



Universiteit
Leiden
The Netherlands

Data Science & Artificial Intelligence

Evaluating the Reliability of Explainable AI Methods for CNN and Vision Transformer Models in COVID-19 Chest X-ray Classification

Talya Seryano

Supervisor:

Christos Athanasiadis

Second Supervisor: Lifeng Han

Bachelor Thesis

Leiden Institute of Advanced Computer Science (LIACS)

www.liacs.leidenuniv.nl

July 1, 2026

Abstract

Deep learning algorithms have shown great performance on medical image classification problems such as COVID-19 detection from chest x-rays. However, transparency and trust in them is limited because of their black-box nature. This thesis looks at how well explainable artificial intelligence(XAI) techniques offer reliable and meaningful insights on the decision-making processes of a Vision Transformer (ViT) and a convolutional neural network (CNNs) for COVID-19 chest X-ray classification. A ResNet-50 CNN and a Vision Transformer were trained and evaluated on a three-class chest X-ray dataset containing COVID-19, Non-COVID and Normal cases. Three approaches of explanations were investigated. These are Local Interpretable Model-agnostic Explanations (LIME), Layer-wise Relevance Propagation (LRP) for both models, while Attention Rollout was evaluated for Vision Transformer. Explanation quality was evaluated using faithfulness, sensitivity and complexity metrics. This evaluation is complemented with qualitative analysis of correctly classified and misclassified samples. An artifact robustness study was also carried out on models and explanation methods to investigate the influence of peripheral image regions through border and corner perturbations. Findings suggest that different explanations have very different strengths in representing model behavior. LRP offered the most balanced performance by showing low sensitivity with greater faithfulness where LIME produced localized but less stable explanations. Despite the stability, Attention Rollout showed very low faithfulness indicating that attention alone may not accurately reflect model reasoning. Additionally, the Vision Transformer showed more resilience to peripheral image perturbations than the CNN. The thesis concludes that LRP offers the most reliable and robust understanding of the decision-making processes of both models in this application. More broadly, these results underline the need of assessing explanation techniques with several supplementary criteria, contributing to the development of more transparent and trustworthy AI systems for medical imaging.

Contents

1	Introduction	1
1.1	AI in Medical Imaging	1
1.2	The Problem: Black-Box models	1
1.3	Explainable AI (XAI)	2
1.4	The Research Gap	3
1.5	The Research Question	4
1.6	Thesis Overview	4
2	Background and Related Work	4
2.1	Deep Learning for Chest X-ray Classification	4
2.2	Explainable AI Methods	6
2.3	Evaluation of Explainability Methods	8
2.4	XAI in Medical Imaging	9
3	Methodology and Experiments	11
3.1	Research Design	11
3.2	Dataset	11
3.3	Deep Learning Models	12
3.3.1	CNN Model: ResNet-50	12
3.3.2	Vision Transformer Model	13
3.4	Explanation Methods	14
3.4.1	LIME	14
3.4.2	Layer-wise Relevance Propagation (LRP)	15
3.4.3	Attention Rollout	16
3.5	Quantitative Evaluation Framework	17
3.5.1	Faithfulness	18
3.5.2	Sensitivity	18
3.5.3	Complexity	18
3.6	Qualitative Evaluation	19
3.7	Artifact Robustness of Models	19
3.8	Artifact Robustness of Explanation Methods	20
3.9	Experimental Setup	21
4	Results and Analysis	21
4.1	Evaluation of Explanation Methods	21
4.2	Class-wise Analysis of CNN Explanations	22
4.3	Qualitative Analysis	23
4.4	Artifact Robustness Study for Models	27
4.5	Artifact Robustness Study for Explanation Methods	28
5	Discussion	31
6	Limitations	33

7 Conclusions and Further Research	34
References	38

1 Introduction

1.1 AI in Medical Imaging

Artificial intelligence (AI) has become a very important tool in medical imaging due to its ability to automate the analysis of large amounts of medical data and support clinical decision-making processes. In recent years, machine learning and deep learning techniques have shown great success in various medical imaging methods for tasks such as disease detection, image classification and segmentation [AR24]. Moreover, AI has been successfully applied across a wide range of imaging modalities, including chest X-rays, computed tomography (CT), magnetic resonance imaging (MRI) and retinal imaging[MB24]. This advancement has been largely fueled by developments in deep learning, which allow models to automatically learn complicated feature representations from picture data instead of depending on manually created features[LBH15]. Among deep learning architectures, Convolutional Neural Networks (CNNs) have been the dominating strategy for medical image analysis due to its capacity to learn hierarchical visual information through convolutional operations [LBH15]. Their strong performance has established CNNs as one of the most widely adopted AI techniques in medical imaging[CEVV23]. The COVID-19 pandemic greatly motivated the use of AI in medical imaging. Chest X-rays became an important imaging modality because they are widely available, cost-effective and expose patients to less radiation than computed tomography (CT) scans [CEVV23]. Deep learning models were shown by various research[CEVV23],[CC24],[KBB23] to be able to appropriately detect COVID-19-related abnormalities from chest X-rays, therefore underlining their possible use as decision-support tools for doctors. Vision Transformers (ViTs) have recently shown themselves to be a good substitute for CNN-based designs[DBK⁺20]. Unlike CNNs, which mostly highlight local picture characteristics with convolutional filters, ViTs separate images into patches and employ self-attention methods to simulate connections across the whole image[DBK⁺20]. Growing curiosity in using Vision Transformers for medical imaging applications including COVID-19 identification, cancer detection and thoracic image analysis stems from this capacity to record global contextual information[ADAS⁺25]. However, despite being very efficient and achieving high predictive accuracy, these models struggle to adapt and get acceptance in the field of healthcare where decisions can lead to disastrous consequences[MB24]. Therefore, it is very important to understand how AI systems arrive at their predictions.

1.2 The Problem: Black-Box models

The major challenge of these outperforming deep learning models in medical imaging is their lack of transparency and interpretability. The complexity of recent deep learning methods contribute to their decision-making processes being hard to understand and access, which has been called the “black-box” problem [XUD⁺19],[MB24]. In high-stakes domains, such as in medicine, where incorrect predictions can negatively affect patient outcomes and safety, there are concerns on the lack of interpretability of the systems [FDAC24][XUD⁺19]. It is extremely important that medical professionals understand and justify why AI- based tools made a specific decision before blindly trusting it. The black-box problem is not only about trust, but also about verification. If a model gives a prediction without an explanation, it is difficult to determine if the conclusion is based on medically relevant data or on patterns unrelated to the disease[RPDMMR25],[MB24]. This is important because excellent classification accuracy does not always indicate that a model has

discovered clinically significant features [DJL20]. Chest radiography models occasionally depend on shortcut signals like acquisition variations, dataset-specific markers or other non-clinical signals instead of concentrating on disease-related anomalies [DJL20], [ZBL⁺18]. Hence, these issues may go unnoticed without clear explanations and may even cause loss of life. Such patterns might help a given dataset perform better, but they often compromise a model’s capacity to generalize to new clinical situations. Chest X-ray classification especially highlights this problem, as great predictive ability does not always mean a model has learned clinically relevant traits. Zech et al. showed that pneumonia diagnosis models sometimes had trouble generalizing across hospitals, which indicated that they had acquired institutional-specific traits in addition to illness-related information [ZBL⁺18]. DeGrave et al. also found that artificial intelligence programs designed for COVID-19 detection typically relied on quick signals instead of actual radiographic data of illness [DJL20]. Their study found that elements like image borders, annotations, laterality markers and other acquisition-related aspects unrelated to COVID-19 pathology affected model predictions. For this reason, understanding whether a model focuses on clinically relevant pulmonary regions or on unintended peripheral features is essential when evaluating the trustworthiness of medical AI systems. These black-box models being unable to provide clear reasonings and explanations for their predictions have made it difficult to trust them. Therefore, increasing the interpretability and explainability of artificial intelligence systems in creating reliable medical decision support systems has become a significant and crucial research topic.

1.3 Explainable AI (XAI)

Deep learning’s expanding application in healthcare has raised the demand for techniques that improve human users’ comprehension of model decisions. Explainable Artificial Intelligence (XAI) has emerged as a novel field of study that focuses on improving the interpretability and transparency of machine learning models. XAI methods’ aim is to make model decisions understandable to humans by providing insights into the factors that influence generated predictions [XUD⁺19]. Explainability is especially important in medical imaging since clinicians must understand and interpret the reasoning behind AI-assisted diagnosis before integrating them into medical practices [RPDMR25]. Furthermore, XAI is also capable of identifying potential model biases, errors and can uncover shortcut learning behavior [XUD⁺19] [MB24]. This is particularly helpful for analyzing chest X-rays, as prior research has demonstrated that deep learning models may occasionally rely on non-clinical image characteristics unrelated to the underlying disease [DJL20]. In light of this, many XAI approaches have been developed. These approaches play a crucial role in explaining deep learning models used for image analysis by highlighting the regions of an image that contribute most to the model’s predictions. Common techniques include perturbation-based, attribution-based and attention-based approaches [MB24]. These aim to provide visual or feature-level explanations for model outputs [BSN⁺23]. Perturbation-based techniques explain predictions by deliberately altering elements of an input and tracking how the model’s output varies. Relevance scores are assigned to input features using attribution-based algorithms, showing the degree to which each item contributes to a prediction. Attention-based methods use transformer models’ attention processes to show which areas of the image are given the most attention during inference [RSG16], [BBM⁺15], [AZ20]. In chest X-ray analysis, techniques such as LIME, Layer-wise Relevance Propagation (LRP) and attention-based methods have been used to visualize model reasoning and improve understanding of model behavior [MB24], [ISST24], [KBB23]. These methods offer a chance to evaluate if predictions

are backed by medically significant evidence and whether possible shortcut behavior may be found by increasing the transparency of model decisions. However, explanation methods are not without limitations. Various methods could provide rather different justifications for the same prediction, so it’s challenging to tell which explanation most closely captures the model’s actual line of thinking process[MB24]. Furthermore, explanations that appear visually plausible are not necessarily faithful representations of model behavior[RPDMR25]. Consequently, evaluating the reliability and robustness of explanation methods remains an important challenge in explainable AI research.

1.4 The Research Gap

Explainable AI is becoming very important in medical imaging but there are still certain limitations in the existing literature. Many studies focus primarily on developing novel explanation strategies or providing explanations in qualitative way[MB24]. However, far less research looks at systematic evaluation of explanation quality and reliability. Hereby, it is still difficult to determine whether hypotheses actually reflect model decision- making process and present a detailed comparative analyzes of different XAI approaches. Recent research[KBB23] has examined explainability especially for Vision Transformers. Komorowski et al. developed a system for assessing ViT explanations employing the metrics of faithfulness, sensitivity and complexity and showed that Layer-wise Relevance Propagation (LRP) yields more trustworthy explanations for chest X-ray classification than LIME and Attention Rollout. The writers also contend that in order to provide uniform standards for ViT explanations in medical imaging, it is crucial to have a solid evaluation criteria. Although this study is a step toward consistent assessment, it focuses only on Vision Transformers and offers no understanding of how explanation techniques work across different models. On the other hand, another recent study[NMHA23] has compared convolutional neural networks and Vision Transformers for medical image classification. Nevertheless, these comparisons mostly highlight predictive performance instead of their explanations’ interpretability and reliability. Therefore, it is still unknown if, when applied to the same clinical task, explanation techniques give reliable and consistent insights across CNNs and ViTs. Another aspect is the detection of shortcut learning and spurious correlations. Medical imaging models that rely on non-clinical image features, dataset artifacts or acquisition-related variables instead of disease-relevant information have demonstrated strong prediction performance in earlier research[DJL20]. Therefore, it is important to investigate whether explanations can reliably reveal reliance on irrelevant image regions. This thesis addresses these gaps by extending existing ViT explainability evaluation [KBB23] beyond a single architecture and performing a direct comparison between a ResNet-50 convolutional neural network and a Vision Transformer on the same COVID-19 chest X-ray classification task. The study evaluates LIME, LRP and Attention Rollout(for ViT) using quantitative metrics of complexity, sensitivity and faithfulness on both models. This study also examines explanation robustness through qualitative evaluations of misclassified samples and peripheral artifact experiments. The peripheral imaging experiments investigate if explanation techniques may uncover potential shortcut learning and highlight model dependence on non-clinical picture regions by introducing artificial perturbations at the edges and corners of chest X-rays. The thesis offers a more comprehensive assessment of explanation reliability across deep learning models for medical imaging through combined evaluation.

1.5 The Research Question

Given the increasing prevalence of artificial intelligence in the analysis of medical images and the difficulties in determining the reliability of these descriptions, this academic study aims to examine and measure the effectiveness and reliability of explainable AI techniques applied to AI systems that analyze medical images. Specifically, the research addresses the explainability of AI used in the analysis of chest X-rays and investigates the extent to which these approaches contribute to making sense of the results produced by deep learning systems. The main question guiding this research is:

To what extent do explanation methods provide reliable and robust insights into the decision-making processes of a Convolutional Neural Network(ResNet-50) and a Vision Transformer model for COVID-19 chest X-ray classification?

To address this question, the following sub-questions are considered:

1. How do explanation methods (LIME, LRP and Attention for Vision Transformers) compare in terms of quantitative evaluation metrics such as complexity, sensitivity and faithfulness between the ResNet-50 and Vision Transformer models?
2. How do different explanation methods qualitatively behave when explaining both correctly classified and misclassified COVID-19 chest X-ray predictions generated by CNN and Vision Transformer models?
3. How robust are the explanation methods to peripheral image perturbations?

1.6 Thesis Overview

This bachelor thesis investigates explainable artificial intelligence methods in the context of chest x-ray imaging and evaluates their effectiveness in providing reliable interpretations of deep learning models. The research is conducted at the Leiden Institute of Advanced Computer Science (LIACS) and is supervised by Christos Athanasiadis. The rest of the thesis is structured as follows. Section 2 presents the background and related work, introducing concepts such as deep learning and explainable AI methods along with evaluation designs. Section 3 describes the methodology of this thesis. Section 4 presents the results and the analysis for the reliability and robustness of XAI in medical imaging. Section 5 and 6 discusses the findings and limitations of the study respectively. Finally, Section 7 concludes the thesis and outlines potential directions for future research.

