



Evaluating the quality of visual explanations on chest X-ray images for thorax diseases classification

Shakiba Rahimiaghdam¹ · Hande Alemdar¹

Received: 23 March 2023 / Accepted: 5 February 2024
© The Author(s) 2024

Abstract

Deep learning models are extensively used but often lack transparency due to their complex internal mechanics. To bridge this gap, the field of explainable AI (XAI) strives to make these models more interpretable. However, a significant obstacle in XAI is the absence of quantifiable metrics for evaluating explanation quality. Existing techniques, reliant on manual assessment or inadequate metrics, face limitations in scalability, reproducibility, and trustworthiness. Recognizing these issues, the current study specifically addresses the quality assessment of visual explanations in medical imaging, where interpretability profoundly influences diagnostic accuracy and trust in AI-assisted decisions. Introducing novel criteria such as informativeness, localization, coverage, multi-target capturing, and proportionality, this work presents a comprehensive method for the objective assessment of various explainability algorithms. These newly introduced criteria aid in identifying optimal evaluation metrics. The study expands the domain's analytical toolkit by examining existing metrics, which have been prevalent in recent works for similar applications, and proposing new ones. Rigorous analysis led to selecting Jensen–Shannon divergence (JS_DIV) as the most effective metric for visual explanation quality. Applied to the multi-label, multi-class diagnosis of thoracic diseases using a trained classifier on the CheXpert dataset, local interpretable model-agnostic explanations (LIME) with diverse segmentation strategies interpret the classifier's decisions. A qualitative analysis on an unseen subset of the VinDr-CXR dataset evaluates these metrics, confirming JS_DIV's superiority. The subsequent quantitative analysis optimizes LIME's hyper-parameters and benchmarks its performance across various segmentation algorithms, underscoring the utility of an objective assessment metric in practical applications.

Keywords Machine learning · Explainable artificial intelligence · Deep neural networks · Medical image classification

1 Introduction

The advent of machine learning-based systems has dramatically transformed numerous complex tasks across various domains, from image object detection to nuanced text processing and advanced speech signal analysis [1, 2]. These advancements, while groundbreaking, have brought to the forefront the critical issue of artificial intelligence (AI) transparency, especially in high-stakes domains such as medicine, justice, and cybersecurity [3–5]. The black

box nature of many AI and machine learning algorithms, characterized by their non-transparent, complex, and nested computations, poses significant challenges in establishing user trust, as decisions made by these systems often lack clear, traceable explanations based on input data [6].

In this context, given the sensitivity of the medical domain, the importance of transparency cannot be overstated, especially when it comes to some difficult tasks such as analyzing chest X-ray (CXR) images. 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 [7, 8]. In response to these challenges, the field of explainable artificial intelligence (XAI) has emerged, advocating for interpretable and transparent AI models. XAI aims to demystify the decision-making processes of AI, making them understandable and justifiable, particularly in critical

✉ Shakiba Rahimiaghdam
shakiba.rahimiaghdam@metu.edu.tr

Hande Alemdar
alemdar@metu.edu.tr

¹ Department of Computer Engineering, Middle East Technical University, 06800 Ankara, Turkey

applications such as medical imaging [9]. While much of the current work in AI for medical applications focuses on feature selection and classification [10, 11] or on improving segmentation algorithms for more accurate image analysis [12], there remains a significant gap in understanding and evaluating the explanatory capabilities of these models. The current study addresses this gap by focusing specifically on the evaluation of visual explanations in medical imaging AI models, wherein the context of this study, *explanation* refers to image superpixels or segments crucial to the classifier's decision.

The gap in the literature and practice has led to the introduction of a robust, comprehensive methodology for the quantitative evaluation of visual explanations generated by AI models in medical imaging. This contribution is diverse and addresses several key challenges in the field. The initial step involves defining a set of criteria essential for the ideal metrics that can accurately measure the quality of explanations, with a special emphasis on localization precision, a critical aspect in medical image interpretation. The exploration extends beyond reevaluating existing metrics in computer vision to introduce new metrics, thereby enriching the analytical toolkit available for researchers and practitioners.

The methodology employed is detailed and thorough. Newly introduced metrics are systematically assessed against the defined criteria in various scenarios, illuminating their effectiveness and applicability in both theoretical and practical applications. This process identifies the most effective metrics that address the challenge of evaluating visual explanation quality and transcends the specific scope of the study to offer broader applicability.

Another key aspect of this methodology is the detailed analysis and enhancement of the local interpretable model-agnostic explanations (LIME) method [13]. The research emphasizes the influence of various segmentation approaches on LIME's efficacy, specifically contrasting them with Quickshift [14], its default algorithm, and explores multiple alternatives to identify the most effective option. This exploration is motivated by the goal of attaining a scalable and computationally efficient assessment, leveraging the meticulously developed metrics within the research's scope.

In summary, the research makes the following original and significant contributions:

- Introduction of a comprehensive set of metrics specifically for evaluating visual explanations in medical image classification.
- Definition and curation of essential characteristics for these metrics, providing a novel framework for their systematic analysis.

- Qualitative analysis of the performance of these metrics on the VinDr-CXR dataset, demonstrating their effectiveness in a real-world context.
- Application of the most suitable metrics to improve the quality and reliability of explanations generated by LIME through quantitative analysis.

The rest of the manuscript is organized as follows. The studies in the literature related to visual explanations are outlined, and the metrics used for explainable AI (XAI) model evaluation are surveyed in Sect. 2. The datasets and the trained classifier used in the study are described in Sect. 3. The proposed approach is explained in detail in Sect. 4. Discussions and results of extensive evaluations on the VinDr-CXR dataset are provided in Sect. 5. Finally, the study is concluded in Sect. 6.

2 Related works

Explanations in machine learning can be presented in different modalities, such as textual, mathematical (or numerical), and visual. Visual explanation methods are often preferred because they facilitate understanding and interpreting the output of black box models. A common method for generating such presentations is through the use of saliency maps. The output of LIME in saliency map format is compared to the ground truth masks annotated by experts in this study. Therefore, the focus is on visual explanations generated by LIME, particularly for medical image data, and metrics evaluating the similarity or difference between saliency maps with an emphasis on localization assessments for multi-label classifiers.

Recent works regarding visual explanations for medical image data are surveyed. Xiang and Wang [15] used LIME on a skin lesion classifier to validate that the trained convolutional neural network (CNN) utilizes relevant information for its diagnoses. Arrieta et al. [9], by examining the qualities of LIME explanations, claimed that existing deep learning models perform better on chest X-ray images than computed tomography (CT) scan images for classifying COVID-19 patients. Rajaraman et al. [16] evaluated, visualized, and explained the behavior of trained models for detecting pneumonia using methods such as LIME, class activation maps (CAM), and its other variations. Ahsan et al. [17] applied LIME to detect essential features distinguishing COVID-19 patients from others using CT and CXR images. In another work, Teixeira et al. [18] utilized LIME and Grad-CAM to evaluate the impact of lung segmentation using U-Net CNN architecture [19] for generating segmentations over three well-known CNN classifiers. They applied explanation models to a test set of

images and aggregated all explanation regions in a heatmap to show the model's most common areas for prediction.

There are several taxonomies in the literature for visual explanation evaluation. As classified by Arrieta et al. [9], outputs are mainly discussed in two broad groups: qualitative and quantitative analysis. Qualitative evaluation involves visual investigation of the outcomes of one or more models for a limited number of random samples. Quantitative analysis is categorized into two sub-groups: model truth-based and ground truth-based [20]. The former investigates model behavior by assessing the algorithm's consistency, correctness, and faithfulness, while the latter justifies the visual quality of explanation maps through comparisons with ground truth maps determined by amateur or expert users. Another classification divides assessment approaches into three classes: application-grounded, human-grounded, and functionally-grounded [21], with the first two resembling ground truth-based methods and the latter similar to model-based approaches.

Qualitative analysis plays a crucial role in assisting users to evaluate the quality of explanations. However, to establish or generalize the explanations' correctness, appropriateness, and precision, more than these analyses are needed. Notable metrics used in model truth-based and ground truth-based evaluations are investigated. Model truth-based metrics generally evaluate explanations' faithfulness, and correctness [22]. According to Alvarez and Jaakkola [23], faithfulness of an explanation refers to whether relevance scores reflect true importance. Measures like drop percentage, increase percentage [24, 25], and iAUC (incremental area under the curve) [26] are employed to evaluate faithfulness. Efforts to evaluate concepts such as stability, separability, identity [27, 28], and consistency in explanations [29] have also been made.

