Discriminating Cognitive Disequilibrium and Flow in Problem Solving: A Semi-Supervised Approach Using Involuntary Dynamic Behavioral Signals

Mononito Goswami,[1,2*] **Lujie Chen,**[2*] **Artur Dubrawski**[2]
[1]Delhi Technological University, New Delhi, India
[2]Auton Lab, Carnegie Mellon University, Pittsburgh, USA
mononito_bt2k16@dtu.ac.in, lujiec@andrew.cmu.edu, awd@cs.cmu.edu

Abstract

Problem solving is one of the most important 21st century skills. However, effectively coaching young students in problem solving is challenging because teachers must continuously monitor their cognitive and affective states, and make real-time pedagogical interventions to maximize their learning outcomes. It is an even more challenging task in social environments with limited human coaching resources. To lessen the cognitive load on a teacher and enable affect-sensitive intelligent tutoring, many researchers have investigated automated cognitive and affective detection methods. However, most of the studies use culturally-sensitive indices of affect that are prone to social editing such as facial expressions, and only few studies have explored involuntary dynamic behavioral signals such as gross body movements. In addition, most current methods rely on expensive labelled data from trained annotators for supervised learning. In this paper, we explore a semi-supervised learning framework that can learn low-dimensional representations of involuntary dynamic behavioral signals (mainly gross-body movements) from a modest number of short time series segments. Experiments on a real-world dataset reveal a significant advantage of these representations in discriminating cognitive disequilibrium and flow, as compared to traditional complexity measures from dynamical systems literature, and demonstrate their potential in transferring learned models to previously unseen subjects.

1 Introduction

One of the fundamental goals of education is to transform students into mature problem solvers who are able to overcome the inherent uncertainty of problems, failed attempts and impasses. For young children, solving challenging non-routine math problems emulates the real life challenges they will encounter later in their lives. Different from routine math exercises (e.g. back-of-chapter exercises), non-routine problems may not have immediate solutions, and thus require innovative thinking, and may often invite a child to ride an "emotional roller-coaster" as the student advances through various stages of problem solving (Chen et al. 2016). Problem solving is a complex affective and cognitive process replete with states of *cognitive disequilibrium* manifested by a mixture of confusion, frustration, indecisiveness or struggle, as well as states of *flow* (Csikszentmihalyi 2013) when one is (or at least is feeling of) moving forward smoothly. The cognitive disequilibrium triggered by conflicts and contradictions in these problem solving processes can be beneficial for learning only if appropriately regulated and resolved (D'Mello and Graesser 2014) (*Facilitative Confusion Hypothesis*), which may be challenging for an inexperienced problem solver whose self-regulation and problem solving skills are in their nascent stages.

Therefore, effectively coaching young students requires teachers to continuously monitor their cognitive and affective states and make real time pedagogical decisions such as when to intervene and how best to do so, especially in social environments with low teacher-student ratios and with limited coaching resources available for each student. Moreover, teachers also have to effectively handle the high cognitive loads of monitoring a diverse cohort students varying significantly in their perception of academic self-efficacy and ability to use of self-regulated learning strategies (Zimmerman and Martinez-Pons 1990). Intelligent Tutoring Systems that attempt to teach problem-solving also face similar challenges. To lessen the cognitive load of teachers and also to improve the effectiveness of intelligent tutoring, we envision a decision support system which can monitor the cognitive and affective states of multiple students simultaneously in real time. The focus of this paper is on the state detection capability of such a system, specifically needed to discriminate between cognitive disequilibrium (CD) and flow states, which are the critical inputs to inform appropriate subsequent interventions.

In this work, we investigate a method designed to discriminate between CD and flow using involuntary behavioral signals that are less prone to social editing, including head and eye movement, which can be non-invasively collected using inexpensive sensors such as cameras. To overcome limited supply of labeled data, while taking advantage of the large supply of unlabeled data, we explore a semi-supervised approach where deep embedding features are derived from unlabeled time series segments, which are then

*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

fed into a supervised learning algorithm. We compare these deep features with a set of baseline complexity measures discussed in dynamical systems literature and note significant improvement in predictive power. Furthermore, our experiments confirm that our semi-supervised model performs reasonably well even with very limited amount of data.

The rest of the paper is organized as follows. Section 2 provides background of our study by discussing its motivation and relation to prior work. Section 3 describes the data collection, methodology, and experiments in detail. Section 4 discusses experimental results, and Section 5 explores their implications. We conclude the paper and present avenues of future work in Section 6.

2 Background and Related Work

Cognitive Disequilibrium and Flow in Problem Solving

In the last few years, the research community has shown keen interest in the affective and cognitive dimensions of learning (D'Mello 2013). Studies such as (D'Mello and Graesser 2014) have shown that children may get *confused* when they are unsure of how to proceed or face challenging impasses. They may also get *frustrated* when they repeatedly make mistakes or important goals are blocked (Kapoor, Burleson, and Picard 2007). At the same time, students may also experience *delight* when they achieve their goals by overcoming problems or enter a *flow* state of intense engagement when the learning goals as well as problem solving paths are clear, and they find an appropriate balance between skills and challenge (Csikszentmihalyi 2013).

