UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

A Meta-analysis of Machine Learning Classification Tools Using rs-fMRI Data For Autism Spectrum Disorder Diagnosis

Caio Pinheiro Santana

Itajubá, Março de 2021

UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

Caio Pinheiro Santana

A Meta-analysis of Machine Learning Classification Tools Using rs-fMRI Data For Autism Spectrum Disorder Diagnosis

Dissertação submetida ao Programa de Pós-Graduação em Ciência e Tecnologia da Computação como parte dos requisitos para obtenção do Título de Mestre em Ciência e Tecnologia da Computação.

Área de Concentração: Matemática da Computação

Orientador: Prof. Dr. Guilherme Sousa Bastos
Coorientador: Prof. Dr. Adler Diniz de Souza

Março de 2021
Itajubá - MG

Agradecimentos

Agradeço primeiramente à Universidade Federal de Itajubá e ao programa de pós-graduação em Ciência e Tecnologia da Computação por disponibilizarem e me aceitarem neste programa de mestrado e à CAPES pela bolsa recebida durante a realização do mesmo.

Agradeço a mim mesmo por ter persistido até a finalização deste trabalho, apesar de todas as dificuldades e contratempos que encontrei pelo caminho e por vezes me fizeram pensar em desistir.

Agradeço ao meu orientador Guilherme Souza Bastos e meu coorientador Adler Diniz de Souza pela oportunidade e confiança, pelos direcionamentos e auxílios, por todo o suporte e compreensão.

Agradeço aos meus pais e meus irmãos, que sempre acreditaram em mim, me apoiaram e aconselharam apesar da distância física entre nós. Agradeço também à minha família de uma forma geral por todo o amor e carinho apesar da minha dificuldade em manter contato.

Agradeço à minha psicóloga por todo o trabalho que realizamos juntos e que me fez não apenas conseguir continuar, mas sair fortalecido, confiante, empolgado e inspirado.

Agradeço aos meus colegas, em especial ao Igor e ao Emerson, pelas trocas e conversas, por todas as horas de trabalho conjunto, pelas sugestões, ajudas, indicações de artigos, revisões e muito mais.

Agredeço à república Blasfêmia e seus moradores, ex-moradores e agregados por terem me acolhido durante o mestrado e me tornado parte da família, por todos os rolês malucos, discussões sobre tópicos aleatórios, aprendizados e amizades, por aturar meus surtos, me incentivar e apoiar, por me aceitarem do jeito que sou. Em especial, agradeço à Anna - minha colega de quarto quase vitalícia nesse período - por me aguentar por tanto tempo, ouvir meus lamentos e conselhos, por sempre ser uma parceira incrível nas mais diversas situações, e à Claudinha, pelas conversas e músicas madrugada a dentro e por ter me ajudado tanto a sobreviver aos períodos de quarentena.

Agradeço à Cia de Dança Corpo-a-corpo e todas as pessoas incríveis que encontrei e conheci por lá, em especial o Luiz, o Aliffi, a Nathy e a Júlia, pelo carinho e acolhimento, pela energia contagiante, pelas viagens, apresentações e competições, pela oportunidade de me expressar através da dança e espairecer, pelas longas horas de conversas aleatórias e risadas, por terem contribuído tanto para que eu mantivesse minha sanidade mental durante os últimos três anos.

Agradeço ao meu grupo de DnD pelas horas de jogatina em universos fantásticos e incríveis que me transportam para outra realidade, aliviando o peso da vida real e despertando em mim uma paixão que seguirá sempre comigo.

Agradeço à Nani por tantos anos de uma amizade maravilhosa, pelos filmes que eu vi praticamente sozinho, pela aleatoriedade infinita, por sempre me puxar para as mais diversas situações inusitadas e divertidas, pelos incontáveis momentos, bons e ruins, que passamos juntos. Agradeço à Carol por estar ao meu lado e me apoiar em alguns dos meus piores momentos, por todo o carinho e tudo que compartilhamos, por manter nossa amizade tão linda quanto sempre foi mesmo após longos períodos sem nos vermos. Agradeço à Érica pela conexão profunda que desenvolvemos, por escutar meus inúmeros áudios gigantescos, pela confiança mútua e por se manter sempre presente mesmo à distância.

Por fim, agradeço ao Gui, que entrou há pouco tempo na minha vida, mas fez toda diferença nessa reta final, por todo carinho, apoio e incentivo, por me apresentar tantas coisas e compartilhar momentos tão especiais comigo, por me inspirar tanto e ser um dos grandes responsáveis pela leveza com a qual consegui enfrentar esse período tão complicado.

Resumo

O Transtorno do Espectro Autista (TEA) é uma condição complexa e heterogênea que afeta o desenvolvimento cerebral e é caracterizada por disfunções cognitivas, comportamentais e sociais. Muito esforço vem sendo feito para identificar biomarcadores baseados em imagens cerebrais e desenvolver ferramentas que poderiam facilitar o diagnóstico do TEA - atualmente baseado em critérios comportamentais, através de um processo longo e demorado. Em particular, o uso de algoritmos de Aprendizado de Máquina para classificação de dados de Imagens de Ressonância Magnética funcional em estado de repouso (rs-fMRI) é promissor, mas há uma necessidade contínua de pesquisas adicionais a respeito da precisão desses classificadores. Assim, este trabalho realiza uma revisão sistemática e meta-análise de modo a resumir e agregar as evidências disponíveis na literatura da área até o momento. A busca sistemática por artigos resultou na seleção de 93 deles, que tiveram seus dados extraídos e analisados através da revisão sistemática. Um modelo meta-analítico bivariado de efeitos aleatórios foi implementado para investigar a sensibilidade e especificidade dos 55 estudos (132 amostras independentes) que ofereceram informação suficiente para serem utilizados na análise quantitativa. Os resultados obtidos indicaram estimativas gerais de sensibilidade e especificidade de 73.8% (95% IC: 71.8-75.8%) e 74.8% (95% IC: 72.3-77.1%), respectivamente, e os classificadores baseados em SVM (do inglês, Support Vector Machine) se destacaram como os mais utilizados, apresentando estimativas acima de 76%. Estudos que utilizaram amostras maiores tenderam a obter piores resultados de precisão, com exceção do subgrupo composto por classificadores baseados em Redes Neurais Artificiais. O uso de outros tipos de imagens cerebrais ou dados fenotípicos para complementar as informações obtidas através da rs-fMRI se mostrou promissor, alcançando especialmente sensibilidades mais altas ($p = 0.002$) em relação aos estudos que utilizaram apenas dados de rs-fMRI (84.7% - 95% IC: 78.5-89.4% - versus 72.8% - 95% IC: 70.6-74.8%). Valores menores de sensibilidade/especificidade foram encontrados quando o número de Regiões de Interesse (ROIs, do inglês Regions of Interest) aumentou. Vale destacar também o desempenho das abordagens utilizando o atlas AAL (do inglês, Automated Anatomical Labelling) com 116 ROIs. Em relação às *features* usadas para treinar os classificadores, foram encontrados melhores resultados nos estudos que utilizaram a correlação de Pearson em conjunto com a transformação Z de Fisher ou outras *features* em comparação ao uso da correlação de Pearson sem modificações. Finalmente, a análise revelou valores da área sob a curva ROC (do inglês, Receiver Operating Characteristic) entre aceitável e excelente. Entretanto, considerando as várias limitações que são indicadas no estudo, mais estudos bem desenhados são necessários para estender o uso potencial desses algoritmos de classificação a ambientes clínicos.

Palavras-chaves: Transtorno do Espectro Autista, Aprendizado de Máquina, rs-fMRI, Meta-análise.

Abstract

The Autism Spectrum Disorder (ASD) is a complex and heterogeneous neurodevelopmental condition characterized by cognitive, behavioral, and social dysfunction. Much effort is being made to identify brain imaging biomarkers and develop tools that could facilitate its diagnosis - currently based on behavioral criteria through a lengthy and time-consuming process. In particular, the use of Machine Learning (ML) classifiers based on resting-state functional Magnetic Resonance Imaging (rs-fMRI) data is promising, but there is an ongoing need for further research on their accuracy. Therefore, we conducted a systematic review and meta-analysis to summarize and aggregate the available evidence in the literature so far. The systematic search resulted in the selection of 93 articles, whose data were extracted and analyzed through the systematic review. A bivariate random-effects meta-analytic model was implemented to investigate the sensitivity and specificity across the 55 studies (132 independent samples) that offered sufficient information for a quantitative analysis. Our results indicated overall summary sensitivity and specificity estimates of 73.8% (95% CI: 71.8-75.8%) and 74.8% (95% CI: 72.3-77.1%), respectively, and Support Vector Machine (SVM) stood out as the most used classifier, presenting summary estimates above 76%. Studies with bigger samples tended to obtain worse accuracies, except in the subgroup analysis for Artificial Neural Network (ANN) classifiers. The use of other brain imaging or phenotypic data to complement rs-fMRI information seem to be promising, achieving specially higher sensitivities ($p = 0.002$) when compared to rs-fMRI data alone (84.7% - 95% CI: 78.5-89.4% - versus 72.8% - 95% CI: 70.6-74.8%). Lower values of sensitivity/specificity were found when the number of Regions of Interest (ROIs) increased. We also highlight the performance of the approaches using the Automated Anatomical Labelling atlas with 116 ROIs (AAL116). Regarding the features used to train the classifiers, we found better results using the Pearson Correlation (PC) Fisher-transformed or other features in comparison to the use of the PC without modifications. Finally, our analysis showed AUC values between acceptable and excellent, but given the many limitations indicated in our study, further well-designed studies are warranted to extend the potential use of those classification algorithms to clinical settings.

Key-words: Autism Spectrum Disorder, Machine Learning, rs-fMRI, Meta-analysis.

List of Figures

List of Tables

List of abbreviations and acronyms

AAL	Automated Anatomical Labelling atlas
ABA	Applied Behavior Analysis
ABIDE	Autism Brain Imaging Data Exchange
AD	Autistic Disorder
ADI-R	Autism Diagnostic Interview - Revised
ADOS	Autism Diagnostic Observation Schedule
ANN	Artificial Neural Network
ASD	Autism Spectrum Disorder
AUC	Area Under the Curve
BOLD	Blood Oxygenation Level-Dependent
CC	Craddock brain atlas
CI	Confidence Interval
CNN	Convolutional Neural Network
DFC	Dynamic FC
DOR	Diagnostic Odds Ratio
DSM	Diagnostic and Statistical Manual of Mental Disorders
DTA	Diagnostic Test Accuracy
EEG	Electroencephalogram
ESDM	Early Start Denver Model
FC	Functional Connectivity
FIQ	Full IQ
fMRI	functional MRI
FN	False Negatives

FP	False Positives
FPR	False Positive Rate
HSROC	Hierarchical SROC
ICA	Independent Component Analysis
ICD	International Classification of Diseases
IQ	Intelligence Quotient
L-SVM	Linear SVM
LDA	Linear Discriminant Analysis
LOO	Leave-one-out
LR	Logistic Regression
$LR+$	Positive Likelihood Ratio
$LR-$	Negative Likelihood Ratio
ML	Machine Learning
MLP	Multilayer perceptron
MRI	Magnetic Resonance Imaging
MTL	Multi-task learning
MV/MT	Multi-view multi-task
MVL	Multi-view learning
NDAR	National Database of Autism Research
NPV	Negative Predictive Value
OOB	Out-of-bag
pAUC	partial AUC
PC	Pearson Correlation
PDD	Pervasive Developmental Disorders
PDD-NOS	PDD Not Otherwise Specified
PPV	Positive Predictive Value

QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies
ReHo	Regional Homogeneity
RF	Random Forest
RoB	Risk of Bias
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
rs-fMRI	resting-state fMRI
RSN	Resting-State Network
SLR	Systematic Literature Review
sMRI	structural MRI
SROC	Summary ROC
SVM	Support Vector Machine
TD	Typical Development
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
UMCD	UCLA Multimodal Connectivity Database
XGB	Extreme Gradient Boosting
y.o.	year/s old

Contents

1 Introduction

The Autism Spectrum Disorder (ASD) is a life-long neurodevelopmental condition associated to the atypical development of the brain. Individuals in this group, in general, present a slow development in certain activities when compared to individuals of Typical Development (TD) - such as speech, motor coordination, and social activities - and difficulties to communicate and relate to others (8, 9).

Typically identified in early childhood, ASD's development is believed to have both genetic and environmental roots (10, 11). Also, despite being considered a neurological disorder, the diagnosis of ASD remains exclusively based on behavioral criteria (12). This may be due to the great heterogeneity within the population, possibly reflecting an enormous amount of different neurodevelopmental etiologies (13, 14).

Epidemiological studies suggest an increase on its global prevalence in recent years and a systematic review published in 2012 estimated it to be about 0.62% (15).

The impact of this condition on the quality of life extends beyond the affected individual to the entire family. Parents of children with ASD report higher levels of stress even when compared to parents of children with other disabilities (16).

Aggravating the problem, the majority of researches regarding autism is based on data from high-income countries. This creates inequities across the world in access to services and supports (17).

On the other hand, given the plasticity of the brain during the first years of life, early detection paired with early treatment would have considerably stronger benefits compared with later treatments (18, 19).

The gold standard diagnosis of ASD is a standardized interview - the Autism Diagnostic Interview-Revised (ADI-R) (20) - in combination with a semi-structured standardized observation - the Autism Diagnostic Observation Schedule (ADOS/-2) (21) - and a differential diagnostic examination by experienced clinician. This is a long and time-consuming process that requires a multi-disciplinary team to assess information from various sources (22, 23).

In recent years, Machine Learning (ML) classifiers - algorithms that predict for each subject to which class it belongs, based on selected features that optimally represent the data for the problem at hand - have been increasingly applied to neuroimaging data for the diagnosis of psychiatric disorders, which includes ASD. Those classification methods hold the promise of increasing diagnostic accuracy and speeding up the diagnostic process (2, 24).

Throughout the different types of neuroimaging data, the resting-state functional Magnetic Resonance Imaging (rs-fMRI) is increasingly used to investigate neural connectivity and identify biomarkers of psychiatric disorders. It is based on spontaneous fluctuations in the Blood Oxygenation Level-Dependent (BOLD) signal obtained through a non-invasive and relatively fast acquisition process. Also, the rs-fMRI is task-free - requiring no active and focused participation of the patient - and the data can be easily combined to generate large databases (2, 25, 26).

Studies using rs-fMRI data have revealed patterns of brain functional connectivity that could serve as biomarkers for classifying depression (27), Parkinson's disease (28), ADHD (29), ASD (30), and even age (31). However, the reproducibility and generalizability of these approaches in research or clinical settings are debatable. There are many potential sources of variation across studies and its effect on diagnosis and biomarker extraction is still poor understood (26, 32).

Other promising paths on ASD diagnosis are currently under investigation. ML classifiers can be applied to different types of neuroimaging data, such as structural MRI (sMRI) (33) and electroencephalogram (EEG) (34). Beyond that, that are approaches not based on brain imaging data, such as urine analysis (35) and eye-tracking algorithms (36).

Systematic reviews and meta-analyses can be combined to evaluate the evidence in a given area of research qualitatively and quantitatively. Their application in the field of diagnostic test accuracy (DTA) is increasing in recent years, aiming to obtain summary estimates of those tests, identify factors that affect their accuracy, and identify areas for further research (7).

Therefore, we conducted a systematic review and meta-analysis of studies that used ML classifiers based on rs-fMRI data to distinguish patients with ASD from individuals of TD. We aimed to critically review the current literature on this area, evaluate the diagnostic accuracy of such classifiers, analyze the association between methodological differences and performance measures, and explore the applicability of those approaches in real-world settings.

This work is organized as follows: Chapter 2 presents the theoretical foundation used to base the study developed in terms of the Autism Spectrum Disorder, Magnetic Resonance Imaging, Machine Learning Classifiers, and Systematic Reviews and Meta-analyses (Sections 2.1, 2.2, 2.3, and 2.4, respectively); Chapter 3 introduces the methodology applied and the steps used to perform the proposed systematic review and meta-analysis; Chapter 4 presents the qualitative and quantitative results, the pertinent discussions and considerations, and the limitations of the work; Chapter 5 closes the study bringing the conclusions obtained.

2 Theoretical Foundation

2.1 Autism Spectrum Disorder

The Autism Spectrum Disorder (ASD) is a neurodevelopmental and multifactorial condition, characterized by difficulties in social interaction, delayed motor and cognitive development, difficulty in communication, high sensitivity to external stimuli, and repetitive behaviors (12).

Typically identified in early childhood, the disorder is currently considered one of the most common childhood morbidities. The symptoms become evident in toddlers and preschoolers and tend to persist throughout life, often becoming more muted (8, 9, 10).

This condition manifests itself in different levels of severity, overlapping normality at one extreme and intellectual disability with brain malfunction at the other. Children and adults with ASD present a variety of symptoms, but no individual manifests all the possible impairments. Also, no single symptom is sufficient for a diagnosis or invalidates it (8).

The presence of comorbidities is common among individuals within the spectrum and can contribute significantly to the impact of the disorder on their quality of life. Those comorbid symptoms can include insomnia, eating and digestive difficulties, allergies, anxiety, inattention, irritability, and behavior difficulties (9, 37).

2.1.1 Characterization

As stated in (38), ASD can be conceptualized as a series of discrete conditions such as those described in earlier versions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) (39) and in the International Classification of Diseases 10th Revision (ICD-10) (40), or as a spectrum disorder with hierarchical levels of severity, as described more recently in the the fifth edition of the DSM (39).

The DSM-IV Text Revision (39) refer to the autism spectrum as pervasive developmental disorders (PDD) and refer to autistic disorder (AD or simply autism) as the classic, more severe version of the condition. Asperger's syndrome refers to ASD children that did not presented delayed speech and whose IQ is at least 70. The PDD not otherwise specified (PDD-NOS) applies to ASD children who do not fulfill criteria for Asperger's syndrome or AD, forming a heterogeneous group generally less severely affected than those who have AD. The childhood disintegrative disorder was also comprised within the ASDs according to this version of the DSM (8, 38).

The fifth edition of the DSM (12) merged the PDD class into a single class of ASD, and a related disorder - social communication disorder - was added. The umbrella category of ASD started to encompass the previously distinct PDDs of autistic disorder, Asperger's disorder, and PDD-NOS (41).

The triad of impairments comprising social interaction, communicative behavior, and repetitive and restricted behaviors was collapsed into two domains, preserving restricted and repetitive behaviors - with the addition of a symptom cluster reflecting sensory difficulties - but merging social and communicative difficulties into a single domain (41).

DSM-V introduced a series of specifiers that provide information about the current presentation of a person meeting criteria for ASD. A first specifier describes whether a known etiological factor - such as medical condition, genetic syndrome, or environmental exposure - is present. The second, a severity specifier ranging from 1 to 3, describes the required level of support and impact on a person's levels of functioning for each domain of symptoms. The third and fourth specifiers indicate whether intellectual impairment or language impairment are present, respectively. The final specifier indicates whether catatonia is present (41).

2.1.2 Prevalence

In 2010, epidemiological data estimated the presence of 52 million cases of autism worldwide, equating to a prevalence of 7.6 per 1000 or one in 132 persons (38). In 2012, the global prevalence of some form of autism was estimated to be about 0.62%, which translates into one child out of 160 with a PDD (15).

There is evidence that the prevalence of ASDs is increasing around the world. This is most likely due to improved awareness and reporting, expansion of diagnostic criteria, enhancement of diagnostic tools, and changes in diagnostic practices, including expansion of developmental screening, increased diagnosis, and diagnostic substitution (9, 15).

According to (38), in children under 5 years of age, ASDs were the leading cause of disability, in terms of years lived with disability, among all mental disorders. Of the 291 diseases and injuries considered in the Global Burden of Disease 2010, the disorder was ranked among the 20 leading causes of disability for the under 5-year age group. In children aged from 5 to 14 years, ASDs were the fourth leading cause of disability out of the mental disorders.

Epidemiological evidence is, however, very limited to date, especially in low-income and middle-income countries. The majority of the individuals with autism live in those countries but the researches about the condition are largely generated and conducted in high-income countries. Barriers to research include financial inaccessibility, and the

need to validate and adapt diagnostic tools across a variety of contexts. Also, geographic, ethnic, cultural, or socioeconomic factors appear to be strongly associated with delays in diagnosis and difficulties to access and utilize services (9, 38).

2.1.3 Etiology

The etiology of ASD is still poorly understood but genetic and environmental factors are believed to account for its development (10, 11).

In (42), a population-based cohort of children born in Sweden was used to calculate the heritability of ASD. This cohort included 37,570 twin pairs, 2,642,064 full sibling pairs, and 432,281 maternal and 445,531 paternal half-sibling pairs, from which 14,516 children were diagnosed with ASD. The best-fitting model included additive genetic and non-shared environmental parameters, given an estimated ASD heritability of 83%.

A meta-analysis to investigate the prenatal, perinatal, and postnatal risk factors for children autism was carried out in (10) based in data from 37,634 autistic children and 12,081,416 non-autistic children enrolled in 17 studies. A variety of factors were associated with autism risk, such as: maternal and paternal age $\geq$ 35 years, gestational hypertension, gestational diabetes, and antepartum hemorrhage for the prenatal period; caesarian delivery, and gestational age $\leq$ 36 weeks for the perinatal period; low birth weight, postpartum hemorrhage, and brain anomaly for the postnatal period.

2.1.4 Impacts of gender, age, and IQ

Several studies indicate gender, age, and IQ differences on autistic symptoms and impairments. For example, (43) and (44) concluded that boys with ASD showed more restricted and repetitive behaviors than girls with ASD; significant though modest effects of IQ and age were found in (45), indicating increased Autism severity with decreasing IQ and age; (43) also found lower socio-communicative symptoms in older compared to younger individuals.

It is well known that ASD show an imbalanced male-female ratio and recent studies suggest values between 2:1 and 5:1 (46, 47, 48). There is also evidence that this ratio is lower in individuals with lower IQ (46, 49). Since most autism studies tend to follow this ratio or to include only male participants, the underrepresentation of females may have lead to an understanding of the disorder that is biased toward males (50). Females are generally diagnosed at a later age, and even with similar levels of severity of autistic traits, males are more likely to receive a diagnosis (51, 52, 53).

2.1.5 Diagnosis

Despite being considered a neurological disorder, the diagnosis of ASD remains exclusively based on behavioral criteria (12). This may be due to the great heterogeneity within the population, possibly reflecting an enormous amount of different neurodevelopmental etiologies (13, 14).

The gold standard diagnosis of ASD is a standardized interview - the Autism Diagnostic Interview-Revised (ADI-R) (20) - in combination with a semi-structured standardized observation - the Autism Diagnostic Observation Schedule (ADOS/-2) (21) - and a differential diagnostic examination by experienced clinician. This is a long and time-consuming process that requires a multi-disciplinary team to assess behavioral, historical, and parent-report information as well as further information from social environment and other health care institutions (22, 23).

The ADOS-2 consists of five modules to be administered based on the individuals' age, level of expressive language, and the appropriateness of assessment materials. Each module provides different tasks including playful elements and activities as well as verbal tasks intended to provide the examiner with information on social, communicative, play and stereotyped behavior. It is a very complex diagnostic instrument that includes a huge amount of different toys, picture books, newspaper, snacks, paper plates, cups, napkins, interview questions, and others (22, 21).

Individuals at a wide range of developmental and language level can be assessed by the ADOS-2: Module 1 for children who do not consistently use phrase speech; Module 2 for children who use phrase speech, but who are not verbally fluent; Module 3 for verbally fluent children and young adolescents; and Module 4 for verbally fluent older adolescents and adults. It also includes a Toddler module to assess children with limited language and age range between 12 and 30 months (22, 21).

The ADI-R is a standardized caregiver interview that appears as a valid instrument independently from age and level of functioning. Its questions intend to distinguish qualitative developmental deviance from developmental delay, but only cutoffs for autism are provided. Recommended diagnoses are based on cutoff scores for communication - different for non-verbal and verbal subjects - social interaction, and restricted/repetitive behavior. In addition, five items are included to clarify whether some developmental abnormality was present before the age of 36 months (54, 55, 20).

However, early detection of ASD - that is, before 24 months - continues to constitute a global challenge. One barrier is that the defining behavioral characteristics of ASD generally emerge during the second year of life with consolidation of the behavioral syndrome by about 24 months of age or later (56, 57, 58).

Aggravating the problem, the majority of tools for screening and diagnosis were

developed in high-income countries and present variable utility outside those settings and contexts. Also, the lack of adequate knowledge of mental health professionals tends to be a barrier for diagnosis (9).

2.1.6 Early intervention and treatments

Given the plasticity of the brain during the first years of life, early detection paired with early treatment would have considerably stronger benefits compared with later treatments, improving the quality of life of the individuals (18, 19).

Early intervention is a broad concept concerned with the delivery of structured and evidence-based support for the child, their family, and their community as a whole. It aims to address the core ASD deficits in communication, social interaction, and restricted, repetitive patterns of behavior through psycho-educational, developmental, and behavioral interventions. However, they are resource- and labour-intensive (9).

Educational therapies are the main form of intervention for ASDs in children. Throughout the different techniques available, the Applied behavior analysis (ABA) and the Early Start Denver Model (ESDM) are best supported by well-designed studies and, thus, considered evidence-based. The former encourages socially significant behaviors through a reinforcement learning technique, whereas the latter is an intervention based in ABA and designed for toddlers with signs of ASD (59).

Medications are not indicated for the treatment of ASD but can help address different symptoms. Also, alternative therapies do exist, but there is generally little research to support their use (59).

2.2 Magnetic Resonance Imaging

The Magnetic Resonance Imaging (MRI) is a sophisticated imaging technique that has evolved as a clinical modality over the past decades. It uses strong magnetic fields to orientate protons of water molecules in a determined direction. Then, by applying radio-frequency energy pulses at the appropriate resonance frequency, the protons leave the field-determined alignment. While they return to the alignment, they emit back a radio-frequency signal that is picked up by the examination machine. Based on the return speed, the technique allows determining the existence of different tissues (60, 25).

This technique had an important contribution for science and medicine by allowing the anatomical observation of in vivo tissues through non-invasive exams. However, MRI could only obtain information regarding tissue's shape and physicochemical characteristics (e.g., white and gray matter). The techniques for tissue functioning observation, especially in the brain, needed the use of contrast, often radioactive, which could be harmful to the

patient's health, preventing a high frequency of this type of exam.

2.2.1 Functional MRI

In (61) it was demonstrated the possibility to adapt the MRI to observe physiological activities in the brain, which has become a relatively inexpensive and non-invasive technique to study brain functioning known as functional MRI (fMRI).

Two concepts were the basis for this technique: the concept that the blood flow increases in a significant and localized way as the brain metabolism increases in regions of activity (62); and the concept that the hemoglobin molecule - which acts in the transport of oxygen to the cells of the body - has a differentiated magnetic behavior on its oxygenated and non-oxygenated forms (63).

The hemoglobin - present in the blood - contains iron, which distorts the magnetic field generated by an MRI machine. Vessels containing oxygenated arterial blood cause little or no distortion to the magnetic field in the surrounding tissue, while capillaries and veins containing blood that is partially deoxygenated distort the magnetic field in their vicinity. The distortion causes a deflection on the orientated protons that can be measured by the machine and this measure is called Blood-Oxygen-Level Dependent (BOLD) signal (64, 25).

As the blood flow increases, the availability of oxygen increases more than its consumption. Therefore, regions with higher brain activity have proportionally more oxygenated than deoxygenated hemoglobin. By measuring the BOLD signal it is possible to infer which brain regions are more active (25).

The BOLD signal is an indirect measure of brain activity since it does not measure directly the neuron activation but the difference in the blood flow of its adjacent region. Thus, the measure obtained is referent to a small brain volume called voxel. The measure is repeated for each voxel at various points in time, resulting in a time-series that represents the variation in voxel activity during the exam (25).

BOLD fMRI allows an image spatial resolution that is of the order of a few millimetres, with a temporal resolution of a few seconds (limited by the haemodynamic response itself) (25). The BOLD signal takes on average two seconds to increase when a stimulus is received and come to its maximum around three to six seconds. It returns to the normal state at a similar rate followed by a slight drop that takes around 12 to 30 seconds to stabilize (1). A representation of a typical BOLD hemodynamic response function is presented in Figure 1.

Since its discovery, BOLD fMRI using task-based paradigms has been critical to the current understanding of brain function (65). It has been used as a tool to delineate active brain regions, but has also been used to identify regional deactivation and shed

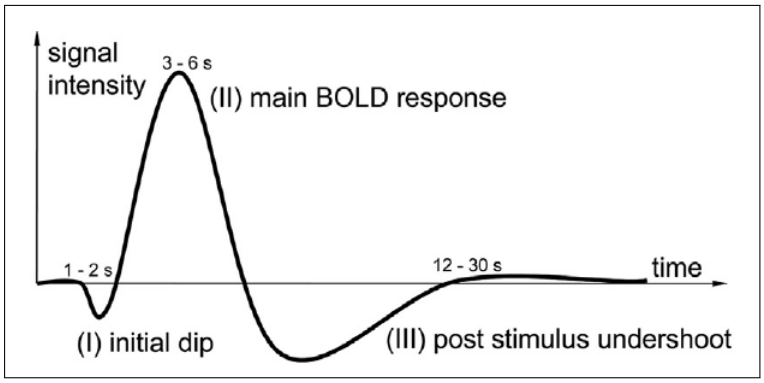


Figure 1 – Representation of a typical BOLD hemodynamic response function (extracted from (1)).

light on the interactions between different cerebral networks (1).

2.2.2 Resting-state fMRI

In recent years, there has been an increasing interest in the application of the fMRI at rest, called resting-state fMRI (rs-fMRI). It investigates synchronous activations between distinct brain regions that occur in the absence of a task or stimulus (65).

Both task-based and resting-state fMRI data are four-dimensional, acquired as a series of volumetric images over time. Each image takes 2-3 s to acquire, and rs-fMRI data is typically acquired during 5-15 min, with the subject asked to "lie still, think of nothing in particular, and not fall asleep" (66).

Since it was discovered that, even with the subject at rest, fMRI time-series from one part of the motor cortex were temporally correlated with other parts of the same functional network, rs-fMRI has been used to study spontaneous fluctuations in the brain activity (67, 66).

Many other brain networks with correlated temporal patterns in the resting condition have subsequently been identified These so called resting-state networks (RSNs) are consistent across subjects and persist even during sleep or under anaesthesia (66).

It is generally accepted that RSNs do reflect networks of brain function. Also, using rs-fMRI data, it is possible to found functional networks previously identified with task-based fMRI, providing useful complements to the inferences made from this type of data (66, 68).

The idea is that RSNs reflect the energy demands of groups of neurons that were wired together via firing with a common functional purpose (69). Besides, the existing evidence indicates that they are core functional networks in the mammalian brain. Therefore, these methods are being applied across multiple fields of neuroscience, improving the understanding of the organization of processing systems in the human brain (68).

2.2.2.1 Preprocessing

A rs-fMRI dataset requires a series of preprocessing steps before RSN analyses can be conducted. They are used to reduce the effects of artefacts (such as head motion and non-neural physiological signals), spatially align the functional data to the subject's structural scan, and may subsequently align the data into a standard-space reference coordinate system (66).

There are some controversial topics within those data preprocessing approaches, which includes whole-brain regression - used to regress out the average time course of the brain - and head-motion correction. Both of them may produce spurious correlations in rs-fMRI analysis and the latter, although less concerning in healthy young adults, poses challenges for the analysis of data acquired from children, older adults, and patients (65).

2.2.2.2 Analysis methods

After the application of the preprocessing approaches, various methods can be used to analyze rs-fMRI data. The first to be developed and applied was the seed-based analysis, which investigates the functional connectivity (FC) or, in other words, the similarities in the BOLD time-series between two or more regions of interest (ROIs) of the brain (65, 28).

Typically, the average BOLD time course of voxels within the ROIs are correlated with the time courses of all other ROIs in the brain and a threshold is determined to identify significant correlations. However, this approach requires a priori selection of ROIs, generally from brain atlases (65).

Figure 2 presents one example of approach to calculate FC measures. Two hundred ROIs were determined using the Craddock brain atlas (70). For each subject, the time-series of voxel activity were averaged over voxels in an ROI, and Pearson correlation coefficients were computed for all pairs of averaged time-series. This resulted in a 200×200 correlation matrix that is also symmetric, thus leaving $19,900$ unique measures per subject. Finally, each entry in the correlation matrix was transformed with Fisher's z-transform to obtain the connectivity values (2).

Another popular approach is the independent component analysis (ICA), a mathematical technique that can be used to spatially identify distinct RSNs. Compared to the previous method, ICA is rather a data-driven approach and does not necessarily need a previous assumption (65, 28).

The analyses can also be performed with the application of graph methods that consider the brain as a complex network formed by a set of nodes connected by edges. For example, ROIs can be represented as nodes and the correlation between them as the connectivity of the edges. Those methods can be used to study the topological organization of brain networks and assess their global and local properties (65, 28).