This study addresses ground truth-based and object localization-based evaluation metrics and approaches, particularly for medical image classifications. Li et al. [22] recommended the pointing game (PG) metric to assess whether explainers correctly localize objects to be recognized. With the assumption that a good saliency map is not required to cover the entire object, PG [30] counts hit numbers if the highest score saliency pixel lies inside the annotated bounding box. Likewise, Sattarzadeh et al. [20] validated the quality of their proposed visual explanation model in comparison with other state-of-the-art attribution-based explanation models using the energy-based pointing game (EBPG), mIoU (mean intersection over union), and bounding box (Bbox)[31]. EBPG extends the concept of PG by considering all pixels in the explanation map for evaluation, measuring the fraction of intersection with the ground truth map normalized by the explanation region [32]. This approach evaluates the precision and denoising capability of the explanations. The weighted intersection

over salient region (WIoSR) is a modified version of the EBPG metric, where saliency scores of detected regions are used as weights in calculating the localization score [26, 33].

Schallner et al. [34] investigated the effect of various segmentation algorithms on LIME performance by utilizing the Jaccard coefficient in a binary classification task for malaria blood smear images. Despite some similarities between this work and the current study, fundamental differences exist. The focus here is on exploring various suitable metrics, introducing, and adapting them to find the most appropriate one for the defined problem of explainability in a more complex classification task for a multi-label multi-class dataset. Additionally, a second, unseen, larger dataset is utilized to replace human intervention with a more efficient and accurate automated system. Significantly, this study clearly highlights the limitations of intersection-based metrics, such as IoU (Jaccard coefficient), mIoU, Bbox, EBPG (IoSR), and WIoSR, in accurately measuring the similarities between two bounding boxes across a range of different scenarios.

3 Thorax disease classification

The focus of this study is on the explanation quality of thorax disease classifier outputs for CXR images. Therefore, dataset preparation and classifier training are essential initial steps in this process. Selecting the appropriate dataset and model for training the CXR image classifier require particular attention, unlike regular image classification tasks. Details about these specific considerations are provided in the subsequent subsections.

3.1 Datasets

The CheXpert [35] and VinDr-CXR [36] datasets are utilized for this research. While numerous chest X-ray datasets exist, many of them suffer from issues like inadequate sample sizes, restricted access, lack of comprehensive annotations, or unreliable annotations. Thus, these two datasets have been chosen for their complementary strengths. The CheXpert dataset, with its ample samples per class and reliable global annotations, is particularly suited for training and classification tasks using advanced deep neural networks. In contrast, the VinDr-CXR dataset, offering precise regional annotations for several classes overlapping with those in CheXpert, is ideal for evaluating generated explanations.

The CheXpert dataset contains 224,316 chest radiographs from 65,240 patients. These images come with chest-related global annotations, which were extracted using an automated, rule-based labeler. This labeler was

designed to categorize radiological findings from free-text radiology reports penned by radiologists. During the label extraction process, additional rules were incorporated to better capture negations and uncertainties present in the reports, thereby enhancing the reliability of the annotations. Each CheXpert data entry encompasses an image ID, the patient's sex and age, the frontal/lateral and AP/PA properties of the X-rays, followed by a hot vector for 14 distinct classes. These classes are labeled with 0 for negative cases, +1 for positive cases, and -1 for uncertain cases. The 14 classes include atelectasis, cardiomegaly, consolidation, edema, pleural effusion, pneumonia, pneumothorax, enlarged cardiomeastinum, lung lesion, lung opacity, pleural other, fracture, support devices, and no finding. Alongside the training set, a validation set of 234 samples is also provided, which have been verified by radiologists, with uncertain labels duly removed.

The VinDr-CXR dataset, comprising 15,000 images, stands out due to the meticulous manual label definitions provided by a group of radiologists. Representative samples from this dataset are displayed in Fig. 1. It includes 22 local labels, which are essentially bounding boxes delineating abnormalities, and 6 global labels identifying detected diseases. The images are presented in digital imaging and communications in medicine (DICOM) format, complete with relevant DICOM tags such as the patient's sex or age. The dataset primarily consists of PA-view CXRs, and any non-relevant X-rays (those depicting other body parts) have been excluded. The local labels comprise a range of conditions, including aortic enlargement, atelectasis, cardiomegaly, calcification, clavicle fracture, consolidation, edema, emphysema, enlarged PA, interstitial lung disease (ILD), infiltration, lung cavity, lung cyst, lung opacity, mediastinal shift, nodule/mass, pulmonary fibrosis, pneumothorax, pleural thickening, pleural effusion, rib fracture, and other lesion. While the global labels cover categories like lung tumor, pneumonia,

tuberculosis, other diseases, chronic obstructive pulmonary disease (COPD), and no finding.

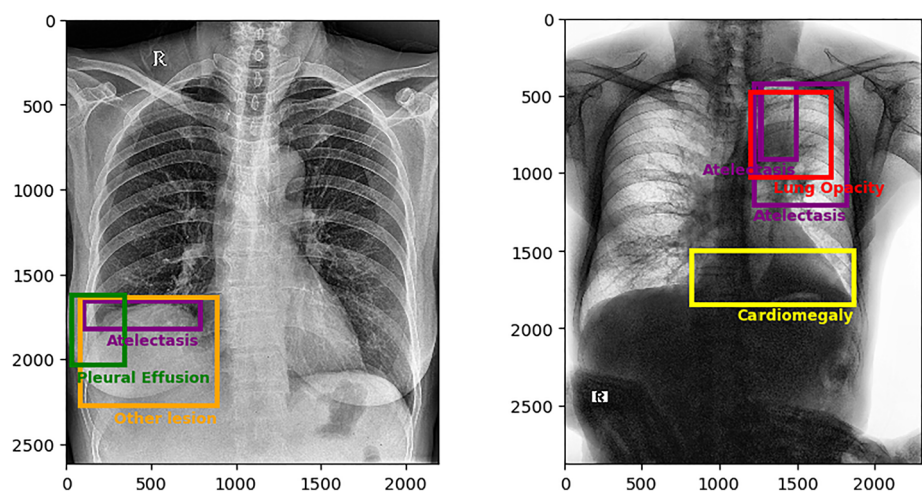
For both datasets, the test sets were reserved for competitions at the time of writing this paper. The distribution of samples across each label category for both datasets is depicted in Fig. 2. The CheXpert dataset's comprehensive sample availability per class renders it apt for training and classification tasks in deep neural networks. In contrast, the VinDr-CXR dataset, with its detailed local annotations for classes overlapping with CheXpert, is well-suited for evaluating the generated explanations.

3.2 Training the classification model

Due to the significant impact of the classifier's result on the quality of explanations, it is crucial for the trained model to identify labels as accurately as possible. Explanations generated for a poorly performing classifier may not be valuable. Data preprocessing techniques are applied to the CheXpert data, where samples with any uncertain label among their classes are excluded. Additionally, only PA and Lateral images are selected for both datasets. As the CheXpert dataset is part of an open competition, the test set is not yet available. Thus, the training phase utilizes 85% of samples in the filtered training set (83,781), reserving the remaining 15% of the samples for validation and testing tasks. A subdivision is made where 10% of the remaining 15% is allocated for validation, and the remaining 5% is combined with the filtered original validation set to create the test set.

