Classifying Learner Behavior from High Frequency Touchscreen Data Using Recurrent Neural Networks

Zachary A. Pardos
UC Berkeley
Berkeley, CA, USA
zp@berkeley.edu

Changran Hu
Tsinghua University
Haidian Qu, Beijing Shi, China
huchangran@gmail.com

Pengqiu Meng
Wuhan University
Wuhan, Hubei, China
mengpengqiu@whu.edu.cn

Michael Neff
UC Davis
Davis, CA, USA
mpneff@ucdavis.edu

Dor Abrahamson
UC Berkeley
Berkeley, CA, USA
dor@berkeley.edu

ABSTRACT
Sensor stream data, particularly those collected at the millisecond of granularity, have been notoriously difficult to leverage classifiable signal out of. Adding to the challenge is the limited domain knowledge that exists at these biological sensor levels of interaction that prohibits a comprehensive manual feature engineering approach to classification of those streams. In this paper, we attempt to enhance the assessment capability of a touchscreen based ratio tutoring system by using Recurrent Neural Networks (RNNs) to predict the strategy being demonstrated by students from their 60Hz data streams. We hypothesize that the ability of neural networks to learn representations automatically, instead of relying on human feature engineering, may benefit this classification task. Our RNN and baseline models were trained and cross-validated at several levels on historical data which had been human coded with the task strategy believed to be exhibited by the learner. Our RNN approach to this historically difficult high frequency data classification task moderately advances performance above baselines and we discuss what implication this level of assessment performance has on enabling greater adaptive supports in the tutoring system.

KEYWORDS
Recurrent neural networks, high frequency data, touchscreen, sensors, tutoring systems, embodied cognition, assessment

1 INTRODUCTION
We explore how patterns in students’ high frequency touchscreen data can be detected in order to classify strategic behaviors exhibited while learning in an embodied mathematics tutor. The particular application is the Mathematics Imagery Trainer for Proportionality (MITp), a teaching tool in which students learn the concept of proportion, in advanced of it being introduced in the standard school curriculum, by moving two touch markers on the screen over time to achieve the desired ratio between the distance from the left hand marker and the bottom of the screen and the distance from the right hand marker and the bottom of the screen. The screen the learner is interacting with turns green as the correct ratio is arrived at. The telemetry data consists of these two input points, sampled at 60Hz, with the students’ activity labeled post-hoc by experts at each time step with the strategy being exhibited. The techniques we bring to bear for classifying these strategies include Long Short-Term Memory models, meant to learn representations of these strategies from raw times-series data, and multinomial logistic regression, meant to serve as a baseline for classification. The goals of this work are to explore the classification of high frequency data using deep learning methods and to extend previous research in order to enable an animated virtual pedagogical agent in the tutor to more effectively guide students through the MITp learning process.

2 BACKGROUND
The use of high frequency sensor data to adapt instruction or contribute to the understanding of the learning process has been used in a variety of contexts. To name only a few, these included using galvanic skin response and other sensors to measure affective states [3], using eye-tracking to study how learners construct knowledge using various graphical representations of concepts [21], and using eye-tracking to diagnose cognitive traits [20]. In these works, the sequence data were pre-processed to a coarser level of frequency before use. The use of these data has been referred to as multimodal learning analytics [4], often to describe data collected from learning contexts in which the learning process is not satisfactorily characterizable from clickstream events or response logs, but rather from other modes of interaction.