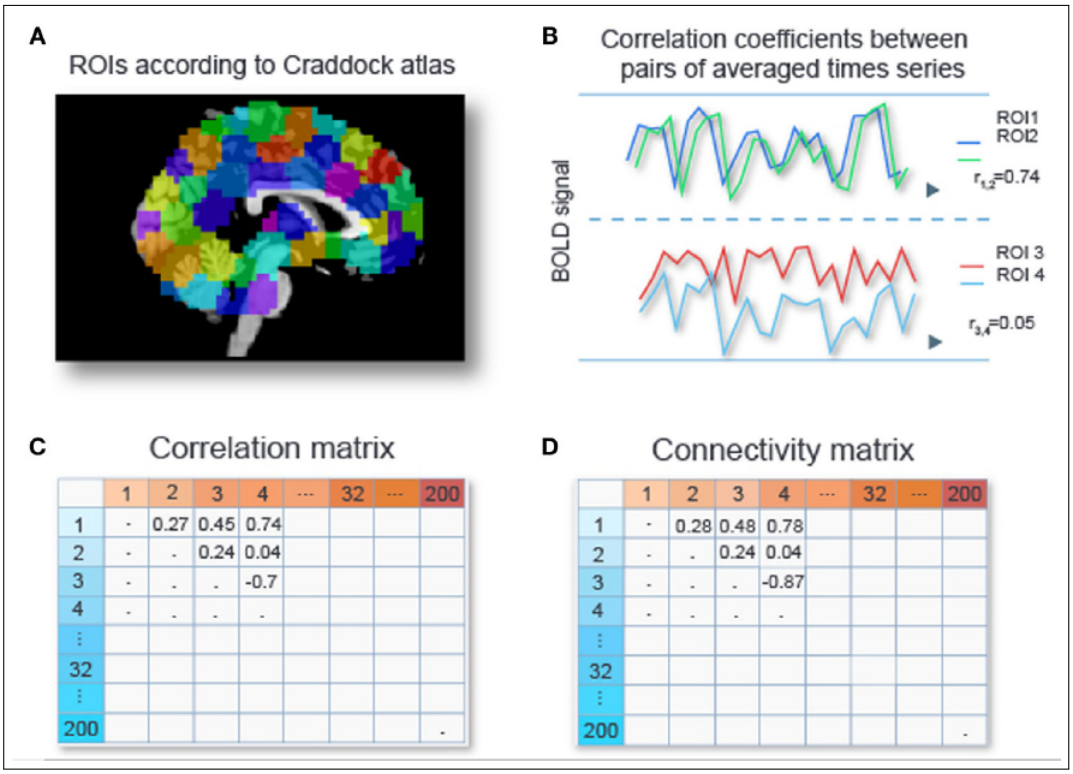


Figure 2 – Example of calculation of functional connectivity measures (extracted from (2)).

Other methods to analyze rs-fMRI data include regional homogeneity (ReHo) and effective connectivity. Knowing that intrinsic brain activity is manifested by clusters rather than single voxels, the ReHo evaluates the similarity between the BOLD signal of given voxels and that of the nearest voxels within a cluster. Effective connectivity, on the other hand, measures the causal and dynamic influence of one region to another and the information flow in particular brain regions (28).

2.2.2.3 rs-fMRI and ASD

The analysis of neuroimaging data has provided a means to explore the neurophysiological bases of ASD. Early studies applied MRI to investigate structural alterations in this disorder. The fMRI has become a common technique to observe functional variations in brain activity while performing a task. Also, the use of rs-fMRI analyses led to the development of resting-state functional connectivity research in ASD (30).

There is great interest in the characterization of brain network-level alterations in patients with ASD and the use of functional connectivity modeling approaches led to two primary theories: the expression of under-connectivity or over-connectivity in the brains of those patients (30).

Under-connectivity is viewed as a decrease in brain connectivity relative to a standard comparison value and can be present in a global level - between different nodes of a

network - or a local one - within a brain region. It indicates that the correlation between time-series signals in different voxels is decreased in affected compared to the unaffected subjects (30).

In the opposed direction, over-connectivity appears when statistically significant correlations are present in the affected subjects but are absent or less pronounced in the unaffected ones (30).

There are reports in the literature indicating both under- and over-connectivity in ASD, depending on whether local or global networks are under examination. Despite this lack of consensus, the use of altered connectivity as a biomarker for ASD diagnosis is still under extensive exploration (30).

2.2.2.4 Brain imaging repositories

As commented before, the rs-fMRI data can be easily combined to generate large databases and repositories (2). Such databases facilitate and make the development of studies using this type of data more accessible.

The most widespread repository of rs-fMRI data of individuals with ASD is the Autism Brain Imaging Data Exchange (ABIDE). ABIDE I (71) was the first imaging repository of rs-fMRI and corresponding structural data of individuals with ASD and of TD aggregated from multiple international institutions. The ABIDE II (72) came next to expand the initiative and accelerate the pace of discovery in the field of autism neuroimaging. Together, these repositories aggregate data from more than 2000 individuals collected across more than 24 international brain imaging laboratories, including a wide diversity in terms of gender, age group, and locality. There is also a preprocessed version of the ABIDE I database (73), comprising preprocessing approaches applied by five different teams using different tools and strategies.

We also highlight the National Database of Autism Research (NDAR[1]) and the UCLA Multimodal Connectivity Database (UMCD) (74). The first makes available fMRI data of at least 283 children and adolescents with ASD and of TD (75). The second is an openly available brain connectivity database that presents data in a preprocessed condition for various neurological diseases, including an ASD repository composed of 79 functional and 94 structural images (76).

2.3 Machine Learning Classifiers

The interpretation of brain imaging experiments require analysis of complex, multivariate data. Therefore, there has been growing interest in the use of machine learning

[1] http://ndar.nih.gov

(ML) algorithms for analyzing fMRI data. They can be used to, for example, decode stimuli, mental states, and behaviors (3). Besides, the application of classification methods to neuroimaging data might increase diagnostic accuracy and speed up the diagnostic process of psychiatric disorders. ML classifiers can detect biomarkers or subtypes of the disorders and also comorbidities, presenting the potential to facilitate the integration of neuroimaging data into clinical practice. (2).

algorithms that predict for each subject to which class it belongs, based on selected features that optimally represent the data for the problem at hand

A classifier is a function that can predict for each given example to which class it belongs, based on selected features that optimally represent the data for the problem at hand. In a neuroimaging setting, the features could be voxels and the class could be the type of stimulus the subject was looking at when the voxel values were recorded (3).

Classification methods have a number of parameters that have to be learned from training data - a set of examples reserved for this purpose. In this way, the classifiers generate a model of the relationship between the features and the class labels in the training set. Denoting an example by the row vector $x = [x_1...x_v]$ and its class label as y, the classifier is a function f that predicts the label $y = f(x)$ (3).

Once trained, the classifier can be tested on a different set of examples, the test data. The idea is that, if the classifier captured the relationship between features and classes, it should be able to predict the classes of examples it hasn't seen before. Thus, the predictions made by the classifier can be compared to the true labels of the samples to determine how well it performed the classification task (3, 2).

The training and testing examples are typically assumed to be independently drawn from an "example distribution". Thus, when judging a classifier on a test set, it is possible to obtain an estimate of its performance on any test set from the same distribution. The most commonly used measure to evaluate a classifier performance is its accuracy - the fraction of samples in the test set for which the class label was correctly predicted (3).

Generally, classifier-based analyses follow a series of specific steps, starting with the conversion of raw data into a set of examples and proceeding through the choice of classifier, training and test sets, and the interpretation of results (3).

2.3.1 Preparing data for classifier training

Before training a classifier, it is necessary to transform the fMRI data into examples by deciding what to use as features, how to extract their values from the data, and what will be predicted (3).

The examples can be created in various ways. In section 2.2.2.2 it was presented some of the methods used to analyze the rs-fMRI data and all of them can be used to obtain examples for classifier training, but many other approaches can be used for this purpose. For example, the average time-series of several voxels in one ROI could be used as a single feature or each voxel's time-series as a different feature (3).

There is an important trade-off associated with the number of examples produced. It is possible to have many noisy examples or fewer, cleaner ones, as a result of averaging images in the same class. Having more examples helps in the training set side since some classifiers require a certain number of examples to obtain good estimates of their parameters. However, some classifiers are particularly appropriated for analysis where few examples are available (3).

There are also some assumptions that should not be violated: there is a source distribution of examples from which they can be drawn; these draws are independent; and the training and test sets are independently drawn from this distribution. If those assumptions are violated, the results obtained could be biased (3).

Another issue to be kept in mind is the desirability of having the same number of examples in each class. When this is not the case, a classifier may tend to focus on the most numerous class and this can affect the interpretation of the accuracy results. For instance, 80% of accuracy may not be very good in a situation where 8 of 10 examples belong to one class and 2 of 10 to the other if the classifier simply predicted all of the examples to belong to the larger class by default. This problem can be alleviated by using performance measures based on sensitivity and specificity (3, 2) (more details on section 2.4.2.2.3).

In neuroimaging studies, there are generally many more features than examples. Therefore, it can be advantageous to select a subset of the available features that are of particular interest. This process is called feature selection and can enhance the accuracy, facilitate visualization of the data, and lead to faster classification (3, 2).

Another option to reduce the number of features is dimensionality reduction techniques. They transform the original feature space into a low-dimensional one, creating a new dataset matrix with the same number of examples but a reduced number of features (3). Thus, rather than selecting certain features, it is possible to work with selected combinations of the original ones. This process, however, can difficult the interpretation of results (2).

A final issue to consider is the preprocessing of the examples. Taking a matrix where each row is an example and each column is a feature, each roll could be normalized to have mean 0 and standard deviation 1. In this case, the idea is to reduce the effect of large, image-wide signal changes, but other preprocessing techniques could be applied in

different scenarios (3).

2.3.2 The classifiers

Various factors can influence the choice of a classifier, which includes the dimensions of the dataset, the feature selection method, the required classification speed, and the statistical properties of the data (2).

When there are many more features than examples, such as in a typical fMRI study, the ML algorithms tend to easily find a function that can classify the examples in the training set. However, this does not necessarily mean that the classifier will have a good performance in the test set, a phenomenon called overfitting (3).

The choice of a simple function among those that do well in the training set can be a way to mitigate the danger of overfitting. It can be done, for example, by having the prediction depend on a linear combination of the features (3).

Beyond their simplicity, linear classifiers have the advantage of each feature affecting the prediction only via its weight and without interacting with other features. Thus, it is easier to measure the influence of each feature on the results obtained, which facilitates interpretation (3).

A more geometric intuition for how linear classifiers work is given in Figure 3. In this simplified model, there are two voxels whose values are used to characterize the examples in each of the two classes. Learning a linear classifier is equivalent to learning a line that separates the points in the two classes (3).

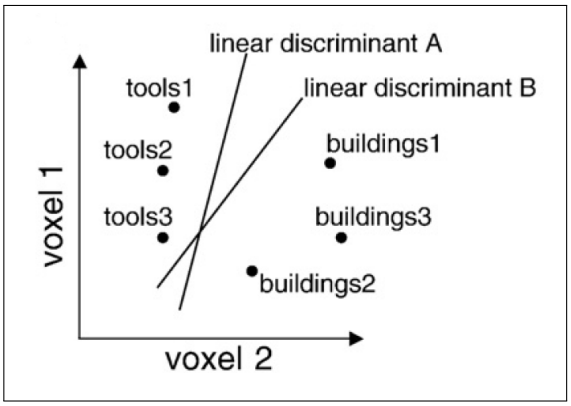


Figure 3 – Representation of the geometric intuition behind linear classifiers - equivalent to learning a line that separates examples in two classes (adapted from (3)).

Many possible linear discriminants can be used for classifying the examples in Figure 3, which should be chosen by the classifier. Also, there may be many possible settings of classifier parameters that lead to good predictions. The process of guiding the procedure that sets the parameters is called regularization, which takes the shape of an

extra parameter (also called hyperparameter) that must be tuned also using the training set (3).

There are more complex classifiers that can let interactions between features and use nonlinear functions. More complex models tend to be more powerful but generally have more hyperparameters. Therefore, they are potentially more capable of explaining noise in the data and overfitting (77). Also, it is not clear that those complex models always provide a significant advantage in practical performance. However, this can be a reflection of the small number of examples available instead of indicating an absence of complicated relationships between features (3).

In the next sections we briefly introduce some of the most known and used classifiers through the rs-fMRI literature.

2.3.2.1 Support Vector Machine

The basic idea of the linear Support Vector Machine (SVM) is to construct a linear decision boundary with the largest possible distance to example points. Thus, instead of minimizing expected empirical loss on the training data, SVM attempts to minimize expected generalization loss (4). It belongs to the category of regularized predictors - a regularization term (C) determines to what extent misclassification of data samples is accepted (2).

If the data are not linearly separable in the original feature space, it is possible to embed the data into a higher-dimensional space using a so-called kernel function to achieve separability (78), as exemplified in Figure 4. Linear SVM, however, have so far been more successful for the classification of ASD based on rs-fMRI data than kernel SVMs (2).

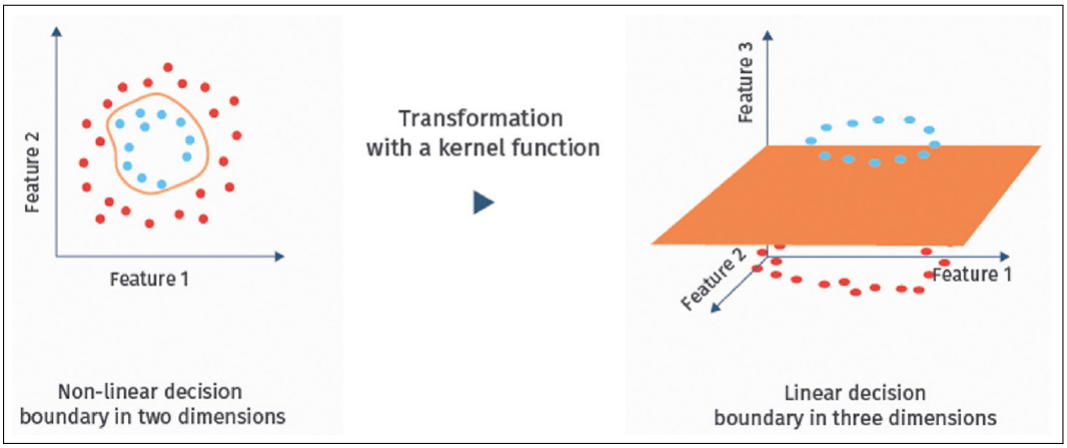


Figure 4 – Example of kernel mapping. The left shows a dataset that cannot be separated linearly in the two feature dimension space, whereas the right shows a three-dimensional embedding, where linear separation is possible (extracted from (2)).

It is well-known that SVMs can handle noisy, correlated features and high-dimensional data sets well (2). Also, SVMs combine the advantages of non-parametric and parametric models, having the flexibility to represent complex functions, but being resistant to over-fitting (4). Hence, they have become one of the most successful classifiers of the recent years, also for the classification of fMRI data (2).

2.3.2.2 Artificial Neural Network

As a branch of the Computational Intelligence area, which takes base from nature to develop intelligent computer-based systems, Artificial Neural Networks (ANNs) model aspects of how human brain works. They are massively parallel distributed processing systems made up of a collection of interconnected neural computing elements, called neurons, that incrementally learn from their environment (usually a complex dataset) to identify essential linear and/or nonlinear trends. A mathematical model for a neuron is presented in Figure 5. So, ANNs provide reliable predictions for new datasets containing even noisy and/or partial information and are widely applied in research because they can model highly nonlinear systems in which the relationship among the variables is unknown or very complex (79, 80, 81, 82). As a simplified model based on the human brain architecture and operation, an ANN has the ability to learn and acquire knowledge. The ANN's learning process is based on examples, starting from a training phase with known information of a problem to pick up knowledge about it. After being properly trained, an ANN can be used to solve unknown or untrained instances of the problem. ANNs have been used in solving problems such as pattern recognition/classification, prediction or function approximation, clustering, image processing, data compression, forecast, optimization, data and signal (time-series) classification etc (79, 80, 81).

The ANN learning system is directly linked to its topology, which is determined by the way the neurons are connected. A simple example of how the neurons can be connected is also presented in Figure 5. The multilayer perceptron (MLP) architecture is the oldest one, and still largely used for researches, been compatible with a large number of training softwares. The MLP consists of three or more sets of neurons, called layers, and each one of those layers receives a input and a bias (a constant value, used to calibrate the calculations) and generates an output. The first layer receives an external input, and its output is used for the next layer, which do the same until the last layer, whose output is a value in a previously determined range or a classification between two or more classes, also previously determined (83).

Many ANN classifiers are also deep-learning methods. Those methods use multiple levels of representation, obtained by composing simple but nonlinear modules that each transform the representation at one level into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex

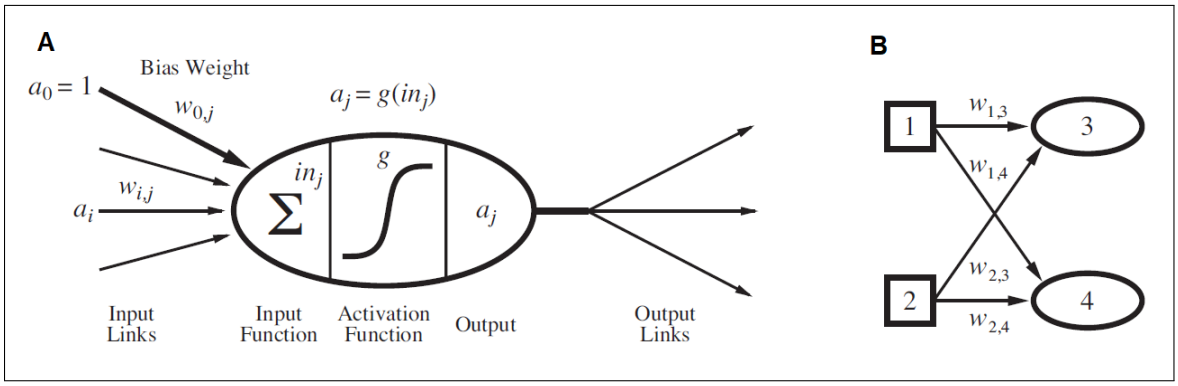


Figure 5 – **A**. A mathematical model for a neuron. Its output activation is $a_j = g(\sum_{i=0}^{n} w_{i,j} a_i)$, where a_i is the output activation of neuron i and $w_{i,j}$ is the weight on the connection from neuron i to this neuron. **B**. A network with two input and two output neurons (adapted from (4)).

functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations (84).

One example of such deep-learning methods is the Convolutional Neural Network (CNN). It is inspired by the visual system's structure and is designed to process data that come in the form of multiple arrays (84). This model is very useful for pattern recognition tasks and is very capable of recognizing similarities from images, even with cardinal position alterations (85).

2.3.2.3 Logistic Regression

Logistic regression (LR) is a mathematical modeling approach that uses the logistic function to describe the relationship of several independent variables to a dichotomous dependent variable (86). Thus, LR models the probability of the dependent variable belonging to a particular category (78).

LR can be viewed as a special case of a generalized linear model, where the log odds is modeled as a linear function of the predictors. A convenient property of this model is that the sizes and signs of the estimated coefficients have a clear interpretation. Also, it is possible to introduce a regularization parameter to the LR, obtaining regularized variants of it (2).

2.3.2.4 Multi-view/Multi-task

Multi-view learning (MVL) is concerned with the problem of ML from data represented by multiple distinct feature sets (87). Conventional ML algorithms concatenate all multiple views into one single view to adapt to the learning setting. However, this concatenation causes overfitting in the case of a small training sample and is not physically

meaningful because each view has a specific statistical property. The MVL introduces one function to model a particular view and jointly optimizes all the functions to exploit the redundant views of the same input data and improve the learning performance (88). Therefore, MVL is a very promising topic with widespread applicability (87).

On the other hand, many practical problems are similar or related to each other. The main goal of multi-task learning (MTL) is to encode the intrinsic relationship among different tasks, tending to learn multiple tasks simultaneously and often obtaining better results (89). Considering the neuroimaging field, the feature learning on each image modality can be denoted as a single task (90), different tasks can correspond to the prediction of different variables (91), or even the scenario of multiple center classification can be addressed with MTL by considering each imaging center as one task (89).

Nevertheless, there are many real-world problems that involve both MVL and MTL. In these problems, a single learning task might have features in multiple views, and different learning tasks might be related to each other through one or more shared views. This forms a more challenging problem that includes both previous approaches, the multi-view multi-task (MV/MT) learning (92).

2.3.2.5 Random Forest

The Random Forest (RF) is a tree-based ensemble machine learning method often used for classification and regression analyses. RF grows many binary decision trees at training time with the aim to increase prediction accuracy via model averaging. Training data are randomly drawn with replacement to construct the trees and excluded data - termed out-of-bag (OOB) sample - are used for testing (14).

The process of splitting the trees using this method forces it to consider only a subset of the predictors available, which can be thought as decorrelating the trees, thereby making the average of the resulting trees less variable and more reliable (78). In addition, the OOB error gives estimates of the generalization error and may remove the need for a set-aside test set (93).

2.3.3 Training and testing

As said before, the classifiers cannot be trained and tested on the same data if the goal is to obtain a useful estimate of the classifier's true accuracy. Also, using just a few examples for testing will not lead to a good estimate (3).

To determine the test accuracy, sensitivity, or specificity, the data are usually split not only once into a training set and a test set, but repeatedly. In particular, the data are randomly split into k disjoint sets of approximately equal size, called folds. Each fold is used once as a test set, while all other folds combined then serve as the training set. This

procedure is called k-fold cross-validation (2). Figure 6 shows an 8-fold cross-validation as an example.

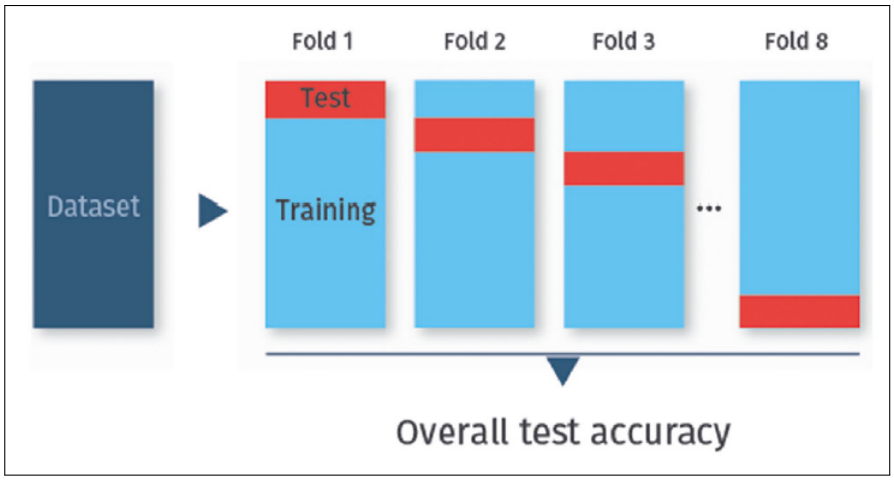


Figure 6 – An example of k-fold cross-validation with 8 folds (extracted from (2)).

The most common cross-validation schemes are leave-one-out cross-validation (LOO cross-validation), where k equals the number of data samples, and 10-fold cross-validation ($k = 10$). For most data sets, 10-fold cross-validation is a good compromise with regard to bias and variance (2).

Cross-validation can also be used in combination with feature selection or the selection of tuning parameters. In this case, nested cross-validation must be used to avoid optimistically biased performance estimates. The idea is to start with a regular k-fold cross-validation, called the outer loop, but in this case each training set is split again into several folds, the inner cross-validation loop used to try out different feature subsets or classifier tuning parameters. The best performing classifier from the inner loop is then applied to the test set from the outer loop to obtain the performance estimates (2).

Ideally, after cross-validation the optimized classifier is applied to an entirely new and independent data set to obtain a better measure of how well it generalizes. However, performing this step assumes the availability of enough data (2).

2.3.4 Evaluating results

When training a classifier, its true accuracy should, ideally, be better than that of a classifier deciding at random. The true accuracy is the probability that the classifier will correct label a new example drawn at random from the same distribution that the training examples came from. The accuracy on the test set is thus an estimate of the true accuracy of a classifier. How precise this estimate is depends on the size of the test set - the fewer examples used, the greater the variability of the estimator (3).

It is common for neuroscientific studies to compare classification accuracy to chance level - the accuracy achieved assuming that it is equally likely for a data sample to fall in one of the existing classes. In the case of a balanced two-class problem, chance level classification accuracy would be equal to 50% (2).

However, chance level accuracies are theoretical values derived from random guessing on data sets of infinite size. Although random guessing will approximate chance level if the dataset is large enough, for small datasets random classification can deliver accuracies strongly deviating from chance level. Instead of comparing classification results to a theoretical chance level, parametric or non-parametric statistical tests can be applied, taking into account the data size (2).

2.4 Systematic Review and Meta-analysis

A systematic review uses systematic methods to identify, select, and critically appraise relevant research and to analyze data from the primary studies included in the review, evaluating a body of evidence in the literature both qualitatively and quantitatively. A meta-analysis can be part of the systematic review and uses statistical methods to integrate the results of multiple primary research studies (7). In other words, meta-analysis is a set of statistical techniques for combining results from two or more separate studies (5).

With the continued publication of primary scientific research studies and the recognition of their importance, the value of systematic reviews and meta-analyses for summarizing results is also being increasingly acknowledged (7). Quantitative aspects have a key role to play in evidence synthesis and, therefore, meta-analysis features in almost every systematic review and continues to undergo rapid development (94).

Due to their methodological rigor, systematic reviews and meta-analyses have become increasingly important in health care and are used to support the development of clinical practice guidelines, keep clinicians up to date with their field and inform clinical decision-making (95, 96).

2.4.1 Systematic Literature Review

A systematic literature review (SLR, also referred as systematic review) is conducted to identify, evaluate and interpret all available research relevant to a particular research question, or topic area, or phenomenon of interest (97). It aims to give a complete, comprehensive and valid picture of the existing evidence. Thus, the identification, evaluation and interpretation must be conducted in a scientifically and rigorous way, adopting a precise, transparent and explicit approach that includes a series of phases (98, 99).

Unless a literature review is thorough and fair, it is of little scientific value. This is the main rationale for undertaking systematic reviews (97). By taking a highly procedural approach to define the problem of study and search the available literature, the technique aims to avoid the danger of selection bias, in which only a subset of studies are found out (100).

The reasons for undertaking a systematic review, as articulated in (97), are: to summarise the existing evidence concerning a treatment or technology; to identify gaps in current research and suggest areas for further investigation; to provide a framework/background and appropriately position new research activities; to examine the extent to which empirical evidence supports/contradicts theoretical hypotheses or assist the generation of new hypotheses.

According to the guidelines presented in (98), a SLR is conducted following three steps that will be presented in the next sections: planning, conducting, and reporting. This process is likely to be highly iterative, with many transitions backwards and forwards among the activities (100).

2.4.1.1 Planning the review

The need for a systematic review originates from the aim to understand the state-of-the-art in an area or to use empirical evidence in strategic decision-making or improvement activities. If there are SLRs available in the field, they should be appraised regarding scope and quality, to evaluate if they are sufficient to meet the current needs for a review (98).

The area of the systematic review and the specific research questions set the focus for the identification and extraction of data from the studies and the analysis to be conducted. Hence, the research questions must be well thought through and defined, taking into account the population in which the evidence is collected, the intervention applied in the empirical study, to what the intervention is being compared, the statistical and practical significance of the outcomes, the context of the study, and the experimental designs to be included (98).

An important part of this procedure is to document the planned activities for conducting the systematic review as a protocol. It facilitates the review of the plan and ensure that decisions are made so as to support a review that is as repeatable and rigorous as possible (100).

2.4.1.2 Conducting the review

Conducting the review means setting the review protocol into practice. This involves several steps that are presented in the next sections.

2.4.1.2.1 Identification of research

The purpose of a systematic search is to identify as many studies on the topic of interest as possible. To achieve this, a comprehensive search strategy should be developed and documented. Also, the final search strategy should be reported in sufficient detail to ensure that its process is repeatable (7, 100).

This step involves specifying search strings and applying them to databases. It may also include manual searches in journals and conference proceedings, as well as systematically searching for primary studies based on references to and from other studies in a process called "snowballing" (98).

The search strategy is a trade-off between finding all relevant primary studies, and not getting an overwhelming number of outcomes that are wrongly considered to be of interest - which must be excluded manually (98).

The search string is developed from the area to be covered and the research questions. Using multiple databases is a necessity to cover all relevant literature, but it also creates duplicates, which must be identified and removed. At the end, the papers found are a sample of the population of all papers on a specific topic. The key issue is that the sample is indeed from the intended population (98).

2.4.1.2.2 Selection of primary studies

The basis for the selection of primary studies is the inclusion and exclusion criteria. The criteria should be developed beforehand, to avoid bias. However, they may have to be adjusted during the course of the selection, since all aspects of inclusion and exclusion are not apparent in the planning stage (98).

During this round of filtering, only primary studies should be selected for inclusion in the systematic review. That is, researchers should analyze only reports of studies that directly examined the research question (100).

The identified set of candidate studies are processed related to the selection criteria. For some studies, it is sufficient to read the title or abstract to judge the paper, while other papers need a more thorough analysis of, for example, the methodology or conclusions to determine its status (98).

As the selection process is a matter of judgments, also with well defined selection criteria, it is advised that two or more researchers assess each paper, or at least a random sample of the papers (98).

Still, the quality of each study included in the analysis must be assessed so that this can be considered when the results from each study are compared and contrasted (100). This topic is discussed in more details in section 2.4.3.4.

2.4.1.2.3 Data extraction

From each study that remains after the selection is performed, the required data for the analysis must be extracted. Thus, a data extraction form is designed to collect the information needed from the primary study reports (98, 100).

The form should be designed based on the research questions. For pure meta-analytical synthesis, the data is a set of numerical values, representing number of subjects, objects characteristics, treatment effects, confidence intervals, etc. For less homogeneous sets of studies, more qualitative descriptions of the primary studies must be included (98).

The data extraction form should be piloted before being applied to the full set of primary studies. If possible, the data extraction should be performed independently by two researchers, at least for a sample of the studies, in order to assess the quality of the extraction procedure (98).

If a primary study is published in more than one paper, only one instance should be counted as a primary study. Mostly, the journal version is preferred, as it is most complete, but both versions may be used in the data extraction. Supporting technical reports, or communication with authors may also serve as data sources for the extraction (98).

2.4.1.2.4 Data synthesis

The guidelines for how the evidence is to be integrated are not much specific. This occurs because the methods which are feasible for each systematic review will depend largely on how much and what type of evidence has been utilized, and on the specific research question under study. Although qualitative measures are allowed, it is recommended to convert each to a quantitative measure if at all possible (100).

Qualitative synthesis refers to a systematic review used to provide the descriptive statistics without statistical pooling, whereas quantitative synthesis refers to a meta-analysis performed to generate summary estimates of a test's diagnostic accuracy (7), which is discussed in section 2.4.2.

It is important to identify whether results from studies are consistent with one another (i.e. homogeneous) or inconsistent (i.e. heterogeneous) (97). Less formal methods for data synthesis include descriptive or narrative synthesis. These methods structure raw evidence and interpretations using tabulation of data, groupings and clustering, or vote-counting as a descriptive tool that brings light to the research question (98).

Independently of the synthesis method, a sensitivity analysis should take place to analyze whether the results are consistent across different subsets of studies (98). This topic is discussed in more details in section 2.4.3.6.

2.4.1.3 Reporting the review

Like any other empirical study, the SLR may be reported to different audiences. In particular, if the purpose of the review is to influence practitioners, the format of the report has to be tailored well to its audience. For academic audiences, the detailed reporting of procedures for the study is critical for the ability to assess and evaluate the quality of the SLR (98).

The process of selecting relevant literature should be presented as a flow chart, which is usually the first figure presented in any systematic review and meta-analysis (7). The reporting ideally includes changes to the study protocol, complete lists of included and excluded primary studies, data on their classification, as well as the raw data derived from each of the primary studies. If space constraints do not allow all details being published, a supporting technical report is recommended to be published online (98).

2.4.2 Meta-analysis

The rationale for meta-analysing is to obtain valid summary estimates and provide information on factors affecting estimates to help readers decide how to generalize results to their settings (101). Obstacles such as time, cost and expert researchers make it difficult to conduct research studies with large samples and meta-analyses make it possible to obtain a large sample size by combining different studies (102). It helps to make sense of apparently conflicting study results, as it identifies which differences are likely to be real, which are explicable by chance, and which can be explained by known differences in study characteristics (5). Also, it can be useful to identify areas for further research.

Given that studies differ in various ways (such as design, sample and results reported), the goal of a meta-analysis is almost invariably to broaden the base of studies in some way, expand the question and study the pattern of answers. Thus, the feasibility of performing a meta-analysis and what kinds of studies should be included depends directly to the specific goals of the analysis to be done (103).

Meta-analysis was traditionally applied to studies that had a form of intervention. However, it can include other outcomes and other types of studies. In health care, it is being applied to diagnostic accuracy studies to measure the ability of a new test to detect the presence or absence of a specific disease or condition (104).

2.4.2.1 Differences between diagnostic test accuracy and intervention meta-analyses

A meta-analysis of intervention summarizes the effectiveness of an experimental intervention compared with a comparator intervention. First, a summary statistic (Yi) is calculated to describe the observed intervention effect for each study. Then a summary

intervention effect estimate (M) is calculated as a weighted average of the effects from the individual studies (105), as follows:

$$M = \frac{\sum W_i Y_i}{\sum W_i},$$

Where W_i is the weight assigned to study i.

