# A Data-Driven Approach To Predict Autism Spectrum Disorders

By:
David C. Anastasiu and Manika Kapoor
San Jose State University, California, USA

# Introduction

- **Objective:** Predict Autism Spectrum Disorders (ASD) and characterize the type of stimuli needed for its detection.


- **Why?**
  - No cure exists
  - Current methods: subjective or responses to single stimulus.

- **How?**
  - Creating machine learning based models using Electrocardiogram (ECG) and Skin Conductance (SC).

# Introduction cont ...

- Time-series component may provide some discriminatory information .
- Analyzing such huge time-series is costly.
- Efficient data preprocessing techniques.

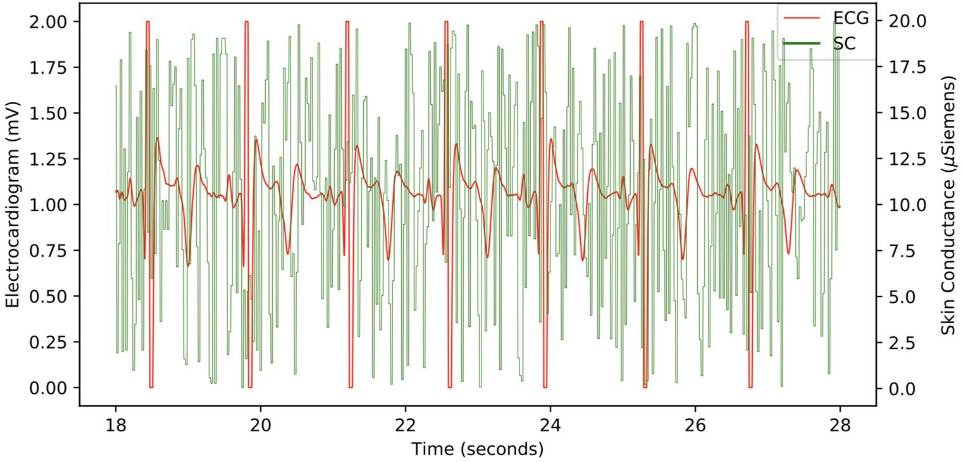



Fig: Sample Dataset

# Background and Related Work

- Based on observation of subjects’ social interaction.
- Subjective (survey-based).
- Social Responsiveness Scale test: not very accurate and tedious.
- Focus on responses to a single stimulus.
- Using physiological responses like ECG using the data collected using computer-based cognitive tasks to train support vector machines (SVM)-based model → predict autism with an accuracy of 82.9%.