In the practice of problem solving education, a teacher would be broadly interested in two cognitive states of the student: (1) *Cognitive Disequilibrium*, characterized by confusion, frustration and indecisiveness, and (2) the *Flow* state, characterized by smooth progression toward the goals. In this paper, we only consider two distinct and broad cognitive states of primary interest to teachers, and therefore we also attribute positive emotions such as curiosity and happiness to the flow state, and pose the problem of detecting cognitive states of a student as a binary classification problem. (D'Mello 2013) previously found that flow, confusion and boredom were the most frequent affective states found in studies employing learning with technology, in support of our choice of the cognitive-affective states to consider.

Automated Detection of Cognitive-Affective States

The problem of identifying cognitive and affective states of students is challenging since these states are loosely-defined psychological constructs embedded in extremely context-sensitive environments (D'Mello and Kory 2015). Many studies have investigated the possibility of detecting affective states automatically, primarily in the context of Intelligent Tutoring Systems. For instance, (Joseph 2005) proposed *Engagement Tracing* to detect the engagement levels of students based on audit logs from an intelligent tutor. Later, (McDaniel et al. 2007) investigated the relationship between facial features and emotions such as confusion, frustration and delight, and found important patterns in the way that learners communicate their emotions through their faces. Most recent literature on affective and cognitive computing has focused on the use of multimodal features. These multimodal affect classifiers have also been shown to be consistently better than their unimodal counterparts (D'Mello and Kory 2012). (D'Mello and Graesser 2012) in their affect-sensitive AutoTutor combined decisions from conversational cues, gross body language and facial feature tracking in order to track the affective and cognitive states of students. (Hussain et al. 2011) used multi-channel physiological signals such as heart-activity, skin conductivity and respiration to detect the learner's affective states during their interaction with AutoTutor. While many of these classifiers have achieved impressive performance, one major limitation is their reliance on data from *expensive* and *intrusive* sensors to monitor body posture (Body Posture Management System), electrocardiogram (ECG), etc. *The expense of these sensors coupled with their intrusive nature preclude their deployment at scale, in common classrooms, and in less developed communities.* Furthermore, many studies have used facial expressions and vocal features as indicators of affect (Camras and Shutter 2010). While facial expressions are widely considered as a language of emotion (Ekman 1994), many studies such as (Kilbride and Yarczower 1983) have highlighted that culture and ethnicity may influence the recognition of emotion by facial expressions. Furthermore, most affective classifiers are trained using supervised machine learning and require a sufficient supply of labeled *"ground-truth"* data from experts and self-reports. Obtaining labeled data free from cultural, reference (Heine et al. 2002) and social desirability (Krosnick 1999) is very hard. In such a scenario, unsupervised representation learning methods may be handy, since they do not require training data and may learn useful features. We illustrate the feasibility of semi-supervised models in the cognitive state detection pipeline through our experiments later in the paper.

The Expressive Power of Gross Body Movements

Many researchers have investigated the role of facial expressions, speech patterns and physiological responses, as indices of cognitive and affective states. (D'Mello, Dale, and Graesser 2012) also pointed out that owing to the numerous degrees of freedom, the human body is a potentially ideal affective communication channel. However, only few studies have focused on *gross-body movements* as predictors of cognitive and affective states, which is surprising due to the embodied nature of affect and cognition (D'Mello, Dale, and Graesser 2012). Gross-body movements are promising predictors of cognitive states because they are mostly involuntary and therefore less prone to social editing in comparison to vocal and facial features. In addition, the human body owing to its large number of degrees of freedom forms a rich affective communication channel (D'Mello, Dale, and Graesser 2012). Existing research utilizing gross body movements as an index of affect, has mostly focused on gestures and specific postures (Coulson 2004), and relied on expensive sensors such as Body Posture Measurement Systems (DMello et al. 2008), which are hard to deploy at large scales in practice. A few years ago, (D'Mello,

Dale, and Graesser 2012) established that body fluctuations in the normal state of mind (cognitive equilibrium) are characterized by *correlated pink noise*, and underwent *whitening* when their participants experienced states of cognitive disequilibrium. Inspired by the findings, we hypothesized that states of cognitive disequilibrium and flow differ in the complexity of the gross body movement signals. By considering gross body movements in addition to facial action units in the form of time series (rather than raw video logs), we ensure that our features are not only privacy-preserving, but also involuntary and therefore less susceptible to social editing. Moreover to the best of our knowledge, no existing work has been able to demonstrate the influence of culture & ethnicity on gross body movements.

3 Data and Methodology

Data Collection and Pre-processing

Our experiments are based on a dataset collected in one-to-one coaching scenarios for math problem solving. Seven children within eight to twelve years of age and their parents were recruited from a local community. Parents were asked to record videos (using a web camera) and pencast videos (using a Livescribe Smartpen) of their children solving a math problem. The cohort comprised of three girls, four boys (two girls were siblings) and their parents (two fathers and four mothers). The dataset consisted of 36 sessions having a cumulative duration of 307 minutes, with a mean duration of 7.9 minutes per session.

