医学诊断如何可解释？贝拉内大学最新《医学诊断中可解释深度学习方法》综述，36页pdf153篇文献概述最新XAI医学诊断进展

原创 专知 专知 2022-05-14 18:26 发表于北京

葡萄牙贝拉内大学最新《医学诊断中可解释深度学习方法》综述，值得关注！

Explainable Deep Learning Methods in Medical Diagnosis: A Survey

CRISTIANO PATRÍCIO and JOÃO C. NEVES, University of Beira Interior, Portugal
LUÍS F. TEIXEIRA, University of Porto, Portugal

The remarkable success of deep learning has prompted interest in its application to medical diagnosis. Even tough state-of-the-art deep learning models have achieved human-level accuracy on the classification of different types of medical data, these models are hardly adopted in clinical workflows, mainly due to their lack of interpretability. The black-box-ness of deep learning models has raised the need for devising strategies to explain the decision process of these models, leading to the creation of the topic of eXplainable Artificial Intelligence (XAI). In this context, we provide a thorough survey of XAI applied to medical diagnosis, including visual, textual, and example-based explanation methods. Moreover, this work reviews the existing medical imaging datasets and the existing metrics for evaluating the quality of the explanations . Complementary to most existing surveys, we include a performance comparison among a set of report generation-based methods. Finally, the major challenges in applying XAI to medical imaging are also discussed.

CCS Concepts: • **Applied computing → Health care information systems**.

Additional Key Words and Phrases: Explainable AI, Explainability, Interpretability, Deep Learning, Medical Image Analysis

ACM Reference Format:
Cristiano Patrício, João C. Neves, and Luís F. Teixeira. 2022. Explainable Deep Learning Methods in Medical Diagnosis: A Survey. 1, 1 (May 2022), 36 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

深度学习的显著成功引发了人们对其在医学诊断中的应用的兴趣。即使最先进的深度学习模型在对不同类型的医疗数据进行分类时达到了人类水平的准确性，但这些模型在临床工作流程中很难被采用，主要是因为它们缺乏可解释性。**深度学习模型的黑盒性提出了设计策略来解释这些模型的决策过程的需要，这导致了可解释人工智能(XAI)这个话题的产生。在此背景下，我们提供了XAI应用于医疗诊断的全面综述，包括可视化、文本和基于示例的解释方法。**此外，这项工作回顾了现有的医学成像数据集和现有的指标，以评估解释的质量。作为对大多数现有综述的补充，我们包含了一组基于报告生成方法之间的性能比较。最后，还讨论了XAI在医学影像应用中的主要挑战。

> https://www.zhuanzhi.ai/paper/f6e90091666dbcaa5b40c1ab82e9703b

引言

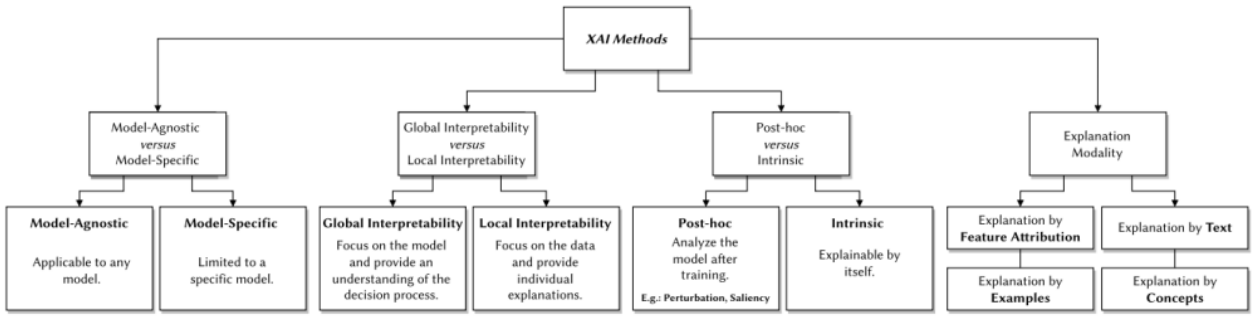
人工智能(AI)领域在过去十年取得的进展，支持了大多数计算机视觉应用的准确性的显著提高。医学图像分析是在对不同类型的医学数据(如胸部X光片[80]、角膜图像[147])进行分类时取得人类水平精确度的应用之一。然而，尽管有这些进展，自动化医学成像在临床实践中很少被采用。Zachary Lipton[69]认为，对这一明显的悖论的解释很简单，医生在不了解决策过程的情况下，永远不会相信算法的决策。**这一事实提出了产生能够解释人工智能算法的决策过程的策略的必要性，随后导致了一个新的研究主题的创建，称为可解释人工智能(XAI)。**根据DARPA[41]的说法，XAI的目标是“在保持高水平的学习性能(预测精度)的同时，产生更多可解释的模型;并使人类用户能够理解、适当、信任和有效地管理新一代人工智能伙伴”。尽管XAI具有普遍适用性，但它在高风险决策(如临床工作流程)中尤其重要，在这种情况下，错误决策的后果可能导致人类死亡。这也得到了欧盟通用数据保护条例(GDPR)法律的证明，该法律要求解释算法的决策过程，使其透明，然后才能用于患者护理[37]。

