ENHANCING THE STABILITY AND QUALITY ASSESSMENT OF VISUAL EXPLANATIONS FOR THORAX DISEASE CLASSIFICATION USING DEEP LEARNING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SHAKIBA RAHIMIAGHDAM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

SEPTEMBER 2023

Approval of the thesis:

ENHANCING THE STABILITY AND QUALITY ASSESSMENT OF VISUAL EXPLANATIONS FOR THORAX DISEASE CLASSIFICATION USING DEEP LEARNING

submitted by **SHAKIBA RAHIMIAGHDAM** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Assoc. Prof. Dr. Hande Alemdar
Supervisor, **Computer Engineering** _____

Examining Committee Members:

Prof. Dr. Pınar Karagöz
Computer Engineering, METU _____

Assoc. Prof. Dr. Hande Alemdar
Computer Engineering, METU _____

Prof. Dr. Ahmet Burak Can
Computer Engineering, Hacettpe University _____

Assoc. Prof. Dr. Emre Akbaş
Computer Engineering, METU _____

Assoc. Prof. Dr. Özlem Durmaz İncel
Computer Engineering, Boğaziçi University _____

Date: 08.09.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Shakiba Rahimiaghdam

Signature :

ABSTRACT

ENHANCING THE STABILITY AND QUALITY ASSESSMENT OF VISUAL EXPLANATIONS FOR THORAX DISEASE CLASSIFICATION USING DEEP LEARNING

Rahimiaghdam, Shakiba
Ph.D., Department of Computer Engineering
Supervisor: Assoc. Prof. Dr. Hande Alemdar

September 2023, 79 pages

Deep learning models are renowned but their complex internal workings often render them opaque. While efforts are underway to amplify their explainability, a substantial gap persists. One pivotal challenge is the unavailability of quantifiable metrics for evaluating visual explanations, leading to a reliance on manual assessments or suboptimal metrics. This restricts scalability, compromises reproducibility, and undermines trustworthiness. The necessity for objective metrics is further underscored during hyperparameters fine-tuning and the evaluation of various Explainable Artificial Intelligence (XAI) models. Another prominent hurdle is the instability exhibited by models like Local Interpretable Model-agnostic Explanations (LIME). The random perturbations intrinsic to LIME lead to inconsistent explanations, eroding trust and obstructing their integration into critical applications.

To navigate these challenges, we unveil a robust method for the objective assessment, refinement, and juxtaposition of visual explanation algorithms, offering a concrete solution to metric inadequacy. Addressing instability, we introduce MindfulLIME. This cutting-edge approach leverages graph-based pruning and uncertainty sampling,

strategically crafting purposeful samples and elevating the reliability and consistency of visual explanations - a discernible advancement over LIME. Our intricate qualitative analysis, applied to unseen random samples from the VinDr-CXR dataset, attests to the preeminence of our selected metric. In a comparative analysis involving multi-label, multi-class diagnosis of thorax diseases, MindfulLIME, evaluated via this refined metric, epitomizes optimal stability and augmented localization accuracy. This underscores its capacity to elevate the trust quotient associated with machine learning applications in the nuanced field of medical imagery.

Keywords: Explainable Artificial Intelligence (XAI), Visual Explanation Evaluation, Machine Learning, Deep Learning Models, Thorax Diseases Diagnosis

ÖZ

DERİN ÖĞRENME KULLANARAK GÖĞÜS HASTALIKLARI SINIFLANDIRMASI İÇİN GÖRSEL AÇIKLAMALARIN KARARLILIK VE KALİTE DEĞERLENDİRMESİNİ GELİŞTİRME

Rahimiaghdam, Shakiba
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Doç. Dr. Hande Alemdar

Eylül 2023 , 79 sayfa

Derin öğrenme modellerinin karmaşık iç işleyişi onları anlaşılmaz hale getirmektedir. Açıklanabilirliklerini artırmaya yönelik çabalardaki önemli zorluklardan biri, görsel açıklamaları değerlendirmek için ölçülebilir ölçümlerin mevcut olmamasıdır. Bu durum, manuel değerlendirmelere veya optimal olmayan ölçümlere güvenilmesine yol açmaktadır. Bu, ölçeklenebilirliği kısıtlar, tekrarlanabilirliği tehlikeye atar ve güvenilirliği zayıflatır. Objektif ölçümlerin gerekliliği, hiperparametrelerin ince ayarı ve çeşitli Açıklanabilir Yapay Zeka (XAI) modellerinin değerlendirilmesi sırasında daha da vurgulanır. Öne çıkan bir diğer engel ise Yerel Yorumlanabilir Model-Agnostik Açıklamalar (LIME) gibi yöntemlerin sergilediği istikrarsızlıktır. LIME yöntemine özgü rastgele karışıklıklar tutarsız açıklamalara yol açarak güvenin aşınmasına ve kritik uygulamalarda kullanılamamasına neden olur.

Bu zorlukların üstesinden gelmek için, görsel açıklama algoritmalarının nesnel değerlendirmesi, iyileştirilmesi ve yan yana getirilmesi için metrik yetersizliğine somut bir çözüm sunan sağlam bir yöntem ortaya koyuyoruz. İstikrarsızlığa çözüm bulmak

için MindfulLIME yöntemini tanıtıyoruz. Bu yeni yaklaşım, grafik tabanlı budama ve belirsizlik örneklemesinden yararlanarak, stratejik olarak amaca yönelik örnekler oluşturur ve görsel açıklamaların güvenilirliğini ve tutarlılığını artırır. Bu, LIME yöntemine göre gözle görülür bir ilerlemedir. VinDr-CXR veri setindeki görünmeyen rastgele örneklere uygulanan nitel analizimiz, seçtiğimiz metriğin üstünlüğünü kanıtlamaktadır. Toraks hastalıklarının çok etiketli, çok sınıflı teşhisini içeren karşılaştırmalı bir analizde, önerilen hassas ölçüm metriği aracılığıyla değerlendirilen MindfulLIME, optimal stabilite ve artırılmış lokalizasyon doğruluğu ortaya koymaktadır. Bu durum, tıbbi görüntülerle makine öğrenimi uygulamalarında güven oranını artırma kapasitesinin altını çizmektedir.

Anahtar Kelimeler: Açıklanabilir Yapay Zeka, Görsel Açıklama Değerlendirmesi, Makine Öğrenmesi, Derin Öğrenme Modelleri, Göğüs Hastalıkları Teşhisi

To my loving spouse, Shahram, my supportive parents, Fariba & Rahim, and the cherished memory of my late grandmothers and aunt, Ameneh,

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Hande Alemdar, for her exceptional guidance, unwavering patience, and continuous motivation throughout the course of this thesis. Her insights, encouragement, and dedication have been invaluable, and without her support, this accomplishment would not have been possible.

I am also deeply appreciative of my former supervisor, Prof. Dr. Adnan Yazici, for the immense support and guidance provided during my PhD journey. His mentorship and direction in shaping this thesis are deeply valued.

I would also like to extend my heartfelt appreciation to the esteemed jury members, Prof. Dr. Pınar Karagöz, Prof. Dr. Ahmet Burak Can, Assoc. Prof. Dr. Emre Akbaş, and Assoc. Prof. Dr. Özlem Durmaz İncel, who dedicated their invaluable time and expertise to review and evaluate this thesis. Their insightful feedback and constructive suggestions have significantly contributed to the refinement of this work.

My gratitude also extends to the exceptional instructors who have shaped my educational path at the University of METU. Each one, including Prof. Dr. Adnan Yazıcı, Prof. Dr. Ahmet Coşar, Prof. Dr. Ali Hikmet Doğru, Dr. Ayşenur Birtürk, Assoc. Prof. Dr. Emre Akbaş, Prof. Dr. Ertan Onur, Prof. Dr. Halit Oğuztüzün, Prof. Dr. İsmail Sengör Altıngövde, Prof. Dr. Murat Manguoğlu, Prof. Dr. Nihan Kesim Çiçekli, Prof. Dr. Pınar Karagöz, Assoc.Prof. Dr. Sinan Kalkan, Assoc. Prof. Dr. Şeyda Ertekin, and Prof. Dr. Volkan Atalay, has imparted invaluable knowledge and guidance that enriched my academic journey.

I am profoundly thankful for the unwavering support of my spouse, Shahram, whose patience, encouragement, and selflessness were instrumental during the challenging phases of this endeavor. His enduring belief in me and willingness to go above and beyond are forever cherished.

The unwavering encouragement and steadfast backing provided by my parents, Fariba and Rahim, deeply move me. Their consistent faith in my abilities serves as an unending wellspring of inspiration and strength, propelling me forward with resolute determination through every phase of my journey.

I extend my heartfelt appreciation to my friends, Zeynep Ertekin, Nurdan Berber, Saeid Sattari, Masoud Sattari, Ezgi Ozyavas, Turgay Pamuklu, Mahila Zamanpour, and Navid Amirirad who have stood by me through the years, offering assistance, motivation, and camaraderie. Their friendship has provided a sense of belonging and encouragement that I deeply treasure.

I am humbled by the collective support and encouragement of all these individuals, without whom this achievement would not have been possible. Thank you for being an integral part of this journey.

With gratitude,
Shakiba Rahimiaghdam

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xii

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xviii

CHAPTERS

1 INTRODUCTION 1

1.1 Research Problem and Approach 1

1.2 Contributions 5

1.3 Structure of the Thesis 6

2 LITERATURE REVIEW 9

2.1 Background 9

2.2 Visual Explanation Evaluation 11

2.3 Enhanced Perturbation-Based XAI Models 14

3 DATA SETS AND TRAINED CLASSIFIER 19

3.1 Datasets 19

3.2 Training The Classifier 21
4 QUANTITATIVE VISUAL EXPLANATION EVALUATION 25
4.1 Desired Criteria 25
4.2 Localization-based Evaluation Metrics 27
4.3 Evaluation Metric Selection 30
4.4 Experimental Results and Discussions 31
4.4.1 Qualitative Analysis 32
4.4.1.1 Segmentation Algorithms Exploration 32
4.4.1.2 Evaluation Metric Exploration 35
4.4.2 Quantitative Analysis 40
5 PURPOSIVE PERTURBATION 43
5.1 LIME 43
5.2 MindfulLIME 45
5.3 Experimental Results and Discussions 52
5.3.1 Qualitative Analysis 52
5.3.2 Quantitative Analysis 55
6 CONCLUSION AND FUTURE WORK 63
REFERENCES 67
CURRICULUM VITAE 77

LIST OF TABLES

TABLES

Table 3.1 Comparing different trained classifiers by their AUC percentages for each class 22

Table 4.1 Summary of the essential parameters for each evaluation metric . . . 31

Table 4.2 Segmentation algorithms parameters 33

Table 4.3 Metric results for the examples violate *Informativeness* and *Localization* shown in Fig. 4.4 38

Table 4.4 Metric results for the examples violate *Coverage* and *Proportionality* shown in Fig. 4.5 38

Table 4.5 Comparison of the LIME method with various segmentation algorithms for the first three predicted classes 41

Table 5.1 Segmentation Algorithms Parameters 57

Table 5.2 Run-Time Comparison: LIME vs. MindfulLIME for various versions and superpixel numbers 62

LIST OF FIGURES

FIGURES

Figure 3.1 Samples of the VinDr-CXR dataset with precise annotated labels 21

Figure 3.2 Distribution of samples within classes for both CheXpert and VinDrCXR datasets 23

Figure 4.1 Exploring Various Segmentation Algorithms: Given superpixels to the LIME generated by different segmentation algorithms 33

Figure 4.2 The first row displays explanations generated by the segmentation algorithm with default parameters, and the second row displays explanations generated for the same segmentation algorithms with the optimized parameter values. The brown (darker) rectangles show the radiologists' annotations for detected lung consolidation, and the orange (lighter) rectangles show the border of explanations (yellow-bordered area) generated by LIME. 34

Figure 4.3 The first row displays explanations generated by the segmentation algorithm with default parameters, and the second row displays explanations generated for the same segmentation algorithms with the optimized parameter values. The dark blue rectangles show the radiologists' annotations for the detected Lung Opacity condition, and the light blue rectangles show the border of explanations (yellow-bordered area) generated by LIME. 35

Figure 4.4 Examples that violate *Informativeness* and *Localization* are shown for a single case with different segmentation algorithms. The dark blue rectangles show the radiologists' annotations for the detected Lung Opacity condition, and the light blue rectangles show the border of explanations (yellow-bordered area) generated by LIME. 36

Figure 4.5 Examples that violate *Coverage* and *Proportionality* are shown for a single case with different segmentation algorithms. The darker orange, green, and blue rectangles show the radiologists' annotations for the detected Cardiomegaly, Pleural Effusion, and Lung Opacity conditions. The lighter orange, green, and blue rectangles show the border of explanations (yellow-bordered area) generated by LIME. Each row shows a different case with three distinct segmentation settings. The left-most examples demonstrate the more desirable result, and the right-most ones show the less desirable explanations. 37

Figure 5.1 A block diagram illustrating the high-level structure of the LIME framework. 44

Figure 5.2 A block diagram illustrating the high-level structure of the MindfulLIME framework. 46

Figure 5.3 Visual Analysis: Comparing generated explanations (the yellow-outlined areas enclosed with light green bounding boxes) by MindfulLIME and LIME to the ground truth annotations (dark green bounding boxes) for Random Sample 1 53

Figure 5.4 Visual Analysis: Comparing generated explanations (the yellow-outlined areas enclosed with light blue bounding boxes) by MindfulLIME and LIME to the ground truth annotations (dark blue bounding boxes) for Random Sample 2 54

Figure 5.5 Exploring Various Segmentation Algorithms: Given superpixels to the LIME generated by different segmentation algorithms 56

Figure 5.6 Stability Comparison: LIME vs. MindfulLIME for various versions and settings 58

Figure 5.7 Quality Comparison: LIME vs. MindfulLIME for various versions and settings 59

Figure 5.8 Run-Time Comparison: LIME vs. MindfulLIME for various versions and superpixel numbers 60

LIST OF ABBREVIATIONS

AHC	Agglomerative Hierarchical Clustering
AI	Artificial Intelligence
AP	Anterior Posterior
AUC	Area Under the ROC Curve
Bbox	Bounding Box
B-LIME	Bootstrap-LIME
CAM	Class Activation Maps
CNN	Convolutional Neural Network
CT	Computed Tomography
CTGAN	Conditional Tabular GAN
CXR	Chest X-ray
DICOM	Digital Imaging and Communications in Medicine
DLIME	Deterministic Local Interpretable Model-Agnostic Explanations
EBPG	Energy-Based Pointing Game
ECG	Electrocardiogram
FCA	Formal Concept Analysis
GradCAM	Gradient-weighted Class-Activation Maps
IoGS	Intersection over Ground truth Saliency
IoSR	Intersection over Salient Region
IoU	Intersection over Union
JS_DIV	Jensen-Shannon Divergence
KL_DIV	Kullback-Leibler Divergence
KNN	K-Nearest Neighbor
LIME	Local Interpretable Model-agnostic Explanations

MAE	Mean Absolute Error
ML	Machine Learning
MPS-LIME	Modified Perturbed Sampling for LIME
OoD	Out-of-Distribution
PA	Posterior Anterior
PG	Pointing Game
PWA	Pixel Wise Accuracy
PWP	Pixel Wise Precision
PWR	Pixel Wise Recall
RBF	Radial Basis Function
SLIC	Simple Linear Iterative Clustering
WIoSR	Weighted Intersection over Salient Region
XAI	Explainable Artificial Intelligence

CHAPTER 1

INTRODUCTION

The advent of machine learning-based systems has revolutionized numerous complex tasks in our lives, ranging from image object detection and text processing to speech signal analysis (Dong et al., 2021; Prasanna & Tripathi, 2022). These systems have demonstrated remarkable performance, often surpassing human-level capabilities. However, their inherent non-transparency, non-linearity, and nested internal computations pose a challenge in establishing user trust (von Eschenbach, 2021). Most machine learning algorithms lack traceability, resembling black boxes where decisions are made without clear explanations based on input data. Hence, it is crucial to have deep learning models that are explainable and transparent in order to enable their widespread utilization across different vertical domains, particularly in critical areas like justice, cybersecurity, and medicine (Collenette et al., 2023; Giudici & Raffinetti, 2022; Jin et al., 2022).

1.1 Research Problem and Approach

While simpler algorithms like decision trees and linear regression are easier to understand due to their simple representations, more complex models like neural networks and random forests are often treated as black boxes. To address this issue, recent literature in the field of Explainable Artificial Intelligence (XAI) proposes interpretable and explainable models as a means to achieve transparency for complex models (Barredo Arrieta et al., 2020). The explainable models can be classified based on criteria such as being model-agnostic or model-specific, local or global, and intrinsic or post hoc (Linardatos et al., 2021). Among all, post hoc model-agnostic

approaches are popular as they aim to explain the decision-making process of any machine learning model. An intrinsically explainable model provides insights directly from a model's parameters or structure, as seen in linear models or decision trees, but they may lack the power required for modern tasks. Post hoc explanations, on the other hand, provide insights after a model is trained and can be model-specific or model-agnostic. Model-agnostic methods, such as LIME (Y. Zhang et al., 2019), Anchors (Ribeiro et al., 2018), and SHAP (Lundberg & Lee, 2017), are popular and do not require knowledge of the internal model structure. In other words, post hoc explanation methods are model-independent and determine the importance of features for predictions.

A key algorithm in this category is LIME, which operates on the principle that, within a local vicinity of the classification boundary, a less complex model, such as logistic regression (the choice of model may vary depending on the data), can yield an explanation that is easier for users to grasp. LIME leverages such simpler models to decode the decision-making process and highlight key features through regression coefficients. It creates simulated data points by perturbing instances and fits a sparse linear model using the responses predicted from these altered points. This multifaceted approach facilitates explanations for a spectrum of classification models, from random forests and gradient boosting trees to neural networks, across diverse input data types such as text, images, or tabular data. Significantly, the explanations LIME provides remain locally faithful to an instance, regardless of the classifier utilized (Y. Zhang et al., 2019).

However, the random perturbation process employed in LIME and similar local explanation methods exhibits certain limitations (Hooker et al., 2021). The non deterministic nature of these algorithms, stemming from the random perturbations, compromises stability, which is a crucial characteristic for interpretable models, particularly in critical domains. Several studies have emphasized the significance of stability, which entails providing consistent explanations for similar inputs, and proposed various methods to enhance this aspect (Alvarez-Melis & Jaakkola, 2018; Khaire & Dhanalakshmi, 2022; Vilone & Longo, 2021). Furthermore, random perturbation techniques can introduce data and label shifts, potentially leading to misleading explanations (Rahnama & Boström, 2019; Slack et al., 2020). Additionally, these

methods often generate perturbed samples from a uniform distribution, overlooking the intrinsic correlation between features. As a result, valuable information for learning local explanation models is lost. Therefore, it becomes essential to ensure proper sampling operations to address and resolve instability, prevent the generation of Out-of-Distribution (OoD) data, and efficiently produce samples of higher quality, particularly in sensitive domains such as medicine.

