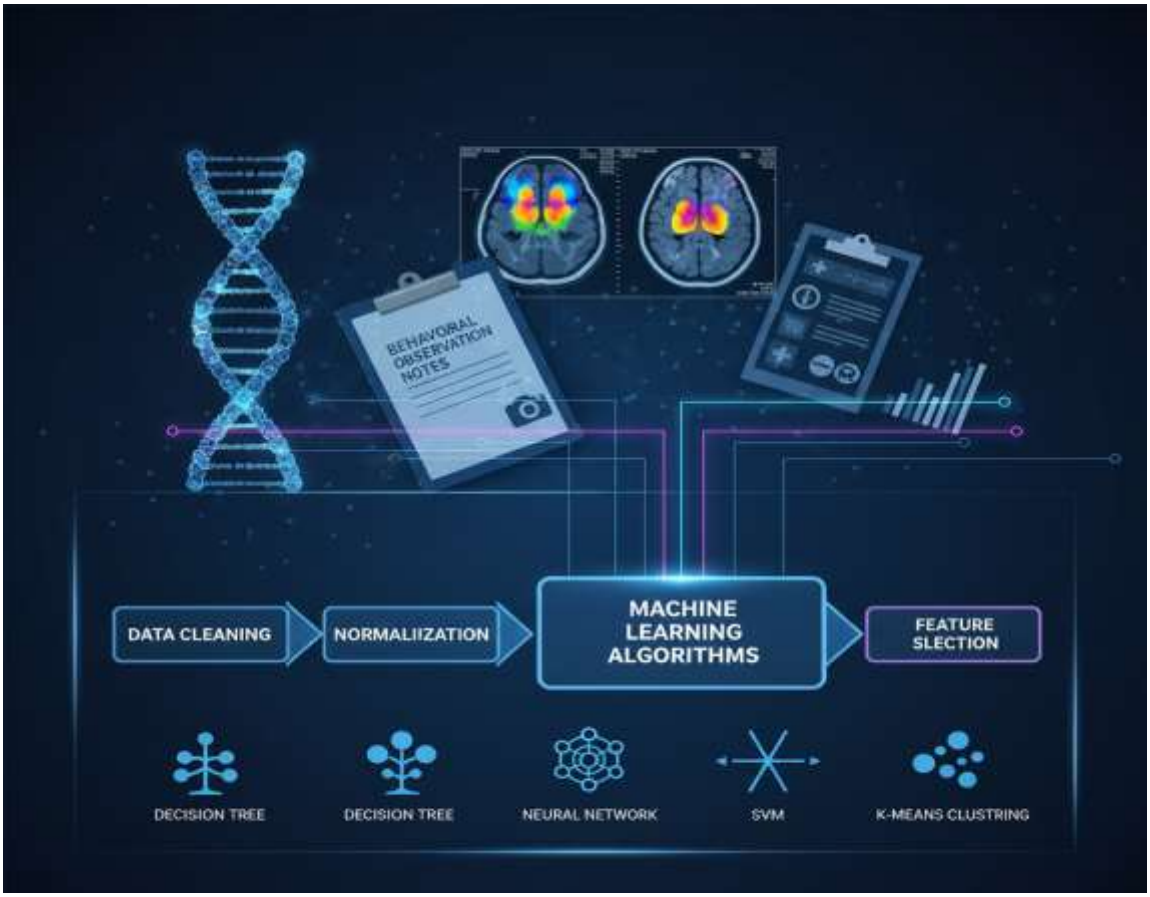
Chapter 2: Data Science and Machine Learning Fundamentals

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computer systems to learn from data, recognize patterns, and make decisions or predictions with minimal human intervention. In contrast to conventional programs that require each rule to be explicitly programmed, ML algorithms create a model using example inputs, or "training data," enabling predictions or decisions to be made without requiring explicit instructions on how to perform the task. In the realm of healthcare, such a paradigm shift is revolutionary, lending itself to a data-driven understanding of complicated conditions, including Autism Spectrum Disorder (Patil et al., 2024). The ML landscape is generally divided into three fundamental frameworks: supervised, unsupervised, and reinforcement learning.



Supervised training is arguably the most widely used ML approach in healthcare and refers to its reliance on labeled training data. In this method, the algorithm is trained on a dataset for which both input features (patient age, gene markers, and behavior scores) and output labels ("ASD" or

"non-ASD") are available. The algorithm aims to discover a mapping from input to output that enables accurate label prediction on new, previously unobserved data. Here, we deal with two types of tasks in supervised learning systems: classification and regression. (5) Classification: classification models predict a discrete (or categorical) outcome, for instance, to classify whether a patient has ASD or not. Famous algorithms used for classification purposes include Decision Trees, which generate a flowchart-like structure to determine the class, and Support Vector Machines (SVMs), which find the optimal hyperplane in n-dimensional space with maximum margin.(GeeksforGeeks, 2025). In contrast, regression models predict a single numeric value, such as a child's developmental age or the severity of their symptoms on a scale.