If it is assumed that each study is estimating exactly the same quantity and the variation between the studies is due to random error, then a fixed-effect meta-analysis is performed (105). Considering inverse-variance weights, $W_i = \frac{1}{V_i}$, where V_i is the within-study error variance of study i (106).

The combination of intervention effect estimates across studies may optionally incorporate an assumption that the studies are estimating different, yet related, intervention effects. For example, the effect size might be higher (or lower) in studies where the participants are older, or more educated, or healthier than in other studies. Thus, it is assumed that the effect sizes in the different studies represent a random sample from a particular distribution (conventionally a normal distribution). This results in a random-effects meta-analysis and allows to address heterogeneity that cannot readily be explained by other factors (105, 106). In this case, the weights would be $W_i = \frac{1}{V_i + T^2}$, where T^2 is the between-study variance.

Likewise, a meta-analysis of diagnostic test accuracy (or DTA meta-analysis) summarizes the ability of a new test, called index test, to detect the presence or absence of a specific disease or condition based on a reference standard (104). However, a DTA meta-analysis presents some specific issues.

Evaluating test accuracy require knowledge of two quantities, the test sensitivity and specificity (see section 2.4.2.2.3). Thus, meta-analysis methods for DTA have to deal with two summary statistics simultaneously rather than one. Also, it has to allow for the trade-off between sensitivity and specificity that occurs between studies (5) (see section 2.4.2.2.4).

Heterogeneity is to be expected in results of test accuracy studies. Therefore, to account for both sensitivity and specificity, the relationship between them, and the heterogeneity in test accuracy, it is necessary to fit hierarchical random effects models, which results in more challenging statistical aspects (5).

2.4.2.2 Key concepts for DTA meta-analysis

2.4.2.2.1 Types of results data

Results from DTA studies can be presented in three data types (5):

- Binary - the test result is reported as positive or negative.
- Ordinal - the test result is reported on a set of ordered categories.
- Continuous or Count - the test result is reported on a continuous scale or as a count.

To be included in a meta-analysis, the results need to be re-categorized as binary by selecting a threshold and presenting the data as a 2×2 table (5).

2.4.2.2.2 The 2×2 table

The data from a primary study can be presented in a 2×2 table showing the cross classification of condition status (result of the reference standard) and test outcome (result of the index test) (5) as can be seen in Table 1.

Table 1 – A 2×2 table

	with condition	without condition
Test positive	a	b
Test negative	c	d

The numbers a and b are the numbers of true positives (TP) and false positives (FP) whereas c and d are the numbers of false negatives (FN) and true negatives (TN), respectively.

2.4.2.2.3 Summary measures of test accuracy

From the 2×2 table, summary measures of test accuracy can be computed either as proportions of the condition positive or negative (such as sensitivity and specificity) or test positive or negative (such as predictive values). The most common summary measures are presented in this section.

The sensitivity (or true positive rate - TPR) is the probability that the index test result will be positive in a case with the condition. Likewise, the specificity is the probability that the index test result will be negative in a case without the condition. Also, the term false positive rate (FPR) is used for the complement of specificity. From Table 1, $sens = \frac{a}{a+c}$, $spec = \frac{d}{b+d}$, and $FPR = 1 - spec = \frac{b}{b+d}$ (5). From all the summary measures, sensitivity and specificity are the most commonly reported (107).

The positive predictive value (PPV) is the probability that a case with a positive index test result has the condition. In the same way, the negative predictive value (NPV) is the probability that a case with a negative index test result do not have the condition. Thus, $PPV = \frac{a}{a+b}$ and $NPV = \frac{d}{c+d}$ (5). The PPV and the NPV are measures that depend

on the prevalence of the target condition. Since the prevalence might vary across studies, using these quantities for meta-analyses is somewhat more complicated (104).

The positive likelihood ratio ($LR+$) describes how many times more likely positive index test results were in the group with the condition compared to the group without the condition. Similarly, the negative likelihood ratio ($LR-$) describes how many times less likely negative index test results were in the group with the condition compared to the group without the condition. If the test is informative, the $LR+$ should be greater then 1 while the $LR-$ should be less then 1. They are defined as $LR+ = \frac{sens}{1-spec} = \frac{a(b+d)}{b(a+c)}$ and $LR- = \frac{1-sens}{spec} = \frac{c(b+d)}{d(a+c)}$ (5). It was advised in (108) to consider meta-analysis of sensitivity and specificity values instead of likelihood rations once summarizing LR across studies may lead to impossible summary estimates for sensitivity and specificity and the approaches found to deal with it properly were too complicated.

The diagnostic odds ratio (DOR) summarizes the diagnostic accuracy of the index test as a single number that describes how many times higher the odds are of obtaining a positive test result in a person with the condition than in a person without the condition. It is defined as $DOR = \frac{LR+}{LR-} = \frac{sens \times spec}{(1-sens)(1-spec)} = \frac{ad}{bc}$. Expressing accuracy in terms of ratios of odds means the measure has little direct clinical relevance, and it is rarely used as a summary statistic in primary studies. Paired summary statistics are more clinically useful as they distinguish between the two dimensions of test accuracy. The DOR does, however, remain an important element in meta-analytic model building (5).

2.4.2.2.4 ROC curves

Binary test outcomes are defined on the basis of a threshold for test positivity and change if the threshold is altered. In the case of sensitivity and specificity, the dependence induces a trade-off between the two quantities, one value increasing whilst the other decreases as the threshold is moved (5).

Figure 7 illustrate this relationship. Each panel uses a different threshold to define test positive. The sensitivities are measured by the proportion of the area under the "diseased" curve to the right of the threshold while the specificities are measured by the proportion of the area under the "non-diseased" curve to the left of the threshold. We can see that as the threshold decreases, the sensitivity increases and the specificity decreases.

Primary studies that evaluate a test at several thresholds can present results as receiver operating characteristic curves (or ROC curves). The ROC curve of a test is the graph of the values of sensitivity and specificity that are obtained by varying the threshold across all possible values. The graph usually plots sensitivity against FPR. The curve for any test moves from the point where sensitivity and FPR are both 1 (all participants are classified as test positive) to the point where both are zero (all participants are classified

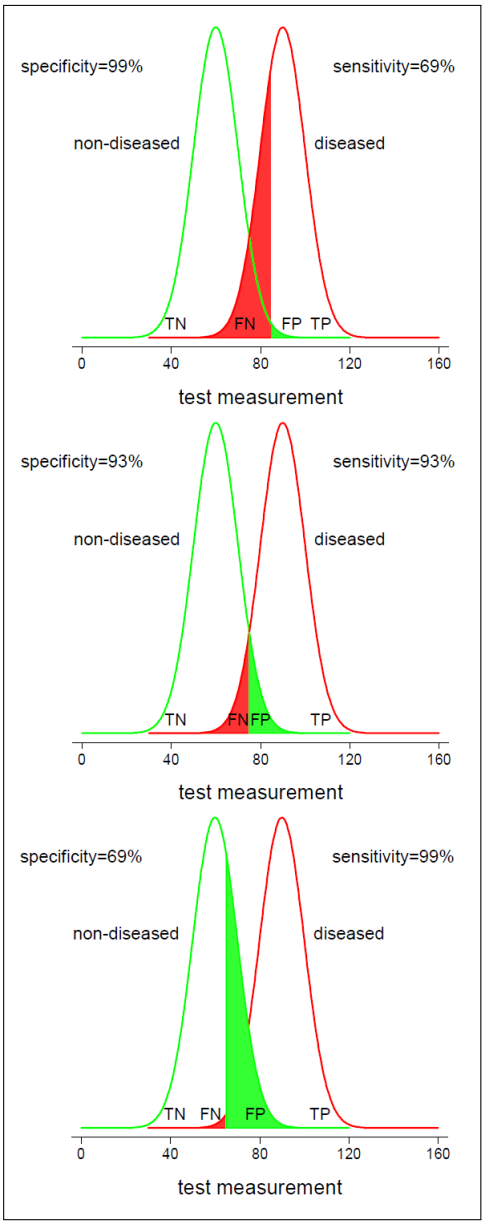


Figure 7 – Relationship between sensitivity, specificity and the threshold (adapted from (5)).

as test negative) (5). An example of ROC curve is shown in Figure 8.

The position of the ROC curve depends on the discriminatory ability of the test. The more accurate the test is, the closer the curve to the upper left hand corner of the ROC plot (where sensitivity and specificity are 1). In a completely uninformative test, the ROC curve would be the upward diagonal of the square (5).

From the ROC curve, the area under the curve (AUC) can be calculated to summarise the accuracy of the test across all possible thresholds. It is the probability that if a pair of individuals with and without the condition is selected at random, the individual with the condition will have a higher test result than the one without the condition. For a perfect test the AUC would be equals 1 while for a completely uninformative one the

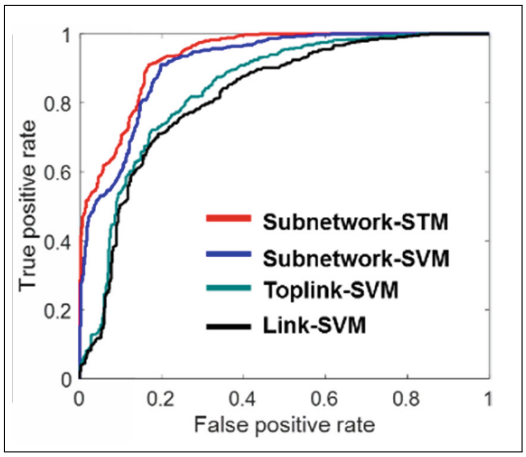


Figure 8 – Example of ROC curves of different techniques to classify ASD versus control (extracted from (6)).

AUC would be equals 0.5 (5).

ROC curves are sometimes described by quoting a point known as Q* (or Q-point) where the ROC curve intersects the diagonal line running from the top left corner to the lower right corner. By definition, at this point the sensitivity and specificity values are equal. However, the use of Q* values often give the wrong impression of the accuracy, particularly if ROC curves are asymmetric, or the study points lie away from the downward diagonal (5).

2.4.2.3 Descriptive Plots

In this section, the two main forms of graphical display used in DTA meta-analyses are presented - summary ROC and forest plots.

2.4.2.3.1 Summary ROC plots

The Summary ROC (SROC) plot (Figure 9.A) is a descriptive plot that merely displays the results of individual studies in ROC space with each study being plotted as a single sensitivity against FPR point and its 95% confidence region (7). Thus, the SROC plot depicts the scatter of the study results (5).

The sizes of the points can be controlled to depict the precision of the estimate or according to their sample sizes. Also, "cross-hairs" can be added to each study point to indicate confidence limits for sensitivity and specificity (5).

It differs from an SROC curve (Figure 9.B) - a statistically estimated meta-analytic summary line in the ROC space (7). SROC curves and summary sensitivity and specificity points can be added to an SROC plot (5).

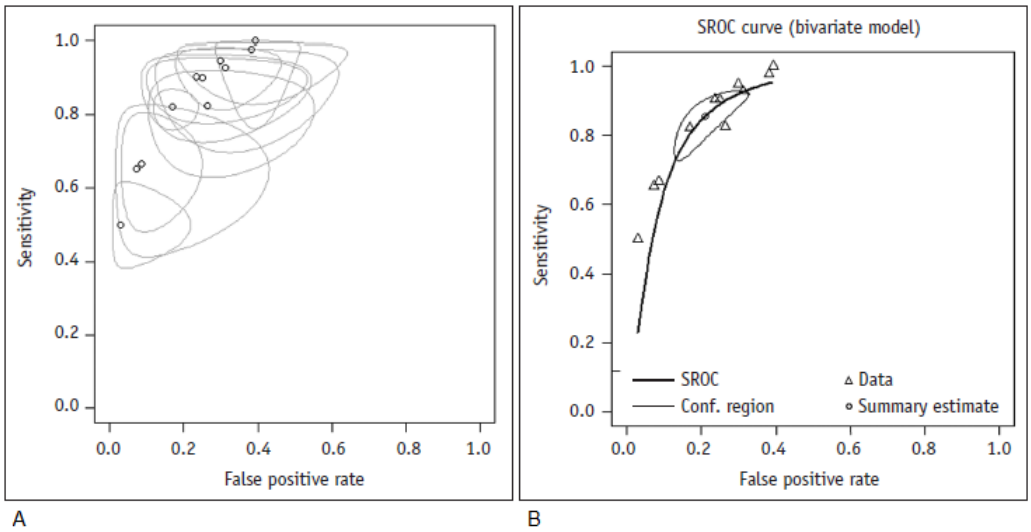


Figure 9 – **A.** Example of SROC plot with the results of individual studies and the 95% confidence regions. **B.** SROC curve with a summary point and its confidence region (adapted from (7)).

2.4.2.3.2 Forest plots

A forest plot is a chart in which the x-axis can be any calculated descriptive statistic with its 95% confidence interval (CI) and the y-axis is the study identifier. The plot is often rendered so that the size of the data points reflects the sample size of each study (7). Commonly, each study is accompanied by its number of TP, FP, FN, and TN, and the value of sensitivity and specificity, together with confidence intervals. Also, summary statistics computed from meta-analyses can be added to forest plots. Whilst it is possible to observe heterogeneity in sensitivity and specificity individually on such plots, it is not as easy to visualise whether there are threshold-like relationships (5).

2.4.2.4 Model fitting

In this section we introduce the statistical methods proposed to synthesize evidence in diagnostic meta-analysis. Section 2.4.2.4.1 briefly presents some less common approaches while the following sections focus on the most common and important ones.

2.4.2.4.1 Less common approaches

Simple pooling

This approach derives a single-summary two-by-two table by adding the numbers of true positives, false positives, true negatives, and false negatives across all studies. Test sensitivity and specificity can then be estimated as though all the data came from a

single study. This is a form of fixed-effect meta-analysis of sensitivity and specificity, ignoring any correlation between them and assuming no between-study heterogeneity (104).

Separate random-effects meta-analysis of sensitivity and specificity

In this approach, the same statistical techniques that would be applied to a random-effects meta-analysis of intervention studies (see section 2.4.2.1) are applied separately for the logit transforms of sensitivity and specificity (109).

The logit transform $logit(p) = ln(\frac{p}{1-p})$ is used to transform probabilities or rates in the unit interval [0, 1] to values on the complete real line. Also, the assumption of a normal distribution between studies is more reasonable on the logit scale (104).

This approach allows for between-study heterogeneity in sensitivity and specificity but ignores their correlation. In addition to summary points and confidence intervals for these points, a summary ROC curve can be obtained from this method using the ratio of the estimated between-study variances (104).

It is also possible to conduct a separate meta-analysis of positive and negative likelihood ratios. However, it ignores the correlation between positive and negative LR as well (5) and is not recommended (as seen in section 2.4.2.2.3).

2.4.2.4.2 Moses-Littenberg SROC curves

One of the earliest methods developed to summarize diagnostic studies is the summary receiver operating characteristic curve (SROC curve). The Moses-Littenberg method (110, 111) provides a simple model for deriving a SROC curve and provides a nice visualization of the relationship of sensitivity and specificity across studies. It is more akin to a fixed effect than a random-effects model as it do not allow for systematic variation between studies apart from the applied threshold. While there are more advanced methods nowadays (explained in section 2.4.2.4.3), this approach is vital to the understanding of most other methods (5, 104).

The basic SROC curve can be obtained as follows. Again, the pairs of sensitivity and specificity estimates from each study are transformed onto logit scale to compute,

$$D_i = logit(sens_i) - logit(1 - spec_i)$$
$$S_i = logit(sens_i) + logit(1 - spec_i)$$

The quantity D_i is the natural logarithm of the DOR,

$$logit(sens_i) - logit(1 - spec_i) = ln\left(\frac{sens_i}{1 - sens_i}\right) - ln\left(\frac{1 - spec_i}{spec_i}\right)$$

$$= ln\left(\frac{sens_i \times spec_i}{(1 - sens_i)(1 - spec_i)}\right) = ln(DOR),$$

while S_i is a quantity related to the overall proportion of positive test results and can be considered as a proxy for test threshold. The relationship between D_i and S_i is expected to be linear.

The simple linear regression model $D_i = \alpha + \beta S_i + e_i$ characterizes how test accuracy, measured by the diagnostic log odds ratio (D_i), varies with S_i. Here, α and β are the model intercept and slop, respectively, and e_i is the error term, which is assumed to be normally distributed.

With the estimates of α and β, the values can be back-transformed to the original scales to obtain the SROC curve,

$$sens(FPR) = \left(1 + exp\left(\frac{-\hat{\alpha}}{1 - \hat{\beta}}\right)\left(\frac{1 - FPR}{FPR}\right)^{\frac{1+\hat{\beta}}{1-\hat{\beta}}}\right)^{-1},$$

where $\hat{\alpha}$ and $\hat{\beta}$ denote the estimates of α and β computed from the data.

The interpretation of the intercept and the slope of the linear regression model is not straightforward. When the DOR does not depend on the threshold S (e.g., $\beta \approx 0$), the intercept would provide a summary estimate of the DOR. When the DOR does vary with S, the coefficient of the slope (β) has no direct interpretation, but has a considerable effect on the shape of the SROC curve (112).

The disadvantage of the DOR as the outcome parameter is that summary estimates of sensitivity and specificity are not directly available. It is only possible to obtain an estimate of one by specifying a value for the other. The Q-point could be used to summarise values of sensitivity and specificity, but it may lead to summary values that are not close, or even outside the range of values from the original studies (112).

Also, Q-points could be used to test for a difference in overall accuracy between diagnostic tests, since comparing the DORs at a specific value of S (zero in this case) would remove the effect of a possible difference in threshold. However, testing at a different value of S could lead to different conclusions if the DOR of one or both tests varies with S (112).

2.4.2.4.3 Hierarquical models

More statistically rigorous approaches based on hierarchical models have been proposed that overcome the limitations of the Moses-Littenberg method. In this section, the hierarchical SROC model of Rutter and Gatsonis (113) and the Bivariate model (112) are presented.

Both hierarchical models involve statistical distributions at two levels. At the first level, a within study variability for both sensitivity and 1-specificity is assumed to follow a binomial distribution. Thus, the numbers of test positives (y_{ij}) from each study (i) in each disease group ($j = 1,\ 2$ considering binary test results) are assumed to follow binomial distributions $y_{ij} \sim B(n_{ij}, \pi_{ij})$, where n_{ij} and π_{ij} respectively represent the total number of tested subjects and the probability of a positive test result. The first level is the same in both models. However, they differ at the higher level when modeling a between-study difference in diagnostic test accuracy beyond that accounted for by sampling variability at the lower level (5, 109).

The Bivariate model and Rutter and Gatsonis HSROC model are mathematically equivalent when no covariates are fitted, but differ in their parametrizations. The former models sensitivity, specificity and the correlation between them directly, whereas the latter models functions of sensitivity and specificity to define a summary ROC curve (5).

Both models could be used to estimate SROC curves, the summary values of sensitivity and specificity, 95% confidence regions of the summary values, and its 95% prediction regions of the SROC curve (within which we may expect the true sensitivity and specificity of a future study to lie (104)). However, the bivariate model is preferred for the estimation of a summary value of sensitivity and specificity, as well as for evaluating how their expected values may vary with study level covariates; whereas, the HSROC model is favored for the estimation of the SROC curve for assessing test accuracy and determining how the curve's position and shape may vary with study level covariates (109). Since the bivariate model is easier to fit and perhaps also easier to understand, it has become the standard approach for meta-analysis of diagnostic studies (104).

Rutter and Gatsonis HSROC model

To overcome the main problem of the basic SROC approach and incorporate systematic variation between studies, Rutter and Gatsonis (113) introduced their hierarchical model, also known as the HSROC (hierarchical SROC) model (104). It assumes that there is an underlying ROC curve in each study with parameters α and β that characterize the accuracy and asymmetry of the curve, in a similar (though technically distinct) way to the α and β in the linear regression method of Moses and Littenberg (5).

At the higher level, the probabilities π_{ij} are to be predicted in a regression model. The logit of π_{ij} should be regressed as:

$$logit(\pi_{ij}) = (\theta_i + \alpha_i X_{ij}) exp(-\beta X_{ij})$$

The dummy variable X_{ij} represents the "true" disease status (coded as -0.5 for

the non-diseased and 0.5 for the diseased). The parameters of the model, which are to be estimated, are θ_i, α_i, and β. The model intercepts θ_i are called cutpoint parameters since they model the trade-off between sensitivity and false-positive rate. The slopes α_i are called accuracy parameters since they model the difference between sensitivity and FPR. The scale parameter β provides for asymmetry in the SROC by allowing accuracy to vary with threshold. Since each study contributes only with one estimate of sensitivity and specificity at a single threshold, it is necessary to assume that the shape of the true underlying ROC curve in each study is the same, hence β cannot be assumed to vary across studies (5, 104, 113).

The between-study variation is, in fact, allowed in the HSROC model by assuming that parameters θ_i and α_i are independently and normally distributed with a mean threshold of Θ and a mean accuracy of Λ with variances σ_θ^2 and σ_α^2, respectively (5, 109):

$$\theta_i \sim N(\Theta, \sigma_\theta^2)$$
$$\alpha_i \sim N(\Lambda, \sigma_\alpha^2)$$

Thus, when no covariates are included, the HSROC model has five parameters: Θ, Λ, β, σ_θ^2, and σ_α^2. The authors of the HSROC model proposed to fit it using fully Bayesian techniques, but it may also be fitted using classical statistical methods. Having estimated the model parameters, the HSROC curve can be computed as (104, 5)

$$sens(FPR) = logit^{-1}((logit(FPR)e^{\frac{\hat{\beta}}{2}} + \hat{\alpha})e^{\frac{\hat{\beta}}{2}})$$

where $logit^{-1}(x) = \frac{e^x}{1+e^x}$ is the inverse of the logit-transform.

Covariates (Z_i) can be taken into account in the HSROC model to explore heterogeneity in test positivity (threshold), position of the curve (accuracy) and shape of the curve (5):

$$logit(\pi_{ij}) = ((\theta_i + \gamma Z_i) + (\alpha_i + \lambda Z_i)X_{ij})exp(-(\beta + \delta Z_i)X_{ij})$$

Hence, the distribution of the random effects for threshold and accuracy would be given by $\theta_i \sim N(\Theta + \gamma Z_i, \sigma_\theta^2)$ and $\alpha_i \sim N(\Lambda + \lambda Z_i, \sigma_\alpha^2)$, respectively.

Bivariate model

The bivariate model preserves the two-dimensional nature of the data throughout the analysis by modeling sensitivity and specificity directly (112). It can be considered an

extension of the separate random-effects approach (section 2.4.2.4.1) but allows for the correlation between sensitivity and specificity (104).

At the higher level, a between-study difference is modeled and the logit-transformed sensitivities and specificities are assumed to have a normal distribution with means μ_A and μ_B, variances σ_A^2 and σ_B^2, respectively, and the covariance σ_{AB} between logit sensitivity and specificity. The combination of two normally distributed outcomes, the logit transformed sensitivities and specificities, while acknowledging the possible correlation between them, leads to the bivariate normal distribution (5, 109, 112):

$$\begin{pmatrix}\mu_{A,i}\\ \mu_{B,i}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_A\\ \mu_B\end{pmatrix}, \Sigma\right) \quad \text{with} \quad \Sigma = \begin{pmatrix}\sigma_A^2 & \rho_{AB}\\ \rho_{AB} & \sigma_B^2\end{pmatrix}$$

where $\rho_{AB} = \frac{\sigma_{AB}}{\sigma_A \sigma_B}$ is the correlation between logit sensitivity and specificity. Thus, the bivariate model also has five parameters: μ_A, μ_B, σ_A^2, σ_B^2, and ρ_{AB}.

Like the HSROC model, the bivariate model can take into account the effect of covariates that affect sensitivity and specificity by replacing the means of μ_A and μ_B with linear predictors in the covariates (109):

$$\begin{pmatrix}\mu_{A,i}\\ \mu_{B,i}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_A + v_A Z_i\\ \mu_B + v_B Z_i\end{pmatrix}, \Sigma\right)$$

2.4.3 Other considerations

2.4.3.1 When does it make sense to perform a meta-analysis?

The questions of whether it makes sense to perform a meta-analysis and what kinds of studies to include must be asked and answered in the context of specific goals and the review question (103).

From a statistical perspective, there is no restriction on the similarity of studies based on the types of participants, interventions, or exposures. However, for the analysis to be meaningful, it is need to pay careful consideration to the diversity of studies in these respects, addressing those technical differences in the analysis (103). Also, the estimates of test accuracy in the individual studies should be relevant and unlikely to be biased (5).

It is an important feature of a meta-analysis that it may (and usually must) address a broader question than those addressed by the primary studies it includes. Thus a certain amount of diversity among the studies is not only inevitable but also desirable. A good meta-analysis will anticipate this diversity and will interpret the findings with attention to the dispersion of results across studies. It may lead to heterogeneous results and this heterogeneity needs to be recognized in the analysis and interpretation (103).

In addition, the studies that are being combined in the analysis should be methodologically rigorous. Meta-analysis of studies at risk of bias may be misleading. If bias is present in individual studies meta-analysis may compound the errors and produce an erroneous result which may be inappropriately interpreted as having credibility (5).

2.4.3.2 Aims of DTA meta-analyses

In general, there are three main types of question that can be addressed in a analysis concerning the accuracy of a test: to conclude what is the accuracy of a test; to analyse how does the accuracy vary with clinical and methodological characteristics; to assess how the accuracy of two or more tests compare (5).

The first case is restricted to characterizing the accuracy of a single test, either estimating an average summary value of sensitivity and specificity or describing how they vary with the threshold by estimating an SROC curve. The second will focus on investigating heterogeneity by analysing whether the observed test accuracy varies between studies according to characteristics associated with the test, settings, participants or methodology of the studies. The last would identify which test (or tests) yields superior test accuracy through a form of subgroup analysis (5).

The goal of some syntheses will be to report the summary effect, but the goal of other syntheses will be to assess the dispersion as well as the mean effect, and the goal of others will be to focus on the dispersion exclusively (103). Also, conducting a meta-analysis can be useful to identify areas for further research as new hypotheses may be generated or it may highlight deficits that need to be addressed in future primary studies before a useful meta-analysis can be done (114).

2.4.3.3 SROC curve versus summary point

It is necessary to make a choice of which summary statistics are to be computed. The inclusion criteria can be narrowly defined and focus on the summary effects by estimating expected values of sensitivity and specificity for the test at a common threshold, or more broadly defined and explore the dispersion and the difference in the results between the studies by estimating the expected ROC curve for a test across many thresholds and investigating heterogeneity (103, 5).

Whilst for some tests there is consensus of what value the positivity threshold should take, more often tests are evaluated at different thresholds in different studies. Estimating a summary point by pooling studies which mix thresholds would produce an estimate that relates to some notional unspecified average of the thresholds, which is clinically unhelpful (5).

The choice of analytical approach is influenced by the variation of thresholds in

the available studies. When there is little consistency in the thresholds used, estimating a SROC curve may be preferred. If there is little variation in threshold between studies attempting to fit a SROC curve will be difficult as the points are likely to be too tightly clustered in ROC space (5).

Finally, it is also reasonable to estimate both SROC curves and summary points, as they may complement each other in providing clinically useful summaries and powerful ways of detecting effects (5).

2.4.3.4 Risk of bias and quality assessment

DTA studies are often subject to bias - a systematic deviation of the study results from the true diagnostic accuracy that typically occurs due to flaws in study design or inappropriate execution of the study. A meta-analysis of study results that contain numerous variations or biases would be of little value. Therefore, it is important to detect possible variations and biases in the research studies included in a meta-analysis and to assess the methodological quality of the studies (7).

For a systematic review of DTA studies, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) (115) is currently recommended. If a study is found to be of poor quality upon QUADAS-2 evaluation it can be excluded from the meta-analysis or its effect on the outcome can be analyzed (104, 7).

This tool assesses study quality in four key domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of the risk of bias, while the first three domains are also assessed in terms of concerns about applicability (the concern that a study does not match the review question) (115).

The patient selection domain assesses distortions in the process of selecting the sample of the study. It is important that the sample recruited is representative of that observed at the point in the care pathway where the test is planned to be used to ensure that it includes a similar spectrum of those both with and without the target condition (116). Inappropriate exclusions of patients may result in distortions of diagnostic accuracy. For example, excluding "difficult to diagnose" patients may result in overoptimistic estimates, while excluding patients with "red flags" for the target condition, who may be easier to diagnose, may lead to underestimation of diagnostic accuracy (115).

Distorted selection may also occur when the inclusion was not done consecutively or randomly, but for example based on the clinician's preferences - which may coincide with the ability of a test to make the right diagnosis (116). Furthermore, studies enrolling patients with known disease and a control group without the condition may exaggerate diagnostic accuracy (115).

The index test domain evaluates if the execution of the index test may lead to

biased results. If its interpretation was not blinded against the results of the reference standard, then both sensitivity and specificity may be overestimated. Also, the assessors of the index test should be blinded for clinical information usually not provided in practice to avoid distorting the accuracy (116, 115).

Another way through which the execution of the index test might be distorted is by selecting the test threshold to optimize sensitivity and/or specificity, which may lead to overoptimistic estimates of test performance (115).

The reference standard domain evaluates if the execution of the reference standard may lead to biased results. In this case, blinding is also an issue. If the results of the index test are already known it might influence the judgement of the reference standard. Moreover, the reference standard might not be accurate enough (116).

The flow and timing domain assess if the overarching process may be distorted. Ideally, results of the index test and reference standard are collected on the same patients at the same time. If there is a delay or if treatment is started between index test and reference standard, misclassification may occur due to recovery or deterioration of the condition. The length of interval leading to a high risk of bias will vary between conditions (115).

Verification bias occurs when not all of the study group receive confirmation of the diagnosis by the same reference standard. Also, all patients who were recruited into the study should be included in the analysis (115).

Signaling questions are used to categorize the risk of bias as low, high, or unclear. If a study is judged "high" or "unclear" in 1 or more domains, then it may be judged "at risk of bias" or as having "concerns regarding applicability" (115).

The QUADAS-2 must be tailored to each review by adding or omitting signaling questions and developing review-specific guidance on how to assess each signaling question and use this information to judge the risk of bias (115).

After applying the tool, reviews should present a summary of the results of the assessment for all included studies. This could include summarising the number of studies that found low, high or unclear risk of bias/concerns regarding applicability for each domain. If studies are found to consistently rate well or poorly on particular signalling questions then reviewers may choose to highlight those (115).

2.4.3.5 Heterogeneity

In DTA reviews large differences are commonly noted between studies, too big to be explained by chance, indicating that actual test accuracy varies between the included studies, or that there is heterogeneity in test accuracy (5).

The general recommendation is to group the studies in categories for graphical illustration and to investigate the relationship between diagnostic accuracy and covariates by meta-regression models. The magnitude of observed heterogeneity is best depicted graphically where such relationships can be observed by the scatter of points and from the prediction ellipse. Statistically, it is generally more efficient to make use of all of the data available across studies when investigating heterogeneity by adding study level covariates to a hierarchical model to identify factors associated with diagnostic test accuracy. (5).

2.4.3.6 Sensitivity analysis

The process of undertaking a systematic review involves a sequence of decisions, and some of them will be somewhat arbitrary or unclear. To demonstrate that the findings obtained are not dependent on such decisions, a sensitivity analysis should be conducted. It is basically a repeat of the primary analysis or meta-analysis, substituting decisions that were arbitrary or unclear for alternative decisions. For example, if the eligibility of some studies in the meta-analysis is dubious because they do not contain full details, sensitivity analysis may involve undertaking the meta-analysis twice: first including all studies, and second, only including those that are definitely known to be eligible (5).

A sensitivity analysis asks the question "Are the findings robust to the decisions made in the process of obtaining and analysing them?", and differs from a subgroup analysis - where the purpose is to investigate how study design and patient characteristics are associated with test accuracy, thus exploring and explaining heterogeneity in test accuracy (5).

Where sensitivity analysis show the overall result and conclusions are not affected by the different decisions made during the review process, the results of the review can be regarded with a higher degree of certainty. Where sensitivity analyses identify particular decisions or missing information that greatly influence the findings of the review, greater resources can be deployed to try and resolve uncertainties and obtain extra information. If this cannot be achieved, the results must be interpreted with an appropriate degree of caution. Such findings may generate proposals for further investigations and future research (5).

3 Methods

Based on the theoretical foundation presented before, in this chapter we present the methodology used to conduct a systematic review and meta-analysis of studies that used ML classifiers based on rs-fMRI data to distinguish patients with ASD from individuals of TD.

The steps adopted followed the structure: 1) definition of objectives and formulation of research questions; 2) search strategy; 3) study selection; 4) data extraction; 5) snowballing; 6) quality assessment; 7) statistical analysis. These steps are explained in the following sections and Figure 10 summarizes our methodology.

