College Life is Hard! - Shedding Light on Stress Prediction for Autistic College Students using Data-Driven Analysis

Tanzima Z. Islam[1], Philip Wu Liang[2], Forest Sweeney[2], Cody Pragner[2], Jayaraman J. Thiagarajan[3], Moushumi Sharmin[2], Shameem Ahmed[2]

1 Texas State University
2 Western Washington University
3 Lawrence Livermore National Laboratory

Autism in USA: Some Statistics

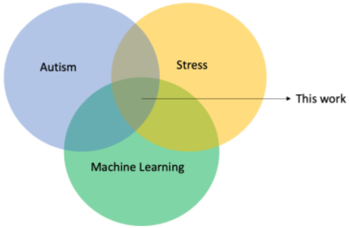
Autism Prevalence in USA	1 in 54
Average to Above-average IQ	44%
Post-Secondary Enrollment	43.9%
Graduation Rate	38.8% (~60% for neurotypical)

What does it mean?

Primary Challenge of an Autistic College Student

Our Goal

- To create a computational model for predicting the onset of stress in autistic college students
- Features: biomarkers collected using commercially available hardware
- In this research, we investigate
    - Heart Rate
    - Sleep
    - Step Count
    - Sound Intensity
    - Light
    - EMA (Ecological Momentary Assessments)



Others vs This Research

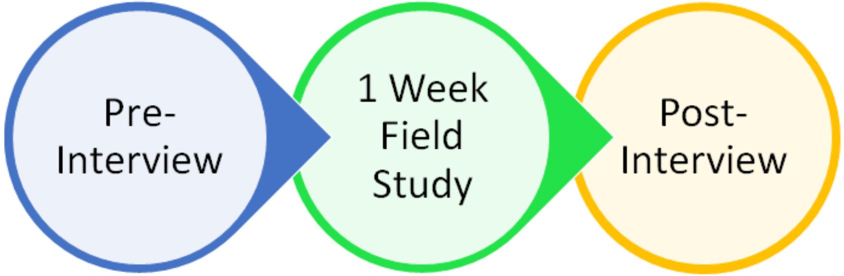
Characteristics	Others	This work
Linear models	☑	☑
Use of complex models	✕	☑
Automated feature extraction	✕ Manual and ad-hoc feature selection. Correlation among features degrades the accuracy of downstream analysis.	☑ Automated feature compression using Information Sieve. Removes inter-correlation from features, which improves the accuracy of downstream analysis.
Need for stress labels	High. Each sample needs to be annotated with stress labels. Hence, for 1000 labeled samples, one would need to interview 1000 participants or query one participant 1000 times.	This work proposes a semi-supervised learning method to estimate stress labels for samples that do not have them. This approach removes the need for 1000s of interviews.

Our Approach

- Collect biomarkers (Data Source)
- Data Preparation
- Feature Engineering
  - Automated
  - Unsupervised learning
- Modeling
  - Predictive modeling using supervised learning
  - Label propagation using semi-supervised learning
- Visualization

Data analysis pipeline

Data Collection Methodology



- **10 Autistic Students**
- **10 Neurotypical Students**

Overall college experiences (Subjective)

Continuous, objective, environmental, physiological, experience data (almost unobtrusive)

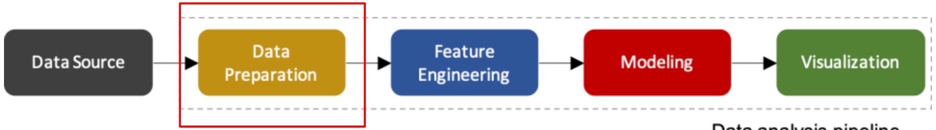
Data verification, in-depth contextualization

Features

- **1,737,625** units of heart rate
- **318,863** units of geo-location
- **315,345** units of step count
- **1,146.3 hours** of sleep (146 days)
- **170,801** units of light data (brightness)
- **1,402,779** units of sound data (amplitude)
- EMA (Ecological Momentary Assessment)
    - Answering **5,336** individual questions