The era of deep-learning has catalyzed the use of recurrent neural networks for a variety of time-series tasks outside of education. This
Thus human reasoning emerges, and is expressed through situated virtual pedagogical agent with a limited ability to guide students and are described below. Distinct strategies were identified in past work as most important to be helpful states to transition through while learning [7]. Three collected during tutoring in previous work and were determined MITp have been deduced through manual analysis of data streams artifacts. The nature of strategies exhibited during interaction with asking of questions that provoke reflection, and introduction of new automatic assessment at the strategy level and to provide subsequent through the process [1]. It remains an open challenge to enable au- to scale up MITp to a large audience. Previous work introduced a action sessions can last an hour and require a trained technician introduce scaffolding artifacts, such as grid lines or numbers. Inter- no- suggestions as they interact. The technician also decides when to explaining how the system works, sitting beside the child and making physically coordinated challenge of moving two hands on a touch screen to make it green, a result which occurs when the ratio of hand heights matches the pre-programmed ratio of 1:2. Through engaging in this embodied-interaction activity and building particular movement schemes related to proportions process, students can develop pre-symbolic quantitative understanding of proportions draws on embodiment theory, which views the mind as extending dynamically through the body into the natural-cultural ecology [2]. Thus human reasoning emerges, and is expressed through situated sensorimotor interactions [1]. The MITp system (Figure 1) poses the left touch and Right touch indicate the locations of the left and most finger and the right most finger on the 2-D coordinate plane of the touchscreen. In practice, only the Y values are used, as it is the ratio of the two heights that affect satisfaction of the task. NA values represent no touch during that 60hz reading.

Figure 1: the Mathematics Imagery Trainer (MITp)

In trying to achieve a green screen background (and thus the 1:2 ratio of cursor positions), they can discover "the higher, the bigger": that the gap between their hands is bigger as they make the same proportion higher on the screen. Moving one hand and then adjusting the other leads to the insight "A-per-B": that when they move their left hand one unit, they must move their right two (for 1:2). When they create continuous green, they learn that the "speed" of the right hand is twice that of the left. If we can determine which strategy the child is deploying, this gives great insight into their learning process and is therefore powerful information in determining the most effective response of a pedagogical agent. For example, it can indicate the need to provide remedial instruction. Successful performance of a strategy can indicate that it is time to encourage the child to explore other strategies. Successful performance of the set of strategies can indicate that it is time to introduce new artifacts and advance to the next stage in the tutorial process.

### 2.1 The Proportionality Tutor App

Our data were obtained from an application called the Mathematics Imagery Trainer for Proportionality (MITp) [12], collected as part of an effort to enable adaptive virtual agent tutoring to children as they use MITp. MITp is an activity design architecture developed to support students in learning the contents of ratio and proportion, an important yet difficult topic for many students. Understanding proportionality involves appreciating multiplicative relations between extensive quantities; a change in one quantity is always accompanied by a change in the other, and these changes are related by a constant multiplier [14, 26]. Our MITp approach to support students in developing multiplicative understanding of proportions draws on embodiment theory, which views the mind as extending dynamically through the body into the natural-cultural ecology [2].

MITp has traditionally been used with a human technician who explains how the system works, sitting beside the child and making suggestions as they interact. The technician also decides when to introduce scaffolding artifacts, such as grid lines or numbers. Interaction sessions can last an hour and require a trained technician to guide every student through the process. This limits the ability to scale up MITp to a large audience. Previous work introduced a virtual pedagogical agent with a limited ability to guide students through the process [1]. It remains an open challenge to enable automatic assessment at the strategy level and to provide subsequent appropriate instruction, valorization, providing of correction hints, asking of questions that provoke reflection, and introduction of new artifacts. The nature of strategies exhibited during interaction with MITp have been deduced through manual analysis of data streams collected during tutoring in previous work and were determined to be helpful states to transition through while learning [7]. Three distinct strategies were identified in past work as most important and are described below.