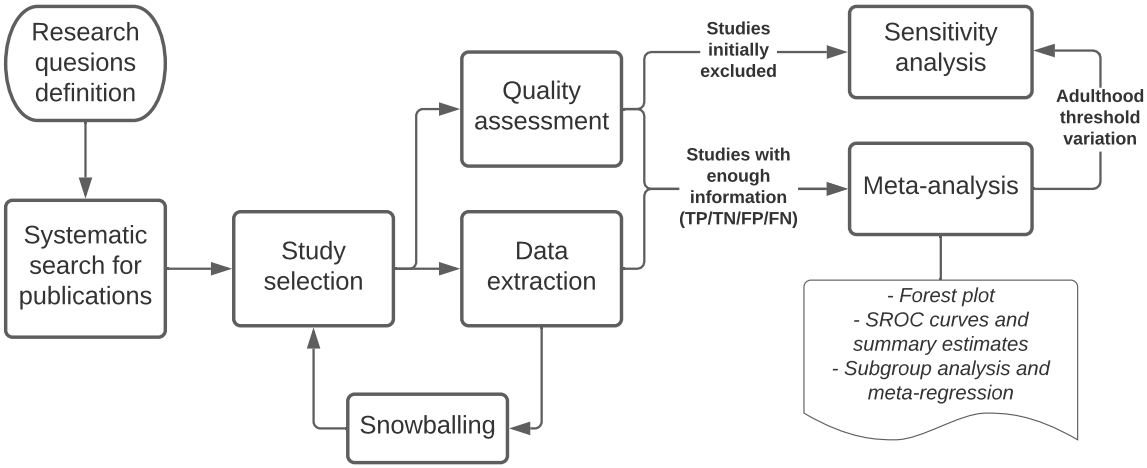


Figure 10 – Fluxogram of the methodology used in the systematic review and meta-analysis.

3.1 Objectives and Research Questions

Objective

To analyze scientific publications through a systematic literature review and meta-analysis.

In order to identify studies on ASD diagnosis.

Regarding the use of rs-fMRI brain images and their classification through ML techniques.

From the point of view of the researchers.

In the academic **context**.

Starting from this proposed objective, the research questions were defined:

- Which ML techniques are used to classify ASD and TD individuals based on rs-fMRI?
- What are the results obtained by the studies using these approaches?
- Which methodological differences are associated to the performance measures obtained throughout the publications?
- The approaches are robust enough to be applied in a clinical setting?
- What are the aspects that still need to be investigated?

3.2 Search strategy

The articles used in this review were found through four digital libraries: Scopus[1], El Compendex[2], PubMed – NCBI[3], and IEEE Xplore[4]. We used Scopus and IEEE libraries since they are large and well-known digital libraries. Scopus is the largest abstract and citation database of peer-reviewed literature whereas IEEE is the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity. In order to expand the search, we also used El Compendex and PubMed NCBI libraries. The first one is focused on engineering whereas the second focus on biomedical literature. Those two last libraries resulted in a large number of duplicated articles (approximately 73% and 91% of the articles, respectively) so we decided to not include other libraries in the search and find out other studies through the snowballing (more details at section 3.5).

The search expression was iteratively defined using keywords considered appropriate. We analyzed the titles and abstracts of the publications found through the searches to define whether they were related or not to the purpose of this study. Based on that, we refined the search expression and obtained the final version presented below:

[1] https://www.scopus.com
[2] https://www.engineeringvillage.com
[3] https://www.ncbi.nlm.nih.gov/pubmed/
[4] https://ieeexplore.ieee.org

```
("Artificial Intelligence" OR "Machine Learning" OR "Artificial Neural Network" OR
↪  "Neural Network" OR "Neural Net" OR "Artificial Neural Net" OR "SVM") AND
↪  ("rs-fMRI" OR "rsfMRI" OR "R-fMRI" OR "fcMRI" OR "Resting-State Functional
↪  Magnetic Resonance Imaging" OR "Resting State Functional Magnetic Resonance
↪  Imaging" OR "Rest State Functional Magnetic Resonance Imaging" OR "functional
↪  connectivity Magnetic Resonance Imaging" OR "Resting-State fMRI" OR "Resting
↪  State fMRI" OR "Rest State fMRI" OR "Resting-State Functional MRI" OR "Resting
↪  State Functional MRI" OR "Rest State Functional MRI" OR "functional connectivity
↪  MRI") AND ("Autism spectrum disorders" OR ASD OR "Autism")
```

It is worth noting that we started using other expressions related to ASD as defined by the DSM-IV (39) in order to possibly include articles published before 2013, when the DSM-V (12) was first published. These expressions were: "Pervasive Development Disorders"; "PDD"; "Autistic Disorder"; "Asperger's Disorder"; "Asperger"; "Childhood Disintegrative Disorder"; "PDD-NOS". However, the addition of these terms only resulted in two new articles that were not related to the purpose of this study. Therefore, we decided to simplify the expression by removing those terms.

The search was carried out in two parts. First, we searched for articles published between January 1, 2010, and December 7, 2018, the date of the final search conducted. The string was applied directly in the digital libraries El Compendex and PubMed. For Scopus, the advanced mode was used and the search was specified for title, abstract and keywords. Likewise, the advanced mode was used for IEEE Xplore, but the search was specified for full text. The start date was defined taking into account that during the tests with the string only one article published before 2010 was found and it did not fulfill the criteria to be included in this study. Furthermore, the use of the snowballing technique (see section 3.5) should retrieve the most relevant papers published before this date.

After the first search, the development of the study experienced some delays. Therefore, a second search was performed to keep the study updated. We searched for articles published between December 7, 2018, and April 3, 2020, the date of the final search conducted. The string was applied to the digital libraries as explained before. The only exception was the IEEE Xplore for which we used the command search instead of the advanced mode and the search was specified for full text and metadata.

3.3 Study selection

First, a triage process was applied to the non-duplicate publications. Three researchers submitted each paper to a selection based on specific inclusion and exclusion criteria previously defined. However, some exclusion criteria needed to be created or adjusted during the selection for better classification. The exclusion criteria are shown in

Table 2, where new and adjusted criteria are explicit, whereas the inclusion criteria are shown in Table 3.

Table 2 – Exclusion criteria

Criteria	**Description**
EC1	Publications in which the search keywords do not appear in the title, abstract and/or text of the publication (excludes the keywords field, the sections thanks, authors biography, bibliographical references and attachments).
EC2	Reviews whose main focus was not the classification between ASD and TD using rs-fMRI and ML.
EC3	Publications that deal with the genetic basis of ASD.
EC4	Publications that seek to define or analyze treatments for ASD.
EC5	Publications that seek to predict the evolution of ASD.
EC6	Publications on ASD that exclusively use other types of brain imaging rather than rs-fMRI.
EC7	Publications that address the classification of ASD and use fMRI, but in specific experiments and not in resting state.
EC8	Publications that mention ASD but it is not the main focus.
EC9	Publications whose objective is to find the relationship between ASD symptoms and the functional connectivity of the brain.
EC10	Publications that seek to understand and characterize the ASD brain network, but do not perform the classification of subjects.
EC11	Publications that are not scientific articles.
EC12	Publications that seek to classify subjects in relation to ASD severity rather than ASD versus TD.
EC13	Publications that address the ASD classification and use rs-fMRI, but do not use ML for classification.
EC14*	Publications that use rs-fMRI, ML, and data from ASD subjects for classification, but do not classify explicitly between ASD and TD.
EC15**	Publications that seek to determine the best modeling choices for ASD classification using rs-fMRI and ML but do not present the classification results directly.
EC16**	Publications whose objective is to find the relationship between functional connectivity and specific activities but do not perform the classification of subjects.
EC17**	Publications related to video-based ASD detection.
EC18**	Articles not published in English.

* Criteria created after starting the process, during the first search or ** during the second search.

The criteria were applied based on the abstracts of the studies. When it was not sufficient, a superficial reading of the entire article was carried out - it is worth noting that this superficial reading was conducted only for a pre-selection of the articles. The papers were selected if at least one of the researchers concluded it should be. Then, the same three

Table 3 – Inclusion criteria

Criteria	**Description**
IC1	Publications that use ML techniques to classify subjects between ASD and TD, based only on rs-fMRI.
IC2	Publications that present guidelines for the application of ML techniques in the classification of brain images, as long as they treat rs-fMRI and ASD and present classification results.
IC3	Publications that use ML techniques to classify individuals between ASD and TD based on rs-fMRI together with other data types.
IC4	Publications that use ML techniques and rs-fMRI to distinguish ASD from other disorders, as long as they also perform the classification between ASD and TD.

researchers performed a new assessment to confirm the selection. In this step, each paper selected was read carefully to determine if it fulfilled three requirements: 1) used rs-fMRI data; 2) performed a classification between ASD and TD; and 3) the classification was performed using a ML technique. If at least one of those requirements was not fulfilled, the article was excluded from the study.

3.4 Data Extraction

A standardized data extraction sheet was used by the researchers to collect data from all included studies. We extracted source and type of the data, sample size, if the study included both males and females, average age and Full IQ (FIQ) of the subjects, preprocessing steps, feature extraction and selection procedures, the validation process, classifiers used, outcomes reported, main results (accuracy, sensitivity, specificity, and measures of TP, TN, FP, and FN), other tests performed, and important brain areas.

We extracted/calculated only one result from each independent sample in a study. Since the majority of the publications presented multiple results from different tests, the main results were selected according to the following criteria: results from the classification method proposed in the article were prioritized; results presenting enough information to conduct the meta-analysis (measures of TP, TN, FP, and FN, number of ASD and TD subjects in the test set) were prioritized; results using only rs-fMRI data were prioritized; results using a hold-out test set, a inter-site (leave-one-site-out) approach or a train/validation/test procedure were prioritized; tests using larger samples were prioritized; if the study presented results using different number of folds for the cross-validation, 10-fold was prioritized (the most common approach); finally, the results with higher accuracy were prioritized.

3.5 Snowballing

Snowballing means to systematically search for primary studies based on references to and from other studies. Since we limited our research to the date of the final search conducted, we only performed a backward snowballing (98). The goal was to broaden the scope of this work and include the maximum number of related articles, especially those before 2010, if any.

As the selected articles were analyzed, we looked for references that could be included in this systematic review according to the inclusion/exclusion criteria. It resulted in a large number of duplicated articles, so we decided not to re-apply the snowballing technique. Also, the new articles found went through the same selection process presented before.

3.6 Quality assessment

Methodological quality was assessed using the QUADAS-2 (115) - the currently recommended tool for a systematic review of DTA studies (104, 7).

The first step to apply the QUADAS-2 is to describe the review question in terms of patients, index test(s), and reference standard and target condition. It will be used as a basis for the analysis, indicating whether any signalling question does not apply to the review or whether any specific issues for the review are not adequately covered by the core signalling questions (115).

In the case of our systematic review, the target population was defined as any TD/ASD subject that underwent an ML classification tool. Thus, there were not many specific criteria regarding the population selected provided that the objective of the test was to classify between ASD subjects and TD subjects.

Regarding the index tests, we were looking for any ML classification tool designed to classify the subjects between ASD and TD using rs-fMRI data. Therefore, our target condition was the Autism Spectrum Disorder.

Finally, the reference standards were the DSM-V, DSM-IV-TR or DSM-IV criteria for ASD. ADOS-2 and ADI-R (or other versions of those tools) were also considered as reference standards.

Next, it is essential to tailor QUADAS-2 to each review by adding or omitting signalling questions and developing review-specific guidance on how to asses each signalling question, using this information to judge the risk of bias and applicability concerns. Once tool content has been agreed, review-specific rating guidance should be developed. The tool should be piloted independently by at least two people. If agreement is good, the tool

can be used to rate all included studies. If agreement is poor, further refinement may be needed (115).

The tool was tailored by two researchers. After defining the signaling questions and review-specific guidance, both authors applied the tool using five articles. The answers to the signalling questions and the risks of bias/applicability were compared, and any disagreement was discussed to reach a consensus.

For the first domain - patient selection - we decided to maintain the three core signalling questions to judge the risk of bias without any modifications. In the first question - was a consecutive or random sample of patients enrolled? - we defined that cases where the individuals came from the ABIDE - even if the specific sites used were presented - or another similar database and no further information was provided the answer should be unclear. As stated in (117), the retrospective collection of large radiologic imaging data by means of clinical referrals is prone to spectrum bias because there are multiple layers of possible patient exclusion between the clinical cohort and the data ultimately available. Thus, without any further information regarding the process by which the subjects were recruited and had their data available on the database, we cannot affirm if a consecutive or random sample was enrolled. For the other cases, the information presented was assessed to define the answer.

The second question - was a case control design avoided? - was rated as no if the study in analysis used data from the ABIDE or another similar database, since those databases provide individuals already defined as with or without the disorder. The other cases had their information assessed to define the answer.

The third question - did the study avoid inappropriate exclusions? - was rated as no if the study selected only individuals in any restricted interval (e.g., only in a small range of age, only high-functioning individuals, only males). Otherwise, the information presented was assessed to define the answer. Several studies (43, 44, 45, 50) indicate gender, age, and IQ differences on autistic symptoms and impairments. Therefore, studies that selected individuals on those restricted intervals may present distortions in the diagnostic accuracy obtained.

The risk of bias related to the patient selection was defined as follows: if the answer to the third question was no, there was a high risk; if only the second question was no there was a low risk or unclear risk - since most of the included studies used data from a database such as the ABIDE; there was a low risk only if at least two answers were yes.

The concerns regarding the applicability of this domain were defined based on the patient characteristics recorded. As our question is very broad regarding the patient sampling, there was a low concern in the majority of the cases.

For the second domain - index test - we did not use any signalling question to

judge the risk of bias. Since the index tests considered in our review are ML algorithms, the question related to blinding the interpreter to the results of the reference standard does not apply. Moreover, the question about the threshold was also omitted since the bias analysis will generally focus on the validation process - the process by which the ML algorithm's accuracy is reported.

If the study used a cross-validation scheme to evaluate the proposed model but no further information was provided, there was a unclear risk of bias. If it explicitly says that a nested cross-validation was used, there was a low risk. Also, if the study used a hold-out set for testing the algorithm, there was a low risk. If it presented the results per number of features, per atlas used or similar but the best model was not applied to an independent set, there was a high risk of bias - since there is a concern that the classifier was adapted to peculiarities of the data-set and the accuracy may be overestimated. Otherwise, the validation process was assessed to determine the risk of bias.

The concerns regarding the applicability of this domain were assessed through a new signaling question - was the test conducted using only rs-fMRI data or information that would be known in a real application? - addressing the situations were the ML algorithms used other types of data beyond rs-fMRI. Since our review question specify the use of rs-fMRI by the ML algorithms, the answer was no if the study used not only rs-fMRI but also other kinds of image data (e.g., sMRI). Regarding the phenotypic information, if only age and/or gender was used the answer was yes, but if IQ, and/or cognitive and behavioral assessments were used the answer was no - once those information may not be available in a real application. Also, acquisition site was considered an information that would be known in a real application.

Based on this signalling question, there was a high concern if the answer to the question was no, a unclear concern if the answer was unclear, and a low concern if the answer was yes.

For the third domain - reference standard - we decided to maintain the two core signalling questions to judge the risk of bias without any modifications. In the first question - is the reference standard likely to correctly classify the target condition? - if the study used one or more of the reference standards defined in the review question the answer was yes. In other cases, the test used was assessed to define its reliability. If the study used a database such as the ABIDE but no further information regarding the reference standard was provided, the answer was yes since we considered the diagnosis obtained by these databases as reliable.

The second question - were the reference standard results interpreted without knowledge of the results of the index tests? - was rated as yes if the sample of the study came from databases such as the ABIDE and the reference standard results were previously interpreted and reported. If the study used an own sample, the information provided

was assessed to determine the answer.

Therefore, if any of the previous questions was answered as no, there was a high risk of bias. Otherwise, if one of the answers was unclear, there was a unclear risk.

The concerns that the target condition as defined by the reference standard may not match the review question were defined based on the information recorded. There was a high risk of bias only if the diagnostic was assessed using other reference standards rather then the ones defined in the review question.

For the fourth domain - flow and timing - we decided to maintain the three core signalling questions to judge the risk of bias without any modifications. In the first question - was there an appropriate interval between index test and reference standard? - we defined that if the study used a database such as the ABIDE the answer was unclear, since it is not clear when the diagnostic and the images were obtained. In other cases, the information provided was assessed.

The second - did all patients received the same reference standard? - and the third - were all patients included in the analysis? - questions were defined directly from the data recorded from the study.

Finally, the risk of bias related to the flow and timing was defined as follows: if the second question was no but all patients received reference standards as defined in the review question it does not indicate necessarily a high risk of bias; For the first or third question, a no answer does indicate a high risk; Otherwise, if at least two answers were yes, the answer was low risk.

The QUADAS-2 tool was applied two times. In the first, all articles were assessed as a whole. In the second, only the articles selected for the meta-analysis were assessed, considering the main results (as defined in section 3.4) used for the statistical analysis. In both cases, the information used to reach the judgement of each of the domains were recorded to make the rating transparent and facilitate discussion.

3.7 Statistical analysis

Studies were eligible for inclusion in the quantitative meta-analysis if measures of TP, TN, FP, and FN were available or if the data allowed for the calculation of these measures. The TP/TN/FP/FN values were extracted or calculated from each independent sample in a study according to the criteria defined in section 3.4.

In order to avoid bias, it is necessary to handle sample overlap between the studies, possibly excluding samples with a large overlap. However, the majority of the studies selected in this review extracted their samples from the ABIDE database. Thereby, we have a lot of potential overlapping samples and, at the same time, there is little information

with respect to the exact individuals used in each study to conclude the real extent of the overlap. The exclusion of all the potential overlapping samples would make it difficult to perform a meta-analysis since only a few results would remain. Furthermore, we can take into consideration that the studies vary considerably regarding characteristics such as the preprocessing, features, and classification techniques used. Thus, we decided to use all the results regardless of the sample overlap.

The statistical analysis was performed using the open source package mada (118) in R Statistics (119). A coupled forest plot of sensitivity and specificity was created using RevMan version 5.3 (120). Microsoft Excel and Lucidchart [5] were also used to create the images presented in the following chapters.

Summary receiver operating characteristic (SROC) curves, summary estimates of sensitivity and specificity and the corresponding 95% confidence intervals (CIs) were calculated by the bivariate model of Reitsma *et al.* (112). Prediction region, area under the curve (AUC) and partial AUC (pAUC) were also obtained. Studies that were visually deviant from the 95% prediction region on the SROC curves were considered heterogeneous (5).

Subgroup analysis and bivariate meta-regression with potential covariables were performed to reduce any heterogeneity noted between the studies. The ML technique used, year of publication, sample size, type of data, source of the sample, atlas used, number of ROIs, QUADAS-2 results, type of features, and sex, IQ and age of the subjects were investigated. Following the approach from (121) and knowing that the bivariate model has five parameters (5), we considered $n = 5$ to be the minimum number of studies to justify a separate meta-analysis. All tests were based on a 2-sided significance level of $p = 0.05$.

In sensitivity analysis, three studies that were initially excluded from the meta-analysis (see section 4.1) were included to verify the robustness of the results. Also, we investigated the effect of the age of the subjects considering different adulthood thresholds (18-21 years old).

Publication bias was not assessed in our analysis, as there are currently no statistically adequate models in the field of meta-analysis of DTA studies and further research is required (5).

[5] <https://www.lucidchart.com>

4 Experiments and Results

4.1 General study characteristics

The searches resulted in 269 publications, including 78 that were duplicates. From those 191 articles, 92 were pre-selected and 17 were excluded in the final assessment, resulting in 75 selected articles.

We found 45 new publications through the snowballing, from which 18 were pre-selected and none was excluded in the final assessment, resulting in 18 new selected articles. Therefore, a total of 93 studies were selected for the systematic review. Figure 11 summarizes our methodology and the details according to the screening stage. Also, Table 4 shows the final selected articles per digital library and inclusion criteria.

Beyond that, we provide access to the tables of publications found and selected, data extraction and quality assessment, and results obtained through the meta-analysis in [1].

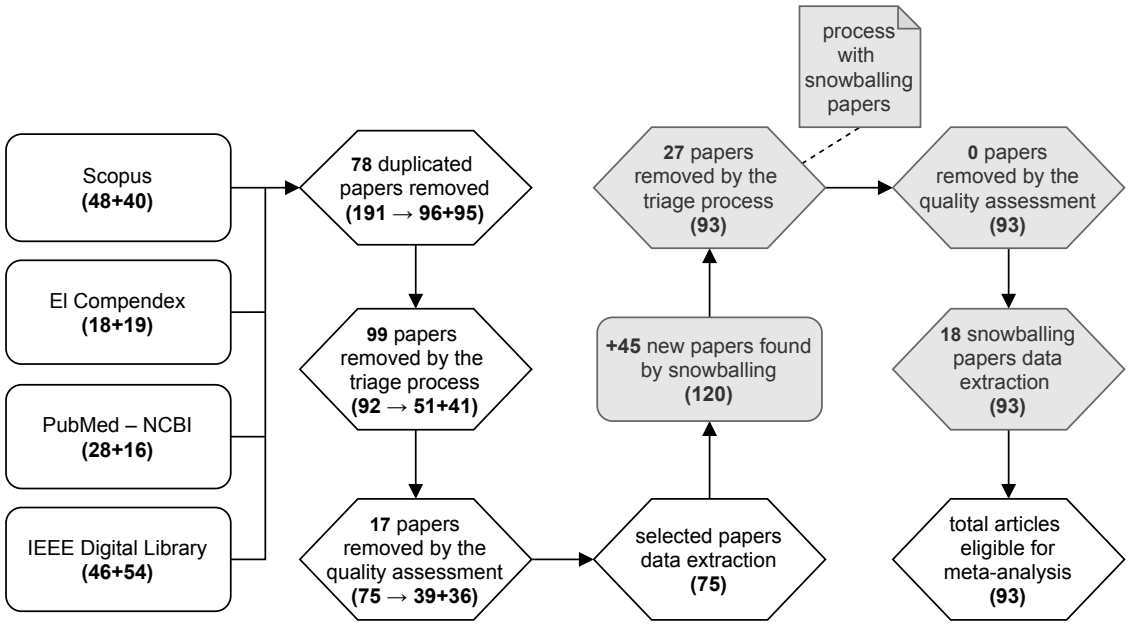


Figure 11 – Screening and selection of studies according to inclusion and exclusion criteria at different stages of the meta-analysis. The numbers between parentheses indicate the total of articles remaining after each step. The numbers separated by + indicate the total of articles from the first and second search, respectively.

[1] <https://drive.google.com/drive/folders/1Ksqa0Ok28OhDhjDVwfaQRZQ_LlmgD-0c?usp=sharing>

Table 4 – Selected Articles

Digital Library	Inclusion Criteria			
	IC1	**IC2**	**IC3**	**IC4**
Scopus	(122) (123) (124) (125)* (75)* (126) (127) (128)* (14) (129)* (130) (131)* (132) (32)* (133)* (134) (135)* (6) (136)* (137) (138)* (139)* (140) (141) (142) (143) (144)* (145) (146)* (147) (148) (149) (150)* (151)* (152)* (153)* (154)* (155)*	(2)*	(156)* (157) (56)* (158)* (76)* (159) (160)* (161)	(162)*
PubMed - NCBI	(163)*	-	(164)	-
EI Compendex	(165)* (166)* (167)* (168)*	-	(169)*	-
IEEE	(170) (171) (172) (173)* (174) (175)* (176)* (177)* (178) (179) (180)* (181)* (182)* (183) (184)	-	(185) (186)* (92)* (187) (188)	-
Snowballing	(26)* (189)* (190)* (191) (192)* (193)* (194) (195) (196)* (197)* (198)* (199)* (200)* (201)*	-	(202)* (203) (204)* (89)*	-

The articles with an * presented enough information and were also selected for the meta-analysis.

All the 93 studies were published between 2013 and 2020 and used samples that varied from 24 to 2352 individuals. The most commonly applied ML techniques for classification were SVM ($n = 33$) and ANN ($n = 30$), followed by studies that used more than one technique (M, $n = 19$). Figure 12 shows the distribution of the selected articles by year and ML technique used whereas Table 5 shows the general characteristics of the included studies.

Almost 85% of the studies ($n = 79$) extracted their samples from versions of the ABIDE, specially ABIDE I preprocessed ($n = 34$) or ABIDE without specifying the version ($n = 34$). The other articles used data from UCLA Multimodal Connectivity Database ($n = 3$), NDAR ($n = 3$), own samples ($n = 3$), own samples and ABIDE ($n = 3$), others ($n = 2$).

The majority of the studies ($n = 73$) used only rs-fMRI data for classification.

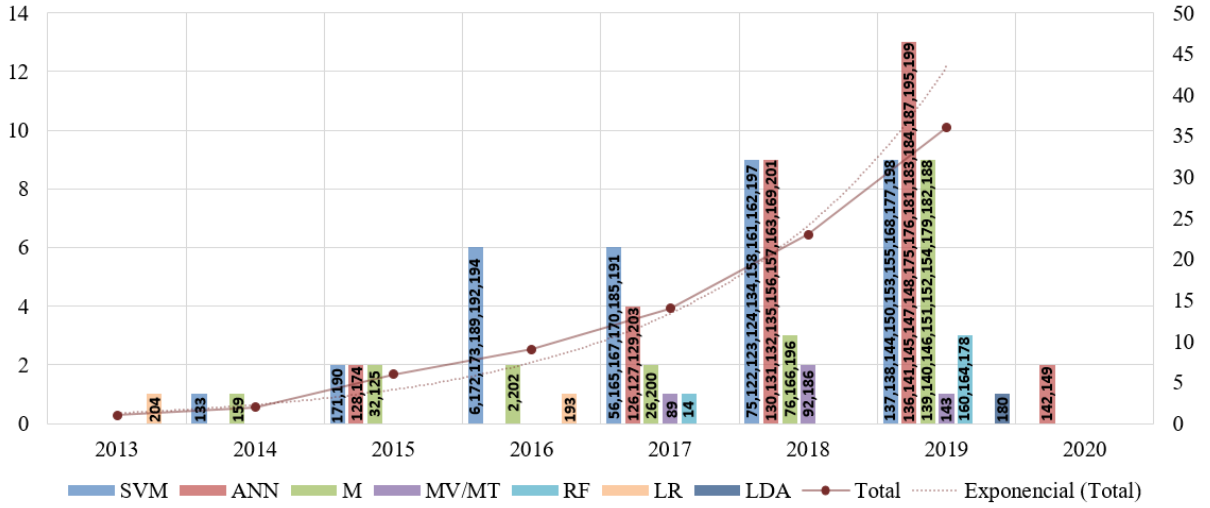


Figure 12 – Distribution of the selected studies by year of publication and type of ML technique used. The numbers inside the bars indicate each article.

Beyond that, some studies used other types of brain imaging data ($n = 11$) or phenotypic data ($n = 9$).

Regarding the subjects characteristics, we found studies that included both males and females ($n = 62$), only male subjects ($n = 5$), and studies that did not present enough information regarding the sex of the selected individuals ($n = 26$). Furthermore, there were samples with subjects both above and below 18 y.o. ($n = 42$), only below 18 y.o. ($n = 20$), only above 18 y.o. ($n = 3$), and studies without enough information ($n = 28$).

From the 93 studies selected for the systematic review, 27 (130, 123, 185, 124, 132, 157, 170, 127, 126, 191, 171, 159, 194, 134, 203, 142, 145, 161, 140, 137, 141, 183, 179, 184, 188, 187, 195) did not report any data regarding sensitivity and/or specificity and were excluded from the meta-analysis. Five articles (149, 148, 147, 122, 178) did not present the exact number of TD and ASD subjects on the sample or on the test set, making it impossible to calculate the TP/TN/FP/FN values. Two articles (172, 174) defined specific percentages of their sample as training or test sets and performed a number of random trials, thus it was not possible to define the exact number of subjects in the test set nor the proportion of ASD and TD subjects. In (164), 7 subjects had corrupted rs-fMRI imaging files, so they were included in the sMRI analysis and excluded from fMRI analysis and from sMRI-fMRI modalities fusion, but the study did not inform from which group (ASD or TD) those subjects were.

Two articles (6, 143) presented their results through bar charts without showing the exact values of sensitivity and specificity, so we decided not to include them in the meta-analysis. In (14), an RF was used as the classifier and measures of sensitivity and specificity were presented only for the external validation data-set. The main results from

Table 5 – General characteristics of the studies selected in the systematic review and of the samples included in the meta-analysis.

Characteristics	Studies (n)	Samples (n)
Total	93	132
ML technique		
SVM	33	54
ANN	30	44
M	19	2
MV/MT	4	15
RF	4	2
LR	2	4
LDA	1	8
Ridge	-	1
XGB	-	1
Affine	-	1
Dataset		
ABIDE (any version)	79	121
ABIDE	*34*	*41*
ABIDE I - preprocessed	*34*	*54*
ABIDE I + ABIDE II	*7*	*26*
ABIDE I	*2*	*0*
ABIDE II	*2*	*0*
UMCD	3	2
NDAR	3	2
Own sample	3	4
Own sample + ABIDE	3	2
Others	2	1
Type of data		
Only rs-fMRI	73	114
rs-fMRI plus other types of brain imaging data	11	14
rs-fMRI plus phenotypic information	9	4
Sex of the subjects		
Males and females	62	80
Not enough information	26	44
Only males	5	8
Age of the subjects		
Both above and below 18 y.o.	42	62
Not enough information	28	25
Below 18 y.o.	20	39
Above 18 y.o.	3	6

the article were obtained trough OOB error, but only the accuracy was reported. Since the results from the validation data-set and the ones obtained using the OOB error presented high variation, we decided not to use the results from this article. However, those three

articles were included in a sensitivity analysis to assess the effect they had on the meta-analysis results.

Finally, 55 studies - published between 2013 and 2019 - provided sufficient data for a quantitative meta-analysis. A total of 132 independent samples were extracted from those studies with sensitivity and specificity ranging from 37.5% to 100% and 20% to 100%, respectively. The techniques used for classification, according to the main results, were SVM ($n = 27$ articles/54 samples), ANN ($n = 13/44$), multiview/multitask learning (MV/MT, $n = 3/15$), Logistic Regression (LR, $n = 3/4$), Random Forest (RF, $n = 2/2$), Linear Discriminant Analysis (LDA, $n = 2/8$), multiple classifiers ($n = 2/2$), Ridge classifier ($n = 1/1$), Extreme Gradient Boosting (XGB, $n = 1/1$), and Affine-Invariant ($n = 1/1$). Figure 13 shows a conceptual map of ML techniques used throughout the articles selected for meta-analysis whereas Table 5 presents the general characteristics of the samples included.

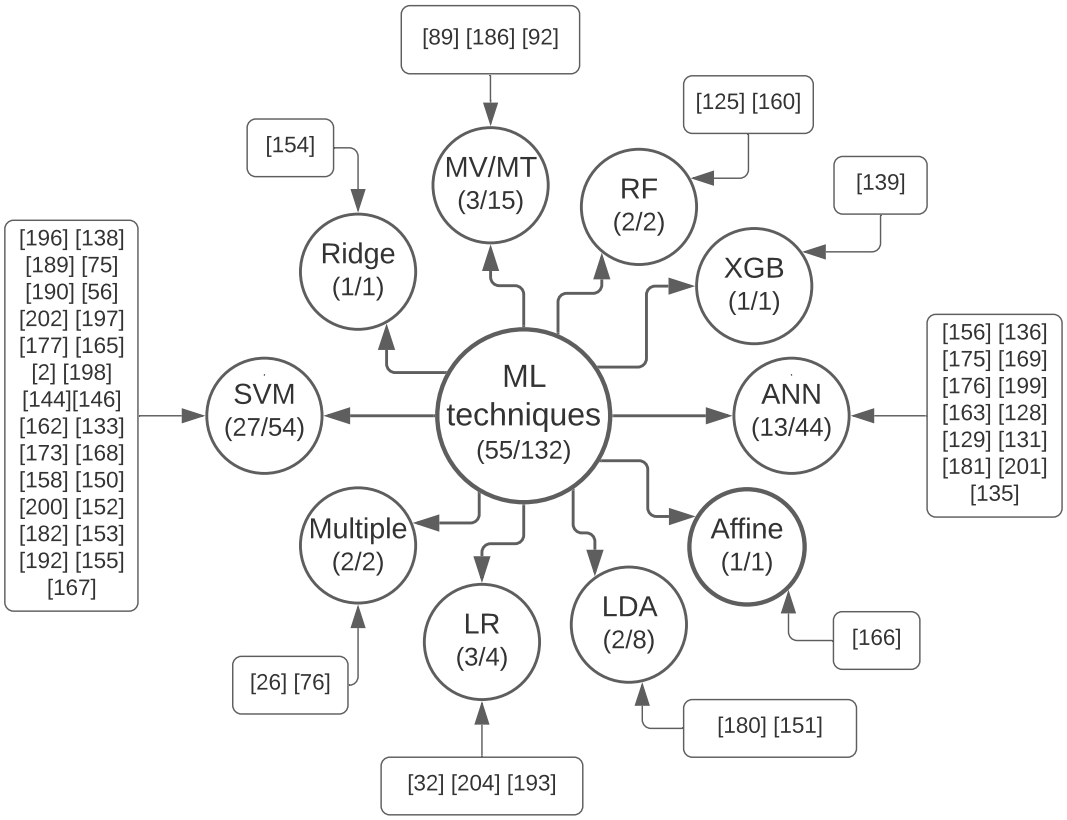


Figure 13 – Conceptual map of ML techniques used throughout the articles selected for meta-analysis (number of articles/number of samples).

4.2 Quality assessment

Figure 14 shows the distribution of the results of QUADAS-2 for RoB and applicability by considering all the selected articles.