Data Preparation

- Align data across various tools
- Compute the minimum, average, maximum, and standard deviation of features within a configurable interval
    - Use these measures as metrics
    - Interval = 5 mins, 30 mins, 60 mins
- Remove data that are 3-standard deviations away from the mean



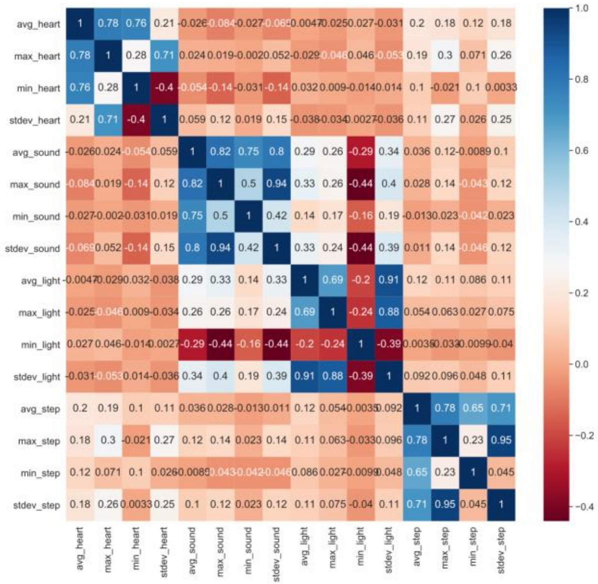
Data analysis pipeline

Observation: Feature aggregation in a 5 minute interval captures the most important information

Feature Engineering

- Observation 1: Features are correlated, which causes downstream analysis to be less accurate

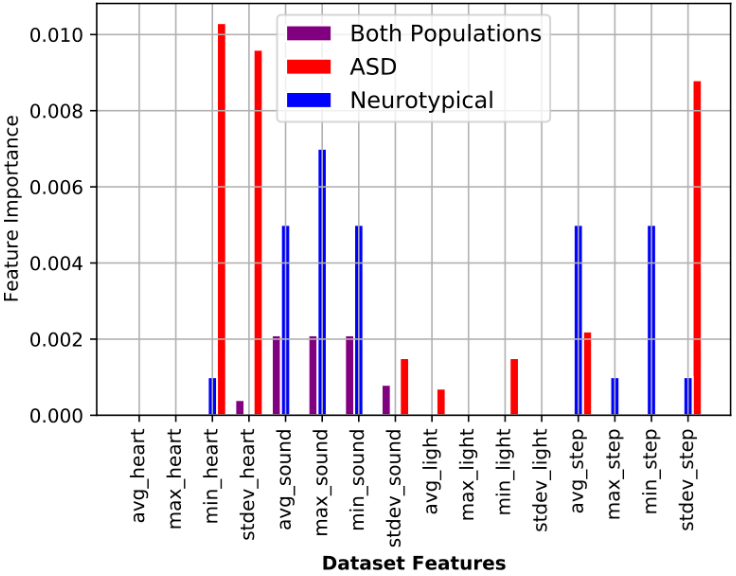
Data analysis pipeline



Feature Engineering

- Features are correlated, which causes downstream analysis to be less accurate
- Observation 1: Some features are more important than others in predicting stress