A number of features were extracted from the dataset along the visual and writing channels. Visual features such as Facial Action Units (FAUs), head and eye gaze orientations were extracted using OpenFace (Baltrušaitis, Robinson, and Morency 2016) at a sampling frequency of 30 Hz. We computed the first and second order derivatives of all visual features with the exception of Facial Action Units using NumPy's gradient function which approximates the gradient of an array using second order accurate central differences in the interior points and second-order accurate one sides in the end points. The writing speed was estimated from Livescribe Echo Smartpen by computing the cumulative distance covered by the tip of the pen and thereafter measuring the change in the "amount of ink" collected in a trailing window of two seconds. The final sets of features used in our study are listed in Table 1.

Ground Truth Labels

In order to validate our results, we annotated non-overlapping 10-second time series segments for states of cognitive disequilibrium or flow. We use those annotations as a proxy for "ground truth" that teachers would rely on in real time decision making. 20% of the video segments from each child were annotated by two independent annotators[1]. Each annotator labeled a ten-second window within a session based on the "perceived" cognitive state of the child, as *cognitive disequilibrium* (1), *neutral* (2), *flow* (3) or *off-task behavior* (-1). The rest of the data was then labeled by one annotator, after a satisfactory inter-rater consensus was reached with Cohen's kappa greater than 0.5.

The choice of 10-second windows was inspired by literature where a number of studies such as (D'Mello, Dale, and Graesser 2012) used fixed size windows for annotating affect. The choice of the window size was also driven by the fact that complexity measures such as Higuchi Fractal Dimension expect stationary time series as input, and while 10s windows (300 time steps at 30 frames-per-second video) are short enough to be considered stationary, they also include sufficient number of time steps to accurately compute the complexity measures. In our experiments, we only used time series segments labeled as cognitive disequilibrium and flow since both the annotators had substantial agreement (average Cohen's kappa = 0.6). From a total of 353[2] time series segments, we could only use 248[3] time series segments for our analysis. The remaining segments were shorter than 10s, too short for computing complexity measures.

Measures of Time Series Complexity

Measures of time series complexity were developed to distinguish regular, chaotic and random behavior. Measures such as Higuchi Fractal Dimension, Approximate Entropy, etc. have been widely used in bio-medical signal processing applications such as electroencephalographic time series analysis (Esteller et al. 2001) and psychology (Pincus and Goldberger 1994).

In their seminal work, (D'Mello, Dale, and Graesser 2012) found that fluctuations in gross body movements in states of cognitive equilibrium are characterized by *correlated pink noise*, and undergo *whitening* when students experience cognitive disequilibrium. White noise is characteristic of random systems having no long or short term correlations between observations, whereas pink noise exhibits both long and short term correlations (D'Mello, Dale, and Graesser 2012). Inspired by those results and the success of complexity measures in analyzing physiological time series (Kantz, Kurths, and Mayer-Kress 2012), we hypothesize that states of cognitive disequilibrium and flow may differ in complexity.

Numerous time series complexity measures have been proposed, but in our study we consider the following six most widely used: *Approximate Entropy* (Pincus 1991), *Sample Entropy* (Richman and Moorman 2000), *Spectral Entropy* (Powell and Percival 1979), *Permutation Entropy* (Bandt and Pompe 2002), *Katz Fractal Dimension* (Esteller et al. 2001), and *Higuchi Fractal Dimension* (Higuchi 1988).

In order to test our hypothesis, we conducted the two-sample Kolmogorov-Smirnov (K-S) test which compares the empirical distribution functions of two samples under the null hypothesis that both are drawn from the same underlying distribution. We carried out a total of 48 univariate two-sample K-S tests, one for each combination of 8 features (*gaze_vel_X, gaze_vel_Y, gaze_acc_X, gaze_acc_Y, head_vel_T, head_vel_R, head_acc_T, head_acc_R*) and 6 complexity measures. The two samples for the test were the

[1] The first and second authors of the paper. The second author has considerable experience in annotating similar datasets.