2 Background and Related Work

2.1 Deep Learning for Chest X-ray Classification

Convolutional Neural Networks (CNNs) are now the most frequently preferred approach when it comes to medical imaging. The main reason for this is that these networks can learn hierarchical feature representations from raw image data. Early breakthroughs in deep learning showed that increasing network depth significantly improves performance in visual recognition tasks since deeper architectures can capture very complex patterns and abstraction [HZRS15]. In light of these developments, CNN based methods in chest x-ray analysis have become widely applied and have

yielded results equivalent to, or even sometimes better than those obtained by physicians [RIZ⁺17]. For instance, the CheXNet model, which is a 121-layer deep convolutional network, was shown to outperform average radiologists in detection of pneumonia on chest X-rays [RIZ⁺17], highlighting the great potential of deep learning systems in clinical decision making . These developments strongly demonstrated the potential of CNNs for automated diagnosis using medical images, particularly in areas such as determining the type and location of thoracic diseases. During the pandemic, especially when the need for rapid and widespread screening increased, Convolutional Neural Networks (CNNs) were frequently used on chest X-ray images for COVID-19 diagnosis. Since the COVID-19 virus leaves distinct traces on lung X-rays, deep learning algorithms showed success in distinguishing between infected and healthy individuals [CC24]. Research has shown that Convolutional Neural Network (CNN) based designs, particularly Residual Networks (ResNets) achieve high performance rates in COVID-19 datasets and demonstrate their success in classifying medical images [CC24]. The experiments on COVID-19 X-ray datasets have reported accuracy levels above 95% [CC24] , indicating that CNNs can serve as reliable tools for assisting diagnosis. These applications highlight once again how important the interpretability is, as clinical experts need to understand which parts of the image influence the model’s decision. More recently, Vision Transformers (ViTs) emerged as a powerful alternative to CNNs for image classification tasks. Unlike CNNs, which rely on convolutional filters to extract local features, ViTs process images as sequences of patches and use a Transformer encoder with self-attention [DBK⁺20]. ViTs can model long-range spatial relationships that local convolutional networks might not catch as well since self-attention let every image patch to interact directly with every other patch [DBK⁺20]. This is very helpful in chest X-rays where diagnostically important results can spread over several lung areas instead of being limited to a single local area [DBK⁺20]. ViT models have already had impressive success in image classification benchmarks and yielded positive results in medical imaging fields such as chest X-ray classification [KBB23] becoming an important subject for explainability research. Moreover, due to being a relatively new architecture, the interpretability of ViT models is less proven compared to CNNs ,which makes them also very relevant for researches examining the reliability of various explanation methods[KBB23]. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) exhibit fundamentally different approaches for analyzing information from images. CNNs rely on convolutional operations that is more focused on local spatial patterns which makes them highly effective at capturing fine-grained features and performing well even on relatively smaller datasets because of their built-in inductive biases [NMHA23]. On the other hand, ViTs use self attention mechanisms to model global relationships across image patches, enabling them to capture long-range dependencies more effectively even though this often requires larger datasets for optimal performance [NMHA23],[DBK⁺20]. Based on comparisons in [NMHA23], medical imaging tasks such as COVID-19 detection from chest X-rays show that both approaches can achieve similarly high performance where CNN-based models such as EfficientNet and ViT-based models like SegFormer reaches nearly same accuracy levels. However, the choice varies depending on the specific application. CNNs generally provide more stable computation using less data, while ViTs offer greater flexibility in understanding the overall context, which can be advantageous for complex visual patterns [NMHA23].

2.2 Explainable AI Methods

The increasing use of deep learning in important fields like medical imaging has highlighted the need for techniques that can help to understand and trust model predictions. Although modern models have great predictive power, their internal decision-making processes remain opaque which has led to the creation of Explainable Artificial Intelligence (XAI) methods providing human-interpretable insights into model behavior [XUD⁺19],[AR24],[ANK⁺26]. How explanations are produced is the foundation of a typical categorization of XAI approaches. This includes whether they are intrinsic or post-hoc, local or global and according to the mechanism used to generate explanations(See Table 1. In [XUD⁺19],[AR24], XAI is often presented as a collection of post-hoc or intrinsic approaches meant to bridge the knowledge between model complexity and human interpretability, hence facilitating both local explanations for individual predictions and global comprehension of model behavior. Post-hoc techniques predominate in medical imaging because they can be applied to complex deep learning models without modifying their architecture [AR24]. They produce explanations like saliency maps or heatmaps that draw attention to crucial areas in an image influencing the model’s prediction [AR24]. Based on the mechanism used to generate explanations, post-hoc methods are commonly grouped into perturbation-based, backpropagation(or attribution)-based and attention-based techniques. These three categories form the basis of the explanation methods investigated in this study. Perturbation-based approaches analyze model sensitivity to input changes. These approaches explain predictions by purposefully altering portions of the input and monitoring the changes in the model output. The underlying assumption is that image regions causing larger prediction changes are more influential for the decision. Occlusion-based methods, for example, eliminate or cover sections of an image to assess their impact on the prediction, therefore identifying important features [AR24]. Similarly, [RSG16] suggested LIME as a method that uses an interpretable surrogate to locally approximate behavior of the model, therefore enabling explanations for particular predictions. Since perturbation-based techniques only need access to a model’s input and outputs, they are model-agnostic and so may be used on a broad spectrum of machine learning models, including CNNs and Vision Transformers, without knowing how they work internally. This adaptability allows them to be used for comparing explanations across totally different model designs. However, earlier research[AR24] has shown that they have a high computational cost and may be unstable because they rely on sampling strategies and perturbation design. The second main group consists of attribution-based approaches, which explain predictions by propagating information backward through a trained neural network to determine the contribution of every input feature to the ultimate prediction. These techniques are model-specific since they need access to the inner calculations of the model. They all give the input relevance scores but they vary in how these scores are computed. Grad-CAM [SCD⁺16] is a popular technique that creates class-discriminative localization maps by calculating the gradients of the projected class with regard to the feature maps of the final convolutional layer. These gradients help to weight the feature maps generating a coarse heatmap emphasizing the areas of the image most important for prediction [SCD⁺16]. Furthermore, by integrating the gradients along a path from a baseline image to the original input, Integrated Gradients instead directly attribute value to the input pixels[STY17]. More complex attribution techniques such as Layer-wise Relevance Propagation (LRP)[BBM⁺15], takes a different approach by redistributing the prediction score backwards across the network layers based on conservation rules. Every layer’s relevance allocated to a neuron is equally distributed to the neurons in the preceding layer depending on their contribution to its activation. This guarantees that the total

relevance is maintained until it reaches the input pixels [BBM⁺15], [CGW20], [KBB23]. Medical imaging uses attribution-based approaches extensively since they produce natural visualizations consistent with how doctors interpret images. But existing literatures [STY17], [CGW20] mentioned that they can be sensitive to noise, architectural decisions and implementation details, which could jeopardize their interpretability and robustness. A third group consists of attention-based techniques. Particularly in transformer-based designs, attention mechanisms offer a natural kind of explanation by showing how several components of the input interact during prediction [KBB23]. Nevertheless, [KBB23], [JW19] caution that although attention weights can be useful, raw attention weights alone should not be interpreted as faithful. This is because information becomes progressively mixed across successive Transformer layers and residual connections [JW19]. Therefore, they might not accurately reflect the actual significance of features or the model’s reasoning processes. This limits their effectiveness as independent explanations. Moreover, Abnar and Zuidema [AZ20] introduced Attention Rollout. It propagates attention weights throughout all Transformer layers while expressly considering residual connections. Attention Rollout estimates the information flow from the input patches to the final prediction by recursively combining attention matrices from subsequent layer rather than viewing the attention of a single layer. Including the cumulative impact of several attention layers usually yields more insightful and consistent visualizations than raw attention maps. More recently, Chefer et al. [CGW20] presented a state-of-art Transformer-specific Layer-wise Relevance Propagation (LRP) framework that creates class-specific explanations by combining relevance propagation with gradient information and attention weights. Unlike raw attention maps, this approach redistributes the model’s prediction score backward through all Transformer layers while maintaining relevance. This produces explanations that more closely represent the model’s decision-making process. Moreover, it manages important Vision Transformer architectural elements, such as skip connections and attention operations, resulting in more accurate attribution maps. More recently, explainability research has shifted towards developing methods that are specifically tailored to the underlying architecture of deep learning models rather than relying on generic explanation techniques. For instance, Jo et al. [JJP25] suggested GMAR, Gradient-Driven Multi-head Attention Rollout, which enhances traditional Attention Rollout [AZ20] by weighting individual attention heads depending on their gradient-derived importance before aggregating attention over layers. GMAR creates explanations that more closely capture the model’s class-specific reasoning and boosts explanation faithfulness over conventional Attention Rollout by understanding that several attention heads contribute differently to a prediction. In parallel, explainability techniques for CNNs still improve. For example, Eigen-CAM uses principal component analysis of convolutional feature maps instead of gradient-based weighting to produce strong localization maps without the need for backpropagation through the target class [MY20]. These advances show that present XAI studies is progressively focused on creating architecture-aware explanation tools that offer more trustworthy and accurate interpretations of deep learning models. All in all, XAI is essential for increasing the clarity, dependability and practical usefulness of deep learning models in real-world situations. In the framework of medical imaging it allows doctors to better grasp model predictions, spot possible mistakes or biases and develop trust in AI-assisted decision-making systems [ANK⁺26]. Each of the existing XAI approaches offers a unique viewpoint on model behavior, with both benefits and drawbacks. These variances underscore the need for methodical evaluation and comparison of explanatory techniques. Thus, the general consensus is that no one technique is enough to fully understand model behavior, especially in critical applications like medical image analysis.

Table 1: Common categorizations of Explainable Artificial Intelligence (XAI) methods.

Categorization		Categories	Description
Timing		Intrinsic / Post-hoc	Intrinsic methods are inherently interpretable by design, whereas post-hoc methods explain the behavior of an already trained model.
Scope		Global / Local	Global explanations describe the overall behavior of a model, while local explanations explain individual predictions.
Model	depen-	Model-specific /	Model-specific methods require access to the model internals, whereas model-agnostic methods can be applied to any predictive model.
dency		Model-agnostic	
Explanation mechanism		Perturbation-, Attribution-, Attention- based	Methods differ according to how explanations are generated. Perturbation-based methods modify the input, attribution-based methods propagate importance scores through the model and attention-based methods analyze attention distributions

2.3 Evaluation of Explainability Methods

In fields like medical imaging where trust and reliability are crucial, evaluating explainable artificial intelligence (XAI) techniques remains a fundamental issue. Unlike traditional machine learning problems, there isn’t usually a single accepted ground-truth explanation against which explanation quality may be directly evaluated[HWB⁺22]. As was emphasized in recent research[HWB⁺22], the lack of a universally agreed upon definition of what constitutes a ”correct” explanation complicates the creation of standard assessment methods. Early methods for evaluating explanations mostly depended on qualitative human judgment, whereby explanations were visually examined or evaluated via user studies. Such techniques are by nature arbitrary, though, and provide no assurance that an explanation accurately depicts the underlying model behavior. Actually, it has been demonstrated by [GRB⁺22] that explanations that seem logical to people might not be connected at all to the way the model makes judgments internally. This restriction has driven the creation of quantitative evaluation methods meant to objectively evaluate explanation quality. Perturbation-based metrics make up one of the main categories of evaluation techniques. They evaluate explanations by altering input features and tracking model output variations. For instance, the explanation is regarded more faithful if deleting extremely pertinent features results in a notable decrease in prediction confidence. Recent research in medical imaging[HVH22], extending this concept, suggests employing insertion and deletion metrics as casual evaluation criteria whereby crucial areas are either progressively deleted or added to the input picture to evaluate their influence on model behavior. As shown in [HVH22], these measures monitor how similarity ratings vary when significant areas are perturbed, therefore offering a quantitative approach to assess whether explanations accurately capture meaningful traits. Stronger correspondence between highlighted areas and model predictions implies improved explanations, as shown by higher insertion scores and lower deletion scores [HVH22], [GRB⁺22]. But the techniques have certain drawbacks as perturbations could produce implausible data distributions and cause misleading assessment outcomes. To deal with this,

prior research[BWM20] offers three main criteria for evaluating feature-based explanations. These are sensitivity, faithfulness and complexity. Faithfulness evaluates if the explanation accurately represents the model’s decision-making process, sensitivity (or robustness) examines whether explanations stay consistent under minor input changes and complexity assesses how simple the explanation is for people to comprehend[BWM20]. These standards have been generally adopted in the literature such as [KBB23] in the context of ViT explanations in medical imaging and serve as the foundation for several current evaluation systems. Another significant line of research[AGM⁺18] highlights the limits of qualitative assessment and has raised concerns on the reliability of current explanatory approaches. Adebayo et al.[AGM⁺18], for example, illustrate how some frequently used saliency techniques generate comparable results even when model parameters or training data are randomly assigned, therefore implying that these explanations could not rely on the learned model at all. Their suggested ”sanity checks”, which are based on model and data randomization, show that visual similarity by itself isn’t enough to ensure meaningful explanations. Robustness is very important, especially in this case. Minor input changes should ideally not cause major explanation changes. In contrast, Alvarez-Melis and Jakkola show that many well-known interpretability methods are somewhat contradictory, providing significantly different justifications for similar inputs[AMJ18]. This instability casts doubt on the reliability of the explanations, particularly in sensitive applications when consistency is essential. This unpredictability is very unsettling in the medical field since it undermines the confidence in artificial intelligence systems used in clinical decision making. The X-MIR research[HVH22] further emphasizes that saliency techniques might occasionally concentrate on irrelevant areas like background artifacts rather than medically important regions, hence underscoring the need for careful review. Its underlined that no one criterion is enough to properly judge explanations. Gevaert et al. show that different evaluation metrics often measure different underlying properties and may lead to conflicting conclusions about which method performs the best[GRB⁺22]. Moreover, since findings on one dataset don’t always apply to others, assessment ought to be context-dependent instead of universal. More methodical assessment systems have been created to help solve these problems. The Quantus toolkit offers a structured compilation of measurements across numerous categories, including faithfulness, robustness and complexity, hence allowing more consistent and repeatable assessment of explanation techniques[HWB⁺22]. Such models encourage a more whole evaluation by combining several criteria rather than depending on one measurement since several uses may need various explanations of quality. The ViT evaluation study[KBB23] finds that various explanation methods might differ greatly in terms of faithfulness and sensitivity with gradient-based approaches. For example, LRP usually offers more trustworthy explanations, though occasionally at the expense of higher complexity. Therefore, evaluation ought to be regarded as a multi-dimensional issue taking several complementing criteria into account. This mix allows for a more thorough and dependable comparison of XAI techniques, especially in medical imaging settings where both interpretability and accuracy are paramount.

