PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITÉ FERHAT ABBAS - SETIF 1



Faculty of Sciences
Computer Science Department

MASTER'S THESIS

FIELD: Mathematics and Computer Science

FIELD OF STUDY: Computer Science

SPECIALTY: Foundations and Engineering of Information and Image (F3I)

Theme

Artificial Intelligence and Computational Approaches for the Identification of Clinical Markers in Autism Spectrum Disorder (ASD)

Presented by:
Tebib Oumaima
Saber Feriel

Directed by:
DR. LAKEHAL ABDERRAHIM

Defense Date : 27/06/2026

Defense Jury :

Dr. Annane Boubakeur
Dr. Touahria Imad Eddine

Promotion : 2025/2026

Contents

List of Figures **VI**

List of Tables **VII**

Abstract **XV**

Résume **XVI**

General Introduction **XIX**

1 Background and General Context **1**

1.1 Introduction 1

1.2 Autism Spectrum Disorder: Definition and Characteristics 2

1.2.1 Definition 2

1.2.2 Neurodevelopmental Nature of ASD 5

1.2.2.1 Early Brain Development 5

1.2.3 The Spectrum Concept 7

1.2.3.1 Variability Across Domains 7

1.2.3.2 Historical Context: 9

1.2.3.3 Why the Spectrum Concept Matters: 9

1.2.4 Genetic and Environmental Effects 9

1.2.4.1 Types of Genetic Variations in ASD 9

1.2.4.2 Gene–Environment Interactions 11

1.2.4.3 ASD Is NOT Caused by Parenting 12

1.2.5 Hereditary and Environmental Effects 12

1.2.5.1 Developmental Changes in Symptom Severity 12

1.2.5.2 Impact of Early Intervention 13

1.2.5.3 Functional Independence Varies Widely 13

1.2.6 Developmental Trajectories 14

1.3 Current Diagnostic Approaches for Autism Spectrum Disorder (ASD) . . 15

1.3.1 Clinical Diagnostic Criteria (DSM-5) 16

1.3.2 Standardized Behavioral Assessment Tools 17

1.3.2.1 Autism Diagnostic Observation Schedule (ADOS) . . . 17

1.3.2.2 Autism Diagnostic Interview – Revised (ADI-R) 17
1.3.2.3 Screening Tools for Early Detection 18
1.3.2.4 Developmental and Cognitive Assessments 18
1.3.3 Challenges and Limitations of Current ASD Diagnostic Methods . 19
1.3.3.1 Subjectivity in Clinical Diagnosis 19
1.3.3.2 Time-Consuming Diagnostic Process 19
1.3.3.3 Delayed Diagnosis and Early Detection Difficulties . . . 19
1.3.3.4 Lack of Biological Markers 19
1.3.3.5 ASD Heterogeneity 19
1.3.3.6 Scalability and Data Complexity 19
1.4 Role of Artificial Intelligence in Autism Spectrum Disorder 20
1.4.1 Early Diagnosis and Screening 20
1.4.2 Medical Imaging Analysis (Neuroimaging) 20
1.4.3 Biomarker Discovery: 20
1.4.4 Genetic and Genomic Analysis 21
1.4.5 Speech and Language Analysis 21
1.4.6 Behavioral Data Analysis 21
1.4.7 Patient Stratification and Personalized Medicine 21
1.4.8 Multimodal Data Integration 21
1.4.9 Assistive Technologies and Intervention 21
1.4.10 Explainable AI (XAI) in ASD 22
1.5 Conclusion 23

2 Literature Review **24**
2.1 Introduction 24
2.2 ASD and Data Diversity 25
2.2.1 Data Types 25
2.2.1.1 Behavioral datasets 25
2.2.1.2 Neuroimaging data (MRI, fMRI) 26
2.2.1.3 Genetic and genomic data 27
2.2.2 Data bases and repositories 29
2.2.2.1 Autism Brain Imaging Data Exchange (ABIDE) 29
2.2.2.2 Simons Foundation Autism Research Initiative (SFARI) 29
2.2.2.3 ASD Bank (TalkBank) 30
2.2.2.4 Open-access datasets from research institutions 30
2.3 Deep Learning Applications in Neurodevelopmental Disorders 32
2.3.1 Early Deep Learning Integration 32
2.3.2 Temporal Modeling with LSTM 32
2.3.3 Large-Scale DNN on ABIDE 33

2.3.4 Hybrid Learning and Representation . . . 33
2.3.5 CNN for Spatial Feature Learning . . . 34
2.3.6 Systematic Review and Critical Insights . . . 34
2.3.7 CNN-Based EEG Analysis for ASD Detection . . . 34
2.3.8 Transformer-Based ASD Classification from Neuroimaging Data . 35
2.3.9 Graph Neural Networks for Brain Connectivity Analysis . . . 35
2.4 Multimodal Data Analysis in ASD . . . 37
2.4.1 Rationale for Multimodal Integration . . . 37
2.4.2 Modalities Used in ASD Research . . . 38
2.4.2.1 Clinical and Behavioral Modalities . . . 38
2.4.2.2 Neuroimaging Modalities . . . 38
2.4.2.3 Genetic and Molecular Modalities . . . 42
2.4.2.4 Developmental and Environmental Data . . . 43
2.4.3 Multimodal Integration Strategies . . . 43
2.4.4 Challenges in Multimodal ASD Analysis . . . 48
2.5 General Discussion . . . 48
2.5.1 Key Studies and Findings . . . 48
2.5.2 Clinical Applicability . . . 49
2.6 Conclusion . . . 49