因此，在将深度学习方法应用于临床实践之前，研究新的策略以提高其可解释性是至关重要的。近年来，对这一课题的研究主要集中在设计间接分析预建模型决策过程的方法。这些方法要么分析输入图像的特定区域对最终预测的影响(基于扰动的方法[77;101]和基于遮挡的方法[151])或检查网络激活(显著性方法[112;153])。这些方法可以应用于任意网络架构，而不需要对模型进行额外的定制，这一事实支持了它们在XAI早期的流行。然而，最近的研究表明，事后策略在解释的重要性方面存在一些缺陷[2;105]。因此，研究人员将他们的注意力集中在能够解释其决策过程本身的模型/架构的设计上。现有的可解释模型被认为在医学成像中特别有用[105]，证明了最近集中于这一范式而不是传统的后特殊策略的医学成像作品数量的增长是合理的[53;144]。尽管近年来固有可解释模型的流行，但现有的关于深度学习应用于医学成像的可解释性的研究并没有全面回顾这一新的研究趋势的进展。此外，专注于解释应用于医学成像的深度学习决策过程的著作数量显著增加，因此有必要对最近一次关于该主题的综述未涵盖的最新方法进行更新调研。

为了解决这些问题，我们全面回顾了可解释深度学习应用于医学诊断的最新进展。特别是，这项综述提供了以下贡献:

- 回顾最近关于医学成像中可解释深度学习主题的调研，包括从每个工作中得出的主要结论，以及对我们调研的比较分析。
- 用于医学成像的深度学习方法可解释性研究中常用的数据集的详尽列表。
- 全面调研最先进的可解释医学成像方法，包括事后模型和固有的可解释模型。

- 对基准可解释性方法常用的度量标准的完整描述，无论是可视化的还是文本的解释。关于文本解释质量的可解释医学成像方法的基准。
- 医学影像中可解释深度学习的未来研究方向



基于文献综述，XAI方法可以根据三个标准进行分类: (i) 模型无关性vs模型具体; （ii）全局可释性与局部可释性; (iii)事后对内在。图1说明了XAI方法的分类法，

医疗诊断中的可解释人工智能方法

正如前面提到的，深度学习模型在部署到现实场景时必须具有透明性和可信赖性。此外，这一要求在临床实践中尤其相关，在临床实践中，不知情的决定可能会将患者的生命置于危险之中。在综述的文献中，已经提出了几种方法来赋予应用于医学诊断的深度学习方法解释性。以下部分总结和分类了应用于医学诊断的可解释模型范围内最相关的工作。此外，我们特别关注内在可解释的神经网络及其在医学成像中的适用性。我们根据解释方式将这些方法分为:(i)特征归因解释，(ii)文本解释，(iii)实例解释，(iv)概念解释，(v)其他解释;受[86]提出的分类学启发。根据所使用的算法、图像形态和数据集分类的综述方法列表见表4。

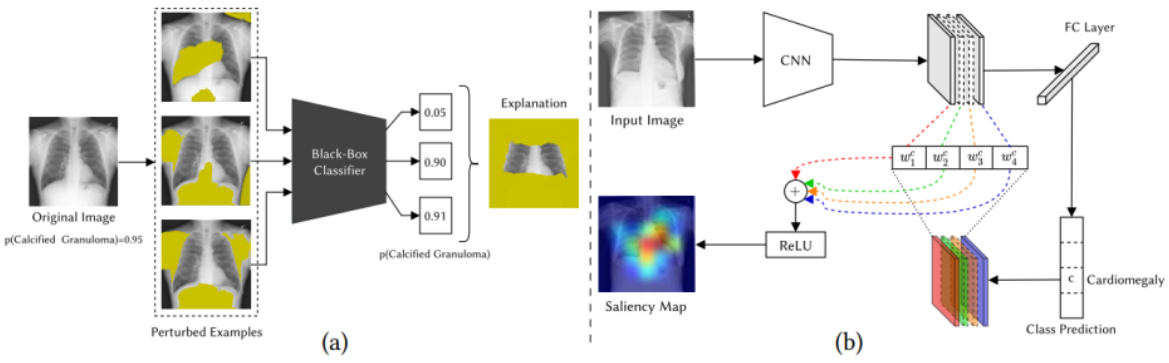


Fig. 3. (a) **Perturbation-based methods**. The input image is randomly perturbed by turning on and off certain pixels, resulting in an image with occluded parts (*Perturbed Examples* in the figure). Then, the perturbed image is fed to the classification model and the prediction confidence is exploited to determine the regions that contributed positively to the class prediction. Therefore, those regions will be considered to obtain the final explanation map (*Explanation* in the figure). (b) **Gradient-based methods**. The input image is fed to the classification model to obtain a class prediction. Then, the gradient is calculated for the score of the class concerning the feature maps of the last convolutional layer. After calculating the importance of the feature map regarding the predicted class, they are weighted with each of respective weight, followed by a ReLU operation to obtain the final saliency map.

阅读原文