Chapter 7
Unravelling the Enigma of Machine Learning Model Interpretability in Enhancing Disease Prediction

Rati Kailash Prasad Tripathi
Assam University, Silchar, India

Shrikant Tiwari
https://orcid.org/0000-0001-6947-2362
Galgotias University, Greater Noida, India

ABSTRACT

Machine learning (ML) models have made significant strides in disease prediction, providing new avenues for early detection and intervention. These models have demonstrated remarkable capabilities in analysing vast and complex datasets to identify patterns and trends that can aid in early diagnosis and treatment. However, opacity of these models often leaves healthcare practitioners and patients in the dark about the reasoning behind their predictions, raising concerns about trust, fairness, and practical adoption of AI-based disease prediction. This review delves into the critical topic of interpretability in ML models for disease prediction, its importance, techniques to achieve it, impact on clinical decision-making, challenges, and implications in healthcare. Urgent issues and moral dilemmas pertaining to model interpretability in healthcare, areas for further research to enhance interpretability of predictive models, and applications are also highlighted. Thus, the chapter provides insights into the applicability of AI-driven models to improve healthcare decision-making and patient outcomes.

1. INTRODUCTION

The integration of machine learning into the healthcare sector represents a monumental leap in the ongoing evolution of medical science and patient care. In recent years, artificial intelligence (AI), par-

DOI: 10.4018/978-1-6684-8531-6.ch007

ticularly machine learning (ML), has emerged as a disruptive force with the potential to revolutionize the healthcare landscape (Davenport & Kalakota, 2019). ML, a subfield of artificial intelligence, endows computer systems with the ability to learn and adapt from data, making predictions and decisions without explicit programming. In the context of healthcare, this translates to an unprecedented capacity to analyse vast and complex medical datasets, extract meaningful insights, and make informed predictions (Bohr & Memarzadeh, 2020).

One of the primary areas where ML excels in healthcare is disease prediction and diagnosis (Marr, 2018). These models can comb through electronic health records, medical images, and genetic data to identify patterns that might elude even the most seasoned healthcare professionals (Miller & Brown, 2018). Early detection of diseases such as cancer, diabetes, neurodegenerative and cardiovascular conditions has the potential to significantly improve patient outcomes, leading to timely interventions and more effective treatments (Ekins, 2019). Additionally, ML contributes to the development of precision medicine (Konieczny & Roerman, 2019). By analysing a patient's genetic makeup and medical history, these models can recommend treatments tailored to an individual's unique characteristics, increasing the likelihood of successful outcomes and reducing adverse effects. Beyond clinical applications, machine learning assists in hospital operations, streamlining administrative processes, optimizing resource allocation, and improving patient management. It is also instrumental in epidemiology, tracking the spread of diseases, and monitoring public health trends (Johnson et al., 2021).

However, the integration of ML in healthcare is not without challenges. Ensuring data privacy, maintaining the security of sensitive medical records, addressing issues of bias and fairness in AI algorithms, and meeting regulatory compliance are critical concerns that demand attention (Murdoch, 2021; Nicholson Price & Glenn Cohen, 2019). This transformation, driven by ML, represents a paradigm shift in healthcare, enabling practitioners to move from a reactive, one-size-fits-all approach to a proactive, patient-centric, and data-driven model of care (Murdoch, 2021; Nicholson Price & Glenn Cohen, 2019). A deeper exploration of the myriad applications, challenges, and ethical considerations that underpin the integration of machine learning into healthcare, ultimately paves the way for a brighter and more data-informed future in medicine and patient well-being (Murdoch, 2021).

This chapter delves into the critical subject of interpretability in ML models for disease prediction, exploring the techniques, challenges, and implications that surround it. The pivotal role of model interpretability in healthcare and disease prediction has been examined herein, stressing the consequences of relying on black-box models in clinical decision-making. Furthermore, to unravel the enigma of ML model interpretability, a range of techniques and methodologies that have been developed to make these models more transparent and interpretable have been explored. Methods such as Local Interpretable Model-agnostic Explanations (LIME), Shapley values (SHAP), and feature importance analysis are discussed, shedding light on how these techniques can provide insights into the decision-making process of complex models.

Besides, the challenges and limitations associated with achieving model interpretability in the healthcare domain has also been confronted. Ethical considerations, particularly those related to fairness, bias, and transparency in predictive models has also been delved into. Furthermore, the tangible impact of interpretable models on clinical decision-making, from more accurate diagnoses to tailored treatment recommendations, ultimately improving patient care has also been considered. Understanding the importance of trust in healthcare, the role of model interpretability in building trust between patients and healthcare providers has been explored. Furthermore, areas of further research and development to enhance the interpretability of machine learning models in disease prediction has also been discussed.