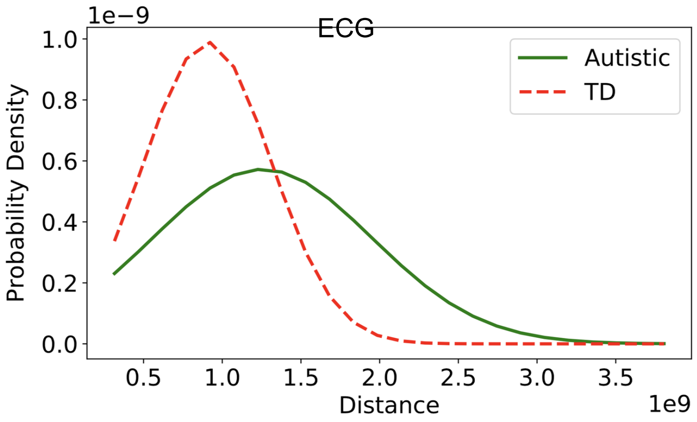
# Dataset

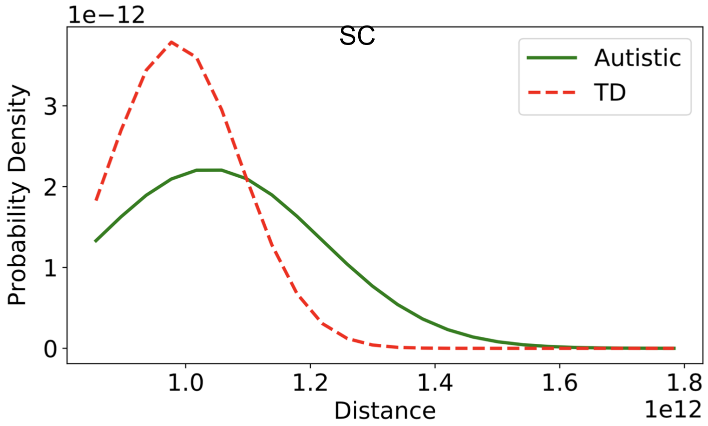
- 25 children (between 5 and 12 years)  of each kind (ASD and TD)..
- Time taken for each protocol → 45 to 90 minutes.
- Included three phases: baseline (3 min), sensory challenge, and recovery (3 min).
- 6 stimuli,  each administered for 3 seconds and was presented at least 8 times.

  Auditory tones, Visual, Auditory siren, olfactory, tactile, vestibular

# Hypothesis and Supporting Evidence

- Autistic children may take longer to return to normal state.
- Compare the data recorded during **baseline and recovery stages**.




Shorter distances between baseline and recovery phase for TD children.

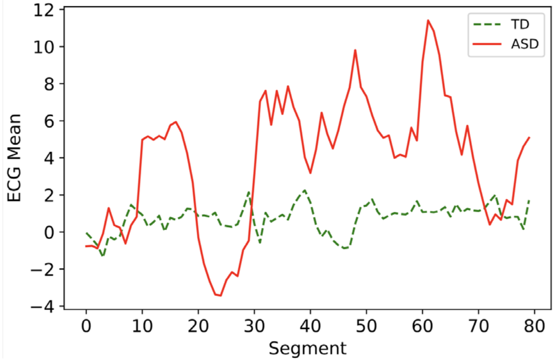
# Feature Extraction Methods

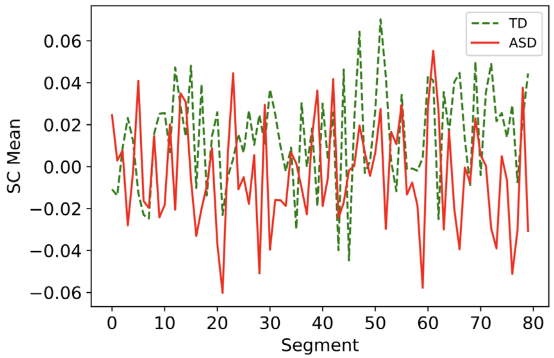
1. **Equal Width Partitioning (EWP):**

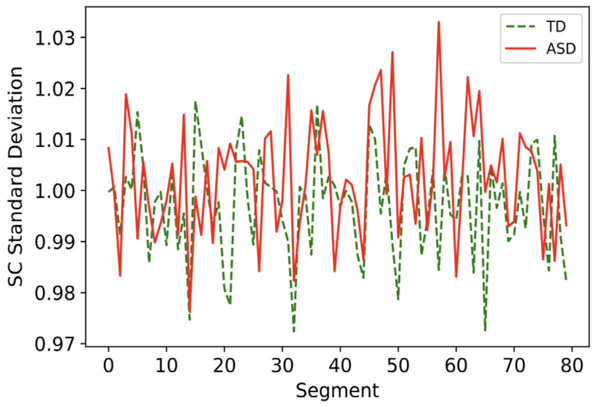
- Split data into 8 parts representing the 8 stimuli .
- Standardized using the mean and standard deviation of the baseline stage.
- Split each stimulus data is split into *"n"* equal parts.
- Two types:
  - **Mean and standard deviation (MSD):** splits represented by mean and standard deviation.
  - **Slope and intercept (SI):** splits represented by slope and intercept.

