Deployment of a Clinical Data Lake for ADHD: Architecture and Preliminary Insights

Sandra García-Ponsoda
Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, Carretera San Vicent del Raspeig, s/n, San Vicent del Raspeig, 03690, Alicante, Spain

Javier Sanchis
Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, Carretera San Vicent del Raspeig, s/n, San Vicent del Raspeig, 03690, Alicante, Spain

Juan Trujillo
Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, Carretera San Vicent del Raspeig, s/n, San Vicent del Raspeig, 03690, Alicante, Spain

Alejandro Maté
Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, Carretera San Vicent del Raspeig, s/n, San Vicent del Raspeig, 03690, Alicante, Spain

Miguel A. Teruel
Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, Carretera San Vicent del Raspeig, s/n, San Vicent del Raspeig, 03690, Alicante, Spain

Abstract

Attention Deficit/Hyperactivity Disorder (ADHD) generates diverse clinical, behavioral, and physiological data that are typically fragmented across systems. To address this, we present the deployment of a clinical data lake within the BALLADEER project, designed to integrate multimodal information, including EEG, physiological signals, and gamified task performance, into a single repository. The architecture preserves raw data formats under a schema-on-read approach, supports synchronization across modalities via event markers and timestamps, and enforces privacy through pseudonymization and controlled access. Standardized acquisition protocols and quality-control checks ensure data reliability, while a consistent file hierarchy facilitates findability and reuse. The dataset currently includes 164 participants aged 6-18 years and incorporates serious games developed within the project, such as Attention Robots and Attention Slackline. At the time of writing, the system is being extended with a new task, Attention Mistakes, demonstrating its adaptability. This deployment establishes a scalable foundation for future Lakehouse extensions, federated deployments, and advanced analytics to support biomarker discovery and decision support in ADHD research.

Keywords

Data Lake, Big Data, Healthcare Analytics, Machine Learning, Attention/Deficit Hyperactivity Disorder.