[2] 198 segments for Cognitive disequilibrium and 158 for flow

[3] 128 segments for Cognitive disequilibrium and 120 for flow

Features	Description	Derivation from OpenFace features
FAUs	Indicate the presence or absence of 18 Facial Action Units	AU01_c, AU02_c, AU04_c, AU05_c, AU06_c, AU07_c, AU09_c, AU10_c, AU12_c, AU14_c, AU15_c, AU17_c, AU20_c, AU23_c, AU25_c, AU26_c, AU28_c, AU45_c
gaze_vel_X	Velocity of eye gaze along X-axis	$(\texttt{gaze\_angle\_x})'$
gaze_vel_Y	Velocity of eye gaze along Y-axis	$(\texttt{gaze\_angle\_y})'$
gaze_acc_X	Acceleration of eye gaze along X-axis	$(\texttt{gaze\_vel\_Y})'$
gaze_acc_Y	Acceleration of eye gaze along Y-axis	$(\texttt{gaze\_vel\_Y})'$
head_vel_T	Translational velocity of head	$\sqrt{(\texttt{pose\_Tx})'^2 + (\texttt{pose\_Ty})'^2 + (\texttt{pose\_Tz})'^2}$
head_vel_R	Rotational velocity of head	$\sqrt{(\texttt{pose\_Rx})'^2 + (\texttt{pose\_Ry})'^2 + (\texttt{pose\_Rz})'^2}$
head_acc_T	Translational acceleration of head	$\sqrt{(\texttt{pose\_Tx})''^2 + (\texttt{pose\_Ty})''^2 + (\texttt{pose\_Tz})''^2}$
head_acc_R	Rotational acceleration of head	$\sqrt{(\texttt{pose\_Rx})''^2 + (\texttt{pose\_Ry})''^2 + (\texttt{pose\_Rz})''^2}$
writing_speed	Speed of writing	–

Table 1: Features used in the study. X are features returned by OpenFace. $(\texttt{X})'$ and $(\texttt{X})''$ are their first and second derivatives.

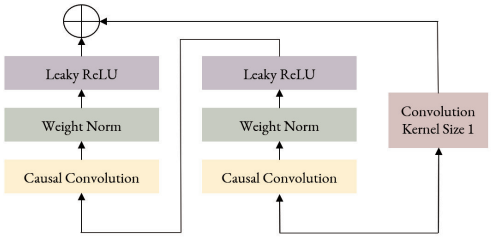

Figure 1: Composition of the i-th layer of the network (Franceschi, Dieuleveut, and Jaggi 2019).

states of cognitive disequilibrium and flow. We also carried out a randomization test (with 1000 runs) and computed the K-S statistics (D) by randomly permuting cognitive state labels. The results of our experiments are discussed in detail in Section 4.

Deep Feature Embedding

The field of affective and cognitive computing relies on supervised learning algorithms (D'Mello, Bosch, and Chen 2018), and is therefore heavily dependent on training data from expert annotators or self-reports by participants of a study. Since most advanced and powerful supervised learning algorithms require substantial amounts of training data to learn reliable decision functions, application of affective computing is severely limited by short supply of trained expert annotators or potentially biased self-reports. To this end, we investigated the utility of an unsupervised representation learning model proposed by (Franceschi, Dieuleveut, and Jaggi 2019), which can be trained on a large amount of unlabeled data to learn potentially useful feature representations. By automatically learning useful features for classifying raw data, representation learning algorithms replace manual feature engineering and allow systems to identify potential discriminators and use them to support a specific predictive task. Very few studies have focused on unsupervised representation learning for time series and (Franceschi, Dieuleveut, and Jaggi 2019) is amongst the few general-purpose representation learning algorithms for time series without any structural assumptions on non-temporal data. Their model can learn representations from multivariate time series segments of varying lengths in a completely unsupervised fashion using a triplet loss function coupled with time-based negative sampling. The model (Figure 1) comprises of a deep neural network with dilated causal convolutions to handle time series (Oord et al. 2016). This model minimizes an unsupervised triplet loss function which assigns similar time series proximate embeddings based on the assumption that they occur in temporal proximity while a distant subseries chosen at random (from either the same time series or a different one) is likely to be dissimilar. Therefore, for a reference time subseries x^{ref}, the paper chooses one of its own subseries as the positive example x^{pos} and another randomly chosen subseries x^{neg} as the negative example. In order to improve the convergence and the stability of the training procedure, the model chooses multiple negative samples independently. The training objective of the model is given by the following equation:

$$C = -log\Big(\sigma\Big(f\big(x^{ref},\theta\big)^T f\big(x^{pos},\theta\big)\Big)\Big) - \sum_{k=1}^{K} log\Big(\sigma\Big(f\big(x^{ref},\theta\big)^T f\big(x_k^{neg},\theta\big)\Big)\Big) \quad (1)$$

where $f(.,\theta)$ is a deep network with parameters θ and σ is the sigmoid function.

The unsupervised representation learning model was trained on 248 time series segments each having 27 features (refer Table 1) over 300 time steps. The unsupervised model returns embeddings of a fixed and pre-determined shape. We trained our models for 4 different output dimensions of $(64, 1)$, $(128, 1)$, $(256, 1)$ and $(512, 1)$ respectively, and found that the model with 64 features performed comparably to more complex ones in the classification task, and we chose to use 64-dimensional embeddings as our featurization.

