Improving Feedback in Massive Open Online Course (MOOC) Learning through EEG Analysis

Amanda Breton • Alex Bzdel • 05.04.2023

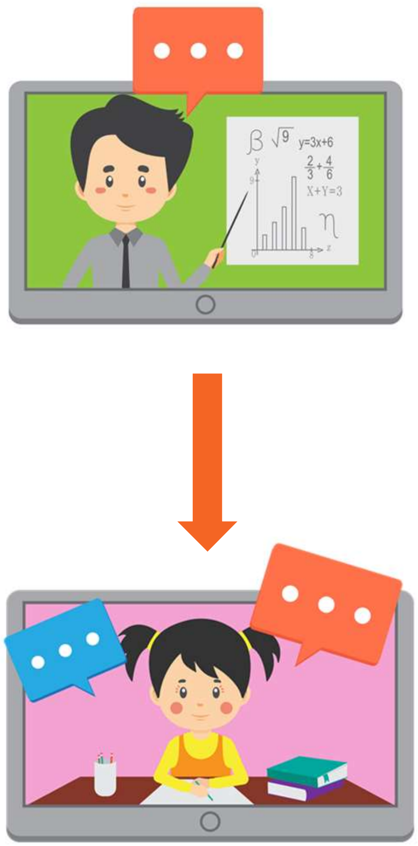
Overview

Original Paper Introduction

Wang, Li, et. al 2011

Goals

- Determine reproducibility
- Improve results with different model(s)
- Improve preprocessing and feature selection



Original Paper

Goals

- Determine if EEG can detect confusion
- Determine if EEG can detect confusion **better than human observers**
- End goal: provide feedback about student confusion level during remote learning

Original Paper - Setup

Design & Data Collection

- 10 students wore a **single-channel MindSet headset**
- Watched videos *assumed to be* confusing or not confusing
    - Confusing = quantum mechanics, stem cell research
    - Not confusing = geometry, algebra
- Self-reported confusion on **scale of 1-7**
    - Videos were also predefined as confusing or not confusing *(second target variable)*

Original Paper - Models

Gaussian Naive Bayes Classifier

- Good for **sparse & noisy** training set
- Two targets (predefined/student-defined confusion)
- Used various features captured from EEG data
  - Not much said about dimensionality reduction or feature selection
  - No scaling of data

Original Paper - Models

Student Specific

- Single student as dataset
- Training on half the student's videos
- Testing on the other half of the student's videos

Student Independent

- Leave one out cross validation:
    - Training on all but 1 student
    - Testing on the left out student

Original Paper Results

Student Specific:

- 67% pre-defined confusion
- 56% user-defined confusion

Student Independent

- 57% pre-defined confusion
- 51% user-defined confusion

Room for Improvement

Data

- **No feature selection**
- Student-defined confusion is convoluted
- **No data standardization**
- VERY small dataset + one sample corrupted
- Age range narrow (24-31)

Model Selection/Usage

- **No testing of various models**
- **No grid search or hyperparameter tuning**

How do we improve??