This work delves into post hoc explanations that utilize perturbations, a widely embraced paradigm for crafting model explanation methods, with a specific focus on image data. Despite significant advancements in the stability of other data modalities, such as text and tabular data, a pronounced gap persists for image data. To address this, we introduce MindfulLIME, an innovative non-random sample generation algorithm engineered expressly to tackle the challenges of stability and reproducibility in repeated runs. The name MindfulLIME mirrors our endeavor to augment LIME, a widely-used model-agnostic explanation method, by infusing it with intelligent sampling strategies that render it more mindful of the stability in the explanations it generates.

While many alternative methods in the XAI arena typically employ random sample perturbations, we deliberately opted for LIME to effectively demonstrate and evaluate the performance of MindfulLIME. Our decision to center on medical images and specifically X-ray images in this study is driven by their pivotal role in medical diagnostics, the complexities and challenges tied to interpreting these images, and the accessibility of publicly available datasets for experimentation. This strategy not only accentuates the specific application of medical images addressed herein but also suggests the potential for MindfulLIME to be efficaciously generalized across other domains.

In this context, *explanation* term refers to the superpixels or segments of the image that lead the classifier to a particular conclusion. A superpixel is a cluster of pixels that share similar characteristics and are processed as a unit. These are the critical segments of the image that the classifier leverages to render its final judgment, and our primary goal is to enhance the stability of these visual explanations. Our approach involves transforming superpixels into an undirected graph, where each seg-

ment represents a vertex, and adjacent superpixels are connected through edges. This transformation facilitates the systematic generation of new samples in each iteration by deactivating a combination of superpixels from the original image using a defined algorithmic procedure. The suitability of each generated sample is then evaluated to decide whether to keep or discard it. To aid in this decision-making process, we employ uncertainty sampling, which involves selecting samples based on the classifier's confidence level. This technique prioritizes samples for which the classifier exhibits lower confidence (uncertainty sampling) Lewis and Gale, 1994 or higher confidence (reversed uncertainty sampling) Cardellino and Alemany, 2015, enabling us to bolster the quality and reliability of our explanations. By focusing on samples when the classifier is more confident, we ensure their informativeness and representation of the data distribution while mitigating the risk of generating out-of-distribution samples through sample filtering based on the classifier's confidence. Additionally, our approach fully incorporates the interplay between features to enhance the quality of explanations.

In ensuring the objective and computationally efficient evaluation of our proposed algorithm, the right metrics play an indispensable role. Another pivotal contribution of this study bridges a gap in the current literature, introducing a robust quantitative evaluation for generated visual explanations. We start by defining criteria essential for the ideal metrics to measure the quality of explanations, particularly in terms of localization. While we delve into metrics already recognized in computer vision, our research also brings forth new metrics, enriching the available repertoire. These metrics are systematically assessed against our criteria in various scenarios, shedding light on their efficacy. We then pinpoint the most effective metric that addresses the broader challenge of evaluating visual explanation quality, transcending the specific scope of our study. Following this, we showcase the practical utility of the selected metric through both qualitative and quantitative analyses. Qualitatively, we assess various metrics, with our chosen one proving its superiority. Quantitatively, the metric aids in refining LIME across diverse segmentation techniques and facilitates a performance comparison with Gradient-weighted Class-Activation Maps (GradCAM), a notable saliency map model in XAI.

Given the sensitivity of the medical domain, the importance of explainability can-

not be overstated, especially when it comes to analyzing Chest X-ray (CXR) images for automated thorax disease diagnosis. Relying solely on human expertise for CXR interpretation poses challenges due to human limitations and the complexity of the task, increasing the risk of misdiagnosis (Fourcade & Khonsari, 2019; van der Velden et al., 2022). To address this issue and provide AI-driven automated tools with explainable decisions, we have developed a framework that leverages two CXR image datasets. For training our classifier, we utilized the CheXpert CXR image dataset (Irvin et al., 2019), which offers a substantial number of samples for each class. To evaluate the spatial precision of explanations, we employed the VinDr-CXR dataset (Nguyen et al., 2020), which provides precise local annotations for the same classes as the CheXpert dataset, annotated by a group of radiologists.

The experimental findings prominently highlight the superiority of our novel algorithm, named MindfulLIME. Notably, MindfulLIME not only achieves a remarkable stability rate of 100%, but also outperforms LIME in terms of precision in localization and efficiency of execution. Furthermore, we delve into the examination of diverse parameters, including segmentation size and quantity, as well as various sample numbers, influencing the quality of explanations generated by the aforementioned algorithms. Our devised framework, coupled with the exceptional performance exhibited by MindfulLIME, effectively showcases the potential of AI-driven automated tools within the realm of thorax disease diagnosis. By furnishing dependable and precise explanations, our framework empowers medical professionals with heightened decision-making acumen and significantly contributes to the advancement of medical research.

1.2 Contributions

We summarize our contributions as follows:

- **Objective Evaluation Framework for Visual Explanations:** We propose criteria tailored to identify the optimal metrics for evaluating visual explainability algorithms. While our methodology integrates established metrics, it also pioneers new ones. After evaluating the metrics' formulations against our criteria,

we discern a leading metric for visual explanation quality. This metric plays a pivotal role in optimizing LIME and facilitates quantitative comparisons with other XAI models.

- **Explanation Stability Enhancement:** Our approach leverages graph representation to systematically select combinations of superpixels, employing a non-random sampling technique. This technique addresses the stability issue in generating explanations while improving the quality of explanations and enhancing interpretability.

- **Explanation Quality Refinement:** We incorporate uncertainty sampling techniques to ensure reliable and trustworthy explanations. By assessing the acceptability of each sample based on the classifier's confidence, we enhance the quality of generated explanations and filter out potential out-of-distribution samples.

- **Application to Medical Diagnosis:** We rigorously assess the effectiveness of our framework within the domain of automated thorax disease diagnosis utilizing chest X-ray images. By contrasting the performance of the conventional LIME with our novel algorithm, MindfulLIME, employing varying superpixel algorithms and sample quantities, we gained valuable insights. Having undergone training on the CheXpert dataset and subsequent evaluation on the unseen VinDr-CXR dataset, MindfulLIME demonstrated an impeccable stability rate of 100% and showcased superior efficiency in terms of localization precision when compared to LIME.

1.3 Structure of the Thesis

The manuscript's organization unfolds as follows: In Chapter 2, we present a concise overview of the LIME algorithm. Alongside this, we outline pertinent studies from the literature pertaining to visual explanations and delve into the exploration of metrics commonly utilized for the evaluation of explainable AI (XAI) models. Moreover, Chapter 2 comprehensively reviews existing literature studies that emphasize enhancements to perturbation-based sampling algorithms, including LIME.

Moving to Chapter 3, we meticulously detail the datasets we have chosen, as well as the optimal classifier we trained for this study. Our dataset selection is carefully tailored to fulfill the dual roles of training the classifier and generating/evaluating visual explanations.

Transitioning to Chapter 4, we embark on our primary contribution, conducting an in-depth exploration of explanation assessment and proposing appropriate localization-based evaluation metrics for visual explanations. This section meticulously expounds on the insights and conclusions drawn from our extensive evaluations.

Advancing to Chapter 5, we introduce our secondary innovative approach, which centers around a perturbation-based XAI model that is deterministic in nature, ensuring stable and consistent explanations. In tandem, we provide a comprehensive explanation of our novel algorithm, MindfulLIME. Concluding this chapter, we present a thorough discussion and showcase the outcomes gleaned from meticulous evaluations, which compare MindfulLIME and LIME. These evaluations encompass a diverse array of assessment criteria rooted in various superpixel algorithms and sample counts.

Finally, Chapter 6 marks the conclusion of our endeavor, wherein we succinctly summarize the conclusions we have drawn. Additionally, we offer valuable insights into potential pathways for future research endeavors.

CHAPTER 2

LITERATURE REVIEW

In this chapter, our aim is to delve into the literature that pertains to our two main contributions. The first centers around the accurate and effective evaluation of visual explanations, which is a prerequisite for discerning the superiority of AI models. The second area of focus involves the introduction of innovative techniques to enhance perturbations-based XAI models. These techniques are meticulously designed to bolster stability and tackle associated challenges. To demonstrate the efficacy of our proposed approaches, we utilize the capabilities of LIME, employing it as a benchmark for illustration. We commence by elucidating the mechanics of LIME, which serves as the foundation for ensuing discussions. Subsequently, we conduct an in-depth survey to examine prevailing methodologies for assessing the effectiveness of visual explanations. Following this, we explore methods that have enhanced perturbation-based XAI algorithms, addressing issues like stability, data shift (caused by out-of-distribution data generation), and the neglect of feature correlations.

2.1 Background

LIME, which stands for Local Interpretable Model-agnostic Explanations, is a modern technique designed to provide understandable and faithful explanations for predictions made by any classifier. Its primary purpose is to demystify how classifiers arrive at their decisions. This is achieved by initially selecting a specific example for which an explanation of the classifier's decision is needed. LIME then creates perturbed samples by introducing controlled changes to the interpretable elements of the chosen instance.

The core idea behind LIME is to understand how a black box model behaves locally in the vicinity of a specific point of interest. Remarkably, LIME achieves this understanding without requiring detailed knowledge about the internal workings of the model. To overcome this challenge, LIME generates a simulated dataset by creating multiple new examples that are close to the original point of interest. These new examples are used to query the black box model and gather predictions. The outcome is a collection of pairs, where each pair consists of an example and the corresponding prediction made by the black box model.

The black box model can take various forms, such as regression or classification. To make the explanation easily understandable, LIME selects a simpler model, termed the explanation model, from a set of models that are interpretable. This selection process involves finding the explanation model that best approximates the behavior of the black box model in the local vicinity of the point of interest.

In essence, LIME makes slight changes to the features of an example and observes how these changes impact the predictions. By systematically modifying one feature at a time, LIME measures the difference between the new prediction and the original prediction for the example. This difference helps quantify the importance of each feature in the decision-making process.

When dealing with image data, LIME simplifies the analysis by grouping pixels into clusters called superpixels. These superpixels represent meaningful regions in the image. During the explanation process, LIME deactivates certain superpixels to observe how the absence of those regions affects the predictions. This approach allows LIME to provide insights into which parts of an image contribute most to a particular classification decision.

In summary, LIME is a technique that dissects how classifiers arrive at their predictions. It achieves this by generating new examples that are similar to the original and observing how changes in those examples impact the predictions. By doing so, LIME provides understandable explanations for complex model behaviors, even when the inner workings of the models are not fully known.

2.2 Visual Explanation Evaluation

Different types of explanations can take various forms, such as text, math (numbers), and visuals. Visual explanations are often preferred since they make it simpler to understand and interpret the results of complex models. One common way to create such visuals is by using saliency maps. In our study, we take the saliency map output from LIME and compare it to expert-annotated ground-truth masks. This lead us to focus on visual explanations generated by LIME, particularly for medical image data. We are particularly interested in metrics that assess the similarity or dissimilarity between saliency maps, with a focus on localizing multi-label classifiers.

We survey recent works concerning visual explanations for medical images. For instance, Xiang and Wang (Xiang & Wang, 2019) used LIME on a skin lesion classifier to verify that the trained Convolutional Neural Network (CNN) correctly uses relevant information for diagnosis. Similarly, Arrieta et al. (Barredo Arrieta et al., 2020) employed LIME explanations to show that deep learning models perform better on chest X-ray images than Computed Tomography (CT) scans for classifying COVID-19 patients. As another example, Rajaraman et al. (Rajaraman et al., 2018) evaluated, visualized, and explained the behavior of trained models for detecting pneumonia by employing popular methods such as LIME, Class Activation Maps (CAM), and its other variations.

Ahsan et al. (Ahsan et al., 2020) applied LIME to detect essential features distinguishing COVID-19 patients from other patients using CT and CXR images. Lastly, in another work, Teixeira et al. (Teixeira et al., 2021) utilized LIME and GradCAM to evaluate the impact of lung segmentation using a U-Net CNN architecture (Ronneberger et al., 2015) for generating segmentations over three well-known CNN classifiers. Since the authors' primary objective is to explore which part of the image effectively contributes to the identification, they applied explanation models to a test set of images. By aggregating all explanation regions in a heatmap, they showed the model's most common areas for prediction.

There are several taxonomies in the literature for visual explanation evaluation. Based on the presented classification by Arrieta et al. (Barredo Arrieta et al., 2020), outputs

are mainly discussed in two broad groups: qualitative and quantitative analysis.

In qualitative evaluation, the outcomes of one or more models are visually investigated for a limited number of random samples. Quantitative analysis can be categorized into two sub-groups: model truth-based and ground truth-based (Sattarzadeh et al., 2021). The former investigates the model behavior by assessing the algorithm's consistency, correctness, and faithfulness. The latter justifies the visual quality of the explanation maps generated through explanation models by comparing them to the ground truth maps determined by amateur or expert users.

According to another classification, assessment approaches can be divided into three classes: application-grounded, human-grounded, and functionally-grounded (DoshiVelez & Kim, 2018). Application-grounded and human-grounded assessments resemble ground truth-based methods, and functionally-grounded methods are similar to model-based ones.

Qualitative analysis plays a crucial role in assisting users in evaluating the quality of the explanations. However, more than these analyses are needed to establish or generalize the explanations' correctness, appropriateness, and precision. Hence, we investigate the notable metrics used in model truth-based and ground truth-based evaluations.

Model truth-based metrics generally evaluate the explanations' faithfulness, and correctness (Li et al., 2021). As Alvarez and Jaakkola claimed (Alvarez-Melis & Jaakkola, 2018), the faithfulness of an explanation refers to whether the relevance scores reflect the true importance. Drop percentage, increase percentage (Chattopadhay et al., 2018; Ramaswamy & Desai, 2020), and incremental Area Under the Curve (iAUC) (Petsiuk et al., 2018) are measures employed to evaluate the faithfulness of a model. Some other efforts have been made to evaluate concepts such as stability, separability, identity (Hailemariam et al., 2020; Sokol & Flach, 2020), and consistency degrees in explanations (Graziani et al., 2020).

This study addresses ground truth-based and object localization-based evaluation metrics and approaches, especially for medical image classifications. Li et al. (Li et al., 2021) recommended the Pointing Game (PG) metric to assess whether the explainer

correctly localizes the object to be recognized. With the assumption that a good saliency map is not required to cover the entire object, PG (J. Zhang et al., 2018) counts hit numbers if the highest score saliency pixel lies inside the annotated bounding box.

Likewise, Sattarzadeh et al. (Sattarzadeh et al., 2021) validated the quality of their proposed visual explanation model in comparison to other state-of-the-art attribution-based explanation models using Energy-Based Pointing Game (EBPG), mean Intersection over Union (mIoU), and Bounding Box (Bbox)(Schulz et al., 2020). Extending the idea of PG, EBPG considers all pixels in the explanation map for evaluation by measuring the fraction of its intersection with the ground truth map normalized by the explanation region (Wang et al., 2020).

In other words, it evaluates the precision and denoising capability of the explanations. Weighted Intersection over Salient Region (WIoSR) is a modified version of the EBPG metric in which saliency scores of the detected regions are used as the weights while calculating the localization score (Petsiuk et al., 2018; Zeiler & Fergus, 2014).

Schallner et al. (Schallner et al., 2020) investigate the effect of various segmentation algorithms on the LIME performance by utilizing the Jaccard Coefficient on a binary classification task for malaria blood smear images. Despite some similarities between this work and the current one, fundamental differences exist.

Our work emphasizes exploring various suitable metrics, introducing, and adapting them to find the most proper one for the defined problem, which is explainability for a more complex classification task for a multi-label multi-class dataset. Besides that, we utilize a second unseen larger dataset for replacing human intervention with a more efficient and accurate automated system. Finally, yet importantly, we show how IoU-based metrics such as Jaccard Coefficient, Bbox, and mIoU perform poorly in measuring the similarities between two bounding boxes when there is no intersection.

2.3 Enhanced Perturbation-Based XAI Models

Perturbation-based techniques have shown promise in explaining black-box machine learning models, but they have several issues that must be addressed. These include generating out-of-distribution data points, which can lead to less reliable classifier outputs and inconsistent results in LIME (Agarwal & Nguyen, 2021). Furthermore, these techniques often fail to consider useful information, such as feature correlations, when generating new samples (Shi et al., 2020).

As a result, the explanations generated by these approaches are relatively less reliable, less trustworthy, and of poor quality. This section discusses relevant works that aim to improve LIME or similar models from various aspects. These approaches encompass various strategies, including identifying more suitable random samples, utilizing additional data, generating synthetic in-distribution datasets, or completely altering the sample generation process to eliminate randomness. Indeed, it is worth noting that these approaches also address distinct issues.

Influenced by stable approximation trees in model distillation (Y. Zhou et al., 2018), Zhou et al. (Z. Zhou et al., 2021) presented S-LIME. This method aims to address the challenge of stability in LIME and similar techniques, though it does not guarantee absolute stability. S-LIME employs a hypothesis testing framework grounded in the central limit theorem to deduce the optimal number of perturbed samples, striving to offer explanations with enhanced stability compared to traditional LIME.

OptiLIME, another framework proposed by Visani et al. (Visani et al., 2022), focuses on maximizing stability and maintaining fidelity to ML model explanations. It achieves this by automatically determining the optimal kernel width value, ensuring the desired fidelity level while optimizing explanation stability.

In contrast to our integrated solution, these techniques require individual execution for each dataset or problem, increasing the workload for each case. While they improve stability, they do not provide a guarantee of a 100% stable solution. Additionally, the run-time efficiency and potential impact on the quality of explanations have yet to be thoroughly examined.

Some approaches generate datasets that preserve the distribution of the training data around the target instance, resulting in stable and accurate explanations. For example, authors (Jiang, 2022) proposed a modified version of LIME for regression models in aerodynamic tabular data. Their approach involves perturbed sample generation that emphasizes differences and ensures a uniform distribution of distances and weights. However, further research is needed to validate this approach with more extensive and diverse datasets, different models, and additional analyses.

A study (Vreš & Šikonja, 2021) proposes advanced sampling methods utilizing data generators to capture the training dataset's distribution effectively. The authors explore various data generators, including Radial Basis Function (RBF) network-based, and variational autoencoder-based approaches (Miok et al., 2019; Robnik-Šikonja, 2016). The focus is on robustness, which refers to the ability of an explanation method to identify biased classifiers in adversarial environments, mainly targeted attacks (Alvarez-Melis & Jaakkola, 2018; Slack et al., 2020). The research highlights the importance of data generators and emphasizes the need for requirements regarding access to training data.

Similarly, Saito et al. (Saito et al., 2020) introduce a method to enhance the robustness of LIME against adversarial attacks. They propose training a Conditional Tabular GAN (CTGAN) model (Xu et al., 2019), a generative adversarial network, to generate realistic synthetic data for LIME's explanation generation process. Experimental results on three real-world datasets demonstrate improved accuracy in detecting adversarial behavior compared to vanilla LIME while maintaining comparable explanation quality.

Generating synthetic data demands particular attention to detail and significant effort, involving creating artificial data points that closely resemble the distribution and characteristics of the original training data. This can be challenging for high-dimensional or complex datasets, requiring computational resources, parameter tuning, and iterative refinement. Integrating synthetic data generation into the explanation process adds complexity and computational overhead, necessitating customized approaches and parameter settings for each dataset. Careful consideration of computational resources, time, and expertise is crucial, along with optimizing implementation to mit-

igate the additional load.