Unsupervised learning is a powerful approach employed when the data is unlabeled, aiming to learn the underlying patterns or structures. Unlike supervised learning, there is no "true" answer to direct the algorithm toward. In the context of ASD research, unsupervised learning is beneficial to discover underlying distinct ASD subtypes. Because of the tremendous heterogeneity associated with ASD, one diagnosis is likely insufficient to drive treatment. Clustering algorithms (e.g., k-means) can be used to cluster subjects based on their behavioral, genetic, and neuroimaging profiles, identifying novel sub-phenotypes that inform personalized interventions (Solek et al., 2024). Such an approach may reveal functional patterns that are not readily apparent to human inspection.

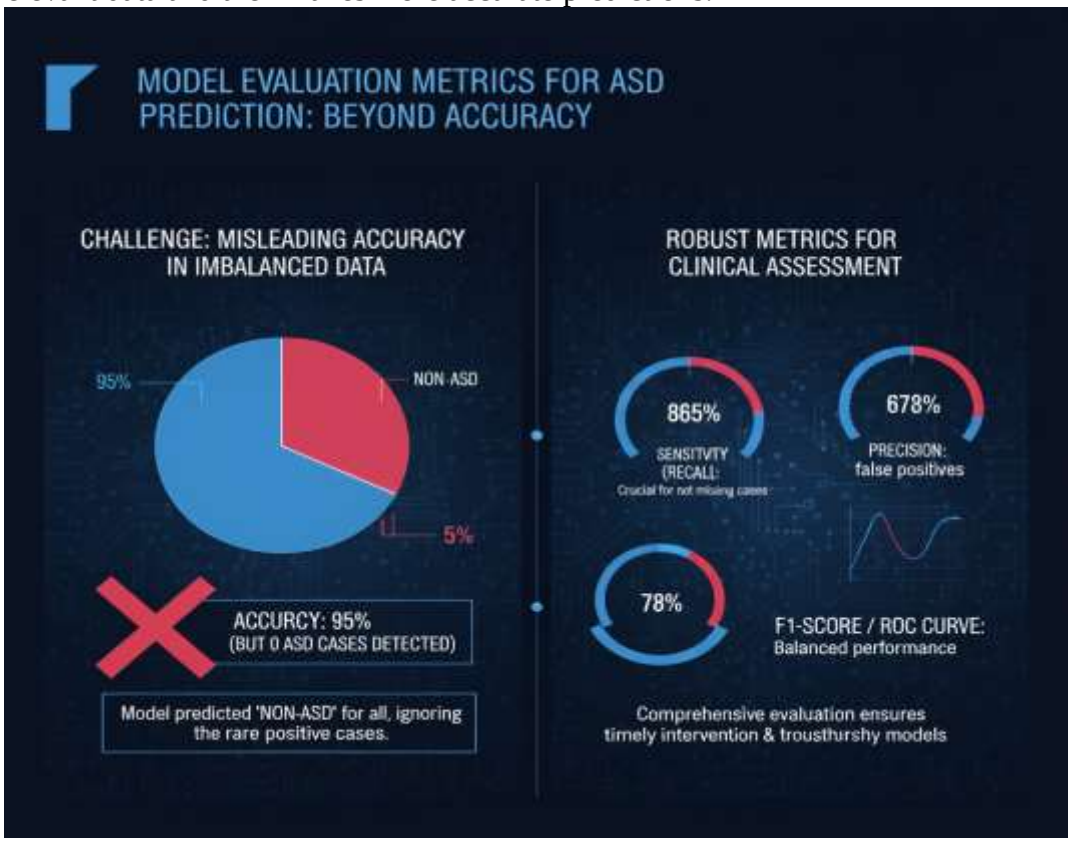
At last, RL is a field in which an agent learns to make a sequence of decisions within an environment such that the total reward is maximized. The agent learns through trial and error, being rewarded for good actions and penalized for bad ones. RL is less frequently used in a diagnostic setting, but it has great potential to advance dynamic and adaptive intervention strategies for autism. For instance, an RL-informed system may process real-time behavioral data from a patient during a therapy session and immediately inform the clinician which strategies work best for this individual, thereby optimizing treatment planning on the fly. This dynamic adaptation model is a significant step beyond static, fixed interventions.