Data Collection

4 NB Models:

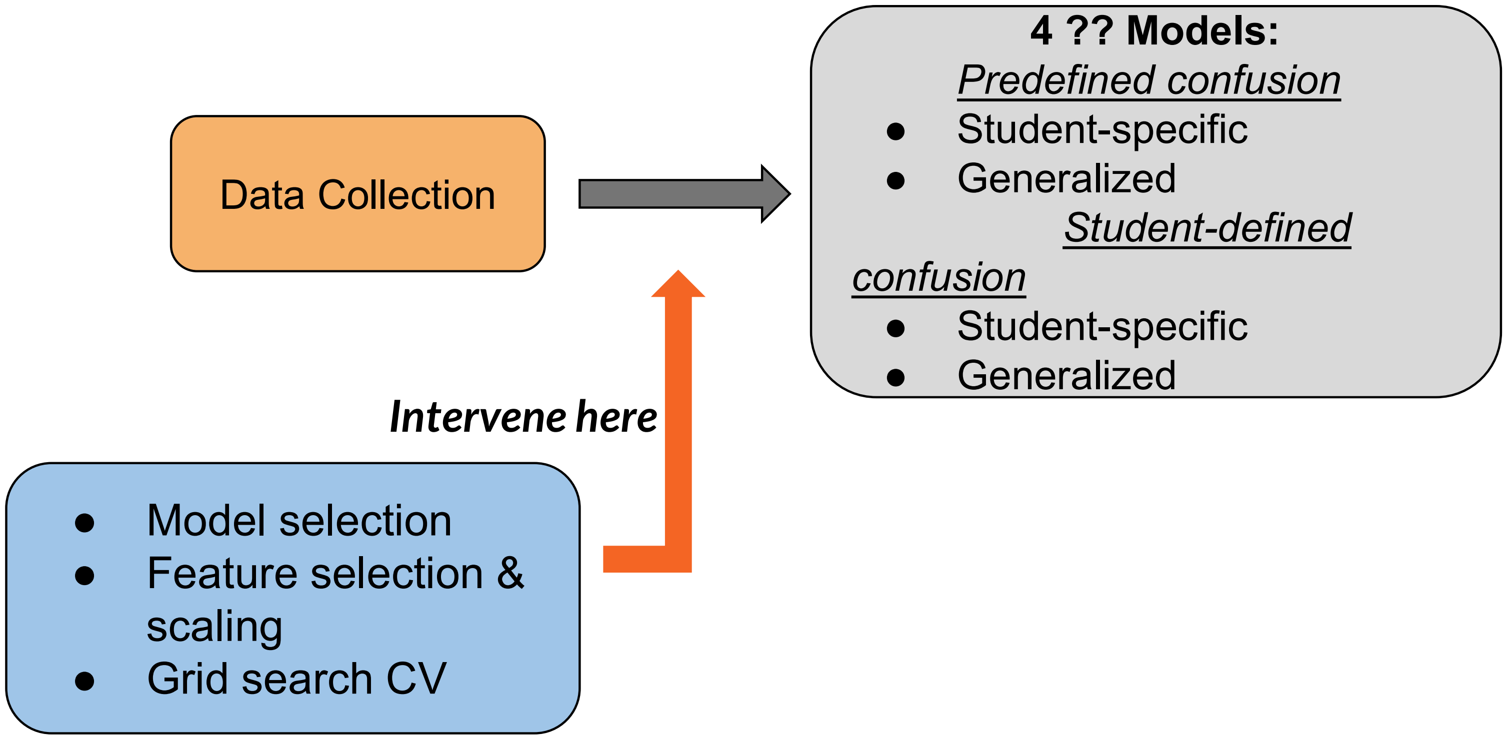
Predefined confusion

- Student-specific
- Generalized

Student-defined confusion

- Student-specific
- Generalized

How do we improve??



