# Enhancing Early Intervention: ML-Based Autism Detection in Children

Technical Project Report & Implementation

Swati Swarup Panda
Anand Naik
Abhik Sundar Sahu

Department of Computer Science and Engineering
International Institute of Information Technology, Bhubaneswar

Under the Guidance of **Dr. Anjali Mohapatra**

December 2025

# Presentation Outline

# What is Autism Spectrum Disorder (ASD)?

## Definition

ASD is a complex neurodevelopmental condition that affects how people communicate and interact with the world. It is defined as a "spectrum" because it affects individuals in varying ways and degrees.

**Core Characteristics (The Triad of Impairments):**

- **Social Communication:** Challenges in using or understanding verbal and non-verbal language (e.g., literal interpretation, delayed speech).
- **Social Interaction:** Difficulties in recognizing social cues, maintaining eye contact, or expressing emotions.
- **Social Imagination:** Repetitive behavior patterns, resistance to change, and restricted interests.

# Global Landscape & The "Detection Gap"

## Epidemiology

- ASD is a pervasive neurodevelopmental condition altering cognitive and social maturation.
- **Prevalence:** Rise from 1 in 150 (early 2000s) to **1 in 36** children (2023 CDC data).
- It has shifted from a rare condition to a global public health priority.

## The "Detection Gap"

- **Biological Reality:** ASD can be reliably diagnosed by 18–24 months.
- **Logistical Failure:** Median diagnosis age remains above 4 years in many developed nations.
- **Bottleneck:** Gold standards (ADOS, ADI-R) are resource-intensive and require specialized training.

# Paper 1: Text-Based Detection (Mukherjee et al., 2023)

## Early Detection of ASD using Traditional Machine Learning Models

**Objective:** Automate detection using parents' text dialogues to address professional shortages.

### Methodology

A two-stage classification process:

1. **Sentiment Analysis:** Trained SVM, Logistic Regression, RF, and KNN to predict if a sentence indicates a "positive" symptom.
2. **Symptom Mapping:** Used Spacy Cosine Similarity Model to map positive sentences to clinical labels (e.g., "Eye Contact Problem").

### Key Results

- **Best Accuracy:** 71% (SVM & Logistic Regression).
- **AUC Score:** 0.77 (SVM).
- **Validation:** Successfully mapped phrases like "Eyes are scrolling" to correct labels.

### Limitations

- **Dependency:** Failure in Stage 1 causes total system failure.
- **Scalability:** Traditional models struggle with larger, complex datasets.

# Paper 2: Image-Based Detection (Ghazal et al., 2023)

Early Detection of Autism in Children Using Transfer Learning

**Objective:** Detect subtle facial biomarkers (e.g., lack of eye contact) using Deep Learning.

## Methodology

- **Dataset:** Kaggle repository of 2,940 facial images (Autistic vs. Non-Autistic).
- **Model: ASDDTLA** (ASD Deep Transfer Learning Algorithm), based on modified **AlexNet**.
- **Technique:** Utilized Transfer Learning (ImageNet weights) to overcome data scarcity.

### Key Results

- **Accuracy:** 87.7% (Superior to standard CNNs at 63.11%).
- **Miss Rate:** Very low (12.3%).
- **Sensitivity:** High sensitivity (87.6%) crucial for medical screening.

### Limitations

- **Subtlety:** Facial cues are extremely subtle, complicating standard model training.
- **Constraints:** Heavily reliant on pre-trained weights due to lack of massive pediatric datasets.

# Paper 3: Feature Optimization (Bala et al., 2022)

## Efficient Machine Learning Models for Early Stage Detection

**Objective:** Reduce diagnosis time by identifying critical behavioral attributes across age groups.

### Methodology

- **Feature Selection:** Applied Boruta, CFS, RIPPER, and Recursive Feature Elimination (RFE) to reduce dimensionality.
- **Classifiers:** Tested 8 classifiers including SVM, Naïve Bayes, and Random Forest.
- **Explainability:** Used **SHAP** values to rank feature importance.

### Key Results

- **Toddlers:** SVM + RIPPER achieved **97.82% accuracy**.
- **Children:** SVM + CFS achieved **99.61% accuracy**.
- **Key Features:** "Age group" (Toddlers) vs. "Understanding feelings" (Adults).

### Limitations

- **Data Structure:** Not trained on multivariate/high-dimensional data structures.
- **Scope:** Focused on "Version-2" datasets; gaps remain in complex data fusion.