Using these output embeddings as feature vectors and manually annotated labels, we trained a random forest classifier to predict the cognitive state (Flow or Cognitive Disequilibrium) of a time series segment. We chose random forests because they are able to learn non-linear and complex decision boundaries, work well with high-dimensional data and can be robust to outliers. The unsupervised representation learning model coupled with a random forest classifier can function as a semi-supervised model, where the former learns embeddings (features) from a large number of time

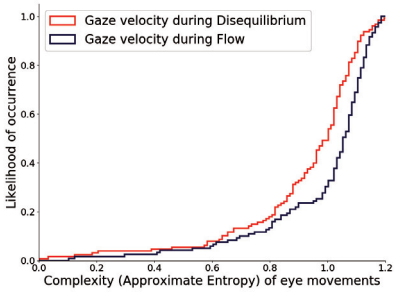

Figure 2: ECDFs of Approximate Entropy (AppEn) values of Gaze velocity in CD (red) & Flow (blue). While CD has slightly lower complexity than flow, this difference is not obvious in regions of low complexity.

series segments in a completely unsupervised fashion, and the latter uses these features and a limited number of annotations to learn a decision function. Such a semi-supervised paradigm can be extremely useful in practice of affective computing, where obtaining vast amounts of unlabeled data is extremely easy, but its annotation can be expensive.

4 Results

Analysis of Time Series Complexity Measures

The results of K-S and randomization tests are illustrated in Table 2. It can be clearly seen that the distributions of complexity measures of gaze velocity significantly differ across states of Cognitive Disequilibrium and Flow. Furthermore, distributions of Approximate and Sample Entropies of *all* behavioural features yielded significant differences between the two states. In order to investigate the directionality of the difference i.e. to answer whether behavioral signals in Cognitive Disequilibrium resulted in higher complexity than Flow or vice versa, we plotted the Empirical Cumulative Distribution Functions (ECDFs) for each complexity measure-feature pair which had a significant difference (Figure 2). The plots reveal that it is much more likely to observe lower complexity values in CD than in Flow. These results are in contrast to the findings of (D'Mello, Dale, and Graesser 2012), which suggested that cognitive disequilibrium is correlated with a whitening of gross-body movement signals. Since whitening of a signal adds to its complexity, then gross body signals in Cognitive Disequilibrium should have higher complexity. However, our results (for instance Figure 2) consistently suggest otherwise. Inspired by these statistical results, we investigate the utility of complexity measures from a multivariate point of view in predicting CD and Flow.

Deep Feature Embedding Results

Figure 3 is a 3-dimensional UMAP (McInnes, Healy, and Melville 2018) visualization of deep features (embeddings) returned by the unsupervised representation learning model. Subplots A and B represent embeddings of baseline non-personalized features and are colored by labels and subjects respectively. As shown, the deep features group the data points into two separate clusters and in most cases the same subjects belongs to the same cluster. In other words, it seems that the deep embeddings have learned mostly the between-subject difference rather than the discrimination between labels. Subplots C and D are results from embedding learned from personalized features (i.e. features for a given subject are normalized using mean and standard deviation of the same subject aggregated across all sessions). As a result, the post-personalization features remove the between-subject variation and thus force the embeddings to learn something different, as can be seen from subplot D. It is however not obvious from subplot C whether the embedding is able to discriminate between labels due to its high dimensional feature space and possibly non-linear decision boundary, which motivates us to feed the embedding results into powerful classifier such as random forest for further evaluation. The next section presents results from these experiments.

Predictive Utility of Deep Features Embedding and Complexity Measures: Multivariate View

We conducted experiments to compare predictive utility of time series complexity measures and deep embedding features by feeding the two different features sets into random forest classifiers. We choose random forest for illustration as one a popular model type capable of learning complex non-linear decision boundaries. In order to test the utility of feature personalization/normalization, we compared the performance of the model using personalized and non-personalized features for both complexity measures and deep embedding features. In addition, we conducted three types of experiments given the hierarchical structure of the data: one subject has multiple sessions (one session is one child solving one problem) and one session has multiple

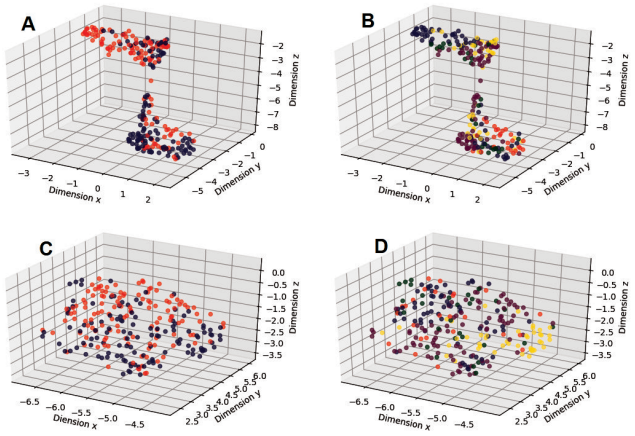

Figure 3: 3D visualization of Deep Feature Embedding: (A) non-personalized features version colored by labels; (B) non-personalized features version colored by subjects; (C) personalized features version colored by labels; (D) personalized features version colored by subjects.