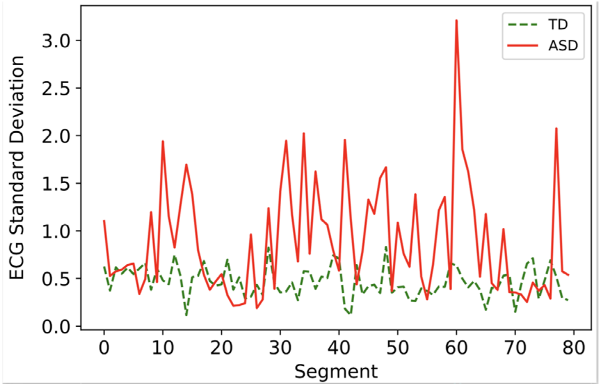
# Feature Extraction Methods

MSD representation:




**Mean and Standard deviation values for the autistic subject is higher and shows more variation than a TD subject.**

# Feature Extraction Methods

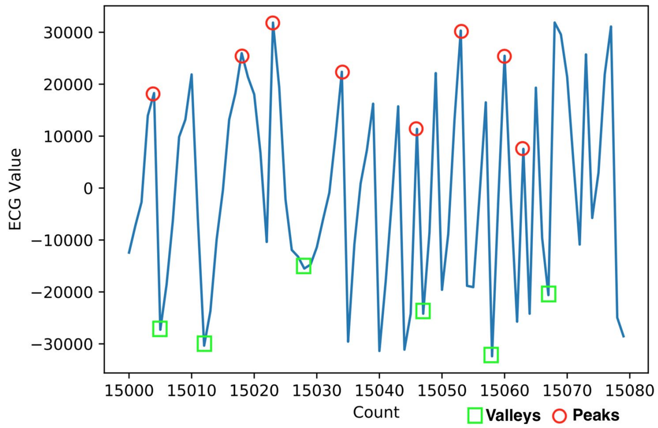
**Slope and Intercept (SI) representation :**

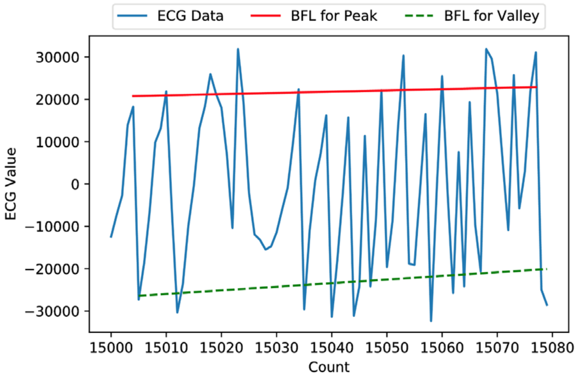
**Maximum (peak) values:** greater than the value of its two neighboring data points.

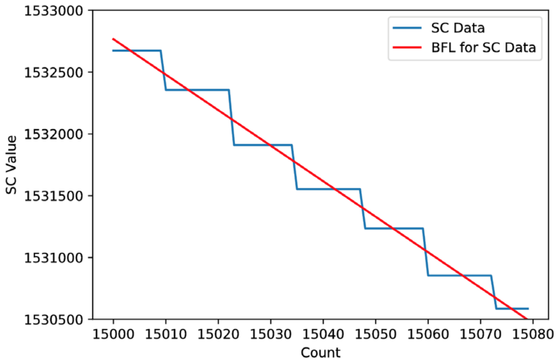
**Minimum values (valleys):** lower than the value of its two neighboring data points.