3 Methodology – Multi-Modal Biomarker Discovery for Autism Spectrum Disorder 51
3.1 Introduction and Motivation . . . 51
3.2 Dataset Description . . . 51
3.3 Time Series Pipeline . . . 52
3.3.1 Preprocessing . . . 53
3.3.2 Feature Extraction . . . 54
3.3.3 Classification Model . . . 55
3.4 Genomic Pipeline . . . 57
3.4.1 Preprocessing and Quality Control . . . 57
3.4.2 Feature Extraction on Genetic Data . . . 59
3.5 fMRI Imaging Pipeline . . . 61
3.5.1 Temporal Window Selection . . . 61
3.5.2 CNN with Attention Mechanism for Classification . . . 62
3.5.3 Explainability and Brain Region Extraction . . . 63
3.6 Cross-Modal Discussion . . . 65
3.7 Conclusion . . . 66

4 Experiments, Results, and Implementation 68

4.1 Introduction . . . 68
4.2 Experimental Setup . . . 69
4.2.1 Hardware . . . 69
4.2.2 Software: . . . 69
4.2.3 Reproducibility Measures . . . 70
4.3 Time Series Pipeline — Details and Results: . . . 71
4.3.1 Dataset Details . . . 71
4.3.2 Preprocessing Steps . . . 72
4.3.3 Results and Analysis . . . 73
4.4 Genomic Pipeline — Details and Results . . . 76
4.4.1 Dataset Details . . . 76
4.4.2 Preprocessing and Quality Control Steps . . . 77
4.4.3 Feature Extraction Results . . . 78
4.4.4 fMRI Imaging Pipeline — Details and Results . . . 80
4.4.5 Dataset Details . . . 80
4.4.6 Preprocessing Steps . . . 81
4.4.7 Temporal Window Selection — Results . . . 82
4.4.8 CNN + Attention Model — Implementation and Results . . . 82
4.4.9 Explainability Results and Brain Region Analysis . . . 86
4.5 Cross-Modal Integration Results . . . 86
4.5.1 Quantitative Comparison of Pipeline Outputs . . . 87
4.5.2 Qualitative Comparison . . . 87
4.5.3 Relationship Between Genomic and Neuroimaging Findings . . . 88
4.6 Discussion . . . 88
4.6.1 Synthesis of Findings . . . 88
4.6.2 Potential Clinical Relevance . . . 88
4.6.3 Future work . . . 89
4.7 Conclusion . . . 89

General Conclusion **90**

Bibliography **93**

Abstract

This abstract discusses the integration of artificial intelligence (AI) and deep learning (DL) and machine laerning (ML) in the identification of clinical markers for Autism Spectrum Disorder (ASD). It emphasizes the use of heterogeneous clinical, behavioral, and neuropsychological data combined with machine learning techniques to improve early diagnosis and patient stratification. Large-scale datasets and recent advances in data collection technologies have enabled the availability of complex ASD-related data, highlighting the need for efficient computational approaches to extract meaningful patterns and biomarkers.

The study highlights the importance of data preprocessing, including quality control, normalization, and integration, to ensure the reliability and robustness of predictive models. Feature selection and dimensionality reduction techniques are applied to handle high-dimensional data, reduce noise, and identify the most discriminative clinical features associated with ASD. The performance of the proposed models is evaluated using standard metrics such as accuracy, precision and recall.

This study develops a deep learning-based framework for identifying and predicting ASD-related clinical markers. The model incorporates advanced architectures, such as Convolutional Neural Networks (CNN) and other machine learning algorithms, and is trained on integrated multimodal datasets. To evaluate the effectiveness of the approach, different strategies are compared, including traditional feature selection methods and AI-driven automated feature extraction techniques.

In conclusion, the experimental results demonstrate that AI-based models outperform conventional statistical approaches in identifying relevant clinical markers and improving diagnostic accuracy. The ability of deep learning models to capture complex patterns and interactions within heterogeneous ASD data contributes significantly to their superior performance. These findings highlight the potential of AI in enhancing early diagnosis, clinical decision-making, and personalized intervention strategies for Autism Spectrum Disorder.

Keywords: Autism Spectrum Disorder (ASD), Artificial Intelligence, Machine Learning, Deep Learning, Feature Extraction, Neuroimaging, Classification.

Résume

Ce résumé aborde l'intégration de l'intelligence artificielle (IA), de l'apprentissage profond (AP) et de l'apprentissage automatique (AA) dans l'identification de marqueurs cliniques des troubles du spectre de l'autisme (TSA). Il met l'accent sur l'utilisation de données cliniques, comportementales et neuropsychologiques hétérogènes, combinées à des techniques d'apprentissage automatique, afin d'améliorer le diagnostic précoce et la stratification des patients. L'abondance des jeux de données et les progrès récents des technologies de collecte de données ont permis d'accéder à des données complexes sur les TSA, soulignant la nécessité d'approches informatiques efficaces pour extraire des modèles et des biomarqueurs pertinents.