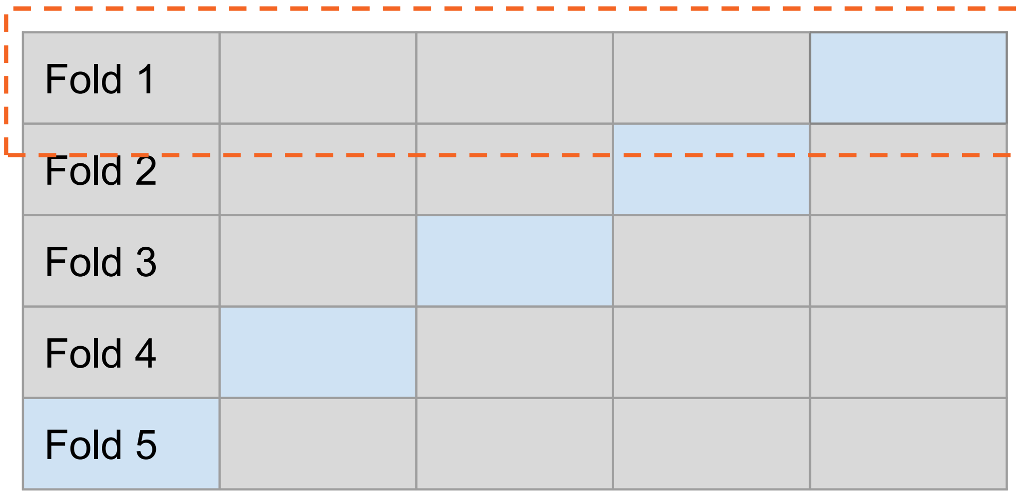
Improving Model Selection and Usage

Models Tested

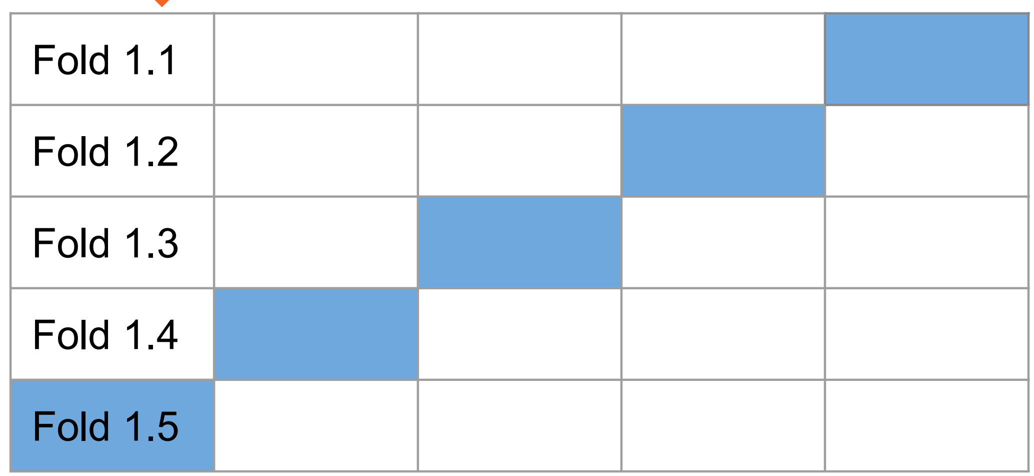
- Logistic regression
- K Nearest Neighbors
- Support Vector Machine
- Random Forest
- Decision Tree

- Why these models?
- Considered the best for classification problems
- Compare their performance with naive bayes

Cross Validation



- First Level: improve accuracy
- 2nd Level: grid search to find optimal parameters



Training Patients	Testing Patient
Grid Search Training Data	Grid Search Testing Data

Model Results - Student Specific

Model	Average Pre-defined Confusion Label Accuracy	Average User-defined Confusion Label Accuracy
Their Naive Bayes	67%	56%
Naive Bayes	55%*	73%
Logistic regression	49%*	65%
K Nearest Neighbors	51%*	79%
Support Vector Machine	41%*	71%*
Random Forest	57%	80%
Decision Tree	58%*	83%*