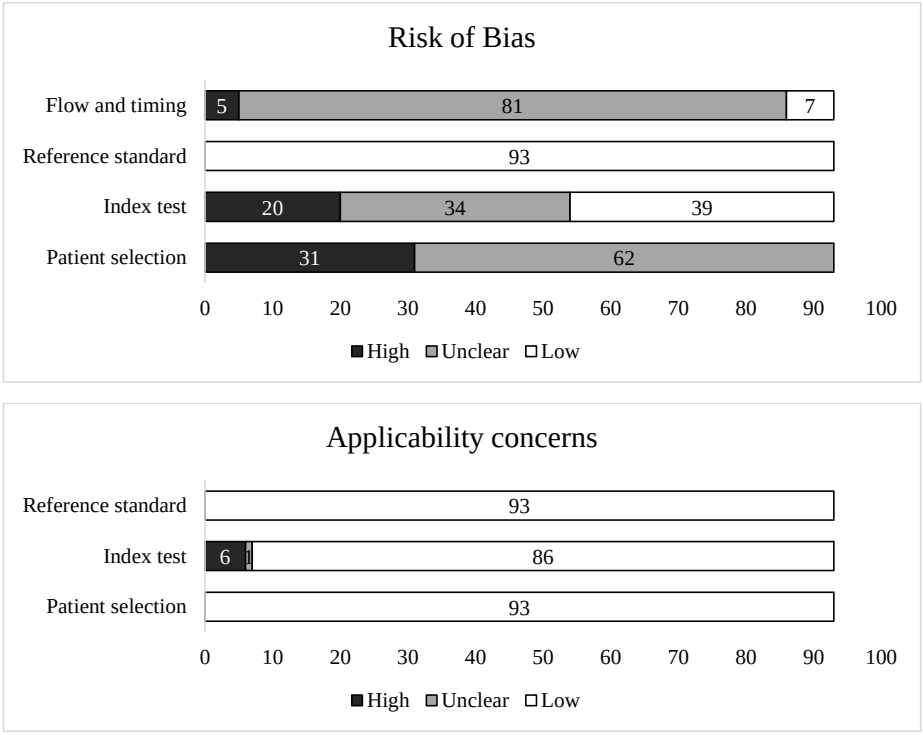


Figure 14 – Risk of bias and applicability concerns by domain in QUADAS-2.

From the graphics we can highlight that from the 93 studies, none was considered to have a low RoB by patient selection domain. Most of the articles were assessed to have an unclear RoB (62 studies) given that they used databases such as the ABIDE and did not present details regarding the recruitment of the subjects nor sufficient details of the characteristics of the subjects selected. The remaining articles (31 studies) were shown to have a high RoB due mainly to the selection of subjects in restricted intervals of age and IQ or the exclusion of female subjects.

The great majority of the studies were considered to have an unclear RoB by flow and timing domain (81 studies) mostly because they did not present the interval between the application of the index test and the reference standard nor sufficient information to conclude if all subjects received the same reference standard.

All of the articles were shown to have a low RoB by reference standard domain given that we considered the reference standards used in databases such as the ABIDE as reliable even if the article did not presented exactly what reference standards were used. For the same reason, all of the articles were assessed to have low concerns regarding applicability by the same domain.

More than half of the studies were considered to have an unclear or high RoB (54 studies) by the index test domain. Also, the great majority of the studies (86 studies) were assessed to have low concerns regarding applicability by the same domain.

Finally, all of the articles were shown to have low concerns regarding applicability

by the patient selection domain.

According to the QUADAS-2 tool, the risk of bias was judge as high in at least 1 category in 42 studies, and 12 studies presented a high risk of bias in at least 2 categories.

Figure 15 show the distribution of the results of QUADAS-2 for RoB and applicability by considering only the studies selected for the meta-analysis.

The only difference in the results between the two applications was in the RoB by index test. Since some of the results in the articles did not have enough information to be used in the meta-analysis, three articles (125, 75, 200) showed low RoB by index test when assessed as a whole but when considering only the results used for the meta-analysis they presented unclear RoB.

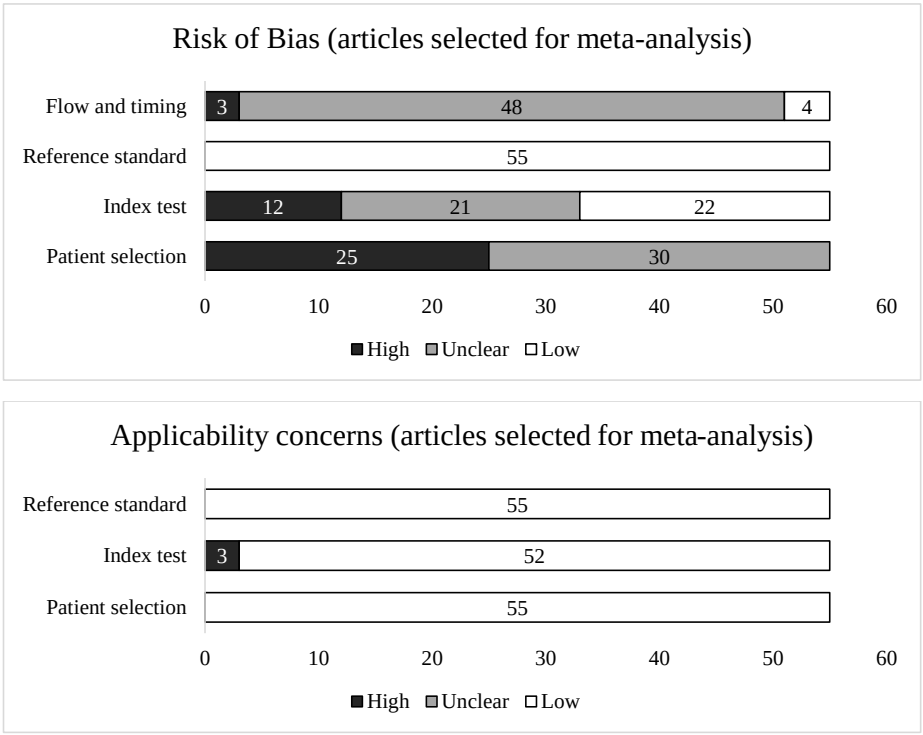


Figure 15 – Risk of bias and applicability concerns by domain in QUADAS-2 for the studies selected for the meta-analysis.

As we can see, the distribution of the results in this analysis was basically the same as those of the first one and the same considerations apply here.

Finally, according to this second analysis, the risk of bias was judge as high in at least 1 category in 29 studies, and 9 studies presented a high risk of bias in at least 2 categories.

4.3 Diagnostic accuracy

Figure 16 shows the forest plot of sensitivity and specificity for all the 132 samples included in the meta-analysis.

Using the bivariate model, machine learning-based classifiers separated ASD from TD with a sensitivity of 73.8% (95% CI: 71.8-75.8%), a specificity of 74.8% (95% CI: 72.3-77.1%), and AUC/pAUC of 0.803/0.765. A SROC curve of the included studies - along with the estimated summary point, confidence region and prediction region - is presented in Figure 17. Of the 132 samples, 40 were outside the 95% predictive region of the SROC curve, indicating heterogeneity.

Considering the type of ML technique used, only SVM and ANN classification tools were used in 5 or more articles. When only the SVM studies were analyzed, we obtained a sensitivity of 76.3% (95% CI: 73.2-79.2%), a specificity of 77.5 (95% CI: 73.7-80.8%), and AUC/pAUC of 0.832/0.748. The ANN studies had a sensitivity of 68.4% (95% CI: 65-71.5%), a specificity of 70.2% (95% CI: 66.2-73.9%), and AUC/pAUC of 0.743/0.582. The SROC curves for the studies using SVM and ANN are presented in Figure 18.

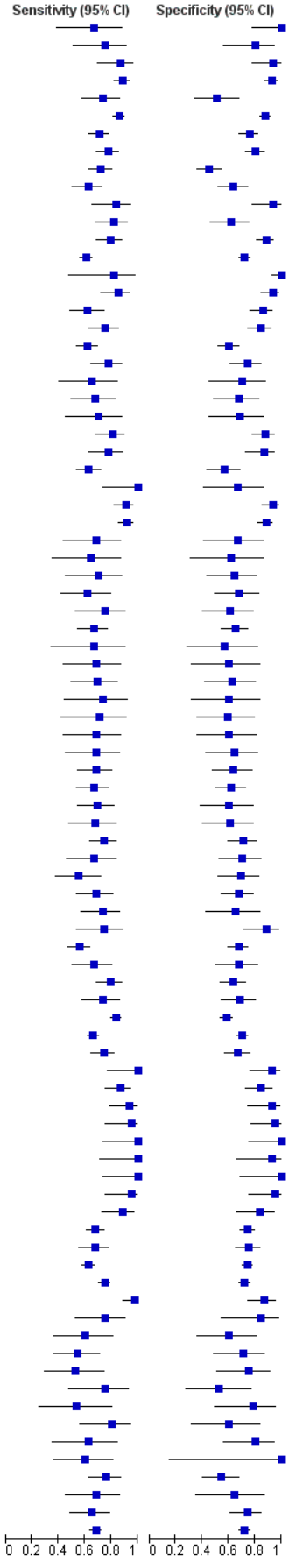
It is clear in Figure 18 that the studies using SVM obtained better sensitivity and specificity than the ANN studies. Also, the summary points of the SROC curves are outside the confidence region of one another. Therefore, we added ML technique as a moderating variable to the bivariate meta-analysis model and found a significant difference between the sensitivities ($p = 0.002$) and the specificities ($p = 0.008$).

We also conducted an analysis considering the subtype of ML technique used. For SVM, only Linear SVM (L-SVM) was used in five or more articles - sensitivity of 73.9% (95% CI: 70.2-77.2%), specificity of 77.5% (95% CI: 73.3-81.2%), and AUC/pAUC of 0.813/0.708 - whereas for ANN the same happened with CNN - sensitivity of 66.7% (95% CI: 63.3-69.9%), specificity of 70.1% (95% CI: 66.3-73.7%), and AUC/pAUC of 0.732/0.565. Thus, we compared L-SVM with other types of SVM and CNN with other types of ANN. In both cases the regression showed no effect on sensitivity or specificity (all $p > 0.1$). Finally, the regression comparing L-SVM with CNN indicated higher sensitivity ($p = 0.009$) and specificity ($p = 0.024$) in L-SVM studies.

Regression with sample size as moderator showed a significant effect on both sensitivity ($p = 0.004$) and specificity ($p < 0.001$) when analysing all the samples together. Figure 19 shows the linear regression models with sample size predicting sensitivity and specificity and indicates that bigger sample sizes tend to obtain worse accuracies.

However, the same analysis segregating the studies per type of ML technique used indicated a significant effect on specificities ($p = 0.001$) and no effect on sensitivities ($p = 0.152$) for the studies using SVM, with worse specificities in studies with larger samples. Also, no significant effect was found for the ANN studies (all $p > 0.1$).

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Uddin et al. (2013) - Salience (Independent)	10	0	5	15	0.67 [0.38, 0.88]	1.00 [0.78, 1.00]
Uddin et al. (2013) - Salience (Own sample)	15	4	5	16	0.75 [0.51, 0.91]	0.80 [0.56, 0.94]
Price et al. (2014) - Multiple	26	2	4	28	0.87 [0.69, 0.96]	0.93 [0.78, 0.99]
Chen et al. (2015) - Diagnostic	112	9	14	117	0.89 [0.82, 0.94]	0.93 [0.87, 0.97]
Dodero et al. (2015) - Kernel (Grassmann)	31	18	11	19	0.74 [0.58, 0.86]	0.51 [0.34, 0.68]
Iidaka (2015) - Resting	268	40	44	288	0.86 [0.82, 0.90]	0.88 [0.84, 0.91]
Plitt et al. (2015) - Functional	105	36	43	112	0.71 [0.63, 0.78]	0.76 [0.68, 0.82]
Chen et al. (2016) - Multivariate	87	25	25	103	0.78 [0.69, 0.85]	0.80 [0.73, 0.87]
Ghiassian et al. (2016) - Using	78	63	30	52	0.72 [0.63, 0.80]	0.45 [0.36, 0.55]
Kassraian et al. (2016) - Promises	48	28	29	49	0.62 [0.51, 0.73]	0.64 [0.52, 0.74]
Ren et al. (2016) - Identifying	25	2	5	28	0.83 [0.65, 0.94]	0.93 [0.78, 0.99]
Wee et al. (2016) - Diagnosis	37	18	8	29	0.82 [0.68, 0.92]	0.62 [0.46, 0.75]
Yahata et al. (2016) - small	59	12	15	95	0.80 [0.69, 0.88]	0.89 [0.81, 0.94]
Abraham et al. (2017) - Deriving	246	130	157	338	0.61 [0.56, 0.66]	0.72 [0.68, 0.76]
Emerson et al. (2017) - Functional	9	0	2	48	0.82 [0.48, 0.98]	1.00 [0.93, 1.00]
Jun et al. (2017) - Region	40	5	7	68	0.85 [0.72, 0.94]	0.93 [0.85, 0.98]
Kam et al. (2017) - Multiple (NYU)	36	10	22	62	0.62 [0.48, 0.74]	0.86 [0.76, 0.93]
Kam et al. (2017) - Multiple (UM)	46	11	15	61	0.75 [0.63, 0.86]	0.85 [0.74, 0.92]
Tejwani (2017) - Autism	91	58	56	88	0.62 [0.54, 0.70]	0.60 [0.52, 0.68]
Wang et al. (2017) - Multi (NYU)	42	15	12	43	0.78 [0.64, 0.88]	0.74 [0.61, 0.85]
Wang et al. (2017) - Multi (STANFORD)	13	6	7	14	0.65 [0.41, 0.85]	0.70 [0.46, 0.88]
Wang et al. (2017) - Multi (UM-1)	23	10	11	21	0.68 [0.49, 0.83]	0.68 [0.49, 0.83]
Wang et al. (2017) - Multi (YALE)	14	7	6	15	0.70 [0.46, 0.88]	0.68 [0.45, 0.86]
Zhu et al. (2017) - tensor (NYU)	46	9	11	65	0.81 [0.68, 0.90]	0.88 [0.78, 0.94]
Zhu et al. (2017) - tensor (UM)	35	6	10	39	0.78 [0.63, 0.89]	0.87 [0.73, 0.95]
Aghdam et al. (2018) - Combination	73	30	43	39	0.63 [0.53, 0.72]	0.57 [0.44, 0.68]
Bernas et al. (2018) - Wavelet	12	6	0	12	1.00 [0.74, 1.00]	0.67 [0.41, 0.87]
Dekhil et al. (2018) - Identifying	71	5	7	73	0.91 [0.82, 0.96]	0.94 [0.86, 0.98]
Dekhil et al. (2018) - Using	113	19	10	141	0.92 [0.86, 0.96]	0.88 [0.82, 0.93]
Heinsfeld et al. (2018) - Identification (CALTECH)	13	6	6	12	0.68 [0.43, 0.87]	0.67 [0.41, 0.87]
Heinsfeld et al. (2018) - Identification (CMU)	9	5	5	8	0.64 [0.35, 0.87]	0.62 [0.32, 0.86]
Heinsfeld et al. (2018) - Identification (KKI)	14	10	6	18	0.70 [0.46, 0.88]	0.64 [0.44, 0.81]
Heinsfeld et al. (2018) - Identification (LEUVEN)	18	11	11	23	0.62 [0.42, 0.79]	0.68 [0.49, 0.83]
Heinsfeld et al. (2018) - Identification (MAX_MUN)	18	11	6	17	0.75 [0.53, 0.90]	0.61 [0.41, 0.78]
Heinsfeld et al. (2018) - Identification (NYU)	50	35	25	65	0.67 [0.55, 0.77]	0.65 [0.55, 0.74]
Heinsfeld et al. (2018) - Identification (OHSU)	8	6	4	8	0.67 [0.35, 0.90]	0.57 [0.29, 0.82]
Heinsfeld et al. (2018) - Identification (OLIN)	13	6	6	9	0.68 [0.43, 0.87]	0.60 [0.32, 0.84]
Heinsfeld et al. (2018) - Identification (PITT)	20	10	9	17	0.69 [0.49, 0.85]	0.63 [0.42, 0.81]
Heinsfeld et al. (2018) - Identification (SBL)	11	6	4	9	0.73 [0.45, 0.92]	0.60 [0.32, 0.84]
Heinsfeld et al. (2018) - Identification (SDSU)	10	9	4	13	0.71 [0.42, 0.92]	0.59 [0.36, 0.79]
Heinsfeld et al. (2018) - Identification (STANF.)	13	8	6	12	0.68 [0.43, 0.87]	0.60 [0.36, 0.81]
Heinsfeld et al. (2018) - Identification (TRINITY)	15	9	7	16	0.68 [0.45, 0.86]	0.64 [0.43, 0.82]
Heinsfeld et al. (2018) - Identification (UCLA)	37	16	17	28	0.69 [0.54, 0.80]	0.64 [0.48, 0.78]
Heinsfeld et al. (2018) - Identification (UM)	44	28	22	46	0.67 [0.54, 0.78]	0.62 [0.50, 0.73]
Heinsfeld et al. (2018) - Identification (USM)	32	10	14	15	0.70 [0.54, 0.82]	0.60 [0.39, 0.79]
Heinsfeld et al. (2018) - Identification (YALE)	19	11	9	17	0.68 [0.48, 0.84]	0.61 [0.41, 0.78]
Huang et al. (2018) - Enhancing	59	22	20	54	0.75 [0.64, 0.84]	0.71 [0.60, 0.81]
Li et al. (2018) - novel (LEUVEN)	18	10	9	24	0.67 [0.46, 0.83]	0.71 [0.53, 0.85]
Li et al. (2018) - novel (UCLA)	20	12	16	27	0.56 [0.38, 0.72]	0.69 [0.52, 0.83]
Li et al. (2018) - novel (UM)	33	21	15	44	0.69 [0.54, 0.81]	0.68 [0.55, 0.79]
Li et al. (2018) - novel (USM)	28	8	10	15	0.74 [0.57, 0.87]	0.65 [0.43, 0.84]
Mastrovito et al. (2018) - Differences	20	3	7	24	0.74 [0.54, 0.89]	0.89 [0.71, 0.98]
Sen et al. (2018) - general	84	52	67	108	0.56 [0.47, 0.64]	0.68 [0.60, 0.75]
Tolan et al. (2018) - Graph	28	12	14	25	0.67 [0.50, 0.80]	0.68 [0.50, 0.82]
Wang et al. (2018) - Sparse (NYU)	62	38	16	67	0.79 [0.69, 0.88]	0.64 [0.54, 0.73]
Wang et al. (2018) - Sparse (UM-1)	31	17	11	37	0.74 [0.58, 0.86]	0.69 [0.54, 0.80]
Wong et al. (2018) - Riemannian	338	194	65	274	0.84 [0.80, 0.87]	0.59 [0.54, 0.63]
Xing et al. (2018) - Convolutional	350	168	177	401	0.66 [0.62, 0.70]	0.70 [0.67, 0.74]
Zhao et al. (2018) - 3D	74	33	26	67	0.74 [0.64, 0.82]	0.67 [0.57, 0.76]
Zhou et al. (2018) - Multi (ASD4 - KKI)	14	2	0	26	1.00 [0.77, 1.00]	0.93 [0.76, 0.99]
Zhou et al. (2018) - Multi (ASD4 - NYU)	47	9	7	49	0.87 [0.75, 0.95]	0.84 [0.73, 0.93]
Zhou et al. (2018) - Multi (ASD4 - UCLA_1)	29	2	2	24	0.94 [0.79, 0.99]	0.92 [0.75, 0.99]
Zhou et al. (2018) - Multi (ASD4 - YALE)	19	1	1	21	0.95 [0.75, 1.00]	0.95 [0.77, 1.00]
Zhou et al. (2018) - Multi (ASD5 - LEUVEN_2)	12	0	0	13	1.00 [0.74, 1.00]	1.00 [0.75, 1.00]
Zhou et al. (2018) - Multi (ASD5 - OHSU)	11	1	0	13	1.00 [0.72, 1.00]	0.93 [0.66, 1.00]
Zhou et al. (2018) - Multi (ASD5 - PITT)	12	0	0	10	1.00 [0.74, 1.00]	1.00 [0.69, 1.00]
Zhou et al. (2018) - Multi (ASD5 - STANFORD)	19	1	1	19	0.95 [0.75, 1.00]	0.95 [0.75, 1.00]
Zhou et al. (2018) - Multi (ASD5 - UM_1)	31	5	4	26	0.89 [0.73, 0.97]	0.84 [0.66, 0.95]
Aghdam et al. (2019) - Diagnosis	143	64	67	185	0.68 [0.61, 0.74]	0.74 [0.68, 0.80]
Brahim et al. (2019) - Classification	50	24	24	74	0.68 [0.56, 0.78]	0.76 [0.66, 0.84]
Chen et al. (2019) - Convolutional	276	123	164	357	0.63 [0.58, 0.67]	0.74 [0.70, 0.78]
Dammu et al. (2019) - Employing	317	125	106	321	0.75 [0.71, 0.79]	0.72 [0.68, 0.76]
Eill et al. (2019) - Functional	45	6	1	41	0.98 [0.88, 1.00]	0.87 [0.74, 0.95]
Ell Gazzar et al. (2019) - Simple (ABIDEII-ETH1)	18	2	6	11	0.75 [0.53, 0.90]	0.85 [0.55, 0.98]
Ell Gazzar et al. (2019) - Simple (ABIDEII-IU1)	12	8	8	12	0.60 [0.36, 0.81]	0.60 [0.36, 0.81]
Ell Gazzar et al. (2019) - Simple (ABIDEII-OILH2)	19	7	16	17	0.54 [0.37, 0.71]	0.71 [0.49, 0.87]
Ell Gazzar et al. (2019) - Simple (ABIDEII-TCD1)	11	5	10	15	0.52 [0.30, 0.74]	0.75 [0.51, 0.91]
Ell Gazzar et al. (2019) - Simple (ABIDEII-USM1)	12	8	4	9	0.75 [0.48, 0.93]	0.53 [0.28, 0.77]
Ell Gazzar et al. (2019) - Simple (CMU)	7	3	6	11	0.54 [0.25, 0.81]	0.79 [0.49, 0.95]
Ell Gazzar et al. (2019) - Simple (LEUVEN2)	16	6	4	9	0.80 [0.56, 0.94]	0.60 [0.32, 0.84]
Ell Gazzar et al. (2019) - Simple (OLIN)	10	4	6	16	0.63 [0.35, 0.85]	0.80 [0.56, 0.94]
Ell Gazzar et al. (2019) - Simple (STANFORD)	12	0	8	2	0.60 [0.36, 0.81]	1.00 [0.16, 1.00]
Ell Gazzar et al. (2019) - Simple (UM1)	42	25	13	30	0.76 [0.63, 0.87]	0.55 [0.41, 0.68]
Ell Gazzar et al. (2019) - Simple (UM2)	15	5	7	9	0.68 [0.45, 0.86]	0.64 [0.35, 0.87]
Ell Gazzar et al. (2019) - Simple (USM)	28	15	15	43	0.65 [0.49, 0.79]	0.74 [0.61, 0.85]
Eslami et al. (2019) - ASD-DiagNet	345	147	160	383	0.68 [0.64, 0.72]	0.72 [0.68, 0.76]



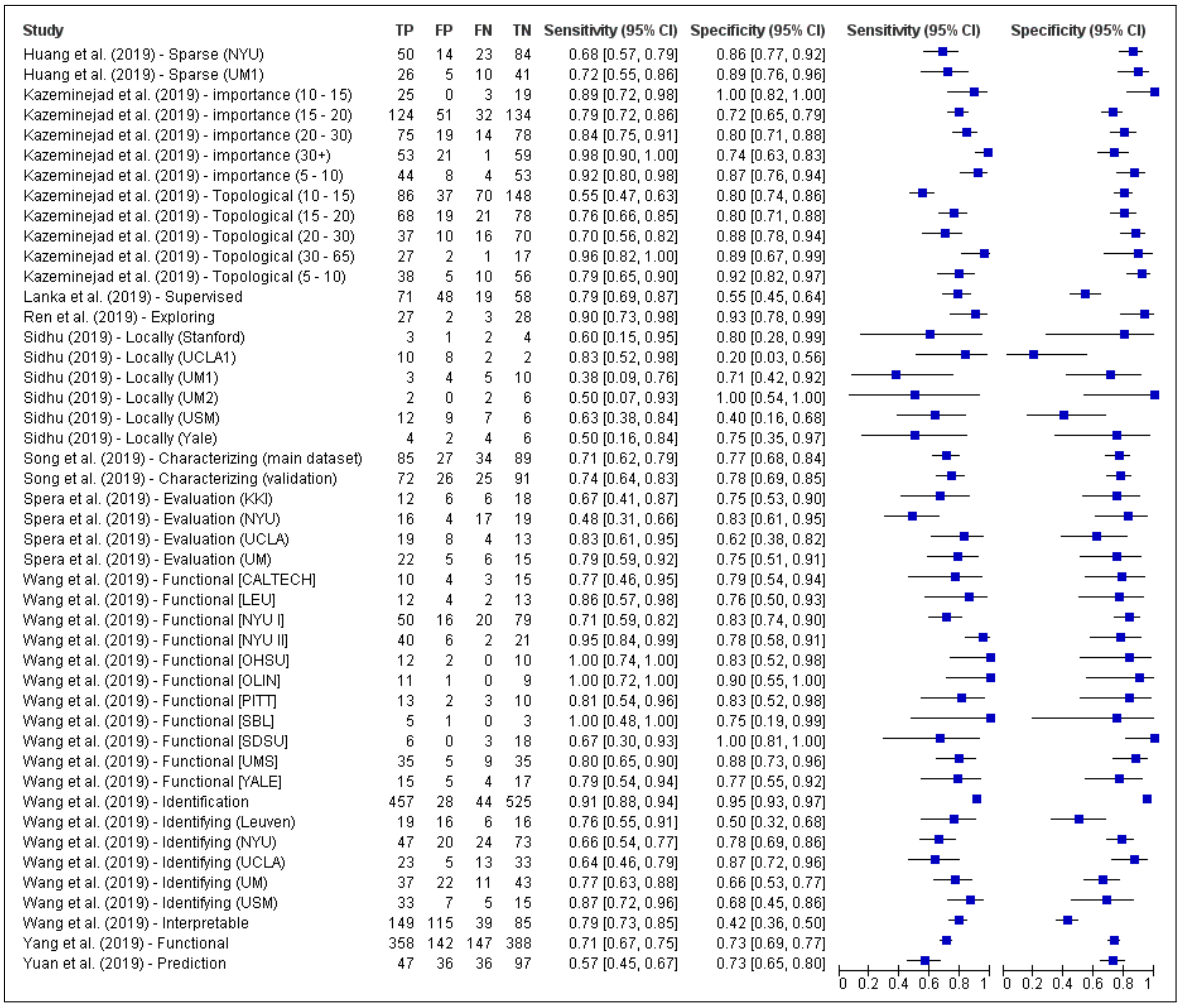


Figure 16 – Paired forest plot of all samples included in the meta-analysis.

Regression with year of publication did not show any effect on sensitivity ($p = 0.250$) or specificity ($p = 0.283$), even segregating per type of ML technique: SVM - sensitivity ($p = 0.913$), specificity ($p = 0.537$); ANN - sensitivity ($p = 0.062$), specificity ($p = 0.242$).

No significant effects of sex (only males against males and females studies) or FIQ (neither considering the mean FIQ nor comparing studies that used only high-functioning subjects with the ones that used high- and low-functioning subjects) on sensitivity or specificity (all $p > 0.1$) were observed.

Regression with the mean age of the subjects did not show any effect on sensitivity or specificity (all $p > 0.1$). However, comparison between samples with subjects under 18 years old and samples composed of individuals both under and above that age showed a significant difference between the specificities ($p = 0.020$) but no effect on sensitivity ($p = 0.225$), indicating higher specificity in studies that used only subjects under 18 y.o. (77.6% - 95% CI: 73-81.6% - versus 70.5% - 95% CI: 66.6-74.1%). Segregating per type of ML technique, only SVM had enough studies to conduct the analysis but the results

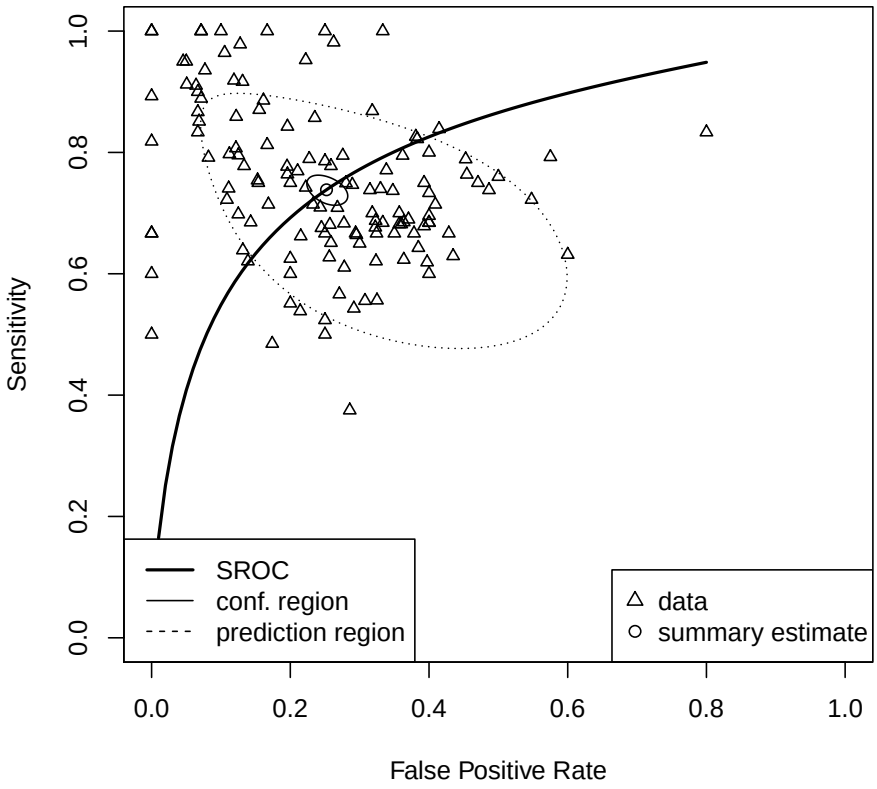


Figure 17 – SROC curve of all the included studies with summary estimate.

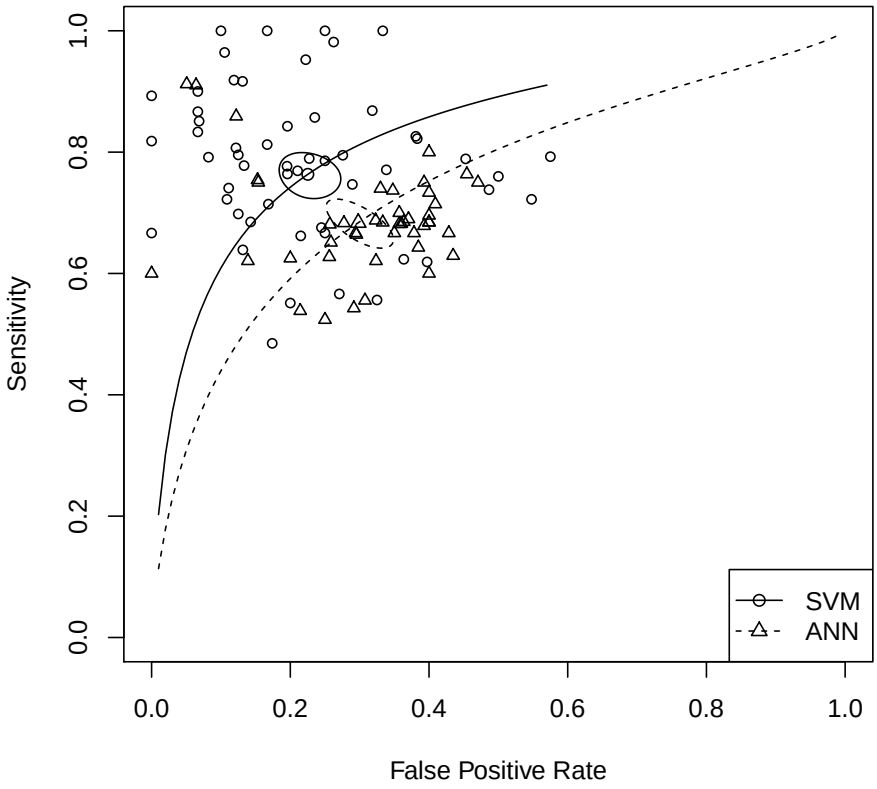


Figure 18 – SROC curves of the studies using SVM and ANN with their summary estimates.