On the other hand, alternative methods take a different approach by utilizing a portion of the training data instead of generating a synthetic dataset. Hall et al. (Hall et al., 2019) propose a technique that involves partitioning the dataset using k-means clustering instead of perturbing data points around the instance being explained. However, the non-deterministic nature of the default k-means implementation, which randomly selects centroids, limits the effectiveness of this approach. Similarly, Hu et al. (Hu et al., 2018) introduce a supervised tree-based method for dataset partitioning.

Building upon these ideas, Zafar et al. (Zafar & Khan, 2021) present a more practical approach called Deterministic Local Interpretable Model-Agnostic Explanations (DLIME) to enhance the interpretability of black box Machine Learning (ML) algorithms. DLIME utilizes Agglomerative Hierarchical Clustering (AHC) and K-Nearest Neighbor (KNN) techniques instead of generating synthetic samples, providing a deterministic variation of LIME. It is important to note that this approach requires access to the training data and may face challenges when applied to image data or datasets with a large number of features due to the inherent complexity involved.

Guided-LIME (Sangroya et al., 2020) represents an alternative approach that leverages a subset of the training data, aiming to refine explanations for machine learning models. This approach employs Formal Concept Analysis (FCA) to enhance the structured sampling of instances, which are then fed to SP-LIME—a sampling technique that emphasizes interpretable features with high importance (Ribeiro et al., 2016). The result is a more comprehensive coverage of the explanation space compared to using SP-LIME alone. By leveraging this structured sampling, Guided-LIME optimizes instance selection in terms of both quantity and quality. When evaluated with the calibration error metric, Guided-LIME demonstrates a closer approximation to original black-box ML models. However, there is no explicit claim about its stability, suggesting potential residual instabilities.

Some approaches use methods to select relevant samples or modify random samples to improve their suitability. Applying criteria like affinity quantification or inlier scores in the filtering step enhances the quality and effectiveness of the explanations. For instance, a solution proposed by the authors (Qiu et al., 2022) addresses the out-

of-distribution problem by incorporating a module that measures the similarity between perturbed data and the original dataset distribution. The authors enhance the robustness of the final explanations by penalizing the influence of unreliable OoD data, leveraging inlier scores and prediction results. While their approach claims compatibility with popular perturbation-based XAI algorithms such as RISE (Petsiuk et al., 2019), OCCLUSION (Zeiler & Fergus, 2014), and LIME, they acknowledge its inapplicability to LIME and identify deficiencies in localization metrics.

Rasouli et al. (Rasouli & Yu, 2019) address limitations in traditional data sampling for local explanation methods. They propose a sampling methodology based on observation-level feature importance, which generates more meaningful perturbed samples than Gaussian or training data distributions. The proposed approach captures feature interactions and creates a perturbed dataset that accurately represents the locality of a given sample. However, further testing is needed to evaluate its stability and efficiency, and its applicability is currently limited to tabular data.

ALIME (Shankaranarayana & Runje, 2019) utilizes a pre-trained autoencoder to transform randomly generated disturbance samples into lower-dimensional latent codes. Instead of relying on Euclidean distance, the similarity between disturbed samples and the input instance is assessed based on the distance between these latent codes. This approach offers an alternative way to measure similarity and capture significant patterns during the explanation process.

Some approaches aim to refine LIME's random sample generation algorithms by prioritizing feature correlations rather than the stability and quality of explanations. Modified Perturbed Sampling for LIME (MPS-LIME) (Shi et al., 2020) stands as a notable example, addressing the inherent limitations of LIME's standard sampling. MPS-LIME delves into feature correlations by converting the superpixel image into an undirected graph, treating perturbed sampling as a clique set construction challenge. Rigorous evaluations on Google's pre-trained Inception neural network (Szegedy et al., 2015) show that MPS-LIME holds an advantage over LIME in areas such as understandability, fidelity, and efficiency. However, it is crucial to note that the method has not been tested for stability, raising questions about its reliability.

As another example in this category, Bootstrap-LIME (B-LIME) (Abdullah et al.,

2023) is tailored to enhance the explainability of deep learning models for Electrocardiogram (ECG) data. Recognizing the inherent temporal relationships within ECG features, B-LIME seamlessly integrates advanced techniques such as heartbeat segmentation and the bootstrapping technique for data resampling. This combination is meticulously crafted to bolster the model's interpretability, emphasizing the temporal dynamics of features. The authors further pioneered a data-generation technique, ensuring explanations resonate with the local vicinity of predictions, and an explanatory framework reflecting the precision of the original ML model. However, a thorough evaluation concerning the approach's quality and stability remains absent from the study, leaving lingering questions about its definitive stability. Predominantly, this method is tailored for temporal signal data.

The proposed algorithm distinguishes itself from previous works by offering a purposeful, deterministic sample generation approach that addresses stability, improves localization precision in bounding boxes, and considers feature correlations. It comprehensively improves on all these factors, significantly contributing to the field of perturbed-based XAI models. Unlike other approaches, it achieves simultaneous improvements in stability, explanation qualities, and run-time efficiency. Additionally, our method does not require additional training data and is simple to implement while enhancing LIME in multiple aspects. Notably, the stability enhancement provided by our method is guaranteed. Our approach transcends the confines of specific data modalities. It can be applied not only to image data but also to any dataset where samples can be converted into a graph of correlated features, making it more versatile than solutions limited to tabular data. Furthermore, we conducted experiments on a larger dataset, utilizing more suitable metrics such as Jensen-Shannon Divergence (JS_DIV) (Samek et al., 2019) to compare bounding box similarities between ground truth data and explanations, even in cases without overlap.

CHAPTER 3

DATA SETS AND TRAINED CLASSIFIER

Our research centers on producing explanations for the outcomes of the thorax disease classifier, specifically within the realm of Chest X-ray (CXR) images. However, our exploration does not solely begin with generating explanations. It starts with the foundational steps of dataset preparation and classifier training. Robust and representative datasets, carefully curated to reflect real-world variations and complexities, play a pivotal role in ensuring the effectiveness of our explanations. A well-trained classifier, accurately and consistently discerning patterns within CXR images, serves as the cornerstone for our explanation generation process. Therefore, the achievement of an accurate classifier enhances the subsequent explanation outcomes with trustworthiness and reliability. As we dive deeper, the upcoming subsections will provide comprehensive insights into the intricacies of the datasets used and the model trained for our study.

3.1 Datasets

This study uses two datasets: CheXpert (Irvin et al., 2019) and VinDr-CXR (Nguyen et al., 2020). These datasets were thoughtfully selected as the most suitable and complementary combination. CheXpert offers sufficient samples per class and reliable global annotations, making it well-suited for training and classification using deep neural networks. Conversely, vinDr-CXR provides precise regional annotations for classes that align with CheXpert, making it a valuable resource for evaluating explanations. CheXpert ensures adequate training samples, while vinDr-CXR provides essential and precise local annotations for evaluation.

The CheXpert dataset comprises 224,316 chest radiographs from 65,240 patients, accompanied by chest-related global annotations. These annotations were extracted using an automated, rule-based labeler that parsed categories from the free-text radiology reports authored by radiologists. To ensure reliability, additional rules were applied to capture negations and uncertainties mentioned in the reports. Each sample within the CheXpert dataset includes the image ID, patient's sex and age, X-ray properties (frontal/lateral and AP/PA), and a hot vector representing 14 distinct classes. Within this vector, 0 labels indicate negative findings, +1 labels represent positive findings, and -1 labels indicate uncertain findings.

The 14 classes encompass Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, Pneumonia, Pneumothorax, Enlarged Cardiomediastinum, Lung Lesion, Lung Opacity, Pleural Other, Fracture, Support Devices, and No Finding. Alongside the training set, a separate validation set of 234 samples is provided, which underwent label approval by radiologists. Consequently, uncertain labels were eliminated from the validation data.

The VinDr-CXR dataset includes 15,000 images within its training set, with labels manually defined by a team of radiologists. Figure 3.1 presents sample images from this dataset. The dataset includes 22 local labels represented by bounding boxes surrounding abnormalities, and six global labels denoting detected diseases. The images are provided in the Digital Imaging and Communications in Medicine (DICOM) format, accompanied by relevant DICOM tags such as patient sex and age. Primarily consisting of Posterior Anterior (PA)-view chest X-rays, the dataset excludes non-relative X-rays from other body parts.

The local labels cover various abnormalities such as Aortic Enlargement, Atelectasis, Cardiomegaly, Calcification, Clavicle Fracture, Consolidation, Edema, Emphysema, Enlarged PA, Interstitial lung disease, Infiltration, Lung Cavity, Lung Cyst, Lung Opacity, Mediastinal Shift, Nodule/Mass, Pulmonary Fibrosis, Pneumothorax, Pleural Thickening, Pleural Effusion, Rib Fracture, and Other Lesion. Additionally, the global labels encompass Lung Tumor, Pneumonia, Tuberculosis, Other Diseases, Chronic Obstructive Pulmonary Disease, and No Finding.

Figure 3.2 presents the distribution of samples for each label category within each

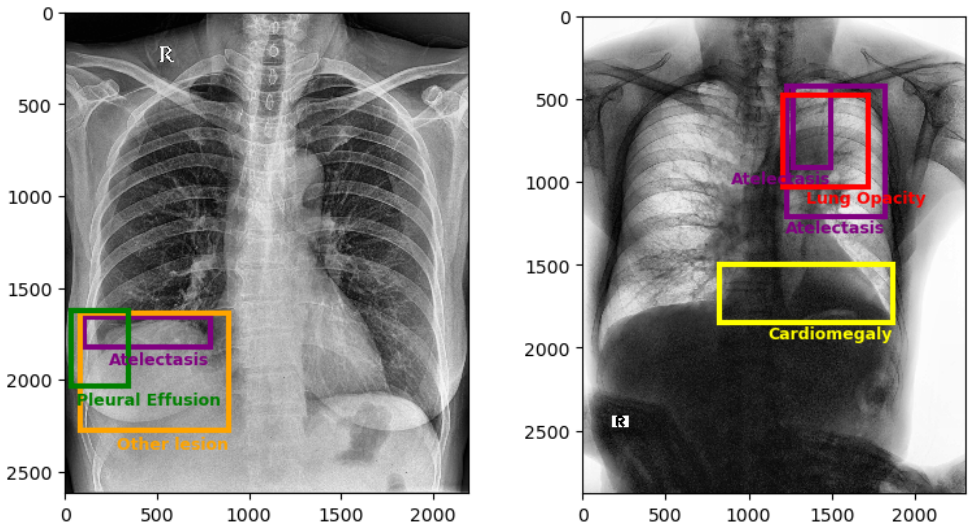


Figure 3.1: Samples of the VinDr-CXR dataset with precise annotated labels

dataset. The CheXpert dataset offers a sufficient number of samples per class, making it well-suited for training and classifying tasks using deep neural networks. Conversely, VinDr-CXR includes precise local annotations for several classes that overlap with the classes found in CheXpert, making it an appropriate choice for evaluating the explanations generated.

3.2 Training The Classifier

The accuracy of the classifier's results greatly influences the explanations' quality. Therefore, our trained model must identify labels accurately. Explanations generated by a poorly performing classifier are often uninformative and lack value. We apply data preprocessing techniques to the CheXpert dataset to ensure data quality. We exclude samples that contain uncertain labels among their classes. Furthermore, we only select PA and Lateral images for both sets. The test set is not yet available because the CheXpert dataset is an ongoing competition. Hence, we conduct the training phase using 85% of the samples from the filtered training set (83,781) while reserving the remaining 15% for validation and testing purposes. Within the remaining 15%, 10% is allocated for validation, and the remaining 5% is combined with the filtered original validation test to form the test set.

Subsequently, we train several well-known CNN models including InceptionResNetV2 (Längkvist et al., 2014), InceptionV3 (Szegedy et al., 2016), DenseNet121 (Huang et al., 2017), ResNet50 (He et al., 2016), MobileNetV2 (Sandler et al., 2018), and VGG16 (Simonyan & Zisserman, 2015). All models are trained on an Nvidia GeForce RTX 3080 GPU, employing a batch size of 16. We utilize the Adam optimizer with a learning rate 0.0001, initializing the models with weights pre-trained on the ImageNet dataset.

In order to determine the most suitable classifier model, we assessed model accuracy using the Area Under the ROC Curve (AUC). Based on the overall performance presented in Table 3.1, InceptionV3 emerged as the top choice with a mean AUC value of 89%. This performance is consistent with findings from literature, specifically (Ke et al., 2021), where InceptionV3 recorded a leading AUC of 0.862 for the CheXpert dataset, surpassing other well-known models such as DenseNet121 (0.859), MobileNetV2 (0.854), and ResNet50 (0.859). As a result, we opted to continue the study with InceptionV3 as our selected classifier model.

Table 3.1: Comparing different trained classifiers by their AUC percentages for each class

Class Name	Inception ResNetv2	Mobile NetV2	VGG16	ResNet50	Dense Net121	Inception V3
Cardiomegaly	79	84	79	79	82	82
Pleural Effusion	90	88	91	91	91	92
Lung Opacity	90	90	88	89	89	94
Consolidation	90	90	91	91	90	91
Atelectasis	75	72	75	76	78	77
Pneumothorax	98	85	81	94	92	95
No Finding	94	94	95	94	93	92

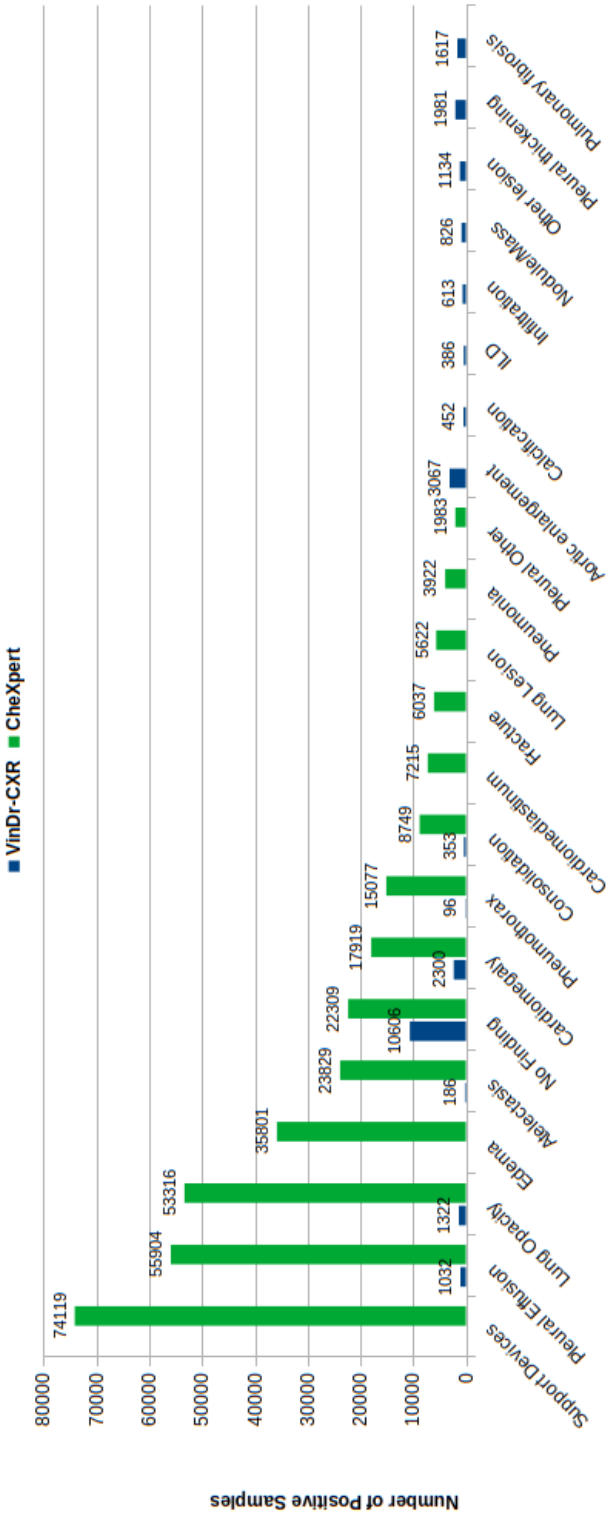


Figure 3.2: Distribution of samples within classes for both CheXpert and VinDrCXR datasets

CHAPTER 4

QUANTITATIVE VISUAL EXPLANATION EVALUATION

The primary objective of this chapter is to present a practical solution for evaluating generated explanations. We address this issue in a step-by-step manner. Firstly, we establish essential criteria that define what we expect from an acceptable explanation when the classifier correctly detects the labels. Secondly, we introduce several metrics suitable for comparing local annotations made by experts with the generated explanations produced by LIME. Lastly, we discuss the drawbacks of each metric in relation to the defined criteria, using their respective formulas, to determine the most suitable approach. Detailed information is provided in the following subsections.

4.1 Desired Criteria

To effectively assess the quality and precision of generated explanations, particularly when the classifier has accurately predicted the label(s), we have established a set of fundamental principles. These principles serve as the foundation for our objective evaluation:

- **Overlap with Ground Truth:** The generated explanations should largely coincide with the ground truth annotations, which means striving for a high true positive rate and a low false positive rate.
- **Multi-label Consideration:** In multi-label scenarios, the explanations should encapsulate multiple regions, akin to what the ground truth annotations illustrate, thereby minimizing false negatives.
- **Minimized Discrepancy:** If an explanation and the ground truth regions do not

overlap, they should still be proximate; the distance between them should be as minimal as possible.

Rooted in these principles, we introduce our tailored criteria—a unique contribution that paves the way for choosing more discerning metrics. These criteria not only reflect the considerations mentioned earlier but also enhance our ability to objectively evaluate and compare different explanation mechanisms:

- ***Informativeness:*** This criterion underscores the necessity for the metric to yield a non-zero value in every conceivable scenario. It is crucial for the metric to encapsulate the relationship between the explanations and ground truth annotations, unaffected by the configuration—whether it's a direct alignment, one region enveloping another, or any diverse arrangement. The metric must consistently yield a discernible value that comprehensively conveys the comparative quality of the explanation relative to the ground truth, ensuring no scenario results in an undefined or zero value.

- ***Localization:*** Centered on spatial considerations, this criterion evaluates the closeness of generated explanations to the ground truth. Explanations in close proximity to the actual ground truth, regardless of overlap, offer valuable contextual clues. An effective metric should acknowledge and appropriately value these near-miss explanations over more distant ones.

- ***Coverage:*** This criterion prioritizes explanations that mirror the ground truth. Explanations closely aligned with the ground truth indicate accurate representation, while those deviating should be penalized, signaling their reduced accuracy.

- ***Multi-target Capturing:*** Recognizing the complexities of multi-label scenarios, this criterion ensures that each target region within a sample is independently assessed. This granular approach guarantees that every annotated area is evaluated appropriately, irrespective of the number of target regions in a sample.

- ***Proportionality:*** Metrics should directly correlate the degree of difference between explanations and the ground truth. A larger disparity should result in

a commensurately higher metric value, guaranteeing that the evaluation accurately mirrors the given situation.

