ov	4935	/	/
var ov	2826	/	/
t ov	14787	t ov	8877

Table 3.32. Analysis of OtP method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.32, it can be observed that the OtP method makes no errors; 401 images have flag 0, while 7 images have flag 1. Only 5 out of 7 images with flag 1 and 116 out of 401 images with flag 0 are able to achieve a margin wider than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are found to be very small. The largest area reported in Table 3.32 is "t ov = 14,787", which corresponds to 6% of the total area of the image.

The ReP and OtP methods are compared, and the results are presented in Table 3.33. The header of the table continues to indicate the number of images correctly classified by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only the images common to both methods are taken into consideration. The margins are analyzed by identifying the method that possesses the greatest number of images with flags 0 and 1 that have the widest margins. The areas are compared, and the method that has the smaller areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.33, it can be observed that the ReP and OtP methods identify the same images with flags 0 and 1. Both methods make no errors. Neither of the two methods exhibits better margins for both images with flags 0 and 1. Only the cells highlighted in green, which correspond to the number of images that meet the condition "ReP < OtP", support the ReP method. In this case, only one area out of ten favors the ReP method.

The Score-CAM images with 0.001 dropout probability exhibit more robust and less uncertain heatmaps when analyzed using the OtP method. This method also has lower prediction uncertainty, but only for images with flag 0.

408 correctly classified images	
share flag 0	401
share flag 1	7
margin gain OtP > ReP	205
margin gain ReP > OtP	4
flag 1 - ReP < OtP	
unFC	3
unTC	3
varFC	3
varTC	5
un ov	3
var ov	1
t ov	3
flag 0 - ReP < OtP	
unFC	199
varFC	200
t ov	196

Table 3.33. Comparison between OtP and ReP method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP" is reported. Just the green entries are those in favour of the ReP method.

408 correctly classified images			
ReP-A			
flag 1		flag 0	
number	6	number	401
margin gain	2	margin gain	111
unFC	13566	/	/
unTC	12854	unTC	13274
varFC	11358	/	/
varTC	13171	varTC	11847
un ov	5347	/	/
var ov	2697	/	/
t ov	13824	t ov	8885

Table 3.34. Analysis of ReP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.34, it can be observed that the ReP-A method commits one error, with 401 images having flag 0, while 6 images have flag 1. Only 2 images with flag 1 out of 6 and 111 images with flag 0 out of 401 manage to achieve a margin wider than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area that appears in Table 3.34 is "t ov = 13,824", which corresponds to 5% of the total area of the image.

408 correctly classified images			
OtP-A			
flag 1		flag 0	
number	6	number	401
margin gain	1	margin gain	114
unFC	13331	/	/
unTC	11563	unTC	13235
varFC	10713	/	/
varTC	13201	varTC	11889
un ov	4661	/	/
var ov	2285	/	/
sovrap	13063	sovrap	8611

Table 3.35. Analysis of OtP-A method. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.35, it can be noted that the OtP-A method commits one error, with 401 images having flag 0, while 6 images have flag 1. Only 1 image with flag 1 out of 6 and 114 images with flag 0 out of 401 manage to achieve a margin wider than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area that appears in Table 3.35 is "unFC = 13,331", which corresponds to 5% of the total area of the image.

The ReP-A and OtP-A methods are compared, and the results are presented in Table 3.36. The table header continues to report the number of images correctly classified by the neural network. It is also verified whether the methods identify the same images with flag 0 and flag 1. Only images common to both methods are taken into consideration. The margins are analyzed to identify which method has the highest number of images with flags 0 and 1 that exhibit the widest margins. The areas are compared, and the method that has the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.36, it can be observed that the ReP-A and OtP-A methods do not identify the same images with flag 1; both methods differently identify one image with flag 1 out of the six identified. Both methods commit one error. The ReP-A method has the best margins for both flag 0 and flag 1 images, and therefore exhibits the lowest prediction uncertainty. The cells highlighted in green, corresponding to the number of images that

408 correctly classified images	
share flag 0	401
share flag 1	5
margin gain ReP-A > OtP-A	230
margin gain ReP-A > OtP-A	4
flag 1 - ReP-A < OtP-A	
unFC	3
unTC	0
varFC	3
varTC	1
un ov	3
var ov	3
t ov	0
flag 0 - ReP-A < OtP-A	
unFC	191
varFC	200
t ov	141

Table 3.36. Comparison between OtP-A and ReP-A method. The number of share images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP-A < OtP-A" is reported. Just the green entries are those in favour of the ReP-A method.

satisfy the condition "ReP-A < OtP-A", favor the ReP-A method. In this case, four areas out of ten lean towards the ReP-A method.

The Score-CAM 0.001 images exhibit low prediction uncertainty when analyzed by the ReP-A method; however, the generated heatmaps are more variable and less certain than those generated by the OtP-A method.

The ReP method produces the most stable and least uncertain heatmaps; only the Score-CAM 0.001 data indicate that the OtP method achieves the most stable and least uncertain heatmaps. Although the ReP method is robust in constructing heatmaps, it generates softmax outputs with high prediction uncertainties. Only the Score-CAM 0.001 data demonstrate that the ReP method is the most optimal for analyzing flag 1 images. Both the ReP and OtP methods do not commit any errors.

The OtP-A method, like the ReP method, generates the most stable heatmaps. Only the Grad-CAM 0.001 data indicate that the ReP-A method produces the most stable and least uncertain heatmaps. Despite the OtP-A method being robust in constructing heatmaps, it generates softmax outputs with high prediction uncertainties. Only the Grad-CAM 0.001 data show that the OtP-A method is the most optimal for analyzing flag 1 images. Both ReP-A and OtP-A methods commit errors; however, the ReP-A method commits one additional error compared to the OtP-A method in the Score-CAM 0.005 data.

Having identified the ReP and OtP-A methods as the two optimal methods for analyzing correctly classified images, a final comparison between these two methods is conducted, with results reported in Tables 3.37, 3.38, 3.39, and 3.40. The headers of the tables present the number of correctly classified images identified by the neural network, the XAI method used, and the dropout probability. The analysis is focused solely on the flag 0 and flag 1 images that the methods have in common, determining which method possesses lower prediction uncertainty and more contained areas of variability, uncertainty, and overlap.

408 correctly classified images	
Grad-CAM 0.005	
share flag 0	394
share flag 1	10
margin gain ReP > OtP-A	228
margin gain ReP > OtP-A	9
flag 1 - ReP < OtP-A	
unFC	4
unTC	4
varFC	5
varTC	3
un ov	2
var ov	3
t ov	4
flag 0 - ReP < OtP-A	
unFC	190
varFC	168
t ov	177

Table 3.37. Comparison between OtP-A and ReP method. The number of share images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP-A" is reported. The yellow entries are not decisive.

The results presented in Table 3.37 demonstrate that both methods identify the same number of flag 0 images, but do not share the same number of flag 1 images. The ReP method identifies four additional flag 1 images, which are considered errors by the OtP-A method. The ReP method achieves lower prediction uncertainty; however, it generates more variable and uncertain heatmaps. The cell highlighted in yellow corresponds to the only area not considered for the final comparison, as this entry indicates a number of images that corresponds to "number of images flag/2". Of the nine remaining areas, none lean towards the ReP method.

The results presented in Table 3.38 indicate that both methods identify the same number of flag 0 images, but do not share the same number of flag 1 images. The ReP

408 correctly classified images	
Grad-CAM 0.001	
share flag 0	401
share flag 1	6
margin gain ReP > OtP-A	227
margin gain OtP-A > ReP	4
flag 1 - ReP < OtP-A	
unFC	3
unTC	4
varFC	3
varTC	3
un ov	3
var ov	1
t ov	5
flag 0 - ReP < OtP-A	
unFC	205
varFC	183
t ov	166

Table 3.38. Comparison between OtP-A and ReP method. The number of share images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP-A" is reported. The green entries are those in favour of the ReP method, while the yellow entries are not decisive.

method, in fact, identifies one additional flag 1 image, which is considered an error by the OtP-A method. The ReP method achieves lower prediction uncertainty for flag 0 images, while the OtP-A method exhibits lower prediction uncertainty for flag 1 images. Since both methods identify an equal number of flag 0 images, the entries in the table that correspond to "number of images flag/2" are highlighted in yellow and are not considered for the final comparison. Only the cells highlighted in green indicate the entries of the areas that favor the ReP method. Among the six remaining areas, both methods exhibit the same performance, generating heatmaps that are minimally variable and uncertain.

The results presented in Table 3.39 indicate that both methods identify the same number of flag 0 images, but do not share the same number of flag 1 images. The ReP method, in fact, identifies two additional flag 1 images, which are considered errors by the OtP-A method. The ReP method achieves lower prediction uncertainty for both flag 0 and flag 1 images. Since both methods identify an equal number of flag 0 and flag 1 images, the entries in the table that correspond to "number of images flag/2" are highlighted in yellow and are not considered for the final comparison. Only the cells highlighted in green indicate the entries of the areas that favor the ReP method. Among the eight remaining areas, only one favors the ReP method.

The results presented in Table 3.40 indicate that both methods identify the same

408 correctly classified images	
Score-CAM 0.005	
share flag 0	392
share flag 1	14
margin gain ReP > OtP-A	262
margin gain ReP > OtP-A	11
flag 1 - ReP < OtP-A	
unFC	7
unTC	2
varFC	6
varTC	7
un ov	8
var ov	4
t ov	6
flag 0 - ReP < OtP-A	
unFC	158
varFC	170
t ov	174

Table 3.39. Comparison between OtP-A and ReP method. The number of share images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP-A" is reported. The green entries are those in favour of the ReP method, while the yellow entries are not decisive.

number of flag 0 images, but do not share the same number of flag 1 images. The ReP method, in fact, identifies one additional flag 1 image, which is considered an error by the OtP-A method. The ReP method achieves low prediction uncertainty for both flag 0 and flag 1 images; however, it generates more variable and uncertain heatmaps. Since both methods identify an equal number of flag 0 and flag 1 images, the entries in the table that correspond to "number of images flag/2" are highlighted in yellow and are not considered for the final comparison. Among the six remaining areas, none favor the ReP method.