Following this, several well-known CNN architectures such as InceptionV3 [37], DenseNet121 [38], VGG16 [39], ResNet50 [40], MobileNetV2 [41], and InceptionResNetV2 [42] are trained. All models are trained on an Nvidia GeForce RTX 3080 GPU, using a batch size of 16 with the Adam optimizer (learning rate of 0.0001) and leveraging the initial weights from the pre-trained models

Fig. 1 Samples of the VinDr-CXR dataset with annotated labels



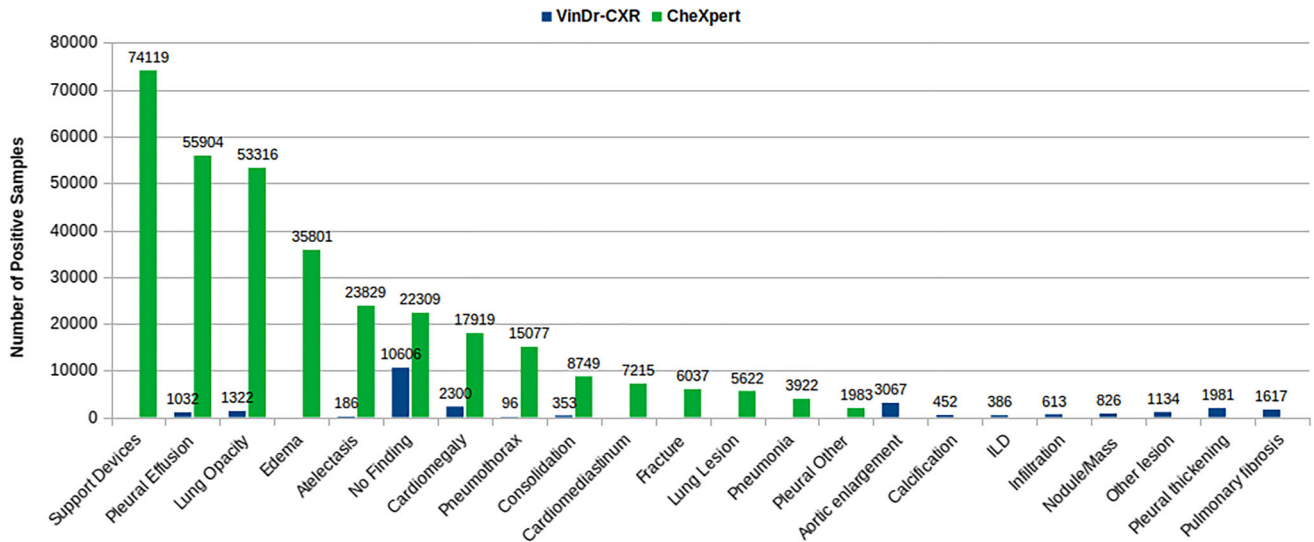


Fig. 2 Distribution of samples within classes for both CheXpert and VinDr-CXR datasets

on the ImageNet dataset. Evaluation results of these algorithms, based on the Area under the ROC Curve (AUC), are presented in Table 1. Given the minor differences in the results, InceptionV3 is selected as the classifier model for the subsequent phases of the study.

4 Methodology

The study is conducted in two phases. The initial phase involves generating explanations using the LIME algorithm to illustrate the inner workings of the trained black box classifier. The subsequent phase focuses on evaluating the efficiency of these explanations. This involves defining essential criteria expected from the evaluation metrics, exploring various localization-aware metrics, and selecting the most appropriate ones. Details of these phases are discussed in the following subsections.

4.1 Generating visual explanations

Visual explanations are generated to understand the reasoning logic of the model trained using the LIME algorithm [13]. LIME is versatile, capable of explaining classification models ranging from random forests and gradient boosting trees to neural networks. It is applicable to various types of input data, including text, images, or tabular data. LIME operates by slightly altering the input and generating new predictions for these modified inputs. By adjusting the value of a variable, LIME calculates the difference between the new prediction and the original data point’s prediction, thereby determining the importance of that variable. In this study, segments of image samples are treated as variables that LIME decides to keep or ignore in each iteration. An extensive analysis is conducted with a large set of samples to explore the role of superpixels in the precision of the explanations.

The quality of the explanations significantly depends on various hyper-parameters that LIME uses, especially the superpixels given to the model as input. Images are defined

Table 1 Comparing different trained classifiers by their AUC percentages for each class

Class name	IRNetV2 ¹	MNetV2 ²	VGG16	ResNet50	DenseNet121	InceptionV3
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

¹InceptionResNetV2

²MobileNetV2

by a set of pixel values, where individual pixels alone do not represent a natural depiction of the scene. Thus, groups of pixels are selected as semantically meaningful features for different studies, particularly in object detection. Segmentation algorithms group pixels based on common properties, such as color, forming what are known as superpixels. By converting individual pixels into superpixels, expressiveness increases while complexity decreases.

Felzenszwalb [43], SLIC [44], and Quickshift [14] are the segmentation algorithms used for comparison in this study. Felzenszwalb is a graph-based algorithm efficient in segmenting images based on the minimum spanning tree principle. Simple linear iterative clustering (SLIC) is a cluster-based superpixel algorithm that employs k-means clustering in a five-dimensional space of color information. Quickshift, the default algorithm of LIME, generates superpixels by approximating kernelized mean-shift and applying the mode-seeking method. As an illustration, Fig. 3 shows the output of these superpixel algorithms on a random sample from the VinDr-CXR dataset.

4.2 Evaluating visual explanations

As previously mentioned, the main focus of the current study is to present a practical solution for the evaluation of generated explanations. This section addresses the issue in a step-by-step manner. Initially, essential criteria are defined, outlining what is expected from an acceptable explanation when the classifier accurately detects the labels. Following this, a comprehensive set of metrics suitable for comparing local annotations made by experts with generated explanations by LIME are introduced. Lastly, the drawbacks of each metric, in relation to the defined criteria, are discussed, taking into account their respective formulas. Detailed information on these aspects is presented in the subsequent subsections.

4.2.1 Desired criteria

Before defining appropriate metric characteristics, assuming the classifier accurately predicts the label(s), it is necessary to specify expectations for an adequate explanation. The factors to be fulfilled while evaluating a generated explanation are as follows:

- **Overlap with Ground Truth:** Generated explanations must have maximum overlap with the ground truth annotations (maximum true positive) and minimum overlap outside the ground truth annotations (minimum false positive).
- **Multi-label Consideration:** Generated explanations should cover multiple regions akin to ground truth annotation in multi-label cases (minimum false negative)
- **Minimized Discrepancy:** In instances of zero intersection between explanation and ground truth regions, the distance between these regions should be minimal.

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

- **Informativeness:** This criterion stresses the importance of the metric producing a nonzero value under any scenario. It is essential that the metric captures the relationship between the explanations and ground truth annotations, irrespective of the configuration—be it direct alignment, one region encompassing another, or any other arrangement. The metric should provide a value that adequately reflects the quality of the explanation in comparison with the ground truth.
- **Localization:** Centered on spatial considerations, this criterion evaluates the closeness of generated explanations to the ground truth. Explanations in close

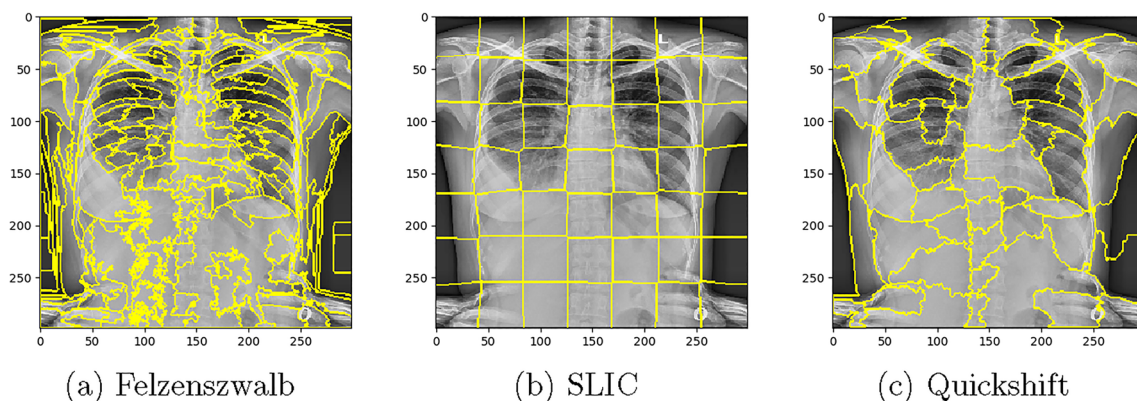


Fig. 3 Given superpixels to the LIME generated by different segmentation algorithms

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 introduced aim to thoroughly and extensively assess the effectiveness of generated explanations. By carefully considering all aspects of explanation quality, from spatial alignment to complexities in multi-label scenarios, these criteria ensure a comprehensive evaluation. This innovative contribution enhances the precision of evaluation metrics and exemplifies the growing sophistication and meticulousness in the field of explainable AI.