Complexity Measures	Gaze Velocity		Gaze Acceleration		Head Velocity		Head Acceleration	
	X	Y	X	Y	Translational	Rotational	Translational	Rotational
Approximate Entropy	**0.278**	**0.155**	**0.243**	**0.243**	**0.172**	**0.231**	**0.167**	**0.200**
Higuchi Fractal Dimension	**0.140**	**0.208**	0.119	0.119	0.122	**0.141**	0.071	0.114
Katz Fractal Dimension	**0.176**	**0.128**	0.085	0.085	**0.182**	**0.230**	**0.151**	**0.222**
Permutation Entropy	**0.215**	**0.218**	**0.283**	**0.283**	0.119	0.082	0.077	0.079
Sample Entropy	**0.259**	**0.166**	**0.178**	**0.178**	**0.177**	**0.224**	**0.190**	**0.210**
Spectral Entropy	**0.136**	**0.196**	**0.157**	**0.157**	**0.174**	**0.133**	0.119	**0.278**

Table 2: Kolmogorov-Smirnov statistics. Values in **bold** indicate statistically significant differences at 5% significance levels. All these values also had empirical p-values < 0.05 resulting from the randomization test. The distribution of complexity measures of *gaze velocity* differs significantly across states of Cognitive Disequilibrium and Flow.

Features	Deep				Complexity			
	Non-personalized		Personalized		Non-personalized		Personalized	
Experiments	Random	LOPO	Random	LOPO	Random	LOPO	Random	LOPO
Precision	0.83 (0.037)	0.81 (0.063)	0.82 (0.035)	0.78 (0.083)	0.74 (0.062)	0.61 (0.099)	0.71 (0.030)	0.69 (0.143)
Recall	0.82 (0.04)	0.7 (0.111)	0.8 (0.050)	0.61 (0.159)	0.71 (0.070)	0.5 (0.138)	0.71 (0.033)	0.59 (0.143)
F1	0.82 (0.04)	0.71 (0.092)	0.8 (0.050)	0.61 (0.137)	0.71 (0.071)	0.48 (0.127)	0.71 (0.033)	0.60 (0.135)
Accuracy	0.82 (0.04)	0.7 (0.111)	0.8 (0.050)	0.61 (0.158)	0.71 (0.070)	0.5 (0.137)	0.71 (0.033)	0.59 (0.143)
AUC	0.83 (0.052)	0.79 (0.058)	0.8 (0.051)	0.74 (0.143)	0.72 (0.056)	0.43 (0.177)	0.71 (0.034)	0.55 (0.133)

Table 3: Performance comparison of deep feature embeddings vs. complexity measures, personalized vs. non-personalized feature sets in Random and LOPO experiments.

time series segments. The first type of experiment ("Random") makes a *random split* between train and test sets[4], ignoring the grouping structures. This type of experiment could yield inflated algorithm performance as the information from the same session and the same subject may appear in both the training and testing sets, allowing the model to succeed by hooking-onto personal characteristics of some distinct subjects. The second type of experiment is conducted by *leaving one session out* ("LOSO") where the test set contains all data from one session (thus the same subject). This setup illustrates a "warm start" where we have data from all other subjects in addition to data from the same test subject, but from different sessions than the left-out test session. The last type is *leave-one person(subject)-out* ("LOPO"), which represents a "cold start" scenario where the model is trying to predict for a completely unseen subject. Due to varying degrees of information sharing between training and test set, we expect the performance will degrade from the upper bound case of random split, to LOSO and to the most conservative (but of most practical utility) LOPO experiments. Figure 4 shows the Area Under Receiver Operating Characteristic Curve (AUC) scores under various experimental conditions, comparing the effect of feature personalization and utility of deep embedding features versus baseline complexity features. The left panel shows the results from non-personalized features while right panel are those with personalized features. There are several interesting findings:

- *Effect of experiment conditions*: We observe a downward trend for both deep features and complexity measures from random split to Leave-One-Person-Out (LOPO, "cold start" condition), suggesting that the supervised model can be trapped to overfit on subject's specifics;
- *Effect of deep embeddings and complexity features*: As shown, deep features seem to have a clear advantage in predictive utility over complexity measures. This advantage is more prominent with non-personalized feature set;
- *Effect of feature personalization:* With complexity measures, personalization shows slight improvement from the non-personalized ones across all experiment conditions. With deep embedding features, it is interesting to note that the performance does not drop as significantly as in non-personalized version. In fact, the LOPO scores reveal a level of performance comparable with the random split evaluation with non-personalized features.

Table 3 presents detailed performance metrics (Precision, Recall, F1 score, Accuracy and Area Under ROC Curve) under different experiment conditions, generally consistent with data in Figure 4.

Towards Semi-supervised Learning: How Much Supervision is Necessary?

We conducted sensitivity analysis to demonstrate the utility of unsupervised embedding in the prediction task. In these set of experiments, we fixed the held-out test set, varied the size of the training set and reported the performance of our semi-supervised approach accordingly. For brevity, we only present results using non-personalized deep embedding features with random split and leave-one-person-out (LOPO) evaluation. As shown in Figure 5, the model was able to achieve reasonable performance even with limited amount of supervision. For random split, the performance drop is more prominent however within reasonable range. For the leave one person out (LOPO) condition, the performance is robust even with very limited amount of labeled data.