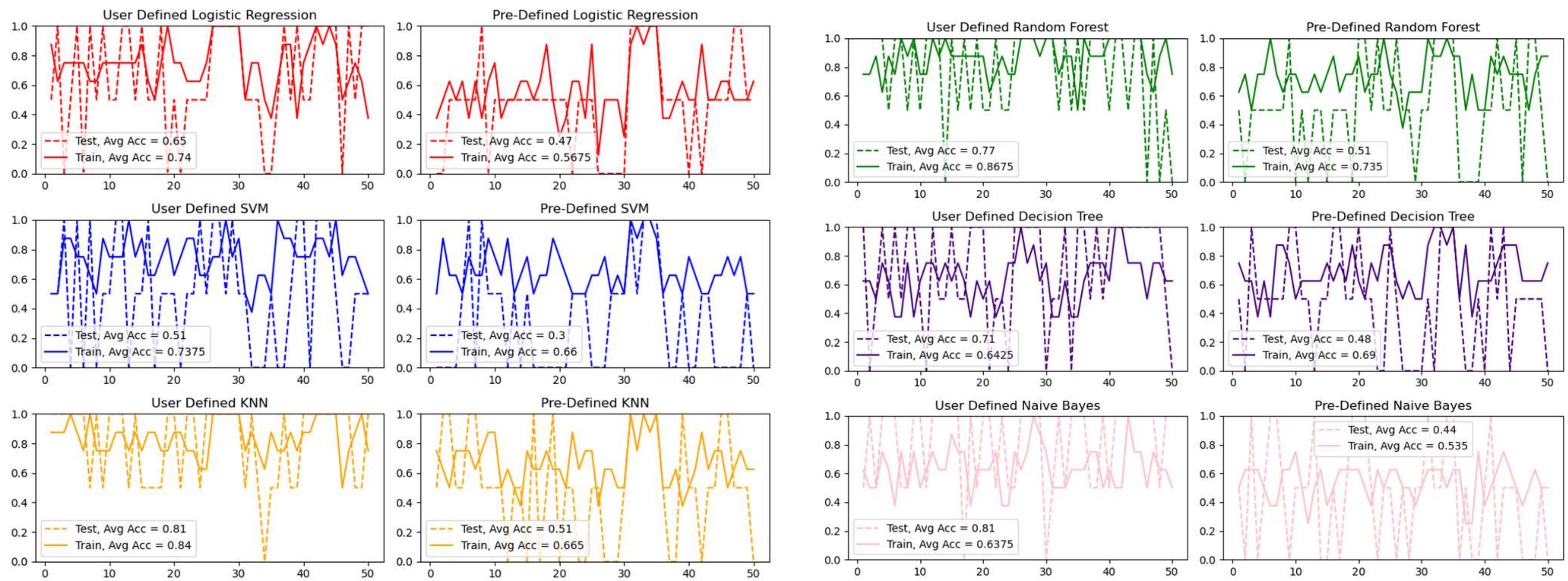
** = Model improved by a combination of feature selection + scaling*

Model Results - Student Independent

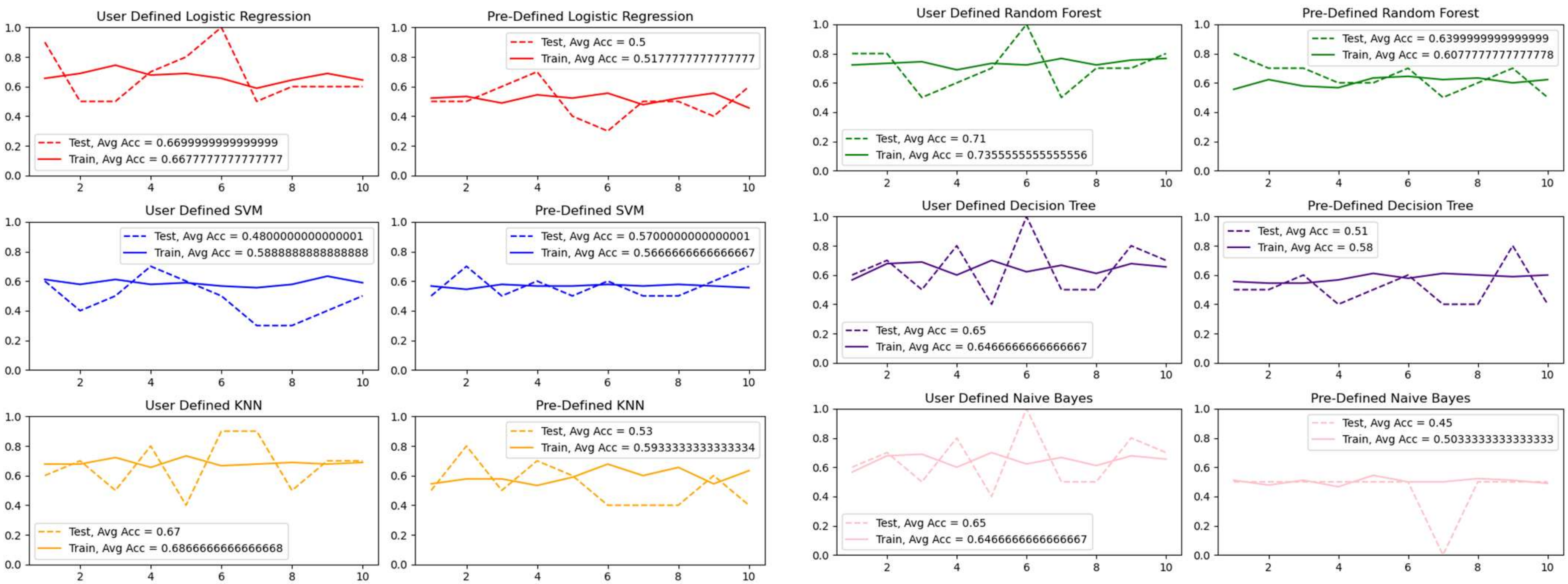
Model	Average Pre-defined Confusion Label Accuracy	Average User-defined Confusion Label Accuracy
Their Naive Bayes	57%	51%
Naive Bayes	50%*	67%*
Logistic regression	50%	67%
K Nearest Neighbors	65%*	70%*
Support Vector Machine	52%*	65%*
Random Forest	65%*	74%
Decision Tree	62%*	72%