4.2.2 Localization-based evaluation metrics

This research explores various metrics to evaluate the quality of generated explanations. By comparing the localization of LIME's explanations to the ground truth annotations of the VinDr-CXR dataset, the best configuration for the LIME method can be determined, or even different explanation methods can be compared in terms of accuracy. It is assumed that both ground truth masks and LIME explanation masks are converted into one-dimensional lists. This approach simplifies handling multi-bounding boxes for both masks. Additionally, to make both maps comparable and ensure consistency, pixel values are converted to binary values. The ground truth mask is denoted 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) is a commonly used metric for error calculation [45]. It measures the distance between the model's saliency mask and the ground truth mask. Formally, it is defined as follows:

$$\text{MAE}(\text{GT}, \text{EX}) = \frac{1}{P} \sum_{i=1}^P |\text{GT}(i) - \text{EX}(i)| \quad (1)$$

Intersection over union (IOU), or Jaccard index is a popular metric in object detection tasks. It quantifies the similarity between two sets by dividing the intersection of the estimated area and ground truth area by their union. An IOU of zero indicates no intersection, while one signifies complete overlap. IOU is calculated pixel wise, taking the intersection of the GT map and the EX map, divided by their union:

$$\text{IOU}(\text{GT}, \text{EX}) = \frac{\sum_{i=1}^P \text{GT}(i) \cap \text{EX}(i)}{\sum_{i=1}^P \text{GT}(i) \cup \text{EX}(i)} \quad (2)$$

Jensen–Shannon divergence (JS_DIV) offers a method to quantify the difference or similarity between two probability distributions [46]. This approach uses the Kullback–Leibler divergence (KL_DIV) measurement, which quantifies the extent to which one probability distribution differs from another [47]. In this context, the GT and EX maps are transformed into probability distributions of the ground truth map and explanation map, respectively. This transformation ensures that the sum of all values in each map equals one. Given two distributions, KL_DIV is calculated as:

$$\text{KL_DIV}(\text{EX}_{\text{pd}}, \text{GT}_{\text{pd}}) = \sum_{i=1}^P \text{EX}_{\text{pd}}(i) \log \left(\frac{\text{EX}_{\text{pd}}(i) + \varepsilon}{\text{GT}_{\text{pd}}(i) + \varepsilon} \right) \quad (3)$$

KL_DIV is identified as a non-symmetrical and nonlinear measure, ranging from zero to infinity. A lower score value between EX_{pd} and GT_{pd} indicates a closer approximation of EX_{pd} to the ground truth GT_{pd} . Essentially, this score reflects the divergence in probabilities: a significant divergence occurs when the probability of an event is high in EX_{pd} but low in GT_{pd} . Conversely, when the probability of an event is low in EX_{pd} and high in GT_{pd} , there is still notable divergence, although it is not as pronounced as in the first scenario.

JS_DIV employs KL_DIV to compute a normalized, symmetrical score in the following manner:

$$\text{JS_DIV}(\text{EX}_{\text{pd}}, \text{GT}_{\text{pd}}) = \frac{\text{KL}(\text{EX}_{\text{pd}}, M) + \text{KL}(\text{GT}_{\text{pd}}, M)}{2} \quad (4)$$

In this context, M represents the average of two probability distributions. JS_DIV offers a smoothed and normalized version of KL_DIV, yielding scores that range from zero (indicating identical distributions) to one (signifying maximally different distributions) when employing the base-2

logarithm. As KL_DIV is a score rather than a metric, its symmetric normalized version, JS_DIV , serves as an appropriate substitute.

Pixel wise precision (PWP), or intersection over salient region (IoSR), calculates the ratio of the intersection between the explanation mask and the ground truth mask over the explanation map [22]. PWP is defined as:

$$PWP(GT, EX) = \frac{\sum_{i=1}^P GT(i) \cap EX(i)}{\sum_{i=1}^P EX(i)} \quad (5)$$

The formula mirrors the structure of the well-known precision metric. It considers the intersection of the ground truth mask with the explanation mask as true positives. The total area of the explanation mask comprises a combination of both true positives and false positives.

Pixel wise recall (PWR), inspired by the recall metric, can also be called as intersection over ground truth saliency (IoGS). This metric calculates the ratio of the intersection between the explanation mask and the ground truth mask over the ground truth area. PWR is calculated using the following formula:

$$PWR(GT, EX) = \frac{\sum_{i=1}^P GT(i) \cap EX(i)}{\sum_{i=1}^P GT(i)} \quad (6)$$

The formula aligns with the concept of the recall metric. In this context, the intersection of the ground truth mask with the explanation mask represents true positives, while the entire area of the ground truth mask encompasses both true positives and false negatives.

Pixel wise accuracy (PWA) shares a similar concept with pixel wise precision (PWP) and pixel wise recall (PWR). Following the conventional definition of accuracy, PWA is defined as the percentage of correctly detected pixels. This means PWA represents the division of the sum of true positives and true negatives (the XNOR of GT and EX) by the sum of all true positives, true negatives, false positives, and false negatives (encompassing the entire image). PWA is calculated using the following formula:

$$PWA(GT, EX, Image) = \frac{\sum_{i=1}^P GT(i) \odot EX(i)}{\sum_{i=1}^P Image(i)} \quad (7)$$

To the best of current knowledge, JS_DIV and the proposed PWR and PWA metrics are being employed in this domain for the first time. Additionally, based on the available literature, apart from MAE and IOU, the other metrics mentioned here are being applied for the first time to evaluate explanations generated by LIME.

4.2.3 Evaluation metric selection

Based on the criteria defined in Sect. 4.2.1 and the metrics presented in Sect. 4.2.2, this section concludes the

methodology by exploring the behavior of each metric in terms of the criteria. This approach helps in narrowing down the proposed metric set to 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 their numerator, fail to calculate the distance between two entirely separated regions. In cases where explanation and ground truth bounding boxes are completely separated, IOU, PWP, and PWR become zero. Hence, intersection-based metrics cannot guarantee the measurement of *informativeness* and *localization* in certain scenarios. Moreover, MAE and PWA, despite considering false positives and false negatives, fail to accurately evaluate more complex cases in terms of *localization* due to their disregard for position.

Regarding *coverage*, both PWP and PWR struggle in some scenarios. PWP fails to evaluate better *coverage* when considering only the explanation region ($\sum_{i=1}^P EX(i)$) in its denominator, and PWR similarly fails when accounting for only the ground truth region ($\sum_{i=1}^P GT(i)$). Additionally, PWR reaches its maximum value when the explanation region completely surrounds the ground truth bounding box, leading to an inaccurate assessment of *coverage*. Conversely, PWP falls short in scenarios where explanation masks are entirely within the ground truth annotations.

In the proposed explainable classifier for multi-label data, due to pixel wise calculation, none of the mentioned metrics violate the *multi-target capturing* criterion, even if more than one region is marked as an explanation region. This pixel wise approach ensures that multiple marked areas in the explanation can be compared simultaneously to multiple annotated ground truth regions.

MAE and PWA, as examples, violate the *proportionality* condition by distributing the difference across the total number of pixels. As a result, even in cases with significant explanation differences, these metrics show only minor changes in their values. Intersection-based metrics such as IOU, PWP, and PWR also violate *proportionality* due to their inability to measure accurately in certain scenarios, as explained above.

The essential characteristics of each evaluation metric applicable to the current problem are summarized in Table 2. This analysis concludes that JS_DIV is the only metric fulfilling all necessary conditions. Detailed results and discussions related to these metrics and the utilized segmentation algorithms are provided in the subsequent section.

Table 2 Summary of the essential characteristics for each evaluation metric

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

5 Experimental analysis and findings

This section thoroughly presents the results of experiments, organized into distinct subsections for clarity and depth. Sect. 5.1 comprehensively details the key parameter settings fundamental to the study's methodology. It emphasizes the configurations of superpixel segmentation algorithms and incorporates important settings of the LIME algorithm, laying a robust groundwork for the analysis that follows. Proceeding to Sect. 5.2, a qualitative analysis is conducted. This analysis closely examines the performance of the segmentation algorithms in both their standard and optimized configurations and evaluates the effectiveness of metrics established earlier in the study, in line with the essential criteria defined. Sect. 5.3 shifts focus to a quantitative analysis, where the effects of various segmentation algorithms and their respective parameter settings are assessed, guided by the metrics selected for this purpose. This structured layout of the section ensures a comprehensive and interconnected understanding of each facet of the study, spanning from the detailed parameter settings to the intricate qualitative and quantitative evaluations.

5.1 Parameter settings

Most LIME parameters remain at default settings; however, an essential adjustment is made to the number of samples used for perturbation. To mitigate LIME's inherent instability with smaller sample sizes, the number of samples is increased from the standard 1000–3000. This enhancement ensures more consistent and reliable interpretations, which is crucial in medical image analysis, where the accuracy and trustworthiness of model explanations are paramount. Despite the higher computational demand, this approach is justified by the need for precision in medical diagnostics.