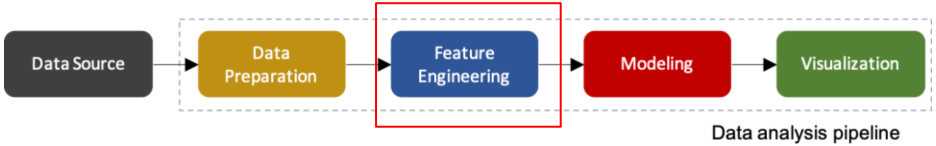
Data analysis pipeline

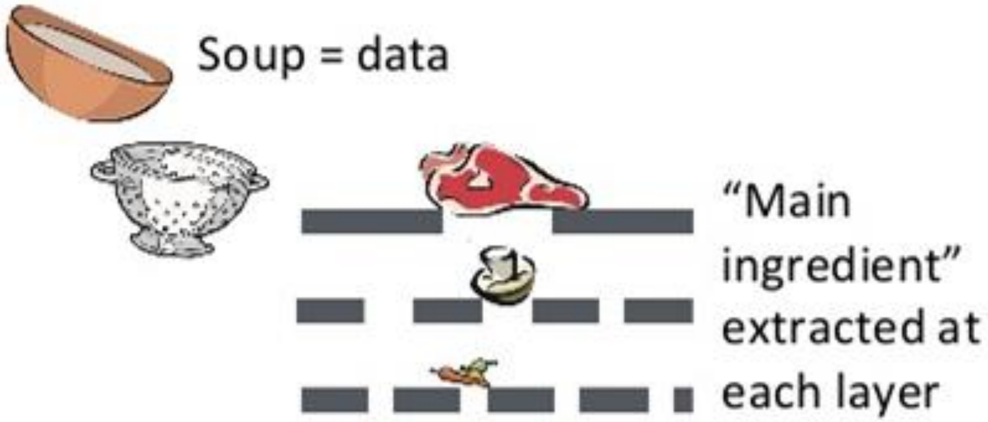


Manual feature selection is ad-hoc, labor-intensive, and error prone

Feature Engineering

- Features are correlated, which causes downstream analysis to be less accurate
- Observation 1: Some features are more important than others in predicting stress
- Our solution:
  - Iteratively extract common information and create latent features.
  - Remove common information from all features to create residuals.





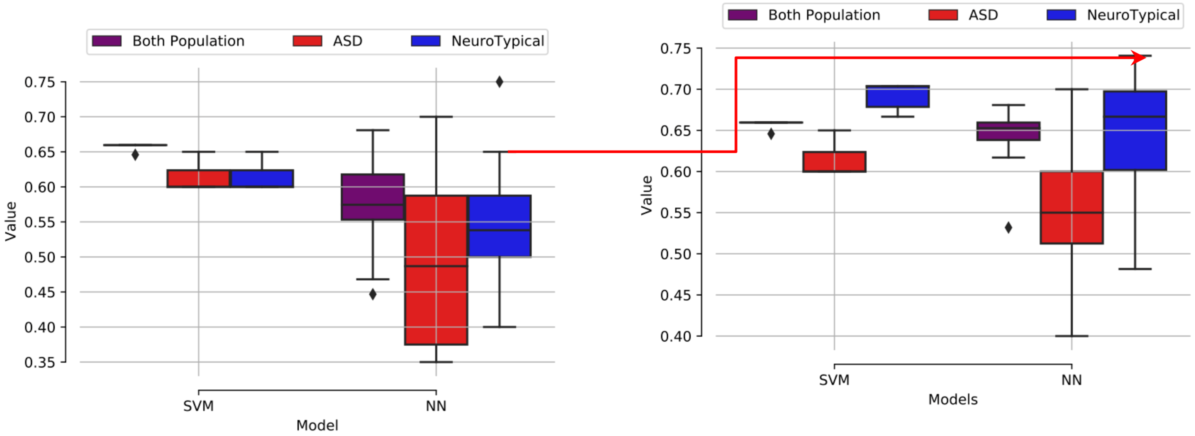
Information sieve, Steeg et. al., ICML'16

Feature Engineering

- Features are correlated, which causes downstream analysis to be less accurate
- Observation 1: Some features are more important than others in predicting stress
- Our solution:
    - Feature extraction using Information Sieve
- Observation 2: Automated feature extraction captures salient characteristics that improve model accuracy.