2.1 Key Data Science Concepts for ASD

The general quality and relevance of the input data are key aspects of a well-functioning model in any machine learning application. In ASD studies, multiple data modalities are combined to form a comprehensive picture of the person. These data sources are the source of both structured (e.g., diagnostic codes, demographic information) and unstructured data types (e.g., full-length clinician-generated notes). Behavioural data may also be extracted using standard assessments (e.g., Autism Diagnostic Observation Schedule, ADOS) or in the context of digital devices, which allow for obtaining quantitative scores regarding performances in social interactions, motor skills, and communication. Genomic information, provided through genome sequencing, offers insights into genetic phenotypes and variants associated with ASD. Finally, neuroimaging information, provided by modalities such as functional Magnetic Resonance Imaging (fMRI) and Electroencephalography (EEG), may provide insights into brain structure and connections among

different regions of interest (Solek et al., 2024). All these data types have different aspects for analysis.

Data Preprocessing. Before you can use your data to train a machine learning model, it must undergo a critical process known as data preprocessing. This is a multiple-step process that cleans and prepares raw data for analysis and model training. Poor data quality, including missing values and inconsistent formats, is the primary challenge in clinical data (Liu et al., 2024). Methods such as mean/median imputation or more complex regression-based approaches are employed to fill in these missing entries. Feature selection is another pivotal step in selecting the most informative features for dimensionality reduction, improving model performance,, and enhancing interpretability. For example, among thousands of genetic markers, only a small subset may truly predict whether someone will be diagnosed with an ASD. Different types of feature selection algorithms are, for example: 1) Filter method: it selects features independently from any supervised learning technique. Mode details can be found. Concatenate all predictors to train a logistic regression model using them. Good preprocessing is vital to ensure the model learns from clean, relevant data and then makes more accurate predictions.



Its performance is evaluated using several quantitative measures. Just accuracy alone is often misleading, especially in the clinical domain when there are far more non-ASD than ASDs (a classic example of what is known as an "imbalanced" dataset). For instance, a model that always predicts “non-ASD” could achieve high accuracy but fail to detect any cases at all. To compensate for this, healthcare apps use more sophisticated metrics. Precision is the ratio between the number of accurate positive predictions and the total number of optimistic predictions made by a model. Sensitivity, also known as Recall, is the ratio of accurate optimistic predictions to all actual positive cases (Neptune. ai, 2024). In the context of ASD diagnosis, high recall is crucial for the model to

avoid missing cases and facilitate timely intervention. The F1 score balances precision and recall, while the ROC curve (and its AUC) provides practical insights into a model's performance across varying classification thresholds, allowing for an informed trade-off between false positives and false negatives (Keylabs, 2024).

2.2 Tools and Technologies

Machine learning in ASD research is driven by a rich ecosystem of utilities, libraries, and open data sources, making the utilization of this technology extremely appealing. Python is now the unquestionable industry leader in data science and machine learning, primarily due to the plethora of libraries it offers, as well as its ease of readability. Key libraries entail Scikit-learn, a flexible and powerful Python library for classic ML algorithms, including support vector machines (SVMs), decision trees, and clustering (Scikit-learn, 2025). For instance, in deep learning applications, which are becoming increasingly significant in processing complex data types such as neuroimaging and video data, implemented libraries like TensorFlow or PyTorch support the construction of highly complex neural networks by enabling their training (GeeksforGeeks, 2025). Python is the primary language used, but many people continue to use the statistical programming language R, which is popular in academic research and has a massive community with numerous libraries for data analysis and visualization. These software packages provide the computational infrastructure for analyzing and modeling the large-scale data generated in ASD studies.

The lack of access to high-quality, diverse data has arguably been the most significant barrier to advancing ASD research. To accommodate this, different projects and repositories for data sharing have been founded. For instance, the Autism Brain Imaging Data Exchange (ABIDE) database is an international initiative that compiles and openly disseminates resting-state fMRI and structural MRI data for a large sample of subjects with and without ASD, along with associated phenotypic descriptors (NYU Data Catalog, 2017). This publicly available resource has significantly facilitated the development and validation of neuroimaging-based predictive models by researchers worldwide. Besides these public databases, researchers also rely on clinical data collected through large-scale studies, such as SPARK (the Simons Foundation Powering Autism Research for Knowledge), as well as published research. Such multi-institutional data sharing efforts are critical to advancing beyond single-site investigations and developing predictive models of ASD that are generalizable and robust.