Slope → variation in trend

Intercept → intensity of signal

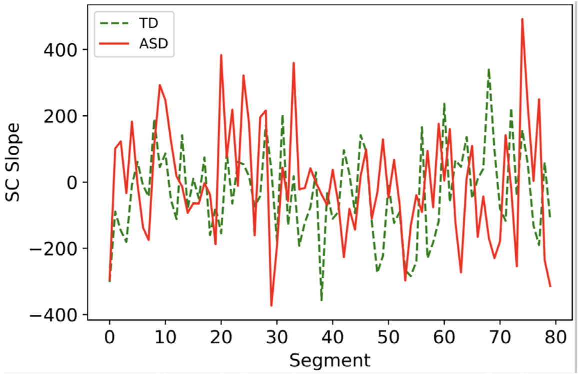


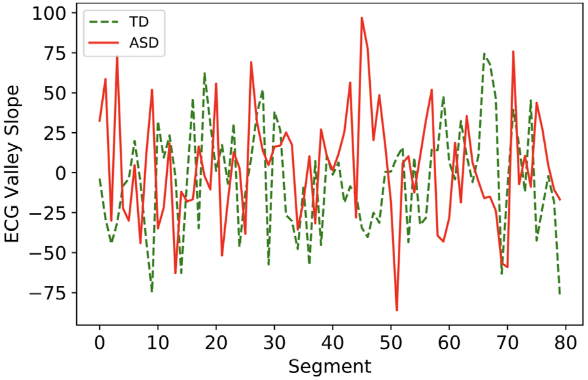



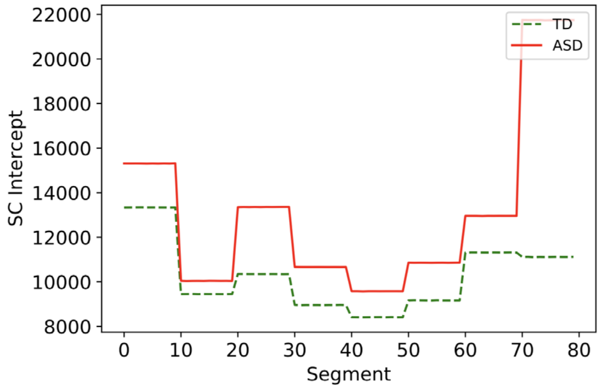



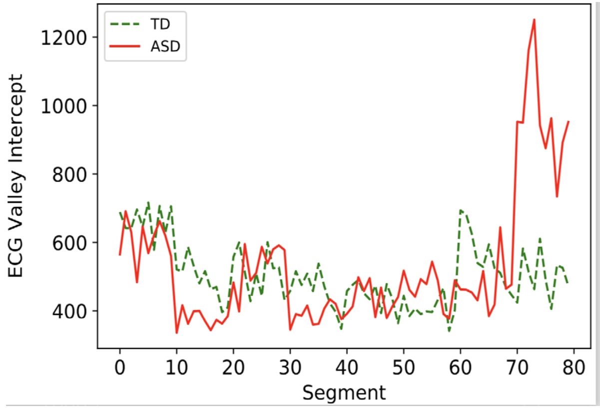
# Feature Extraction Methods

SI representation



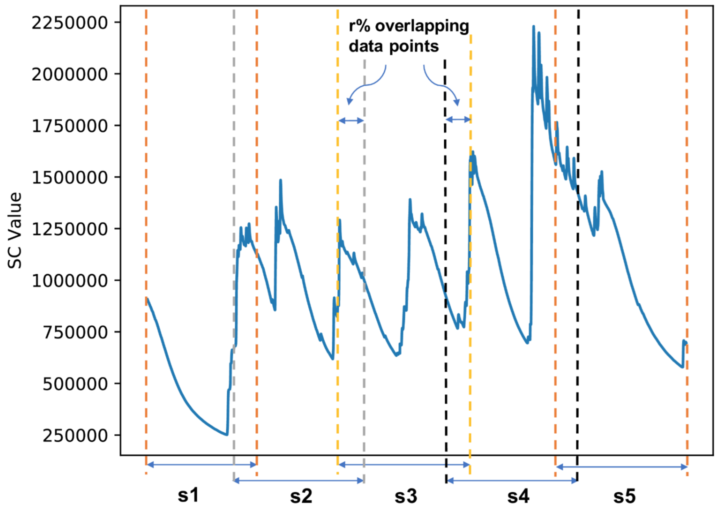




**Higher variation and intensity in autistic child**

# Feature Extraction Methods

2. Dynamic Time Warping (DTW):

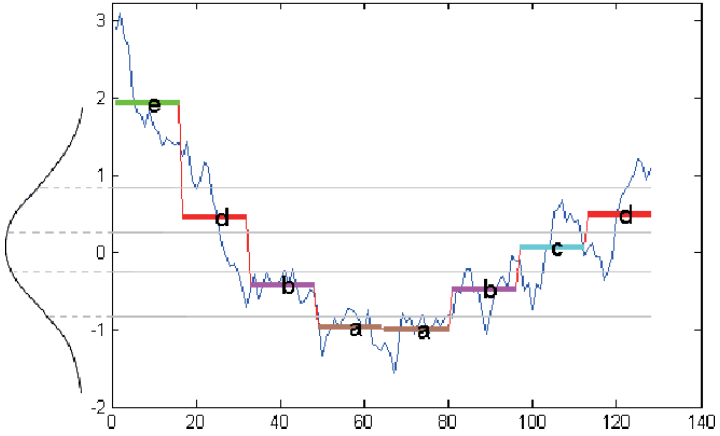
- Calculate the DTW Euclidean distance between ECG and SC data.
- **Challenge:**
  - Large size 2D matrix computation,
- Pairwise distances are used to create a KNN-based machine learning model.



# Feature Extraction Methods

3. Symbolic Representation of Time Series:

- Time series represented using Symbolic Aggregate approXimation (SAX)
- Compute pairwise Euclidean DTW distances, between all the subjects.
- Create a KNN-based machine learning classification model to predict autism in children.



# Methods for Developing Prediction Models

- 10-fold cross-validation
- Using only ECG data, or only SC data, or both ECG and SC data.
- Eight different types of models:
    - Decision Tree (DT),
    - K-Nearest Neighbor (KNN),
    - Support Vector Machine (SVM)
    - Naive Bayes (NB),
    - Random Forest (RF),
    - XGBoost (XGB),