---

**DATASET**

We used chronological time-stamped sequences of interactions of students with the tutor that record the student touch locations as they progress through instructional prompts with the tutor. The dataset contains 49 students’ csv files, of which only 5 were labeled, seen in Table 1. The length of each data file ranged from 39,736 to 176,283 time slices, lasting from 11 minutes to 48 minutes and sampled at 60 hz. For the training of our models, we used only the 5 labeled students, whose descriptive stats can bee seen in Table 4. Two students did not make it to the last of three phases of the tutor session, and thus never demonstrated the SP (speed) strategy. While we could have filtered this out, low subject count and missing values are common place real-world classification tasks and thus, we chose to include the label to keep the task authentic and because it was an important label to predict in the pedagogical scheme of the tutor. While filtering out students with such missing data may be common place when more subjects are on hand, further reducing our subject pool would not have been tenable. The data shown in Table 1 included:

1. Left touch and Right touch indicate the locations of the left most finger and the right most finger on the 2-D coordinate plane of the touchscreen. In practice, only the Y values are used, as it is the ratio of the two heights that affect satisfaction of the task. NA values represent no touch during that 60hz reading.
2. Color indicates how close the screen is to the goal color of green. When the color value is 1, this means that the student has achieved the target ratio of 1:2 between the heights of her two fingers.
Table 1: Data Examples

(A) Raw Data from one student. This is example showing one reading per second for demonstrating purposes. The actual data we use is sampled at 60 readings per second.

<table>
<thead>
<tr>
<th>Left Touch</th>
<th>Right Touch</th>
<th>Color</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-3.0 2.9)</td>
<td>(3.4 4.2)</td>
<td>0.561</td>
<td>08:13.000</td>
</tr>
<tr>
<td>(-3.0 3.1)</td>
<td>(3.4 8.7)</td>
<td>0.772</td>
<td>08:14.017</td>
</tr>
<tr>
<td>(-3.0 3.1)</td>
<td>(3.4 9.3)</td>
<td>1.000</td>
<td>08:15.033</td>
</tr>
</tbody>
</table>

(B) Data format after processing

<table>
<thead>
<tr>
<th>Left-y</th>
<th>Right-y</th>
<th>Color</th>
<th>ID</th>
<th>Prompt</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.9</td>
<td>4.2</td>
<td>0.561</td>
<td>EL</td>
<td>1003</td>
<td>D</td>
</tr>
<tr>
<td>3.1</td>
<td>8.7</td>
<td>0.772</td>
<td>EL</td>
<td>1003</td>
<td>D</td>
</tr>
<tr>
<td>3.1</td>
<td>9.3</td>
<td>0.940</td>
<td>EL</td>
<td>1003</td>
<td>D</td>
</tr>
</tbody>
</table>

Table 2: Instructional Prompts

<table>
<thead>
<tr>
<th>Prompt ID</th>
<th>Text displayed to students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1003</td>
<td>“Your goal is to make the screen green”</td>
</tr>
<tr>
<td>1004</td>
<td>“You make the screen green by moving the cursors”</td>
</tr>
<tr>
<td>1005</td>
<td>“You can move the cursor up and down like this”</td>
</tr>
</tbody>
</table>

Table 3: Label Definitions. The labels marked as strategy labels are various ways in which students can find or maintain the proper ratio between two fingers and thus keep the screen green.

<table>
<thead>
<tr>
<th>Label</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Drag one hand</td>
<td>Keeping one hand still, exploring with the other.</td>
</tr>
<tr>
<td>AB</td>
<td>A per B</td>
<td>Step-wise movement</td>
</tr>
<tr>
<td>SP</td>
<td>Speed</td>
<td>Simultaneous movement.</td>
</tr>
<tr>
<td>NT</td>
<td>No Touch</td>
<td></td>
</tr>
<tr>
<td>TT</td>
<td>Initial Touch</td>
<td>Release Touch</td>
</tr>
<tr>
<td>TL</td>
<td>The Ladder</td>
<td>Drag Both Hands</td>
</tr>
<tr>
<td>T, NV</td>
<td>Tuning, Not Visible</td>
<td></td>
</tr>
<tr>
<td>H, O</td>
<td>Horizontal movement, Other</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Descriptive stats of the 5-labeled students showing the number of time slices coded with each of the three strategy labels