2.4 XAI in Medical Imaging

In medical imaging explainable artificial intelligence(XAI) is highly important since AI models are used progressively in clinical settings where a wrong prediction might impact patient diagnosis, treatment and patient safety. The need for interpretability in healthcare is greater than in many other fields since doctors must understand, challenge and maybe reject model outputs before apply-

ing them in practice[ATE18]. Holzinger et al. contend that medical AI systems should make their judgments clear and retraceable, given that medical specialists deal with complicated and varied data and must be able to understand how and why a decision was reached[HBPK17]. According to Tjoa and Guan, explainability is closely linked to accountability in medical AI because clinical responsibility ultimately remains with healthcare professionals rather than the AI system[TG19]. XAI should thus not only make models more understandable but also aid in evaluating if predictions are based on clinically meaningful evidence in medical imaging. Chest X-ray interpretation is an obvious illustration of this need. Deep learning models have demonstrated excellent results in radiology applications, including pneumonia detection in CheXNet[RIZ⁺17], where a DenseNet model was trained on ChestX-ray14 dataset and used class activation maps to highlight significant image areas. CheXNet’s[RIZ⁺17] heatmaps were utilized to highlight areas that helped to identify disorders such as pneumonia, nodules, masses and pleural effusion. Similar explanation-based techniques have also been applied for the categorization of chest X-rays for COVID-19. Rather than treating the model as a black box, COVID-Net[WW20] included explainability to audit whether the network was using meaningful information from chest X-rays. LRP was also applied to the classification of COVID-19 chest X-rays by Bassi and Attux[BA21], who claimed that heatmaps can reveal which specific regions contributed to network classification and whether non-clinical components such as text on the X-rays affected the predictions. More recent study[NMHA23] comparing Vision Transformers and CNNs for COVID-19 detection also show that explainability can assist clinicians by highlighting the image regions that most influenced a prediction. These visual explanations enable radiologists to confirm whether the model concentrates on clinically significant lung structures or on possibly confusing artifacts, therefore increasing confidence in AI-assisted decision-making rather than offering an automatic diagnosis. Still, using explanations in medical imaging has serious drawbacks and can be misleading if they are not carefully evaluated. Even if they are partial, unstable or not true to the model’s actual reasoning, visual explanations could seem clinically convincing[LWB⁺19]. This is especially true for medical imaging, where a model may pick up artifacts from picture acquisition, hospital source or preprocessing rather than disease-related characteristics. DeGrave et al.[DJL20] demonstrated this issue explicitly in the classification of COVID-19 chest X-rays. Many models that seemed to work well used shortcuts like laterality markers, image borders, patient placement and other acquisition-related artifacts instead of actual COVID-19 pathology[DJL20]. Their saliency maps occasionally highlighted lung fields, but they also often highlighted areas outside the lungs, indicating that high accuracy alone is insufficient to demonstrate that a medical AI system has learned clinically relevant characteristics. Another limitation is that explanation methods themselves may lack consistency or faithfulness. Kindermans et al.[KHA⁺17] demonstrate that certain saliency approaches might generate visually realistic maps regardless of their sensitivity to the trained model or the data-generating process, implying that visual inspection alone might be unstable. It was shown that certain saliency techniques are sensitive to straightforward input changes that don’t affect the model’s prediction. Even when the model makes the exact same choice, the explanation may differ and this raises questions about the robustness and reliability of these techniques. Attention-based explanations also require careful consideration. Attention weights explain the information flow across image patches, but they don’t always show which areas had the greatest impact on the final prediction. Jain and Wallace [JW19] showed that different attention distributions can result in identical model outputs, indicating that attention by itself is not a reliable explanation of model reasoning. This matter is very important for Vision Transformers because attention maps are mostly appealing due to their ease of visualization

even though they do not adequately convey the logic of the model. Because of this, medical imaging research has moved beyond only visualizing explanations to methodically evaluating their quality. According to Komorowski et al., saliency maps are frequently presented in medical imaging studies without being quantitatively tested, which restricts their reliability and quality[KBB23]. Their research uses faithfulness, sensitivity and complexity to assess ViT explanations for chest X-ray classification, demonstrating that explanation quality varies between classes and approaches.

Therefore, although many researches have investigated CNN and Vision Transformer models for the classification of chest X-rays and suggested a number of explanation techniques, systematic comparison of explanation reliability across model types is still lacking especially utilizing both quantitative metrics, qualitative and robustness analysis.

3 Methodology and Experiments

3.1 Research Design

Using a comparative experimental research design, this study examines the reliability and interpretability of XAI techniques in medical image classification. The study looks at how several explanation techniques shed insight into the decision-making processes of deep learning models and concentrates on chest X-ray analysis for COVID-19. Two inherently different model architectures: a convolutional neural network (CNN) and a Vision Transformer (ViT)[KBB23] are compared to address this, along with several explanation techniques symbolizing various XAI methodologies. The study uses both quantitative and qualitative evaluation methods. Qualitative analysis is utilized to visually examine explanation behavior on correctly classified and misclassified chest X-ray images, while quantitative analysis is used to evaluate explanation quality through established evaluation metrics of complexity, sensitivity and faithfulness[HWB⁺22]. Moreover, an artifact robustness study is conducted by introducing border and corner perturbations to chest X-ray images to assess whether models and explanation method’s predictions relied on peripheral image information. This study seeks to offer an structured assessment of the trustworthiness of current XAI approaches in the context of medical imaging by combining architectural comparison with quantitative evaluation, qualitative analysis and robustness viewpoints.

3.2 Dataset

This study uses the COVID-QU-Ex dataset[TCK⁺21], a publicly available chest X-ray dataset developed for COVID-19 classification tasks and previously used in explainable AI research[KBB23]. The dataset consists of 33,920 chest X-ray images, including 11,956 COVID-19 cases, 11,263 non-COVID infections (viral or bacterial pneumonia) and 10,701 healthy chest X-rays. This three-class classification setting was considered particularly suitable for the present study, as distinguishing COVID-19 from other respiratory infections and healthy cases presents a more clinically realistic and challenging task than binary classification. To ensure consistency with prior work[KBB23] and maintain reproducibility, the dataset authors’ recommended 65/15/20 split was adopted for the training, validation and test sets, respectively. Generating and evaluating explanations, particularly for perturbation-based methods such as LIME, is computationally intensive. Therefore,

all explainability tests were performed on the same balanced selection of 300 test images, with 100 samples per class, to guarantee an equitable comparison across architectures and explanation techniques.

3.3 Deep Learning Models

3.3.1 CNN Model: ResNet-50

ResNet-50 (See Figure 1) is used in this work as a convolutional baseline to classify chest X-rays. Because of their capacity to automatically learn hierarchical spatial features from image data, convolutional neural networks (CNNs) have been extensively used in medical image analysis, making them a good baseline for comparison with transformer-based methods. ResNet-50 was chosen for its proven image classification performance[HZRS15] and its residual learning architecture, which reduces vanishing gradient problems in deeper networks by adding skip connections between levels. Weights from the ImageNet dataset were used to initialize the CNN model, then fine-tuned for the three-class COVID-19 chest X-ray classification. With this transfer learning technique, the model develops on broad visual elements discovered from a big dataset and customize them to fit the features of chest X-ray images. Using pretrained weights also lowers the amount of training needed and benefits over building a deep architecture from scratch for medical imaging data. The first fully connected classification layer was replaced with a new output layer with three neurons, one for each target class: Normal, Non-COVID pneumonia, COVID-19. To meet ResNet-50’s input requirements, all input chest X-ray pictures were resized to 224×224 pixels and then normalized using the mean and standard deviation values from ImageNet. The Adam optimizer was used to train the model with a learning rate of 1×10^{-4} . Using the cross-entropy loss function improved classification accuracy. Using a batch size of 8 for training and 32 for evaluation, training was carried out for a maximum of 10 epochs. To guarantee methodological consistency with the Vision Transformer experiment[KBB23], the same official dataset split was employed, which included distinct training, validation and testing sets. Model selection was guided by validation accuracy. The checkpoint exhibiting the best validation performance was kept for next explanation building and evaluation.

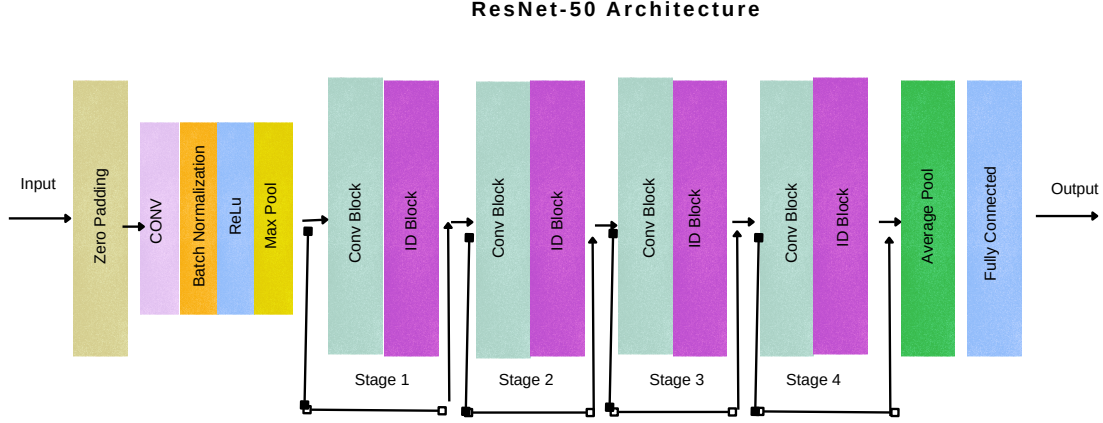


Figure 1: Overview of the ResNet-50 architecture used in this study. The network begins with an initial convolutional layer followed by four residual stages containing convolutional and identity blocks. Skip connections enable residual learning by adding the input of a block directly to its output, facilitating gradient propagation during training. The network concludes with global average pooling and a fully connected layer adapted for three-class chest X-ray classification.

3.3.2 Vision Transformer Model

To offer a comparison with the convolutional baseline, the transformer-based model used in this research was a Vision Transformer (ViT) design (See Figure 2). Vision Transformers divide images into a sequence of fixed-size patches that are treated as input tokens for a transformer encoder, unlike convolutional neural networks, which process images through local convolutional filters. The model may be able to capture long-range dependencies and global contextual interactions throughout the image by using self-attention techniques, which might be helpful in medical imaging applications where diagnostically relevant patterns are spatially dispersed. The implementation was adapted from the explainability benchmarking repository by Komorowski et al.[KBB23] and is used as the experimental starting point for this work. This research employs the base-sized Vision Transformer(ViT) with a patch size of 16×16 . The first classification head was changed for the three-classes, which included normal, non-COVID and COVID-19 cases. The same official COVID-QU-Ex training, validation and testing splits were applied to guarantee uniformity with the CNN experiments. Light data augmentation was applied during training by means of random resized cropping and random rotation $\pm 15^\circ$ to increase generalization ability. Validation and Testing Images were center cropped to undergo deterministic preprocessing. Normalization of images was done using mean and standard deviation values unique to the dataset. AdamW optimizer with a learning rate of 3×10^{-4} was used to conduct training with batch sizes of 32 for training, validation and testing for a maximum of 20 epochs. Model selection was guided by validation accuracy. The checkpoint with the highest validation performance was kept for further explanation generation and quantitative evaluation. Vision Transformer acts as the transformer-based comparison model

in this thesis, allowing examination of both predictive behavior and explainability features relative to the CNN baseline.

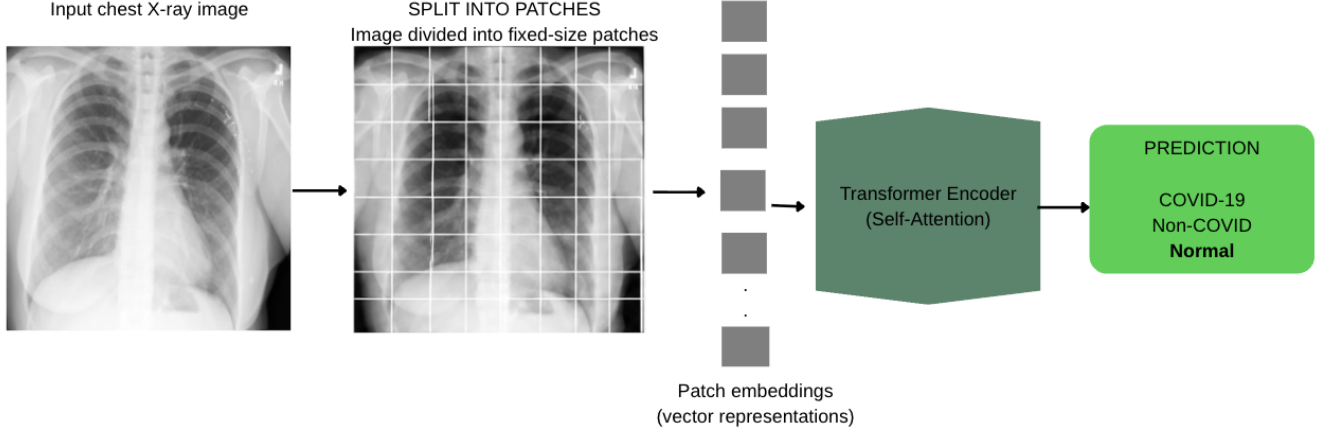


Figure 2: Simplified overview of the Vision Transformer (ViT) architecture used in this study. The input chest X-ray is partitioned into fixed-size image patches, which are transformed into patch embeddings and processed by transformer encoder layers employing self-attention. The resulting representation is then used to predict one of the three diagnostic classes: COVID-19, Non-COVID or Normal.