** = Model improved by a combination of feature selection + scaling*

Evaluating Overfitting and generalizability: Student Specific



Evaluating Overfitting and generalizability: Student Independent

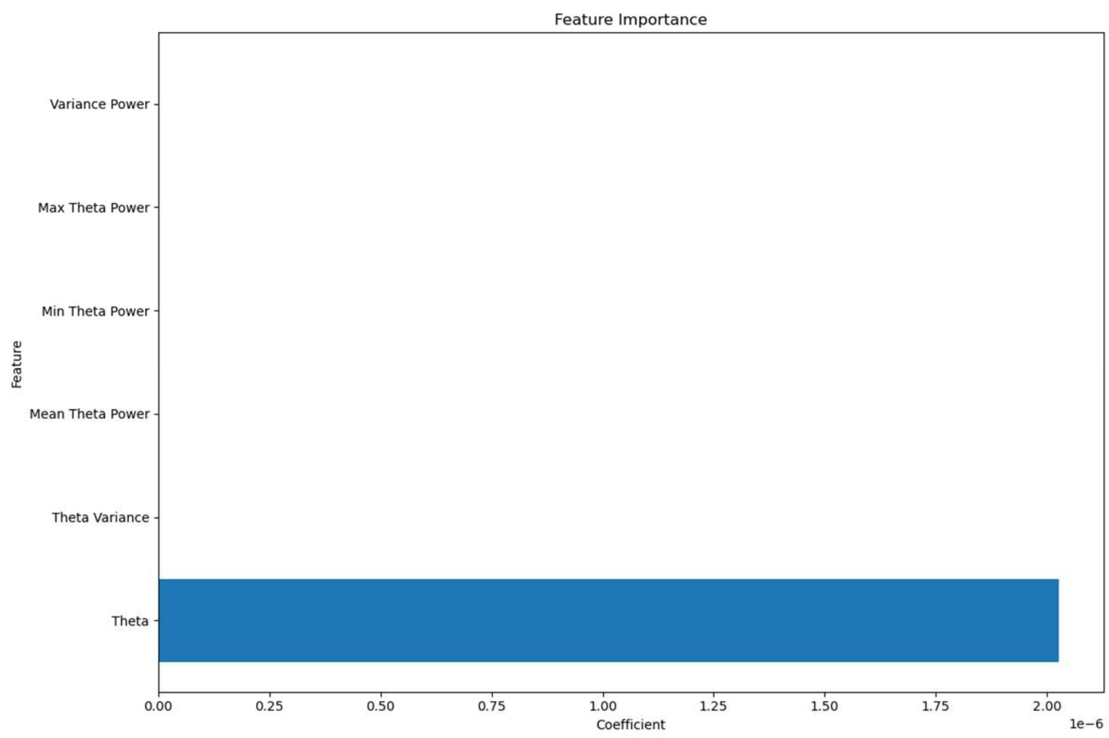


Improving Data

How can we improve the data?