<table>
<thead>
<tr>
<th></th>
<th>AB</th>
<th>D</th>
<th>SP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>EL</td>
<td>1740</td>
<td>15780</td>
<td>5940</td>
<td>23460</td>
</tr>
<tr>
<td>ER</td>
<td>5100</td>
<td>3960</td>
<td>0</td>
<td>9060</td>
</tr>
<tr>
<td>KN</td>
<td>6180</td>
<td>2700</td>
<td>8040</td>
<td>16920</td>
</tr>
<tr>
<td>MS</td>
<td>4680</td>
<td>6120</td>
<td>4140</td>
<td>14940</td>
</tr>
<tr>
<td>ND</td>
<td>14640</td>
<td>13500</td>
<td>0</td>
<td>28140</td>
</tr>
</tbody>
</table>

4 METHODS

In this section we describe a multinomial logistic baseline for classification, the LSTM classification model, and a variety of cross-validations used to evaluate the models which correspond to different ways in which the models could be applied in the tutor.

4.1 Multinomial Logistic Regression

We used a standard multinomial logistic regression model to serve as the baseline classifier. We used six very simple features as the input features to the model at every time slice: (1) Left hand-y coordinate (2) Right hand-y coordinate (3) Student ID represented as a one-hot dummy variable (4) Whether or not at least one finger was touching the screen (5) Instructional prompt represented as a one-hot dummy variable (6) Boolean representing if the student had reached the goal state of exhibiting the 1:2 ratio (within the tutor’s specified margin of error). At every time slice, these features were used to classify the distinct label associated by an expert with the behavior exhibited at that time.

This is a baseline approach as it represents the degree to which a class can be predicted by the student, prompt, and instantaneous value of the left and right finger. The RNN-based method, described in the next subsection, is hypothetically better equipped to identify a pattern characterized from a temporal series of finger movements.

4.2 Long Short-Term Memory Classification

Long Short-Term Memory is an augmentation to the classic Recurrent Neural Network (RNN) model [17] and was first proposed by Sepp Hochreiter and Jurgen Schmidhuber [11]. It maintains a hidden state and a “longer-term” cell state. Through its architecture [10], it has been demonstrated as being able to classify patterns based on longer sequences than an RNN; diminishing the phenomenon known as the vanishing gradient. In our classification task, we choose an LSTM as the model of choice with a times series of left and right hand coordinates as the inputs and the hand coded labels as the categorical outputs.

We used the 60Hz sampled x, y position of both hands (4 features in total) as input, and used the labels (in one-hot representation) as ground truth. The model will predict the current label with current input every time step or every entire sequence, depending on the specific cross-validation setting discussed in the evaluation section.

Because labels only exist for part of the whole time series, we use a default value -1 to pad all the time when no label exists. A custom loss function is defined to only calculate the loss when the
We enumerate the following expectations for the classification

5 EVALUATION

We constructed a set of experiments which evaluated different
generalizing properties of the models. All experiments were tested
on both the expanded label set and the set only containing the
strategy labels. In addition to training the models on the entire
student sequence, we also tried partitioning the sequences into
different segments (or chops) in order to reduce the length of the
sequences being trained on and to explore if the model would better
generalize when training on larger batches of sequences instead
of fewer longer sequences. We also varied the frequency of the
predictions being made between (1) predicting the label at every
time slice (60hz) and (2) only predicting the presence of the label
in the sequence at the very end of the sequence, as defined by
the segmenting. When predicting at the sequence level, a binary
prediction was made for every label for every sequence. Accuracy
was calculated based on the aggregate performance of all binary
predictions. The topology for this sequence prediction model was
the same as the time slice LSTM except that instead of a softmax
over all the labels at every time slice, there were instead independent
sigmoid outputs for each label occurring at the very last time slice
of the sequence. Five-fold cross-validation is conducted at both the
student level and sequence chop level for all experiments. The data
were separated into 5 folds and each fold served as the test set once,
with the rest serving as the training set. The results from the 5
phases were averaged to produce a single accuracy metric.