3.4 Explanation Methods

3.4.1 LIME

The perturbation-based explanation approach applied in this study was Local Interpretable Model-agnostic Explanations (LIME)[RSG16](See Figure 3). LIME is a model-agnostic technique that does not rely on the internal architecture of the model being explained, therefore it is appropriate for direct comparison across the Vision Transformer and ResNet-50 CNN. Rather than looking at internal activations or gradients, LIME explains predictions by looking at how the model’s output changes when parts of the input image are systematically changed. LIME starts by breaking down the input image into understandable areas known as superpixels, which group adjacent pixels into visually coherent region. Then, by deliberately masking combinations of these areas and asking the model for its estimates on each perturbed sample, it produces several perturbed versions of the original image. Based on the model’s responses, LIME fits a simple local surrogate model that estimates the behavior of the original classifier in the vicinity of that particular image. Areas whose deletion causes a notable shift in prediction are regarded as locally relevant for the model’s decision. Because of this local approximation, LIME is especially helpful for interpreting specific forecasts rather than the general decision-making processes of the model. To guarantee equitable comparison, LIME explanations were generated for both the ViT and CNN models using the same hyperparameter configuration and evaluation procedure. For every image, 500 perturbed samples were produced, whereby masked areas were substituted with a baseline zero value. For each image, LIME was configured to explain only the model’s predicted class, ensuring only positively contributing regions were kept. The last explanation contained the two most

powerful superpixel regions, the rest of the original image was kept to allow interpretation during visualization. This setup was selected to generate brief and visually interpretable explanations while preserving uniformity across architectures so that noted variations might be mostly ascribed to model performance rather than explanation settings.

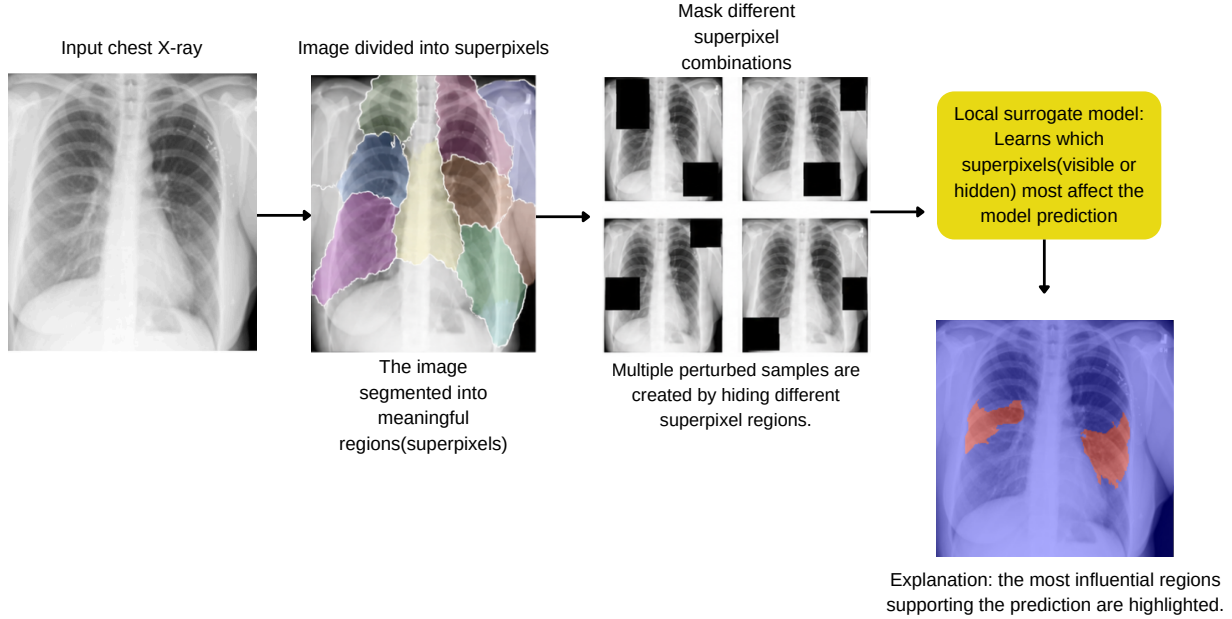


Figure 3: Simplified overview of the LIME explanation process. First, the input chest X-ray is segmented into superpixels representing meaningful image regions. Multiple perturbed samples are then generated by masking different combinations of superpixels and obtaining the corresponding model predictions. These predictions are used to fit a local surrogate model that approximates the behavior of the original classifier in the vicinity of the input image. Finally, the most influential regions are highlighted to indicate which parts of the image contributed most strongly to the predicted class.

3.4.2 Layer-wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation (LRP)(See Figure 4) was used as a gradient-based attribution method in this study to provide architecture-specific explanations for both the CNN and Vision Transformer models. Unlike methods based on perturbations, LRP explains a prediction by following relevance back across the network itself. Each neuron’s, and each input pixel’s contribution to the model’s final prediction is represented by this relevance score. Starting with the model’s output for a particular prediction, relevance scores are passed layer by layer toward the input image, therefore distributing the prediction score among the relevant neurons according to set propagation rules. The final attribution map shows which input areas helped the model’s prediction, therefore directly breaking down the prediction in terms of input relevance. LRP’s main benefit is that

it uses the model’s internal structure instead of depending on outside perturbation, therefore keeping explanations near to the learned decision process. The transmission of relevance through the model’s computational path allows for more accurate and finely-grained attribution, particularly when subtle spatial factors affect classification decisions in fields like medical imaging. LRP explanations were produced for the Vision Transformer by using the transformer attribution method, TransLRP[CGW20], given in the explainability benchmarking repository[KBB23] that served as the basis for this study. This technique uses relevance propagation to transformer designs by distributing relevance across patch embeddings, transformer encoder layers and self-attention mechanisms to help to assign prediction relevance back to the first input image. Unlike CNNs, whose relevance mostly comes from hierarchical convolutional feature maps, transformer-based relevance propagation has to take token interactions and attention-based information flow into account.

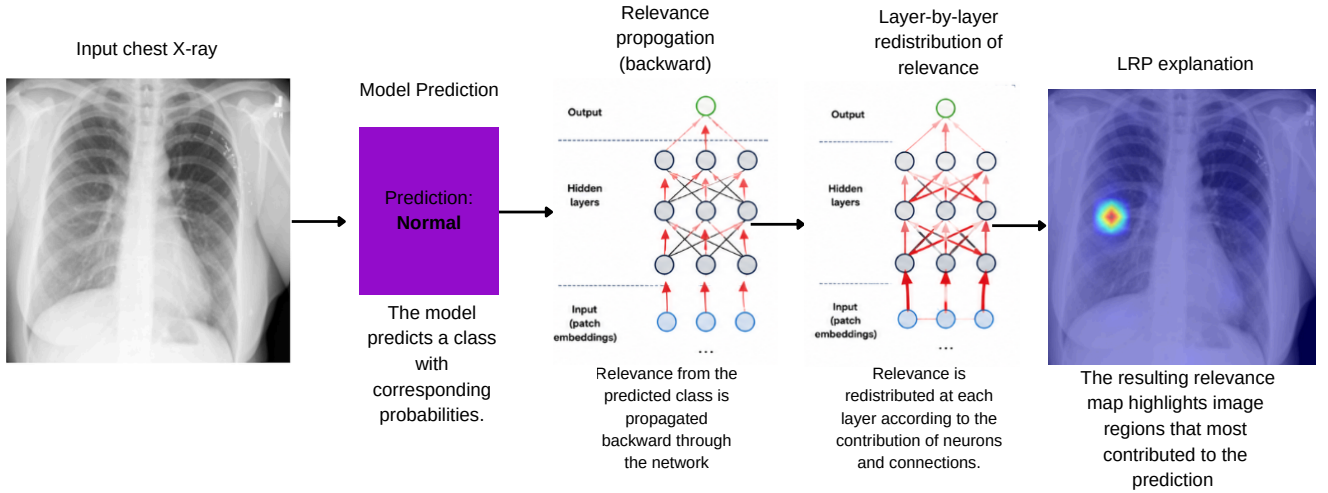


Figure 4: Simplified overview of the Layer-wise Relevance Propagation (LRP) process. Following a model prediction, the prediction score is propagated backwards through the network. At each layer, relevance is redistributed according to the contribution of individual neurons and connections to the final prediction. This process continues until the input layer is reached, producing a relevance map that highlights the image regions that contributed most strongly to the model’s decision.

3.4.3 Attention Rollout

The Vision Transformer’s attention-based explanation method was Attention Rollout(See Figure 5). Attention Rollout drives explanations right from the self-attention mechanism of the transformer architecture. Since Vision Transformers divide an image into a series of patches, the attention weights capture how information is propagated between image regions during prediction. This method

recursively aggregates attention information across transformer layers using Attention Rollout technique. This way, one can follow the contribution of image patches from input to last classification token. The generated attention map offers a visual representation of the parts of the image that get the most accumulated attention during the processing pipeline of the model. Attention aggregation was considered in multiple ways. Using mean, minimum and maximum aggregation over various attention heads, attention maps were produced according to the approach offered in [KBB23]. Quantitative evaluation (Table 2) revealed notable differences among these variants. The maximum aggregation variant was chosen for qualitative analysis and visual comparison with LIME and LRP since it generated the most localized attribution maps and avoided the very spread-out explanations seen for several other aggregation methods.

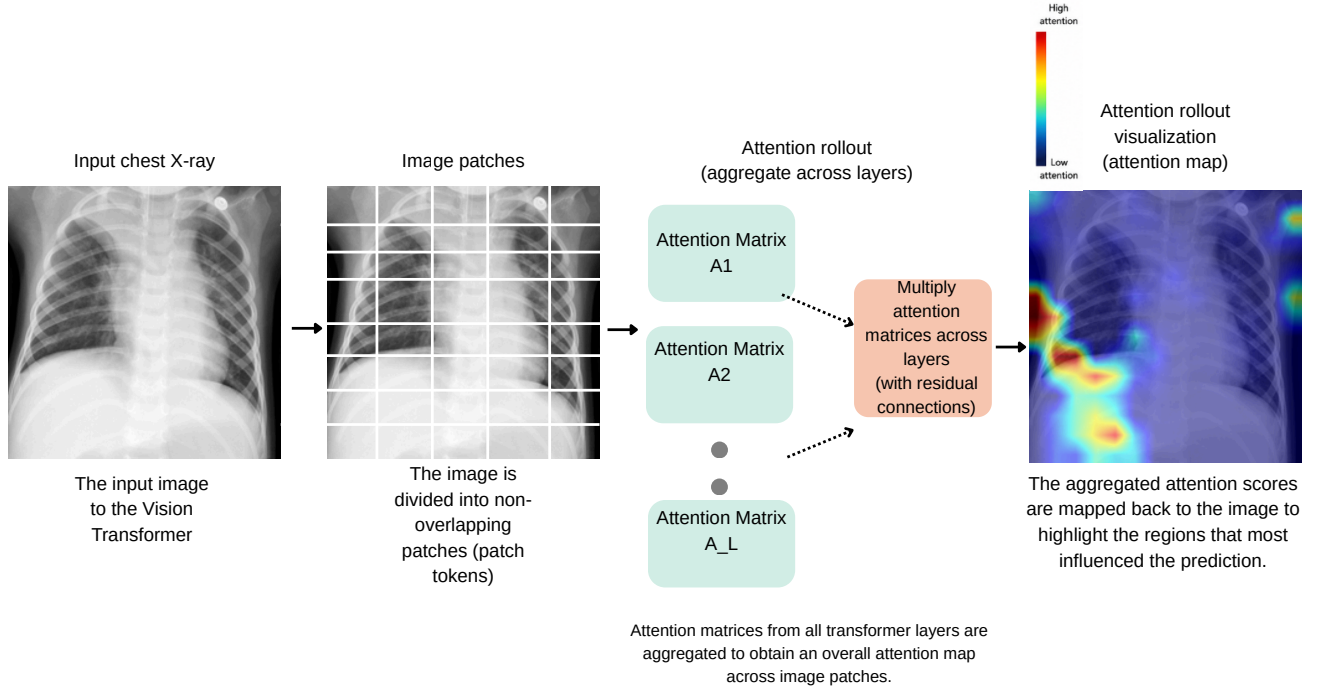


Figure 5: Simplified overview of the Attention Rollout process used to generate explanations from the Vision Transformer. The input chest X-ray is first divided into image patches and processed by multiple transformer layers employing self-attention. Attention matrices from all transformer layers are aggregated to obtain an overall attention map across image patches. The resulting attention scores are then mapped back to the image space, producing a visualization that highlights the regions that most strongly influenced the model’s prediction.

3.5 Quantitative Evaluation Framework

To judge explanation quality beyond visual examination, quantitative analysis was carried out for explanation methods as well as per-class analysis. This study evaluates explanations using three criteria : sensitivity, faithfulness and complexity. All metrics were computed using the Quantus[HWB⁺22] framework.

3.5.1 Faithfulness

Faithfulness measures the extent to which an explanation reflects the features that genuinely influence a model’s prediction. A faithful explanation should give areas of the image that really affect the model’s output high importance and low importance to areas with little contribution to the ultimate result. In this thesis, faithfulness was evaluated using the Faithfulness Correlation [BWM20] metric. The metric examines how attribution significance relates to prediction changes under feature perturbation. For every explanation, important image areas are gradually altered and the effect on model confidence is noted. A positive correlation indicates that areas of the explanation method considered most important have a significant impact on their prediction. On the other hand, small or negative values imply that the highlighted areas are not very linked with the decision-making process of the model. To guarantee a fair comparison, faithfulness was determined for all explanation techniques with same settings across both model types.

3.5.2 Sensitivity

Sensitivity looks at how stable an explanation is under minor input changes. If a little change in input image causes significant changes in explanations generated, an explanation approach is deemed sensitive. Although some fluctuation is normal, very sensitive explanations could be difficult to believe since the highlighted areas could shift easily even as the model prediction stays mostly unchanged. Therefore, sensitivity helps understand how robust and consistent a certain explanation technique is.

In this thesis, sensitivity was measured using the Average Sensitivity [BWM20] metric. Formally, it can be expressed as

$$\text{AvgSensitivity}(x) = \frac{1}{N} \sum_{i=1}^N \frac{\|E(x) - E(x + \delta_i)\|}{\|\delta_i\|}, \quad (1)$$

where $E(x)$ denotes the explanation generated for input image x , δ_i represents a small perturbation applied to the input, and N is the number of perturbations considered. The metric compares explanations derived from somewhat perturbed copies of the same image with explanations produced from the initial image. The average difference between these explanations is then determined across several perturbation samples. Explanations that remain fairly unchanged under perturbation get lower sensitivity ratings and those that change significantly get greater scores. Quantus produced several modified copies of every image for the analysis and noted how much the related explanation maps shifted. All explanation techniques and both models were subjected to the same assessment process. Lower sensitivity values point to more consistent explanations and show that the explanation approach always finds the same locations even if the input changes only slightly. Lower sensitivity ratings are thus usually preferred since they indicate more robustness and consistency of the produced explanations.