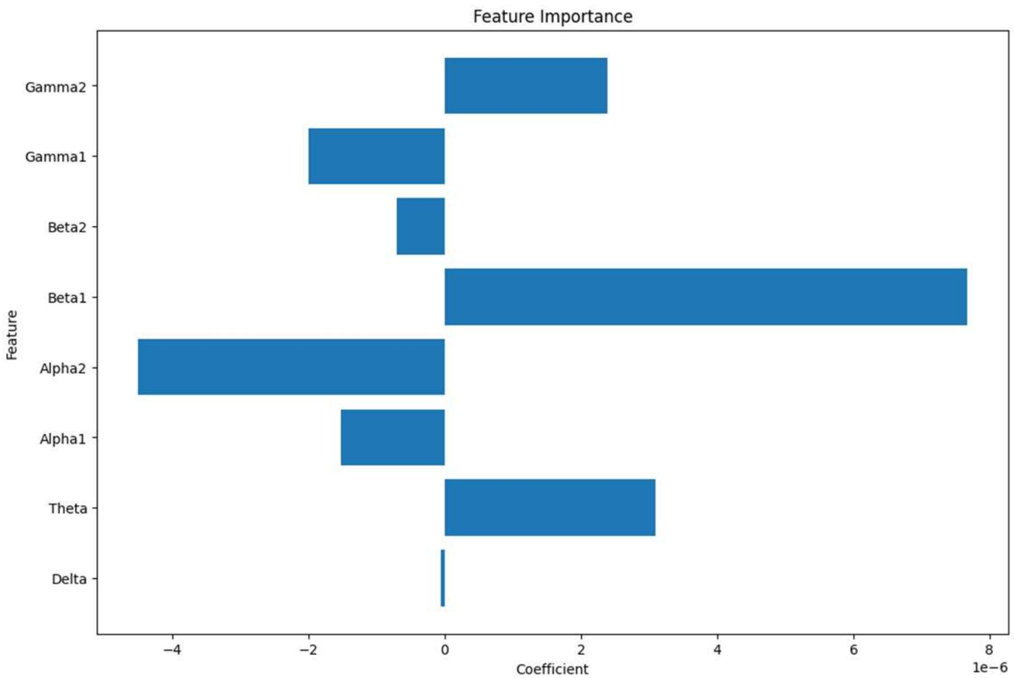
- Paper's researchers speculated that theta signal played important role
  - In neuroscience theta wave correlated with :
    - Memory, learning and spatial navigation
  - Can we generate more useful features from the theta band?
- Can we improve the features used in the models with better/more feature engineering?
- Will data normalization help with performance and overfitting?