Furthermore, the LIME method is evaluated using feature settings of 1, 2, and 4. The term feature in the context of the LIME algorithm refers to the number of patches, or superpixels, that are identified as explanations. This range is chosen to explore the algorithm's performance across a

spectrum of explanation granularity, from coarser to finer details, providing important insights into how different levels of explanation detail affect interpretability and accuracy in medical diagnostics.

For the comprehensive exploration of the effects of various superpixel segmentation algorithms on explanation quality, this study employs two configurations for each algorithm: the default setting and an optimized version. This decision is pivotal for the effective application of the LIME method, extending beyond the standard Quickshift algorithm to incorporate Felzenszwalb and SLIC in both their standard and optimized forms. These algorithms are specifically chosen for their unique properties, significantly contributing to the efficacy of LIME, particularly in the complex domain of medical imaging. The focus is to ascertain how the performance of these segmentation algorithms influences the quality of explanations.

The optimization process for each segmentation algorithm starts with a grid search over 50 random samples from the VinDr-CXR dataset, aimed at identifying a range of potentially effective parameter settings. Subsequently, an analysis oriented toward the general demands of medical imaging identifies optimal parameters for each algorithm:

- For the optimized Felzenszwalb algorithm, a scale of 600 is selected to balance between segment size and detail; a lower scale tends to over-segment the image. A minimum segment size of 200 prevents overly fragmented results, and a sigma of 0.2 smoothens the image for clearer segmentation boundaries.
- The optimized SLIC algorithm uses an `n_segments` value of 50 to determine the approximate number of equally sized superpixels, striking a balance between detail capture and computational efficiency. A compactness of 80 leads to more square-shaped superpixels, aiding in uniformity, and a sigma of 20 reduces image noise, crucial in medical imaging.
- For the optimized Quickshift algorithm, a `kernel_size` of 4 captures finer details in chest X-rays. The `max_dist` of 100 dictates the cut-off distance for merging clusters, allowing for effective grouping of pixels. A ratio of 0.6

ensures a balance between color-space proximity and image-space proximity in forming superpixels, and a sigma of 0.3 aids in creating well-defined superpixels.

These optimized parameters for each segmentation algorithm, detailed in Table 3, are selected to enhance the LIME method's ability to generate precise and interpretable explanations for thoracic disease classification.

The quantitative analysis evaluates the impact of these optimized segmentation algorithms and LIME settings on explanation quality. Conducted on a random sample of 2000 images from the VinDr-CXR dataset, this analysis balances statistical significance with computational manageability. It provides a robust dataset large enough to yield meaningful insights into the efficiency of various segmentation algorithms in enhancing LIME-generated explanations.

5.2 Qualitative analysis

5.2.1 Segmentation algorithms exploration

Following the detailed explanation of the optimization process and parameter settings for the segmentation algorithms, the impact of these settings is visually demonstrated through specific case studies. Figures 4 and 5 showcase the practical effects of both default and optimized settings on the segmentation algorithms. These figures also compare the explanations generated under various settings to the ground truth annotations, highlighting the influence of superpixel algorithmic adjustments.

In Fig. 4, the brown rectangles represent radiologists' annotations for detected lung consolidation conditions, while the orange rectangles illustrate explanations generated by LIME. The first row of the figure shows explanations generated using default parameter settings for each superpixel algorithm. The second row presents

explanations for the same segmentation algorithms with optimized parameter values. By comparing each figure in the top row with its corresponding one in the bottom row, the impact of optimization and parameter tuning on each segmentation algorithm becomes evident. Moreover, better *coverage* of optimized SLIC and optimized Felzenszwalb compared to optimized Quickshift is also demonstrated.

Figure 5 similarly illustrates a case study with multiple annotations. Dark blue rectangles represent radiologists' annotations for detected lung opacity conditions, while light blue rectangles display explanations generated by LIME. The impact of parameter tuning on superpixel algorithms in terms of segmentation accuracy and efficiency is evident in Fig. 5, showcasing its effectiveness even in multi-label scenarios.

5.2.2 Evaluation metric exploration

Regardless of the outcomes of the segmentation algorithms and their impact on the explanation quality and accuracy, this section visually demonstrates and investigates the behavior of the metrics introduced previously (as per the criteria defined in Sect. 4.2.1). This analysis is conducted through samples shown in Figs. 6 and 7, with corresponding results presented in Tables 4 and 5, respectively. In both figures, darker rectangles indicate radiologists' annotations for detected conditions, while lighter ones outline the borders of explanations (yellow-bordered area) generated by LIME. Each row in these figures presents a different case with three distinct segmentation settings. The left-most examples in each row depict more desirable results, whereas the right-most ones illustrate less desirable explanations.

Informativeness: Intersection-based metrics are limited in their ability to measure distance or position between two regions in some cases. As an example, when there is a complete separation between explanation and ground truth bounding boxes, metrics such as IOU, PWP, and PWR yield a value of zero, as illustrated in Fig. 6 and Table 4. Hence, intersection-based calculations sometimes lack *informativeness*.

Localization: A comparison of Fig. 6a–c illustrates what constitutes a better *localization* measurement. When there is no overlap between explanations and ground truth regions, an ideal metric should assign higher values to regions that are in closer proximity. Despite the larger false positive areas in the explanations shown in Fig. 6b, these are more informative than those in Fig. 6c, as at least one of the regions is nearer to the ground truth annotations, as detailed in Table 4. Intersection-based metrics like IOU, PWP, and PWR are not equipped to assess such scenarios. Moreover, metrics like MAE and PWA account for false

Table 3 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

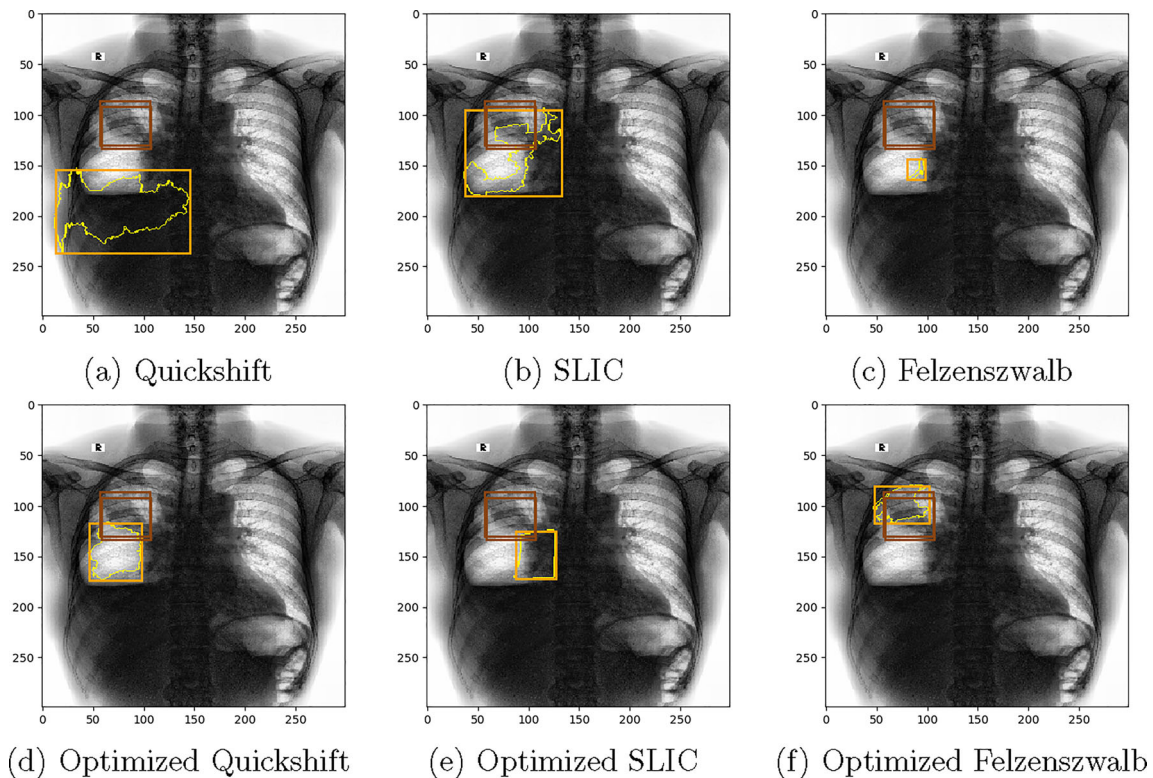


Fig. 4 The first row displays explanations generated by the segmentation algorithm with default parameters, and the second row displays explanations generated for the same segmentation algorithms with the optimized parameter values. The brown (darker) rectangles show the

radiologists' annotations for detected lung consolidation, and the orange (lighter) rectangles show the border of explanations (yellow-bordered area) generated by LIME (Color figure online)

positives and false negatives but do not consider their spatial positioning, leading to limitations in accurately evaluating complex cases as shown in Fig. 6b and c (see Table 4).