3.5.3 Complexity

Complexity assesses how focused or spread out an explanation is by counting the number of attribution values above a given importance threshold. While values below the threshold are regarded as insignificant, attribution values above this threshold are deemed significant for the explanation. Complexity therefore shows how much of the input picture an explanation technique

is highlighting. In this thesis, complexity was evaluated using the Effective Complexity. The metric uses the number of features with relevance values beyond the given threshold($\epsilon = 0.1$) to help to project the total complexity of the explanation. The resulting score was normalized by the total number of image pixels, producing values between 0 and 1. Explanations with little complexity concentrate on a somewhat limited number of picture areas(sparse) and explanations with great complexity spread importance across a bigger part of the image appearing more diffuse. Given the great number of highlighted areas, very complicated explanations could be challenging to understand, so complexity is especially relevant when explanations are given visually. Explanations with extremely low complexity are simpler to examine, but could oversimplify the model’s thinking process by emphasizing only a small subset of relevant characteristics. Furthermore, complexity should not be interpreted alone but together with sensitivity and faithfulness. A good explanation should perfectly balance the interpretability, stability and alignment with the decision process of the model.

3.6 Qualitative Evaluation

The quality of explanations can be objectively measured using quantitative metrics. However, they cannot tell if the highlighted areas represent important anatomical structures or if the explanations center on visually likely regions of the chest X-ray. The spatial features of the created explanation maps were therefore investigated to balance the quantitative analysis with a qualitative assessment. The qualitative study examined visual inspection of explanation maps produced by LIME, LRP and Attention Rollout. Through a systematic inspection procedure, representative chest X-ray images were chosen from the evaluation subset to include both correct and incorrect predictions as well as instances from each class. Four factors were analyzed. Explanations for accurately classified and misclassified samples were first compared to determine whether explanation behavior differed between successful and failed predictions. Secondly, comparison of whether LIME, LRP and Attention Rollout highlighted same picture areas for the same prediction as well as agreement between explanation techniques was evaluated. Thirdly, misclassified samples were investigated in order to see whether explanation maps might offer insight into possible model failures, like inconsistent attribution patterns or peripheral region attention. At last, qualitative observations were matched with the quantitative evaluation metrics to see if visual features of the explanations matched observed outcomes in sensitivity, complexity and faithfulness. The qualitative assessment was not meant to establish diagnostic accuracy or clinical validity. Rather, its goal was to offer a visual evaluation of explanation behavior and to help to understand the quantitative results better by looking at how the explanation techniques represented the decision-making processes of the CNN and Vision Transformer models.

3.7 Artifact Robustness of Models

An artifact study for models was done to see whether the model predictions depended on peripheral image information. Although interpretation techniques show which areas might affect a prediction, they do not directly test if a model uses maybe unimportant image characteristics. So, controlled image perturbations were applied in order to asses the robustness of the CNN and Vision Transformer predictions. Two alternative approaches were taken into account. First, setting pixel values to zero helped to cover a 20-pixel border all around the image. Using 40x40 pixel square regions,

the four picture corners were masked in the second setup. These perturbations were chosen since peripheral areas could unintentionally affect model projections by containing acquisition-related artifacts, boundaries, markers or other non-clinical data. For every picture in the evaluation subset, predictions were made for the altered forms as well as the initial one. Changes in predicted class labels and changes in confidence for the originally predicted class were recorded. In addition, prediction change rates were analyzed for both in general and per-class. A model that can be consistent under these perturbations and is less reliant on peripheral picture data is hence more robust to possible shortcut features. The artifact ablation study was utilized as a complementary study to the quantitative and qualitative assessment. The experiments directly test the effect of removing image regions on the model’s predictions. These evaluations together help to better understand model interpretability and robustness.

3.8 Artifact Robustness of Explanation Methods

An artifact robustness analysis was conducted to evaluate the stability of explanation methods under controlled image perturbations. This analysis aims to understand how explanation methods behave when potentially irrelevant image regions are removed. Same as the study for models, two strategies were applied to the chest x-ray images. First, a 20-pixel border surrounding the image was masked by setting the corresponding pixel values to zero. Then, the four image corners were masked using 40×40 pixel square regions. Explanations were created for every image in the evaluation subset including the original, border masked and corner masked variations. All of the examined explanation techniques were subjected to this process. The generated explanation maps were then contrasted to help understand how significantly the explanations were altered with the artifact removal. Two different metrics were used to assess explanatory robustness. First, Pearson correlation was applied to evaluate the general resemblance of the first and altered explanation maps. Given an original attribution map (A) and an altered attribution map (B), Pearson correlation is defined as

$$\rho = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^n (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^n (B_i - \bar{B})^2}} \quad (2)$$

where A_i and B_i denote the attribution values at pixel i , and $\bar{A}$ and $\bar{B}$ represent the mean attribution values of the respective maps. Higher correlation number shows that perturbation left the explanation mostly unaltered. Second, the top 10% most relevant attribution areas’ Intersection over Union (IoU) was determined to help to assess whether the explanations still highlighted the same significant image areas following artifact removal. The Intersection over Union (IoU) was computed as:

$$IoU = \frac{|R_A \cap R_B|}{|R_A \cup R_B|} \quad (3)$$

where R_A and R_B denote the sets of pixels belonging to the top 10% highest attribution values in the original and altered explanations respectively. Apart from assessing explanation similarity, the altered images’ faithfulness and complexity were recomputed in this experiment. This helped to determine whether explanation quality was maintained.

3.9 Experimental Setup

All experiments were implemented in Python using the PyTorch deep learning framework. The CNN experiments were built upon a ResNet-50 implementation, while the Vision Transformer experiments were based on the open-source framework developed by Komorowski et al.[KBB23], which provides implementations for training, generating explanations and evaluating explanation for ViT. The original repository was adapted to support the experiments performed in this thesis. The experiments were executed on a GPU-enabled computing environment to efficiently generate explanations for the evaluation dataset. Model training, explanation generation and metric computation were performed using identical preprocessing pipelines and evaluation settings to ensure a fair comparison between architectures and explanation methods. All experiments were conducted on the same subset of 300 test images with equal numbers of classes. For reproducibility the implementation used in this study is available at: <https://github.com/3876853/bachelor-thesis-talya-seryano>

4 Results and Analysis

On the chest X-ray dataset, both ResNet-50 CNN and Vision Transformer showed good classification accuracy, therefore suggesting that both models could learn significant discriminative patterns for COVID-19, Non-COVID and Normal image classification. The CNN had a best validation accuracy of 94.65% and a matching test accuracy of 96.07% . The Vision Transformer had a best validation accuracy of 93.78% . The common subset of 300 samples with equal number of COVID-19, Non-COVID and Normal images was used for explanation analysis to guarantee direct comparison across models and explanation approaches. Both models kept enough prediction power to enable meaningful explanation analysis.

4.1 Evaluation of Explanation Methods

Table 2 presents the quantitative evaluation results for the explanation methods across both model architectures. Several consistent patterns can be observed across the evaluated metrics. LIME generated the lowest complexity numbers for both the Vision Transformer and CNN models. This shows that LIME produced sparsest explanations across both models. As also visible in Figure 6, LIME explanations typically focused on a limited number of localized image regions, resulting in visually compact attribution maps. Though often visually basic and simple to understand, such justifications also showed the highest sensitivity numbers among all the tested techniques. LIME explanations differed greatly under tiny perturbations for both models, therefore indicating less explanation stability. This behavior is consistent with the perturbation-based nature of LIME, where the explanation depends on locally sampled image perturbations and the surrogate model fitted to those samples. Furthermore, faithfulness ratings for LIME stayed rather low, especially for the Vision Transformer. For the Vision Transformer, the difference was particularly noticeable since ViT LIME showed much lower faithfulness than CNN LIME, therefore implying that perturbation-based local surrogate explanations could be less fit with transformer-based decision representations. This means that although LIME explanations were understandable and less complex, they weren't always closely in line with the characteristics most affecting the model predictions and were less stable. In contrast, LRP-based explanations showed reduced sensitivity along with increased faithfulness across both models, therefore implying that relevance propagation more closely caught the areas affecting

model predictions. Though LRP explanations usually had more complexity than LIME, the greater density of the explanations seemed to point to more accurate and trustworthy attribution patterns instead of graphically sparse but unstable highlights. Relatively low sensitivity and more faithfulness together point to LRP’s explanations being more consistent and more indicative of the underlying decision-making processes of the models. The Attention Rollout explanation method for the Vision Transformer showed inconsistent behavior. Though the sensitivity values for the variants were low, which suggests excellent stability under perturbation, their faithfulness ratings stayed consistently low. Rather than focused on diagnostically significant areas, attention explanations’ results seemed to point out to highlights that are spread out and somewhat evenly distributed throughout the image. The Attention Mean approach especially generated extremely high complexity scores that pointed to highly diffuse attribution patterns across big image areas. While attention visualizations provide an intuitive view of how information flows through the transformer, these findings suggest that attention alone should not be interpreted as a direct explanation of the model’s decision-making process. Consequently, the quantitative results show that LRP gave the most balanced explanation quality across both architectures by low sensitivity together with reasonably high faithfulness. By contrast, LIME created simpler but less reliable explanations, whereas Attention Rollout exhibited limited faithfulness even with their stability. These results already point to important differences in how the Vision Transformer and CNN examine chest X-ray data.

Table 2: Quantitative comparison of explanation methods across the CNN (ResNet-50) and Vision Transformer (ViT) on the shared 300-sample evaluation subset. Lower complexity and sensitivity indicate simpler and more stable explanations, while higher faithfulness indicates stronger alignment with model predictions.

Model	Explanation Method	Complexity ↓	Sensitivity ↓	Faithfulness ↑
CNN	LIME	0.083	0.216	0.092
CNN	LRP	0.239	0.078	0.148
ViT	LIME	0.095	0.218	0.053
ViT	LRP	0.224	0.061	0.145
ViT	Attention (Mean)	0.936	0.055	0.005
ViT	Attention (Min)	0.622	0.084	0.020
ViT	Attention (Max)	0.174	0.069	0.012

4.2 Class-wise Analysis of CNN Explanations

The per-class evaluation results for the CNN explanations, shown in Table 3, reveal several consistent differences between LIME and LRP among Non-COVID, Normal and COVID-19 classes. Although the general patterns found in the combined analysis are still evident at the class level, the class-wise analysis offers more understanding of how explanation behavior varies with the underlying image features. For LIME, explanation complexity was consistently low throughout all classes, hence suggesting localized explanations with attribution concentrated in relatively few regions of the image. The COVID-19 class had the highest sensitivity values. The numbers progressively dropped for the Normal class. This could indicate that the image regions highlighted by LIME for COVID-19 images were less robust to small input perturbations, causing greater changes in the resulting explanations.

Visually, LIME explanations typically centered on small areas that changed significantly between samples (See Figure 6). Explanations from LRP revealed more different cross-class behavior. LRP created denser attribution maps with less sensitivity and higher faithfulness across all classes than LIME. COVID-19 samples showed the highest complexity scores. CNN relevance propagation tended to spread attribution across more of the chest area for COVID-related forecasts. These explanations often covered more general lung structures instead of separate hotspots. This suggests the CNN’s dependency on more spatially distributed data to detect COVID-19 patterns. The class-wise faithfulness findings also show that the predicted category determined the explanation reliability variation. In particular, Normal samples usually showed greater faithfulness values than COVID-19 samples for both LIME and LRP explanations. This might mean that the explanations for Normal images were more closely related with the elements really affecting the CNN predictions. Normal chest X-rays may have fewer overlapping pathological patterns than those of COVID-19 and Non-COVID patients. By contrast, lower faithfulness ratings for diseased groups could indicate that the model’s reasoning process is more challenging to localize regularly when pathological results vary or are spread out. The class-wise study shows generally that the fundamental prediction category as well as the explanation technique itself affected explanation behavior. While more dense attribution patterns often seemed more robust and more closely matched the prediction behavior of the CNN, sparse explanations did not always correspond to consistent or faithful explanations.

Table 3: Per-class quantitative evaluation of CNN explanation methods.