5.1 Sequence segmentation

Besides using the original whole sequence, we chop the sequence
into segments in two ways (1) Chop by prompts: during the tutor
sessions, students were given prompts to direct them to adjust their
movements. Accordingly, there is a column named ‘prompt’ illust-
trating some specific instructions given at that time period. A chop
segmented by prompt runs from the start of a prompt to the start
of the next prompt, normally including some time after the prompt
during which the student is interacting with the system and the
tutor is silent. This sequence chop approach produced 162 total
segments of various lengths. (2) Chop by labels: domain experts
labeled the behavior from recordings of sessions based on the move-
ment pattern exhibited by the student. Chop by label segmented
the sequences using contiguous labels and produced 1239 chops
of various lengths. We note that this level of segmentation would
not practically be available in a real-world scenario as it requires
knowledge of the label beforehand. Nevertheless, this can serve as
a test of smaller sequence length segmentation for classification.

5.2 Hypotheses

We enumerate the following expectations for the classification
results:

- H1: Compared to logistic regression, the LSTM has the abil-
ity to learn the chronological information in the sequence.
Therefore, we expect that the LSTM models will perform
better on average than logistic regression.
- H2: Different students may exhibit strategies in different
ways. Therefore, we anticipate that models that have trained
on some portion of the student they are predicting (sequence
level cross-validation) will perform better than their respec-
tive student level cross-validation experiments.
- H3: We assume predicting the label at the frequency of ev-
ery sequence will be more accurate than predicting every
timestep due to the former being the easier scenario since
the model need only predict if a label occurred and not the
temporal sequencing in which they occurred within the se-
quence.
- H4: After restricting to the strategy label set, the classifica-
tions should be easier to learn compared to the expanded
label set since there will be less opportunity for similar labels
to be confused for one another.

6 RESULTS

Results in terms of accuracy of the majority class baseline (B),
multinomial logistic regression (LR), and Long short-term memory
model1 (LSTM) are shown for predicting the strategy label set
(Table 5) and the extended label set (Table 6).

In all but two of the experiments, the LSTM outperformed
the baseline models, mostly confirming H1. Focusing on the results of
the strategy label set which were cross-validated at the student level
and made predictions of labels at the moment-by-moment (time
step) frequency, we see that the majority class and logistic models
predicted the same across all sequence chop experiments. This was
because the training of neither baseline is affected by the chop
level - the logistic was trained using instantaneous independent
readings from each time slice, whereas the LSTM’s hidden state,
and thus predictions, are affected by inputs from previous time
slices within the chop. The LSTM most benefited from training
and predicting using the entire student’s sequence, not training on
partitions (chops) of the sequence. This indicates that the LSTM was
able to leverage signal from previous prompts in making predictions
of labels in the current prompt. The LSTM scored 20% above the
logistic and 51% above majority class in this evaluation category.
When dealing with many more classes in the extended label set,
an improvement over logistic was only seen when training and
predicting on more granular label or prompt segmented sequences.

Contrary to H2, the sequence level cross-validation outper-
formed its student level counterpart in only a few experiments (strat-
edy labels/sequence cv/sequence freq./prompt chop and extended
labels/sequence cv/time step freq./no label chop). This perhaps sug-
gests that there is less importance to pickup on a "signature" of a
student being predicted by observing some of her behaviors in the
training set.

As expected, classification at the label frequency of sequence
(H3) was by in large the easier classification task, only perform-
ing worse than its by-time-step counterpart in one out of the 10

1The LSTM model trained with a 256 node hidden layer and RMSprop optimizer
consistently performed best in the hyperparameter search.
experiments across label sets (strategy labels/sequence freq./label chop).

Finally, it can also be observed that in comparing the 10 LSTM results of the strategy label set to the expanded label set, in only one experiment does the expanded label set perform better (student cv/sequence freq./prompt chop), mostly confirming H4.