Coverage: An effective metric should assign higher values to cases with maximum intersection between explanations and ground truth masks, and minimal separation between these regions. The limitations of PWP and PWR metrics in measuring *coverage* are evident in several cases in Fig. 7. Each row in Fig. 7 presents a different case with three distinct segmentation settings. The left-most examples in each row depict more desirable results, while the right-most ones represent less desirable explanations. The shortcoming of PWR in evaluating optimal *coverage* is evident in Fig. 7a–f. PWR struggles to assess better *coverage*, particularly when there is a trade-off between higher true positives and lower false positives (Fig. 7b, c and Table 5). For instance, Fig. 7b shows a smaller true positive area than Fig. 7c, but its explanation is more accurate, as it marks a specific region closer to the expert's annotation. Furthermore, PWR reaches its highest value of one when the explanation region completely encircles the ground truth bounding box, as seen in Fig. 7d–f, where different explanations generated by various segmentations

appear identical in PWR terms. Conversely, PWP does not accurately measure *coverage* when explanation masks are entirely within ground truth annotations. As shown in the third row of Fig. 7, the most accurate explanation in Fig. 7g receives a lower PWP value, whereas the least accurate one in Fig. 7i scores the highest PWP value of one (refer to Table 5).

Multi-target capturing: This study aims to provide an explainable classifier for multi-label data. With pixel wise calculations, the metrics mentioned do not violate the *multi-target capturing* criterion, even when multiple regions are marked as explanation areas.

Proportionality: There should be a positive correlation between the distances of ground truth and explanation masks and the values of the metrics, reflecting the situation proportionally. Metrics like MAE and PWA exemplify a violation of this condition by averaging the difference across the total number of pixels. Consequently, significant differences in a single sample may result in only minimal changes in their values, even when there are notable variances between those cases. For instance, despite the significant differences between Fig. 7g (the most desirable explanation) and Fig. 7i (one of the least favorable), there

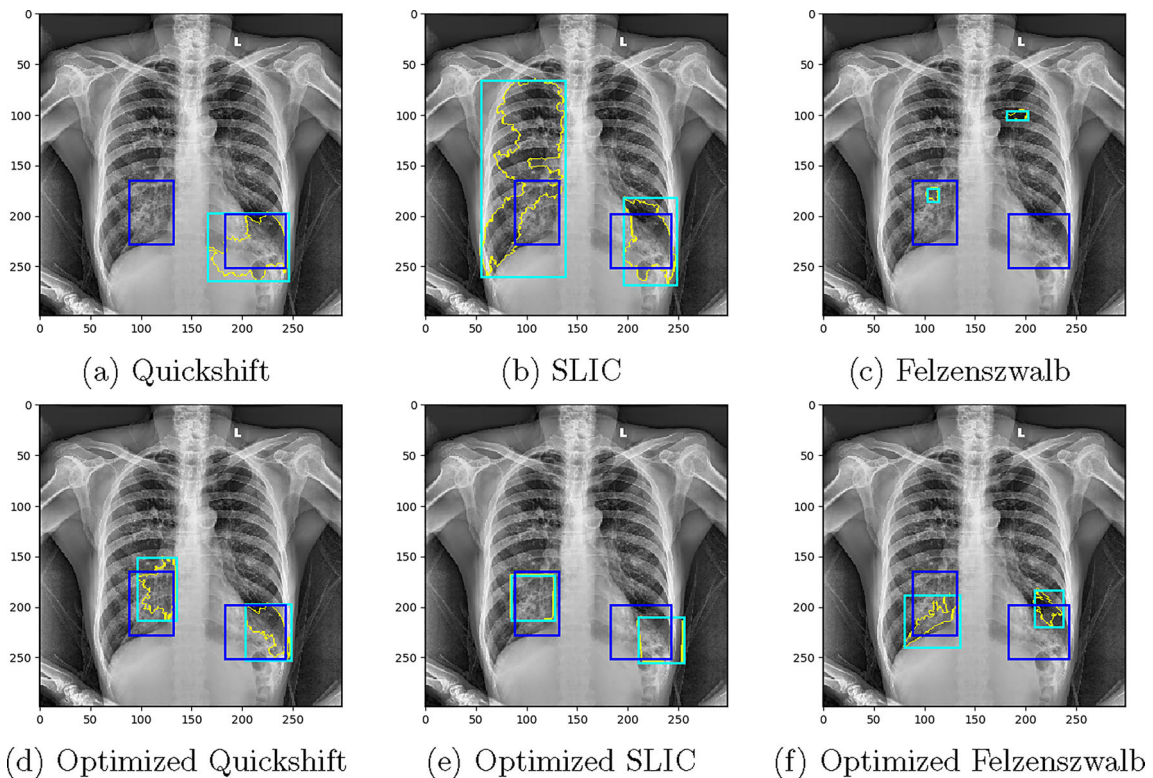


Fig. 5 The first row displays explanations generated by the segmentation algorithm with default parameters, and the second row displays explanations generated for the same segmentation algorithms with the optimized parameter values. The dark blue rectangles show the

radiologists' annotations for the detected lung opacity condition, and the light blue rectangles show the border of explanations (yellow-bordered area) generated by LIME (Color figure online)

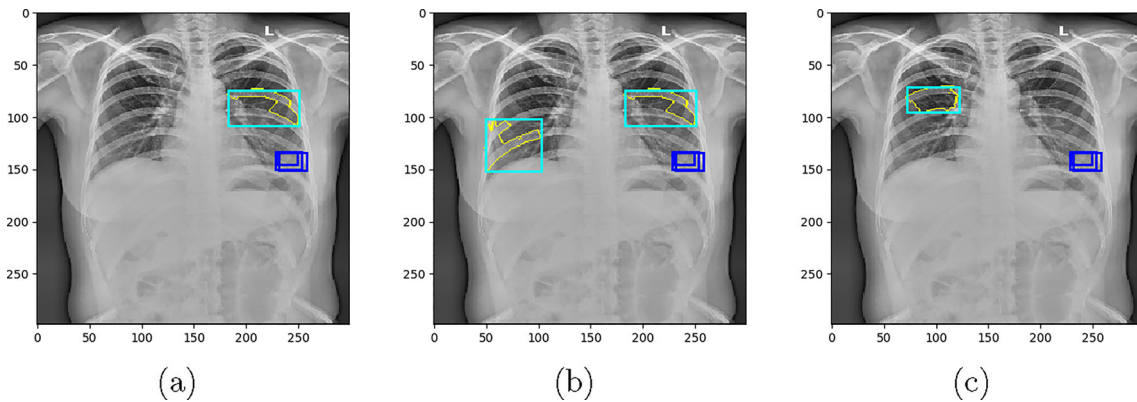


Fig. 6 Examples that violate *informativeness* and *localization* are shown for a single case with different segmentation algorithms. The dark blue rectangles show the radiologists' annotations for the

detected lung opacity condition, and the light blue rectangles show the border of explanations (yellow-bordered area) generated by LIME (Color figure online)

are only negligible variations in values reported by either MAE or PWA for both instances (as seen in Table 5). Furthermore, in situations with no overlap between explanation and ground truth annotations, intersection-based metrics (IOU, PWP, and PWR) fail to represent *proportionality*, as they yield zero as their final value.