Method	Class	Complexity ↓	Sensitivity ↓	Faithfulness ↑
LIME	COVID-19	0.085	0.242	0.076
LIME	Non-COVID	0.086	0.212	0.076
LIME	Normal	0.077	0.194	0.123
LRP	COVID-19	0.363	0.068	0.130
LRP	Non-COVID	0.223	0.076	0.132
LRP	Normal	0.131	0.089	0.182

4.3 Qualitative Analysis

Further insight into how the various explanation techniques work for both correct and unsuccessful predictions is given in Figures 6 and 7. The qualitative study shows how these explanations diverge in application and whether they correspond to clinically significant image areas even if the quantitative metrics point to general explanation quality. For correctly classified samples, LRP explanations generally appeared more structured and anatomically focused than both LIME and Attention Rollout explanations. LRP regularly emphasized areas inside the lung fields. It usually generated attribution maps spread over pertinent thoracic structures rather than just small image spots in the CNN and for Vision Transformer. It mostly seemed to identify localized regions of relevance that corresponds to anatomically meaningful areas. LIME explanations were much less frequent and usually focused on a limited number of local areas. This sparsity helps with the visual interpretation but the highlighted areas sometimes seemed fragmented or confined to just a small part of the image. Attention Rollout explanations showed less localization than LRP. Some focus maps showed areas within the lung fields, however, the highlighted areas were

usually broader and less structured than the related LRP explanations. These findings support the quantitative ones, which revealed that LRP had the best balance between faithfulness and sensitivity and that Attention Rollout had poor faithfulness despite low sensitivity. Misclassified examples reveal the differences even more clearly. In the scenario when the CNN misclassified a COVID-19 image as Normal while the Vision Transformer projected the correct class, the CNN explanations seemed weak and poorly localized, especially for LRP. By contrast, the equivalent ViT interpretations centered more visibly on areas inside the lung fields. This implies that the transformer model caught diagnostically important image characteristics for this sample that the CNN couldn't really highlight. The case shows how different explanations can reveal model-specific failure modes outside the prediction itself. In the sample where the Vision Transformer made a mistake prediction but the CNN correctly classified the image, a different pattern shows itself. Even with the wrong prediction, the ViT explanations here remained somewhat focused. This observation draws attention to a major restriction of qualitative interpretation: an explanation that appears visually plausible is not necessarily a faithful representation of the model's decision-making process. An explanation might accurately highlight the lung region while continuing to rely on picture characteristics that are not clinically significant or contribute little to the prediction. Consequently, this is why explanation quality should not be judged only by visual appearance but should be considered together with quantitative evaluation metrics. The last misclassification case, in which both models gave wrong predictions, helps to understand the situation even more. Even though the model's classifications were both wrong, several explanation techniques nonetheless emphasized areas inside the lungs instead of clear picture noise or background features. This suggests that model errors do not necessarily arise from completely irrelevant attention patterns. Rather, the models might be reacting to picture features that resemble disease-related characteristics but finally inadequate for differentiating between diagnostic groups. Such occurrences show the fundamental difficulty of interpreting model failures in medical imaging, where tiny and overlapping visual contrasts exist between classes. Direct comparison of the explanation techniques shows noticeable behavioral changes. Usually pointing out just a few image areas, LIME explanations were usually the most coarse and localized. LRP explanations were more spatially coherent and showed greater agreement with anatomical structures visible in the chest X-rays. Attention Rollout explanations showed attribution patterns that often differed from those produced by LRP and LIME. The highlighted regions were not consistently aligned with the attribution patterns identified by the other explanation methods. These qualitative results generally support the quantitative results given before. LRP's higher faithfulness was reflected in more meaningful attribution patterns, higher sensitivity of LIME corresponded to more variable and fragmented explanations. Low faithfulness of Attention Rollout explanations corresponded with their highlighted areas not being necessarily the regions most responsible for the final prediction. Therefore, the integration of qualitative assessment and quantitative measurements offers a more thorough knowledge of how the CNN and Vision Transformer reach their predictions and why they sometimes fail.

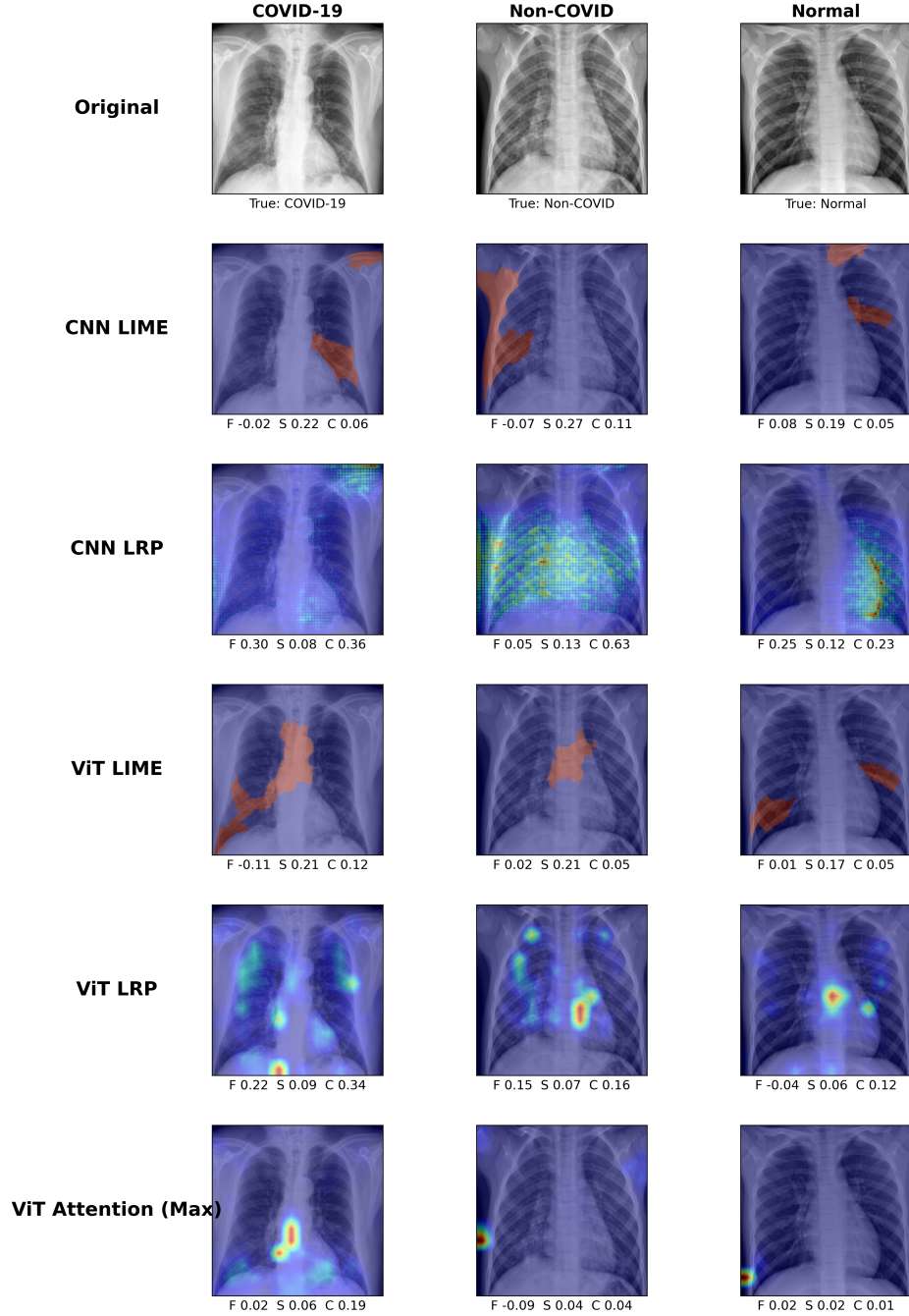


Figure 6: Qualitative comparison of explanation methods for correctly classified chest X-ray samples from the COVID-19, Non-COVID and Normal classes. Explanations are shown for the CNN (ResNet-50) and Vision Transformer models using LIME, LRP and Attention Rollout. For each explanation, F denotes faithfulness, S sensitivity and C complexity for the individual sample. Negative faithfulness values indicate inverse correlation between attributed importance and prediction confidence changes after perturbation.

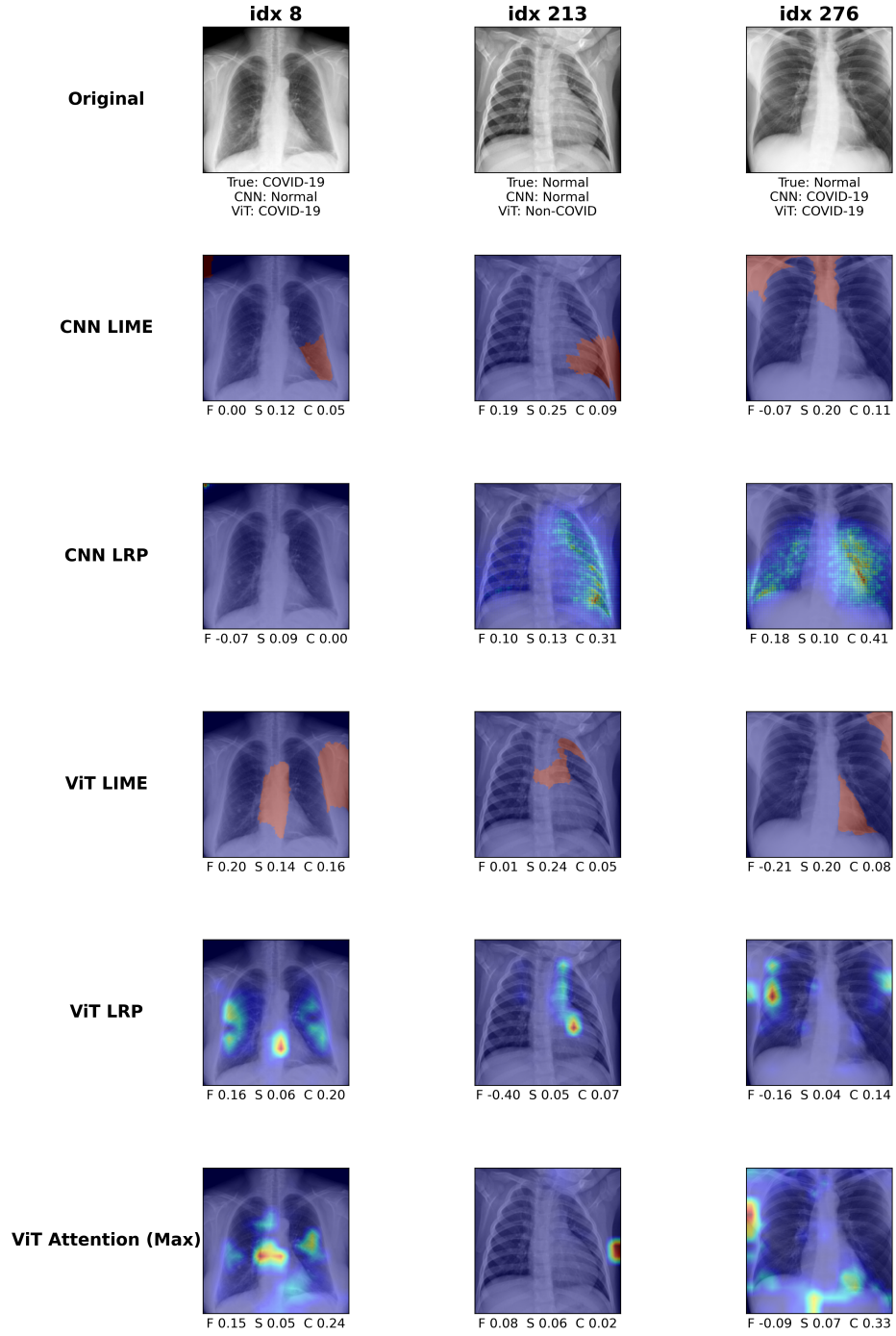


Figure 7: Qualitative analysis of misclassified chest X-ray samples. Three representative failure cases are shown: (left) a CNN misclassification where the Vision Transformer predicted correctly, (middle) a Vision Transformer misclassification where the CNN predicted correctly, and (right) a case where both models failed. F denotes faithfulness, S sensitivity and C complexity for the individual explanation instance.

4.4 Artifact Robustness Study for Models

Beyond the explanation data, the artifact study offers another robustness view. The earlier investigations assessed the quality of the produced explanation maps. This study examined how much each model’s predictions relied on peripheral image areas that might include non-clinical artifacts including borders, markers or acquisition-related signals. As seen in Table 4, the Vision Transformer suffered very much less from both border and corner perturbations than the CNN. This is because CNN’s higher prediction change rates and bigger reductions in prediction confidence following image masking. The same perturbation had far less impact on the ViT, masking the image borders resulted in a significantly lower percentage of ViT predictions to change. A comparable pattern was seen for corner masking whereby the ViT predictions stayed mostly constant but the CNN once more showed a strong sensitivity to the areas removed. This shows how the ViT seems more resilient and robust to changes outside the core thoracic area while the CNN depended more heavily on peripheral picture information. The per-class findings in Table 5 indicate that this sensitivity varied across classes. Normal and Non-COVID samples for the CNN were far more affected by border and corner masking than COVID-19 samples. This means that the CNN’s choices for non-COVID and healthy patients could have relied more heavily on image-level signals outside the major lung area. On the other hand, the ViT showed continuously low prediction change rates across all classes which suggests a more stable response to the same perturbations. This difference is notable in the setting of medical imaging, where a model can depend on peripheral or acquisition-specific artifacts and could do well on a dataset but struggle to generalize consistently. This behavior is shown in Fig. 8 for a typical Normal sample. For the CNN, covering outlying areas caused the forecast from Normal to COVID-19 to be altered. However, for the ViT, the prediction stayed the same under both perturbations. By demonstrating that the CNN’s classification threshold was more sensitive to data left out from the border and corner areas, this case supports the quantitative results. This study offers proof in this evaluation subset that the CNN was more reliant on peripheral image areas than the ViT. The artifact robustness study results generally support the larger interpretation of the explanation analysis. Even though its predictions were more sensitive to peripheral perturbations, the CNN created helpful and accurate LRP explanations. Despite the fact that there are restrictions on explanations based on Attention Rollout, the ViT proved to be more robust in this study. This implies that interpretability consists of two different but connected components: explanation quality and robustness. While sensitive to non-clinical image areas, a model can generate useful justifications.

Table 4: Artifact robustness comparison between the CNN and Vision Transformer on the shared 300-sample evaluation subset. Lower prediction change rates and confidence drops indicate greater robustness to potential image artifacts.

Model	Perturbation	Prediction Change Rate ↓	Mean Confidence Drop ↓
CNN	Border Ablation (20 px)	0.333	0.281
CNN	Corner Ablation (40×40)	0.223	0.190
ViT	Border Ablation (20 px)	0.073	0.060
ViT	Corner Ablation (40×40)	0.027	0.016

Table 5: Per-class artifact robustness comparison between the CNN and Vision Transformer on the shared 300-sample evaluation subset. Values indicate the proportion of predictions that changed after applying artificial image perturbations. Lower values indicate greater robustness.