    - DTW-based KNN (DTW-KNN), and
    - SAX-based KNN (S-KNN).

**Using MSD and SI features**

- **Base Models:** Separate models created for different stimuli.

# Methods for Developing Prediction Models: Ensemble Models

- **Majority Votes:**
  - Final prediction -> majority predicted class
  - Sameweight to all models.
- **Weighted Prediction :**
  - Weight → probability of prediction
  - +ve value → TD; -ve value → ASD

$$y_w = \sum_{i=1}^{8} p_i w_i$$

- **Stochastic Gradient Descent (SGD)**
  - Linear combination of the weight vector and predictions from different stimulus models.
  - SGDto learn weight vector .

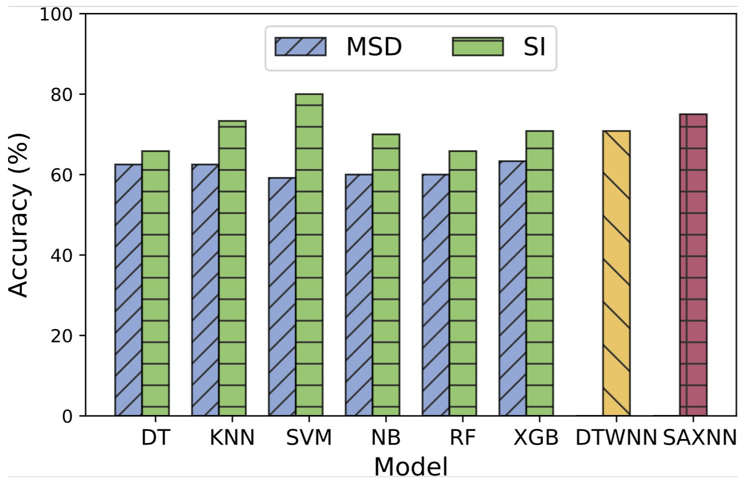
# Effectiveness Results : Base Models

- The models created using SI features perform better than those created using MSD features.
- SAX features perform better than DTW distances.

BEST BASE MODEL ACCURACY VALUES USING EACH STIMULUS

| | Accuracy(%) | Model | Data Used |
|---|---|---|---|
| Baseline | 75.83 | SAXNN | SC |
| Auditory (Tones) | **80.00** | SVM | Both |
| Visual | **80.00** | XGB | SC |
| Auditory (Siren) | 77.50 | RF | ECG |
| Olfactory | 77.50 | SAXNN | SC |
| Tactile | 74.17 | SAXNN | SC |
| Vestibular | 78.33 | RF | Both |
| Recovery | 73.33 | SAXNN | Both |

Best performing Base model: 80.00%




Base model comparison for Auditory (Tones)