Comparison with Deep Supervised Learning

We also compared the performance of our semi-supervised model with *ResNet* (He et al. 2016). (Fawaz et al. 2019) in

[4]Random split results are reported over 5-folds of cross validation throughout the paper.

a recent and comprehensive survey found that ResNet can significantly outperform other deep learning approaches in classifying time series on the UCR/UEA and MTS archives. In addition, they found encouraging results (comparable predictive performance and significantly less training & testing time) while comparing ResNet to other state-of-the-art time series classification algorithms such as HIVE-COTE (Lines, Taylor, and Bagnall 2016).

	Semi-supervised		Supervised (ResNet)	
Experiments	Random	LOPO	Random	LOPO
Precision	0.83 (0.037)	0.81 (0.063)	0.81 (0.016)	0.77 (0.074)
Recall	0.82 (0.04)	0.7 (0.111)	0.81 (0.023)	0.78 (0.071)
F1	0.82 (0.04)	0.71 (0.092)	0.81 (0.024)	0.77 (0.069)
Accuracy	0.82 (0.04)	0.7 (0.111)	0.81 (0.022)	0.78 (0.072)
AUC	0.83 (0.052)	0.79 (0.058)	0.8 (0.016)	0.73 (0.081)

Table 4: Performance comparison of semi-supervised-model & ResNet.

Table 4 compares ResNet and the deep semi-supervised model introduced above on several performance metrics. For brevity, we only use the non-personalized feature set for two types of experiments: *random split* and *leave-one-person-out*. The results reveal that while ResNet achieved higher accuracy (0.78%) in the LOPO experiments, there was no significant difference (within 95% confidence interval) between the models in terms of AUC in both evaluation scenarios.

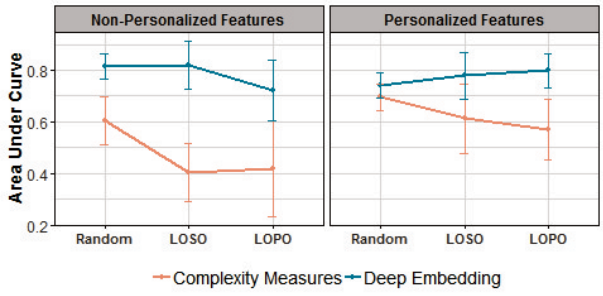

Figure 4: Area Under ROC Curve (AUC) with 95% confidence interval, varying experimental conditions, feature penalization choices and featurization techniques (deep embedding vs. complexity measures).

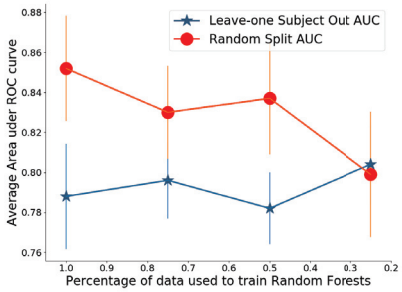

Figure 5: Area Under Curves (AUCs) with varying amount of training data, comparing random split and LOPO experiments.

5 Discussion

In this paper, we explored a semi-supervised framework to model the dynamics of involuntary behavioral signals collected using inexpensive sensors in order to discriminate between cognitive disequilibrium and flow as the primary input for decision making by human teachers or intelligent tutoring systems. Experimental results with a modestly sized multi-modal multi-sensor dataset, collected from young children practicing problem solving in a naturalistic environment, reveal several insights. Firstly, in comparison to time series complexity measures commonly cited in dynamical systems literature, we find that the deep feature embedding approach is able to identify plausible discriminators between those two states of interest more effectively than considered alternatives, when coupled with a random forest classifier. Secondly, we notice that this deep representation was able to effectively generalize from training subjects to previously unseen subjects, as demonstrated by its robust performance with leave-one-person-out experiments, and the advantage is even more pronounced with personalized features. Thirdly, sensitivity analysis with the semi-supervised framework shows that with deep embeddings features, the model is able to learn effective discrimination with even a small number of labeled data points, and the resulting performance is comparable with a potent fully supervised deep learning alternative which often requires large extents of supervision. When further validated with a more diverse set of subjects, the proposed approach has the promise to scale up practicality of the task of cognitive and affective state detection that is often bottle-necked by high costs of label acquisition even with abundant unlabeled data. Practically relevant capability of generalization to unseen subjects is also encouraging as the proposed approach would often be expected to work well with out-of-sample subjects in the real world use cases.