5.3 Quantitative analysis

To evaluate the impact of various segmentation algorithms on LIME explanations, a comprehensive analysis is conducted using a random set of 2000 images from the VinDr-CXR dataset. For each variation of LIME, the study assesses the effectiveness using the JS_DIV metric, calculated for the top three predicted classes. The analysis

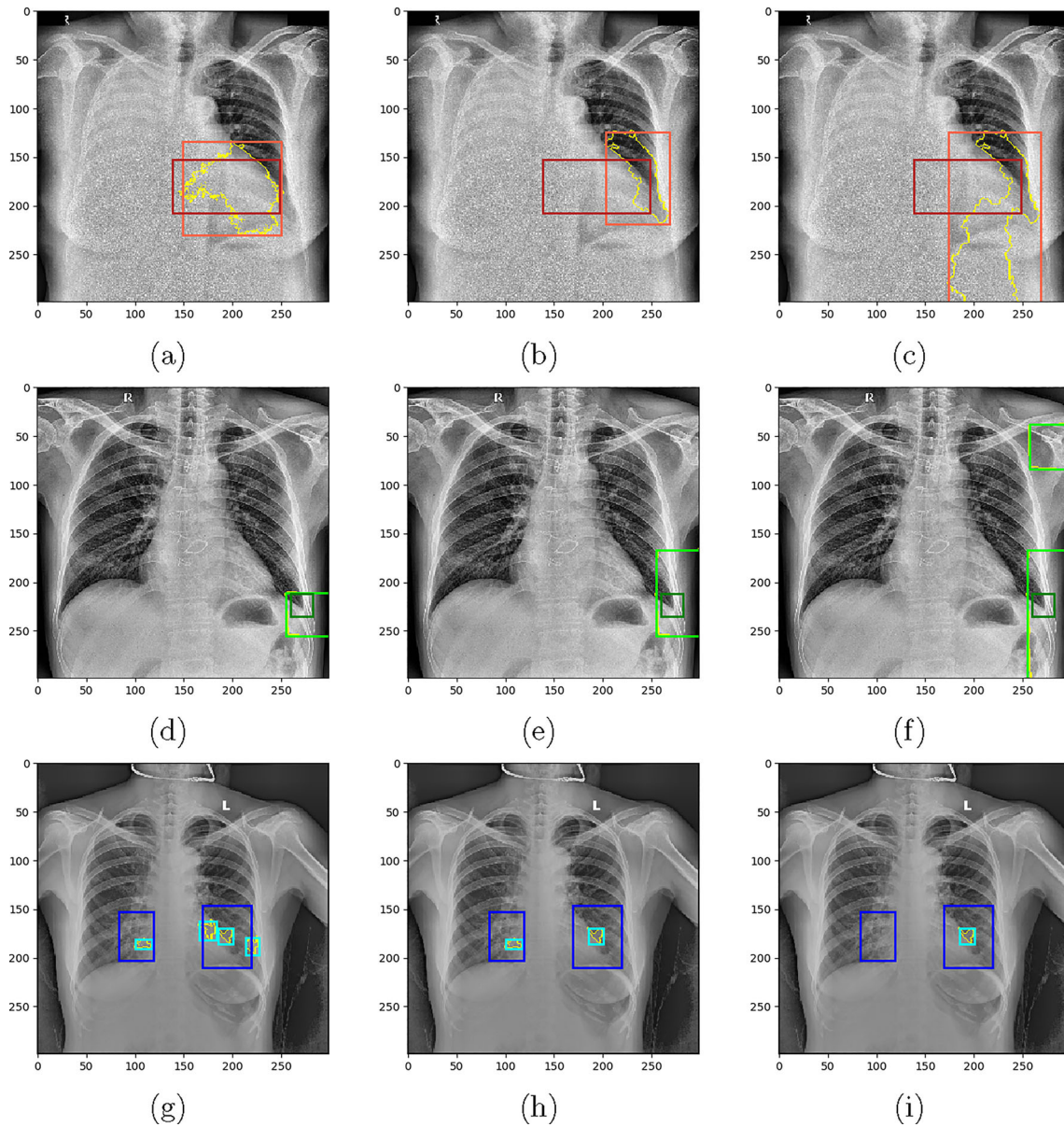


Fig. 7 Examples that violate *coverage* and *proportionality* are shown for a single case with different segmentation algorithms. The darker orange, green, and blue rectangles show the radiologists’ annotations for the detected cardiomegaly, pleural effusion, and lung opacity conditions. The lighter 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 (Color figure online)

Table 4 Metric results for the examples violate *informativeness* and *localization* shown in Fig. 6

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

spans multiple feature settings, and the performance of each segmentation algorithm is evaluated by averaging the metric values across feature settings 1, 2, and 4, to gain a comprehensive understanding of their impact. The results

of this analysis, presented in Table 6, enable the identification of the best segmentation configuration for each prediction scenario.

Table 5 Metric results for the examples violate *coverage* and *proportionality* shown in Fig. 7

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

Table 6 Comparison of the LIME method with various segmentation algorithms for the first three predicted classes

	Segmentation algorithm	JS_DIV
1st prediction	Optimized SLIC	0.558
	Default SLIC	0.576
	Optimized Felzenszwalb	0.559
	Default Felzenszwalb	0.565
	Optimized Quickshift	0.607
	Default Quickshift	0.619
2nd prediction	Optimized SLIC	0.562
	Default SLIC	0.582
	Optimized Felzenszwalb	0.565
	Default Felzenszwalb	0.568
	Optimized Quickshift	0.615
	Default Quickshift	0.629
3rd prediction	Optimized SLIC	0.603
	Default SLIC	0.612
	Optimized Felzenszwalb	0.607
	Default Felzenszwalb	0.608
	Optimized Quickshift	0.624
	Default Quickshift	0.635

Based on the JS_DIV value, identified as the most suitable metric, it is evident that LIME employing optimized SLIC segmentation consistently demonstrates superior results across all predictions. This consistency suggests that the SLIC algorithm, particularly in its optimized form, is well-suited for tasks in medical imaging, offering a balance of precision and interpretability crucial in this field. The performance of optimized Felzenszwalb and its default form, ranking as the second and third best, respectively, also highlights the significant impact of parameter optimization on algorithm efficiency.

An interesting observation is the decreasing JS_DIV value from the first to the third prediction. This trend reflects the relation between classifier confidence and the quality of explanations; as the classifier's confidence

diminishes in its subsequent predictions, the quality of explanations, as measured by JS_DIV, also tends to decrease. This aspect underscores the importance of classifier reliability in the context of explainable AI, particularly in medical diagnostics where decision-making clarity is paramount.

The performance trends of these algorithms, especially the contrast between the optimized Quickshift and other algorithms, point toward nuanced considerations in selecting and tuning segmentation algorithms for specific tasks. The study's findings indicate that while parameter optimization generally improves algorithm performance, its effectiveness varies with the algorithm's inherent characteristics and the nature of the task.

Reflecting on these results within the broader scope of medical image analysis, the choice of segmentation algorithms and their configuration plays a pivotal role in the practical application of AI models. These insights can guide future algorithm selections, contributing to more reliable and interpretable deep learning tools in healthcare.

As the study concludes, it is essential to acknowledge its methodological complexities and the challenges encountered. The process of systematic selection and rigorous assessment of various metrics against the introduced criteria was integral to enhancing the transparency and interpretability of AI models in medical diagnostics. This study's approach, from an extensive literature survey to the development of a novel evaluation framework, illustrates the blend of technical expertise and innovation required in advancing the field of XAI. Future work could focus on exploring other segmentation algorithms, experimenting with different datasets, or employing alternative evaluation metrics to further this research domain.

6 Conclusions and future work

In this study, significant advancements in XAI for medical imaging, especially in thoracic disease classification, have been achieved. A comprehensive framework for assessing

visual explanation quality in deep learning models has been developed, establishing novel criteria like *informativeness*, *localization*, *coverage*, *multi-target capturing*, and *proportionality*, and identifying JS-DIV as the most effective metric. Both qualitative and quantitative analyses have been conducted, offering insights into the behavior of superpixel segmentation algorithms in LIME, exploring various metrics, and emphasizing the significance of metric selection for enhanced interpretability of AI models. The research has revealed limitations of popular metrics such as IOU, MAE, and PWP (IoSR) in fully satisfying established criteria, highlighting their inadequacy in capturing key aspects like *informativeness*, *localization*, and *coverage*. The study also uncovers LIME's dependency on segmentation parameters and the correlation between classifier confidence and the quality of explanations. In conclusion, this research advances the transparency and interpretability of AI models in medical diagnostics, paving the way for more reliable and interpretable deep learning tools in healthcare. Future work will focus on developing advanced visualization techniques for more accessible AI insights and patient-centric models for personalized medical decision explanations. Additionally, the study's methodologies will be adapted for broader applications in sectors like finance and autonomous vehicles, testing their effectiveness across various scenarios.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability The CheXpert dataset [35] analyzed during the current study is a public chest X-ray dataset available at <https://stanfordaimi.azurewebsites.net/datasets/8cbd9ed4-2eb9-4565-afc-111cf4f7ebe2> upon request and with the permission and under the license of Stanford University Dataset Research Use Agreement. Similarly, the VinDr-CXR dataset [36] that supports the findings of this study can be accessed through <https://physionet.org/content/vindr-cxr/1.0.0/> in case of fulfilling the requirements.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript. This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this