Data analysis pipeline

Up to ~9% improvement



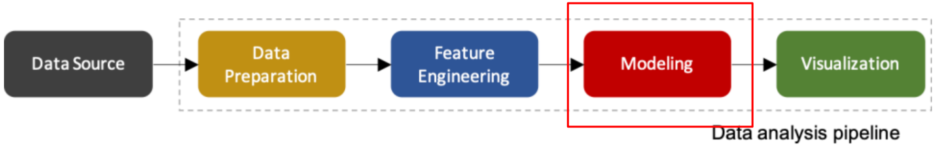
Manual feature selection

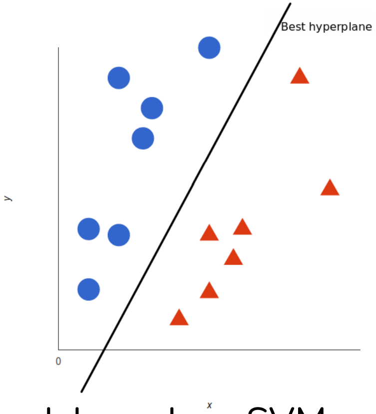
Automated feature selection

Modeling -- Predictive

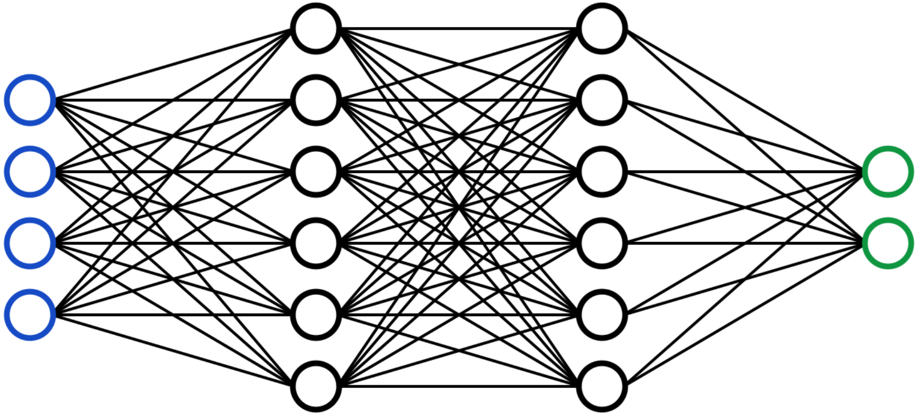
- Both linear and non-linear modeling techniques
- Target -> stress label (stressed or not)

f(features) = target





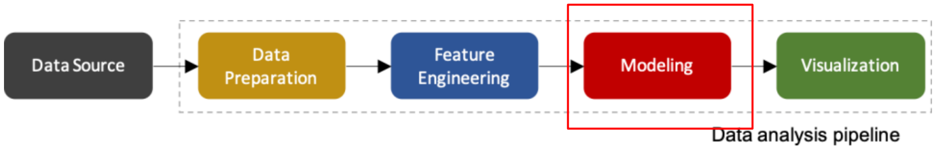
Linear models such as SVM are good for features that have linear relationship with the target



Complex models such as neural network is able to predict complex relationships

Modeling -- Generative

- Goal: Automatically infer stress labels for samples that are not labeled
- Evaluate based on the labeled samples alone



Labeled and unlabeled samples → KNN → Build neighborhoods with both labeled and unlabeled samples → CAMLP → Inferred labels for the unlabeled samples such that error is minimized