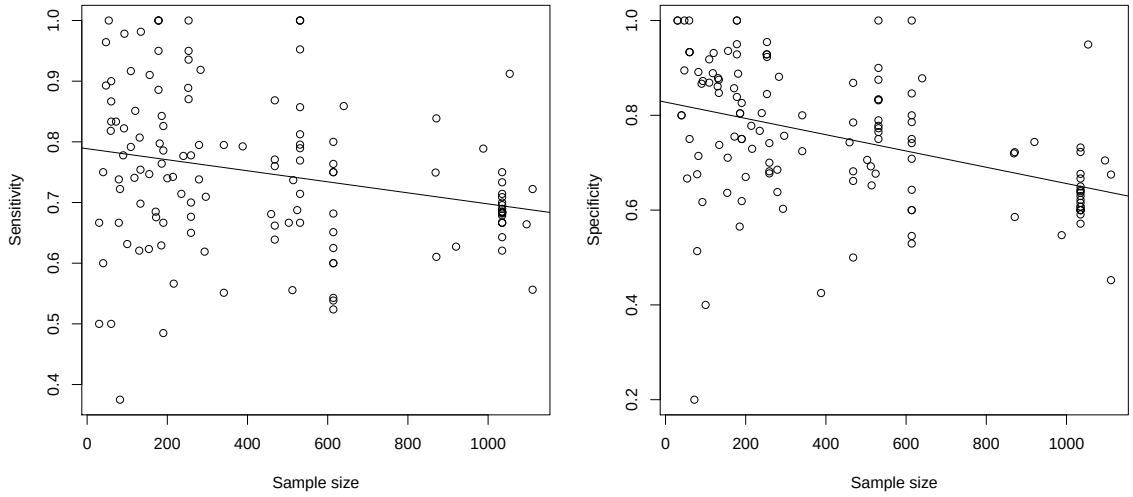


Figure 19 – Linear regression models with sample size predicting sensitivity (left) and specificity (right) for all the studies.

did not show any effect on sensitivity ($p = 0.790$) or specificity ($p = 0.427$). Sensitivity analysis considering other age thresholds (19, 20, 21 y.o.) resulted in the same conclusions.

Regression considering the database or source of the sample (ABIDE, ABIDE I preprocessed or ABIDE I + ABIDE II) indicated a significant effect on the sensitivity when comparing ABIDE with ABIDE I preprocessed ($p = 0.046$) or ABIDE I + ABIDE II ($p = 0.043$) and in both cases the ABIDE group presented higher sensitivity (77.1% - 95% CI: 73.2-80.6% - versus 72% - 95% CI: 69-74.9% - and 69.2% - 95% CI: 65.8-72.4% - respectively). All the other analysis indicated no significant effects on sensitivity or specificity (all $p > 0.1$). However, the same analysis with SVM samples (ABIDE or ABIDE I preprocessed) did not indicate any effect on sensitivity ($p = 0.756$) or specificity ($p = 0.731$).

We conducted another analysis comparing the studies that used any version of ABIDE with studies that used databases or samples other than ABIDE (Own sample, NDAR, UMCD). The regression indicated higher sensitivity ($p = 0.024$) and specificity ($p = 0.045$) in studies that used databases or samples other than ABIDE (Sensitivity: 81.8% - 95% CI: 73.4-88.1% - versus 73.2% - 95% CI: 71.1-75.2%; Specificity: 83% - 95% CI: 72.5-90% - versus 74.1% - 95% CI: 71.6-76.5%).

Regression with type of data as moderator (only rs-fMRI or rs-fMRI plus other data types) showed a significant difference between the sensitivities ($p = 0.002$) and the specificities ($p = 0.047$), indicating higher sensitivity and specificity in studies that used other types of data together with rs-fMRI (Sensitivity: 84.7% - 95% CI: 78.5-89.4% - versus 72.8% - 95% CI: 70.6-74.8%; Specificity: 81% - 95% CI: 74.1-86.3% - versus 73.9% - 95% CI: 71.3-76.4%).

Regression with atlas as moderator (Automated Anatomical Labelling with 90

ROIs [AAL90], AAL116 or Craddock with 200 ROIs [CC200]), indicated that: studies using AAL90 obtained better specificity (74.9% - 95% CI: 68.7-80.1%; $p = 0.001$) than studies using CC200 (64.4% - 95% CI: 60.7-67.9%) but no significant effect was observed on the sensitivity ($p = 0.56$); studies using AAL116 obtained better sensitivity (77.7% - 95% CI: 73.7-81.2%; $p = 0.001$) and specificity (78.2% - 95% CI: 72.8-82.9%; $p < 0.001$) than studies using CC200 (Sensitivity: 68% - 95% CI: 65.4-70.4%); there was no significant effect on sensitivity ($p = 0.054$) or specificity ($p = 0.397$) between studies using AAL 116 or 90. Figure 20 shows the SROC curves for the studies using AAL90, AAL116 or CC200.

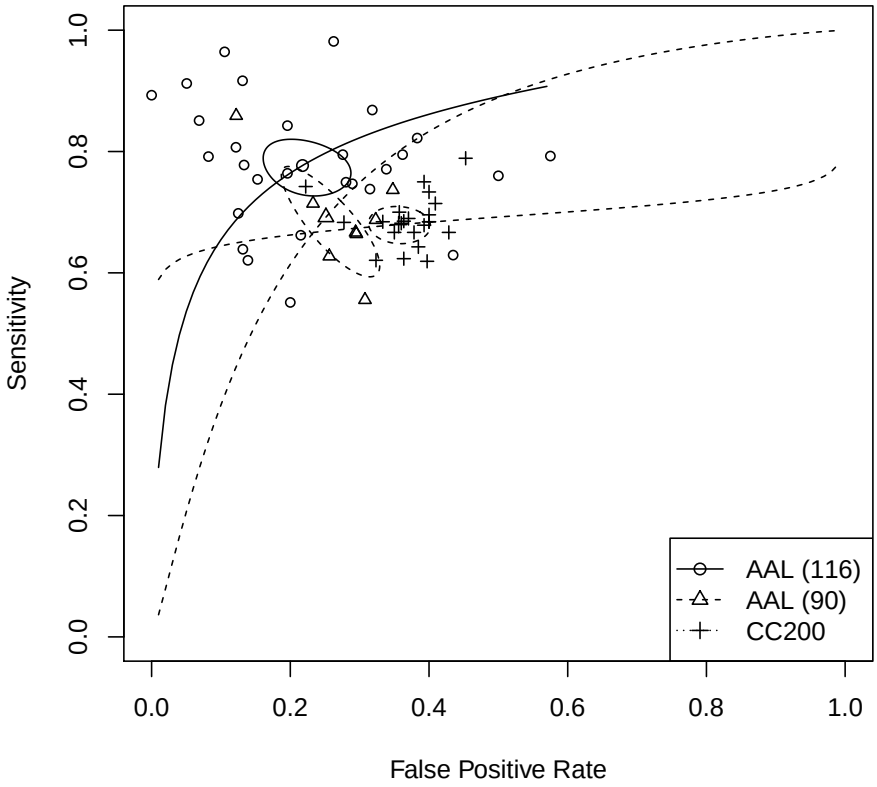


Figure 20 – SROC curves of the studies using AAL90, AAL116 or CC200 with their summary estimates.

Regression considering the number of ROIs used showed significant effects on sensitivity ($p = 0.043$) and specificity ($p = 0.018$). When segregating per ML technique, there was a significant effect on sensitivity ($p = 0.029$) for the SVM studies but no effect on specificity ($p = 0.089$) whereas for the ANN studies there was a significant effect on specificity ($p = 0.016$) but no effect on sensitivity ($p = 0.557$). For all the significant effects, the linear regression models indicated lower values of sensitivity/specificity as the number of regions increased (Figure 21).

Regression with type of feature used as moderator (Pearson Correlation [PC], PC (Fisher-transformed) or others), indicated that: studies using PC (Fisher-transformed) obtained better sensitivity (76.7% - 95% CI: 71.3-81.3%; $p = 0.001$) and specificity (81% - 95% CI: 75.6-85.4%; $p < 0.001$) than studies using PC (Sensitivity: 68.9% - 95% CI:

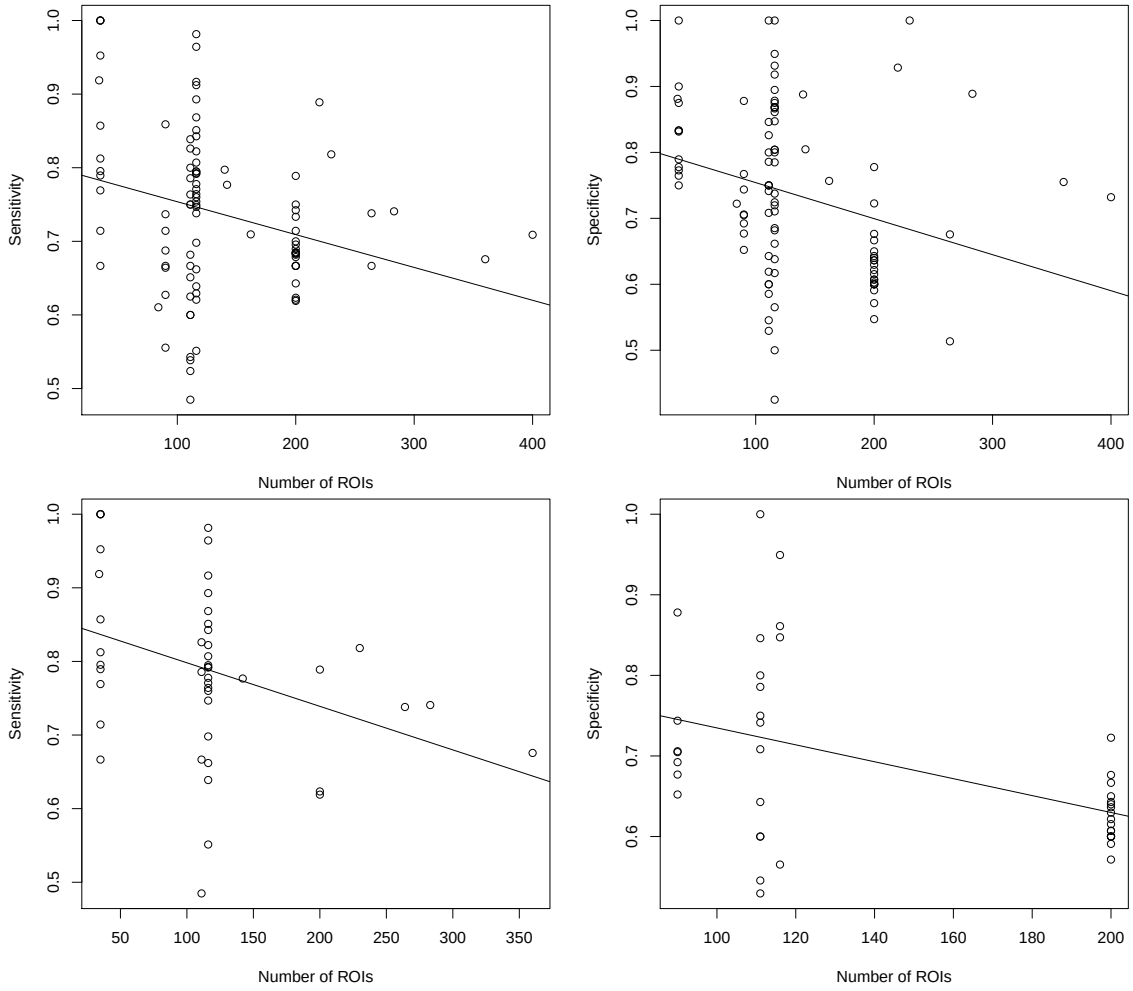


Figure 21 – Linear regression models with number of ROIs predicting sensitivity and specificity. The upper graphics refer to all the studies whereas the down graphics refer to SVM studies (left) and ANN studies (right).

66.8-70.9%; Specificity: 68.3% - 95% CI: 64.3-72.1%); similarly, studies using other features obtained better sensitivity (73.5% - 95% CI: 70.6-76.2%; $p = 0.031$) and specificity (74.7% - 95% CI: 71-78%; $p = 0.024$) than studies using PC; there was no significant effect on sensitivity ($p = 0.173$) or specificity ($p = 0.072$) between studies using PC (Fisher-transformed) and other features. Figure 22 shows the SROC curves for the studies using PC, PC (Fisher-transformed) or other features.

Regression considering the number of domains with low RoB in QUADAS-2 results (one or two) showed a significant difference between the specificities ($p < 0.001$) but no effect on sensitivity ($p = 0.236$), indicating higher specificity in studies that had only one domain with a low risk of bias (78.4% - 95% CI: 75.5-81.1% - versus 69.6% - 95% CI: 65.9-73%).

Analysis considering the RoB of the index test obtained from QUADAS-2 (high, unclear or low) indicated that: studies with unclear RoB obtained better sensitivity (80.5% - 95% CI: 76.1-84.2%; $p = 0.001$) than studies with high RoB (70.1% - 95% CI: 66.3-73.5%)

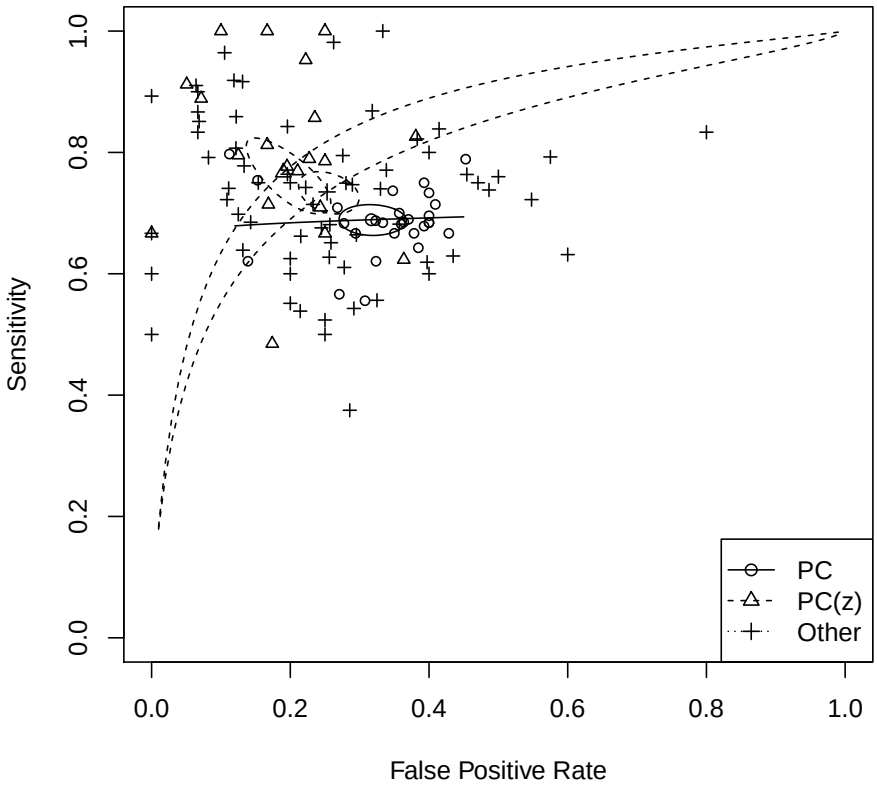


Figure 22 – SROC curves of the studies using PC, PC (Fisher-tranformed) or other features with their summary estimates.

but no significant effect was observed on the specificity ($p = 0.127$); studies with high RoB obtained better specificity (76.6% - 95% CI: 73.2-79.8%; $p = 0.011$) than studies with low RoB (69.9% - 95% CI: 66.2-73.3%) but no significant effect was observed on the sensitivity ($p = 0.373$); studies with unclear RoB obtained better sensitivity ($p = 0.003$) and specificity (81.3% - 95% CI: 76.9-85%; $p < 0.001$) than studies with low RoB (Sensitivity: 72.1% - 95% CI: 69.6-74.4%).

We conducted another analysis splitting the studies with low RoB between the ones that performed a temporal or geographic validation and the ones that performed a split-sample validation (117). The conclusions were basically the same with exception of the comparison between the studies with high RoB and the ones with low RoB (split-sample) that did not indicate a significant effect on sensitivity ($p = 0.291$) or specificity ($p = 0.148$). Also, there was no effect on sensitivity ($p = 0.451$) or specificity ($p = 0.441$) when comparing the studies that used a split-sample validation and the ones that used a temporal or geographic validation.

Sensitivity analysis including three more articles (6, 143, 14) - four new samples - that were initially excluded from the meta-analysis (as cited in Section 4.1) indicated no significant change in overall sensitivity (73.7% - 95% CI: 71.7-75.6%) or specificity (75% - 95% CI: 72.7-77.2%).

We repeated all the tests that could be impacted by the addition of those articles.

In general, their influence on the outcome was minor but in three analyses the conclusions from the meta-regression differed.

The regression with type of data as moderator showed the same conclusion for the sensitivities ($p = 0.002$) but no significant effect between the specificities ($p = 0.057$).

The regression considering the database or source of the sample indicated no significant effect on sensitivity when comparing ABIDE with ABIDE I preprocessed ($p = 0.097$) or ABIDE I + ABIDE II ($p = 0.087$).

Regression with atlas as moderator showed a significant effect on sensitivity ($p = 0.045$) between studies using AAL 116 or 90, indicating higher sensitivity in studies that used the AAL 116 (78% - 95% CI: 74.1-81.4% - versus 69.2% - 95% CI: 61.4-76%).

Appendix A presents the main results from the meta-analysis in a tabulated way, whereas Appendix B and Appendix C show the same for the sensitivity analysis of the age threshold and the sensitivity analysis including three more articles, respectively.

4.4 Discussion

4.4.1 ML techniques and sample size

As we can see from Figure 12, the number of publications in the area has been increasing over the years following an exponential trend and SVM was the most used classification technique, especially until 2018. Also, from the 54 samples that used SVM for classification, 33 of them used a linear SVM.

However, we can observe an increasing number of studies using ANN techniques to perform classification, making it the second most used technique in absolute numbers and the first when considering only articles published in 2019.

Throughout the selected articles that used ANNs as classifiers, many of them used deep-learning methods. In fact, the ANN technique most used in the selected publications was the Convolutional Neural Network - 16 out of the 44 samples using ANN for classification.

Even though the use of ANN techniques in the field is increasing in the last years, regression considering the type of classification algorithm used indicated better results for SVM against ANN on both sensitivity and specificity. Also, the analysis comparing L-SVM and CNN resulted in the same conclusions. In all those cases, the difference was about 7 percentage points.

More complex models tend to be more powerful but generally have more hyperparameters. Therefore, they are potentially more capable of explaining noise in the data and overfitting (77). Also, it is not clear that those complex models - such as nonlinear classi-

fiers - always provide a significant advantage in practical performance. However, this can be a reflection of the small number of examples available instead of indicating an absence of complicated relationships between features (3).

We conducted a regression analysis to investigate the effect of the sample size on the results. Considering all the samples selected for the meta-analysis we found a tendency of worse results on both sensitivity and specificity by increasing the sample size. This same trend was also observed on specificities when considering only the SVM studies.

By increasing the number of subjects used to train a classifier we end up with a more diverse and variable sample, often composed of data from different acquisition sites. Thus, this tendency of obtaining lower accuracies when using bigger samples is not a surprise.

However, using only the ANN studies, we could not find any significant effect of the sample size on the results. Considering that great part of the ANN methods were also deep-learning methods and that more complex models may demand larger samples in order to avoid overfitting, our analysis suggests that ANN techniques may have an advantage when dealing with larger samples in comparison to SVM techniques.

It is also interesting to mention that other classification techniques were applied throughout the publications. In special, Random Forests were used as classifiers in two studies (125, 160) and both of them obtained results ranging from 87 to 97% of sensitivity and specificity when considering the Out-of-bag (OOB) error - that is estimated internally and may eliminate the need for a set-aside testing set. However, those techniques were not used in enough articles to be included in the analysis and need more investigation.

4.4.2 Subjects characteristics

As commented before, several studies indicate gender, age, and IQ differences on autistic symptoms and impairments. Also, there is an imbalanced male-female ratio in ASD and an underrepresentation of females in research and clinical practice. Therefore, it is essential to understand how those variables may affect the classification accuracy to eventually obtain a clinically useful ML diagnostic tool.

In our analysis, we could not find any significant effect of the sex of the subjects or their FIQ on sensitivities or specificities. However, we must highlight some questions.

The regression considering the sex of the subjects compared the articles that used only male subjects with the ones without a sex restriction - whose samples were composed of males and females. It would be interesting to include studies that selected only female subjects in the analysis, but none of the articles did so.

It is worth noticing that only eight samples (from 5 articles) were composed of

males, far less than the eighty samples (from 37 articles) with males and females. Moreover, the number of females within those samples may be small compared to the number of males.

From the articles selected in the systematic review, only (191) and (161) performed tests considering different categories of gender. They obtained higher classification accuracies for females compared to males, even though the number of training samples for females were significantly lower.

Regarding the FIQ of the individuals, we performed tests considering the mean FIQ and comparing the samples composed of high-functioning subjects against the ones with both high- and low-functioning subjects. However, we found a lack of information in the publications: from the 55 studies (132 samples) only 11 (18 samples) were included in the first and 22 (42 samples) in the second test. Also, the number of low-functioning individuals within those samples may be small compared to the number of high-functioning ones.

Regression with the mean age of the subjects did not show any effect on sensitivity or specificity but less than half of the articles presented enough information to be included in this analysis (24 articles, 61 samples).

Analysis considering an adulthood threshold indicated higher specificity in studies that used only subjects under 18 y.o. when compared with studies without this restriction. We would like to include publications that selected only individuals above 18 y.o. in the analysis, but only 3 articles (6 samples) presented that characteristic.

This result is in accordance with (172), in which lower classification performances on the full ABIDE dataset were obtained when compared to tests using only the adolescents. Also, (174) found a clear improvement in the classification performance when adult males and adolescent males were considered separately, achieving a significantly better classification accuracy when considering only the adult males - similar to what was concluded in (191).

The same analysis using only the SVM studies showed no significant effect on sensitivity or specificity. It may be that the effect previously observed is due to other factors beyond the age of the subjects but it is worth mentioning that only 17 studies (37 samples) were included in the analysis considering only the SVM classification tools.

We would like to perform tests considering other age ranges - such as segregate the data using the Piaget (205) or Freud (206) stages of development - to better analyse the effect of subject's age on the classification accuracy, specially in the childhood. However, only few articles used samples with such a low age. Furthermore, analysing the ABIDE - the most used database - we found a low proportion of child subjects: the ABIDE I do not include individuals below 7 years old whereas ABIDE II do not include individuals

below 5 years old; in addition, from the subjects that were available in both databases, less than 20% were bellow 10 years old (about 100 subjects in each one).

The lack of younger individuals in those studies raises some questions. Those classifiers may be, in fact, detecting the consequences in terms of brain circuitry alterations of living with ASD instead of identifying the true roots of the disorder (32, 207).

The only article in this review that used a sample of subjects below 24 months of age was (56). Their sample was composed of prospective neuroimaging of 59 6-month-old infants with a high familial risk for developing ASD. Each children received a research clinical best-estimate diagnosis of ASD at 24 months of age - 11 were diagnosed with ASD and 48 did not have ASD. The functional connectivity features were chosen within each training sample as showing a brain-behavior correlation with 24-month scores on measures of social behavior, language, motor development, and repetitive behavior. Using a linear SVM and a nested leave-one-out cross-validation, they obtained a sensitivity of 81.8% and a specificity of 100%, requiring only information from the left-out children 6-month-old rs-fMRI scan for the classification.

As we could see, there is a lack of information regarding the characteristics of the subjects and samples included in many of the studies. Aggravating the problem, as highlighted in (125), variables such as IQ, symptom severity, and handedness are missing for some sites of ABIDE. Therefore, we highly recommend for new studies in this field to present such information, when available, in order to enable new and more robust analyses in the future.

4.4.3 Sources of the samples

As we saw, the majority of the studies used versions of the ABIDE - 79 out of the 93 publications selected for the systematic review and 121 out of the 132 independent samples extracted for the meta-analysis.

As a heterogeneous and complex disorder, any ASD cohort is likely composed of ill-understood subtypes with different brain features. The use of large samples, such as provided by the ABIDE, can be helpful to address those issues (125).

Studies based on smaller datasets from a single site are composed of more homogeneous participants, reducing the generalizability of those models. Therefore, large multi-site datasets are needed to include a greater diversity of participants and to obtain diagnostic systems more reliable, robust and that generalize better to new data, revealing common features that contribute to classification (202, 189, 208, 209). Taking it all in consideration, it is not a surprise that the ABIDE databases have been so widely used.

We conducted analyses comparing the different versions of ABIDE used throughout the studies. At first, we found a significant effect on the sensitivity, indicating higher

sensitivity in studies that used the ABIDE (without specifying which version) compared to the ones that used ABIDE I preprocessed or ABIDE I + ABIDE II. However, in the sensitivity analysis or considering only the SVM studies we could not find any effect on sensitivity or specificity. Thus, our analysis indicated no significant difference in the accuracies between the different versions of ABIDE.

It is noteworthy that the ABIDE group is in fact composed of samples from the ABIDE I, ABIDE I preprocessed or ABIDE II, but the studies in this group did not specify which version they were using. Moreover, we would like to compare samples from databases or sources other than ABIDE in this analysis but there were not enough articles in each group.

We performed another analysis comparing the studies that used any version of ABIDE with studies that used databases or samples other than ABIDE - obtained from the NDAR, the UMCD or proprietary samples. The regression indicated higher sensitivity and specificity in the studies that did not use the ABIDE.

There was a great imbalance between the groups in this analysis (121 samples using ABIDE and 9 samples using other sources) but the result obtained is in line with what would be expected considering that the ABIDE aggregates data from many more individuals and sites than the other sources. The greater diversity of participants makes it difficult to obtain high accuracies whereas the use of a single dataset leads to overestimated results. Even using the ABIDE dataset it is possible to achieve higher classification accuracies by selecting smaller subsets of the data (202, 33).

It is also important to highlight the imbalance that exist within those samples and databases. As stated in (17, 15), most research on ASD comes from high-income countries - specially from North America, Europe and Japan - but the majority of individuals with autism live in low- and middle-income countries.

We face the same problem here. All the ABIDE, NDAR and UMCD participant data were collected at institutions from North America and Europe. Moreover, the proprietary samples from the studies selected for the meta-analysis were all from those same regions with exception of the study conducted in (193) that used a multi-site dataset from Japan. Thus, from the 132 independent samples used in the meta-analysis only 1 was not from North America or Europe and all of them were from high-income countries.

It is of utmost importance that more diverse datasets and studies be created and conducted. The applicability of the results obtained so far needs to be tested and confirmed across different cultures and social classes. Also, larger and more diverse samples would allow studies using restricted samples (such as low-motion data or samples composed entirely of female subjects) to obtain more reliable and robust results by selecting a bigger number of participants.

4.4.4 Features definition

Even though the focus of our research was the ML diagnostic tools that used rs-fMRI for classification, some of the selected studies used other types of data together with rs-fMRI. Their aim was to obtain better results by complementing the information available. In general, we found two types of complementary data: phenotypic information such as age and sex; other brain imaging data, specially sMRI.

As there were not enough articles in each of these subgroups, we compared the studies that used only rs-fMRI data with the ones that used any other data type together with rs-fMRI. Initially, our results indicated a higher sensitivity and specificity in studies that used other types of data to complement their features. In the sensitivity analysis we could not find a significant effect for the specificities, but the conclusion for the sensitivities remained the same.

Different types of brain images provide different views of the same brain and may reveal hidden evidences of ASD that are not available by using a single imaging modality (89). As an example, the best result obtained in (156) came from the combination of rs-fMRI with gray and white matter features. Also, in (202), the use of personal characteristic data together with structural or functional images resulted in better classification accuracy and the models required fewer features compared to the ones based only on structural or functional images.

We must highlight though that investigation of the effect of combining different types of data in the classification is not the main objective of this study.

Beyond the definition of the type of data to be used for classification, most of the studies defined ROIs to reduce the dimensionality of their features. Those ROIs can be selected from a priori atlases or estimated from the data being analyzed (26).

From the 132 samples used in the meta-analysis, 101 used a priori atlases or combinations of them. Also, as stated in (130), atlases are often selected arbitrarily in the rs-fMRI community. Therefore, we conducted a regression analysis with the atlas used as moderator.

Only three atlases were used in enough articles to be included in the analysis. They were the AAL in versions of 90 and 116 ROIs (210) and the CC atlas with 200 ROIs (70). Both versions of the AAL obtained significant better results in comparison to the CC200 - specially the AAL116 - but there were no significant difference within them. The sensitivity analysis, however, showed a significant effect on sensitivity between studies using AAL 116 or 90, indicating better results when using the AAL116. We must also highlight that in both cases the p-values of the comparison between the versions of the AAL were close to the threshold of 0.05 and those results should be taken with caution.

Still following that line, we made a regression analysis considering the number of

ROIs used throughout the studies. Our results indicated smaller accuracies as the number of regions used increased - more specifically, worse sensitivities for the SVM studies and worse specificities for the ANN studies.

The use of atlases with more ROIs generally results in more features available to be used by the classifier. However, using a large number of features relative to the number of data samples can cause classifiers to overfit by adapting to peculiarities of the dataset (2). That may be the reason why our analysis indicated better results in studies that used atlases with less ROIs.

In (26), many pipeline options for extracting predictive biomarkers were evaluated. They were composed of four steps: region estimation, time-series extraction, matrix estimation, and classification. Their results indicated that the choice of atlas had the greatest impact on prediction accuracy and that MSDL (211) - a data-driven approach - or using a priori atlases led to maximal performance. Also, their analysis of the effect of the number of ROIs found the best results between 40 and 100 ROIs.

From the defined ROIs, a subject's functional connectivity (FC) pattern can be estimated by calculating the correlations between all pairs of averaged time-series (2). As stated in (212), resting-state FC has almost exclusively been estimated using Pearson Correlation (PC) and we found it to be the most used feature throughout the selected studies, both on its original version and its Fisher-transformed one - 28 and 20 samples, respectively.

However, the PC faces some limitations - such as not taking into account the temporal structure of the rs-fMRI signal - and other methods that address those limitations might improve FC estimates (212).

The main focus of part of the articles was exactly the definition and extraction of the features to be used by the classifier. Those features presented great variety, coming from different points of view and in different levels of complexity.

For example, in (175, 139), dynamic functional connectivity (DFC) was proposed to capture the time-varying information of the brain FC by computing sliding window correlations from the PC Fisher-transformed coefficients, whereas (133, 167) used the sliding window technique based on different methods to calculate the FC.

Graph theory was used in (76, 198, 144) to extract features from the rs-fMRI data. On the other hand, in (156, 176), the mean time-series of each ROI from the rs-fMRI data were directly used as features together with deep learning techniques for classification.

The variety of choices in data processing adds to the variability of the results obtained in the studies of the field (26, 213). Therefore, we conducted a regression analysis with the type of feature used as moderator. It was not possible to include a lot of subgroups due to the limited number of articles that used each of those features. So, we compared

the results obtained using the PC and PC Fisher-transformed against each other and against the other types of features.

Our results indicated a significant advantage on both sensitivity and specificity to the studies using the Fisher-transformed version of the PC against the studies using it without modifications. Similarly, the studies using other features showed the same advantage against the PC ones. Finally, even though the studies using the PC Fisher-transformed obtained better summary estimates for sensitivity and specificity, their comparison against the studies using other features did not indicate any significant effect - what is not a surprise taking in consideration the great variety in the latter group.

Based on those results, we recommend for future studies in this field that intend to use the PC as a feature to consider using it in the Fisher-transformed version.

4.4.5 QUADAS-2 analyses

Bias and variation are often present in diagnostic test accuracy studies. Therefore, they need to be detected and assessed in the studies included in a meta-analysis to understand the validity of the results obtained (7, 214).

In the results obtained through the QUADAS-2 (see Figure 15) we found a great number of studies with high RoB on the patient selection domain, basically due to the selection of subjects in restricted intervals of age and IQ or the exclusion of female subjects. The effect of those variables on the classification results was already discussed in Section 4.4.2.

Beyond that, we can see that many studies were assessed to have an unclear RoB, specially on the patient selection and flow and timing domains. This reinforces the necessity to present more detailed information regarding the characteristics of the subjects and samples included in the publications in this field.

We conducted some analysis using the QUADAS-2 results to better understand them. The first one considered the number of domains with low RoB in each study. We could only include two categories in this analysis - one or two domains with low RoB - and the results indicated higher specificity in studies with only one low RoB domain.

This is not surprising since more domains with a low RoB indicates more reliable results whereas domains with an unclear or high RoB may indicate a bias on the results - in this case overestimating the specificity.

Taking in consideration that the index test domain was the only domain to present enough articles in each of the categories, we conducted a regression analysis with its results as moderator.

As expected, studies with high or unclear RoB obtained significantly better results

than studies with low RoB. We also found better results for the studies with unclear RoB in comparison to those with high RoB. We can suppose that a great part of those studies assessed as unclear should have been assessed as high, but there was not enough information to conclude that.

This also indicates a clear bias of overestimation - at least for the specificities - on the studies that do not apply their best models to independent sets after testing different numbers of features or atlases.

For the last analysis, we separated the low RoB category into the articles that performed a temporal or geographic validation and the ones that performed a split-sample validation (117). Even though the latter obtained better summary estimates than the former, we did not find a significant effect between their sensitivities nor specificities.

From the clinical standpoint, complete external validation (temporal or geographic) is preferred. Split sample validation would not accurately assess the generalizability of a model whereas geographic validation is helpful for this purpose since it may be performed with different technical parameters at different sites (117).