The ReP method generates softmax outputs with low prediction uncertainties. Only the Grad-CAM 0.001 data suggest that the ReP method is not optimal for analyzing flag 1 images. Although the method produces solid softmax outputs, the heatmaps generated by this method are more unstable and uncertain than those generated by the OtP-A method. The Grad-CAM 0.005 and Score-CAM 0.001 data show that none of the entries in the table favor the ReP method. The Score-CAM 0.005 data indicate that only one entry in the table favors the ReP method, while the Grad-CAM 0.001 data are inconclusive since both methods exhibit the same performance. Although the ReP method generates more variable and uncertain heatmaps, it does not commit any errors. All flag 0 and flag 1 images conform to their respective classifications, whereas the OtP-A method consistently

408 correctly classified images	
Score-CAM 0.001	
share flag 0	401
share flag 1	6
margin gain ReP > OtP-A	220
margin gain ReP > OtP-A	5
flag 1 - ReP < OtP-A	
unFC	2
unTC	1
varFC	3
varTC	1
un ov	3
var ov	3
t ov	3
flag 0 - ReP < OtP-A	
unFC	197
varFC	194
t ov	152

Table 3.40. Comparison between OtP-A and ReP method. The number of share images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "ReP < OtP-A" is reported. The yellow entries are not decisive.

commits errors.

As was identified for the misclassified images, the ReP method also proves to be the most optimal for analyzing correctly classified images.

3.3 Second objective

Having identified the ReP method as the most optimal for analyzing images, the next step is to determine which XAI method and what dropout probability allow the XCiT network to achieve the best performance. The analysis proceeds by identifying the optimal dropout probability for each XAI method. Once the dropout probability is identified, the corresponding XAI method to be used with the XCiT neural network is determined. The results are presented in tables that indicate the XAI method under analysis and whether the results pertain to misclassified or correctly classified images. Only the images common to both probabilities are taken into consideration. The probability that exhibits the highest number of flag 0 and flag 1 images, reflecting the behavior of each flag as described in Section 3.1, is identified, along with the probability that generates heatmaps with contained variability, uncertainty, and overlap areas.

The organization of the tables remains unchanged even when the dropout probability is identified and the XAI methods are compared against one another. The analysis is presented by separating the misclassified images from the correctly classified ones.

3.3.1 Validation set misclassified images

Tables 3.41 and 3.42 present the comparison between the probabilities of 0.001 and 0.005 for the Grad-CAM and Score-CAM methods, respectively. The analyzed method is indicated in the table header. The probability that offers optimal margins is the one that enables the identification of a high number of flag 0 images with a final margin lower than the initial margin, as well as a high number of flag 1 images with a final margin higher than the initial margin.

25 misclassified images	
Grad-CAM	
share flag 0	9
share flag 1	4
margin decrease 0.005 < 0.001	9
margin gain 0.005 > 0.001	4
flag 1 - 0.005 < 0.001	
unFC	0
unTC	2
varFC	1
varTC	1
un ov	2
var ov	0
t ov	1
flag 0 - 0.005 < 0.001	
unFC	4
varFC	5
t ov	4

Table 3.41. Comparison between 0.005 and 0.001 dropout probabilities for Grad-CAM XAI method. The number of share images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "0.005 < 0.001" is reported. The green entry is in favour of the 0.005 probability, while the yellow entries are not decisive.

The results presented in Table 3.41 indicate that the same method does not identify the same number of flag 0 and flag 1 images. When the probability of 0.001 is used, more flag 0 images are identified - images that have the GT as the second most probable class in the softmax output of the ReP method. However, this probability identifies fewer flag 1 images, which means fewer images are classified with the GT as the final classification. The probability of 0.005 allows for the identification of a higher number of flag 1 images;

however, this probability incurs four more errors than the 0.001 probabilty. The 0.005 probability successfully identifies more flag 0 images with final margins lower than the initial margins, as well as more flag 1 images with final margins higher than the initial margins. Since both probabilities identify an equal number of flag 1 images, any entries in the table that report a number of images corresponding to "number of flag images/2" are highlighted in yellow and not considered in the final comparison. Only the cell highlighted in green indicates a preference for the 0.005 method. Among the remaining eight areas, only one area shows that the 0.005 probability generates heatmaps that are less variable and uncertain.

25 misclassified images	
Score-CAM	
share flag 0	10
share flag 1	5
margin decrease 0.005 < 0.001	8
margin gain 0.005 > 0.001	4
flag 1 - 0.005 < 0.001	
unFC	2
unTC	1
varFC	1
varTC	0
un ov	3
var ov	2
t ov	0
flag 0 - 0.005 < 0.001	
unFC	6
varFC	5
t ov	4

Table 3.42. Comparison between 0.005 and 0.001 dropout probabilities for Score-CAM XAI method. The number of share images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "0.005 < 0.001" is reported. The green entries are in favour of the 0.005 probability, whle the yello entries are not decisive.

The results presented in Table 3.42 indicate that the same method does not identify the same number of flag 0 and flag 1 images. When the probability of 0.001 is used, more flag 0 images are identified, but fewer flag 1 images. The probabilty of 0.005 is able to correct the prediction of one additional image compared to the probability of 0.001; however, it incurs three more errors than the probability of 0.001. Since both probabilities identify an equal number of flag 0 images, any entries in the table that report a number of images corresponding to "number of flag images/2" are highlighted in yellow and not considered in the final comparison. Only the cells highlighted in green favor the 0.005

probability. Among the remaining nine areas, only two identify the 0.005 probability as the one that generates heatmaps that are less variable and uncertain.

The dropout probability of 0.001 applied to each XAI method allows the ReP method to make fewer errors, even though it generates softmax outputs with high prediction uncertainties. The probability of 0.001 is the one that produces the most stable and least uncertain heatmaps.

Having identified the dropout probability of 0.001 as the one with the best performance for each XAI method, the two XAI methods are compared to determine which is more suitable for use with the XCiT network. The comparison between Score-CAM 0.001 and Grad-CAM 0.001 is presented in Table 3.43.

25 misclassified images	
share flag 0	15
share flag 1	4
margin decrease Grad-CAM < Score-CAM	9
margin gain Grad-CAM = Score-CAM	2
flag 1 - Score-CAM < Grad-CAM	
unFC	2
unTC	2
varFC	2
varTC	2
un ov	3
var ov	2
t ov	2
flag 0 - Score-CAM < Grad-CAM	
unFC	6
varFC	5
t ov	7

Table 3.43. Comparison between Score-CAM and Grad-CAM methods with 0.001 dropout probability. The number of share images, the XAI method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "Score-CAM < Grad-CAM" is reported. The green entries are those in favour of the Score-CAM method, while the yellow entries are not decisive

The results presented in Table 3.43 indicate that each XAI method does not identify the same number of flag 0 and flag 1 images. The Grad-CAM 0.001 method identifies one additional flag 0 image compared to the Score-CAM 0.001 method, thus identifying one more image that has the GT class as the second most probable class in the softmax output of the ReP method. In contrast, the Score-CAM 0.001 method identifies two additional flag 1 images, allowing it to correct the predictions of more images than the Grad-CAM 0.001 method. Furthermore, the Grad-CAM method incurs one more error than the Score-CAM method.

The margin analysis shows that the Grad-CAM method obtains more flag 0 images with final margins lower than the initial ones. However, both XAI methods exhibit the same performance on the softmax outputs for flag 1 images. Since both XAI methods identify an equal number of flag 1 images, any entries in the table that report a number of images corresponding to "number of flag images/2" are highlighted in yellow and not considered in the final comparison. Only the cells highlighted in green favor the Score-CAM method. Among the remaining four areas, only one leans towards the Score-CAM method.

3.3.2 Validation set correctly classified images

Tables 3.44 and 3.45 present a comparison between the probabilities 0.001 and 0.005 for the Grad-CAM and Score-CAM methods, respectively. The analyzed method is indicated in the header of each table. The probability that possesses the optimal margins is the one that allows for the identification of a high number of flag 0 images with final margins lower than the initial margins, as well as a high number of flag 1 images with final margins higher than the initial margins.

408 correctly classified images	
Grad-CAM	
share flag 0	392
share flag 1	5
margin gain 0.001 > 0.005	332
margin gain 0.001 > 0.005	3
flag 1 - 0.005 < 0.001	
unFC	0
unTC	2
varFC	1
varTC	1
un ov	1
var ov	0
t ov	1
flag 0 - 0.005 < 0.001	
unFC	161
varFC	127
t ov	131

Table 3.44. Comparison between 0.005 and 0.001 dropout probabilities for Grad-CAM XAI method. The number of share images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "0.005 < 0.001" is reported.

The results presented in Table 3.44 demonstrate that the same method does not identify the same number of flag 0 and flag 1 images. When the probability 0.001 is used, more flag 0 images are identified, indicating that more images continue to be classified within the GT class even when the MCD method is applied. However, it identifies a lower number of flag 1 images, resulting in fewer images being misclassified when the MCD method is applied. Neither of the probabilities commits any errors. The probability 0.001 is capable of achieving final margins that exceed the initial margins for flag 0 images but does not achieve the same performance for flag 1 images. Among the total of 10 areas, none identifies the probability 0.005 as the one that generates heatmaps with low variability and uncertainty.

408 correctly classified images	
Score-CAM	
share flag 0	390
share flag 1	5
margin gain 0.001 > 0.005	344
margin gain 0.001 > 0.005	3
flag 1 - 0.005 < 0.001	
unFC	2
unTC	2
varFC	1
varTC	4
un ov	1
var ov	2
t ov	2
flag 0 - 0.005 < 0.001	
unFC	162
varFC	120
t ov	151

Table 3.45. Comparison between 0.005 and 0.001 dropout probabilities for Score-CAM XAI method. The number of share images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "0.005 < 0.001" is reported. The green entry is in favour of the 0.005 probability.