In summary, the criteria we have presented are designed to rigorously and comprehensively evaluate the efficacy of generated explanations. By meticulously addressing every facet of explanation quality, from spatial alignment to the nuances of multi-label scenarios, our criteria ensure a holistic assessment. This novel contribution not only bolsters the accuracy of evaluation metrics but also stands as a testament to the evolving sophistication and rigor in the domain of explainable AI.

4.2 Localization-based Evaluation Metrics

We explore several metrics to evaluate the quality of the generated explanations. By comparing the localization of LIME's explanations to the ground truth annotations of the VinDr-CXR dataset, we can determine the best configuration for the LIME method or even compare it with other explanation methods in terms of accuracy. Both ground truth masks and LIME explanation masks are converted to one-dimensional lists to facilitate pixel-wise measurements, especially for dealing with multi-bounding boxes for both masks. Additionally, to make the maps comparable and ensure consistency, we convert the pixel values to binary values. In this context, we denote the ground truth mask as GT, the achieved explanation map as EX, and the total number of pixels as P. The evaluation metrics used in the study are as follows:

Mean Absolute Error (MAE): A common metric for error calculation (Mohseni et al., 2021). It measures the distance between the model saliency mask and the ground truth mask. Formally, it can be defined as:

$$MAE(GT, EX) = \frac{1}{P} \sum_{i=1}^{P} GT(i) - EX(i) \tag{4.1}$$

Intersection over Union (IOU) or Jaccard Index: A popular metric for object detection tasks. It indicates the similarity of two sets by dividing the intersection of the

estimated area and ground truth area by their union. An IOU of zero indicates no intersection, and one indicates that two areas are entirely overlapped. In our case, IOU is calculated pixel-wise by taking the intersection of the GT map and the EX map, divided by their union:

$$IOU(GT, EX) = \frac{\sum_{i=1}^{P} GT(i) \cap EX(i)}{\sum_{i=1}^{P} GT(i) \cup EX(i)} \tag{4.2}$$

Jensen-Shannon Divergence (JS_DIV): Another way to quantify the difference or similarity between two probability distributions (Samek et al., 2019). It utilizes the Kullback-Leibler Divergence (KL_DIV) measurement, which quantifies how much one probability distribution differs from another probability distribution (Bylinskii et al., 2019). We convert GT and EX maps to probability distributions of the ground-truth map and explanation map, respectively, so that the sum of all values in each map equals one. Then, given two probability distributions, KL_DIV is calculated as:

$$KL_DIV(EX_{pd}, GT_{pd}) = \sum_{i=1}^{P} EX_{pd}(i) log(\frac{EX_{pd}(i) + \varepsilon}{GT_{pd}(i) + \varepsilon}) \tag{4.3}$$

KL_DIV is a non-symmetrical and non-linear measure that varies from zero to infinity. If the score value between EX_{pd} and GT_{pd} is lower, then the EX_{pd} has a better approximation of the ground truth GT_{pd}. In other words, the intuition for this score is that when the probability for an event from EX_{pd} is large, but the probability for the same event in GT_{pd} is small, there is a considerable divergence. Reversely, when the probability of an event from EX_{pd} is small and the probability of GT_{pd} is large, there is also a large divergence, but not as significant as the first case.

JS_DIV uses the KL_DIV to calculate a normalized score that is symmetrical as follows

$$JS_DIV(EX_{pd}, GT_{pd}) = \frac{KL(EX_{pd}, M) + KL(GT_{pd}, M)}{2} \tag{4.4}$$

where M is the average of two probability distributions. JS_DIV provides a smoothed and normalized version of KL_DIV, with scores between zero (identical) and one

(maximally different) when using the base-2 logarithm. KL_DIV is a score rather than a metric; hence, it can be substituted with its symmetric normalized version, JS_DIV.

Pixel Wise Precision (PWP) or Intersection over Salient Region (IoSR): This metric calculates the ratio of the intersection between the explanation mask and the ground truth mask over the explanation map (Li et al., 2021). PWP is defined as:

$$PWP(GT, EX) = \frac{\sum_{i=1}^{P} GT(i) \cap EX(i)}{\sum_{i=1}^{P} EX(i)} \tag{4.5}$$

The formula has the same structure as the popular precision metric, where the intersection of the ground truth mask with the explanation mask indicates true positives. The whole area of the explanation mask represents the combination of the true positives and false positives.

Pixel Wise Recall (PWR) or Intersection over Ground truth Saliency (IoGS): This metric is inspired by the recall metric and calculates the ratio of the explanation mask and the ground truth mask intersection over the ground truth area. PWR is calculated as:

$$PWR(GT, EX) = \frac{\sum_{i=1}^{P} GT(i) \cap EX(i)}{\sum_{i=1}^{P} GT(i)} \tag{4.6}$$

The formula has the same idea as the recall metric, where the intersection of the ground truth mask with the explanation mask indicates true positives, and the whole area of the ground truth mask combines true positives and false negatives.

Pixel Wise Accuracy (PWA): Following the same idea as PWP and PWR, PWA is defined as the percentage of correctly detected pixels. In other words, PWA is the division of the sum of true positives and true negatives (XNOR of GT and EX) to the sum of all true positives, true negatives, false positives, and false negatives (the whole image):

$$PWA(GT, EX, Image) = \frac{\sum_{i=1}^{P} GT(i) \odot EX(i)}{\sum_{i=1}^{P} Image(i)} \tag{4.7}$$

To the best of our knowledge, JS_DIV, PWR, and PWA are employed for the first time in this domain. Additionally, except for MAE and IOU, the other mentioned metrics are applied for the first time to evaluate explanations generated by LIME.

4.3 Evaluation Metric Selection

In this section, we conclude our study by examining the behavior of each metric in terms of the defined criteria in Section 4.1 and the metrics presented in Section 4.2. Our goal is to limit the proposed metric set to a minimum, including only the most suitable ones.

Intersection-based metrics such as IOU, PWP, and PWR (which have $\sum_{i=1}^{P} GT(i) \cap EX(i)$ in the numerator) fail to calculate the distance between two entirely separated regions. In cases where there is a complete separation between the explanation and ground-truth bounding boxes, IOU, PWP, and PWR become zero. Therefore, intersection-based metrics cannot guarantee measuring *Informativeness* and *Localization* in some scenarios. Additionally, MAE and PWA consider false positives and false negatives regardless of their position. Consequently, they also fail to evaluate more complex cases accurately in terms of *Localization*.

Regarding *Coverage*, PWP and PWR fail to evaluate *Coverage* effectively for some cases due to considering only the explanation region ($\sum_{i=1}^{P} EX(i)$) or the ground-truth region ($\sum_{i=1}^{P} GT(i)$) in their denominators. PWR struggles to evaluate *Coverage* while trading off between higher true positive and lower false positive values. Moreover, PWR reaches its highest value (one) when the explanation region surrounds the ground-truth bounding box. In the opposite situation, PWP fails to evaluate *Coverage* when the explanation masks are entirely inside the ground-truth annotations.

In this study, we propose an explainable classifier for multi-label data. Due to pixel-wise calculation, none of the mentioned metrics violate the *Multi-target Capturing*

criterion if more than one region is marked as an explanation region. By considering every pixel value in the calculations, we ensure that multiple marked areas of explanation can simultaneously be compared to multiple annotated ground-truth regions.

MAE and PWA are examples of violating the *Proportionality* condition by distributing the difference to the total number of pixels. Therefore, while having explanations with significant differences for a single sample, there are trivial changes in their values, even if notable differences between those cases exist. Intersection-based metrics (IOU, PWP, and PWR) violate the *Proportionality* due to the lack of measurement when no overlap exists between explanations and ground-truth annotations.

Table 4.1 summarizes the essential parameters for each evaluation metric applicable to the current problem. Based on this analysis, we can confidently state that JS_DIV is the only metric that satisfies all the necessary conditions. The results and discussions related to these metrics and the segmentation algorithms used are provided in the following section.

Table 4.1: Summary of the essential parameters for each evaluation metric

	JS_DIV	IOU	MAE	PWP	PWR	PWA
Informativeness	✓	×	✓	×	×	✓
Localization	✓	×	×	×	×	×
Coverage	✓	✓	✓	×	×	✓
Multi-target Capturing	✓	✓	✓	✓	✓	✓
Proportionality	✓	×	×	×	×	×

4.4 Experimental Results and Discussions

We divide our experimental results into two subsections. In the initial subsection, we visually evaluate the performance of superpixel segmentation algorithms (Quickshift (Vedaldi & Soatto, 2008), Felzenszwalb (Felzenszwalb & Huttenlocher, 2004), and SLIC (Achanta et al., 2012)) against their optimized counterparts. Additionally, we analyze and illustrate the quality of the metrics presented in the preceding section,

referencing the essential criteria defined earlier and examining them through various examples. In the subsequent subsection, we present assessment results based on the metric chosen earlier. This encompasses evaluations for different segmentation settings employed by LIME and a comparative analysis between the optimal LIME setting and another XAI technique, GradCAM. As a result, we quantitatively identify the best segmentation algorithm among multiple LIME variants for the top three predicted labels and compare it with GradCAM's visual explanations in terms of localization precision.

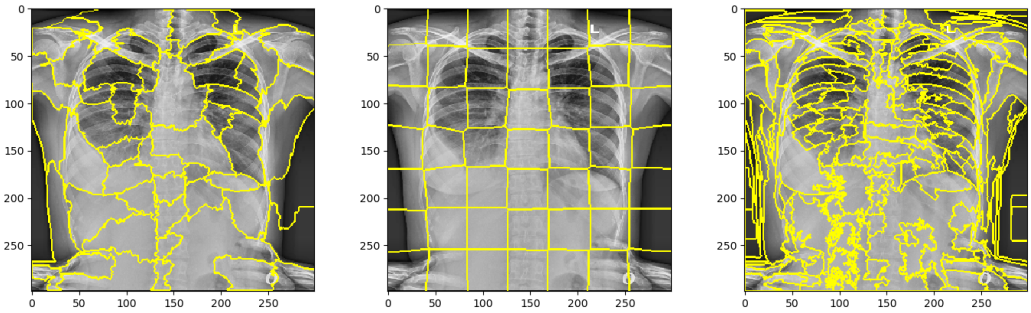
4.4.1 Qualitative Analysis

4.4.1.1 Segmentation Algorithms Exploration

Before the performance evaluation of various superpixel segmentation algorithms formally, it could be practical to display their behavior within some case studies. For each algorithm, two versions are used: the default and optimized configurations. Thus, apart from LIME's default segmentation (Quickshift), we use Optimized Quickshift, Felzenszwalb, Optimized Felzenszwalb, Simple Linear Iterative Clustering (SLIC), and Optimized SLIC.

Felzenszwalb is an efficient graph-based algorithm that utilizes the minimum spanning tree. SLIC, on the other hand, is a cluster-based superpixel algorithm that applies k-means clustering in a five-dimensional space of color information. Quickshift, the default algorithm in LIME, generates superpixels using an approximation of kernelized mean-shift through the mode-seeking method. To visually demonstrate these algorithms, Figure 4.1 shows the output of each algorithm on a random sample from the VinDr-CXR dataset.

We examine segmentation algorithms manually to find their optimized version. Hence, we perform a grid search on each parameter's range of values for each segmentation method on 50 randomly selected VinDr-CXR data samples. Then, based on how detailed segmentation each demonstrates on X-ray images, we decide which parameter set is the best. The tuned parameters used for each segmentation approach are presented in Table 4.2.



(a) Quickshift (b) SLIC (c) Felzenszwalb

Figure 4.1: Exploring Various Segmentation Algorithms: Given superpixels to the LIME generated by different segmentation algorithms

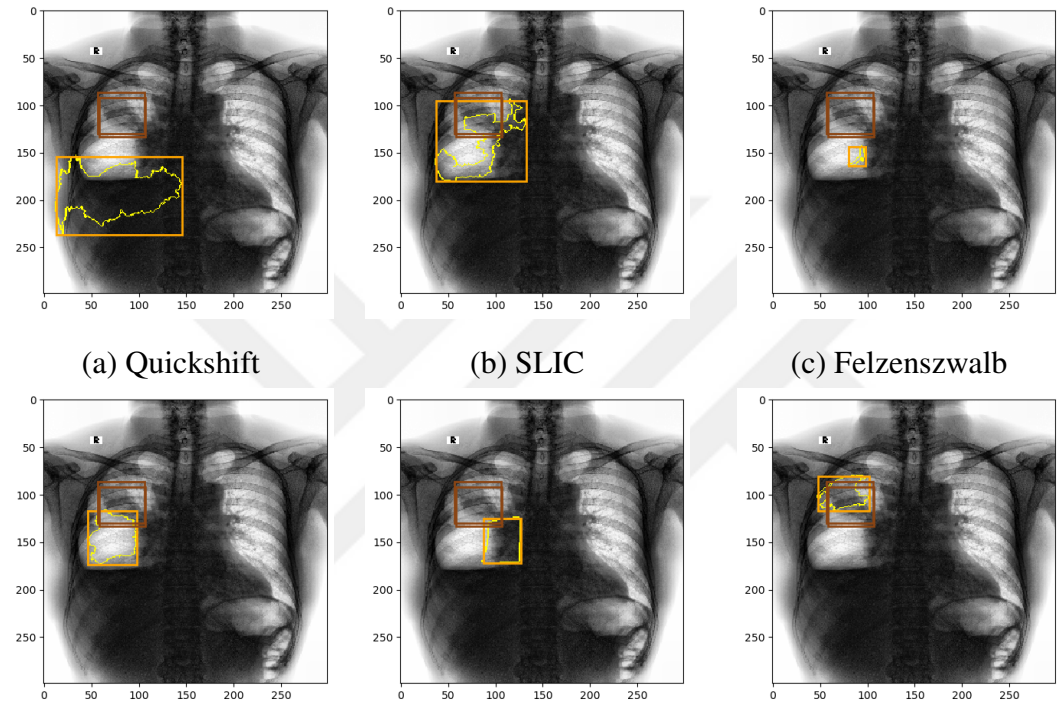
Table 4.2: Segmentation algorithms parameters

Segmentation Algorithms	Parameters	Value
Optimized Felzenszwalb	scale	600
	min_size	200
	sigma	0.2
Optimized SLIC	n_segments	50
	compactness	80
	sigma	20
Optimized Quickshift	kernel_size	4
	max_dist	100
	ratio	0.6
	sigma	0.3

Since our primary focus is on the effect of the segmentation performance on the explanations, except for superpixel algorithms, we use the default values for other LIME parameters. The LIME default value for sample numbers for the perturbation is 1000. However, for more robust results, we use 3000 samples. Moreover, we compute the LIME results for 1, 2, and 4 features and report the average.

In Fig. 4.2, the brown rectangles show the radiologists' annotations for the detected lung consolidation condition, and the orange rectangles show the explanations gen-

erated by LIME. The first row displays explanations generated by the segmentation algorithm with default parameter settings of each superpixel algorithm. And the second row displays explanations generated for the same segmentation algorithms with the optimized parameter values. The effect of optimization and parameter tuning on each segmentation algorithm can be noticed by comparing each figure at the top with its corresponding one below. In addition, better coverage of Optimized SLIC and Optimized Felzenszwalb in comparison with the Optimized Quickshift is exemplified.



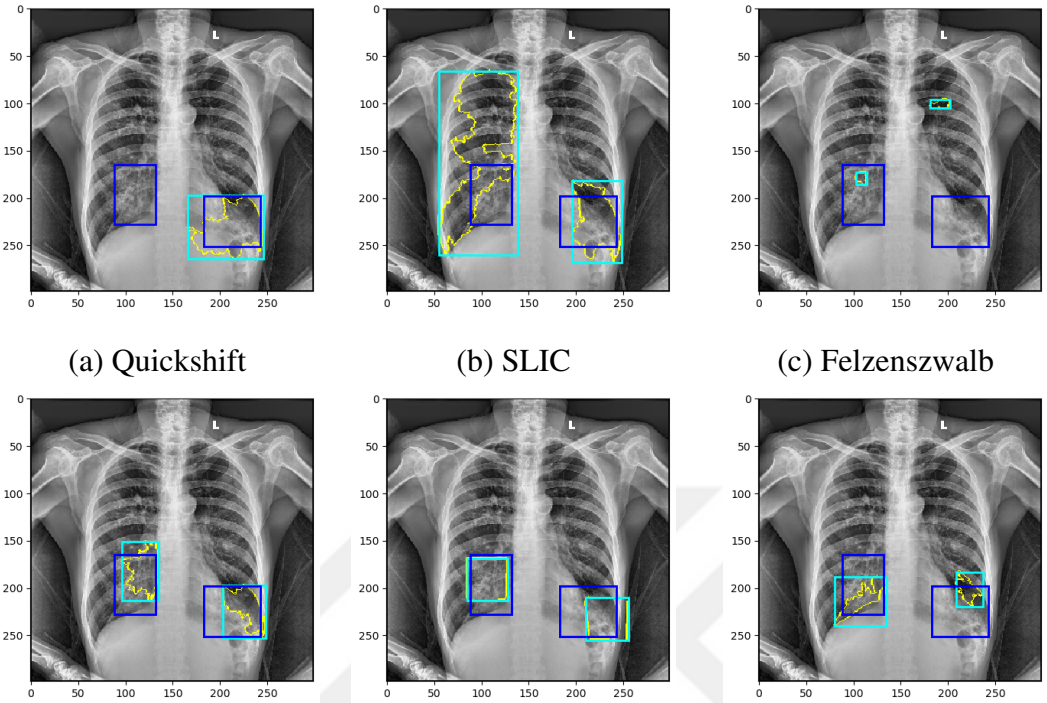
(a) Quickshift (b) SLIC (c) Felzenszwalb

(d) Optimized Quickshift (e) Optimized SLIC (f) Optimized Felzenszwalb

Figure 4.2: The first row displays explanations generated by the segmentation algorithm with default parameters, and the second row displays explanations generated for the same segmentation algorithms with the optimized parameter values. The brown (darker) rectangles show the radiologists' annotations for detected lung consolidation, and the orange (lighter) rectangles show the border of explanations (yellow-bordered area) generated by LIME.

Similarly, Fig. 4.3 illustrates another case study with multiple annotations. The dark blue rectangles show the radiologists' annotations for the detected Lung Opacity condition, and the light blue rectangles show the explanations generated by LIME. The

effect of parameter tuning for superpixel algorithms on the accuracy and efficiency of segmentations can be clearly observed in Fig. 4.3 (even in multi-label cases).