6.0.1 Notable Null Results. Additional methods were attempted to improve the classification but did not enhance the results. They were abandoned after the early or middle stages of evaluation and are reported here for posterity. We tried incorporating additional input features into the LSTM such as the color of the screen, the current prompt, and student identifier but no improved accuracy was found. We also tried different preprocessing procedures, such as downsampling the raw data to one reading per second (instead of 60) so that the entire sequence length could be reduced and so patterns could be better identified at a less granular level. We also tested different model structures, including multiple LSTM hidden layers, Phased LSTM [18], different activation functions, GRUs, and simple RNNs, but no improvements were seen.

6.0.2 Discussing the Suitability of these Results for LSTM Integration into the MITp App. It can be seen in the confusion matrices (Fig. 2) that both models do a reasonable job of correctly predicting the D label when D was actually employed, but the LSTM classification is much better (91% LSTM vs. 72% logistic). Unfortunately, both models have a low true positive rate for the other labels; 18% (logistic) or 16% (LSTM) for AB and almost 0% for SP, a label which did not appear in two out of the five students’ sequences. This is because of the high false positive rate for D (55% for logistic and 54% for LSTM). Essentially, both models report D much more frequently than the ground truth data. D is one of the earliest strategies students use when finding proportions in MITp, so the false positive rate is not overly problematic for integration into the tutor. It tells the system that the child may be doing something more simple than they are actually doing, but this will lead to simply repeating instructions to try to create a higher level behavior, which is less pedagogically problematic than skipping ahead prematurely, and repetition may even offer learning benefits.

A particularly interesting finding is the false positive rate for AB. This is 73% for logistic regression, but only 42% for LSTM, so when the LSTM predicts AB, it is much more likely to be correct. Detecting correct performance of AB is particularly useful for an autonomous tutor design as AB is the second strategy taught to a child while using the system and successful performance of AB is a pre-cursor for moving on to teaching SP. While an approximately 60% correct rate for AB labels is not sufficient on a single instance for LSTM). Essentially, both models report D much more frequently than the ground truth data. D is one of the earliest strategies students use when finding proportions in MITp, so the false positive rate is not overly problematic for integration into the tutor. It tells the system that the child may be doing something more simple than they are actually doing, but this will lead to simply repeating instructions to try to create a higher level behavior, which is less pedagogically problematic than skipping ahead prematurely, and repetition may even offer learning benefits.

A particularly interesting finding is the false positive rate for AB. This is 73% for logistic regression, but only 42% for LSTM, so when the LSTM predicts AB, it is much more likely to be correct. Detecting correct performance of AB is particularly useful for an autonomous tutor design as AB is the second strategy taught to a child while using the system and successful performance of AB is a pre-cursor for moving on to teaching SP. While an approximately 60% correct rate for AB labels is not sufficient on a single instance to decide the child is ready to move on, when this occurs multiple times and is combined with knowledge of where the system is in the overall tutoring process, it provides a useful signal.

7 CONCLUSIONS

Methodologically, the application of LSTMs to our dataset of 60hz touchscreen sensor data was successful in realizing a moderate gain in classification performance. It achieved a 47.1% accuracy in predicting the moment-by-moment strategy being employed by the student, compared to 39.29% accuracy when using logistic regression with simple features and 31.2% when classifying based on the majority class of the training set. If, instead of predicting the strategy at every moment, the classification predicts if the strategy was ever employed by the student, the accuracy climbs to 86.7% with an RNN, compared to 66.6% using the majority observation (or non-observation) of each class across students in the training set.