article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Dong S, Wang P, Abbas K (2021) A survey on deep learning and its applications. *Comput Sci Rev* 40:100379. <https://doi.org/10.1016/j.cosrev.2021.100379>
- Prasanna DL, Tripathi SL (2023) Machine and deep learning techniques for text and speech processing. In: Ghai D, Tripathi SL, Saxena S, Chanda M, Alazab M (eds) *Machine learning algorithms for signal and image processing*. Wiley, New York, pp 115–128. <https://doi.org/10.1002/9781119861850.ch7>
- Collette J, Atkinson K, Bench-Capon T (2023) Explainable AI tools for legal reasoning about cases: a study on the European Court of Human Rights. *Artif Intell* 317:103861. <https://doi.org/10.1016/j.artint.2023.103861>
- Giudici P, Raffinetti E (2022) Explainable AI methods in cyber risk management. *Qual Reliab Eng Int* 38(3):1318–1326. <https://doi.org/10.1002/qre.2939>
- 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 Mech Dis*. <https://doi.org/10.1002/wsbm.1548>
- Eschenbach WJ (2021) Transparency and the black box problem: why we do not trust AI. *Philos Technol* 34(4):1607–1622. <https://doi.org/10.1007/s13347-021-00477-0>
- Fourcade A, Khonsari RH (2019) Deep learning in medical image analysis: a third eye for doctors. *J Stomatol Oral Maxillofac Surg* 120(4):279–288. <https://doi.org/10.1016/j.jormas.2019.06.002>
- Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med Image Anal* 79:102470. <https://doi.org/10.1016/j.media.2022.102470>
- 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 artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Sahlol AT, Yousri D, Ewees AA, Al-Qaness MAA, Damasevicius R, Elaziz MA (2020) Covid-19 image classification using deep features and fractional-order marine predators algorithm. *Sci Rep* 10(1):15364. <https://doi.org/10.1038/s41598-020-71294-2>
- Yousri D, Abd Elaziz M, Abualigah L, Oliva D, Al-qaness MAA, Ewees AA (2021) Covid-19 x-ray images classification based on enhanced fractional-order cuckoo search optimizer using heavy-tailed distributions. *Appl Soft Comput* 101:107052. <https://doi.org/10.1016/j.asoc.2020.107052>
- Elaziz MA, Ewees AA, Yousri D, Alwerfali HSN, Awad QA, Lu S, Al-Qaness MAA (2020) An improved marine predators algorithm with fuzzy entropy for multi-level thresholding: real world example of covid-19 CT image segmentation. *IEEE Access*

- 8:125306–125330. <https://doi.org/10.1109/ACCESS.2020.3007928>
13. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144. <https://doi.org/10.1145/2939672.2939778>
 14. Vedaldi A, Soatto S (2008) Quick shift and kernel methods for mode seeking. In: European conference on computer vision, pp 705–718. https://doi.org/10.1007/978-3-540-88693-8_52. Springer
 15. Xiang A, Wang F (2019) Towards interpretable skin lesion classification with deep learning models. *AMIA Annu Symp Proc* 2019:1246–1255
 16. 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. *Appl Sci (Switz)*. <https://doi.org/10.3390/app8101715>
 17. Ahsan MM, Gupta KD, Islam MM, Sen S, Rahman ML, Hossain MS (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/10.3390/make2040027>
 18. Teixeira LO, Pereira RM, Bertolini D, Oliveira LS, Nanni L, Cavalcanti GDC, Costa YMG (2021) Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors*. <https://doi.org/10.3390/s21217116>
 19. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015. https://doi.org/10.1007/978-3-319-24574-4_28
 20. Sattarzadeh S, Sudhakar M, Lem A, Mehryar S, Plataniotis KN, 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. In: 34th AAAI Conference on artificial intelligence
 21. DoshiVelez F, Kim B (2018) Considerations for evaluation and generalization in interpretable machine learning. In: Explainable and interpretable models in computer vision and machine learning, pp 3–17. https://doi.org/10.1007/978-3-319-98131-4_1
 22. Li X-H, Shi Y, Li H, Bai W, Cao CC, Chen L (2021) An experimental study of quantitative evaluations on saliency methods. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining, pp 3200–3208. <https://doi.org/10.1145/3447548.3467148>
 23. Alvarez Melis D, Jaakkola T (2018) Towards robust interpretability with self-explaining neural networks. In: Advances in neural information processing systems, pp 31
 24. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter conference on applications of computer vision (WACV). IEEE, pp 839–847. <https://doi.org/10.1109/WACV.2018.00097>
 25. Ramaswamy HG, Desai S (2020) Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization. In: 2020 IEEE winter conference on applications of computer vision (WACV), pp 972–980. <https://doi.org/10.1109/WACV45572.2020.9093360>
 26. Petsiuk V, Das A, Saenko K (2018) RISE: randomized input sampling for explanation of black-box models
 27. Sokol K, Flach P (2020) Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 56–67. <https://doi.org/10.1145/3351095.3372870>
 28. Hailemariam Y, Yazdinejad A, Parizi RM, Srivastava G, Dehghantanha A (2020) An empirical evaluation of AI deep explainable tools. In: 2020 IEEE Globecom workshops (GC Wkshps). IEEE, pp 1–6. <https://doi.org/10.1109/GCWkshps50303.2020.9367541>
 29. Graziani M, Lompech T, Müller H, Andrearczyk V (2020) Evaluation and comparison of CNN visual explanations for histopathology. In: Explainable agency in artificial intelligence at AAAI21, pp 195–201
 30. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S (2018) Top-down neural attention by excitation backprop. *Int J Comput Vis* 126(10):1084–1102. <https://doi.org/10.1007/s11263-017-1059-x>
 31. Schulz K, Sixt L, Tombari F, Landgraf T (2020) Restricting the flow: information bottlenecks for attribution. In: International conference on learning representations
 32. 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. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 111–119. <https://doi.org/10.1109/CVPRW50498.2020.00020>
 33. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer, pp 818–833
 34. Schallner L, Rabold J, Scholz O, Schmid U (2020) Effect of superpixel aggregation on explanations in LIME—a case study with biological data. In: Cellier P, Driessens K (eds) *Mach Learn Knowl Discov Databases*. Springer, Cham, pp 147–158
 35. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghighi 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. *Proc AAAI Conf Artif Intell* 33:590–597. <https://doi.org/10.1609/aaai.v33i01.3301590>
 36. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, Le DD, Pham CM, Tong HTT, Dinh DH, Do CD, Doan LT, Nguyen CN, Nguyen BT, Nguyen QV, Hoang AD, Phan HN, Nguyen AT, Ho PH, Ngo DT, Nguyen NT, Nguyen NT, Dao M, Vu V (2020) VinDr-CXR: an open dataset of chest X-rays with radiologist’s annotations, 1–10
 37. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
 38. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708. <https://doi.org/10.1109/CVPR.2017.243>
 39. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: 3rd International conference on learning representations, ICLR
 40. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778. <https://doi.org/10.1109/CVPR.2016.90>
 41. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern

- recognition, pp 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
42. Längkvist M, Karlsson L, Loutfi A (2014) A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit Lett* 42:11–24. <https://doi.org/10.1016/j.patrec.2014.01.008>
 43. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. *Int J Comput Vis* 59(2):167–181. <https://doi.org/10.1023/B:VISI.0000022288.19776.77>
 44. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans Pattern Anal Mach Intell* 34(11):2274–2282. <https://doi.org/10.1109/TPAMI.2012.120>
 45. Mohseni S, Block JE, Ragan E (2021) Quantitative evaluation of machine learning explanations: a human-grounded benchmark. In: 26th International conference on intelligent user interfaces, pp 22–31. <https://doi.org/10.1145/3397481.3450689>
 46. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R (2019) Explainable AI: interpreting, explaining and visualizing deep. Learning. <https://doi.org/10.1007/978-3-030-28954-6>
 47. Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F (2019) What do different evaluation metrics tell us about saliency models? *IEEE Trans Pattern Anal Mach Intell* 41(3):740–757. <https://doi.org/10.1109/TPAMI.2018.2815601>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.