(a) Quickshift (b) SLIC (c) Felzenszwalb

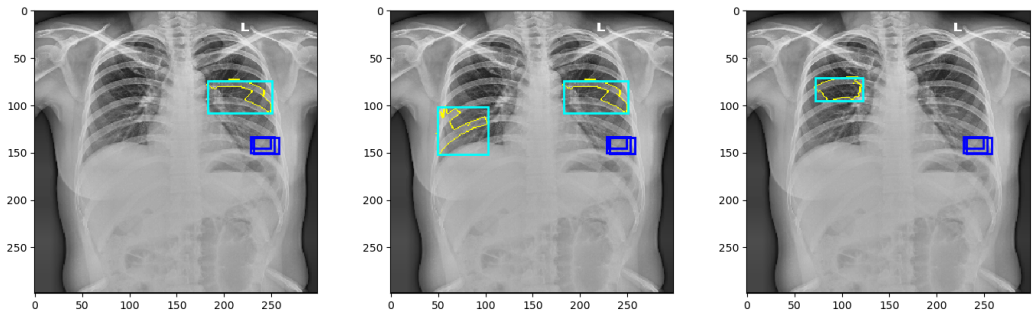
(d) Optimized Quickshift (e) Optimized SLIC (f) Optimized Felzenszwalb

Figure 4.3: The first row displays explanations generated by the segmentation algorithm with default parameters, and the second row displays explanations generated for the same segmentation algorithms with the optimized parameter values. The dark blue rectangles show the radiologists' annotations for the detected Lung Opacity condition, and the light blue rectangles show the border of explanations (yellow-bordered area) generated by LIME.

4.4.1.2 Evaluation Metric Exploration

Regardless of the segmentation algorithm outcomes and their role in the explanation quality and accuracy, in this part, the behavior of the metrics introduced in the previous section (according to the defined criteria in previous chapter) is visually demonstrated and investigated through samples shown in Fig. 4.4, Fig. 4.5 with the results presented in Table 4.3, and Table 4.4 respectively. In both figures, the darker rectangles show the radiologists' annotations for the detected condition, and the lighter

ones show the border of explanations (yellow-bordered area) generated by LIME. Each row shows a different case with three distinct segmentation settings. The left-most examples demonstrate the more desirable result, and the right-most ones show the less desirable explanations.



(a) (b) (c)

Figure 4.4: Examples that violate *Informativeness* and *Localization* are shown for a single case with different segmentation algorithms. The dark blue rectangles show the radiologists' annotations for the detected Lung Opacity condition, and the light blue rectangles show the border of explanations (yellow-bordered area) generated by LIME.

Informativeness: As mentioned, intersection-based metrics such as IOU, PWP, and PWR fail to calculate the distance between two entirely separated regions. In the complete separation between explanation and ground-truth bounding boxes, IOU, PWP, and PWR become zero, as shown in Fig. 4.4 and Table 4.3. Therefore intersection-based calculations lack *Informativeness* in some cases.

Localization: By comparing Fig. 4.4(a), Fig. 4.4(b), and Fig. 4.4(c), the meaning of better *Localization* measurement can be inferred. In case of no overlap between explanations and the ground-truth regions, the proper metric should generate higher values for closer regions. Although explanations in Fig. 4.4(b) contain larger false positive areas, they are still more helpful than the one in Fig. 4.4(c) since at least one of the two regions is closer to the ground-truth annotations (Table 4.3). Intersection-based metrics such as IOU, PWP, and PWR fail to evaluate such situations. In ad-

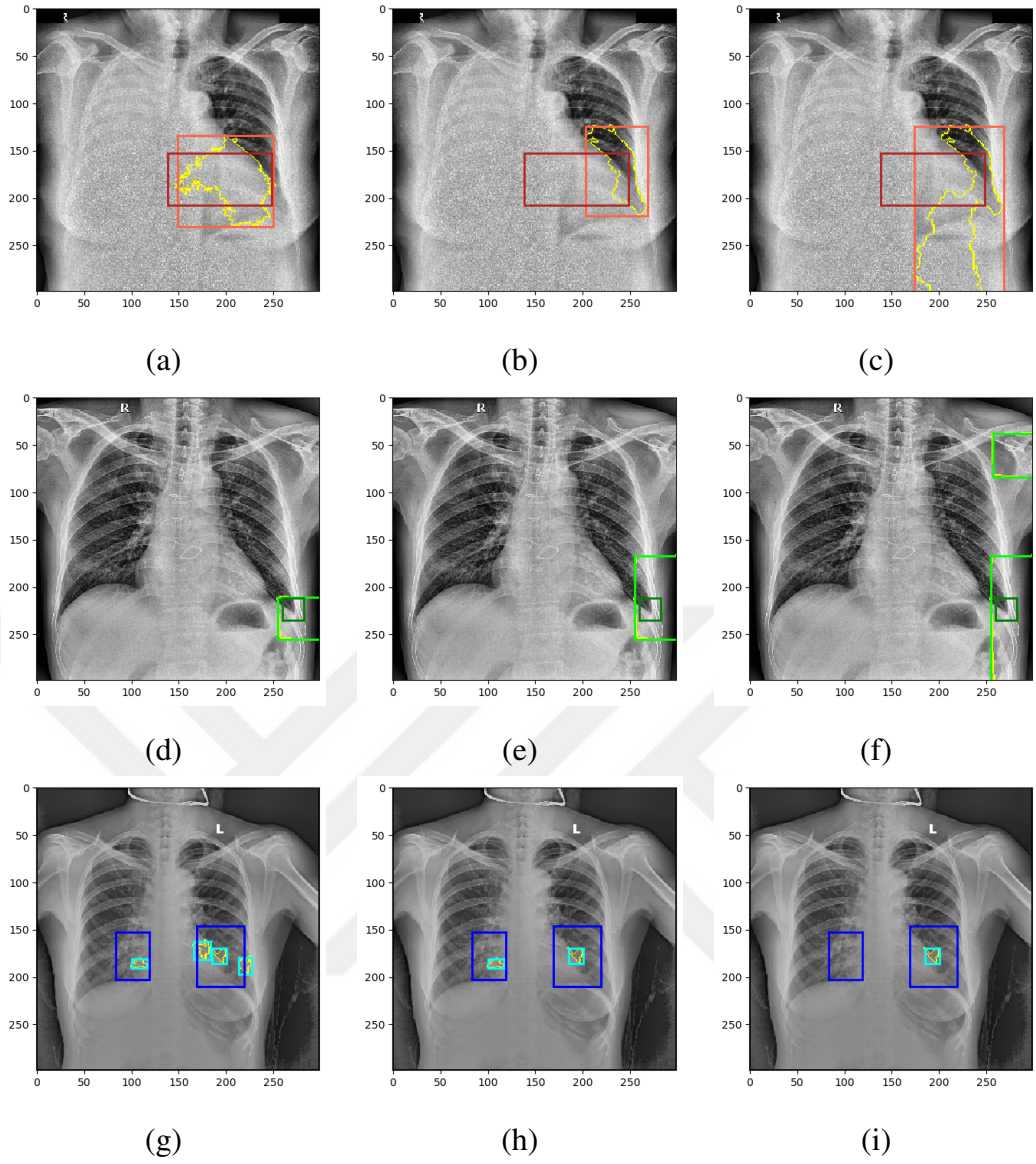


Figure 4.5: Examples that violate *Coverage* and *Proportionality* are shown for a single case with different segmentation algorithms. The darker orange, green, and blue rectangles show the radiologists' annotations for the detected Cardiomegaly, Pleural Effusion, and Lung Opacity conditions. The lighter orange, green, and blue rectangles show the border of explanations (yellow-bordered area) generated by LIME. Each row shows a different case with three distinct segmentation settings. The left-most examples demonstrate the more desirable result, and the right-most ones show the less desirable explanations.

dition, MAE and PWA consider false positives and false negatives regardless of the position. Consequently, they also fail to evaluate more complicated cases similar to Fig. 4.4(b) and Fig. 4.4(c) correctly (Table 4.3).

Table 4.3: Metric results for the examples violate *Informativeness* and *Localization* shown in Fig. 4.4

Sub Figure	JS_DIV ↓	IOU ↑	MAE ↓	PWP ↑	PWR ↑	PWA ↑
a	0.693	0	0.013	0	0	0.987
b	0.734	0	0.023	0	0	0.977
c	0.806	0	0.015	0	0	0.985

Table 4.4: Metric results for the examples violate *Coverage* and *Proportionality* shown in Fig. 4.5

Sub Figure	JS_DIV ↓	IOU ↑	MAE ↓	PWP ↑	PWR ↑	PWA ↑
a	0.23	0.496	0.04	0.781	0.576	0.96
b	0.56	0.099	0.075	0.341	0.122	0.925
c	0.576	0.092	0.123	0.155	0.185	0.877
d	0.358	0.278	0.015	0.278	1	0.985
e	0.482	0.139	0.035	0.139	1	0.965
f	0.565	0.069	0.076	0.069	1	0.924
g	0.575	0.066	0.053	0.792	0.068	0.947
h	0.616	0.036	0.054	1	0.036	0.946
i	0.674	0.019	0.055	1	0.019	0.945

Coverage: the proper metric should give higher values for cases with the maximum intersection between explanations and the ground-truth masks and the minimum separation of these two regions. Poor performance of PWP and PWR metrics in measuring the *Coverage* is noticeably observable in several cases in Fig. 4.5. Each row of Fig. 4.5 shows a different case with three distinct segmentation settings. The left-most examples demonstrate the more desirable result, and the right-most ones show

the less desirable explanations. PWR's deficiency regarding the best *Coverage* is shown in Fig. 4.5(a-c) and Fig. 4.5(d-f). PWR fails to evaluate better *Coverage* while trading off between higher true positive and lower false positive (Fig. 4.5(b-c) and Table 4.4). Although Fig. 4.5(b) shows a smaller true positive area, the explanation is more acceptable rather than the one in Fig. 4.5(c) since it marks a more specific region that is closer to the expert's annotation. Moreover, PWR is equal to its highest value (one) when the explanation region surrounds the ground-truth bounding box. For example, different explanations generated by different segmentations in Fig. 4.5(d), Fig. 4.5(e), and Fig. 4.5(f) are entirely similar and even best in terms of PWR value. In the opposite situation, PWP fails to evaluate the *Coverage* while the explanation masks are entirely inside the ground-truth annotations. As shown in the third row of Fig. 4.5, the best explanation shown in Fig. 4.5(g) has a lower value of PWP, and the worst one shown in Fig. 4.5(i) has the highest value of PWP as one (Table 4.4).

Multi-target Capturing: in this study, we intend to offer an explainable classifier for multi-label data. Due to pixel-wise calculation, none of the mentioned metrics violate *Multi-target Capturing* if more than one region is marked as an explanation region.

Proportionality: There should be a positive correlation between ground truth and explanation mask distances and the metric values reflecting the situation proportionally. MAE and PWA are examples of violating this condition by distributing the difference to the total number of pixels. Therefore, while having significant differences for a single sample, there are trivial changes in their values, even if notable differences between those cases exist. As an example, although there are noteworthy differences between Fig. 4.5(g) (the most desirable explanation) and Fig. 4.5(i) (almost the most unfavorable one), there are negligible differences between values reported by either MAE or PWA for both cases (Table 4.4). In case of no overlap between explanation and ground-truth annotations, it can be inferred that intersection-based metrics (IOU, PWP, and PWR) fail, reflecting the *Proportionality* due to generating zero as their final value.

4.4.2 Quantitative Analysis

To examine the effect of various segmentation algorithms on the LIME explanations, we perform extensive analysis on a random selection of 2000 images belonging to the unseen VinDr-CXR dataset. The analysis result is presented in Table 4.5. LIME uses Quickshift by default as the segmentation algorithm. Apart from this algorithm, we apply SLIC and Felzenszwalb segmentation algorithms to compare the effect of different segmentation settings on the explanation qualities. We consider two versions for each segmentation algorithm, default and optimized (Table 4.2). For each variation of LIME, we calculate the JS_DIV value for the top three predicted classes with the probability of more than 25%. We then apply the same calculations for feature settings 1, 2, and 4. The number of patches (superpixels) given as output by the LIME algorithm is called a feature and should be set manually. We take the average of all metric values for each segmentation algorithm to measure the overall performance for features 1, 2, and 4. Then for each prediction, we select the finest segmentation setting for LIME as shown in Table 4.5.

Based on the JS_DIV (selected as the most proper metric previously) value, LIME using optimized SLIC segmentation has the best result for all predictions. Moreover, optimized Felzenszwalb and default Felzenszwalb show the best second and third results, respectively. The JS_DIV value for each metric decreased from the first prediction to the third one. It is expected since the second predicted class has a lower probability value (the classifier is less confident), and the third one has the lowest among all three.

In another significant experiment presented in Table 4.5, we compare LIME, with its various segmentation configurations, to GradCAM. GradCAM, by leveraging gradient information from within models having convolutional layers, typically offers more precise localization maps. However, with strategic optimization of LIME, particularly when paired with the SLIC and Felzenszwalb segmentation algorithms, we observe superior explanatory results than those generated by GradCAM. The reason behind this improvement lies in the intrinsic design of these segmentation algorithms. Both SLIC and Felzenszwalb generate coherent and homogeneous segments that capture meaningful structures within images. This capability assists LIME in crafting expla-

Table 4.5: Comparison of the LIME method with various segmentation algorithms for the first three predicted classes

	Segmentation Algorithm	JS_DIV
1st Prediction	LIME + Optimized SLIC	0.558
	LIME + Default SLIC	0.576
	LIME + Optimized Felzenszwalb	0.559
	LIME + Default Felzenszwalb	0.565
	LIME + Optimized Quickshift	0.607
	LIME + Default Quickshift	0.619
	GradCAM	0.567
2nd Prediction	LIME + Optimized SLIC	0.562
	LIME + Default SLIC	0.582
	LIME + Optimized Felzenszwalb	0.565
	LIME + Default Felzenszwalb	0.568
	LIME + Optimized Quickshift	0.615
	LIME + Default Quickshift	0.629
	GradCAM	0.574
3rd Prediction	LIME + Optimized SLIC	0.603
	LIME + Default SLIC	0.612
	LIME + Optimized Felzenszwalb	0.607
	LIME + Default Felzenszwalb	0.608
	LIME + Optimized Quickshift	0.624
	LIME + Default Quickshift	0.635
	GradCAM	0.611

nations that not only localize features accurately but also resonate better with human interpretability, making them more comprehensive than GradCAM's outputs.

In this chapter, we have outlined specific criteria to pinpoint an optimal metric for comparing explanation annotations to their intended ground truth. While examining various existing metrics, we bring some metrics into this domain for the first time. Each metric is rigorously assessed based on its formula, behavior, and the cri-

teria we set forth. From this selection, our chosen metric, originally prominent in the computer vision domain and novel to the XAI domain, distinguishes itself with its precision and reliability. Qualitative analyses underscore its accuracy, while our quantitative research further highlights its utility in refining models, specifically like LIME, compared to higher-performing models such as GradCAM. As we progress, this metric paves the way for the next phase, where we use it to assess the quality of explanations generated by our proposed algorithm and contrast it with LIME.

CHAPTER 5

PURPOSIVE PERTURBATION

In this section, we outline the random sampling algorithm of LIME, followed by a detailed overview of our MindfulLIME algorithm, which employs a purposive sample generation-based approach. Our enhanced perturbation-based XAI algorithm framework is crafted to function with any data that can be transformed into a graph of correlated features. Specifically, in this paper, we underscore its application and validation employing a dataset of X-ray images, wherein each segment of an image is treated as a feature and represented as a vertex in the graph.

5.1 LIME

LIME, which stands for Local Interpretable Model-agnostic Explanations, is an innovative method designed to explain the predictions of various types of machine learning models in an interpretable and faithful manner. Before diving into the LIME framework, it is essential to understand the initial processing of the input image. Image classification tasks typically involve working with tensors with three color channels per pixel. However, interpreting such tensors can be challenging and computationally intensive. This is where superpixel-based representations come into play. Images are composed of individual pixels, but it is important to note that a single pixel may not convey a complete and accurate representation of the depicted scene. Therefore, selecting groups of pixels that collectively represent semantically meaningful features becomes necessary, particularly for tasks like object detection. Segmentation algorithms are designed to group pixels based on common properties such as color, forming superpixels. Superpixels are non-overlapping segments that are easier to

work with than individual pixels. By converting pixels into superpixels, we increase expressiveness while reducing complexity. In the context of the LIME framework, superpixels are treated as features.

Figure 5.1 presents a high-level block diagram illustrating the LIME framework. The initial step involves selecting a specific instance for which we want to explain the classifier's decision. LIME then generates perturbed samples by introducing random disturbances to the features of the selected instance, as generated in the preceding step. Specifically, this involves *turning on and off different superpixels*. LIME aims to gain insights into the local behavior of a black box model f around the point of interest x, even without knowing its internal structure.

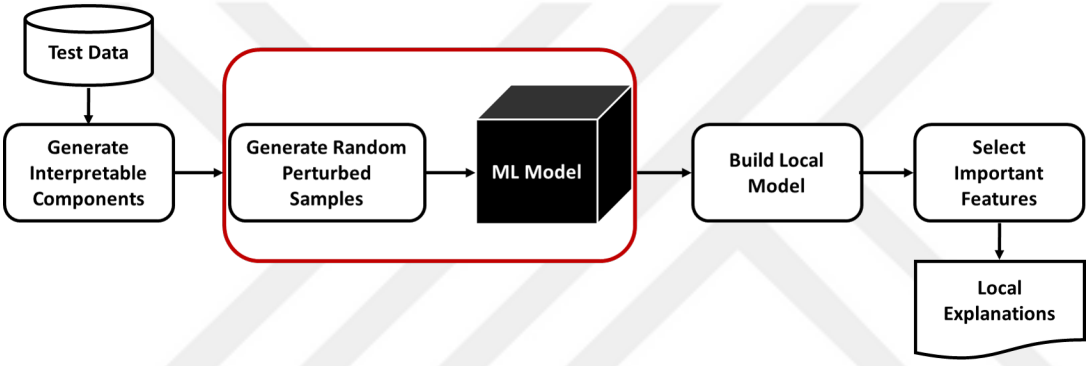


Figure 5.1: A block diagram illustrating the high-level structure of the LIME framework.

Given the lack of detailed information about the inner workings of the black box model f, LIME tackles this challenge by randomly generating samples in the proximity of x and querying the model to obtain predictions for each generated sample (red rectangle in Figure 5.1). This process creates a simulated dataset $D = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where y_i represents the class probabilities outputted by f for x_i. The number of perturbations used in the sampling process is determined by the hyperparameter n. The model f can take various forms, such as regression (with $y_i \in R$) or classification (with $y_i \in 0, 1$ or $y_i \in [0, 1]$ for probability outputs). To provide an explanation, LIME chooses a linear model g (denoted as the local model in Figure 5.1) from a collection of function spaces G, by solving an optimization problem:

$$\underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) \tag{5.1}$$

Where f refers to the original (or black box) model, x represents the original data point on which the explanation will be performed, and g corresponds to the model deduced by LIME, which can vary from a small linear model to a decision tree or a rule list, contingent on the nature of the data. This selection is based on the widely held belief that, while not universally interpretable across all data types, simpler models are often easier for users to understand. The proximity measure π quantifies the distance between an instance x_i and x, defining the locality around x. It assigns weights to x_i (perturbed instances) based on their distance from x. The first term of the formula shows the measure of the unfaithfulness of g in approximating f in the locality defined by π. This is known as the locality-aware loss in the original paper. The second term represents a measure of the complexity of the explanation model g. The coefficients of the linear model g indicate the contribution of each superpixel towards the prediction of each class.