Our analysis did not indicate a significant difference between the results using each of these types of validation. However, we must highlight that there were only 5 articles (35 samples) with low RoB by the index test domain that performed a complete external validation and it is still the more reliable approach to assess generalizability.

4.4.6 Clinical validity

At present, ASD diagnosis is based on behavioral criteria, being vulnerable to subjectivity and interpretative bias. Also, less experienced clinicians seem to have more problems with the challenges of this complex diagnostic process (22).

In (55), the diagnostic utility and discriminative ability of the ADOS-G and the ADI-R were assessed using a clinical population of children. The results indicated approximately 75% of agreement with the qualified multidisciplinary team diagnoses and most inconsistencies were false-positives.

The accuracy and validity of the ADOS-2 and ADI-R in diagnosing ASD in adults without intellectual disability were evaluated in (54). The original algorithm of ADOS-2 Module 4 obtained 85.9% and 82.9% whereas its revised algorithm obtained 87.2% and 74.3% of sensitivity and specificity, respectively. On the other hand, the ADI-R obtained 43.1% of sensitivity and 94.7% of specificity. The authors also highlight the importance of training and experience while assessing ASD diagnosis in adults.

Considering all the studies selected for the meta-analysis, we found summary sensitivity and specificity of 73.8% and 74.8%, respectively, for the ASD diagnosis using

rs-fMRI and ML classifiers. Also, the AUC/pAUC of 0.803/0.765 indicates values between acceptable and excellent according to the classification proposed in (215) (0.5: no discrimination; 0.7–0.79: acceptable; 0.8–0.89: excellent; ≥0.9 outstanding). If we look at the analysis considering only the SVM studies, those results were even better, with sensitivity and specificity above 76% and AUC of 0.832.

Even within the articles that presented lower RoB and, therefore, more reliable results, the AUC values were acceptable. It was the case of the studies with two domains with low RoB (AUC of 0.762) or with low RoB on the index test domain that performed a complete external validation (AUC of 0.744). Those results are promising, but we must highlight some questions.

Also, as commented before, the articles included in our analysis presented great variety of features extracted and selected, classifiers used, and even the validation approach applied. Thus, the summary estimates that we obtained show the overall potential of those procedures, but do not indicate a specific one that could be applied in clinical practice. It would even be possible to use different classifiers for subjects with different characteristics - such as sex and age - similarly to what happens with the different modules of ADOS-2.

Taking the variety of neurodevelopmental etiologies that are believed to exist within the ASD population, there may not be an exceptional biomarker to diagnose the disorder (13, 14). Perhaps the classifiers must consider different biomarkers for different etiologies, partitioning the ASD into more than a single class (216).

There are many questions that need to be assessed to define the clinical validity of those procedures. It includes the underrepresentation of females in research and clinical practice, the effects of subject's IQ and age, the lack of such information in many studies, and the necessity of larger and more diverse samples to confirm the generalizability of the classification tools.

Moreover, verification of the performance of the models in an epidemiologically well-defined clinical cohort that adequately represents the target population is of paramount importance for clinical verification of a diagnostic or predictive machine learning model (117).

4.4.7 Limitations

There are several limitations in our study. Considering the great heterogeneity within the selected publications, the summary estimates obtained through the meta-analysis have to be interpreted with caution and in light of the methodologic quality of them. From the 55 articles used in the quantitative analysis, none presented a low RoB by the patient selection domain whereas only 4 of them were assessed to have a low RoB

by the flow and timing domain. Besides, more than half of the studies were considered to have an unclear or high RoB by the index test domain.

Most studies provided only limited information regarding the patients samples and their clinical characteristics. However, detailed information about the participants' disease status, symptoms, current medication, history of interventions or comorbidities is crucial for evaluating the potential of the proposed models to be applied in clinical practice (217, 125). Thus, the impact of those variables on classification accuracy needs to be better investigated.

The studies included in our analysis identified ASD-distinctive brain patterns as compared to healthy volunteers. Nevertheless, it is critical to investigate the patterns of brain abnormalities that differentiate between different psychiatric disorders. Also, the results obtained in this meta-analysis do not apply to individuals below 5 years of age since almost none of the studies included individuals with such low age.

Another problem to be kept in mind is the sample overlap between the studies, especially considering the lack of information on the patient selection process and the large number of studies that used the ABIDE database. Thus, it is not clear to which extent this overlap could bias the results obtained.

In addition, there are some methodological steps that were not investigated in our analyses such as the data preprocessing and feature selection procedures. Those aspects still need to be assessed to define their effects on classification accuracy.

5 Conclusion

Considering the increasing number of publications in this field, systematic reviews and meta-analyses such as this one are a necessity to better understand the results obtained and the limitations so far, indicating promising pathways and questions that still need to be addressed - specially considering that this is the first meta-analysis focused on ASD classification using rs-fMRI and ML techniques, to the best of our knowledge.

Our results indicated a significant better accuracy for SVM classifiers in comparison to the ANN ones. However, the use of ANN techniques - specially deep-learning models - is increasing and our analysis suggests that they may have an advantage when dealing with larger samples in comparison to SVM techniques.

The use of other types of data to complement rs-fMRI information seem to be promising, achieving specially higher sensitivities when compared to rs-fMRI data alone. Yet, other analyses focused on this topic should be conducted.

Lower values of sensitivity/specificity were found when the number of ROIs increased. Also, the performance of the approaches using the AAL116 to define their ROIs stood out in comparison to the ones that used the AAL90 or CC200.

Regarding the features used to train the classifiers, we found better results using the PC Fisher-transformed or other features in comparison to the use of the PC without modifications.

The overall sensitivity and specificity estimates were approximately 74% and above 76% when considering only SVM classifiers - with excellent AUC values in both cases. Even within the articles that presented lower RoB and, therefore, more reliable results, the AUC values were acceptable.

However, given the many limitations indicated in our study and the poor methodological quality found in a great part of the selected articles - as indicated by the QUADAS-2 assessment - further well-designed studies are warranted to extend the potential use of those classification algorithms to clinical settings and the results presented here should be taken with caution.

It is important to highlight that all of the selected studies were from high income countries and there was a lack of information in many of them - specially regarding the characteristics of the subjects and samples. Therefore, we highly encourage more diverse datasets and studies to be created and conducted, presenting more complete information to enable more robust analysis in the future.

Finally, it is not clear to what extent those classification techniques could be used

for early diagnosis. So far, the promising results obtained are referent to the diagnosis of older children, adolescents, and adults.

Appendix

APPENDIX A – Main results from the meta-analysis

			Summary Estimate with 95% CI	p-value(s)
Overall		Sensitivity	0.738 (0.718 - 0.758)	
		Specificity	0.748 (0.723 - 0.771)	
Type of ML technique	SVM	Sensitivity	0.763 (0.732 - 0.792)	
		Specificity	0.775 (0.737 - 0.808)	
	ANN	Sensitivity	0.684 (0.650 - 0.715)	p = 0.002
		Specificity	0.702 (0.662 - 0.739)	p = 0.008
Sub-type (SVM)	L-SVM	Sensitivity	0.739 (0.702 - 0.772)	
		Specificity	0.775 (0.733 - 0.812)	
	SVM - other	Sensitivity	0.799 (0.737 - 0.850)	p = 0.179
		Specificity	0.767 (0.688 - 0.830)	p = 0.619
Sub-type (ANN)	CNN	Sensitivity	0.667 (0.633 - 0.699)	
		Specificity	0.701 (0.663 - 0.737)	
	ANN - other	Sensitivity	0.703 (0.658 - 0.744)	p = 0.107
		Specificity	0.712 (0.655 - 0.764)	p = 0.726
L-SVM x CNN	L-SVM	Sensitivity	0.739 (0.702 - 0.772)	
		Specificity	0.775 (0.733 - 0.812)	
	CNN	Sensitivity	0.667 (0.633 - 0.699)	p = 0.009
		Specificity	0.701 (0.663 - 0.737)	p = 0.024
Sample size	Overall	Sensitivity		p = 0.004
		Specificity		p < 0.001
	SVM subgroup	Sensitivity		p = 0.152
		Specificity		p = 0.001
	ANN subgroup	Sensitivity		p = 0.414
		Specificity		p = 0.124
Year of publication	Overall	Sensitivity		p = 0.250
		Specificity		p = 0.283
	SVM subgroup	Sensitivity		p = 0.913
		Specificity		p = 0.537
	ANN subgroup	Sensitivity		p = 0.062
		Specificity		p = 0.242

Continues on next page

			Summary Estimate with 95% CI	**p-value(s)**
Sex	Males and females	Sensitivity	0.745 (0.721 - 0.769)	
		Specificity	0.731 (0.701 - 0.760)	
	Only males	Sensitivity	0.688 (0.618 - 0.751)	p = 0.176
		Specificity	0.724 (0.660 - 0.781)	p = 0.934
Mean FIQ	ASD	Sensitivity		p = 0.456
		Specificity		p = 0.993
	TD	Sensitivity		p = 0.654
		Specificity		p = 0.567
FIQ	High- and low-functioning	Sensitivity	0.728 (0.684 - 0.768)	
		Specificity	0.694 (0.635 - 0.747)	
	Only high-functioning	Sensitivity	0.697 (0.642 - 0.746)	p = 0.544
		Specificity	0.724 (0.650 - 0.788)	p = 0.501
Mean age	ASD	Sensitivity		p = 0.386
		Specificity		p = 0.352
	TD	Sensitivity		p = 0.333
		Specificity		p = 0.196
Age	Under 18 y.o.	Sensitivity	0.767 (0.718 - 0.810)	
		Specificity	0.776 (0.730 - 0.816)	
	Under and above 18 y.o.	Sensitivity	0.723 (0.695 - 0.750)	p = 0.225
		Specificity	0.705 (0.666 - 0.741)	p = 0.020
Age (SVM)	Under 18 y.o.	Sensitivity	0.736 (0.663 - 0.798)	
		Specificity	0.762 (0.692 - 0.820)	
	Under and above 18 y.o.	Sensitivity	0.733 (0.689 - 0.773)	p = 0.790
		Specificity	0.726 (0.658 - 0.785)	p = 0.427
Database	ABIDE	Sensitivity	0.771 (0.732 - 0.806)	
		Specificity	0.772 (0.720 - 0.817)	
	ABIDE I preprocessed	Sensitivity	0.720 (0.690 - 0.749)	p = 0.046*
		Specificity	0.721 (0.685 - 0.755)	p = 0.130*
	ABIDE I + ABIDE II	Sensitivity	0.692 (0.658 - 0.724)	p = 0.043*; p = 0.631**
		Specificity	0.733 (0.684 - 0.777)	p = 0.447*; p = 0.630**

Continues on next page

			Summary Estimate with 95% CI	**p-value(s)**
Database (only SVM)	ABIDE	Sensitivity	0.752 (0.687 - 0.807)	
		Specificity	0.763 (0.663 - 0.840)	
	ABIDE I preprocessed	Sensitivity	0.760 (0.707 - 0.805)	p = 0.756
		Specificity	0.764 (0.706 - 0.814)	p = 0.731
ABIDE (any) x others	ABIDE (any)	Sensitivity	0.732 (0.711 - 0.752)	
		Specificity	0.741 (0.716 - 0.765)	
	Others	Sensitivity	0.818 (0.734 - 0.881)	p = 0.024
		Specificity	0.830 (0.725 - 0.900)	p = 0.045
Data type	Only rs-fMRI	Sensitivity	0.728 (0.706 - 0.748)	
		Specificity	0.739 (0.713 - 0.764)	
	rs-fMRI + other data types	Sensitivity	0.847 (0.785 - 0.894)	p = 0.002
		Specificity	0.810 (0.741 - 0.863)	p = 0.047
Atlas	AAL116	Sensitivity	0.777 (0.737 - 0.812)	
		Specificity	0.782 (0.728 - 0.829)	
	AAL90	Sensitivity	0.692 (0.614 - 0.760)	p = 0.054’
		Specificity	0.749 (0.687 - 0.801)	p = 0.397’
	CC200	Sensitivity	0.680 (0.654 - 0.704)	p = 0.001’; p = 0.560”
		Specificity	0.644 (0.607 - 0.679)	p < 0.001’; p = 0.001”
Number of ROIs	Overall	Sensitivity		p = 0.043
		Specificity		p = 0.018
	SVM subgroup	Sensitivity		p = 0.029
		Specificity		p = 0.089
	ANN subgroup	Sensitivity		p = 0.557
		Specificity		p = 0.016
Feature	PC	Sensitivity	0.689 (0.668 - 0.709)	
		Specificity	0.683 (0.643 - 0.721)	
	PC(z)	Sensitivity	0.767 (0.713 - 0.813)	$p = 0.001^x$
		Specificity	0.810 (0.756 - 0.854)	$p < 0.001^x$
	Other	Sensitivity	0.735 (0.706 - 0.762)	$p = 0.031^x$; $p = 0.173^y$
		Specificity	0.747 (0.710 - 0.780)	$p = 0.024^x$; $p = 0.072^y$

Continues on next page

			Summary Estimate with 95% CI	**p-value(s)**
Low RoB	One	Sensitivity	0.751 (0.720 - 0.780)	
		Specificity	0.784 (0.755 - 0.811)	
	Two	Sensitivity	0.719 (0.694 - 0.744)	p = 0.236
		Specificity	0.696 (0.659 - 0.730)	p < 0.001
RoB of the index test	Low	Sensitivity	0.721 (0.696 - 0.744)	
		Specificity	0.699 (0.662 - 0.733)	
	Unclear	Sensitivity	0.805 (0.761 - 0.842)	p = 0.003[a]
		Specificity	0.813 (0.769 - 0.850)	p < 0.001[a]
	High	Sensitivity	0.701 (0.663 - 0.735)	p = 0.373[a]; p = 0.001[b]
		Specificity	0.766 (0.732 - 0.798)	p = 0.011[a]; p = 0.127[b]
RoB of the index test (split)	Low (temporal / geographic)	Sensitivity	0.704 (0.673 - 0.733)	
		Specificity	0.679 (0.642 - 0.713)	
	Low (split-sample)	Sensitivity	0.730 (0.691 - 0.765)	p = 0.451[1]
		Specificity	0.721 (0.651 - 0.782)	p = 0.441[1]
	Unclear	Sensitivity	0.805 (0.761 - 0.842)	p = 0.005[1]; p = 0.024[2]
		Specificity	0.813 (0.769 - 0.850)	p < 0.001[1]; p = 0.010[2]
	High	Sensitivity	0.701 (0.663 - 0.735)	p = 0.715[1]; p = 0.291[2]; p = 0.001[3]
		Specificity	0.766 (0.732 - 0.798)	p = 0.001[1]; p = 0.148[2]; p = 0.127[3]
				End of Table

* compared to ABIDE; ** compared to ABIDE I preprocessed; ’ compared to AAL116; ” compared to AAL90; [x] compared to PC; [y] compared to PC(z); [a] compared to low; [b] compared to unclear; [1] compared to low (temporal/geographic); [2] compared to low (split-sample); [3] compared to unclear.

APPENDIX B – Results from the sensitivity analysis for the adulthood threshold

			Summary Estimate with 95% CI	**p-value(s)**
Age (18)	Under 18 y.o.	Sensitivity	0.767 (0.718 - 0.810)	
		Specificity	0.776 (0.730 - 0.816)	
	Under and above 18 y.o.	Sensitivity	0.723 (0.695 - 0.750)	p = 0.225
		Specificity	0.705 (0.666 - 0.741)	p = 0.020
Age (19)	Under 19 y.o.	Sensitivity	0.776 (0.728 - 0.817)	
		Specificity	0.781 (0.738 - 0.818)	
	Under and above 19 y.o.	Sensitivity	0.723 (0.695 - 0.750)	p = 0.102
		Specificity	0.705 (0.666 - 0.741)	p = 0.009
Age (20)	Under 20 y.o.	Sensitivity	0.768 (0.726 - 0.806)	
		Specificity	0.787 (0.751 - 0.819)	
	Under and above 20 y.o.	Sensitivity	0.723 (0.695 - 0.750)	p = 0.103
		Specificity	0.705 (0.666 - 0.741)	p = 0.002
Age (21)	Under 21 y.o.	Sensitivity	0.768 (0.726 - 0.806)	
		Specificity	0.787 (0.751 - 0.819)	
	Under and above 21 y.o.	Sensitivity	0.723 (0.695 - 0.750)	p = 0.103
		Specificity	0.705 (0.666 - 0.741)	p = 0.002
Age (18 - SVM)	Under 18 y.o.	Sensitivity	0.736 (0.663 - 0.798)	
		Specificity	0.762 (0.692 - 0.820)	
	Under and above 18 y.o.	Sensitivity	0.733 (0.689 - 0.773)	p = 0.790
		Specificity	0.726 (0.658 - 0.785)	p = 0.427
Age (19 - SVM)	Under 19 y.o.	Sensitivity	0.758 (0.688 - 0.817)	
		Specificity	0.776 (0.715 - 0.827)	
	Under and above 19 y.o.	Sensitivity	0.733 (0.689 - 0.773)	p = 0.811
		Specificity	0.726 (0.658 - 0.785)	p = 0.250
Age (20 - SVM)	Under 20 y.o.	Sensitivity	0.763 (0.702 - 0.815)	
		Specificity	0.778 (0.727 - 0.822)	
	Under and above 20 y.o.	Sensitivity	0.733 (0.689 - 0.773)	p = 0.659
		Specificity	0.726 (0.658 - 0.785)	p = 0.186

Continues on next page

			Summary Estimate with 95% CI	**p-value(s)**
Age (21 - SVM)	Under 21 y.o.	Sensitivity	0.763 (0.702 - 0.815)	
		Specificity	0.778 (0.727 - 0.822)	
	Under and above 21 y.o.	Sensitivity	0.733 (0.689 - 0.773)	p = 0.659
		Specificity	0.726 (0.658 - 0.785)	p = 0.186
				End of Table

APPENDIX C – Results from the sensitivity analysis including three more articles

			Summary Estimate with 95% CI	p-value(s)
Overall		Sensitivity	0.737 (0.717 - 0.756)	
		Specificity	0.750 (0.727 - 0.772)	
Type of ML technique	SVM	Sensitivity	0.766 (0.734 - 0.794)	
		Specificity	0.776 (0.739 - 0.809)	
	ANN	Sensitivity	0.684 (0.650 - 0.715)	p = 0.001
		Specificity	0.702 (0.662 - 0.739)	p = 0.006
Sub-type (SVM)	L-SVM	Sensitivity	0.739 (0.702 - 0.772)	
		Specificity	0.775 (0.733 - 0.812)	
	SVM - other	Sensitivity	0.802 (0.743 - 0.850)	p = 0.134
		Specificity	0.769 (0.696 - 0.830)	p = 0.678
Sample size	Overall	Sensitivity		p = 0.005
		Specificity		p < 0.001
	SVM subgroup	Sensitivity		p = 0.125
		Specificity		p = 0.001
Year of publication	Overall	Sensitivity		p = 0.193
		Specificity		p = 0.288
	SVM subgroup	Sensitivity		p = 0.740
		Specificity		p = 0.613
Sex	Males and females	Sensitivity	0.741 (0.717 - 0.764)	
		Specificity	0.735 (0.705 - 0.762)	
	Only males	Sensitivity	0.688 (0.618 - 0.751)	p = 0.211
		Specificity	0.724 (0.660 - 0.781)	p = 0.881
FIQ	High- and low-functioning	Sensitivity	0.724 (0.681 - 0.764)	
		Specificity	0.696 (0.639 - 0.748)	
	Only high-functioning	Sensitivity	0.697 (0.642 - 0.746)	p = 0.618
		Specificity	0.724 (0.650 - 0.788)	p = 0.520
Mean age	ASD	Sensitivity		p = 0.464
		Specificity		p = 0.332
	TD	Sensitivity		p = 0.409
		Specificity		p = 0.181

Continues on next page

			Summary Estimate with 95% CI	**p-value(s)**
Age	Under 18 y.o.	Sensitivity	0.767 (0.718 - 0.810)	
		Specificity	0.776 (0.730 - 0.816)	
	Under and above 18 y.o.	Sensitivity	0.719 (0.691 - 0.744)	p = 0.161
		Specificity	0.712 (0.675 - 0.746)	p = 0.032
Database	ABIDE	Sensitivity	0.762 (0.724 - 0.796)	
		Specificity	0.776 (0.729 - 0.816)	
	ABIDE I preprocessed	Sensitivity	0.720 (0.690 - 0.749)	p = 0.097*
		Specificity	0.721 (0.685 - 0.755)	p = 0.072*
	ABIDE I + ABIDE II	Sensitivity	0.692 (0.658 - 0.724)	p = 0.087*
		Specificity	0.733 (0.684 - 0.777)	p = 0.345*
Database (only SVM)	ABIDE	Sensitivity	0.761 (0.698 - 0.815)	
		Specificity	0.767 (0.676 - 0.838)	
	ABIDE I preprocessed	Sensitivity	0.760 (0.707 - 0.805)	p = 0.920
		Specificity	0.764 (0.706 - 0.814)	p = 0.815
ABIDE (any) x others	ABIDE (any)	Sensitivity	0.730 (0.710 - 0.750)	
		Specificity	0.744 (0.720 - 0.767)	
	Others	Sensitivity	0.818 (0.734 - 0.881)	p = 0.022
		Specificity	0.830 (0.725 - 0.900)	p = 0.049
Data type	Only rs-fMRI	Sensitivity	0.726 (0.705 - 0.746)	
		Specificity	0.742 (0.717 - 0.766)	
	rs-fMRI + other data types	Sensitivity	0.847 (0.785 - 0.894)	p = 0.002
		Specificity	0.810 (0.741 - 0.863)	p = 0.057
Atlas	AAL116	Sensitivity	0.780 (0.741 - 0.814)	
		Specificity	0.784 (0.731 - 0.829)	
	AAL90	Sensitivity	0.692 (0.614 - 0.760)	p = 0.045’
		Specificity	0.749 (0.687 - 0.801)	p = 0.375’
	CC200	Sensitivity	0.680 (0.654 - 0.704)	p < 0.001’
		Specificity	0.644 (0.607 - 0.679)	p < 0.001’
Number of ROIs	Overall	Sensitivity		p = 0.027
		Specificity		p = 0.019
	SVM subgroup	Sensitivity		p = 0.028
		Specificity		p = 0.085

Continues on next page

			Summary Estimate with 95% CI	**p-value(s)**
Feature	PC	Sensitivity	0.689 (0.668 - 0.709)	
		Specificity	0.683 (0.643 - 0.721)	
	PC(z)	Sensitivity	0.757 (0.703 - 0.805)	$p = 0.004^{x}$
		Specificity	0.808 (0.756 - 0.850)	$p < 0.001^{x}$
	Other	Sensitivity	0.734 (0.706 - 0.760)	$p = 0.032^{x}$; $p = 0.298^{y}$
		Specificity	0.751 (0.717 - 0.783)	$p = 0.014^{x}$; $p = 0.087^{y}$
Low RoB	One	Sensitivity	0.748 (0.718 - 0.776)	
		Specificity	0.786 (0.758 - 0.811)	
	Two	Sensitivity	0.719 (0.694 - 0.744)	$p = 0.306$
		Specificity	0.696 (0.659 - 0.730)	$p < 0.001$
RoB of the index test	Low	Sensitivity	0.721 (0.696 - 0.744)	
		Specificity	0.699 (0.662 - 0.733)	
	Unclear	Sensitivity	0.799 (0.757 - 0.835)	$p = 0.005^{a}$
		Specificity	0.812 (0.772 - 0.847)	$p < 0.001^{a}$
	High	Sensitivity	0.697 (0.659 - 0.731)	$p = 0.278^{a}$; $p = 0.001^{b}$
		Specificity	0.767 (0.734 - 0.797)	$p = 0.009^{a}$; $p = 0.107^{b}$
RoB of the index test (split)	Low (temporal / geographic)	Sensitivity	0.704 (0.673 - 0.733)	
		Specificity	0.679 (0.642 - 0.713)	
	Low (split-sample)	Sensitivity	0.730 (0.691 - 0.765)	
		Specificity	0.721 (0.651 - 0.782)	
	Unclear	Sensitivity	0.799 (0.757 - 0.835)	$p = 0.008^{1}$; $p = 0.036^{2}$
		Specificity	0.812 (0.772 - 0.847)	$p < 0.001^{1}$; $p = 0.007^{2}$
	High	Sensitivity	0.697 (0.659 - 0.731)	$p = 0.591^{1}$; $p = 0.227^{2}$; $p = 0.001^{3}$
		Specificity	0.767 (0.734 - 0.797)	$p = 0.001^{1}$; $p = 0.135^{2}$; $p = 0.107^{3}$
				End of Table

* compared to ABIDE; ’ compared to AAL116; ^{x} compared to PC; ^{y} compared to PC(z); ^{a} compared to low; ^{b} compared to unclear; ^{1} compared to low (temporal/geographic); ^{2} compared to low (split-sample); ^{3} compared to unclear.

Bibliography

1 SIERO, J. C.; BHOGAL, A.; JANSMA, J. M. Blood oxygenation level–dependent/functional magnetic resonance imaging: underpinnings, practice, and perspectives. *PET clinics*, Elsevier, v. 8, n. 3, p. 329–344, 2013. 7, 23, 24

2 KASSRAIAN-FARD, P. et al. Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Frontiers in psychiatry*, Frontiers, v. 7, p. 177, 2016. 7, 16, 17, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35, 36, 67, 87

3 PEREIRA, F.; MITCHELL, T.; BOTVINICK, M. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, Elsevier, v. 45, n. 1, p. S199–S209, 2009. 7, 28, 29, 30, 31, 34, 35, 82

4 RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. [S.l.]: Malaysia; Pearson Education Limited,, 2016. 7, 31, 32, 33

5 MACASKILL, P. et al. Chapter 10: analysing and presenting results. *Cochrane handbook for systematic reviews of diagnostic test accuracy version*, v. 1, n. 0, 2010. 7, 36, 40, 41, 42, 43, 44, 45, 46, 47, 49, 50, 51, 52, 53, 54, 55, 65

6 ZU, C. et al. Identifying high order brain connectome biomarkers via learning on hypergraph. In: SPRINGER. *International Workshop on Machine Learning in Medical Imaging*. [S.l.], 2016. p. 1–9. 7, 45, 67, 68, 80

7 KIM, K. W. et al. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part i. general guidance and tips. *Korean journal of radiology*, v. 16, n. 6, p. 1175–1187, 2015. 7, 17, 36, 38, 39, 40, 45, 46, 53, 61, 88

8 RAPIN, I.; TUCHMAN, R. F. Autism: definition, neurobiology, screening, diagnosis. *Pediatric Clinics of North America*, Elsevier, v. 55, n. 5, p. 1129–1146, 2008. 16, 18

9 HAHLER, E.-M.; ELSABBAGH, M. Autism: A global perspective. *Current Developmental Disorders Reports*, Springer, v. 2, n. 1, p. 58–64, 2015. 16, 18, 19, 20, 22

10 WANG, C. et al. Prenatal, perinatal, and postnatal factors associated with autism: a meta-analysis. *Medicine*, Wolters Kluwer Health, v. 96, n. 18, 2017. 16, 18, 20

11 HERTZ-PICCIOTTO, I. et al. The charge study: an epidemiologic investigation of genetic and environmental factors contributing to autism. *Environmental health perspectives*, National Institute of Environmental Health Sciences, v. 114, n. 7, p. 1119–1125, 2006. 16, 20

12 ASSOCIATION, A. P. et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. [S.l.]: American Psychiatric Pub, 2013. 16, 18, 19, 21, 58

13 GESCHWIND, D. H.; STATE, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *The Lancet Neurology*, Elsevier, v. 14, n. 11, p. 1109–1120, 2015. 16, 21, 90

14 JAHEDI, A. et al. Distributed intrinsic functional connectivity patterns predict diagnostic status in large autism cohort. *Brain connectivity*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 7, n. 8, p. 515–525, 2017. 16, 21, 34, 67, 68, 80, 90

15 ELSABBAGH, M. et al. Global prevalence of autism and other pervasive developmental disorders. *Autism research*, Wiley Online Library, v. 5, n. 3, p. 160–179, 2012. 16, 19, 85

16 HAYES, S. A.; WATSON, S. L. The impact of parenting stress: A meta-analysis of studies comparing the experience of parenting stress in parents of children with and without autism spectrum disorder. *Journal of autism and developmental disorders*, Springer, v. 43, n. 3, p. 629–642, 2013. 16

17 DURKIN, M. S. et al. Autism screening and diagnosis in low resource settings: challenges and opportunities to enhance research and services worldwide. *Autism Research*, Wiley Online Library, v. 8, n. 5, p. 473–476, 2015. 16, 85

18 WEBB, S. J. et al. The motivation for very early intervention for infants at high risk for autism spectrum disorders. *International journal of speech-language pathology*, Taylor & Francis, v. 16, n. 1, p. 36–42, 2014. 16, 22

19 ROGERS, S. J. et al. Autism treatment in the first year of life: a pilot study of infant start, a parent-implemented intervention for symptomatic infants. *Journal of autism and developmental disorders*, Springer, v. 44, n. 12, p. 2981–2995, 2014. 16, 22

20 RUTTER, M.; LECOUTEUR, A.; LORD, C. Autism diagnostic interview revised (adi-r). *Los Angeles: Western Psychological Services*, 2003. 16, 21

21 LORD, C. et al. Autism diagnostic observation schedule, (ados-2) modules 1-4. *Los Angeles, California: Western Psychological Services*, 2012. 16, 21

22 KAMP-BECKER, I. et al. Diagnostic accuracy of the ados and ados-2 in clinical practice. *European child & adolescent psychiatry*, Springer, v. 27, n. 9, p. 1193–1207, 2018. 16, 21, 89

23 FALKMER, T. et al. Diagnostic procedures in autism spectrum disorders: a systematic literature review. *European child & adolescent psychiatry*, Springer, v. 22, n. 6, p. 329–340, 2013. 16, 21

24 LITJENS, G. et al. A survey on deep learning in medical image analysis. *Medical image analysis*, Elsevier, v. 42, p. 60–88, 2017. 16

25 MATTHEWS, P. M.; JEZZARD, P. Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, BMJ Publishing Group Ltd, v. 75, n. 1, p. 6–12, 2004. 17, 22, 23

26 ABRAHAM, A. et al. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage*, Elsevier, v. 147, p. 736–745, 2017. 17, 67, 86, 87

27 SUNDERMANN, B.; BEVERBORG, M. Olde lütke; PFLEIDERER, B. Toward literature-based feature selection for diagnostic classification: a meta-analysis of resting-state fmri in depression. *Frontiers in human neuroscience*, Frontiers, v. 8, p. 692, 2014. 17

28 TAHMASIAN, M. et al. A systematic review on the applications of resting-state fmri in parkinson's disease: does dopamine replacement therapy play a role? *Cortex*, Elsevier, v. 73, p. 80–105, 2015. 17, 25, 26

29 MILHAM, M. P. et al. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, Frontiers, v. 6, p. 62, 2012. 17

30 HULL, J. V. et al. Resting-state functional connectivity in autism spectrum disorders: A review. *Frontiers in psychiatry*, Frontiers, v. 7, p. 205, 2017. 17, 26, 27

31 DOSENBACH, N. U. et al. Prediction of individual brain maturity using fmri. *Science*, American Association for the Advancement of Science, v. 329, n. 5997, p. 1358–1361, 2010. 17

32 PLITT, M.; BARNES, K. A.; MARTIN, A. Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage: Clinical*, Elsevier, v. 7, p. 359–366, 2015. 17, 67, 84

33 KATUWAL, G. J. et al. The predictive power of structural mri in autism diagnosis. In: IEEE. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. [S.l.], 2015. p. 4270–4273. 17, 85

34 BOSL, W. J.; TAGER-FLUSBERG, H.; NELSON, C. A. Eeg analytics for early detection of autism spectrum disorder: a data-driven approach. *Scientific reports*, Nature Publishing Group, v. 8, n. 1, p. 1–20, 2018. 17

35 SUGANYA, V.; GEETHA, A.; SUJATHA, S. Urine proteome analysis to evaluate protein biomarkers in children with autism. *Clinica Chimica Acta*, Elsevier, v. 450, p. 210–219, 2015. 17

36 VARGAS-CUENTAS, N. I. et al. Developing an eye-tracking algorithm as a potential tool for early diagnosis of autism spectrum disorder in children. *PloS one*, Public Library of Science San Francisco, CA USA, v. 12, n. 11, p. e0188826, 2017. 17

37 MOYAL, W. N.; LORD, C.; WALKUP, J. T. Quality of life in children and adolescents with autism spectrum disorders: what is known about the effects of pharmacotherapy? *Pediatric Drugs*, Springer, v. 16, n. 2, p. 123–128, 2014. 18

38 BAXTER, A. J. et al. The epidemiology and global burden of autism spectrum disorders. *Psychological medicine*, Cambridge University Press, v. 45, n. 3, p. 601, 2015. 18, 19, 20

39 ASSOCIATION, A. P. *Diagnostic and statistical manual of mental disorders. 4th edition. text revision: DSM IV–TR*. [S.l.]: American Psychiatric Pub, 2000. 18, 58

40 ORGANIZATION, W. H. et al. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*. [S.l.]: World Health Organization, 1993. v. 2. 18

41 VOLKMAR, F. R.; MCPARTLAND, J. C. From kanner to dsm-5: autism as an evolving diagnostic concept. *Annual review of clinical psychology*, Annual Reviews, v. 10, p. 193–212, 2014. 19