The results presented in Table 3.45 indicate that the same method does not identify the same number of flag 0 and flag 1 images. When the probability 0.001 is utilized, more flag 0 images are identified, meaning that more images continue to be classified within the GT class even when the MCD method is applied. Conversely, a lower number of flag 1 images is identified, resulting in fewer images being misclassified when the MCD method is applied. Neither of the probabilities commits any errors. The probability 0.001 is capable of achieving final margins that exceed the initial margins for both flag 0 and flag 1 images. Among the total of 10 areas, only one identifies the probability 0.005 as the one that generates heatmaps with low variability and uncertainty.

Both dropout probabilities, when applied to each XAI method, do not commit any errors. The probability 0.001 allows the ReP method to generate softmax outputs with low prediction uncertainties. Only the data from Score-CAM 0.005 exhibit better margins, specifically for flag 1 images. The probability 0.001 is also the one that generates the most stable and least uncertain heatmaps.

Having identified the dropout probability of 0.001 as having the best performance for each XAI method, the two XAI methods are compared with each other to determine which is more suitable for use with the XCiT network. The comparison between Score-CAM 0.001 and Grad-CAM 0.001 is presented in Table 3.46.

408 correctly classified images	
share flag 0	399
share flag 1	5
margin gain Score-CAM $>$ Grad-CAM	202
margin gain Score-CAM $>$ Grad-CAM	3
flag 1 - Score-CAM $<$ Grad-CAM	
unFC	1
unTC	3
varFC	1
varTC	1
un ov	4
var ov	3
t ov	2
flag 0 - Score-CAM $<$ Grad-CAM	
unFC	186
varFC	131
t ov	197

Table 3.46. Comparison between Score-CAM and Grad-CAM methods. The number of share images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "Score-CAM < Grad-CAM" is reported. The green entries are those in favour of the Score-CAM method

The results presented in Table 3.46 demonstrate that each XAI method identifies the same number of flag 0 and flag 1 images. No XAI method commits any errors. The Score-CAM 0.001 method achieves final margins that exceed the initial margins for both flag 0 and flag 1 images. Among the 10 areas compared, only 3 areas, which correspond to the cells highlighted in green, favor the Score-CAM method.

The Score-CAM method, when used with a dropout probability of 0.001, is the XAI method that allows the ReP method to commit fewer errors for misclassified images, whereas the ReP method does not make errors for correctly classified images. Additionally, Score-CAM is the method that achieves softmax outputs from the ReP method with margins greater than the initial margins for correctly classified images, although the margins do not exhibit high performance for misclassified flag 0 images. Score-CAM demonstrates the same performance as Grad-CAM on misclassified flag 1 images. However, this XAI method generates heatmaps that are variable and uncertain.

3.3.3 Discussion on validation set

The ReP method has been identified as the most optimal method for manipulating images. This method has allowed for the identification of heatmaps and softmax outputs generated by the Score-CAM method as the most robust and least uncertain, respectively, when used with a dropout probability of 0.001. Data from the validation set indicate that the Grad-CAM method is not effective when employed with the XCiT network. It is now possible to preliminarily characterize the behavior of the network.

The validation set comprises 433 images, of which 408 are correctly classified while 25 are misclassified. The confusion matrix for the classification of the XCiT network is presented in Fig. 3.1.

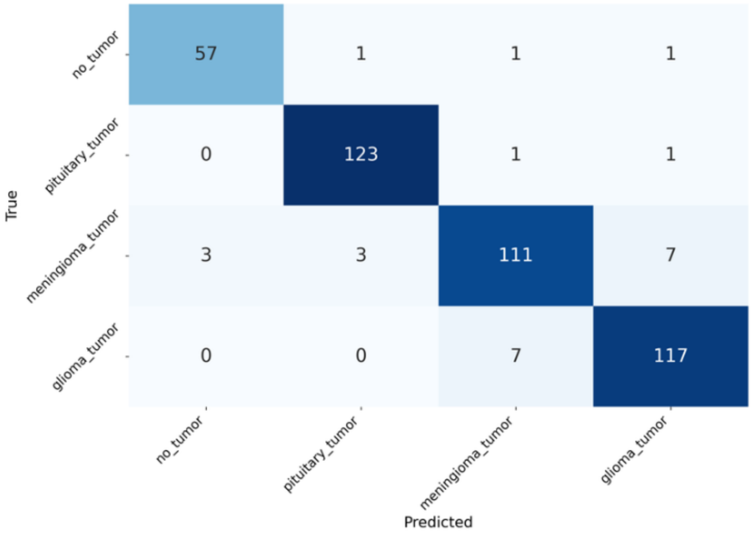

Figure 3.1. This is the confusion matrix of the XCiT net. On the y-axis there are GT labels divided into: no tumor, pituitary tumor, meningioma tumor, glioma tumor. On the x-axis there are the prediction for each class image.

The confusion matrix demonstrates that the network correctly classifies the majority of the images. Only three images from the "no tumor" class are misclassified into each of the three tumor classes. Images classified as "pituitary tumor" are misclassified only twice into the "meningioma tumor" and "glioma tumor" classes. Images from the "meningioma tumor" class are misclassified seven times into the "glioma tumor" class, three times into the "pituitary tumor" class, and three times into the "no tumor" class. Images from the "glioma tumor" class are misclassified seven times into the "meningioma tumor" class.

These data indicate that the network is highly robust and capable of distinguishing well between an image that presents a tumor and one that does not, as evidenced by the data from the "pituitary tumor" and "glioma tumor" classes. The network misclassifies images that do not present a tumor three times, while it incorrectly considers images from the "meningioma tumor" class as "no tumor" three additional times. After employing the ReP method, the confusion matrix is modified, and the new matrix is presented in Fig. 3.2.

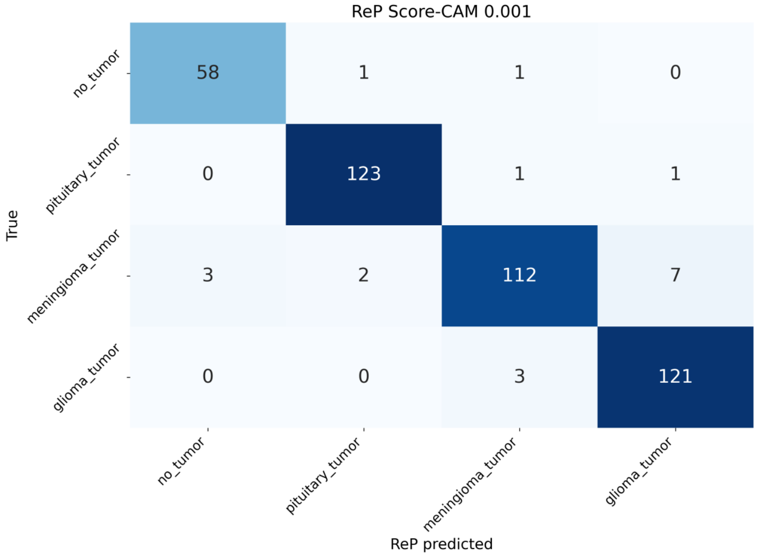

Figure 3.2. This is the confusion matrix after applying the ReP method. On the y-axis there are GT labels divided into: no tumor, pituitary tumor, meningioma tumor, glioma tumor. On the x-axis there are the prediction of ReP softmax for each class image.

By applying the ReP method to the images misclassified by the XCiT neural network, it has been possible to correct the classification of some of these images. In fact, Fig. 3.2 shows six images that are correctly classified in addition to those in the confusion matrix presented in Fig. 3.1. In certain cases, the method achieves a margin that exceeds the original, resulting in a very low prediction uncertainty. This occurs only for misclassified images that have a flag of 1, as described in Section 3.1. The Score-CAM 0.001 data presented in Table 3.10 indicate that this happens for two of the six identified images with flag 1. On average, these flag 1 images may exhibit a percentage variation of up to 227%. The percentage variation is calculated using equation 3.1.

$$\frac{x_f - x_i}{x_i} \times 100 \tag{3.1}$$

In which x_f represents the margin calculated using the softmax from the ReP method, as described in Section 2.5.3, and x_i represents the original margin calculated from the softmax provided by the neural network. Of these two flag 1 images, the image gg(704) achieves the highest percentage variation in the margin, specifically 539%. The initial margin, calculated using the softmax from the XCiT network, is 0.049. As described in Section 2.5.3, the heatmap provided by the neural network is highly uncertain; however, with the application of the ReP method, the proposed heatmap is associated with a softmax that ranks the GT class first and has a final margin of 0.312.

The remaining four flag 1 images continue to possess the GT class as the classification proposed by the method, but they do not have a margin exceeding the original. On average, the percentage variation of the margins is -66%, where:

- The image gg(717) exhibits the smallest percentage variation of -36%, changing from an original margin of 0.246 to a final margin of 0.157;
- The image gg(712) displays the largest percentage variation of -95%, transitioning from an original margin of 0.182 to a final margin of 0.008.

The margin is an essential element for characterizing the heatmaps. Not all misclassified flag 1 images that become correctly classified possess softmax outputs and, consequently, heatmaps that are certain towards the GT class. Of the 25 images misclassified by the XCiT network, only 2 have heatmaps that are certain towards the GT class, while 4 exhibit slightly higher uncertainty towards the GT class.

The identified flag 0 images total 15, and these images have the GT class as the second most probable class in the softmax presented by the ReP method. The method is unable to correct the prediction for these images; however, it succeeds in increasing the probability associated with the GT class while decreasing the margin. This occurs for 13 out of the 15 flag 0 images. The average percentage variation of the margins is -2%, where:

- The image m3(230) exhibits the smallest percentage variation of -0.017%, changing from an original margin of 0.9537 to a final margin of 0.9535;
- The image p(746) demonstrates the largest percentage variation of -11%, changing from an original margin of 0.843 to a final margin of 0.753.

These percentages indicate that the heatmaps of the 13 images remain very certain regarding the predicted class. In this case, the flag 0 identified a set of images that do not exhibit high prediction uncertainty, contrary to the hypothesis in Section 3.1.

The ReP method, as stated in Section 2.5.3, may incur errors: a misclassified flag 1 image might not possess the GT class as the highest probability, or a flag 0 image might not have the GT class as the second-highest probability within the softmax. The data in Table 3.10 indicate that four errors were committed. These images remain misclassified by the neural network. In Fig. 3.2, it is evident that some misclassifications persist as reported in Fig. 3.1.