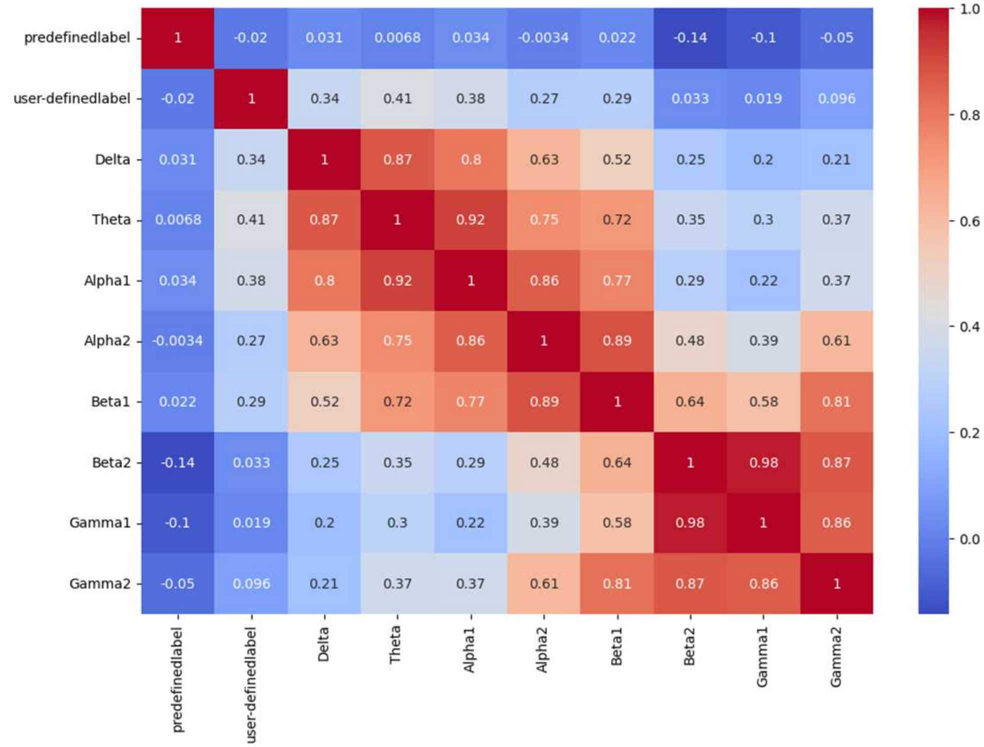
Investigating Theta Features



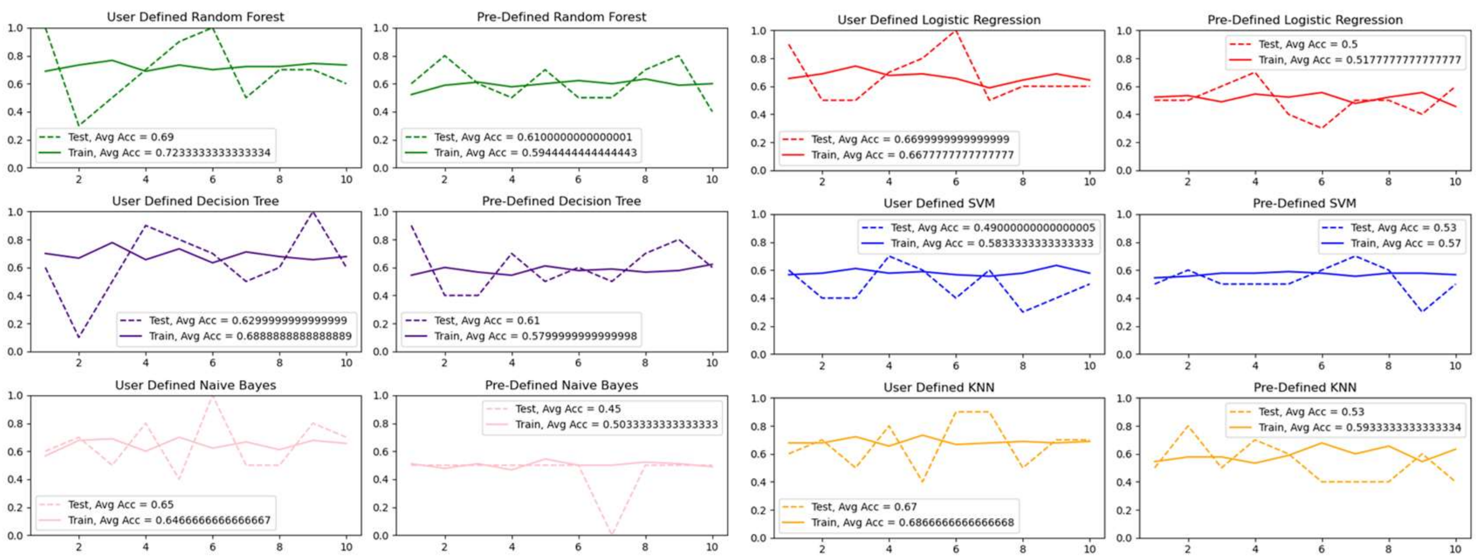
- Generated statistical features from theta band
- Performed LASSO Regression fit to determine if feature were important in classification
- **Theta features not significant**

Investigating Feature Importance





Does data regularization improve model performance and overfitting?



- Improved model performance?
  - Yes
- Improved fit?
  - Yes

Conclusions and Discussion

Best Performing Model(s)

- Model:
  - Student Specific: Decision Tree (27% inc.)
  - Student Independent: Random Forest (23% inc)
- Means of channels are features
- Regularization of data? Yes

Best Performing Model + Data Combo: Results

Student Specific:

- % pre-defined confusion
- % user-defined confusion

Student Independent:

- % pre-defined confusion
- % user-defined confusion

How we would design the for MOOC Feedback

Modeling:

- Label: student/user defined confusion
- Model: Random Forest
- Data normalization

Data Collection:

- Collect more data
- Take into account major(s) of students for video selection
- Wider age range of students (avg 27.9)

Thank you!
Any Questions?