Table 5: Accuracy of Strategy Labels

<table>
<thead>
<tr>
<th>CV_by</th>
<th>label frequency</th>
<th>sequence chop</th>
<th>B</th>
<th>LR</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>student</td>
<td>by time step</td>
<td>no chop</td>
<td>31.2</td>
<td>39.29</td>
<td>47.1</td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>39.29</td>
<td>45.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>prompt chop</td>
<td>39.29</td>
<td>42.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq.</td>
<td>by time step</td>
<td>no chop</td>
<td>66.6</td>
<td>NA</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>47.1</td>
<td>NA</td>
<td>46.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prompt chop</td>
<td>71.7</td>
<td>NA</td>
<td>79.5</td>
<td></td>
</tr>
<tr>
<td>by seq.</td>
<td>no chop</td>
<td>31.2</td>
<td>40.6</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>15.9</td>
<td>24.88</td>
<td>39.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prompt chop</td>
<td>43.4</td>
<td>43.5</td>
<td>37.1</td>
<td></td>
</tr>
<tr>
<td>seq.</td>
<td>no chop</td>
<td>47.5</td>
<td>NA</td>
<td>33.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>73.7</td>
<td>NA</td>
<td>89.8</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Accuracy of Extended Labels

<table>
<thead>
<tr>
<th>CV_by</th>
<th>label frequency</th>
<th>sequence chop</th>
<th>B</th>
<th>LR</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>student</td>
<td>by time step</td>
<td>no chop</td>
<td>13.9</td>
<td>17.9</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>17.9</td>
<td>18.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>prompt chop</td>
<td>17.9</td>
<td>21.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq.</td>
<td>by time step</td>
<td>no chop</td>
<td>82.9</td>
<td>x</td>
<td>88.6</td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>15.2</td>
<td>NA</td>
<td>26.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prompt chop</td>
<td>75.5</td>
<td>NA</td>
<td>90.5</td>
<td></td>
</tr>
<tr>
<td>by seq.</td>
<td>no chop</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>22.2</td>
<td>23.8</td>
<td>24.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prompt chop</td>
<td>18.7</td>
<td>22.7</td>
<td>18.5</td>
<td></td>
</tr>
<tr>
<td>seq.</td>
<td>no chop</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>label chop</td>
<td>21.2</td>
<td>NA</td>
<td>26.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prompt chop</td>
<td>77.0</td>
<td>NA</td>
<td>87.0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Confusion matrices for the logistic and LSTM models based on the strategy label set and experiment with student level CV, with time step label frequency, and trained on the full student sequence (no chop).
8 DISCUSSION AND FUTURE WORK

Students devise a broad variety of sensorimotor schemes for enacting a target movement [2, 7, 8, 12]. Knowing what schemes students are employing is critical for supporting their learning, and yet determining these schemes has been a challenging engineering task. The difficulty that we encountered in modeling students’ schemes thus corroborates expectations coming from constructivist and enactivist theories, viz. that the mind and thus learning is highly individualistic and thus difficult to model in terms of generalizable qualities. Nonetheless, progress was made in establishing the improved performance of LSTM models which have now produced predictions that would be actionable in the MITp application. Most notably, the accuracy of AB labels provides an a signal the tutor can use in determining if the child is ready to advance to the more complicated Speed strategy.

There are several future directions that may prove profitable. The lessons in the MITp tutor go through several phases: guided exploration, using the A-per-B strategy, and then using the Speed strategy. Including information about the current phase of the tutoring process in the analysis may improve labeling prediction as different labels are more likely in different phases. The importance of accuracy for different labels also varies depending on the phase, so results could be more effectively interpreted if this information is included. A different approach would be to include additional multi-modal input data. Eye-tracking analyses may offer a promising direction [8], if applied in real-time and in concert with the touchscreen stream. Incorporating eye-tracking could also take us beyond the hand coded ‘strategy’ labels, which represent what a student is performing, objectively, as opposed to the interpretation of how a student is orienting toward the enactment of a movement. It is expected that with a larger set of labeled students and the ability to limit the imbalance of the Speed class, classification accuracy would improve to levels which would justify deeper integration of these models in the tutor. Finally, data from the unlabeled set of students could be brought to bear in order to explore, instead of classify, the common patterns of learning behavior exhibited during each phase of tutoring. These patterns could then be reconciled with subject matter expert tagged strategies to deepen our understanding of the mechanics of learning in this context.

ACKNOWLEDGMENTS

Support for this work was partially provided by the National Science Foundation through grant awards 1320029, 1321042, and 1547055.

REFERENCES