By employing the ReP method to analyze the images correctly classified by the XCiT network, 401 images are identified that are consistently classified within the GT class, even when the MCD method is applied, flag 0, alongside 7 images that exhibit at least one incorrect classification when the MCD method is applied, flag 1, as shown by the Score-CAM 0.001 results in Table 3.31. For 113 flag 0 images, the method successfully increases the margin while reducing prediction uncertainty, achieving an average percentage variation of the margins of 0.33%. The image with the highest percentage variation, at 22%, is the image m3(162), which changes from an original margin of 0.388 to a final margin of 0.473.

The remaining 288 flag 0 images continue to have the GT class as their prediction; however, their margin is lower than the original. The average percentage variation of the margins for these images is -0.24%, where:

- The image m3(166) exhibits the greatest percentage variation of -14%, transitioning from an original margin of 0.876 to a final margin of 0.751;
- The image gg(814) shows the smallest percentage variation, approximately on the order of 10^{-6}, with the initial margin being similar to the final margin at 0.998.

All identified images as flag 0 present low prediction uncertainty towards the GT class, even though the margin for some of these is less than the original provided by the XCiT network.

Among the 7 identified flag 1 images, the method achieves a final margin exceeding the original for 5 of them, resulting in an average percentage variation of 39%. The image with the greatest percentage variation is the image m3(170), which shows a variation of 145%, changing from an original margin of 0.073 to a final margin of 0.178.

The remaining 2 images exhibit lower margins compared to the originals, with an average percentage variation of -10%, where:

- The image m3(175) shows the largest percentage variation at -12%, changing from an original margin of 0.549 to a final margin of 0.486;
- The image gg(706) has the smallest percentage variation at -8%, changing from an original margin of 0.186 to a final margin of 0.171.

In this instance, not all identified flag 1 images exhibit low prediction uncertainty towards the GT class; these results arise not only from the use of misclassified heatmaps, which contribute to the ReP method but also from the original prediction uncertainty of the neural network.

The constructed and trained network demonstrates a very robust localization, with the areas of variability, uncertainty, and overlap reported in Tables 3.10 and 3.31 being quite limited, reflecting the robustness of the interpretation. The XCiT network is highly confident in its predictions, except for 6 misclassified flag 1 images, for which it was possible to modify the final prediction. Only 2 of these images exhibit low uncertainty towards the GT class, while the remaining 4 exhibit high uncertainty towards the GT class.

3.3.4 Test set misclassified images

Having characterized the trained network, its behavior is analyzed on the test set images. Both Score-CAM and Grad-CAM images are examined using only the ReP method, with a particular focus on the dropout probability of 0.005 for Grad-CAM. This re-evaluation aims to assess the performance trends on the test set and to confirm that Score-CAM with a dropout rate of 0.001 is indeed the most optimal approach for the XCiT network.

The analysis of the test set images is divided between misclassified and correctly classified images. The data presented follows the same format as that of the validation set outlined in Section 3.3.1 and 3.3.2. All images are categorized into flag 0 and flag 1 according to the rules specified in Section 3.1. Only those images that conform to the classification of each flag are reported, indicating the areas of variability and uncertainty in the heatmaps, as well as the prediction uncertainty. All images that do not align with the classification of their respective flag are considered errors of the ReP method.

The dropout probabilities for each XAI method and the methods themselves are compared again. The comparison tables report the number of common flag 0 and flag 1 images, as well as which images exhibit the best margin behavior. Various areas are compared to determine which method and dropout probability yield the most robust and least uncertain heatmaps.

Tables 3.47 and 3.48 present data for the Grad-CAM method. The headers of the tables indicate the number of misclassified images identified by the neural network and the dropout probability used with the XAI method. The analysis focuses solely on flag 0 and flag 1 images that conform to their respective flag classifications, reporting the number of images that reflect margin behavior, along with the areas of variability, uncertainty, and overlap expressed in pixels.

77 misclassified images			
Grad-CAM 0.005			
flag 1		flag 0	
number	1	number	13
margin gain	1	margin decrease	8
unFC	16687	unFC	13845
unTC	13391	/	/
varFC	14014	varFR	10972
varTC	8384	/	/
un ov	6864	/	/
var ov	2851	/	/
t ov	17013	t ov	9636

Table 3.47. Analysis of Grad-CAM 0.005. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.47, it can be observed that the ReP method incurs 63 errors but is only able to correct the prediction of a single image. Meanwhile, 13 images continue to list the GT class as the second most probable class in the softmax of the method. The only identified flag 1 image also achieves a margin wider than the original, while 8 out of the 13 flag 0 images attain a margin lower than the original.

Given the total number of pixels, the areas of variability, uncertainty, and overlap are notably small. The largest area reported in Table 3.47 is "t ov = 17,013", which corresponds to 6% of the total image area.

In Table 3.48, it can be noted that the ReP method incurs 42 errors but is only able

77 misclassified images			
Grad-CAM 0.001			
flag 1		flag 0	
number	1	number	24
margin gain	1	margin decrease	18
unFC	13949	unFC	13584
unTC	12552	/	/
varFC	11006	varFR	10280
varTC	9351	/	/
un ov	5762	/	/
var ov	3202	/	/
t ov	14671	t ov	8952

Table 3.48. Analysis of Grad-CAM 0.001. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

to correct the prediction of a single image, while 24 images continue to list the GT class as the second most probable class in the softmax of the method. The only identified flag 1 image also achieves a margin wider than the original, whereas 18 out of the 24 flag 0 images attain a margin lower than the original.

Given the total number of pixels in the image, the areas of variability, uncertainty, and overlap are quite small. The largest area reported in Table 3.48 is "t ov = 14,671", which corresponds to 6% of the total image area.

The dropout probabilities are compared, and the results are presented in Table 3.49. The table header continues to indicate the number of images misclassified by the neural network and the analyzed XAI method. It is also verified whether the probabilities identify the same flag 0 and flag 1 images. Only the common images are analyzed.

The margins are examined to identify which dropout probability has the highest number of flag 0 images with a lower margin and the highest number of flag 1 images with a greater margin. The areas are compared, and the probability with the smallest areas of variability and uncertainty on the heatmaps is identified.

In Table 3.49, it can be observed that the dropout probabilities identify the same flag 1 image, but not the same flag 0 images. Additionally, Grad-CAM with a dropout probability of 0.005 incurs 11 more errors than Grad-CAM with a dropout probability of 0.001. The 0.001 dropout probability has the highest number of flag 0 images that exhibit a final margin lower than the initial margin, as well as the highest number of flag 1 images that show a final margin greater than the initial margin. Only the cells highlighted in green, which correspond to the number of images satisfying the condition "0.005 < 0.001", favor the 0.005 probability. Out of the 10 areas compared, 5 favor the 0.005 probability.

In this case, the 0.001 probability has the best margins for each flag, contrary to what

77 misclassified images	
Grad-CAM	
share flag 0	13
share flag 1	1
margin decrease 0.001 < 0.005	9
margin gain 0.001 > 0.005	1
flag 1 - 0.005 < 0.001	
unFC	0
unTC	0
varFC	0
varTC	1
un ov	0
var ov	1
t ov	0
flag 0 - 0.005 < 0.001	
unFC	8
varFC	6
t ov	6

Table 3.49. Comparison between 0.005 and 0.001 dropout probability for Grad-CAM XAI method. The number of shared images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "0.005 < 0.001" is reported. Just the green entries are those in favor of 0.005 probability.

was observed in the validation set, where the data in Table 3.41 indicated that the 0.005 probability had the most optimal margins. In the validation set, the comparison of areas identified the 0.001 probability as generating more stable and less uncertain heatmaps, whereas now no probability generates more stable and less uncertain heatmaps.

Tables 3.50 and 3.51 refer to the data from the Score-CAM method. The table headers include the number of misclassified images identified by the neural network and the dropout probability used with the XAI method. The analysis focuses solely on flag 0 and flag 1 images that conform to the classification of their respective flags, reporting the number of images that reflect margin behavior and the areas of variability, uncertainty, and overlap expressed in pixels.

In Table 3.50, it can be observed that the ReP method incurs 62 errors but is only able to correct the prediction of a single image, while 14 images continue to have the GT class as their second most probable class in the method's softmax output. The only identified flag 1 image is able to achieve a margin wider than the original, whereas 8 out of 14 flag 0 images obtain a margin lower than the original. Given the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.50 is "t ov = 21,251", which corresponds to 8% of the total area of

77 misclassified images			
Score-CAM 0.005			
flag 1		flag 0	
number	1	number	14
margin gain	1	margin decrease	8
unFC	16342	unFC	12607
unTC	16819	/	/
varFC	15107	varFR	12663
varTC	15652	/	/
un ov	7222	/	/
var ov	3567	/	/
t ov	21251	t ov	8995

Table 3.50. Analysis of Score-CAM 0.005. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

the image.

77 misclassified images			
Score-CAM 0.001			
flag 1		flag 0	
number	1	number	23
margin gain	1	margin decrease	11
unFC	15644	unFC	11604
unTC	16076	/	/
varFC	16375	varFR	11334
varTC	14102	/	/
un ov	6471	/	/
var ov	5486	/	/
t ov	21238	t ov	7764

Table 3.51. Analysis of Score-CAM 0.001. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.51, it can be observed that the ReP method incurs 53 errors but is only able to correct the prediction of a single image, while 23 images continue to have the GT class as their second most probable class in the method's softmax output. The only identified flag 1 image is also able to achieve a margin wider than the original, whereas 11 out of 23 flag 0 images obtain a margin lower than the original. Given the total number of pixels in the image, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.51 is "t ov = 21,238", which corresponds to 8% of the total area of the image.

The dropout probabilities are compared, and the results are presented in Table 3.52. The header of the table continues to report the number of misclassified images identified by the neural network and the XAI method analyzed. It is also verified whether the probabilities identify the same flag 0 and flag 1 images. Only the common images are analyzed. The margins are examined to determine which dropout probability has the highest number of flag 0 images with lower margins, and the highest number of flag 1 images with higher margins. The areas are compared, and the probability with the smallest areas of variability and uncertainty on the heatmaps is identified.