Model	Perturbation	COVID-19 ↓	Non-COVID ↓	Normal ↓
CNN	Border Ablation (20 px)	0.07	0.38	0.55
CNN	Corner Ablation (40×40)	0.05	0.23	0.39
ViT	Border Ablation (20 px)	0.04	0.06	0.12
ViT	Corner Ablation (40×40)	0.03	0.01	0.04

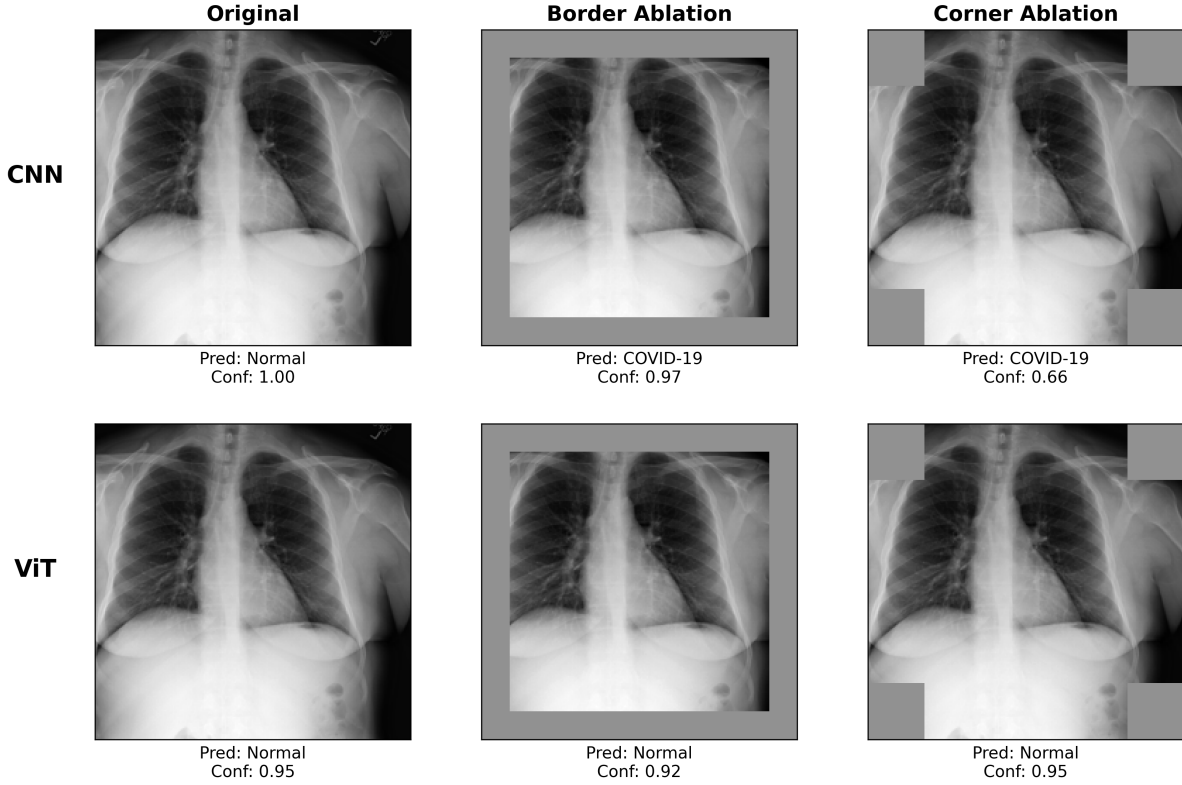


Figure 8: Qualitative artifact ablation analysis for a representative Normal chest X-ray sample . The original image, border masked and corner masked variants are shown for both the CNN and Vision Transformer models. The CNN prediction changes from Normal to COVID-19 after both perturbations, accompanied by substantial confidence changes, whereas the Vision Transformer prediction remains stable under the same perturbations. This example visually illustrates the greater robustness of the Vision Transformer to peripheral image perturbations and supports the quantitative artifact robustness results reported in Table 4.

4.5 Artifact Robustness Study for Explanation Methods

Table 6 presents the results of the artifact robustness experiment while Figures 9 and 10 provide qualitative examples illustrating how the explanations changed with border and corner masking.

The results show clear differences between the evaluated explanation methods in terms of robustness when peripheral image regions are removed. Among the Vision Transformer explanation approaches, ViT-LRP showed the highest robustness. It kept reasonably consistent complexity and faithfulness numbers while also reaching the highest explanatory similarity ratings across both border and corner masking. Figure 9 also shows this behavior as the main areas emphasized by ViT-LRP remain essentially unaltered across all settings. Though little changes in attribution intensity exist, the overall explanatory structure stays the same. These results indicate that ViT-LRP mostly uses image areas pertinent to the model’s prediction instead of information found in peripheral image artifacts. Particularly when the overlap of important areas taken into account, ViT-LIME also showed a somewhat high level of robustness. The IoU ratings show that the technique mostly still found similar image regions post artifact removal. Still, the reduced correlation values compared to ViT-LRP imply a more major shift in the attribution ratings given to these areas. Qualitatively, while the overall placement of key areas stays constant, the shape and distribution of the explanations usually differ across the settings. Thus, even if ViT-LIME seems to be able to highlight fixed areas of interest, the explanations themselves are less consistent than those generated by ViT-LRP. In contrast, Attention Rollout was much more sensitive to artifact removal. Compared to the other Vision Transformer explanation techniques, the quantitative data reveal significantly reduced similarity ratings and more complexity changes. Figure 9 helps to explain this behavior where highlighted areas change after both border and corner removal. In some instances, focus is shifted towards totally other parts of the picture, therefore suggesting that variations in the image greatly impact the generated explanations. The findings imply that attention-based images could not always match reliable or consistent interpretations of model behavior. Though the differences among approaches were less notable, the CNN-based explanation techniques showed a rather consistent general pattern. As seen in Table 6, CNN-LRP usually produced better explanation similarity than CNN-LIME. Figure 10 also shows this trend. While CNN-LRP explanations continue to highlight comparable anatomical areas following artifact removal, CNN-LIME shows changes in the highlighted regions. However, while CNN-LRP kept high faithfulness, it’s complexity increased, indicating a more diffuse explanation after the artifact removal. Nevertheless, both CNN explanation methods maintained a clear overlap between the original and altered explanations. This shows that the CNN model is less sensitive to the removal of peripheral image content than the Vision Transformer explanations based on Attention Rollout. When all findings taken into account, it shows that among both designs LRP offers the most strong explanations. Though LIME usually keeps the locations of key image areas, the degree and distribution of the related attributions are more prone to variation. The least robust approach seems to be Attention Rollout as it shows significant changes in explanation structure when non-diagnostic image areas are eliminated. These results show more proof that explanation techniques vary in their capacity to produce consistent explanations under little input picture perturbations as well as in their faithfulness and complexity.

Table 6: Artifact robustness evaluation of explanation methods. Pearson correlation and IoU compare explanations before and after artifact removal. Complexity and faithfulness are reported for the original explanations and after border and corner masking. Higher correlation, IoU and faithfulness indicate greater robustness, while lower complexity indicates sparser explanations.

Method	Correlation $\uparrow$		IoU $\uparrow$		Complexity $\downarrow$			Faithfulness $\uparrow$		
	Border	Corner	Border	Corner	Orig.	Border	Corner	Orig.	Border	Corner
CNN LIME	0.178	0.374	0.680	0.711	0.083	0.080	0.087	0.074	0.038	0.050
CNN LRP	0.473	0.582	0.403	0.536	0.239	0.403	0.400	0.165	0.158	0.149
ViT LIME	0.467	0.653	0.709	0.818	0.095	0.085	0.091	0.057	0.070	0.070
ViT LRP	0.729	0.845	0.489	0.616	0.224	0.185	0.219	0.147	0.125	0.139
ViT Attention Rollout	0.192	0.571	0.166	0.396	0.174	0.317	0.205	0.011	0.075	0.022

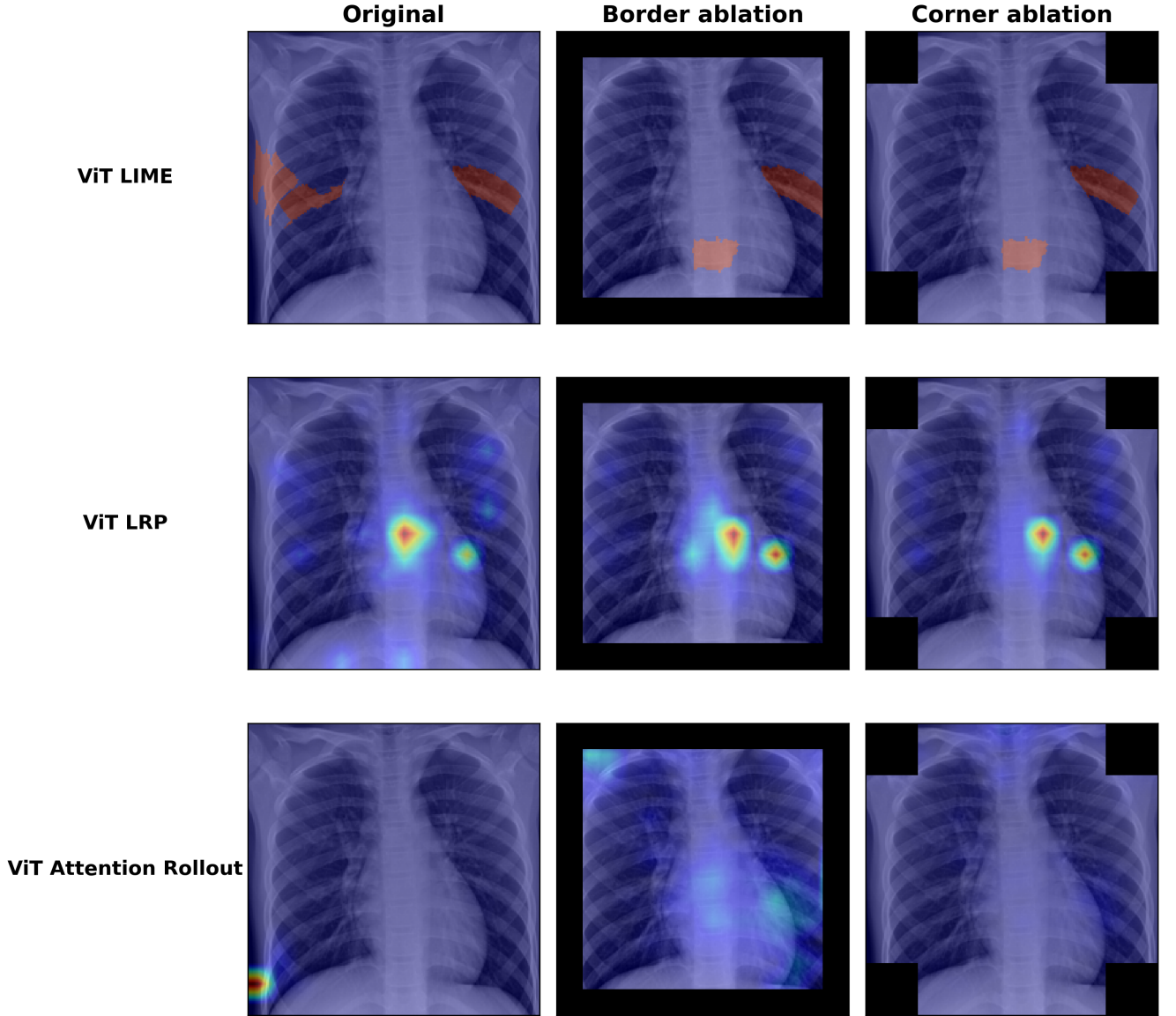


Figure 9: Representative examples illustrating the effect of border and corner artifact removal on explanation maps generated by ViT-LIME, ViT-LRP and Attention Rollout. ViT-LRP produces relatively stable explanations across all conditions, whereas Attention Rollout exhibits substantial changes in highlighted regions following artifact removal. These qualitative observations are consistent with the quantitative robustness metrics reported in Table 6.

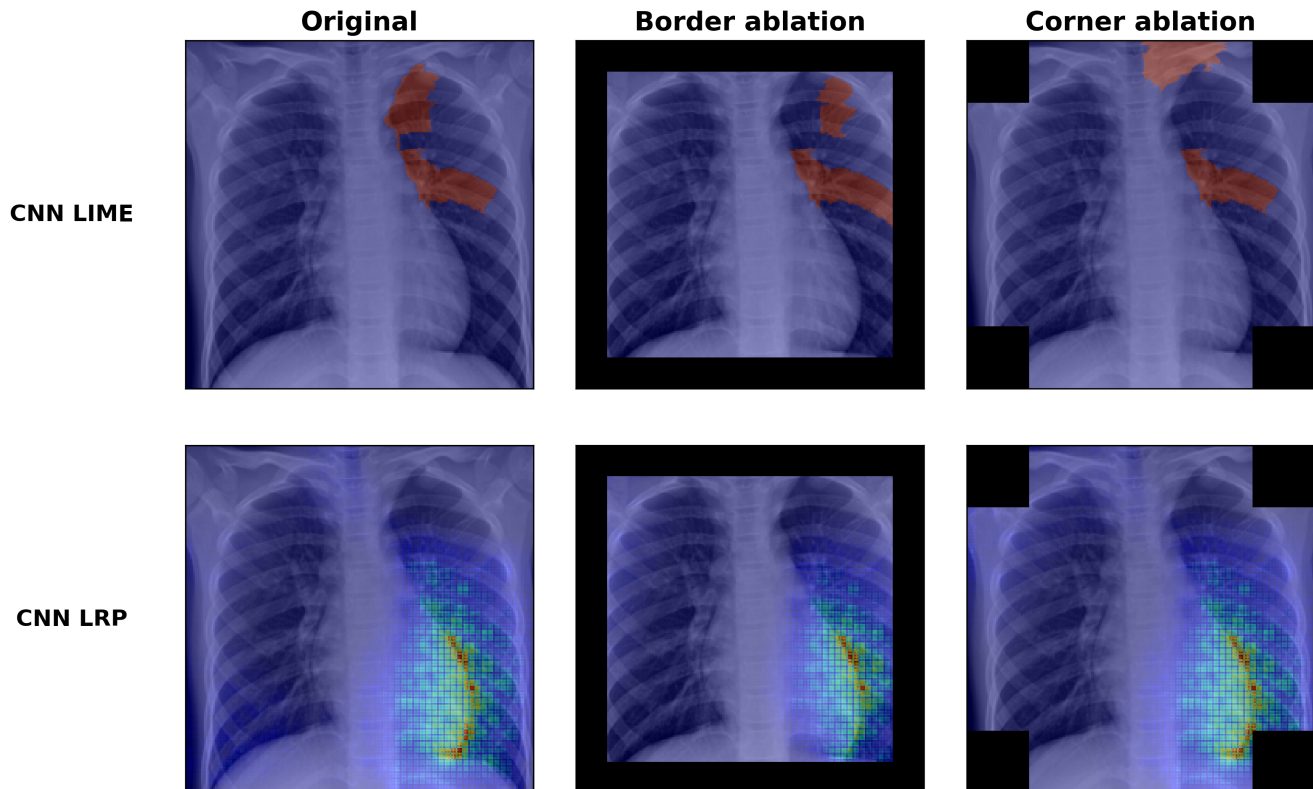


Figure 10: Representative examples illustrating the effect of border and corner artifact masking on explanation maps generated by CNN-LIME and CNN-LRP. While both methods continue to highlight similar anatomical regions after artifact removal, CNN-LRP produces more stable explanations with fewer changes in attribution patterns. In contrast, CNN-LIME exhibits larger shifts in highlighted regions, despite generally preserving the same areas of interest. These qualitative observations are consistent with the quantitative robustness metrics reported in Table 6.