L'étude souligne l'importance du prétraitement des données, incluant le contrôle qualité, la normalisation et l'intégration, pour garantir la fiabilité et la robustesse des modèles prédictifs. Des techniques de sélection de caractéristiques et de réduction de dimensionnalité sont appliquées pour traiter les données de grande dimension, réduire le bruit et identifier les caractéristiques cliniques les plus discriminantes associées aux TSA. La performance des modèles proposés est évaluée à l'aide de métriques standard telles que l'exactitude, la précision et le rappel.

Cette étude développe un cadre basé sur l'apprentissage profond pour identifier et prédire les marqueurs cliniques liés aux troubles du spectre de l'autisme (TSA). Le modèle intègre des architectures avancées, telles que les réseaux de neurones convolutifs (CNN) et d'autres algorithmes d'apprentissage automatique, et est entraîné sur des ensembles de données multimodaux intégrés. Afin d'évaluer l'efficacité de l'approche, différentes stratégies sont comparées, notamment les méthodes traditionnelles de sélection de caractéristiques et les techniques d'extraction automatique de caractéristiques pilotées par l'IA.

En conclusion, les résultats expérimentaux démontrent que les modèles basés sur l'IA surpassent les approches statistiques conventionnelles pour identifier les marqueurs cliniques pertinents et améliorer la précision du diagnostic. La capacité des modèles d'apprentissage profond à appréhender les schémas et interactions complexes au sein de données hétérogènes relatives aux troubles du spectre de l'autisme (TSA) contribue significativement à leurs performances supérieures. Ces résultats soulignent le potentiel de l'IA pour améliorer le diagnostic précoce, la prise de décision clinique et les stratégies

d'intervention personnalisées pour les TSA.

Mots-clés: Trouble du spectre autistique (TSA), intelligence artificielle, apprentissage automatique, apprentissage profond, extraction de caractéristiques, neuro-imagerie, classification

ملخص

يتناول هذا الملخص دمج بين الذكاء الاصطناعي والتعلم العميق والتعلم الالي في تحديد المؤشرات السريرية لاضطراب طيف التوحد(ASD) . وتؤكد على استخدام بيانات سريرية وسلوكية وعصبية نفسية متنوعة، بالإضافة إلى تقنيات التعلم الآلي، لتحسين التشخيص المبكر وتصنيف المرضى. وقد أتاحت مجموعات البيانات الضخمة والتطورات الحديثة في تقنيات جمع البيانات إمكانية الوصول إلى بيانات معقدة متعلقة باضطراب طيف التوحد، مما يبرز الحاجة إلى مناهج حسابية فعالة لاستخلاص الأنماط والمؤشرات الحيوية ذات الدلالة. وتُسلط الدراسة الضوء على أهمية المعالجة المسبقة للبيانات، بما في ذلك مراقبة الجودة وتوحيد القيم والتكامل، لضمان موثوقية نماذج التنبؤ وقوتها. وتُطبق تقنيات اختيار الميزات وتقليل الأبعاد للتعامل مع البيانات عالية الأبعاد، وتقليل التشويش، وتحديد السمات السريرية الأكثر تمييزًا المرتبطة باضطراب طيف التوحد. ويتم تقييم أداء النماذج المقترحة باستخدام مقاييس معيارية مثل الدقة والضبط والاستدعاء. وتطور هذه الدراسة إطار عمل قائم على التعلم العميق لتحديد المؤشرات السريرية المتعلقة باضطراب طيف التوحد والتنبؤ بها. يشتمل النموذج على بنى متقدمة، مثل الشبكات العصبية الالتفافية (CNN) وخوارزميات تعلم الآلة الأخرى، ويتم تدريبه على مجموعات بيانات متعددة الوسائط متكاملة. ولتقييم فعالية هذا النهج، تمت مقارنة استراتيجيات مختلفة، بما في ذلك أساليب اختيار الميزات التقليدية وتقنيات استخراج الميزات الآلية المدعومة بالذكاء الاصطناعي. في الختام، تُظهر النتائج التجريبية أن النماذج القائمة على الذكاء الاصطناعي تتفوق على الأساليب الإحصائية التقليدية في تحديد المؤشرات السريرية ذات الصلة وتحسين دقة التشخيص. وتساهم قدرة نماذج التعلم العميق على التقاط الأنماط والتفاعلات المعقدة ضمن بيانات اضطراب طيف التوحد غير المتجانسة بشكل كبير في أدائها المتفوق.

تُبرز هذه النتائج إمكانات الذكاء الاصطناعي في تعزيز التشخيص المبكر، واتخاذ القرارات السريرية، واستراتيجيات التدخل الشخصية لاضطراب طيف التوحد.

الكلمات المفتاحية: اضطراب طيف التوحد، الذكاء الاصطناعي، التعلم الآلي، التعلم العميق، استخلاص الميزات، التصوير العصبي، التصنيف.