77 misclassified images	
Score-CAM	
share flag 0	12
share flag 1	1
margin decrease $0.005 < 0.001$	9
margin gain $0.005 > 0.001$	1
flag 1 - $0.005 < 0.001$	
unFC	0
unTC	0
varFC	1
varTC	0
un ov	0
var ov	1
t ov	0
flag 0 - $0.005 < 0.001$	
unFC	4
varFC	2
t ov	6

Table 3.52. Comparison between 0.005 and 0.001 probability for Score-CAM XAI method. The number of shared images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "$0.005 < 0.001$" is reported. Just the green entries are those in favour of 0.005 probability, while the yellow entries are not decisive.

In Table 3.52, it can be observed that the dropout probabilities identify the same flag 1 image, but do not identify the same flag 0 images. Furthermore, Score-CAM with a dropout rate of 0.005 incurs 9 more errors than Score-CAM with a dropout rate of 0.001. The dropout rate of 0.005 has the highest number of flag 0 images that exhibit a final margin lower than the initial margin, and it also has the highest number of flag 1 images with a final margin exceeding the initial margin. Only the cells highlighted in green, which correspond to the number of images meeting the condition "0.005 < 0.001", favor the 0.005 probability. Since the number of identified flag 0 images is equal, all cells reporting a number of images corresponding to "number of images flag/2" are highlighted in yellow and are not considered in the final comparison. Out of the 9 identified areas, 2 favor the 0.005 probability.

As observed in the validation set, referenced in Table 3.42, the 0.005 probability possesses the best margins for each flag. However, this probability generates unstable and highly uncertain heatmaps, a trend that is also found in the test set. The 0.001 probability continues to generate stable and low-uncertainty heatmaps, and this performance trend on the heatmaps has been maintained in the test set. The Score-CAM method using the 0.001 probability continues to fail to achieve low prediction uncertainties on the test set, while the Grad-CAM method using the same probability achieves better margin performance compared to the validation set.

The explainability methods are compared to verify whether Score-CAM remains the most performant XAI method, with the results presented in Table 3.53. The header of the table reports the number of misclassified images identified by the neural network. It is also checked whether the two methods identify the same flag 0 and flag 1 images, with only the common images being considered. The margins are analyzed to identify which method has the highest number of flag 0 images with lower margins and the highest number of flag 1 images with higher margins. The areas are compared, and the method with the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.53, it can be observed that the two methods identify the same flag 1 image but do not identify the same flag 0 images. The Score-CAM method incurs one more error than the Grad-CAM method. The Grad-CAM method has the highest number of flag 0 images that exhibit a final margin lower than the initial margin, and it also possesses the highest number of flag 1 images that have a final margin greater than the initial margin. Only the cells highlighted in green, which correspond to the number of images that meet the condition "Score-CAM < Grad-CAM", favor the Score-CAM method. Out of the 10 compared areas, 2 favor the Score-CAM method.

The Score-CAM method demonstrates slightly lower performance on the test set compared to the validation set. During the validation phase, as referenced in Table 3.43, the Grad-CAM method had the best margins only for flag 0 images, while no method had the best margins for flag 1 images. In the test set, Grad-CAM achieves optimal margins even for flag 1 images. In the validation set, many entries were not considered as they did not favor any method, and only one area out of the remaining three preferred the Score-CAM method. In the test set, there are no entries that do not tend toward any method, and the areas indicate that the Grad-CAM method generates the most stable and least uncertain heatmaps.

3.3.5 Test set correctly classified images

Tables 3.54 and 3.55 refer to the data from the Grad-CAM method. The headers of the tables indicate the number of correctly classified images identified by the neural network and the dropout probability used with the XAI method. The analysis focuses exclusively on flag 0 and flag 1 images that adhere to the respective classifications, reporting

77 misclassified images	
share flag 0	21
share flag 1	1
margin decrease Grad-CAM < Score-CAM	15
margin gain Grad-CAM > Score-CAM	1
flag 1 - Score-CAM < Grad-CAM	
unFC	0
unTC	0
varFC	0
varTC	0
un ov	0
var ov	0
t ov	0
flag 0 - Score-CAM < Grad-CAM	
unFC	14
varFC	5
t ov	13

Table 3.53. Comparison between Score-CAM and Grad-CAM method. The number of shared images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "Score-CAM < Grad-CAM" is reported. Just the green entries are those in favor of the Score-CAM method.

the number of images that reflect margin behavior, as well as the areas of variability, uncertainty, and overlap expressed in pixels.

In Table 3.54, it can be observed that the ReP method does not make any errors, with 301 images possessing flag 0, while 16 images possess flag 1. Only 3 out of 16 flag 1 images and 30 out of 301 flag 0 images achieve a wider margin than the original. Considering the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.54 is "t ov = 13,133", which corresponds to 5% of the total image area.

In Table 3.55, it can be observed that the ReP method does not make any errors, with 313 images possessing flag 0, while 4 images possess flag 1. Only 2 out of 4 flag 1 images and 103 out of 313 flag 0 images achieve a wider margin than the original. Considering the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area reported in Table 3.55 is "t ov = 14,214", which corresponds to 5% of the total image area.

The dropout probabilities are compared, and the results are presented in Table 3.56. The table header continues to display the number of images correctly classified by the neural network and the analyzed XAI method. It is also verified whether the probabilities identify the same flag 0 and flag 1 images. Only the common images are taken into consideration. The margins are analyzed to identify which dropout probability has the

317 correctly classified images			
Grad-CAM 0.005			
flag 1		flag 0	
number	16	number	301
margin gain	3	margin gain	30
unFC	11395	/	/
unTC	12549	unTC	12413
varFC	7832	/	/
varTC	8941	varTC	10804
un ov	7795	/	/
var ov	3506	/	/
t ov	13133	t ov	9353

Table 3.54. Analysis of Grad-CAM 0.005. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

317 correctly classified images			
Grad-CAM 0.001			
flag 1		flag 0	
number	4	number	313
margin gain	2	margin gain	103
unFC	11532	/	/
unTC	12854	unTC	12485
varFC	10021	/	/
varTC	9756	varTC	10063
un ov	6084	/	/
var ov	3546	/	/
t ov	14214	t ov	8831

Table 3.55. Analysis of Grad-CAM 0.001. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

greatest number of flag 0 and flag 1 images with margins exceeding the original. The areas are compared, and the probability that possesses the smallest areas of variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.56, it can be noted that the dropout probabilities do not identify the same flag 0 and flag 1 images. The probability 0.001 has more flag 0 images, indicating a greater number of images that are not misclassified when the MCD method is applied, and it

317 correctly classified images	
Grad-CAM	
share flag 0	301
share flag 1	4
margin gain 0.001 > 0.005	274
margin gain 0.001 > 0.005	4
flag 1 - 0.005 < 0.001	
unFC	2
unTC	3
varFC	1
varTC	3
un ov	0
var ov	1
t ov	2
flag 0 - 0.005 < 0.001	
unFC	145
varFC	103
t ov	115

Table 3.56. Comparison between 0.005 and 0.001 dropout probability for Grad-CAM XAI method. The number of shared images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "0.005 < 0.001" is reported. Just the green entries are those in favour of 0.005 probability, while the yellow entries are not decisive.

also has fewer flag 1 images, which means a lower number of images that are misclassified when the MCD method is applied. No dropout probability shows errors from the ReP method.

The probability 0.001 exhibits the best margins for both flag 0 and flag 1 images. Only the cells highlighted in green, which correspond to the number of images that meet the condition "0.005 < 0.001", favor the probability 0.005. Since the number of identified flag 1 images is even, all cells that report a number of images corresponding to "number of flag images/2" are highlighted in yellow and are not considered in the final comparison. Of the remaining 8 areas, 2 favor the probability 0.005.

As was the case in the validation set, referring now to the data in Table 3.44, the probability 0.001 remains the one with the least prediction uncertainty. In the validation set, all 10 areas confirmed that probability 0.001 generates the most stable and least uncertain heatmaps. This trend is also observed in the test set; however, two areas do not fall into the final comparison, and one area prefers the dropout probability of 0.005.

Table 3.57 and Table 3.58 refer to the data from the Score-CAM method. The table headers display the number of correctly classified images identified by the neural network and the dropout probability used with the XAI method. The analysis focuses solely on flag 0 and flag 1 images that comply with their respective classifications, reporting the number of images that reflect the margin behavior and the areas of variability, uncertainty, and overlap expressed in pixels.

317 correctly classified images			
Score-CAM 0.005			
flag 1		flag 0	
number	17	number	300
margin gain	1	margin gain	44
unFC	12669	/	/
unTC	13487	unTC	13326
varFC	12825	/	/
varTC	13014	varTC	12616
un ov	5375	/	/
var ov	3613	/	/
t ov	15214	t ov	9010

Table 3.57. Analysis of Score-CAM 0.005. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

In Table 3.57, it can be observed that the ReP method does not commit any errors; 300 images have a flag 0, while 17 images have a flag 1. Only 1 flag 1 image out of 17 and 44 flag 0 images out of 300 manage to achieve a margin wider than the original. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area shown in Table 3.57 is "t ov = 15,214", which corresponds to 6% of the total area of the image.

In Table 3.58, it can be observed that the ReP method does not commit any errors; 312 images have a flag 0, while 5 images have a flag 1. No flag 1 image is able to achieve a margin wider than the original, whereas 102 flag 0 images out of 312 manage to improve their margin. Knowing the total number of pixels, the areas of variability, uncertainty, and overlap are very small. The largest area shown in Table 3.58 is "t ov = 16,548", which corresponds to 6% of the total area of the image.

The dropout probabilities are compared, and the results are reported in Table 3.59. The header of the table continues to indicate the number of correctly classified images identified by the neural network and the XAI method used. It is also verified whether the probabilities identify the same flag 0 and flag 1 images. Only the common images are considered. The margins are analyzed to determine which dropout probability has the highest number of flag 0 and flag 1 images with margins greater than the originals. The areas are compared, and the probability that has the least variability, uncertainty, and overlap on the heatmaps is identified.