# Effectiveness Results : Ensemble Models

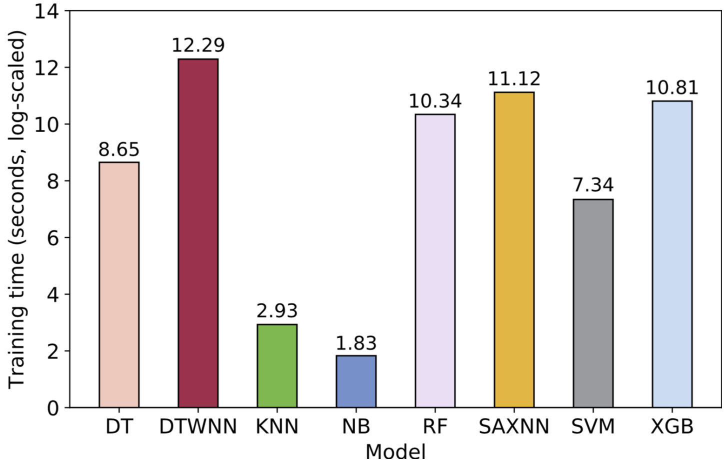
- SI features perform better than those created using MSD.
- Best model: XGB using SI features and SGD→ 93.33% accuracy→ 13.33% more than base model
- Best accuracy achieved by the DTWNN models was 77.50%.

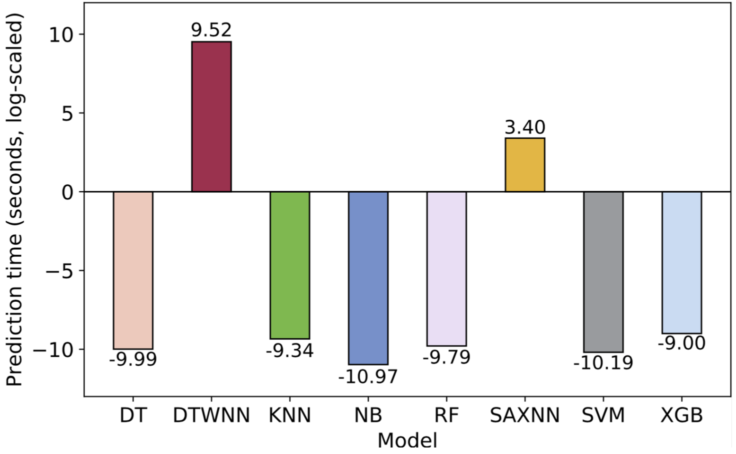
BEST ENSEMBLE MODEL ACCURACY VALUES

| | Accuracy(%) | Ensemble Type | Data Used |
|---|---|---|---|
| DT | 92.50 | SGD | Both |
| KNN | 81.67 | SGD | SC |
| SVM | 87.50 | SGD | Both |
| NB | 88.33 | SGD | SC |
| RF | 89.17 | SGD | Both |
| **XGB** | **93.33** | **SGD** | **Both** |
| DTWNN | 77.50 | SGD | Both |
| SAXNN | 92.50 | SGD | ECG |

# Efficiency Results

- XGB: 49,300 sec to train, 1.23e−4 sec to predict (third slowest).
- DTWNN: 4.4 times longer to train, $10^8$ times longer to predict in comparison to the SAXNN..
- DT → second highest accuracy (92.50%) → predicts 7 times faster than XGB.



# Future Work

- Plan to work on developing additional time series-based analysis techniques to characterize the SNS and PsNS changes over the time of the SCP.
- An unsupervised optimization procedure will be used to automatically identify prototypical SNS and PsNS states (protos).
- Then, nearest neighbor classification models can be built using this proto sequence representation .

# Conclusions

- Autistic children are affected to a higher degree by some stimuli as compared to TD children and take longer to recover.
- The feature extraction methods we developed were both effective and efficient in analyzing multivariate time series with over 2 million values.
- XGB-based model using SI features achieved the best performance (93.33% accuracy) taking only a millisecond to predict samples.
- While DTW is one of the best approaches to compare time series data in general, it does not perform well when working with very large time series data.

# Questions?

# Thank You!