ENSURING CONFIDENCE IN NEURAL NETWORK DECISIONS IN MEDICAL DIAGNOSTICS BASED ON VISUAL DATA

E. Manziuk, T. Skrypnyk, T. Lukmanov, O. Kyrychenko
Khmelnytskyi National University, Ukraine

Abstract. The paper presents a novel method for medical image analysis that combines the high accuracy of deep learning models and the interpretability of logical models. The proposed approach involves training a convolutional neural network (CNN) for accurate image classification, applying a spatial attention mechanism to localize important features, and constructing an interpretable Decision Rule Network (DRN) based on these features. The DRN is a set of logical rules linking feature values to diagnoses, allowing for transparent decision-making.
The method was evaluated on brain MRI scans, achieving high accuracy with the CNN (>95%) and interpretability with the DRN. The authors emphasize the importance of achieving consistency between the CNN and DRN decisions for specific clinical cases, ensuring trust and compliance with ethical and regulatory requirements in medical AI applications.
Keywords: medical image analysis, convolutional neural networks, explainable AI, Decision Rule Network, interpretability.

Introduction

Accelerating development of information systems is accompanied by the expansion of their practical applications, including the integration of intelligent systems - both relatively simple algorithmic decision-making systems and artificial intelligence (AI) systems.

The use of intelligent systems in practical tasks required the development of normative documents to regulate the requirements and limitations of AI application in order to ensure safety, prevent harm, etc. The General Data Protection Regulation (GDPR) and the EU Ethical Principles for the management, development and use of AI have been adopted. They recognize the user's right to receive an explanation of decisions made by AI systems.

Medical AI systems are one of the areas of practical application that are subject to all the necessary requirements of safety, reliability and criticality. AI in medicine is undoubtedly necessary and important, but its use should not be limited only to improving the efficiency of certain safe tasks. The medical field is an area of critical decision-making.

There are challenges to the use of AI in medicine:

1. Risk of poor decisions with serious consequences for the patient's health.
2. Deficiencies in error detection systems.
3. The need for interpretability for physicians and patients to understand the reasons for decision making.
4. The need to incentivize the development of interpretable machine learning models for medicine.
5. The issue of accountability for misdiagnoses or decisions.
6. Ethical issues and confidentiality.
7. The need for training and adaptation of AI systems.

The purpose of this work is to develop a method that will allow to identify and analyze the attributes that influence the decision-making of an AI system in the diagnosis of clinical diseases. This will help to move from abstract models to practical application of intelligent systems in medicine, taking into account ethical aspects, legal compliance and trust building.

Related Works

The problem of trust in AI systems has become relevant due to the acceleration of their practical implementation and has revealed new aspects that need to be addressed. In the medical field, trust has two important aspects: social and technical. From a social point of view, trust is

an important aspect of patient-physician interaction. The patient is in a vulnerable state and has to seek external help by trusting the doctor.

The use of AI systems introduces a new dimension to patient-physician interaction, which may weaken overall trust [1]. Despite high results, the level of trust in AI systems is low, so they cannot be considered reliable. In situations where patients have to rely on AI systems to make important medical decisions, this may lead to a decrease in trust in clinical practice [2].

A number of studies have focused on building AI systems that meet a set of trust requirements. The concept of trust based on ethical principles has been studied [3]. Metrics have been proposed for assessing trust in AI using the explanatory power of expert systems. Concerns have been raised about the potential dangers of "black box" algorithms, limiting the practical application of AI to medical aspects where transparency and interpretability of decisions are not required.

The literature review showed that despite active research in the field of explainable AI (XAI) and decision visualization of deep learning models for medical image analysis tasks, full interpretability of decision logic at the level of medical concepts and rules has not yet been achieved.

Thus, there is a need to develop new methods for medical image analysis that combine the high accuracy of pathology detection achieved by deep learning and the interpretability of decisions at the level of medical terms and rules.

Proposed Method

This paper proposes a method that combines the advantages of deep learning for accurate medical image analysis and logic models to ensure interpretability of decisions.

The method includes the following steps:

1. Build a complex convolutional neural network (e.g., VGG-16) for image analysis and pathology classification. This network provides high accuracy, but its decisions are difficult to interpret.

2. Applying the spatial attention mechanism to the output feature maps of the convolutional network. This allows to select key image regions important for classification.

3. Development of a simplified interpretable model based on selected features - Decision Rule Network (DRN). DRN is a logical model in the form of a set of interpretable rules linking feature values to diagnosis.

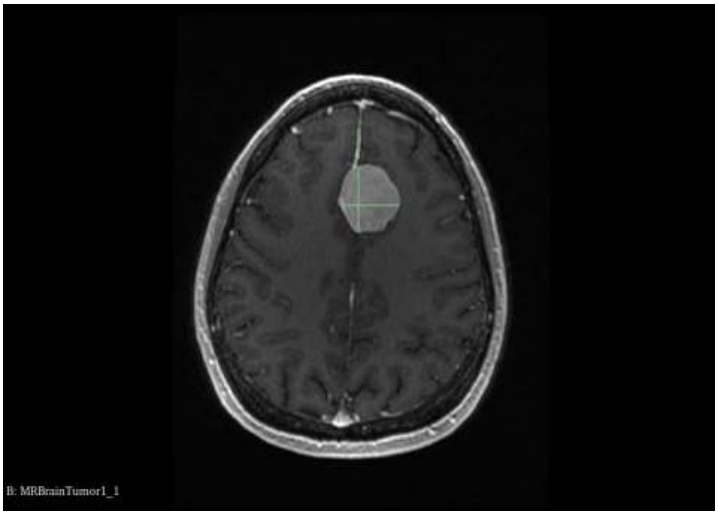

Fig. 1. A depiction of the human brain captured through magnetic resonance imaging (MRI)

Require local decision consistency between the complex and simplified models for specific clinical cases. This ensures consistency of conclusions and allows clinicians to understand the logic of decision making.