In Table 3.59, it can be observed that the dropout probabilities do not identify the same flag 0 and flag 1 images. The probability of 0.001 has more flag 0 images, indicating

317 correctly classified images			
Score-CAM 0.001			
flag 1		flag 0	
number	5	number	312
margin gain	0	margin gain	102
unFC	13336	/	/
unTC	14850	unTC	12817
varFC	15341	/	/
varTC	13812	varTC	11205
un ov	5018	/	/
var ov	5638	/	/
t ov	16548	t ov	8256

Table 3.58. Analysis of Score-CAM 0.001. The analysis is divided between flag 1 and flag 0 images. The "number" row refers to the number of images that belong to each flag group. The number of images that respects the behaviour of the margin is following. Each flag in characterized by the wideness of the analyzed areas in pixel.

a greater number of images that are not misclassified when the MCD method is applied, and it also has fewer flag 1 images, which corresponds to a lower number of images that are misclassified when the MCD method is applied. No dropout probability yields errors from the ReP method. The probability of 0.001 offers the best margins for flag 0 images, while it is the probability of 0.005 that possesses the best margins for flag 1 images. Only the cells highlighted in green, corresponding to the number of images that meet the condition "0.005 < 0.001", favor the probability of 0.005. Out of the 10 areas compared, 6 are in favor of the probability of 0.005.

Referring to the validation set data presented in Table 3.45, the probability of 0.001 remains the one with the least prediction uncertainty for flag 0 images, while flag 1 images continue to exhibit the least prediction uncertainty only when the dropout probability is 0.005. This trend in margin behavior persists in the test set. For the validation set, the probability of 0.001 generates the most stable and least uncertain heatmaps; however, test set data indicate that the probability of 0.005 generates the most stable and least uncertain heatmaps. The drop in performance observed for test set images also affects the Score-CAM method, which, when used with the probability of 0.001, is no longer the optimal XAI method for analyzing correctly classified images.

The explainability methods are compared to determine if Score-CAM continues to be the most effective XAI method, with the data presented in Table 3.60. The header of the table indicates the number of correctly classified images identified by the neural network. It is also verified whether the two methods identify the same flag 0 and flag 1 images. Only the common images are taken into account. The margins are analyzed to identify which method has the highest number of flag 0 and flag 1 images with the best margins. The areas are compared, and the method that has the least variability, uncertainty, and overlap in the heatmaps is identified.

317 correctly classified images	
Score-CAM	
share flag 0	300
share flag 1	5
margin gain 0.001 > 0.005	253
margin gain 0.005 > 0.001	3
flag 1 - 0.005 < 0.001	
unFC	🟩 4
unTC	🟩 3
varFC	🟩 3
varTC	🟩 5
un ov	1
var ov	🟩 5
t ov	🟩 4
flag 0 - 0.005 < 0.001	
unFC	128
varFC	83
t ov	103

Table 3.59. Comparison between 0.005 and 0.001 dropout probability for Score-CAM XAI method. The number of shared images, the probability that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "0.005 < 0.001" is reported. Just the green entries are those in favour of 0.005 probability.

In Table 3.60, it can be observed that the two methods do not identify the same flag 0 and flag 1 images. The Score-CAM method identifies one additional flag 1 image, meaning an image that is not misclassified when the MCD method is applied, and it identifies one fewer flag 0 image, indicating one less image that is misclassified when the MCD method is applied. The Score-CAM method exhibits the least prediction uncertainty for flag 1 images, but it does not achieve the same result for flag 0 images. Only the cells highlighted in green, corresponding to the number of images that meet the condition "Score-CAM < Grad-CAM", favor the Score-CAM method. Of the 10 areas compared, 3 are in favor of the Score-CAM method.

The performance of the Score-CAM method on the test set is slightly lower compared to its performance on the validation set, as indicated in Table 3.46. In the test set, the Score-CAM method is no longer the one that has the least prediction uncertainty for both flag 0 and flag 1 images; it achieves the least prediction uncertainty only for flag 1 images. The Score-CAM method continues to generate more variable and uncertain heatmaps, as observed in the validation set.

317 correctly classified images	
share flag 0	311
share flag 1	3
margin gain Grad-CAM > Score-CAM	163
margin gain Score-CAM > Grad-CAM	3
flag 1 - Score-CAM < Grad-CAM	
unFC	1
unTC	1
varFC	1
varTC	0
un ov	2
var ov	0
t ov	1
flag 0 - Score-CAM < Grad-CAM	
unFC	172
varFC	128
t ov	186

Table 3.60. Comparison between Score-CAM and Grad-CAM method. The number of shared images, the method that has the best margins for each flag are reported. Each area is confronted and the number of images that verify the condition "Score-CAM < Grad-CAM" is reported. Just the green entries are those in favour of the Score-CAM method.

Chapter 4

Discussion

The Score-CAM method was identified as the most optimal approach to use with the XCiT network based on the data obtained from the validation set. However, the performance of this XAI method experiences a significant drop, and it no longer serves as the method that generates stable and low-uncertainty heatmaps, as well as softmax outputs with low prediction uncertainties. When using the Score-CAM method with a dropout probability of 0.001, the correctly classified images exhibit more unstable and uncertain heatmaps compared to the results obtained from the validation set, while the misclassified images maintain consistent performance across both sets.

The test set consists of 394 images, of which 317 are correctly classified and 77 are misclassified. The confusion matrix for the classification performed by the XCiT network is presented in Fig. 4.1.

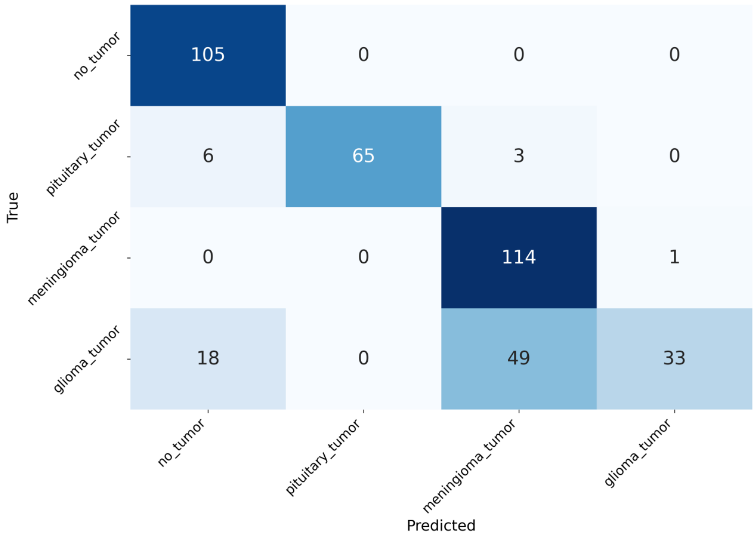

Figure 4.1. This is the confusion matrix of the XCiT net. On the y-axis there are GT labels divided into: no tumor, pituitary tumor, meningioma tumor, glioma tumor. On the x-axis there are the prediction for each class image.

The confusion matrix illustrates that the network correctly classifies the majority of the images, particularly as follows:

- All images without tumors are accurately classified, whereas in the validation set, the network made three errors;
- Images of pituitary tumors are misclassified as meningioma tumors and no tumors. The misclassification in the no tumor class did not occur in the validation set, where misclassifications happened in the glioma tumor class;
- Only one image from the meningioma tumor class is misclassified, with the network assigning it to the glioma class;
- Images of glioma tumors are misclassified more frequently than they are correctly classified; this trend was not evident in the validation set. In the validation set, these images were misclassified as meningioma tumors, but in the test set, 18% of these images are classified as no tumors, and 49% are classified as meningioma tumors.

In the validation set, only three images belonging to the meningioma tumor class are classified as no tumor. In the test set, however, as many as 18 images from the glioma tumor class and six images from the pituitary tumor class are classified as no tumor. These classification errors are unacceptable because, theoretically, a neural network should misclassify an image from one tumor class as belonging to a different tumor class. This data indicates that the network has failed to generalize and has experienced overfitting. After analyzing the images using the ReP method, the confusion matrix is modified, and the new matrix is presented in Fig. 4.2.

By applying the method to the images misclassified by the XCiT neural network, it was possible to correct the classification of only one image belonging to the glioma tumor class. This image was originally classified as meningioma tumor, but it is now correctly classified within the glioma tumor class. As reported in Table 3.51, this image, which is image(309), also achieves a margin higher than the original. The percentage variation of the margin is 14,641%, increasing from an original margin of 0.041 to a final margin of 0.647. The softmax provided by the XCiT network exhibits high prediction uncertainty; however, with the application of the ReP method, this image now demonstrates low prediction uncertainty regarding the GT class.

There are 23 images identified as flag 0, 11 of which possess a margin lower than the original, indicating a higher probability for the GT class. The average percentage variation of the margins is -2%, where:

- Image image(311) has the smallest percentage variation of -0.0009%, with the initial margin being similar to the final margin at 0.996;
- Image image(228) shows the greatest percentage variation of -14%, decreasing from an original margin of 0.886 to a final margin of 0.758.

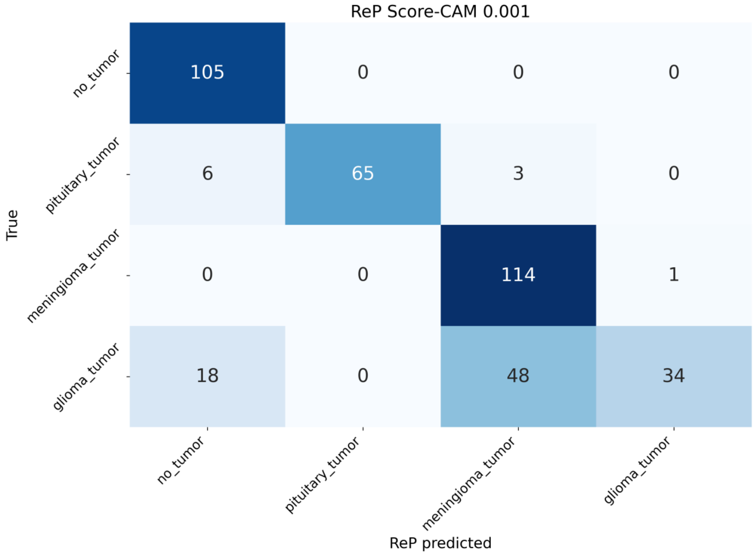

Figure 4.2. This is the confusion matrix after applying the ReP method. On the y-axis there are GT labels divided into: no tumor, pituitary tumor, meningioma tumor, glioma tumor. On the x-axis there are the prediction of ReP softmax for each class image.