To summarize, LIME introduces slight modifications to the features of the input data point and generates new predictions based on these modified samples. By systematically adjusting the value of a variable, it calculates the discrepancy between the new prediction and the original data point's prediction. This information is then used to determine the importance of that particular variable. In the context of this study, the variables that LIME selectively retains or disregards in each iteration correspond to the features of the generated samples, which are segments of image instances. For image data, the explanation provided by LIME typically involves highlighting the superpixels associated with the top positive and negative coefficients of the linear model, indicating which parts of the image have the greatest positive and negative influence on the model's prediction.(Garreau & Mardaoui, 2021)

5.2 MindfulLIME

In this subsection, we present our purposive perturbed sampling algorithm, which primarily aims to address the instability of LIME output. As mentioned earlier, LIME

generates samples randomly, posing risks such as output instability, generation of out-of-distribution data, or even creating nonsensical samples that disrupt the learning of local explanation models. In contrast, our algorithm adopts a deterministic approach, ensuring stability in the generated output. A deterministic algorithm follows a specific set of rules or instructions, consistently producing the same output for a given input. We control the sampling process by employing a non-random algorithm, allowing us to retain samples and consider specific conditions during their generation selectively. This capability significantly enhances the quality of the generated samples efficiently.

Figure 5.2 illustrates a high-level block diagram showcasing the MindfulLIME framework. In this framework, the instance that needs an explanation is represented as a graph, where the superpixel segments are depicted as vertices, and only the adjacent segments are connected by edges. Superpixels are generated using popular segmentation algorithms, such as Quickshift (Vedaldi & Soatto, 2008), Felzenszwalb (Felzenszwalb & Huttenlocher, 2004), and SLIC (Achanta et al., 2012), the details and comparisons of which will be provided in 5.3 Section. We denote this graph as $G = (V, E)$, where V represents the set of vertices (superpixels) and E represents the set of undirected edges. The cardinalities of V and E are denoted as $|V| = S$ (indicating the number of superpixels) and $|E|$, respectively.

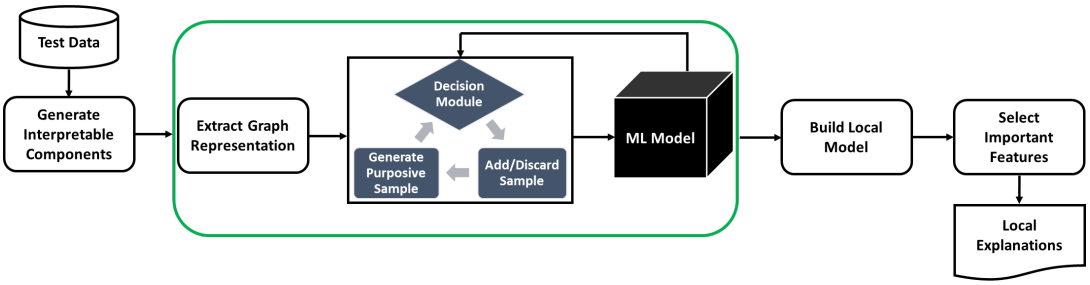


Figure 5.2: A block diagram illustrating the high-level structure of the MindfulLIME framework.

Our modified sampling solution is presented in the following algorithms. The core idea is to systematically analyze the graph of superpixels and generate new samples by selectively deactivating a specific subset of superpixels while leaving the rest activated. We represent the superpixel mask as a binary vector, denoted as $mask$, with a size of S, where each index corresponds to an individual superpixel, denoted as s_i.

A value of 1 indicates an activated superpixel, while 0 represents a deactivated one. In this context, *activation* refers to maintaining the superpixel in its original state, implying that the pixel values remain unchanged. Conversely, *deactivation* involves substituting the pixel values with the mean of all pixels belonging to that specific superpixel, effectively neutralizing its distinct features.

There are two phases within Algorithm 1, *PurposiveSampleGeneration*. In the initial phase, we generate masks where only one superpixel is deactivated per mask. This process aims to identify the individual superpixels that have a negative impact on the certainty of the classifier. Our objective is to determine which superpixels, in this phase, or a combination of superpixels in the next phase, when deactivated, cause the data point to deviate from the expected distribution. In the context of this study, when deactivating one or more superpixels decreases the certainty of the trained classifier, it reduces the reliability of the classifier's output for the corresponding generated sample. This reduction in certainty serves as an indicator of an out-of-distribution generated sample, the exact nature of which will be elaborated upon later in this section.

Assuming superpixels are represented as vertices, we generate V masks in this step, where each mask has only one deactivated vertex with a value of 0, while the rest of the vertices have a value of 1. The new mask is then applied to the original image, and the modified image is passed to the *DecisionModule*, Algorithm 2. The new mask is either discarded or retained depending on the decision the *DecisionModule* returns. If the new mask is discarded, the deactivated vertices and their connected edges are removed from the original graph. When the new mask is retained, a new sample tuple is generated and appended to the *Generated Sample Table (G)*, which includes mask values. Each *sample* has three properties: *path*, *value*, and *processed*. The *path* property consists of a list of edges (v_i, v_j) that represent the path and the order in which the deactivated vertices in the corresponding sample are processed. The *value* property represents the mask value of the sample. The binary *processed* property indicates whether the sample has been processed or not, based on whether the second vertex v_j in the last appended edge (v_i, v_j) in its *path* has been visited or not.

Algorithm 1 PurposiveSampleGeneration

Require: *Undirected Edge List (E), Vertex List (V), original_image, threshold*

Ensure: *Generated Sample Table (G)*

$G \leftarrow Table()$
for $v \leftarrow 1$ **to** V **do**
    $mask \leftarrow [1] * V.count$
    $mask[v] \leftarrow 0$
    $sample_image \leftarrow apply_mask(original_image, mask)$
    $decision \leftarrow DecisionModule(sample_image, threshold)$
    if $decision == True$ **then**
        Initialize new_sample and append it to G
    else
        $V.remove(v)$
        for (x, y) **in** E **do**
            if ($x == v$ **or** $y == v$) **then**
                $E.remove(x, y)$
            end if
        end for
    end if
end for
for $sample$ **in** $G.itertuples()$ **do**
    if $sample.processed == False$ **then**
        $v \leftarrow second_vertex(last_edge(sample.path))$
        for (x, y) **in** E **do**
            if ($x == v$ **or** $y == v$) **then**
                if ((x, y) **or** (y, x) **not in** $sample.path$) **then**
                    $mask \leftarrow sample.mask$
                    $mask[x], mask[y] \leftarrow 0, 0$
                    $sample_image \leftarrow apply_mask(original_image, mask)$
                    $decision \leftarrow DecisionModule(sample_image, threshold)$
                    if $decision == True$ **then**
                        Initialize new_sample and append it to G
                    end if
                end if
            end if
        end for
        $sample.processed \leftarrow True$
    end if
end for

Algorithm 2 DecisionModule

Require: *sample_image, threshold*

1: $probability \leftarrow Classifier(sample_image)$
2: **if** $probability > threshold$ **then**
3: **return** $True$
4: **end if**
5: **return** $False$

A subgraph consisting of the retained vertices from the first phase is passed as input to the second phase. In the second phase of Algorithm 1, we iterate over each generated sample, which is appended to the *Generated Sample Table (G)* (outer loop). If it has not been processed before, we retrieve the second vertex from the last appended edge in its *path* property for each sample. We then examine the unvisited adjacent vertices of the current vertex. In each iteration of the nested loop, we select one of the adjacent vertices, denoted as v_j, and generate a new mask by modifying the previous mask. This modification involves deactivating the vertex v_j in the new mask. Similar to the first phase, a new sample with the value of the new mask is generated and retained if determined to be appropriate by the *DecisionModule*. Otherwise, it is ignored. This process continues until no new samples can be generated and all the generated samples are processed.

The *DecisionModule*, Algorithm 2, plays the main role in determining whether the newly generated sample should be retained or discarded. Its primary objective is to identify and select in-distribution data points, which are characterized by higher confidence from our trained classifier. We employ a technique known as uncertainty sampling to filter out potential out-of-distribution data points. Uncertainty sampling is a machine learning and data annotation approach that prioritizes samples for further analysis or labeling based on their level of uncertainty. It utilizes probabilistic models, such as classifiers or predictive models, to estimate the uncertainty or confidence associated with each sample. By actively seeking out uncertain samples, uncertainty sampling aims to enhance the model's performance by focusing on areas where additional labeled data or targeted analysis would likely yield the greatest benefits.

There are various methods of uncertainty sampling, including entropy-based sampling, margin sampling, and confidence-based sampling. Entropy-based sampling

selects samples based on the entropy of the predicted class probabilities. Higher entropy indicates higher uncertainty, as the model is less confident in its predictions. Margin sampling, on the other hand, considers the difference between the top two predicted probabilities. A smaller margin suggests higher uncertainty, as the model is less decisive between the top two classes. Lastly, confidence-based sampling is a straightforward approach that includes samples if their predicted probability for the selected class exceeds a predefined confidence threshold. In this work, we applied this method in a reverse manner by focusing on selecting samples for which the model exhibits high confidence while disregarding those with lower probabilities. The confidence is calculated as the predicted probability that a given sample belongs to a particular class, derived directly from the output of our trained classifier.

Given that our problem involves multi-class, multi-label classification, each sample can belong to multiple classes, with each class potentially having multiple annotations. As a result, we execute the LIME algorithm separately for each class among the top three predicted probabilities to which the sample belongs. In Algorithm 2, we input the sample image into the trained classifier to calculate the probability of it belonging to the desired class. We need a criterion to determine which samples the model is more confident about and categorize as in-distribution data points. In this context, confidence is not just about the model making a prediction but about the degree of certainty associated with that prediction. Confidence-based sampling becomes a pivotal element in this scenario, especially considering the limited information available. It enables us to differentiate samples with a high degree of certainty from those with lower confidence. Setting specific thresholds for this confidence, as elaborated in the following paragraph, aids in systematically identifying and selecting these in-distribution data points.

To select thresholds, we calculate the average probabilities for each class over a small random unseen subset of the CheXpert dataset, specifically when the classifier correctly detects the correct label. These average probabilities serve as thresholds in our sample generation algorithm for each class within the top three classes to which the original image belongs. Therefore, if the probability of a newly generated unlabeled sample belonging to the desired class exceeds the respective threshold, we consider it to have a lower likelihood of being an out-of-distribution data point, making it more

suitable among some other samples.

This approach ensures that the order in which superpixel combinations are tested does not affect the stability of the outcome or efficiency of the algorithm. Moreover, it is important to note that while our algorithm does generate and evaluate multiple combinations of superpixels, it does so in an efficient manner. In the initial phase, the algorithm quickly narrows down the set of relevant superpixels by evaluating individual superpixels, and in the second phase, the algorithm intelligently generates and tests combinations of adjacent superpixels, rather than exhaustively testing all possible combinations.

Our method not only reduces the number of samples that need to be generated and evaluated but also leverages the spatial relationships between superpixels to generate more informative explanations. In comparison to LIME, which requires generating a much larger number of samples to accurately estimate the local decision boundary, our algorithm terminates with fewer samples while achieving better or comparable results. Our experiments indicate that our algorithm outperforms LIME in terms of efficiency, while consistently delivering stable results. Detailed results of these experiments will be provided further in Section 5.3.

The time complexity of the LIME random perturbation technique is influenced by various factors, including the complexity of the underlying model being explained, the number of perturbed samples, and the number of features or dimensions in the data. In this study, the features correspond to superpixels. The time complexity of the LIME random perturbation technique can be approximated as $O(d \times n \times f)$, where d represents the number of features (superpixels), n represents the number of perturbed samples, and f represents the complexity of fitting the interpretable model. This approximation assumes that the complexity of LIME is directly proportional to the number of perturbed samples n. Considering that the number of features d is generally much lower than the number of samples n ($|d| << |n|$), we can simplify the complexity of LIME as $O(n)$, where n represents the number of samples.

MindfulLIME's complexity is influenced by the number of features (superpixels), denoted as d, and the depth of the graph exploration, referred to as the *level*. Here, *level* signifies the number of iterations we run, where in each iteration, the exploration

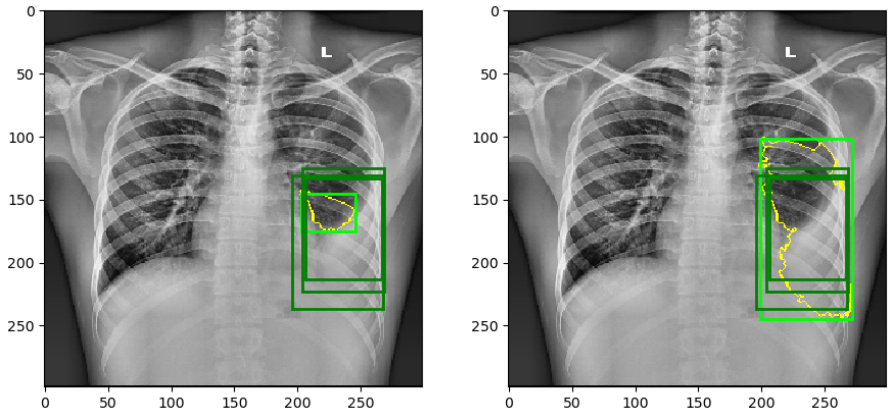
delves one node deeper into the graph, examining adjacent nodes. Initially, d samples are created, each containing a single 0 for the corresponding superpixel, resulting in an initialization complexity of $O(d)$. In the subsequent phase, individual adjacent superpixels to the previously visited ones are considered during each iteration (level). In the worst-case scenario, for every d previously generated samples, k new samples are generated. As a result, the complexity of MindfulLIME (i levels) can be estimated as $O(dk^i)$. The values of d and k can vary depending on the selected superpixel algorithms, highlighting the need to investigate further their impact on the performance and efficiency of MindfulLIME. This investigation will be conducted in the following section.

5.3 Experimental Results and Discussions

This section presents the results of the proposed MindfulLIME algorithm through two subsections: qualitative and quantitative analysis. In the qualitative analysis, we visually explore the algorithm's performance using two random samples, offering valuable insights into the effectiveness of MindfulLIME. The quantitative analysis compares the performance of LIME and MindfulLIME, focusing on stability, localization precision, and run-time efficiency. Various settings, including segmentation algorithms, different sample numbers, and features, comprehensively assess the algorithms' performance in diverse contexts.

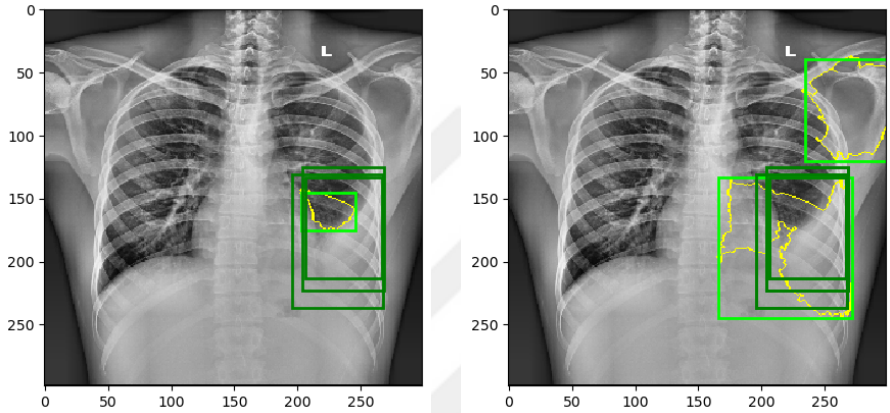
5.3.1 Qualitative Analysis

The behavior of MindfulLIME compared to LIME is demonstrated in Figures 5.3 and 5.4. Figure 5.3 showcases random sample 1 with a detected Pleural Effusion condition (annotations highlighted in green), while Figure 5.4 features random sample 2 with a detected Lung Opacity condition (annotations highlighted in blue). Each figure displays the ground-truth annotations (represented by darker-colored rectangles) in contrast to the explanations generated by LIME or MindfulLIME. These explanations are visualized as lighter-colored rectangles with yellow borders, indicating the selected superpixels as output from the algorithms. We would like to emphasize



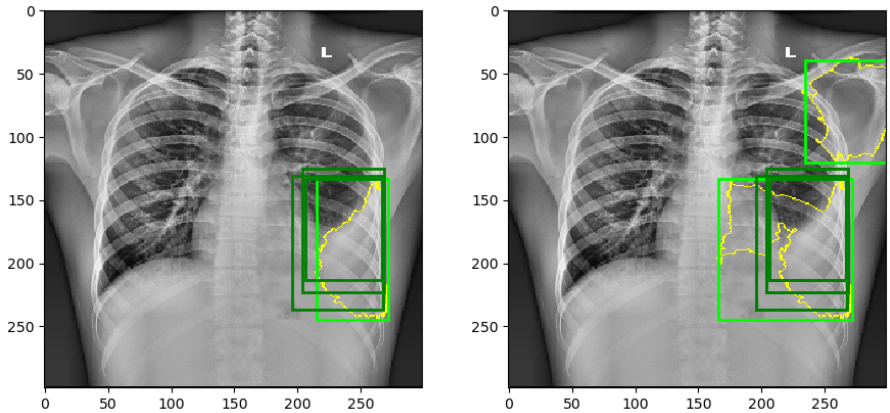
(a) MindfulLIME (1 Feature)

(b) MindfulLIME (4 Features)



LIME 1000 Samples (1 Feature)

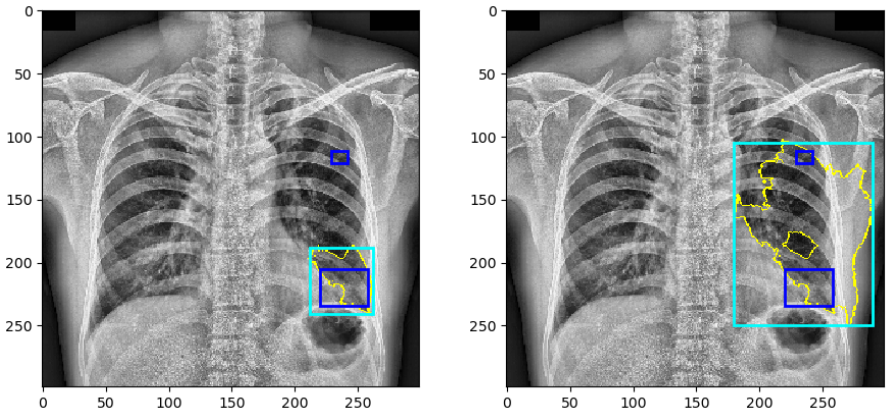
LIME 1000 Samples (4 Features)



LIME 4000 Samples (1 Feature)

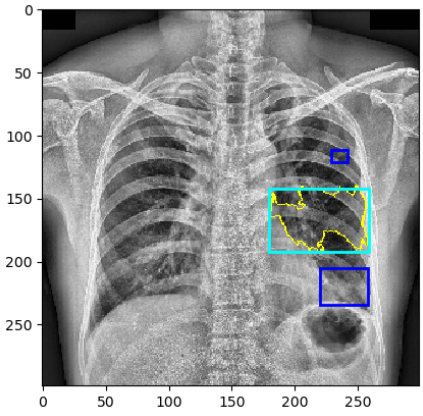
LIME 4000 Samples (4 Features)

Figure 5.3: Visual Analysis: Comparing generated explanations (the yellow-outlined areas enclosed with light green bounding boxes) by MindfulLIME and LIME to the ground truth annotations (dark green bounding boxes) for Random Sample 1

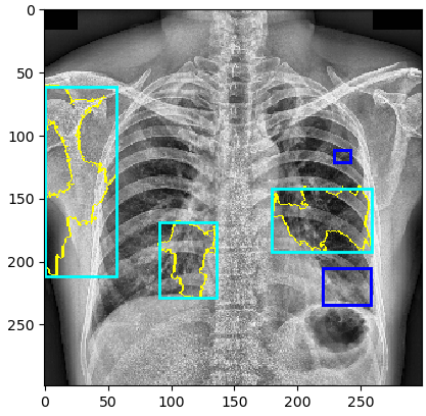


(a) MindfulLIME (1 Feature)

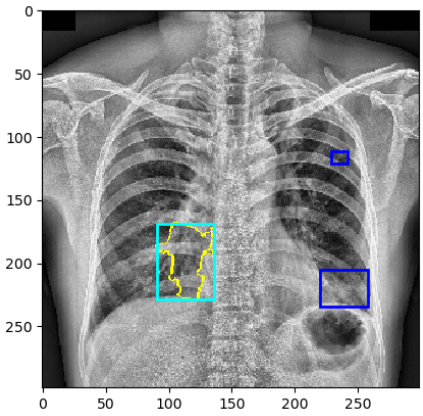
(b) MindfulLIME (4 Features)



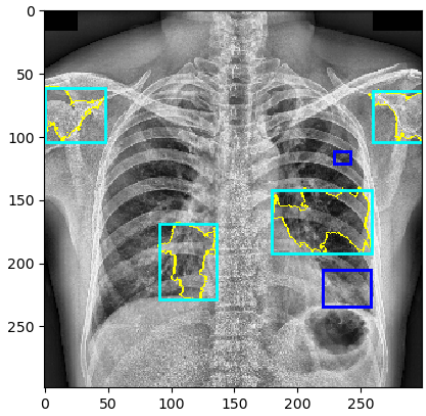
LIME 1000 Samples (1 Feature)



LIME 1000 Samples (4 Features)



LIME 4000 Samples (1 Feature)



LIME 4000 Samples (4 Features)

Figure 5.4: Visual Analysis: Comparing generated explanations (the yellow-outlined areas enclosed with light blue bounding boxes) by MindfulLIME and LIME to the ground truth annotations (dark blue bounding boxes) for Random Sample 2

that our trained classifier model generalizes well and is not simply memorizing the training data, as evidenced by its performance on a separate, unseen dataset used for evaluation.

