Title: Explainable AI for medical imaging: deep learning CNN ensemble for classification of estrogen receptor status from breast MRI.

Paper link: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11314/2549298/Explainable-AI-for-medical-imaging--deep-learning-CNN-ensemble/10.1117/12.2549298.full

Summary:

ABSTRACT

We applied explainable artificial intelligence (XAI) techniques to a deep-learning convolutional neural network (DCNN) trained to classify estrogen receptor status (ER+ vs ER-) in breast images. The DCNN learned relevant features from the spatial and dynamic domains, but there were differences in the contributing features.

1. INTRODUCTION

Using deep-learning predictive models as black-box classifiers is problematic because models can exploit irrelevant patterns in the data to maximize the performance metrics. Using explainable AI (XAI) algorithms, we generated attribution maps for breast DCE-MRI data highlighting features contributing to the output score.

In this study, we trained a DCNN to classify estrogen receptor status from DCE-MRI breast volumes to aid in the molecular classification of breast cancer.

2.1 Data sets

A data set of T1-weighted DCE-MRI scans of 148 patients with stage 2 and 3 invasive ductal carcinoma undergoing neoadjuvant chemotherapy was used to train and test a DCNN.

2.2 Spatial and dynamic domains

We first excluded the pectoral muscle and background external to the breast, then applied White stripe normalization on the DCE-MRI volumes at peak contrast, and extracted the ROI from three time points of the original MRI scans without homogenization.

2.3 Dual-domain DCNN structure and transfer learning

We developed a dual-domain DCNN architecture to utilize the spatial and dynamic components of DCE-MRI volumes for classifying ER status. The model consisted of two identical DCNNs creating a two-input mode mirroring the design of AlexNet that won the ImageNet LSVRC-2012 contest.

2.4 Explainable AI

Attribution maps were generated using the backpropagation-based integrated gradients attribution method and the Smoothgrad algorithm. The integrated gradients attribution method assigns more influential pixels with higher values, thus connecting the complicated path from the predicted output score to the patterns observed at every layer of the DCNN.

3. RESULTS

The DCNN ensemble was trained on the classification task of ER status from DCE-MRI breast volumes across 6 imaging centers and obtained correct prediction scores >0.9.

3.1 Learning relevant features

In this section, we discuss the expected relevant features learned by the DCNN. We found that the DCNN distinguished the tumor from surrounding fatty, dense, and scattered tissue for both the spatial and dynamic ROI components.

3.2 Differences in learning between spatial and dynamic domains

We observed that the DCNN learned from fatty tissue surrounding the tumor in the spatial domain more frequently than in the dynamic domain, and that it distinguished the tumoral tissue from fatty and dense tissue in the dynamic domain more effectively than in the spatial domain.

3.3 Learning from irrelevant features from pre-processing

In this section, we discuss a possible flaw in the pre-processing of the ROIs that may have misguided the network's training. By eliminating these irrelevant features, it may be possible to improve the DCNN's ability to generalize to the test set.

4. DISCUSSION AND CONCLUSION

We applied explainable AI to deep learning on our medical imaging data set to identify artifacts and improve the model. We also presented a novel dual-domain DCNN that can be generalized to multi-domain and/or multi-modality for other medical image analysis tasks.

The future direction of this work should aim to remove the preprocessing artifacts of the breast ROIs, re-train the model, and deploy it to determine if the performance improves.

This work focuses on understanding what the DCNN learns behind its output decision, which may help in the development of methods to minimize corruption in their learning and advance AI toward intelligent decision support tools in medicine.