These percentages indicate that the heatmaps of the 23 images remain highly certain regarding the predicted class. The flag 0, as observed in the validation set, has identified a series of images that exhibit low prediction uncertainty toward the predicted class.

The ReP method has made a greater number of errors compared to those committed in the validation set: as reported in Table 3.51, these 53 errors correspond to misclassified images that continue to remain misclassified.

Upon analyzing the images correctly classified by the XCiT network using the ReP method, 312 images are identified that remain classified in the GT class even when the MCD method is applied (flag 0), while 5 images exhibit at least one incorrect classification when the MCD method is applied (flag 1), as shown in the results presented in Table 3.58. For 102 flag 0 images, the method successfully increases the margin while reducing prediction uncertainty, achieving an average percentage variation of the margins equal to 0.15%. The image with the highest percentage variation, equal to 4%, is image(107), which increases from an original margin of 0.718 to a final margin of 0.744.

The remaining 210 flag 0 images continue to have the GT class as their prediction; however, the margin is lower than the original. The average percentage variation of the margins for these images is -0.4%, where:

- The image image(150) exhibits the largest percentage variation, which is -21%, decreasing from an initial margin of 0.775 to a final margin of 0.615;

- The image image(187) shows the smallest percentage variation, which is -0.0002%, with the initial margin being similar to the final margin, at 0.997.

All images identified as flag 0 exhibit low prediction uncertainty toward the GT class, although the margin for some of these is lower than the original margin provided by the XCiT network.

Of the 5 identified flag 1 images, none are able to achieve a margin greater than the original. All images experience a deterioration of the margin; the average percentage variation of these margins is -35%, where:

- The image image(336) has the largest percentage variation at -54%, decreasing from an original margin of 0.483 to a final margin of 0.220;
- The image image(109) exhibits the smallest percentage variation at -14%, decreasing from an original margin of 0.422 to a final margin of 0.364.

In this case, not all identified flag 1 images exhibit low prediction uncertainty toward the GT class.

The XCiT network maintains low prediction uncertainty during the classification of images in the test set. Whether the network classifies correctly or misclassifies, the prediction uncertainty for the predicted class remains very low. This is also evident from the number of flag 0 images: it is rare for the network to classify an image differently, even when the MCD method is applied. However, the network displays considerable uncertainty when an image is classified differently using the MCD method, particularly for flag 1 images. When this occurs, the network provides highly uncertain predictions; the ReP method emphasizes and characterizes the network's difficulty in identifying the final prediction. The XCiT network exhibits a very robust localization phase, as indicated by the contained areas of variability, uncertainty, and overlap; nevertheless, it has suffered from overfitting, as demonstrated by the margins.

The overfitting is likely also attributable to the suboptimal dataset: as noted in Section 2.1.1, the training set consists of 2,437 images, while the validation set contains 433 images and the test set has 394 images. The sets are imbalanced in size, even though the percentages chosen to divide the images into training and validation sets are optimal for a relatively small dataset like the one used in this thesis. As shown in the confusion matrix presented in Fig. 3.1, the classes are also imbalanced: 60 images belong to the no tumor class, 125 images to the pituitary tumor class, 124 images to the meningioma tumor class, and 124 images to the glioma tumor class. The no tumor class is not well represented at all, accounting for only 14% of the entire validation set.

The confusion matrix shown in Fig. 4.1 illustrates how overfitting has occurred in the no tumor and meningioma tumor classes. In the test set, the network does not classify any images belonging to the no tumor class even once, while it misclassifies only one image from the meningioma tumor class. These two classes exhibit significantly fewer errors compared to those made by the network in the validation set. The pituitary tumor and glioma tumor classes demonstrate the poorest performance. The test set is smaller than the validation set; however:

- The no tumor class contains 75% more images than the number of images present in the validation set;
- The pituitary tumor class contains 41% fewer images than the number of images present in the validation set;
- The meningioma tumor class contains 7% fewer images than the number of images present in the validation set;
- The glioma tumor class contains 19% fewer images than the number of images present in the validation set.

Despite the fact that the Score-CAM method experiences a drop in performance such that it is no longer considered the optimal XAI method to use with the XCiT neural network, the probability of 0.001 continues to be identified as the most optimal dropout probability. Moreover, in the test set, when used in conjunction with the Grad-CAM method to analyze misclassified images, this probability yields margin performance that exceeds those identified in the validation set.

Chapter 5

Conclusion

This Master's Thesis incorporated the aspect of uncertainty in XAI methods with the aim of enhancing the transparency of the neural network's decision-making process in the classification of brain MRI images.

Currently, studies on neural networks in the literature focus on three main directions: designing and developing neural networks with ever-improving performance, associating increasingly innovative XAI methods with neural networks to generate more accurate heatmaps and provide the end-user with a clearer idea of the spatial region the network will use for subsequent classification, and finally, exploring methods that can be applied to neural networks to gain an indication of the network's uncertainty. For years, the literature has been working to develop XAI methods and uncertainty estimation techniques to make the "black box" behavior of neural networks more interpretable for end-users.

These three aspects are not always interconnected: studies mainly focus on either XAI methods applied to neural networks or uncertainty estimation methods for neural networks. However, XAI methods and uncertainty estimation remain two aspects that have not yet been fully integrated. This thesis successfully combined XAI methods with uncertainty estimation, further characterizing the behavior of the neural network used in this project.

The first objective of this thesis was to quantify the variability and uncertainty of the most important features extracted by the neural network. The neural network used in this project is XCiT, a model that is computationally efficient while performing on par with other networks in the literature. The XAI methods implemented were Grad-CAM and Score-CAM, while the uncertainty estimation method used was MCD, applied with two different dropout probabilities. Since it is impossible to know which features the network extracts or their values, and applying the MCD method turns off some neurons in the XCiT network, meaning the features used for new predictions may differ and have different values (even if these cannot be discovered), four methods were devised to manipulate the heatmaps and softmax outputs provided by the neural network. All heatmaps and softmax outputs produced by XCiT, including those generated when the MCD method is applied, were used. These four methods are based on hypotheses and provide a final heatmap and softmax at the end of their manipulation. These heatmaps are more representative than those provided by the neural network before the MCD method is applied, showing only the spatial region consistently identified and used to make the final prediction. Variability areas and uncertainty areas can be identified on the heatmaps: the former refers to the variability of the most important features identified by the neural network, while the latter refers to errors that manipulation methods may introduce. It is also possible to quantify whether these areas overlap.

The uncertainty of the most important features is "hidden" in the softmax outputs provided by the network. From the softmax, prediction uncertainty can be extracted by calculating the margin: the distance between the predicted class and the GT class.

The first objective of this thesis was to identify which of the four devised methods demonstrated the best performance. The search for the optimal method introduced the concept of flags: each image classified by the neural network exhibits its own behavior before and after analysis using the devised methods. Each flag has specific rules for calculating prediction uncertainty and identifies certain areas on the heatmaps proposed by the methods. The best method for analyzing all images was found to be the method called ReP.

The second objective of this work was to determine which XAI method and which dropout probability yielded the best performance from the XCiT network. This master's thesis not only introduces a series of methods to characterize the variability and uncertainty of the most important features extracted by the neural network but also addresses the challenge of identifying, as accurately as possible, the most optimal XAI method and dropout probability for use with the neural network in this study.

In the validation set, using only the identified ReP method, the data showed that the 0.001 dropout probability was more effective than the 0.005 probability. The XAI method found to be most suitable for use with the XCiT network was Score-CAM. When combined with a 0.001 dropout probability, the Score-CAM method was able to generate softmax outputs with low prediction uncertainties, although it produced variable and uncertain heatmaps.

The test set data were analyzed using only the ReP method to identify areas of variability, uncertainty, and overlap on the heatmaps provided by the method, and to quantify prediction uncertainty based on the softmax outputs. The test set data indicated both a performance drop and overfitting. The Score-CAM method was no longer the optimal XAI method for use with the XCiT network. In particular, the performance drop of the Score-CAM method had a significant impact on the softmax outputs, resulting in highly uncertain heatmaps for the predicted class. Despite the XAI method no longer being validated by the validation set, the 0.001 dropout probability continued to exhibit the best performance, even achieving better results than in the validation set.

5.1 Limitations

Despite the success of this thesis in identifying the variability and uncertainty of the most important features, the primary limitation lies in the manipulation methods for heatmaps and softmax that were developed.

One of the underlying assumptions for constructing these methods is the necessity of having the GT class. While this is not an issue during the training phase of the neural network, in real-world applications, the GT class is not available.

This work is able to characterize the localization and interpretation phases of a neural network only when the GT class is available, meaning the methods developed can only be applied to research problems, not to practical applications. In a real-world application, it is not yet possible to manipulate the heatmaps and softmax outputs provided by the neural network to obtain a more representative heatmap of the most important features, nor is it possible to correct the predictions of some misclassified images.

Without access to the GT class in real-world applications, it is not possible to identify only the most important features extracted by the neural network, nor to characterize the uncertainty of the prediction. In practical applications, one can only characterize the behavior of the network's localization and interpretation phases and deliver the final output, along with its identified strengths and weaknesses, for use in real-world scenarios.

Another issue arising from these methods is the need to identify the optimal analysis method for heatmaps and softmax each time a new study on medical image classification is initiated. There is no available data to definitively confirm that the method identified in this work is the most effective for other neural networks and XAI methods.

