Chapter 16
Trustworthy AI in Healthcare:
Insights, Challenges, and the Significance of Overfitting in Predicting Mental Health

Partha Sarathi Bishnu
https://orcid.org/0000-0001-5143-6195
Birla Institute of Technology, Mesra, India

ABSTRACT

The rapid integration of artificial intelligence (AI) into medical informatics, particularly in the context of mental health data, can bring about significant transformations in healthcare decision-support systems. However, ensuring that AI gains widespread acceptance and is regarded as reliable in healthcare requires addressing critical issues concerning its robustness, fairness, and privacy. This chapter presents a comprehensive study that delves into the urgent need for dependable AI in medical informatics, explicitly focusing on collecting mental health data using sensors. The authors put forth a methodological framework combining cutting-edge AI techniques, leveraging deep learning models such as recurrent neural networks (RNN), including variants like LSTM and GRU, and ensemble techniques like random forest, AdaBoost, and XGBoost. Through a series of experiments involving healthcare decision support systems, the authors underscore the pivotal role of model overfitting in establishing trustworthy AI systems.

1. INTRODUCTION

In recent years, the intersection of Artificial Intelligence (AI) and Medical Informatics (MI) has held the promise of transforming healthcare by augmenting decision support systems (Bajwa et al., 2021). However, the adoption of AI in this critical domain necessitates the development of Trustworthy AI systems. Trustworthy AI entails the creation of AI systems that not only demonstrate technical proficiency but also adhere to ethical principles, ensuring transparency, fairness, robustness, and privacy protection in their operations (Barocas et al., 2013). Medical Informatics encompasses the interdisciplinary field of study that focuses on applying information technology, deep learning and data science to healthcare management, decision-making, and patient care. The critical nature of healthcare decisions underscores

DOI: 10.4018/979-8-3693-5261-8.ch016

the need for Trustworthy AI in Medical Informatics. Trustworthy AI is imperative to prevent erroneous medical recommendations, protect sensitive patient data, and enhance the overall quality of healthcare delivery (Diana et al., 2007). Achieving Trustworthy AI in Medical Informatics is challenging due to biased decision-making with diverse patient populations. Biased training data can lead to inaccurate predictions, especially for underrepresented groups. Model overfitting worsens this, hindering generalization. Transparency and robustness are crucial for handling unexpected scenarios, with overfitting complicating bias mitigation. Addressing overfitting is vital for improving AI reliability and reducing biased decision-making in medical informatics, especially with diverse patients. The paper's objectives can be succinctly outlined as follows: Firstly, our primary aim is to elucidate the fundamental principles of Trustworthy AI in the realm of medical informatics, with a strong emphasis on pivotal concepts like transparency, fairness, robustness, and privacy preservation, which have received extensive attention in existing literature. Secondly, we underscore the paramount significance of addressing the model overfitting issue, even when models exhibit high accuracy. Our specific focus revolves around the utilization of mental health data collected from diverse sensors, including those capturing facial expressions, body postures, ECG (Electrocardiogram) signals, and skin conductance, for the purpose of predicting an individual's mental health status. To achieve this, we employ advanced Deep Learning (DL) techniques, encompassing various Recurrent Neural Network (RNN) architectures such as simple RNN, LSTM (Hochreiter and Schmidhuber, 1997), and GRU (Cho et al., 2014), as well as the latest ensemble techniques like Random Forest (Breiman, 2001), AdaBoost (Freund and Schapire, 1997), and XGBoost (Chen and Guestrin, 2016). These techniques are applied to relevant sensor datasets, and we rigorously conduct experiments, presenting results with com- prehensive recommendations that emphasize the development of trustworthy AI by addressing accuracy and overfitting issues, which are of paramount importance.

The chapter is structured as follows: Section 2 delves into the fundamental concept of Trustworthy AI in Healthcare. In Section 3, we explore the significance of the overfitting issue in the context of Trustworthiness in AI Systems, using mental health sensor data as a case study. Lastly, in Section 4, we provide our concluding remarks.

2. TRUSTWORTHY AI IN HEALTHCARE

In this discourse, we delve into the overarching notion of Trustworthy AI, the intersection of Healthcare and AI, the imperative demand for Trustworthy AI within the realm of Medical Informatics, the inherent constraints faced by AI in the context of health- care, and finally, we present a robust methodology geared towards the establishment of Trustworthy AI in the healthcare domain.

2.1 Trustworthy AI: Concepts, Characteristics, and Ethical Principles

Trustworthy AI refers to Artificial Intelligence systems and technologies that excel in their technical capabilities, adhere to ethical principles, and exhibit specific characteristics to ensure transparency, fairness, reliability, privacy protection, and accountability in their operations. Trustworthy AI fosters trust and confidence among users and stakeholders while mitigating the potential adverse effects and risks associated with AI applications in various domains, including healthcare, finance, and transportation (Bajwa et al., 2021; Floridi, 2020). Trustworthy AI encompasses several fundamental characteristics for successfully integrating into various domains. First and foremost, transparency is essential, ensuring that