The left columns depict the results for the top 1 feature, while the right columns display the results for the top 4 features. The first row presents the output of MindfulLIME, while the second and third rows represent the results of LIME using 1000 and 4000 samples, respectively. When comparing the results of LIME and MindfulLIME for two visual random samples, it becomes evident that MindfulLIME excels in multiple aspects. Firstly, MindfulLIME consistently identifies the top 1 feature with the highest similarity with the actual bounding box. Additionally, as the number of top features increases, MindfulLIME effectively prioritizes and presents more relevant features at the forefront. These findings highlight the enhanced performance and prioritization capabilities of MindfulLIME compared to LIME.

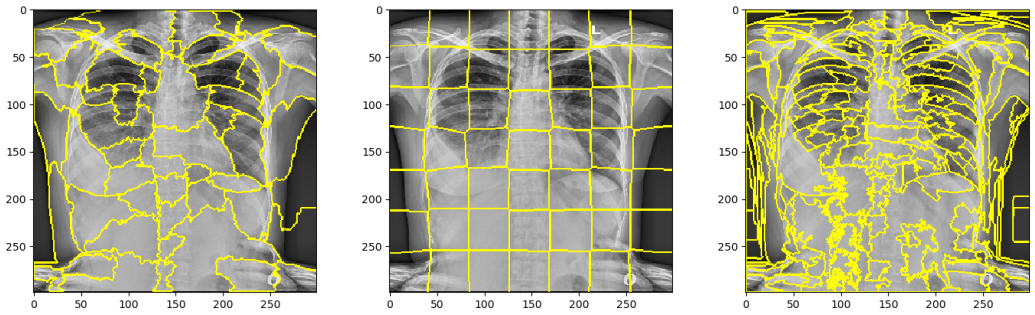
5.3.2 Quantitative Analysis

We conduct experiments using a set of 200 test instances, running each for ten iterations to demonstrate the instability inherent in explanations generated by LIME. The experiments aim to explain predictions made by our trained chest X-ray classifier using both LIME and MindfulLIME. Our evaluation focuses not only on stability but also on precision and run-time efficiency, enabling a comprehensive comparison of the results.

The experiments involve evaluating three versions of MindfulLIME: MindfulLIME (2 levels), MindfulLIME (3 levels), and MindfulLIME (4 levels). Through experiments across different level settings, we examine the trade-off between performance and efficiency, gaining insights into the impact of levels on the algorithm's effectiveness and efficiency. We maintain the default values for most LIME parameters and settings, except for the generated random sample number and the segmentation algorithms used. We specifically focus on reporting the results of LIME for the most significant 1 and 4 features. The feature count, representing the number of patches (superpixels) generated by the LIME algorithm, necessitates manual configuration.

The default value for the number of perturbed samples in LIME is 1000. However, we also conduct experiments with 2000 and 4000 samples to investigate their impact on stability and other aspects. Additionally, we explore the effects of various popular segmentation algorithms apart from LIME's default algorithm (Quickshift). Each superpixel algorithm exhibits variations in shape, size, and number of generated segments. We aim to evaluate the influence of different superpixel shapes and numbers on precision and run-time efficiency.

We compare Quickshift with two additional segmentation algorithms: Felzenszwalb and SLIC. To visually demonstrate these algorithms, Figure 5.5 shows the output of each algorithm on a random sample from the VinDr-CXR dataset. We tune the parameters of the two additional segmentation algorithms to optimize their performance. We identify the most appropriate parameter values by conducting a comprehensive grid search on the parameter space of each segmentation method, using 50 randomly selected VinDr-CXR data samples. This thorough parameter tuning process enables us to tailor the segmentation algorithms to our specific data, ensuring they provide detailed and accurate results on X-ray images. The optimized parameter configurations for each segmentation approach are presented in Table 5.1.



(a) Quickshift (b) SLIC (c) Felzenszwalb

Figure 5.5: Exploring Various Segmentation Algorithms: Given superpixels to the LIME generated by different segmentation algorithms

We also determine the thresholds for sample generation by calculating the average probabilities for each class over a small random subset of the CheXpert dataset. Specifically, we focus on instances where the classifier accurately detects the labels of interest. The selected classes and their corresponding thresholds for the experiments

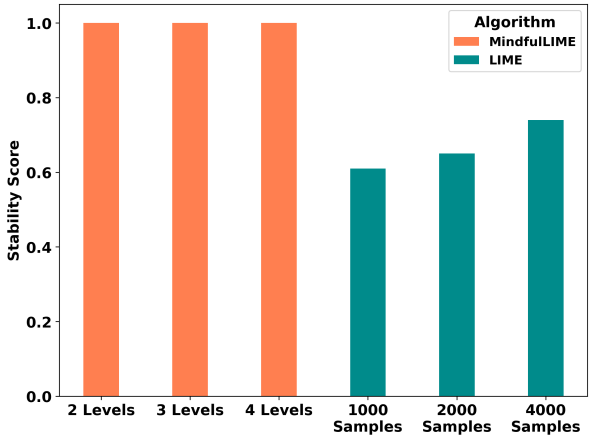
Table 5.1: Segmentation Algorithms Parameters

Segmentation Algorithms	Parameters	Value
SLIC	n_segments	50
	compactness	80
	sigma	20
Felzenszwalb	scale	600
	min_size	200
	sigma	0.2

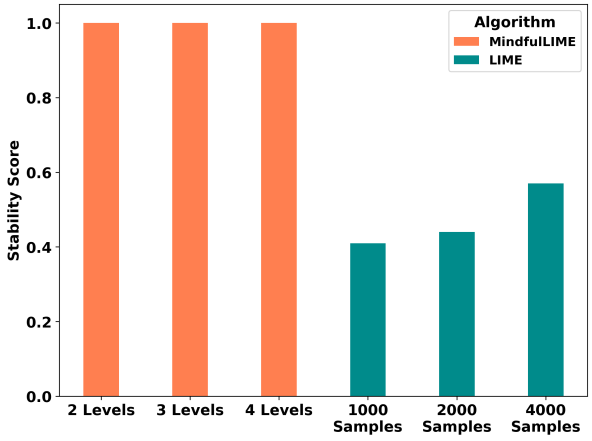
are as follows: Cardiomegaly (0.42), Pleural Effusion (0.44), Lung Opacity (0.19), Consolidation (0.25), Atelectasis (0.16), and Pneumothorax (0.79). These thresholds serve as the criteria for identifying confident in-distribution data points during the sample generation process (Decision Module).

Figure 5.6 illustrates the stability results obtained for different versions of LIME and MindfulLIME using various superpixel algorithms. For each sample instance, we conducted the explanation generation task 10 times. We calculated the Intersection over Union (IOU) in each iteration between the ground truth and the generated explanation. The average IOU across all iterations was then computed as a quantitative measure of stability. Notably, only MindfulLIME achieved a perfect stability score of 100%, while all other versions of LIME exhibited instability. Although increasing the number of random samples generated in LIME improved stability to some extent, the overall instability problem persisted. Regardless of the segmentation algorithm employed, MindfulLIME consistently demonstrated complete stability, confirming its expected effectiveness.

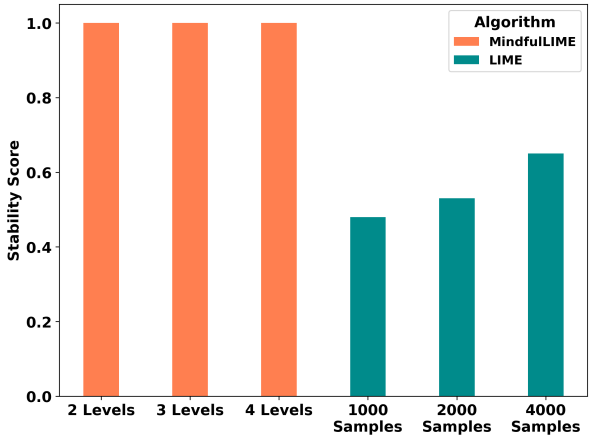
Moving on to Figure 5.7, it presents the localization similarity between the ground truth annotations and the explanations generated by different variations of LIME and MindfulLIME, considering various segmentation algorithms and feature settings. A higher localization precision score indicates greater consistency between the explanation generated by an algorithm and the actual annotated data in terms of bounding box distance. Since each sample instance in our data can contain multiple regions



(a) Quickshift

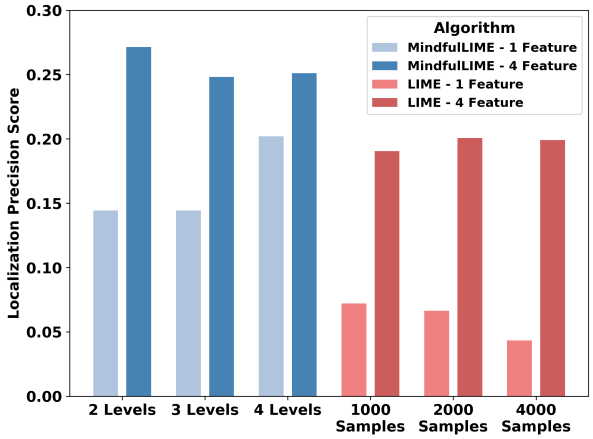


(b) SLIC

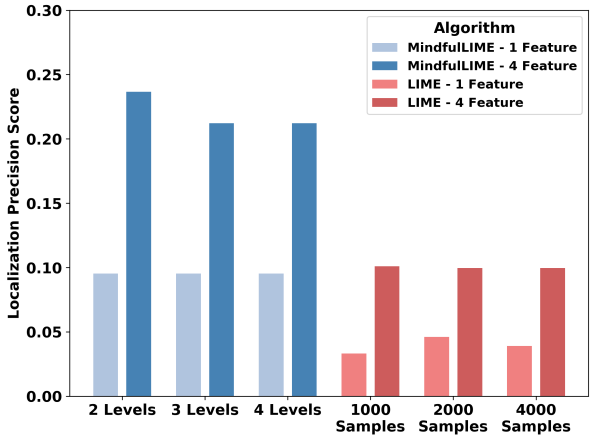


(c) Felzenszwalb

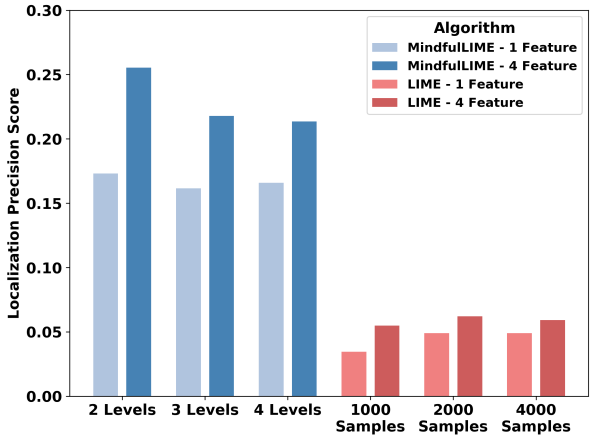
Figure 5.6: Stability Comparison: LIME vs. MindfulLIME for various versions and settings



(a) Quickshift



(b) SLIC



(c) Felzenszwalb

Figure 5.7: Quality Comparison: LIME vs. MindfulLIME for various versions and settings

marked by doctors for the same detected condition, we conducted experiments with the top 1 feature and the top 4 features to observe the effect of increasing the number of top superpixels in the generated explanations.

In all settings, MindfulLIME demonstrates higher localization precision or, in other words, higher explanation quality than LIME. Except for the Quickshift algorithm with the top 1 feature setting, MindfulLIME (2 Levels) consistently produces the best results in all other cases. Additionally, it is evident that increasing the number of top features to 4 leads to an increase in localization precision for both algorithms, which can be attributed to the multi-label nature of samples and the granularity of the marked annotations in actual data.

The run-time of LIME and MindfulLIME is compared with different numbers of superpixels generated by various segmentation methods, as shown in Figure 5.8 and Table 5.2. Each segmentation algorithm produces varying numbers, sizes, and arrangements of segments within an image. For instance, as shown in Table 5.2, SLIC divides images into 49 superpixels, Quickshift into 66, and Felzenszwalb into 158 on average.

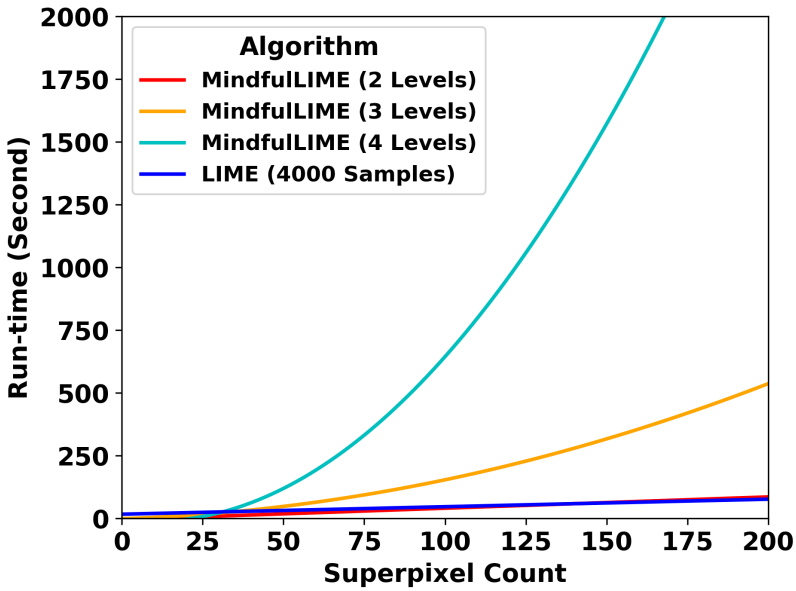


Figure 5.8: Run-Time Comparison: LIME vs. MindfulLIME for various versions and superpixel numbers

While LIME's run-time primarily depends on the number of perturbed samples, Mind-

fulLIME's complexity is influenced by the number of superpixels (which affect the maximum adjacent superpixels) and the algorithm's depth level. As the depth level increases in MindfulLIME, the impact of the number of superpixels, especially the maximum adjacent superpixels, intensifies. This is due to the exponential increase in explored branches and subsequent processes. The effect of this intensification is demonstrated in Table 5.2, where superpixel algorithms that produce denser and more segments, such as Felzenszwalb, require additional computations and lead to longer processing times.

Run-time efficiency holds significance when discussing LIME or MindfulLIME. In practical applications, especially in critical sectors like healthcare, timely outputs can streamline operations and decision-making processes. These algorithms, when operating seamlessly, can swiftly provide insights for individual data points, making them more user-friendly and applicable in real-world scenarios. Moreover, during the model's development or tuning stages, there may be situations where evaluations need to span across entire datasets. Here, efficiency in execution can dramatically shorten the time required for comparisons and adjustments. Concluding, while the core emphasis is often placed on the quality and stability of explanations, the relevance of run-time efficiency should not be underestimated.

MindfulLIME proves to be a practical choice, achieving a favorable balance between stability, precision, and efficiency. It demonstrates these qualities without needing deeper levels or intricate computations, particularly by employing less complex superpixel algorithms like SLIC and Quickshift with a higher granularity of segments. Our findings highlight the superiority of MindfulLIME over LIME in various segmentation settings, as it generates fewer high-quality samples within an acceptable processing time. Overall, MindfulLIME guarantees stability and improves localization precision, making it a valuable tool for generating reliable and accurate explanations in machine learning models.

Concluding our discussion, stability in explainable AI, especially in sensitive areas like healthcare, is crucial not just for consistency but also for establishing trust and ensuring safe decision-making. When experts use AI systems, they expect reliable and consistent reasoning behind decisions. An unstable model, despite its accuracy,

Table 5.2: Run-Time Comparison: LIME vs. MindfulLIME for various versions and superpixel numbers

Algorithm	Version	Average Superpixel Count	Average Run-time (Second)	Average Generated Samples
LIME	1000 Samples	49	11.6	1000
		66	12.0	1000
		158	13.8	1000
	2000 Samples	49	23.2	2000
		66	23.9	2000
		158	27.6	2000
	4000 Samples	49	46.2	4000
		66	47.7	4000
		158	54.9	4000
MindfulLIME	2 Levels	49	14.5	173
		66	21.7	261
		158	69.4	834
	3 Levels	49	40.5	513
		66	79.8	1016
		158	349.0	4196
	4 Levels	49	91.7	1183
		66	258.0	3275
		158	1763.0	17836

can cast shadows of uncertainty and confusion, complicating the decision-making process. Moreover, stable explanations streamline model refinement and enable clear accountability trails. Notably, our proposed algorithm, MindfulLIME, achieves this stability without compromising the quality of explanations. In fact, it enhances both stability and the richness of the insights, ensuring users get the best of both worlds.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Despite the computational power some machine learning-based models offer, their non-transparent structure limits the domains on which such models can be applied. On one side, these models can be significantly beneficial for solving complex problems. Conversely, their black-box nature makes their decisions unreliable to experts. Therefore, in this study, we present an explainable AI solution for the problem of thorax disease classification. Our proposed framework comprises two primary phases: the exploration of appropriate evaluation metrics and the selection of the most suitable one, along with the proposal of a stable yet efficient algorithm for generating samples in perturbation-based XAI models.