42 SANDIN, S. et al. The heritability of autism spectrum disorder. *Jama*, American Medical Association, v. 318, n. 12, p. 1182–1184, 2017. 20

43 TILLMANN, J. et al. Evaluating sex and age differences in adi-r and ados scores in a large european multi-site sample of individuals with autism spectrum disorder. *Journal of autism and developmental disorders*, Springer, v. 48, n. 7, p. 2490–2505, 2018. 20, 62

44 WIJNGAARDEN-CREMERS, P. J. V. et al. Gender and age differences in the core triad of impairments in autism spectrum disorders: a systematic review and meta-analysis. *Journal of autism and developmental disorders*, Springer, v. 44, n. 3, p. 627–635, 2014. 20, 62

45 MAYES, S. D.; CALHOUN, S. L. Impact of iq, age, ses, gender, and race on autistic symptoms. *Research in Autism Spectrum Disorders*, Elsevier, v. 5, n. 2, p. 749–757, 2011. 20, 62

46 LOOMES, R.; HULL, L.; MANDY, W. P. L. What is the male-to-female ratio in autism spectrum disorder? a systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, Elsevier, v. 56, n. 6, p. 466–474, 2017. 20

47 MAENNER, M. J. et al. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2016. *MMWR Surveillance Summaries*, Centers for Disease Control and Prevention, v. 69, n. 4, p. 1, 2020. 20

48 JENSEN, C. M.; STEINHAUSEN, H.-C.; LAURITSEN, M. B. Time trends over 16 years in incidence-rates of autism spectrum disorders across the lifespan based on nationwide danish register data. *Journal of autism and developmental disorders*, Springer, v. 44, n. 8, p. 1808–1818, 2014. 20

49 FOMBONNE, E. Epidemiology of pervasive developmental disorders. *Pediatric research*, Nature Publishing Group, v. 65, n. 6, p. 591–598, 2009. 20

50 LAI, M.-C. et al. Sex/gender differences and autism: setting the scene for future research. *Journal of the American Academy of Child & Adolescent Psychiatry*, Elsevier, v. 54, n. 1, p. 11–24, 2015. 20, 62

51 RUSSELL, G.; STEER, C.; GOLDING, J. Social and demographic factors that influence the diagnosis of autistic spectrum disorders. *Social psychiatry and psychiatric epidemiology*, Springer, v. 46, n. 12, p. 1283–1293, 2011. 20

52 GIARELLI, E. et al. Sex differences in the evaluation and diagnosis of autism spectrum disorders among children. *Disability and health journal*, Elsevier, v. 3, n. 2, p. 107–116, 2010. 20

53 BEGEER, S. et al. Sex differences in the timing of identification among children and adults with autism spectrum disorders. *Journal of autism and developmental disorders*, Springer, v. 43, n. 5, p. 1151–1156, 2013. 20

54 FUSAR-POLI, L. et al. Diagnosing asd in adults without id: accuracy of the ados-2 and the adi-r. *Journal of autism and developmental disorders*, Springer, v. 47, n. 11, p. 3370–3379, 2017. 21, 89

55 MAZEFSKY, C. A.; OSWALD, D. P. The discriminative ability and diagnostic utility of the ados-g, adi-r, and gars for children in a clinical setting. *Autism*, Sage Publications Sage CA: Thousand Oaks, CA, v. 10, n. 6, p. 533–549, 2006. 21, 89

56 EMERSON, R. W. et al. Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Science translational medicine*, American Association for the Advancement of Science, v. 9, n. 393, p. eaag2882, 2017. 21, 67, 84

57 GUTHRIE, W. et al. Early diagnosis of autism spectrum disorder: stability and change in clinical diagnosis and symptom presentation. *Journal of Child Psychology and Psychiatry*, Wiley Online Library, v. 54, n. 5, p. 582–590, 2013. 21

58 OZONOFF, S. et al. A prospective study of the emergence of early behavioral signs of autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, Elsevier, v. 49, n. 3, p. 256–266, 2010. 21

59 TCHACONAS, A.; ADESMAN, A. Autism spectrum disorders: a pediatric overview and update. *Current opinion in pediatrics*, LWW, v. 25, n. 1, p. 130–143, 2013. 22

60 GEUNS, R.-J. M. V. et al. Basic principles of magnetic resonance imaging. *Progress in cardiovascular diseases*, Elsevier, v. 42, n. 2, p. 149–156, 1999. 22

61 OGAWA, S. et al. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, National Acad Sciences, v. 87, n. 24, p. 9868–9872, 1990. 23

62 ROY, C. S.; SHERRINGTON, C. S. On the regulation of the blood-supply of the brain. *The Journal of physiology*, Wiley-Blackwell, v. 11, n. 1-2, p. 85, 1890. 23

63 PAULING, L.; CORYELL, C. D. The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 22, n. 4, p. 210–216, 1936. 23

64 BUXTON, R. B. *Introduction to functional magnetic resonance imaging: principles and techniques*. [S.l.]: Cambridge university press, 2009. 23

65 LEE, M. H.; SMYSER, C. D.; SHIMONY, J. S. Resting-state fmri: a review of methods and clinical applications. *American Journal of neuroradiology*, Am Soc Neuroradiology, v. 34, n. 10, p. 1866–1872, 2013. 23, 24, 25

66 SMITH, S. M. et al. Functional connectomics from resting-state fmri. *Trends in cognitive sciences*, Elsevier, v. 17, n. 12, p. 666–682, 2013. 24, 25

67 BISWAL, B. et al. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, Wiley Online Library, v. 34, n. 4, p. 537–541, 1995. 24

68 COLE, D. M.; SMITH, S. M.; BECKMANN, C. F. Advances and pitfalls in the analysis and interpretation of resting-state fmri data. *Frontiers in systems neuroscience*, Frontiers, v. 4, p. 8, 2010. 24

69 LEWIS, C. M. et al. Learning sculpts the spontaneous activity of the resting human brain. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 106, n. 41, p. 17558–17563, 2009. 24

70 CRADDOCK, R. C. et al. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, Wiley Online Library, v. 33, n. 8, p. 1914–1928, 2012. 25, 86

71 MARTINO, A. D. et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, Nature Publishing Group, v. 19, n. 6, p. 659–667, 2014. 27

72 MARTINO, A. D. et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, Nature Publishing Group, v. 4, n. 1, p. 1–15, 2017. 27

73 CRADDOCK, C. et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, v. 7, 2013. 27

74 BROWN, J. A. et al. The ucla multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis. *Frontiers in neuroinformatics*, Frontiers, v. 6, p. 28, 2012. 27

75 DEKHIL, O. et al. Using resting state functional mri to build a personalized autism diagnosis system. In: IEEE. *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. [S.l.], 2018. p. 1381–1385. 27, 67, 72

76 TOLAN, E.; ISIK, Z. Graph theory based classification of brain connectivity network for autism spectrum disorder. In: SPRINGER. *International Conference on Bioinformatics and Biomedical Engineering*. [S.l.], 2018. p. 520–530. 27, 67, 87

77 ARBABSHIRANI, M. R. et al. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, Elsevier, v. 145, p. 137–165, 2017. 31, 81

78 JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. 31, 33, 34

79 RAJASEKARAN, S.; PAI, G. V. *Neural Networks, Fuzzy Systems and Evolutionary Algorithms: Synthesis and Applications*. [S.l.]: PHI Learning Pvt. Ltd., 2017. 32

80 KELLER, J. M.; FOGEL, D. B.; LIU, D. *Fundamentals of computational intelligence: neural networks, fuzzy systems, and evolutionary computation*. [S.l.]: John Wiley & Sons, 2016. 32

81 SAMARASINGHE, S. *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. [S.l.]: Auerbach publications, 2016. 32

82 AMATO, F. et al. *Artificial neural networks in medical diagnosis*. [S.l.]: Elsevier, 2013. 32

83 Wilamowski, B. M. Neural network architectures and learning algorithms. *IEEE Industrial Electronics Magazine*, v. 3, n. 4, p. 56–63, Dec 2009. ISSN 1932-4529. 32

84 LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. 33

85 BENGIO, Y. et al. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, Now Publishers, Inc., v. 2, n. 1, p. 1–127, 2009. 33

86 KLEINBAUM, D. G. et al. *Logistic Regression*. [S.l.]: Springer, 2002. 33

87 SUN, S. A survey of multi-view machine learning. *Neural Computing and Applications*, Springer, v. 23, n. 7-8, p. 2031–2038, 2013. 33, 34

88 XU, C.; TAO, D.; XU, C. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013. 34

89 WANG, J. et al. Multi-task diagnosis for autism spectrum disorders using multi-modality features: A multi-center study. *Human brain mapping*, Wiley Online Library, v. 38, n. 6, p. 3081–3097, 2017. 34, 67, 86

90 JIE, B. et al. Manifold regularized multitask feature learning for multimodality disease classification. *Human brain mapping*, Wiley Online Library, v. 36, n. 2, p. 489–507, 2015. 34

91 ZHANG, D. et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, Elsevier, v. 59, n. 2, p. 895–907, 2012. 34

92 ZHOU, D. et al. Multi-task multi-view learning based on cooperative multi-objective optimization. *IEEE Access*, v. 6, p. 19465–19477, 2018. ISSN 2169-3536. 34, 67

93 BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. 34

94 SCHWARZER, G.; CARPENTER, J. R.; RÜCKER, G. *Meta-analysis with R*. [S.l.]: Springer, 2015. v. 4784. 36

95 MOHER, D. et al. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, American College of Physicians, v. 151, n. 4, p. 264–269, 2009. 36

96 MOHER, D. et al. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, BioMed Central, v. 4, n. 1, p. 1, 2015. 36

97 KEELE, S. et al. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007. 36, 37, 39

98 WOHLIN, C. et al. *Experimentation in software engineering*. [S.l.]: Springer Science & Business Media, 2012. 36, 37, 38, 39, 40, 61

99 TRANFIELD, D.; DENYER, D.; SMART, P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, Wiley Online Library, v. 14, n. 3, p. 207–222, 2003. 36

100 SHULL, F.; SINGER, J.; SJØBERG, D. I. *Guide to advanced empirical software engineering*. [S.l.]: Springer, 2007. 37, 38, 39

101 IRWIG, L. et al. Meta-analytic methods for diagnostic test accuracy. *Journal of clinical epidemiology*, Elsevier, v. 48, n. 1, p. 119–130, 1995. 40

102 ÇOĞALTAY, N.; KARADAĞ, E. Introduction to meta-analysis. In: *Leadership and organizational outcomes*. [S.l.]: Springer, 2015. p. 19–28. 40

103 BORENSTEIN, M. et al. When does it make sense to perform a meta-analysis. *Introduction to meta-analysis*, John Wiley & Sons, Chichester, United Kingdom, p. 357–64, 2009. 40, 51, 52

104 BIONDI-ZOCCAI, G. *Diagnostic Meta-Analysis: A Useful Tool for Clinical Decision-Making*. [S.l.]: Springer, 2018. 40, 41, 43, 47, 49, 50, 51, 53, 61

105 COLLABORATION, C. et al. *Cochrane handbook for systematic reviews of interventions*. [S.l.]: Cochrane Collaboration, 2008. 41

106 BORENSTEIN, M. et al. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, Wiley Online Library, v. 1, n. 2, p. 97–111, 2010. 41

107 HARBORD, R. M. et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of clinical epidemiology*, Elsevier, v. 61, n. 11, p. 1095–1103, 2008. 42

108 ZWINDERMAN, A. H.; BOSSUYT, P. M. We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in medicine*, Wiley Online Library, v. 27, n. 5, p. 687–697, 2008. 43

109 LEE, J. et al. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-part ii. statistical methods of meta-analysis. *Korean journal of radiology*, v. 16, n. 6, p. 1188–1196, 2015. 47, 49, 50, 51

110 LITTENBERG, B.; MOSES, L. E. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Medical Decision Making*, Sage Publications Sage CA: Thousand Oaks, CA, v. 13, n. 4, p. 313–321, 1993. 47

111 MOSES, L. E.; SHAPIRO, D.; LITTENBERG, B. Combining independent studies of a diagnostic test into a summary roc curve: data-analytic approaches and some additional considerations. *Statistics in medicine*, Wiley Online Library, v. 12, n. 14, p. 1293–1316, 1993. 47

112 REITSMA, J. B. et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, Elsevier, v. 58, n. 10, p. 982–990, 2005. 48, 50, 51, 65

113 RUTTER, C. M.; GATSONIS, C. A. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in medicine*, Wiley Online Library, v. 20, n. 19, p. 2865–2884, 2001. 48, 49, 50

114 IRWIG, L. et al. Guidelines for meta-analyses evaluating diagnostic tests. *Annals of internal medicine*, American College of Physicians, v. 120, n. 8, p. 667–676, 1994. 52

115 WHITING, P. F. et al. Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, American College of Physicians, v. 155, n. 8, p. 529–536, 2011. 53, 54, 61, 62

116 LEEFLANG, M. et al. *Sources of bias - Lesson 6.1: Cochrane Collaboration DTA Online Learning Materials*. [S.l.]: The Cochrane Collaboration, 2014. <http://training.cochrane.org>. Videocast (32 slides, 26min, sound, colour). 53, 54

117 PARK, S. H.; HAN, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, Radiological Society of North America, v. 286, n. 3, p. 800–809, 2018. 62, 80, 89, 90

118 DOEBLER, P.; HOLLING, H. Meta-analysis of diagnostic accuracy with mada. *R Packag*, v. 1, p. 15, 2015. 65

119 TEAM, R. C. *R: A Language and Environment for Statistical Computing*. 2012. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.r-project.org/>. 65

120 COMMUNITY, C. *Review manager (RevMan). version 5.3*. Accessed January 06, 2020. <https://community.cochrane.org/help/tools-and-software/revman-5/revman-5-download>. 65

121 KAMBEITZ, J. et al. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, Nature Publishing Group, v. 40, n. 7, p. 1742–1751, 2015. 65

122 BHAUMIK, R. et al. Predicting autism spectrum disorder using domain-adaptive cross-site evaluation. *Neuroinformatics*, Springer, v. 16, n. 2, p. 197–205, 2018. 67, 68

123 BI, X.-a. et al. Analysis of asperger syndrome using genetic-evolutionary random support vector machine cluster. *Frontiers in physiology*, Frontiers Media SA, v. 9, 2018. 67, 68

124 BI, X.-a. et al. Classification of autism spectrum disorder using random support vector machine cluster. *Frontiers in genetics*, Frontiers, v. 9, p. 18, 2018. 67, 68

125 CHEN, C. P. et al. Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism. *NeuroImage: Clinical*, Elsevier, v. 8, p. 238–245, 2015. 67, 72, 82, 84, 91

126 DVORNEK, N. C. et al. Identifying autism from resting-state fmri using long short-term memory networks. In: SPRINGER. *International Workshop on Machine Learning in Medical Imaging*. [S.l.], 2017. p. 362–370. 67, 68

127 GUO, X. et al. Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Frontiers in neuroscience*, Frontiers, v. 11, p. 460, 2017. 67, 68

128 IIDAKA, T. Resting state functional magnetic resonance imaging and neural network classified autism and control. *Cortex*, Elsevier, v. 63, p. 55–67, 2015. 67

129 KAM, T.-E.; SUK, H.-I.; LEE, S.-W. Multiple functional networks modeling for autism spectrum disorder diagnosis. *Human brain mapping*, Wiley Online Library, v. 38, n. 11, p. 5804–5821, 2017. 67

130 KHOSLA, M. et al. 3d convolutional neural networks for classification of functional connectomes. *arXiv preprint arXiv:1806.04209*, 2018. 67, 68, 86

131 LI, H.; PARIKH, N. A.; HE, L. A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. *Frontiers in neuroscience*, Frontiers, v. 12, p. 491, 2018. 67

132 LIAO, D.; LU, H. Classify autism and control based on deep learning and community structure on resting-state fmri. In: IEEE. *Advanced Computational Intelligence (ICACI), 2018 Tenth International Conference on*. [S.l.], 2018. p. 289–294. 67, 68

133 PRICE, T. et al. Multiple-network classification of childhood autism using functional connectivity dynamics. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2014. p. 177–184. 67, 87

134 SARTIPI, S.; KALBKHANI, H.; SHAYESTEH, M. G. Ripplet ii transform and higher order cumulants from r-fmri data for diagnosis of autism. In: IEEE. *Electrical and Electronics Engineering (ELECO), 2017 10th International Conference on*. [S.l.], 2017. p. 557–560. 67, 68

135 ZHAO, Y. et al. 3d deep convolutional neural network revealed the value of brain network overlap in differentiating autism spectrum disorder from healthy controls. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2018. p. 172–180. 67

136 AGHDAM, M. A.; SHARIFI, A.; PEDRAM, M. M. Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional neural networks. *Journal of digital imaging*, Springer, v. 32, n. 6, p. 899–918, 2019. 67

137 BI, X.-a. et al. The genetic-evolutionary random support vector machine cluster analysis in autism spectrum disorder. *IEEE Access*, IEEE, v. 7, p. 30527–30535, 2019. 67, 68

138 BRAHIM, A.; HASSANI, M. H. E.; FARRUGIA, N. Classification of autism spectrum disorder through the graph fourier transform of fmri temporal signals projected on structural connectome. In: SPRINGER. *International Conference on Computer Analysis of Images and Patterns*. [S.l.], 2019. p. 45–55. 67

139 DAMMU, P. S.; BAPI, R. S. Employing temporal properties of brain activity for classifying autism using machine learning. In: SPRINGER. *International Conference on Pattern Recognition and Machine Intelligence*. [S.l.], 2019. p. 193–200. 67, 87

140 DSOUZA, A. M.; ABIDIN, A. Z.; WISMÜLLER, A. Classification of autism spectrum disorder from resting-state fmri with mutual connectivity analysis. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*. [S.l.], 2019. v. 10953, p. 109531D. 67, 68

141 EL-GAZZAR, A. et al. A hybrid 3dcnn and 3dc-lstm based model for 4d spatio-temporal fmri data: An abide autism classification study. In: *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*. [S.l.]: Springer, 2019. p. 95–102. 67, 68

142 GUPTA, S. et al. Ambivert degree identifies crucial brain functional hubs and improves detection of alzheimer's disease and autism spectrum disorder. *NeuroImage: Clinical*, Elsevier, v. 25, p. 102186, 2020. 67, 68

143 HUANG, F. et al. Multi-template based auto-weighted adaptive structural learning for asd diagnosis. In: SPRINGER. *International Workshop on Machine Learning in Medical Imaging*. [S.l.], 2019. p. 516–524. 67, 68, 80

144 KAZEMINEJAD, A.; SOTERO, R. C. Topological properties of resting-state fmri functional networks improve machine learning-based autism classification. *Frontiers in neuroscience*, Frontiers, v. 12, p. 1018, 2019. 67, 87

145 KHOSLA, M. et al. Ensemble learning with 3d convolutional neural networks for functional connectome-based prediction. *Neuroimage*, Elsevier, v. 199, p. 651–662, 2019. 67, 68

146 LANKA, P. et al. Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. *Brain Imaging and Behavior*, Springer, p. 1–39, 2019. 67

147 RAJESH, G.; PANNIRSELVAM, S. Lucid ant colony optimization based denoiser for effective autism spectrum disorder classification. *International Journal of Advanced Science and Technology*, v. 28, n. 17, p. 865 – 876, Dec. 2019. 67, 68

148 SAIRAM, K. et al. Computer aided system for autism spectrum disorder using deep learning methods. *International Journal of Psychosocial Rehabilitation*, v. 23, n. 01, 2019. 67, 68

149 SHERKATGHANAD, Z. et al. Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in Neuroscience*, Frontiers Media SA, v. 13, 2019. 67, 68

150 SPERA, G. et al. Evaluation of altered functional connections in male children with autism spectrum disorders on multiple-site data optimized with machine learning. *Frontiers in psychiatry*, Frontiers, v. 10, p. 620, 2019. 67

151 MARTIAL, E. E. T.; HU, L.; YUQING, S. Characterising and predicting autism spectrum disorder by performing resting-state functional network community pattern analysis. *Frontiers in human neuroscience*, Frontiers, v. 13, p. 203, 2019. 67

152 WANG, C.; XIAO, Z.; WU, J. Functional connectivity-based classification of autism and control using svm-rfecv on rs-fmri data. *Physica Medica*, Elsevier, v. 65, p. 99–105, 2019. 67

153 WANG, J. et al. Interpretable feature learning using multi-output takagi-sugeno-kang fuzzy system for multi-center asd diagnosis. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2019. p. 790–798. 67

154 YANG, X.; ISLAM, M. S.; KHALED, A. A. Functional connectivity magnetic resonance imaging classification of autism spectrum disorder using the multisite abide dataset. In: IEEE. *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. [S.l.], 2019. p. 1–4. 67

155 YUAN, D.; ZHU, L.; HUANG, H. Prediction of autism spectrum disorder based on imbalanced resting-state fmri data using clustering oversampling. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. *Tenth International Conference on Signal Processing Systems*. [S.l.], 2019. v. 11071, p. 110710W. 67

156 AGHDAM, M. A.; SHARIFI, A.; PEDRAM, M. M. Combination of rs-fmri and smri data to discriminate autism spectrum disorders in young children using deep belief network. *Journal of digital imaging*, Springer, p. 1–9, 2018. 67, 86, 87

157 DVORNEK, N. C.; VENTOLA, P.; DUNCAN, J. S. Combining phenotypic and resting-state fmri data for autism classification with recurrent neural networks. In: IEEE. *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. [S.l.], 2018. p. 725–728. 67, 68

158 SEN, B. et al. A general prediction model for the detection of adhd and autism using structural and functional mri. *PloS one*, Public Library of Science, v. 13, n. 4, p. e0194856, 2018. 67

159 ZHOU, Y.; YU, F.; DUONG, T. Multiparametric mri characterization and prediction in autism spectrum disorder using graph theory and machine learning. *PLoS One*, Public Library of Science, v. 9, n. 6, p. e90405, 2014. 67, 68

160 EILL, A. et al. Functional connectivities are more informative than anatomical variables in diagnostic classification of autism. *Brain connectivity*, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 9, n. 8, p. 604–612, 2019. 67, 82

161 SARTIPI, S.; SHAYESTEH, M. G.; KALBKHANI, H. Diagnosing of autism spectrum disorder based on garch variance series for rs-fmri data. In: IEEE. *2018 9th International Symposium on Telecommunications (IST)*. [S.l.], 2018. p. 86–90. 67, 68, 83

162 MASTROVITO, D.; HANSON, C.; HANSON, S. J. Differences in atypical resting-state effective connectivity distinguish autism from schizophrenia. *NeuroImage: Clinical*, Elsevier, v. 18, p. 367–376, 2018. 67

163 HEINSFELD, A. S. et al. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, Elsevier, v. 17, p. 16–23, 2018. 67

164 DEKHIL, O. et al. A personalized autism diagnosis cad system using a fusion of structural mri and resting-state functional mri data. *Frontiers in Psychiatry*, Frontiers Media SA, v. 10, 2019. 67, 68

165 JUN, E.; SUK, H.-I. Region-wise stochastic pattern modeling for autism spectrum disorder identification and temporal dynamics analysis. In: SPRINGER. *International Workshop on Connectomics in Neuroimaging*. [S.l.], 2017. p. 143–151. 67

166 WONG, E. et al. Riemannian regression and classification models of brain networks applied to autism. In: SPRINGER. *International Workshop on Connectomics in Neuroimaging*. [S.l.], 2018. p. 78–87. 67

167 ZHU, Y. et al. A tensor statistical model for quantifying dynamic functional connectivity. In: SPRINGER. *International Conference on Information Processing in Medical Imaging*. [S.l.], 2017. p. 398–410. 67, 87

168 REN, Y.; WANG, S. Exploring functional connectivity biomarker in autism using group-wise sparse representation. In: *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*. [S.l.]: Springer, 2019. p. 21–29. 67

169 DEKHIL, O. et al. Identifying personalized autism related impairments using resting functional mri and ados reports. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2018. p. 240–248. 67

170 CHAITRA, N.; VIJAYA, P. A. Comparing univalent and bivalent brain functional connectivity measures using machine learning. In: *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*. [S.l.: s.n.], 2017. p. 1–5. 67, 68

171 DODERO, L. et al. Kernel-based classification for brain connectivity graphs on the riemannian manifold of positive definite matrices. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. [S.l.: s.n.], 2015. p. 42–45. ISSN 1945-7928. 67, 68

172 MAHANAND, B. S. et al. An enhanced effect-size thresholding method for the diagnosis of autism spectrum disorder using resting state functional mri. In: *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*. [S.l.: s.n.], 2016. p. 1–6. 67, 68, 83

173 REN, Y. et al. Identifying autism biomarkers in default mode network using sparse representation of resting-state fmri data. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. [S.l.: s.n.], 2016. p. 1278–1281. ISSN 1945-8452. 67

174 VIGNESHWARAN, S. et al. Using regional homogeneity from functional mri for diagnosis of asd among males. In: *2015 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2015. p. 1–8. ISSN 2161-4407. 67, 68, 83

175 CHEN, Z.; JI, J.; LIANG, Y. Convolutional neural network with an element-wise filter to classify dynamic functional connectivity. In: IEEE. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.], 2019. p. 643–646. 67, 87

176 GAZZAR, A. E. et al. Simple 1-d convolutional networks for resting-state fmri based classification in autism. In: IEEE. *2019 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2019. p. 1–6. 67, 87

177 HUANG, F. et al. Sparse low-rank constrained adaptive structure learning using multi-template for autism spectrum disorder diagnosis. In: IEEE. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. [S.l.], 2019. p. 1555–1558. 67

178 LI, J. et al. Deep forest with cross-shaped window scanning mechanism to extract topological features. In: IEEE. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.], 2019. p. 688–691. 67, 68

179 MOSTAFA, S.; TANG, L.; WU, F.-X. Diagnosis of autism spectrum disorder based on eigenvalues of brain networks. *IEEE Access*, IEEE, v. 7, p. 128474–128486, 2019. 67, 68

180 SIDHU, G. Locally linear embedding and fmri feature selection in psychiatric classification. *IEEE journal of translational engineering in health and medicine*, IEEE, v. 7, p. 1–11, 2019. 67

181 WANG, C. et al. Identification of autism based on svm-rfe and stacked sparse auto-encoder. *IEEE Access*, IEEE, v. 7, p. 118030–118036, 2019. 67

182 WANG, M. et al. Identifying autism spectrum disorder with multi-site fmri via low-rank domain adaptation. *IEEE Transactions on Medical Imaging*, IEEE, v. 39, n. 3, p. 644–655, 2019. 67

183 ZHANG, M. et al. Comparison of neural networks' performance in early screening of autism spectrum disorders under two mri principles. In: IEEE. *2019 International Conference on Networking and Network Applications (NaNA)*. [S.l.], 2019. p. 338–343. 67, 68

184 ZHAO, Y. et al. Two-stage spatial temporal deep learning framework for functional brain network modeling. In: IEEE. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. [S.l.], 2019. p. 1576–1580. 67, 68

185 CRIMI, A. et al. Case-control discrimination through effective brain connectivity. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. [S.l.: s.n.], 2017. p. 970–973. ISSN 1945-8452. 67, 68

186 WANG, J. et al. Sparse multiview task-centralized ensemble learning for asd diagnosis based on age- and sex-related functional connectivity patterns. *IEEE Transactions on Cybernetics*, p. 1–14, 2018. ISSN 2168-2267. 67

187 ANIRUDH, R.; THIAGARAJAN, J. J. Bootstrapping graph convolutional neural networks for autism spectrum disorder classification. In: IEEE. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2019. p. 3197–3201. 67, 68

188 MELLEMA, C. et al. Multiple deep learning architectures achieve superior performance diagnosing autism spectrum disorder using features previously extracted from structural and functional mri. In: IEEE. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. [S.l.], 2019. p. 1891–1895. 67, 68

189 CHEN, H. et al. Multivariate classification of autism spectrum disorder using frequency-specific resting-state functional connectivity—a multi-center study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, Elsevier, v. 64, p. 1–9, 2016. 67, 84

190 DODERO, L. et al. Kernel-based analysis of functional brain connectivity on grassmann manifold. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2015. p. 604–611. 67

191 SUBBARAJU, V. et al. Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging: A spatial filtering approach. *Medical image analysis*, Elsevier, v. 35, p. 375–389, 2017. 67, 68, 83

192 WEE, C.-Y.; YAP, P.-T.; SHEN, D. Diagnosis of autism spectrum disorders using temporally distinct resting-state functional connectivity networks. *CNS neuroscience & therapeutics*, Wiley Online Library, v. 22, n. 3, p. 212–219, 2016. 67

193 YAHATA, N. et al. A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature communications*, Nature Publishing Group, v. 7, p. 11254, 2016. 67, 85

194 ZHU, Y. et al. Reveal consistent spatial-temporal patterns from dynamic functional connectivity for autism spectrum disorder identification. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2016. p. 106–114. 67, 68

195 BENGS, M.; GESSERT, N.; SCHLAEFER, A. 4d spatio-temporal deep learning with 4d fmri data for autism spectrum disorder classification. *arXiv preprint arXiv:2004.10165*, 2020. 67, 68

196 BERNAS, A.; ALDENKAMP, A. P.; ZINGER, S. Wavelet coherence-based classifier: A resting-state functional mri study on neurodynamics in adolescents with high-functioning autism. *Computer methods and programs in biomedicine*, Elsevier, v. 154, p. 143–151, 2018. 67

197 HUANG, H. et al. Enhancing the representation of functional connectivity networks by fusing multi-view information for autism spectrum disorder diagnosis. *Human brain mapping*, Wiley Online Library, v. 40, n. 3, p. 833–854, 2019. 67

198 KAZEMINEJAD, A.; SOTERO, R. C. The importance of anti-correlations in graph theory based classification of autism spectrum disorder. *bioRxiv*, Cold Spring Harbor Laboratory, p. 557512, 2019. 67, 87

199 SAEED, F. et al. Asd-diagnet: A hybrid learning approach for detection of autism spectrum disorder using fmri data. *Frontiers in Neuroinformatics*, Frontiers, v. 13, p. 70, 2019. 67

200 TEJWANI, R. et al. Autism classification using brain functional connectivity dynamics and machine learning. *arXiv preprint arXiv:1712.08041*, 2017. 67, 72

201 XING, X.; JI, J.; YAO, Y. Convolutional neural network with element-wise filters to extract hierarchical topological features for brain networks. In: IEEE. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.], 2018. p. 780–783. 67

202 GHIASSIAN, S. et al. Using functional or structural magnetic resonance images and personal characteristic data to identify adhd and autism. *PloS one*, Public Library of Science, v. 11, n. 12, p. e0166934, 2016. 67, 84, 85, 86

203 PARISOT, S. et al. Spectral graph convolutions for population-based disease prediction. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2017. p. 177–185. 67, 68

204 UDDIN, L. Q. et al. Salience network–based classification and prediction of symptom severity in children with autism. *JAMA psychiatry*, American Medical Association, v. 70, n. 8, p. 869–879, 2013. 67

205 PIAGET, J. Part i: Cognitive development in children–piaget development and learning. *Journal of research in science teaching*, ERIC, v. 40, 2003. 83

206 FREUD, A. Observations on child development. *The psychoanalytic study of the child*, Taylor & Francis, v. 6, n. 1, p. 18–30, 1951. 83

207 KARMILOFF-SMITH, A. Challenging the use of adult neuropsychological models for explaining neurodevelopmental disorders: Develop ed versus develop ing brains: The 40th sir frederick bartlett lecture. *Quarterly Journal of Experimental Psychology*, SAGE Publications Sage UK: London, England, v. 66, n. 1, p. 1–14, 2013. 84

208 SEGALL, J. M. et al. Voxel-based morphometric multisite collaborative study on schizophrenia. *Schizophrenia bulletin*, Oxford University Press, v. 35, n. 1, p. 82–95, 2009. 84

209 BISWAL, B. B. et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 107, n. 10, p. 4734–4739, 2010. 84

210 TZOURIO-MAZOYER, N. et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, Academic Press, v. 15, n. 1, p. 273–289, 2002. 86

211 ABRAHAM, A. et al. Extracting brain regions from rest fmri with total-variation constrained dictionary learning. In: SPRINGER. *International conference on medical image computing and computer-assisted intervention*. [S.l.], 2013. p. 607–615. 87

212 LINKE, A. et al. Dynamic time warping outperforms pearson correlation in detecting atypical functional connectivity in autism spectrum disorders. *NeuroImage*, Elsevier, v. 223, p. 117383, 2020. 87

213 CARP, J. The secret lives of experiments: methods reporting in the fmri literature. *Neuroimage*, Elsevier, v. 63, n. 1, p. 289–300, 2012. 87

214 JONES, C. M. et al. Guidelines for diagnostic tests and diagnostic accuracy in surgical research. *Journal of Investigative Surgery*, Taylor & Francis, v. 23, n. 1, p. 57–65, 2010. 88

215 JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398. 90

216 IAKOUCHEVA, L. M.; MUOTRI, A. R.; SEBAT, J. Getting to the cores of autism. *Cell*, Elsevier, v. 178, n. 6, p. 1287–1298, 2019. 90

217 DEVILLÉ, W. L. et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC medical research methodology*, Springer, v. 2, n. 1, p. 9, 2002. 91