5 Discussion

The aim of this study was to investigate the extent to which explanation methods provide reliable and robust insights into the decision-making processes of convolutional neural networks and Vision Transformer models for COVID-19 chest X-ray classification. The findings show generally that the technique employed greatly affects explanation quality and cannot be properly represented by one assessment measure. Some techniques provided explanations that were fairly faithful, stable and visually appealing but others showed flaws when evaluated from several angles. These results underscore the need of assessing explanation techniques with a combination of quantitative and qualitative techniques instead of depending on a single performance measure. LRP regularly showed the best general performance across both models. It reached among the best faithfulness ratings quantitatively and kept somewhat low sensitivity. The explanations qualitatively often stressed parts of the lung fields and stayed focused on anatomically pertinent structures. Additionally, the artifact robustness study demonstrated that the removal of peripheral picture content had relatively little effect on LRP explanations. This behavior may be explained by LRP redistributing prediction

importance via the network itself, which connects the explanation to the internal calculations of the model. LIME showed a mix of strengths and weaknesses. The technique often yielded explanations with somewhat low complexity, suggesting that only a small portion of image areas was emphasized as significant. From an interpretability standpoint, this might be helpful since localized explanations are usually simpler for people to understand. Still, the sensitivity and robustness scores point out that the interpretations were less steady than those generated by LRP. The artifact robustness study revealed that LIME regularly found comparable areas after artifact removal, but the attribution values allotted to those regions sometimes varied significantly. The perturbation-based nature of LIME can help to clarify this behavior. Because the technique uses a surrogate model to estimate the model locally, comparatively minor changes to the input image might affect the local approximation. Therefore this can affect the produced explanation. Although LIME seems to be able to pinpoint relevant areas of interest, the precise significance assigned to those areas should be taken into account with caution. The results for Attention Rollout were quite intriguing. People often like attention-based explanations because they come right from the transformer architecture and are somewhat simple to see. The findings, nevertheless, indicate that this seeming simplicity could be deceiving. In the artifact robustness study, Attention Rollout very much underperformed and obtained very lower faithfulness ratings than the other techniques that are considered. The qualitative illustrations also showed that the highlighted areas might vary greatly with rather little changes to the image. One theory is that attention weights may not always reflect feature importance. Attention mechanisms were not originally intended to act as explanatory tools, even if they clarify how information is spread across image patches. Consequently, the derived focus maps could represent the model’s information flow without always indicating which areas of the image led to the ultimate result. These results agree with earlier research[JW19] asking whether interpretation of attention is explanatory. One interesting finding is that model robustness and explanation quality did not always coincide. The artifact robustness study for models revealed that the Vision Transformer was more robust to the erasure of peripheral image information than the CNN, therefore indicating its predictions were less reliant on non-diagnostic areas of the image. The reason for that might be that the CNN picked up correlations between peripheral image features and the target classes which impacted its predictions. On the other hand, the global self-attention mechanism of the Vision Transformer could combine information from the whole image and lower dependence on any one peripheral area. However, Vision Transformer’s robustness did not always result in improved explanations across every explanation technique. Even though derived directly from the transformer architecture, Attention Rollout performed much worse while ViT-LRP produced robust and faithful explanations. Despite the fact that CNN was more sensitive to border and corner perturbations, CNN-LRP produced explanations that were rather faithful, consistent and visually concentrated on anatomically plausible areas. These results indicate that model robustness and explanation quality are connected but separate features. A model may be robust yet provide inaccurate explanations. Furthermore, a model that shows some sensitivity to peripheral image material may still produce helpful explanations if matched with a suitable explanation technique. One key discussion of this study is that various evaluation criteria sometimes capture different perspectives of explanation quality. Methods with little complexity, for instance, did not always provide great faithfulness. Also, methods that kept the placement of important areas did not always preserve the total attribution structure. Likewise, visually reasonable explanations weren’t always backed by the numbers from analysis. These findings support the theory that explanation quality is a multi-dimensional idea with several angles. Judging explanations based

only on qualitative analysis or just one quantitative measure may give a partial picture of their reliability. Generally, these results show that LRP offers the most compelling compromise among robustness, stability and faithfulness across the two models. LIME gives clear and somewhat sparse explanations, but seems more susceptible to changes in the input picture. Though directly based on the transformer design, Attention Rollout was the least reliable explanation technique in this study. Therefore, these results show that explanation techniques shouldn't be regarded as interchangeable tools. Instead, thorough assessment is needed before utilizing them to make generalizations about model behavior in medical imaging uses.

6 Limitations

This work offers a thorough comparison of explanation methods for CNN and Vision Transformer models in COVID-19 chest X-ray classification but a number of limitations should be taken into account when interpreting the results. Initially, a single CNN and Vision Transformer architecture, as well as a single dataset, were used in the experiments. The observed results might not apply to different architectures, datasets or medical imaging applications, despite the fact that these models are representative and extensively used. Consequently, the findings should be viewed as evidence within the examined environment rather than universal implications regarding explainability methodologies. There are intrinsic limitations to the examined explanation approaches as well. LIME approximates model behavior using a local surrogate model derived from perturbed samples. Although this method offers clear and understandable explanations, the attribution maps that are produced may be influenced by the segmentation settings, perturbation strategy and random sampling. As a result, the same image may have multiple explanations depending on the runs or parameter choices. Furthermore, the explanation might not adequately convey the model's global decision-making process since the surrogate model only approximates the original classifier locally. Despite having the best overall performance in our investigation, LRP has certain drawbacks. The generated explanations depend on the relevance propagation rules used to redistribute relevance through the network. The explanations are somewhat implementation-dependent because different propagation mechanisms may result in different attribution maps. Furthermore, rather than focusing on whether the model depends on clinically significant features, LRP illustrates how relevance moves across the model. For this reason, a faithful and robust relevance map may still highlight features that are unrelated to the underlying pathology if such features influence the model's prediction. There are additional challenges with Attention Rollout. The approach makes the assumption that feature importance may be inferred from attention weights. However, rather than offering explanations, attention mechanisms were initially intended to aid in the processing of information. Consequently, attention maps might not correctly depict the variables that influence a model's prediction. Based on the lower faithfulness and robustness discovered in this work, attention-based visualizations should be interpreted critically and may not always provide reliable insight into model reasoning. Another limitation concerns the evaluation of explainability itself. Explanation quality lacks a definitive ground truth, in contrast to classification performance. As a result, rather than assessing an explanation's actual truth, the metrics used assess desirable attributes like faithfulness, stability and robustness. Furthermore, an explanation may receive high quantitative scores without reflecting evidence that is clinically meaningful. Similarly, the qualitative analysis inevitably involves a degree of subjective interpretation. Therefore, the findings of this study should be viewed as

evidence regarding the relative reliability of the evaluated explanation methods.

7 Conclusions and Further Research

This thesis investigated the extent to which explanation methods provide reliable and robust insights into the decision-making processes of convolutional neural networks and Vision Transformer models for COVID-19 chest X-ray classification. To address this question, a ResNet-50 CNN and a Vision Transformer were evaluated using multiple explanation methods and a combination of quantitative, qualitative and robustness-based evaluation techniques. The findings show that the chosen explanation technique greatly influences explanation quality and cannot be well recorded by one assessment measure. LRP offered the best general balance of stability, robustness and faithfulness across the two models. One probable cause of this performance is LRP’s direct redistribution of prediction relevance across the network to generate explanations that stay quite related to the internal processes of the model. Conversely, LIME produced logical and localized explanations but was more sensitive to perturbations, which indicates that its local surrogate modeling approach might add more variation. The least reliable explanations came from Attention Rollout, which suggests that merely using attention weights may not provide a faithful depiction of the factors driving model predictions. The study also found that model robustness and explanation quality are connected but different ideas. The Vision Transformer was more robust to peripheral picture perturbations than the CNN but the explanation quality still mostly relied on the explanation technique rather than the underlying design. These results underline the need of assessing explanations from several angles before forming opinions about model behavior. Future studies may look at whether the noted effects apply to more datasets, designs and medical imaging applications such as MRI scans. A more thorough evaluation of explainability strategies might be possible with a more extensive comparison that included techniques like SHAP, Integrated Gradients and Grad-CAM variations. Beyond evaluating additional methods, future work should also focus on improving the reliability of existing explanation approaches. The sensitivity found for LIME implies that in order to lower explanation variability, more stable perturbation procedures or surrogate modeling techniques could be needed. Attention Rollout’s poor performance emphasizes the need for transformer-specific explanation techniques that go beyond simple attention weights and more precisely represent the mechanism driving model predictions. Future research may find a viable path in hybrid systems that integrate transformer representations with attribution-based techniques. At last, adding expert clinical evaluation and radiologist notes will clarify if explanation techniques emphasize clinically significant data besides meeting good quantitative performance.

References

- [ADAS⁺25] Sanad Aburass, Osama Dorgham, Jamil Al Shaqsi, Maha Abu Rumman, and Omar Al-Kadi. Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases. *J. Imaging Inform. Med.*, 38(6):3928–3971, December 2025.
- [AGM⁺18] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. October 2018.
- [AMJ18] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. June 2018.
- [ANK⁺26] Fahad Ahmed, Naila Sammar Naz, Sunawar Khan, Ateeq Ur Rehman, Waleed M Ismael, and Muhammad Adnan Khan. Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges. *BMC Med. Imaging*, 26(1):37, January 2026.
- [AR24] João Abrantes and Pouria Rouzrokh. Explaining explainability: The role of XAI in medical imaging. *Eur. J. Radiol.*, 173(111389):111389, April 2024.
- [ATE18] Muhammad Aurangzeb Ahmad, Ankur Teredesai, and Carly Eckert. Interpretable machine learning in healthcare. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 447–447, 2018.
- [AZ20] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. May 2020.
- [BA21] Pedro Bassi and Romis Attux. A deep convolutional neural network for COVID-19 detection using chest x-rays. 2021.
- [BBM⁺15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10(7):e0130140, July 2015.
- [BSN⁺23] Katarzyna Borys, {Yasmin Alyssa} Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, {Christoph M.} Friedrich, and Felix Nensa. Explainable ai in medical imaging: An overview for clinical practitioners - saliency-based xai approaches. *European journal of radiology*, 162:110787, May 2023. Publisher Copyright: Copyright © 2023 Elsevier B.V. All rights reserved.
- [BWM20] Umang Bhatt, Adrian Weller, and José M F Moura. Evaluating and aggregating feature-based model explanations. May 2020.
- [CC24] Mateusz Cedro and Marcin Chlebus. Beyond the black box: Do more complex deep learning models provide superior XAI explanations? May 2024.

- [CEVV23] Marios Constantinou, Themis Exarchos, Aristidis G Vrahatis, and Panagiotis Vlamos. COVID-19 classification on chest x-ray images using deep learning methods. *Int. J. Environ. Res. Public Health*, 20(3):2035, January 2023.
- [CGW20] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. December 2020.
- [DBK⁺20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. October 2020.
- [DJL20] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv*, October 2020.
- [FDAC24] Miguel Fontes, João Dallyson Sousa De Almeida, and António Cunha. Application of example-based explainable artificial intelligence (xai) for analysis and interpretation of medical imaging: A systematic review. *IEEE Access*, 12:26419–26427, 2024.
- [GRB⁺22] Arne Gevaert, Axel-Jan Rousseau, Thijs Becker, Dirk Valkenburg, Tijl De Bie, and Yvan Saeys. Evaluating feature attribution methods in the image domain. February 2022.
- [HBPK17] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable AI systems for the medical domain? December 2017.
- [HVV22] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1544–1554, 2022.
- [HWB⁺22] Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. February 2022.
- [HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. December 2015.
- [ISST24] Tanzina Taher Ifty, Saleh Ahmed Shafin, Shoeb Mohammad Shahriar, and Tashfia Towhid. Explainable lung disease classification from chest x-ray images utilizing deep learning and XAI. April 2024.
- [JJP25] Sehyeong Jo, Gangjae Jang, and Haesol Park. Gmar: Gradient-driven multi-head attention rollout for vision transformer interpretability. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 582–587, 2025.
- [JW19] Sarthak Jain and Byron C Wallace. Attention is not explanation. February 2019.

- [KBB23] Piotr Komorowski, Hubert Baniecki, and Przemysław Biecek. Towards evaluating explanations of vision transformers for medical imaging. April 2023.
- [KHA⁺17] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. November 2017.
- [LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.
- [LWB⁺19] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. February 2019.
- [MB24] Dost Muhammad and Malika Bendeche. Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. *Comput. Struct. Biotechnol. J.*, 24:542–560, December 2024.
- [MY20] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-CAM: Class activation map using principal components. August 2020.
- [NMHA23] Saad I Nafisah, Ghulam Muhammad, M Shamim Hossain, and Salman A AlQahtani. A comparative evaluation between convolutional neural networks and vision transformers for COVID-19 detection. *Mathematics*, 11(6):1489, March 2023.
- [RIZ⁺17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. November 2017.
- [RPDMR25] Tim Rätz, Aurélie Pahud De Mortanges, and Mauricio Reyes. Explainable AI in medicine: challenges of integrating XAI into the future clinical routine. *Front. Radiol.*, 5:1627169, August 2025.
- [RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should I trust you?”: Explaining the predictions of any classifier. February 2016.
- [SCD⁺16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. October 2016.
- [STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. March 2017.
- [TCK⁺21] Anas M Tahir, Muhammad E H Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. COVID-19 infection localization and severity grading from chest x-ray images. *Comput. Biol. Med.*, 139(105002):105002, December 2021.

- [TG19] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): Towards medical XAI. July 2019.
- [WW20] Linda Wang and Alexander Wong. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. 2020.
- [XUD⁺19] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges*, pages 563–574. 09 2019.
- [ZBL⁺18] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.*, 15(11):e1002683, November 2018.