Another limitation of this thesis is the dataset used. The Brain MRI dataset not only contains images that lack standardized dimensions, requiring additional work to achieve optimally sized and high-resolution images, but it is also neither sufficiently large for a study involving a deep neural network nor balanced in terms of its sets and the number of images in each class. The dataset is relatively small, containing only 3,264 images, and is skewed toward the training set, leaving only 13% of the total number of images in the validation set and 12% in the test set. The dataset's classes - no tumor, pituitary tumor, meningioma tumor, and glioma tumor - are also imbalanced. Specifically, the validation set contains a very small number of images in the no tumor class compared to the other three classes, while the test set has a limited number of images belonging to the pituitary tumor and glioma tumor classes.

The inadequacy of the dataset allowed the network to develop a very robust localization phase, as shown by the limited areas of variability, uncertainty, and overlap identified in the heatmaps, but it also led to overfitting. The interpretation phase failed to generalize from the images used for network construction and validation, resulting in lower performance on the test set.

5.2 Future works

The work conducted in this Master's Thesis can be applied to the analysis of any neural network and any XAI method in the field of medical image classification, provided the network architecture allows for the application of the MCD method and the GT classes are available.

Despite the various applications of this work, a key future direction is the identification of one or more methods capable of characterizing the variability and uncertainty of the most important features without the constraint of having the GT class. The methods developed in this thesis could be adapted to function without the GT class, but the assumptions underlying them would need further revision to be applicable in real-world settings. These methods were designed with both the localization phase and the interpretation phase in mind, along with the associated errors that may arise. For practical applications, it is necessary to simplify the manipulation of heatmaps by focusing solely on the localization phase. Additionally, it is crucial to conduct a preliminary analysis of the heatmaps to identify whether there is an indicator that can distinguish between correctly classified heatmaps, discarding the misclassified ones. By following this approach, it is no longer possible to identify the most important features used by the neural network, but only those features present in the heatmaps that are not discarded.

Thus, while it is possible to modify the developed methods to quantify areas of variability and uncertainty in the heatmaps, it is no longer feasible to quantify prediction uncertainty. In this work, the prediction uncertainty identified through softmax is achievable because the GT class is known. Without this data in real-world applications, characterizing the interpretation phase remains highly challenging within AI. The margin can still be utilized, but many of the considerations made in this thesis would no longer apply. The real challenge in analyzing neural networks lies in the difficulty of examining the interpretation phase, which remains the most complex and still largely a "black box."

Bibliography

[1] Z. Senousy, M. M. Abdelsamea, M. M. Mohamed, and M. M. Gaber, “3E-Net: Entropy-based elastic ensemble of deep convolutional neural networks for grading of invasive breast carcinoma histopathological microscopic images,” *Entropy*, p. 620, 2021.

[2] M. S. H. Shovon, M. Mridha, K. M. Hasib, S. Alfarhood, M. Safran, and D. Che, “Addressing uncertainty in imbalanced histopathology image classification of her2 breast cancer: An interpretable ensemble approach with threshold filtered single instance evaluation (sie),” *IEEE Access*, pp. 122 238–122 251, 2023.

[3] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, “A deep learning approach based on explainable artificial intelligence for skin lesion classification,” *IEEE Access*, pp. 113 715–113 725, 2022.

[4] F. Mohammad and S. Al Ahmadi, “Alzheimer's disease prediction using deep feature extraction and optimization,” *Mathematics*, p. 3712, 2023.

[5] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, “AI for radiographic COVID-19 detection selects shortcuts over signal,” *Nature Machine Intelligence*, pp. 610–619, 2021.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why should i trust you?" Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” *The Lancet Digital Health*, pp. e745–e750, 2021.

[8] C. A. Ellis, R. L. Miller, and V. D. Calhoun, “An approach for estimating explanation uncertainty in fMRI dFNC classification,” in *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2022, pp. 297–300.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, pp. 60–88, 2017.

[10] R. K. Singh, R. Gorantla, S. G. R. Allada, and P. Narra, “SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability,” *Plos one*, p. e0276836, 2022.

[11] Z. Senousy, M. M. Gaber, and M. M. Abdelsamea, “AUQantO: Actionable Uncertainty Quantification Optimization in deep learning architectures for medical image classification,” *Applied Soft Computing*, p. 110666, 2023.

[12] M. Ferrante, T. Boccato, and N. Toschi, “BayesNetCNN: incorporating uncertainty in neural networks for image-based classification tasks,” *arXiv preprint arXiv:2209.13096*, 2022.

[13] N. C. Codella, C.-C. Lin, A. Halpern, M. Hind, R. Feris, and J. R. Smith, “Collaborative human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images,” in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1.* Springer, 2018, pp. 97–105.

[14] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso, “Explainable deep CNNs for MRI-based diagnosis of Alzheimer's disease,” in *2020 International Joint Conference on Neural Networks (IJCNN).* IEEE, 2020, pp. 1–8.

[15] E. M. Kenny, C. Ford, M. Quinn, and M. T. Keane, “Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies,” *Artificial Intelligence*, p. 103459, 2021.

[16] Z. Senousy, M. M. Abdelsamea, M. M. Gaber, M. Abdar, U. R. Acharya, A. Khosravi, and S. Nahavandi, “MCUa: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification,” *IEEE Transactions on Biomedical Engineering*, pp. 818–829, 2021.

[17] A. Mobiny, A. Singh, and H. Van Nguyen, “Risk-aware machine learning classifier for skin lesion diagnosis,” *Journal of clinical medicine*, p. 1241, 2019.

[18] X. Zhang, L. Han, L. Han, H. Chen, D. Dancey, and D. Zhang, “sMRI-PatchNet: A novel efficient explainable patch-based deep learning network for Alzheimer's disease diagnosis with structural MRI,” *IEEE Access*, 2023.

[19] J. E. Arco Martín, J. Ramírez Pérez De Inestrosa, F. J. Martínez Murcia, J. M. Gorriz Sáez *et al.*, “Uncertainty-driven ensembles of multi-scale deep architectures for image classification,” 2022.

[20] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, “Attention gated networks: Learning to leverage salient regions in medical images,” *Medical image analysis*, pp. 197–207, 2019.

[21] L. Gaur, M. Bhandari, T. Razdan, S. Mallik, and Z. Zhao, “Explanation-driven deep learning model for prediction of brain tumour status using MRI image data,” *Frontiers in genetics*, p. 822666, 2022.

[22] G. Alicioglu and B. Sun, “A survey of visual analytics for explainable artificial intelligence methods,” *Computers & Graphics*, pp. 502–520, 2022.

[23] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, 2017.

[24] T. Gomez and H. Mouchère, “Computing and evaluating saliency maps for image classification: a tutorial,” *Journal of Electronic Imaging*, pp. 020 801–020 801, 2023.

[25] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Medical Image Analysis*, p. 102470, 2022.

[26] H. G. Ramaswamy *et al.*, “Ablation-cam: Visual explanations for deep convolutional

network via gradient-free localization," in *proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 983–991.

[27] R. A. Zeineldin, M. E. Karar, Z. Elshaer, · . J. Coburger, C. R. Wirtz, O. Burgert, and F. Mathis-Ullrich, "Explainability of deep neural networks for MRI analysis of brain tumors," *International journal of computer assisted radiology and surgery*, pp. 1673–1683, 2022.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[29] K. Kowsari, R. Sali, L. Ehsan, W. Adorno, A. Ali, S. Moore, B. Amadi, P. Kelly, S. Syed, and D. Brown, "Hmic: Hierarchical medical image classification, a deep learning approach," *Information*, p. 318, 2020.

[30] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[31] R. K. Singh, R. Gorantla, S. G. Allada, and N. Pratap, "SkiNet: a deep learning solution for skin lesion diagnosis with uncertainty estimation and explainability," *arXiv preprint arXiv:2012.15049*, 2020.

[32] T. Evans, C. O. Retzlaff, C. Geißler, M. Kargl, M. Plass, H. Müller, T.-R. Kiehl, N. Zerbe, and A. Holzinger, "The explainability paradox: Challenges for xAI in digital pathology," *Future Generation Computer Systems*, pp. 281–296, 2022.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[34] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," *Frontiers in aging neuroscience*, p. 194, 2019.

[35] F. Eitel, K. Ritter, and A. D. N. I. (ADNI), "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9*. Springer, 2019, pp. 3–11.

[36] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[37] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[38] H. Uzunova, J. Ehrhardt, T. Kepp, and H. Handels, "Interpretable explanations of

black box classifiers applied on medical images by meaningful perturbations using variational autoencoders," in *Medical Imaging 2019: Image Processing*, vol. 10949. SPIE, 2019, pp. 264–271.

[39] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, 2018.

[40] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[41] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421*, 2018.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[43] W. Aldhahi and S. Sull, "Uncertain-cam: Uncertainty-based ensemble machine voting for improved covid-19 cxr classification and explainability," *Diagnostics*, p. 441, 2023.

[44] M. Combalia, F. Hueto, S. Puig, J. Malvehy, and V. Vilaplana, "Uncertainty estimation in deep neural networks for dermoscopic image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 744–745.

[45] K. Hauser, A. Kurz, S. Haggenmueller, R. C. Maron, C. von Kalle, J. S. Utikal, F. Meier, S. Hobelsberger, F. F. Gellrich, M. Sergon *et al.*, "Explainable artificial intelligence in skin cancer recognition: A systematic review," *European Journal of Cancer*, pp. 54–69, 2022.

[46] M. Abdar, M. Samami, S. D. Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U. R. Acharya, V. Makarenkov *et al.*, "Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning," *Computers in biology and medicine*, p. 104418, 2021.

[47] J. Shen and H.-W. Shen, "An information-theoretic visual analysis framework for convolutional neural networks," *arXiv preprint arXiv:2005.02186*, 2020.

[48] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems*, pp. 20 014–20 027, 2021.

[49] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[50] R. Seoh, "Qualitative analysis of monte carlo dropout," *arXiv preprint arXiv:2007.01720*, 2020.

[51] M. Abdar, M. A. Fahami, S. Chakrabarti, A. Khosravi, P. Pławiak, U. R. Acharya, R. Tadeusiewicz, and S. Nahavandi, "BARF: A new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification," *Information Sciences*, pp. 353–378, 2021.