# Implementation Strategy: Base Paper Adaptation

## Base Paper: Mukherjee et al. (2023)

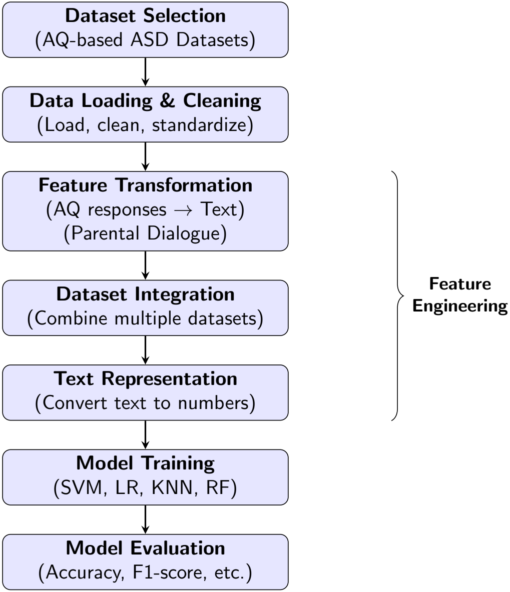
- **Core Concept:** Utilization of parental text dialogues to identify ASD symptoms using sentiment analysis and symptom mapping.
- **Limitation Addressed:** Traditional tabular data often lacks the semantic nuance found in natural language.

## Our Adaptation Strategy

Instead of relying solely on existing text datasets (which are scarce), we implemented a **Feature Transformation** pipeline:

- We utilize structured Q-CHAT/AQ datasets as the source from the UCI Machine Learning repository for ASD.
- **Novel Step:** We algorithmically generate "Parental Dialogue" text from these binary/numeric scores (e.g., A1 Score $\rightarrow$ "Child avoids eye contact").
- This enables us to apply NLP techniques (TF-IDF) to tabular data, mimicking the Base Paper's text-centric approach.

# Methodology Pipeline



# Data Acquisition

## Dataset Sources

The study utilizes publicly available autism screening datasets collected using standardized behavioral questionnaires.

- **Child Dataset:** Age range 4–11 years.
- **Adolescent Dataset:** Age range 12–16 years.
- **Total Instances:** 395 combined records.

## Data Characteristics

- Screening based on AQ-10 / Q-CHAT-10 questionnaires
- Binary behavioral responses capturing ASD-related traits
- Labeled target variable indicating ASD traits (YES / NO)

# Data Cleaning

## Preprocessing Objectives

Raw screening data requires careful cleaning and standardization to ensure consistency and reliability prior to feature engineering.

- Removal of incomplete, duplicate, and inconsistent records
- Standardization of AQ-based screening feature names (A1–A10)
- Verification and correction of binary screening responses (0/1)
- Encoding of categorical target labels into binary format (ASD / Non-ASD)

## Outcome

The data cleaning process ensures a noise-free, consistent dataset, enabling reliable feature transformation and model training.

# Feature Engineering

## Motivation

AQ-based screening data is structured and binary, which limits its semantic expressiveness for machine learning models.

## Feature Engineering Steps

- **Feature Transformation**
  - AQ-10 responses are converted into semantic parental observation text.
- **Dataset Integration**
  - Processed datasets from multiple age groups are merged into a unified dataset.
- **Text Representation**
  - Parental dialogue is transformed into numerical feature vectors using TF-IDF.

# Model Training

## Learning Objective

To classify individuals into ASD and non-ASD categories using engineered textual features.

- Support Vector Machine (SVM)
- Logistic Regression
- K-Nearest Neighbors (KNN)
- Random Forest

## Training Strategy

- Stratified **80–20 train–test split** to preserve class balance
- Uniform feature representation across all classifiers

# Model Evaluation

Using parental dialogue generated from AQ/Q-CHAT-10 screening scores, we evaluated multiple machine learning classifiers.

Table: Performance Metrics on Validation Set

| Algorithm | Accuracy | F1-score | Sensitivity |
|---|---|---|---|
| Linear SVC | 91.14% | 0.91 | 0.90 |
| Logistic Regression | 84.81% | 0.85 | 0.83 |
| KNN | 81.01% | 0.80 | 0.78 |
| Random Forest | 100% | 1.00 | 1.00 |

## Analysis

- **Comparison with Prior Work:** The observed accuracy is higher than that reported in the reference paper, which can be attributed to the use of a smaller and more controlled dataset.
- **Model Behavior:** Tree-based models such as Random Forest achieved very high accuracy due to their ability to capture deterministic patterns introduced during feature engineering.

# Conclusion & Future Scope

## Conclusion

- Presented a structured machine learning pipeline for ASD screening using AQ/Q-CHAT-based behavioral data.
- Demonstrated that transforming structured screening responses into semantic parental dialogue improves classification performance.
- Showed that traditional machine learning models, supported by effective feature engineering, can achieve strong screening accuracy.

## Future Scope

- **Clinical Validation:** Test the approach on larger and real-world clinical datasets.
- **Data Expansion:** Include additional age groups and more diverse screening data.
- **Algorithm Exploration:** Investigate advanced classifiers such as XGBoost for improved performance on larger datasets.

# Thank You