6 Conclusion

Effective coaching of problem solving requires real time monitoring of students' cognitive and affective states, which can be challenging in societal environments with limited teaching resources. This paper tackles this challenge with a semi-supervised framework designed for automatic detection of two critical states of students during problem solving: Cognitive Disequilibrium and Flow. The discrimination model learns from involuntary behavioral signals that are less prone to social editing than more common alternatives, and that can be feasibly collected using inexpensive sensors. We empirically demonstrated the utility of the proposed approach and shown that it could work well even with a modest amount of data and limited supervision. When fully developed into a working system, we envision that the proposed methodology can play a role in augmenting human teacher's perceptual capability in the classroom as well as in improving the effectiveness of intelligent tutoring systems.

References

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. IEEE.

Bandt, C., and Pompe, B. 2002. Permutation entropy: a natural complexity measure for time series. *Physical review letters* 88(17):174102.

Camras, L. A., and Shutter, J. M. 2010. Emotional facial expressions in infancy. *Emotion review* 2(2):120–129.

Chen, L.; Li, X.; Xia, Z.; Song, Z.; Morency, L.-P.; and Dubrawski, A. 2016. Riding an emotional roller-coaster: A multimodal study of young child's math problem solving activities. *International Educational Data Mining Society*.

Coulson, M. 2004. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior* 28(2):117–139.

Csikszentmihalyi, M. 2013. *Flow: The psychology of happiness*. Random House.

D'Mello, S., and Graesser, A. 2012. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(4):23.

D'Mello, S., and Graesser, A. C. 2014. Confusion. In *International handbook of emotions in education*. Routledge. 299–320.

D'Mello, S., and Kory, J. 2012. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 31–38. ACM.

D'Mello, S., and Kory, J. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47(3):43.

DMello, S.; Jackson, T.; Craig, S.; Morgan, B.; Chipman, P.; White, H.; Person, N.; Kort, B.; el Kaliouby, R.; Picard, R.; et al. 2008. Autotutor detects and responds to learners affective and cognitive states. In *Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems*, 306–308.

D'Mello, S.; Bosch, N.; and Chen, H. 2018. Multimodal-multisensor affect detection. In *The Handbook of Multimodal-Multisensor Interfaces*, 167–202. Association for Computing Machinery and Morgan & Claypool.

D'Mello, S.; Dale, R.; and Graesser, A. 2012. Disequilibrium in the mind, disharmony in the body. *Cognition & emotion* 26(2):362–374.

D'Mello, S. 2013. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology* 105(4):1082.

Ekman, P. 1994. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique.

Esteller, R.; Vachtsevanos, G.; Echauz, J.; and Litt, B. 2001. A comparison of waveform fractal dimension algorithms. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 48(2):177–183.

Fawaz, H. I.; Forestier, G.; Weber, J.; Idoumghar, L.; and Muller, P.-A. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33(4):917–963.

Franceschi, J.-Y.; Dieuleveut, A.; and Jaggi, M. 2019. Unsupervised scalable representation learning for multivariate time series. *arXiv preprint arXiv:1901.10738*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heine, S. J.; Lehman, D. R.; Peng, K.; and Greenholtz, J. 2002. What's wrong with cross-cultural comparisons of subjective likert scales?: The reference-group effect. *Journal of personality and social psychology* 82(6):903.

Higuchi, T. 1988. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena* 31(2):277–283.

Hussain, M. S.; AlZoubi, O.; Calvo, R. A.; and DMello, S. 2011. Affect detection from multichannel physiology during learning sessions with autotutor. In *International Conference on Artificial Intelligence in Education*, 131–138. Springer.

Joseph, E. 2005. Engagement tracing: using response times to model student disengagement. *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology* 125:88.

Kantz, H.; Kurths, J.; and Mayer-Kress, G. 2012. *Nonlinear analysis of physiological data*. Springer Science & Business Media.

Kapoor, A.; Burleson, W.; and Picard, R. W. 2007. Automatic prediction of frustration. *International journal of human-computer studies* 65(8):724–736.

Kilbride, J. E., and Yarczower, M. 1983. Ethnic bias in the recognition of facial expressions. *Journal of Nonverbal Behavior* 8(1):27–41.

Krosnick, J. A. 1999. Survey research. *Annual review of psychology* 50(1):537–567.

Lines, J.; Taylor, S.; and Bagnall, A. 2016. Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. In *2016 IEEE 16th international conference on data mining (ICDM)*, 1041–1046. IEEE.

McDaniel, B.; D'Mello, S.; King, B.; Chipman, P.; Tapp, K.; and Graesser, A. 2007. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pincus, S. M., and Goldberger, A. L. 1994. Physiological time-series analysis: what does regularity quantify? *American Journal of Physiology-Heart and Circulatory Physiology* 266(4):H1643–H1656.

Pincus, S. M. 1991. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences* 88(6):2297–2301.

Powell, G., and Percival, I. 1979. A spectral entropy method for distinguishing regular and irregular motion of hamiltonian systems. *Journal of Physics A: Mathematical and General* 12(11):2053.

Richman, J. S., and Moorman, J. R. 2000. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology* 278(6):H2039–H2049.

Zimmerman, B. J., and Martinez-Pons, M. 1990. Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of educational Psychology* 82(1):51.