We trained various renowned neural network architectures on the CheXpert dataset, identifying InceptionV3 as the most efficient for our multi-label classifier. Applying the LIME method with diverse segmentation configurations, we generated explanations for unseen samples from the VinDr-CXR dataset. Highlighting a core contribution, we elucidate the process of achieving the ideal metric for evaluating generated explanations, crafting criteria based on the desired quality of explanations. We present an extensive range of metrics and, using the set criteria, evaluate the effectiveness of each, detailing our selection of the most fitting metric. Through qualitative analysis, we observed the behavior of our chosen metric compared to others, confirming its superiority in terms of accuracy and reliability. Our quantitative analysis not only reveals that employing the right superpixel algorithms can yield more precise and efficient explanations, bypassing LIME's default, but also showcases the utility of having such a metric in comparing diverse XAI models. We demonstrated that, with the optimized LIME, we can achieve results surpassing even those of advanced

models like GradCAM.

We then introduce MindfulLIME, an engineered algorithm to optimize both stability and performance without compromising efficiency. It proposes a purposive sample generation algorithm to eradicate randomness, employs a graph-based algorithm to consider correlated features, and applies uncertainty sampling techniques to precisely localize explanations. Together, these components address the instability issues associated with existing models like LIME which using random perturbation sampling, guaranteeing the production of more reliable and interpretable explanations for black-box models.

Despite advancements in the stability of other data modalities, a notable gap remains for image data. Our rigorous training and evaluation on the CheXpert and VinDr-CXR datasets, respectively, confirm that MindfulLIME not only attains 100% stability but also efficiently surpasses LIME in localization precision while generating fewer samples. This achievement demonstrates MindfulLIME's potential to enhance the interpretability of black-box models, particularly in critical domains. These extensive experiments not only bridge this gap but also underline the model's potential for broader applications beyond the medical imaging domain addressed in this work. The selection of these datasets implies that MindfulLIME can be effectively generalized across other domains, given the challenges associated with interpreting such images. By overcoming the fundamental limitations of LIME, MindfulLIME ensures stable and accurate explanations, thereby amplifying the interpretability of black-box models. This positions MindfulLIME as a valuable, durable, and indispensable tool for the future.

While MindfulLIME has successfully achieved significant stability and shown encouraging results, there are multiple avenues for future research to enhance its performance and efficiency. Crucial areas for improvement include investigating alternative sample filtering techniques and incorporating domain knowledge into the sampling process, specifically designed to identify In-Distribution samples. This action directly influences the pruning policies of the graph, thereby optimizing the model's efficiency. Additionally, we are exploring the integration of various feature selection methods to create even more precise and interpretable explanations. Although

the current study focuses on medical images as a case study, MindfulLIME's design is versatile, making it applicable to any data sample convertible into a graph of correlated features. Nevertheless, we plan to apply MindfulLIME to a wider array of datasets and domains to determine its generalizability and applicability in diverse contexts. These anticipated enhancements will be the focus of our future work, and we are confident they will contribute to further optimizing MindfulLIME's performance.

REFERENCES

Abdullah, T. A. A., Zahid, M. S. M., Ali, W., & Hassan, S. U. (2023). B-lime: An improvement of lime for interpretable deep learning classification of cardiac arrhythmia from ecg signals. *Processes*, *11*(2). https://doi.org/10.3390/pr11020595

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, *34*(11), 2274–2282. https://doi.org/https://doi.org/10.1109/TPAMI.2012.120

Agarwal, C., & Nguyen, A. (2021). Explaining Image Classifiers by Removing Input Features Using Generative Models. In H. Ishikawa, C.-L. Liu, T. Pajdla, & J. Shi (Eds.), *Computer vision – accv 2020* (pp. 101–118). Springer International Publishing.

Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M. L., & Hossain, M. S. (2020). Study of different deep learning approach with explainable AI for screening patients with COVID-19 symptoms: Using CT scan and chest X-ray image dataset. https://doi.org/https://doi.org/10.3390/make2040027

Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7786–7795.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What do different evaluation metrics tell us about saliency models? *IEEE transactions on pat-*

tern analysis and machine intelligence, *41*(3), 740–757. https://doi.org/https://doi.org/10.1109/TPAMI.2018.2815601

Cardellino, C., & Alemany, L. A. (2015). Reversing uncertainty sampling to improve active learning schemes.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE winter conference on applications of computer vision (WACV)*, 839–847. https://doi.org/https://doi.org/10.1109/WACV.2018.00097

Collenette, J., Atkinson, K., & Bench-Capon, T. (2023). Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights. *Artificial Intelligence*, *317*, 103861. https://doi.org/https://doi.org/10.1016/j.artint.2023.103861

Dong, S., Wang, P., & Abbas, K. (2021). A survey on deep learning and its applications. *Computer Science Review*, *40*, 100379. https://doi.org/https://doi.org/10.1016/j.cosrev.2021.100379

DoshiVelez, F., & Kim, B. (2018). Considerations for evaluation and generalization in interpretable machine learning. *Explainable and interpretable models in computer vision and machine learning*, 3–17. https://doi.org/https://doi.org/10.1007/978-3-319-98131-4_1

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International journal of computer vision*, *59*(2), 167–181. https://doi.org/https://doi.org/10.1023/B:VISI.0000022288.19776.77

Fourcade, A., & Khonsari, R. H. (2019). Deep learning in medical image analysis: A third eye for doctors. *Journal of Stomatology, Oral and Maxillofacial Surgery*, *120*(4), 279–288. https://doi.org/https://doi.org/10.1016/j.jormas.2019.06.002 55th SFSCMFCO Congress

Garreau, D., & Mardaoui, D. (2021). What does lime really see in images? *Proceedings of the 38th International Conference on Machine Learning*, *139*, 3620–3629. https://proceedings.mlr.press/v139/garreau21a.html

Giudici, P., & Raffinetti, E. (2022). Explainable AI methods in cyber risk management. *Quality and Reliability Engineering International*, *38*(3), 1318–1326. https://doi.org/https://doi.org/10.1002/qre.2939

Graziani, M., Lompech, T., Müller, H., & Andrearczyk, V. (2020). Evaluation and comparison of cnn visual explanations for histopathology. *Explainable Agency in Artificial Intelligence at AAAI21*, 195–201.

Hailemariam, Y., Yazdinejad, A., Parizi, R. M., Srivastava, G., & Dehghantanha, A. (2020). An empirical evaluation of ai deep explainable tools. *2020 IEEE Globecom Workshops (GC Wkshps*, 1–6. https://doi.org/https://doi.org/10.1109/GCWkshps50303.2020.9367541

Hall, P., Gill, N., Kurka, M., Phan, W., & Bartz, A. (2019). Machine Learning Interpretability with H2O Driverless AI: First Edition Machine Learning Interpretability with H2O Driverless AI. (June), 1–40. http://docs.h2o.ai

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. https://doi.org/https://doi.org/10.1109/CVPR.2016.90

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance.

Hu, L., Chen, J., Nair, V. N., & Sudjianto, A. (2018). Locally interpretable models and effects based on supervised partitioning (lime-sup).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708. https://doi.org/https://doi.org/10.1109/CVPR.2017.243

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D., Halabi, S., Sandberg, J., Jones, R., Larson, D., Langlotz, C., Patel, B., Lungren, M., & Ng, A. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 590–597. https://doi.org/https://doi.org/10.1609/aaai.v33i01.3301590

Jiang, E. (2022). UniformLIME: A Uniformly Perturbed Local Interpretable Model-Agnostic Explanations Approach for Aerodynamics. *Journal of Physics: Conference Series*, *2171*(1). https://doi.org/10.1088/1742-6596/2171/1/012025

In this paper, we propose an original modification to LIME by employing a

new perturbed sample generation method for aerodynamic tabular data in regression model, which makes the differences between perturbed samples and the input instance vary in a larger range. We make several comparisons with three subtasks and show that our proposed method results in better metrics.

Jin, D., Sergeeva, E., Weng, W.-H., Chauhan, G., & Szolovits, P. (2022). Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, *14*. https://doi.org/https://doi.org/10.1002/wsbm.1548

Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., & Rajpurkar, P. (2021). Chextransfer: Performance and parameter efficiency of imagenet models for chest x-ray interpretation. *Proceedings of the Conference on Health, Inference, and Learning*, 116–124. https://doi.org/10.1145/3450439.3451867

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences*, *34*(4), 1060–1073. https://doi.org/https://doi.org/10.1016/j.jksuci.2019.06.012

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, *42*, 11–24. https://doi.org/https://doi.org/10.1016/j.patrec.2014.01.008

Lewis, D. D., & Gale, W. A. (1994). A Sequential Algorithm for Training Text Classifiers. In B. W. Croft & C. J. van Rijsbergen (Eds.), *Sigir '94* (pp. 3–12). Springer London.

Li, X.-H., Shi, Y., Li, H., Bai, W., Cao, C. C., & Chen, L. (2021). An experimental study of quantitative evaluations on saliency methods. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3200–3208. https://doi.org/https://doi.org/10.1145/3447548.3467148

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 1–45. https://doi.org/10.3390/e23010018

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Miok, K., Nguyen-Doan, D., Zaharie, D., & Robnik-Šikonja, M. (2019). Generating data using monte carlo dropout. *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 509–515. https://doi.org/10.1109/ICCP48234.2019.8959787

Mohseni, S., Block, J. E., & Ragan, E. (2021). Quantitative evaluation of machine learning explanations: a human-grounded benchmark. *26th International Conference on Intelligent User Interfaces*, 22–31. https://doi.org/https://doi.org/10.1145/3397481.3450689

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T. T., Dinh, D. H., Do, C. D., Doan, L. T., Nguyen, C. N., Nguyen, B. T., Nguyen, Q. V., Hoang, A. D., Phan, H. N., Nguyen, A. T., Ho, P. H., … Vu, V. (2020). VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations, 1–10.

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models.

Petsiuk, V., Das, A., & Saenko, K. (2019). RisE: Randomized input sampling for explanation of black-box models. *British Machine Vision Conference 2018, BMVC 2018*, (November).

Prasanna, D. L., & Tripathi, S. L. (2022). Machine and Deep-Learning Techniques for Text and Speech Processing. In *Machine learning algorithms for signal and image processing* (pp. 115–128). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781119861850.ch7

Qiu, L., Yang, Y., Cao, C. C., Zheng, Y., Ngai, H., Hsiao, J., & Chen, L. (2022). Generating Perturbation-based Explanations with Robustness to Out-of-Distribution Data. *WWW 2022 - Proceedings of the ACM Web Conference 2022*, 3594–3605. https://doi.org/10.1145/3485447.3512254

Rahnama, A. H. A., & Boström, H. (2019). A study of data and label shift in the lime framework.

Rajaraman, S., Candemir, S., Kim, I., Thoma, G., & Antani, S. (2018). Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences (Switzerland)*, *8*(10). https://doi.org/https://doi.org/10.3390/app8101715

Ramaswamy, H. G., & Desai, S. (2020). Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 972–980. https://doi.org/https://doi.org/10.1109/WACV45572.2020.9093360

Rasouli, P., & Yu, I. C. (2019). Meaningful data sampling for a faithful local explanation method. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11871 LNCS*, 28–38. https://doi.org/10.1007/978-3-030-33607-3_4

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. https://doi.org/https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11491

Robnik-Šikonja, M. (2016). Data generators for learning systems based on rbf networks. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(5), 926–938. https://doi.org/10.1109/TNNLS.2015.2429711

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. https://doi.org/https://doi.org/10.1007/978-3-319-24574-4_28

Saito, S., Capel, N., Chua, E., & Hu, R. (2020). Improving LIME Robustness with Smarter Locality Sampling. *arXiv*

We propose to make LIME more robust by training a generative adversarial network to sample more realistic synthetic data which the explainer uses to generate explanations. Our.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (2019). *Explainable ai: Interpreting, explaining and visualizing deep learning*. https://doi.org/https://doi.org/10.1007/978-3-030-28954-6

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE conference*

on computer vision and pattern recognition, 4510–4520. https://doi.org/https://doi.org/10.1109/CVPR.2018.00474

Sangroya, A., Rastogi, M., Anantaram, C., & Vig, L. (2020). Guided-lime: Structured sampling based hybrid approach towards explaining blackbox machine learning models. *CIKM (Workshops)*.

Sattarzadeh, S., Sudhakar, M., Lem, A., Mehryar, S., Plataniotis, K. N., Jang, J., Kim, H., Jeong, Y., Lee, S., & Bae, K. (2021). Explaining convolutional neural networks through attribution-based input sampling and block-wise feature aggregation. *34th AAAI Conference on Artificial Intelligence*.

Schallner, L., Rabold, J., Scholz, O., & Schmid, U. (2020). Effect of Superpixel Aggregation on Explanations in LIME – A Case Study with Biological Data. In P. Cellier & K. Driessens (Eds.), *Machine learning and knowledge discovery in databases* (pp. 147–158). Springer International Publishing.

Schulz, K., Sixt, L., Tombari, F., & Landgraf, T. (2020). Restricting the Flow: Information Bottlenecks for Attribution. *International Conference on Learning Representations*.

Shankaranarayana, S. M., & Runje, D. (2019). ALIME: Autoencoder Based Approach for Local Interpretability. In H. Yin, D. Camacho, P. Tino, A. J. Tallón-Ballesteros, R. Menezes, & R. Allmendinger (Eds.), *Intelligent data engineering and automated learning – ideal 2019* (pp. 454–463). Springer International Publishing.

Shi, S., Zhang, X., & Fan, W. (2020). A modified perturbed sampling method for local interpretable model-agnostic explanation.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR*.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. https://doi.org/10.1145/3375627.3375830

Sokol, K., & Flach, P. (2020). Explainability fact sheets: a framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference*

on Fairness, Accountability, and Transparency, 56–67. https://doi.org/https://doi.org/10.1145/3351095.3372870

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Teixeira, L. O., Pereira, R. M., Bertolini, D., Oliveira, L. S., Nanni, L., Cavalcanti, G. D. C., & Costa, Y. M. G. (2021). Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors*, *21*(21). https://doi.org/https://doi.org/10.3390/s21217116

van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, 102470. https://doi.org/https://doi.org/10.1016/j.media.2022.102470

Vedaldi, A., & Soatto, S. (2008). Quick shift and kernel methods for mode seeking. *European conference on computer vision*, 705–718. https://doi.org/https://doi.org/10.1007/978-3-540-88693-8_52

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion*, *76*(100), 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Visani, G., Bagli, E., & Chesani, F. (2022). Optilime: Optimized lime explanations for diagnostic computer algorithms.

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Vreš, D., & Šikonja, M. R. (2021). Better sampling in explanation methods can prevent dieselgate-like deception.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vi-*

sion and pattern recognition workshops, 111–119. https://doi.org/https://doi.org/10.1109/CVPRW50498.2020.00020

Xiang, A., & Wang, F. (2019). Towards interpretable skin lesion classification with deep learning models. *AMIA, Annual Symposium proceedings*, *2019*, 1246–1255.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In *Proceedings of the 33rd international conference on neural information processing systems*. Curran Associates Inc.

Zafar, M. R., & Khan, N. (2021). Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction*, *3*(3), 525–541. https://doi.org/10.3390/make3030027

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – eccv 2014* (pp. 818–833). Springer International Publishing.

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., & Sclaroff, S. (2018). Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, *126*(10), 1084–1102. https://doi.org/https://doi.org/10.1007/s11263-017-1059-x

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "why should you trust my explanation?" understanding uncertainty in lime explanations.

Zhou, Y., Zhou, Z., & Hooker, G. (2018). Approximation trees: Statistical stability in model distillation.

Zhou, Z., Hooker, G., & Wang, F. (2021). S-lime: Stabilized-lime for model explanation. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2429–2438. https://doi.org/10.1145/3447548.3467274

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Rahimiaghdam, Shakiba

EDUCATION

Degree	Institution	Year of Graduation
Ph.D. (Computer Engineering)	Middle East Technical University	2023
M.S. (Computer Engineering)	Middle East Technical University	2015
B.S. (Software Engineering)	Alzahra University	2011

PROFESSIONAL EXPERIENCE

- **E-Kalite, Ankara, Turkey:** April 2018 - April 2019 (Data Analyst)
  - First Project: Conducted an analysis of financial data supplied by Turkey's Health Ministry, involving data cleaning, time series analysis application, and prediction generation
  - Second Project: Examined patients' Electronic Health Records (EHR) provided by Turkey's Health Ministry, entailing data cleaning, preprocessing, and the employment of traditional machine learning algorithms for early prediction of severe diseases, including colon cancer
- **TUBITAK Project, Ankara, Turkey:** November 2016 - March 2018 (Graduate Research Assistant)

  - Generating a comprehensive, large-scale audio-visual dataset from outdoor surveillance videos
  - Executing data preprocessing, annotation, and labeling tasks
  - Introducing a hybrid neural network-based framework tailored for precise object classification and detection

- **Iran Afkar Group, Tehran, Iran:** February 2011 - April 2012 (Software Engineer)

  - Conducting a feasibility study for the development plans of Tehran Clinic's Hospital HIS (Hospital Information System)
  - Identifying requirements, exploring types of subsystems, and understanding physician needs
  - Analyzing data and process flow diagrams, and preparing technical comments and reply sheet letters to facilitate communication among clients, designers, and the programming team

TECHNICAL SKILL

- **Concepts/Courses:** Data Science/Analytic, Machine Learning Techniques, Supervised Learning, Deep Neural Networks, Convolutional Neural Network, Explainable Artificial Intelligence, Data Mining, Database Management System, Image Processing, Computer Vision Techniques, Information Retrieval Systems, Web Search Engine Design Concepts

- **Environments/Tools:** Microsoft Office, Microsoft Visual Studio, PyCharm, Weka

- **Programming Languages:** Python (Keras, TensorFlow, Pandas, NumPy, Scikit-learn, Matplotlib), R, SQL, C, C++, C#

PUBLICATIONS

Conference Publication

- **ACM SAC 2016**
  Personalized Time-Aware Outdoor Activity Recommendation System, Shakiba Rahimiaghdam, Pinar Karagoz, Alev Mutlu, 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016

Under Review Journals

- Evaluating The Quality of Visual Explanations on Chest X-ray Images for Thorax Diseases Classification, Shakiba Rahimiaghdam, Hande Alemdar, Journal of Neural Computing and Applications (Under Review), 2023
- MindfulLIME: A Stable Solution for Explanations of Machine Learning Models with Enhanced Localization Precision - A Medical Image Case Study, Shakiba Rahimiaghdam, Hande Alemdar, Expert Systems: The Journal of Knowledge Engineering (Revision Submitted), 2023