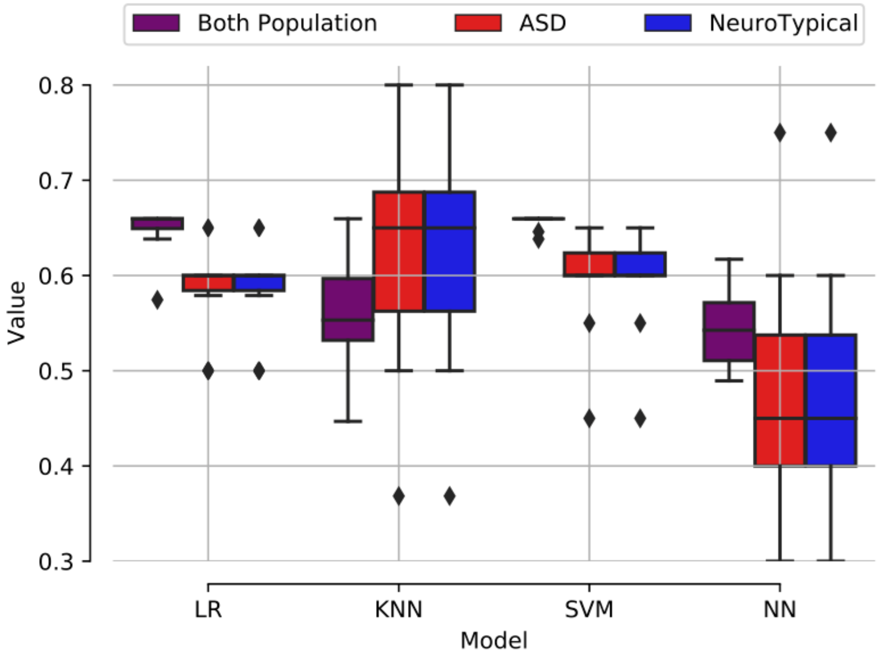
Euclidean distance

Error calculated based on how many originally labeled samples are correctly labeled by the algorithm

Experimental Setup

- Jupyter notebook for data analysis
- Results presented in this talk:
  - Which model performs better?
  - Does adding new samples help?
  - Feature extraction using Information Sieve improves model performance (presented earlier)
  - Overhead of analysis -- negligible

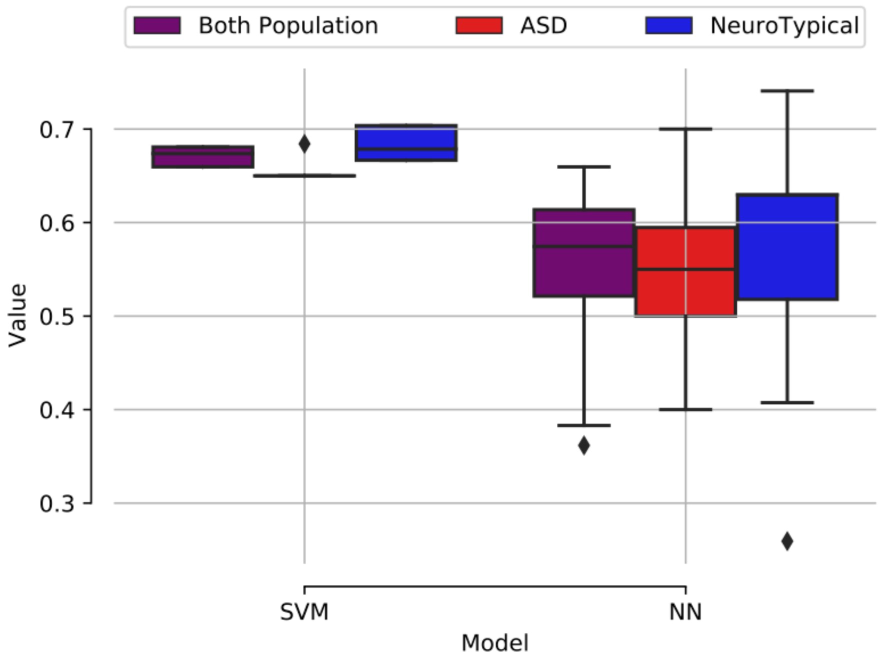
Result 1: Which Model Performs Better?



Observation: SVM performs the best with less variability and higher average accuracy

Implication: Collected biomarkers have simple linear relationship with each other. Hence, NN overfits

Result 2: Usefulness of generated labels in predictive modeling



Observation: Accuracy improves by 7% and 15% for SVM and NN, respectively, compared to the baseline (using 471 labels).

Implications:

- This approach can increase the number of samples and improve model accuracy.
- Data can now be used for deep learning, which requires a lot of labeled samples.

Impact and What Next?

- Neurodiverse students live in fear and anxiety
- Better prediction leads to more accurate detection and faster intervention
- Faster intervention leads to more control over situation for neurodiverse students
- On-time help can make a positive difference in the lives of neurodiverse college students