ASEC 2024 Conference

The 5th International Electronic Conference on Applied Sciences

04–06 December 2024 | Online

MDPI

Integration and Standardization of Heterogeneous Autism Data

Kauã Lima, Vagner Silva, Gabriel Souza, Matheus Nascimento, Jean Turet
Federal University of Alagoas
Group of Engineering in Decision-Making and Artificial Intelligence

INTRODUCTION & AIM

Autism Spectrum Disorder (ASD) is a complex condition that affects neurological and social development, impacting millions of people worldwide. In Brazil, the number of students diagnosed with ASD enrolled in schools reached approximately 636,000 in 2023, reflecting a 48% increase compared to previous years, according to the 2023 School Census (IBGE, 2023). This growth highlights increased awareness and diagnosis but also underscores the need for strategies to enhance support and understanding of the disorder, particularly given its diverse clinical and behavioral manifestations.

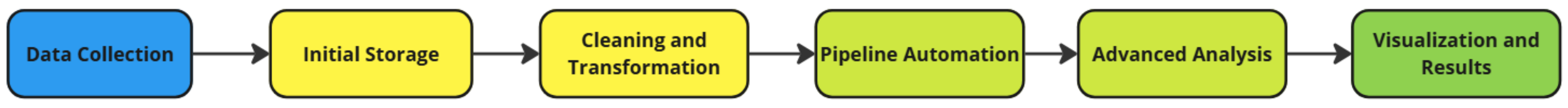
Despite advancements in research and support services, one of the greatest challenges lies in the integration and joint analysis of ASD-related data. These data originate from heterogeneous sources, such as medical records, monitoring devices (IoT), clinical evaluations, and qualitative questionnaires. The lack of standardization and interoperability hampers the generation of meaningful insights and limits the development of personalized interventions. To address this issue, this project proposes the development of an automated pipeline to collect, clean, transform, and integrate these data, enabling more accurate analyses and supporting evidence-based decision-making.

METHOD

The methodology of this project was designed to integrate and analyze heterogeneous data related to Autism Spectrum Disorder (ASD) through an automated pipeline, covering processes from data collection to advanced analysis. The primary goal was to ensure standardization, quality, and efficiency in processing data from various sources, such as medical records, monitoring devices, questionnaires, and clinical reports. Below, we present the pipeline steps that enabled effective data integration and analysis.

1. **Data Collection**
   Various sources were utilized, including medical records, IoT devices, questionnaires, and clinical reports. Initial storage was conducted on **Amazon S3**, ensuring scalability (**Amazon, 2023**).
2. **Data Cleaning and Transformation**
   Data were standardized using **Pandas** and **Google BigQuery**, ensuring unified formats and terminologies to facilitate analysis (**McKinney, 2010; Google, 2023**).
3. **Pipeline Automation**
   Apache Airflow was employed to automate the process, ensuring continuous and efficient execution of all pipeline stages (**Apache Software Foundation, 2023**).
4. **Advanced Analysis**
   Behavioral pattern identification and correlations between clinical and monitoring data were performed. Additionally, sentiment analysis was applied to questionnaires using **Scikit-learn**, **TensorFlow**, and (**Tableau, 2023)** for modeling and visualization.

With the accompanying flowchart, the process can be visualized more clearly.



RESULTS & DISCUSSION

Data Standardization and Quality

The implementation of an automated structure for data collection and processing resulted in consistent, complete, and high-quality information. Standardizing formats and terminologies was crucial for integrating diverse data sources, reducing inconsistencies, and enhancing the reliability of analyses. This improvement provided a solid foundation for predictive modeling and insight generation.

Advances in Predictive Models

With standardized data, machine learning models demonstrated significantly improved performance in terms of precision, recall, and F1-score. These models effectively predicted specific behaviors and supported the early identification of patterns associated with ASD, aiding in diagnosis and patient follow-up. Integrating clinical data with IoT device information proved particularly effective in generating behavioral insights.

Intuitive Visualizations and Insights

The use of Tableau for data visualization enabled clear and intuitive representations of the results. These charts and dashboards facilitated data interpretation by healthcare professionals, allowing for the identification of trends and patterns that were previously unclear. Sentiment analysis in questionnaires, for instance, highlighted important emotional aspects, complementing clinical and monitoring data.

Impact on Care and Public Policies

The integration and advanced analysis of data enabled more personalized and effective interventions for patients. Additionally, the insights generated contributed to evidence-based public policy formulation aimed at ASD. The availability of standardized and high-quality data provided a comprehensive view of the social and clinical impact of ASD, aiding in resource allocation and the development of support programs.

Conclusions on Data Integration

The results demonstrate that integrating heterogeneous sources not only enhances the understanding of ASD but also significantly expands the possibilities for analysis. The project surpassed the state of the art in terms of precision, accessibility, and social impact, establishing a model that can be replicated for other conditions and contexts.

CONCLUSION

The project established a robust infrastructure for integrating and standardizing heterogeneous ASD data, enabling advanced analyses and practical insights for healthcare professionals and public managers. In addition to providing a solid foundation for more accurate predictive models, the system facilitated the identification of behavioral patterns, the correlation of clinical and monitoring data, and the understanding of emotional aspects through sentiment analysis. These advancements have the potential to transform clinical follow-up, personalize interventions, and support the development of more effective and targeted public policies for individuals with ASD and their families.

FUTURE WORK / REFERENCES

•Amazon. (2023). *Amazon S3*. Retrieved from https://aws.amazon.com/s3
• IBGE. (2023). *Censo Escolar 2023*.
•McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*.
•Google. (2023). *Google BigQuery*. Retrieved from https://cloud.google.com/bigquery
•Apache Software Foundation. (2023). *Apache Airflow*. Retrieved from https://airflow.apache.org
•Abadi, M., et al. (2016). TensorFlow: A System for Large-Scale Machine Learning. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI 16)*.
•Tableau. (2023). *Tableau Software*. Retrieved from https://www.tableau.com