Iterations are performed to build an interpretable DRN classifier for a local clinical case:

1. Extracting deep features from training set using convolutional network.
2. Selection of relevant features for the current case.
3. Constructing DRNs on these features.

The cycle is repeated until the consistency of convolution network and DRN decisions for a given clinical case is achieved.

Experimental Results

The proposed method was applied to analyze MR images of the brain for pathologies. A publicly available dataset containing 3D MRI scans and corresponding diagnoses made by expert radiologists was used.

In the first stage, a VGG-16 convolutional neural network was trained on the dataset. The network demonstrated high accuracy of pathology classification - more than 95%. However, its decisions are difficult to interpret due to the large size and high complexity of the model.

An interpretable DRN model was then constructed based on the features extracted by VGG-16 using a spatial attention mechanism. DRN was a set of logical rules linking feature values to diagnoses in the form "If {condition}, then {diagnosis}".

The accuracy of DRN on the test set was 76%, which is lower than that of VGG-16, but its decisions were much more interpretable. Analysis of the extracted rules allowed us to identify features that play a key role in the diagnosis of various pathologies on MRI.

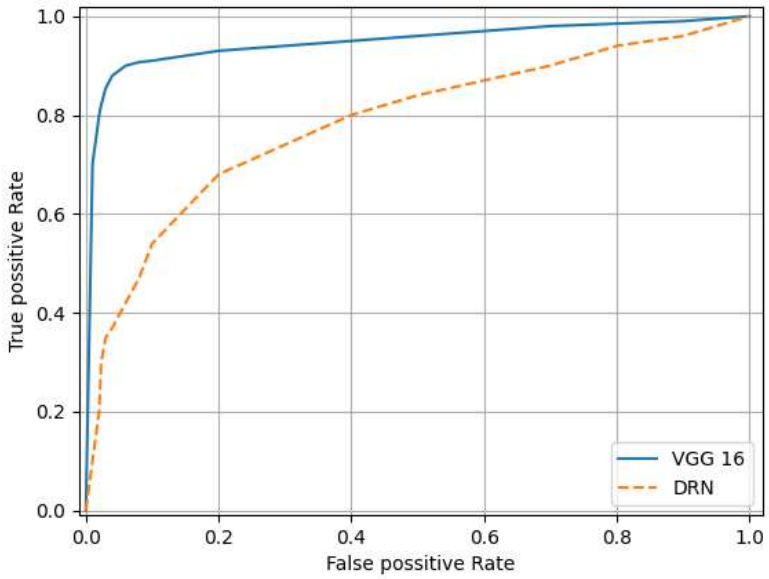

Fig. 2. ROC curve representing the classification performance of neural networks VGG-16 and DRN

For example, such signs as the presence of foci of increased intensity in periventricular regions, white matter structurelessness, loss of corpus callosum volume, etc. have proved to be important for the diagnosis of multiple sclerosis.

Visualization of these features on baseline MRI scans allowed expert clinicians to test the validity of the DRN findings and assess its consistency with clinical guidelines.

To ensure consistency between the accurate but uninterpretable VGG-16 and the simpler but explainable DRN, a requirement for local consistency of decisions for specific clinical cases was introduced. Through an iterative process of relevant feature sampling and DRN tweaking, a situation was reached where both models produced consistent diagnoses on images from a given case.

Thus, the proposed method has effectively combined the high accuracy of deep learning and interpretability of logic models in the task of medical image analysis. This may contribute to a safer and more responsible implementation of AI in clinical practice.

Conclusions

A novel approach to medical image analysis is presented that combines the advantages of deep learning for high-precision classification and logic models for interpretability of decisions. The key components are a convolutional neural network for extracting informative features, an attention mechanism for localizing these features, and an interpretable logic model in the form of a set of rules for explaining diagnoses.

Experimental results on the task of analyzing MRI brain scans demonstrate high accuracy of pathology classification by convolutional network (>95%) along with the ability to understand the reasons for decision making by visualizing important features and logical rules in the interpreted model.

A balance is achieved between the high performance of state-of-the-art deep learning technologies and the ability to explain the system's findings in a manner consistent with the principles of trust, fairness, privacy, and other ethical requirements for AI.

The proposed approach paves the way for the safe and responsible application of artificial intelligence technologies in the critical field of medicine, while fully complying with ethical and regulatory requirements.

References

1. Hatherley J.J.. Limits of trust in medical AI. Journal of Medical Ethics. 2020. Vol. 46, No. 7, P. 478-481. doi: 10.1136/medethics-2019-105935.
2. Durán J.M., Jongsma K.R.. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of Medical Ethics. 2021. Vol. 47, No. 5, P. 329-335. doi: 10.1136/medethics-2020-106820.
3. Hasani N., Morris M.A., Rahmim A. et al. Trustworthy Artificial Intelligence in Medical Imaging. PET Clinics. 2022. Vol. 17, No. 1, P. 1-12. doi: 10.1016/j.cpet.2021.09.007.