Chapter 2
Introduction to XAI and Clinical Decision Support

Thomas M. Connolly
https://orcid.org/0000-0002-4276-7301
DS Partnership, UK

Mario Soflano
https://orcid.org/0000-0003-0758-1509
Glasgow Caledonian University, UK

Petros Papadopoulos
https://orcid.org/0000-0002-8110-7576
University of Strathclyde, UK

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) offer significant opportunities in healthcare for innovation with its ability to solve cognitive problems normally requiring human intelligence. However, the potential of ML in healthcare has not been realised to date, with limited existing reports of the clinical and cost benefits that have arisen from their real-world in clinical practice. This is due to the lack of understanding about how some ML models operate and ultimately the way they come to make decisions. Explainable AI (XAI) has emerged as a response to this problem investigating methods and techniques that provide insights into the outcome of an ML model and present it in qualitative understandable terms or visualisations to the stakeholders of the model. This chapter introduces XAI and provides some examples of its use within healthcare.

INTRODUCTION

Artificial intelligence (AI) and, in particular, Machine Learning (ML) offer significant opportunities for innovation with its ability to solve cognitive problems normally requiring human intelligence. Considerable progress has been seen in many industries with the use of these technologies. In healthcare, ML

DOI: 10.4018/978-1-6684-5092-5.ch002

is viewed as a potential strategic tool that could contribute to improving clinical decision making (eg., diagnosis, screening and treatment), service organisation (eg., flow optimization, triage and resource allocation) and patient management.

ML has been applied to a wide range of medical and biological problems such as diagnosing diabetic retinopathy (Reddy et al., 2020), heart failure (Guo et al., 2020), asthma and COPD (Kaplan et al., 2021); predicting recurrence of cancer (Macyszyn et al., 2015), autism subtyping by clustering comorbidities (Parlett-Pelleriti et al., 2022), drug discovery (Stephenson et al., 2019), vaccine development (Chen et al., 2020) and clinical research in chronic diseases, such as AIDS (Bisaso et al., 2017) and Alzheimer's disease (Mirzaei & Adeli, 2022). It has been used widely in medical imaging with examples such as detecting thoracic disease through chest radiographs (Hwang et al., 2019), detecting cancer in mammograms (Wu et al., 2019), identifying brain tumours on MRI images (Amin et al., 2019), identifying skin cancers (Kassem et al., 2020) and polyp detection from colonoscopy (Wang et al., 2018). It has also been used in areas such as protein folding (Jumper et al., 2021) and genomics (Whalen et al., 2022).

Despite the enormous volume of research on ML in healthcare, its potential has not been realised to date, with limited existing reports of the clinical and cost benefits that have arisen from real-world use of ML algorithms in clinical practice. This is due to complex clinical, ethical and legal questions arising from the lack of understanding about how some ML models operate and ultimately the way they come to make decisions. This is commonly referred to as the "black-box" or opaque problem (Adadi & Berrada, 2018; Arrieta et al., 2020). As it is the responsibility of clinicians to give the best care to each patient, they need to be confident that ML systems can be trusted, however, this is limited by the black-box nature of some ML systems, particularly neural networks. Moreover, patients have a right to be informed about the risks, benefits or potential alternatives for any medical decision and again this is limited by the black-box nature of ML.

Explainable AI (XAI) has emerged as a response to this problem investigating methods and techniques that provide insights into the outcome of an ML model and present it in qualitative understandable terms or visualisations to the stakeholders of the model. This chapter discusses the principles of XAI and how XAI has been implemented within CDSSs. Chapter 11 will provide a systematic literature review of XAI and CDSSs.

BACKGROUND

Explainable AI aims to explain the way that AI systems work. At a high-level, we can distinguish between two types of models:

- models that are inherently explainable - simple, transparent and easy to understand, sometimes referred to as white-box or transparent models;
- models that are black-box in nature and require explanation through separate, replicating (surrogate) models that mimic the behaviour of the original model.

White-box systems include: