

Applying the Deep Learning Method for Simulating Outcomes of Educational Interventions

Hyemin Han

University of Alabama

Kangwook Lee*

Korea Advanced Institute of Science and Technology

Firat Soylu*

University of Alabama

Author Note

Hyemin Han, Educational Psychology Program, University of Alabama. Box 870231, Tuscaloosa AL, 35478, USA. Telephone: 1-205-348-0746. Fax: 1-205-348-0683. Email: hyemin.han@ua.edu

Kangwook Lee, The School of Electrical Engineering, Korea Advanced Institute of Science and Technology. 291 Daehak-ro, Yeseong-gu, Daejeon 34141, South Korea. Tel: 1-42-350-3402. Fax: 1-42-350-3410. Email: kw1jjang@gmail.com

Firat Soylu, Educational Psychology Program, University of Alabama. Box 870231, Tuscaloosa AL, 35478, USA. Telephone: 1-205-348-6267. Fax: 1-205-348-0683. Email: fsoylu@ua.edu

Correspondence concerning this article should be addressed to Hyemin Han, Educational Psychology Program, University of Alabama, Box 870231, Tuscaloosa, AL 35487, USA. Email: hyemin.han@ua.edu

* These authors contributed equally to this work

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

Abstract

Predicting outcomes of educational interventions before investing in large-scale implementation efforts in school settings is essential for educational policy-making. However, due to time and resource limitations, conducting longitudinal, large-scale experiments testing outcomes of interventions in authentic settings is difficult. Here, we introduce the deep learning method as a way to address this issue and illustrate the use of the deep learning method for the prediction of intervention outcomes through a MATLAB implementation. The presented deep learning method extracts predictable patterns from an empirical dataset to simulate large-scale intervention outcomes. Findings from our simulations suggest that the deep learning applied simulation model can predict intervention outcomes significantly more accurately compared to the traditional regression analysis methods.

Keywords: deep learning, machine learning, computer simulation, neural network, educational intervention, outcome prediction, policy-making

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

1. Introduction

Different educational interventions based on findings from psychological studies have been developed to promote academic motivation and social adjustment among children and adolescents [1–5]. Some of these interventions have successfully produced positive long-term, large-scale behavioral effects among diverse populations [6, 7]. Given the investment required for a long-term, large-scale intervention, educators should carefully predict such effects before applying the interventions in the real world. To do so, first, findings from lab studies need to be replicated in authentic contexts, to examine whether similar intervention effects can be found in real-life contexts. However, limited time and resources make it difficult to conduct such replication studies. Moreover, research ethics is also an issue, since interventions with null or negative effects can significantly affect students' long-term development [8, 9].

Analyzing educational intervention data and interpreting results for educational applications present unique challenges. In an educational intervention there are usually many independent variables across different levels. For example, in a school intervention, student characteristics (e.g., academic success measures, demographics, interests and attitudes), teacher characteristics (e.g., education, experience), and school-related factors (e.g., location, school type, infrastructure) might all be factors affecting implementation outcomes. Given this complexity, mainstream, particularly parametric, data analysis approaches can be prone to statistical errors.

To address aforementioned issues, we introduce a computational method for educators to predict longitudinal intervention outcomes. The predictive models target helping with decision making in the implementation of large-scale interventions. To achieve this goal, we use machine learning tools. Machine Learning enables computational algorithms to learn hidden patterns,

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

called *models*, and structures from observed data, and to predict unobserved data. Among various machine learning models, *Deep Learning models* are the most popular choices of today due to high accuracy of their prediction performance. Deep learning models are applied in various computational tasks requiring high accuracy, for example image classification, speech recognition, and language translation [10].

The use of deep learning to predict outcomes of educational interventions aligns with current perspectives in psychology and education that consider predictive accuracy as an important measure in evaluating how well a theory or model can account for the phenomenon studied, and with calls for wider use of machine learning methods in predictive models [11]. Traditional statistical models (e.g., regression, correlation) are prone to two main weaknesses: The first one, *overfitting*, refers to incorporating noise or causal relations specific to the sample (but not to the population) into the model. While an overfitted model can best explain the data at hand, it usually is not the one that best predicts future behavior [12]. The second, *p-hacking* is finding statistically significant effects that were not part of the a priori hypotheses [13]. Checking for statistical effects without a priori theorizing, and neither sufficiently controlling for multiple comparisons nor reporting statistical tests that did not produce significant effects are now widely cited as problematic. Given the multitude of data mining methods, one can almost surely find significant effects in most data sets, however these effects are likely to be due to sample-specific noise and are not generalizable. These issues with traditional statistical modeling have led to a “replication crisis,” especially in psychology and medicine, due to lack of replicability of findings [14]. The solutions proposed to address this crisis include more stringent protocols and conventions for research [13], for example Bayesian statistics and meta-analysis [15, 16], and use of predictive machine learning methods [11].

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

We introduce the deep learning method for educational research, with an illustrative case study that addresses the aforementioned issues. In the case study, we employed a recently developed multi-purpose deep learning toolkit, MATLAB's Deep Learning Toolbox [17]. We compared the predictive power of the deep learning tool with that of the traditional analysis methods, such as regression-based [18]. Such traditional methods would not perform well if the relationships between independent and dependent variables are nonlinear and complex [19, 20]. Thus, we expected that deep learning methods would show higher prediction accuracy by capturing more complex relationships across variables [21].

To test the performance of the deep learning method in the prediction of intervention outcomes, first, we created a simulation model by using data from an actual intervention experiment, collected from a classroom study examining the influence of different types of moral stories on students' prosocial behavior. Second, we trained the model through an iterative learning process to improve the prediction accuracy. Third, prediction accuracy, determined by comparing predicted and actual outcome variables, was evaluated once the iterative learning procedure was completed [22, 23]. Finally, we compared prediction accuracy and robustness against overfitting between the deep learning and traditional regression methods.

2. Material and Methods

2.1. Learning dataset

We used a dataset from a classroom intervention study by Han et al. [24]. This study compared the effectiveness of stories of close other moral exemplars, such as peer exemplars, and those of extraordinary exemplars, such as historic figures, among eighth graders [24]. 107 eighth graders (50 females) at a middle school located in Seoul Metropolitan area in Korea participated in the study.

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

The students were randomly assigned to two groups: peer exemplar (55 students) and historic figure groups (52 students). During an eight-week intervention session, students discussed stories of moral exemplars for an hour per week. The students assigned to the peer exemplar condition were asked to discuss their peer moral exemplars, such as friends or family members. Those in the historic figure group discussed historic moral figures, such as Mother Teresa and Martin Luther King. The study used students' engagement in voluntary service activity as an outcome variable. The study compared change in service engagement between pre-intervention and post-intervention time points. The pre-test engagement was surveyed before the beginning of the intervention period, and the post-test survey was conducted twelve weeks from the pre-test survey. In addition, to control for any possible effects due to individual differences in participants' mindset regarding moral development, the study measured their moral growth mindset based on their volunteer service engagement [25]. Finally, the study surveyed how participants responded to intervention materials and activities, by collecting data on moral elevation, perceived moral excellence, and perceived difficulty to emulate presented moral behaviors after the end of the intervention period. These variables were used to examine how students' responses influenced change in their service engagement. The overall descriptive statistics regarding the nature of the collected dataset are presented in Table 1.

<Place Table 1 about here>

2.2. Measures

2.2.1. Voluntary service engagement

Han et al. [24] used a previously developed voluntary service participation survey form [26–28] that measured service engagement in four domains: 1. Religion-related, 2. Charity-related, 3. Art-related, and 4. Child-adolescent-student-related service activity. Each question

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

asked the frequency of students' engagement in a specific domain during the last two months; their answer was quantified using a 5-point scale ("1. Never," "2. Once or twice," "3. A few times," "4. Almost every week" "5. At least once a week"). Each item was designed to assess one's engagement in a specific individual voluntary service activity domain.

In the case of the pre-test voluntary service engagement, we used the calculated average score as a composite variable for service engagement. However, in the case of the post-test voluntary service engagement, we converted the average score into a binary variable because the deep learning method was originally developed to predict categorical outcomes [29]. If a participant's average post-test voluntary service engagement score was 1, which represents "never," we assigned "0 (did not participate)." If the calculate average score was greater than 1, then we assigned "1 (participated)" to this case. Table 2 demonstrates participants who participated in service activities (1) and who did not (0) at the pre- and post-tests. Out of 107 participants, 27 participants (25.37%) changed their participation status (0 to 1 or 1 to 0).

<Place Table 2 about here>

2.2.2. Moral growth mindset

Before the intervention period, to examine whether a participant possessed either a growth or fixed mindset in the domain of morality, Han et al. [24] assessed their moral growth mindset with a questionnaire. Han et al. [25] reported that the presence of moral growth mindset, which is associated with a belief that one can be a morally better person through intentional efforts (e.g., participating in prosocial activities), was significantly associated with an increase in volunteering. Thus, we decided to control for this factor in our current analyses. To measure this construct, Han et al. [24] used a six-item questionnaire in Korean language inquiring whether a participant has a growth or fixed mindset (e.g., "no matter who you are, you can significantly

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

improve your morality and character”). Each item was rated on a seven-point Likert scale anchored at 1 = strongly disagree and 7 = strongly agree. The calculated Cronbach’s α was .73 indicating acceptable reliability.

2.2.3. *Responses to intervention activities*

After the end of the intervention period, Han et al. [24] surveyed students’ responses. They focused on how strongly each student was elevated by presented exemplars (moral elevation), how the presented exemplars were perceived to be morally excellent (perceived moral excellence) and difficult to emulate (perceived difficulty for emulation) by asking three questions (i.e., “how strongly were you emotionally (morally) touched by stories?” “did you think that persons presented in stories were morally excellent and better compared to yourself?” “did you think that it would be difficult to emulate the activities of persons presented in stories?”). An answer to each question was anchored to a 4-point Likert scale (1 = strongly disagree or extremely unlikely – 4 = strongly agree or extremely likely).

2.2.4. *Descriptive statistics and correlation analysis*

We summarized descriptive statistics, mean, standard deviation, skewness, and kurtosis values, in order to help readers better understand the nature of the dataset used for deep learning. In addition, we examined correlation among dependent and independent variables in order to show the relationships across the factors.

In addition to the conventional correlation analysis, based on the null hypothesis significance testing, which is prone to several statistical issues, such as relatively high frequency of false positives [30], we performed additional Bayesian t-tests with JASP [31]. Recently, there have been debates about whether the conventional threshold for P -values, $p < .05$, is a reliable and valid threshold for statistical decision-making [32]. In particular, we acknowledge that

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

Bayesian inference provide a more reliable way of testing the strength of evidence supporting hypotheses in educational research, compared to the null hypothesis significance testing [15]. Hence, in addition to P -values indicating significance in conventional correlation analysis, we refer to Bayes factor (BF) values from Bayesian correlation analysis, indicating the strength of evidence supporting an alternative hypothesis (H_1 : the presence of an actual difference) and the null hypothesis (H_0 :the absence of a difference). $2\log\text{BF}$ (twice the natural logarithm of BF) < 2 indicates absence of any significant evidence against H_0 , $2 \leq 2\log\text{BF} < 6$ indicates presence of positive evidence against H_0 , $6 \leq 2\log\text{BF} < 10$ indicates presence of strong evidence against H_0 , and $2\log\text{BF} \geq 10$ indicates presence of very strong evidence against H_0 according to conventional guidelines for Bayesian statistics [33].

2.3.Prediction modeling using deep learning

2.3.1. *Deep learning*

We can apply machine learning algorithms to develop a data-driven prediction model, by training the model with experimental data. Among various machine learning algorithms, an artificial neural network with multiple layers of neurons, or simply *deep learning*, is most widely used due to its superior performance in many classical applications, for example image classification, object recognition, and speech recognition. The deep architecture of deep learning corresponds to a hierarchy of features, factors, or concepts, where higher-level features are defined from lower-level ones, and the same lower-level features can help to define diverse higher-level features [34].

Using MATLAB's Deep Learning Toolbox [17], we trained a two-layer convolutional neural network for predicting the intervention outcomes. More precisely, we used the pre-test variables (i.e., service engagement, moral growth mindset, gender, intervention type, responses

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

to intervention activity) as inputs, and the post-test outcome indicator (i.e., whether engaged in service activities at the post-test) as an output (see Figure 1 for the network model structure).

<Place Figure 1 about here>

2.3.2. *Basic concepts and terminologies in the deep learning*

Before we introduce our deep learning model, we briefly describe some technical terminologies. An Artificial Neural Network (ANN) is a network of neurons. In this network, each neuron is a simple computational unit that receives inputs and generates outputs based on some simple rules (functions). Each neuron has an input and an output port, and the value of each neuron is obtained as follows. Consider a neuron, say A . Denote the neurons that are connected to the input port of neuron A by B_1, B_2, \dots, B_k . Then, the value of the neuron A is determined as:

$$v_A = f(\sum B_i W_i + c)$$

where W_i is the weight of the edge connecting neuron B_i and neuron A , and c is a bias. Here, $f(\cdot)$ is called the activation function, which determines the final value of the neuron given the input signals. A typical choice of the activation function is:

$$f(x) = \max(0, x)$$

which is called the ReLU (Rectified Linear Unit) function.

In most applications of deep learning, one considers ANNs where neurons are grouped together to form a ‘layer’, and these layers are connected in systematic ways. For instance, a layer is called ‘fully-connected’ if all the neurons of the previous layer are connected to the input port of each neuron of the layer. Another important layer structure is convolutional layers. Roughly speaking, a small number of the neurons of the previous layer are connected to each of

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

the convolutional layers, and all the neurons of the convolutional layer share the same weights for the edges connected to their input ports.

For notational simplicity, it is common to denote the values of all neurons in a single layer by a single variable. For instance, if the neurons are arranged in a single dimension, one can define a vector, each element of which represents the value of each neuron. If the neurons are arranged in two dimensions, for example as it is for images, a matrix can be used to represent the values of the array. If the neurons are aligned in a high-dimensional space, one can use a high-dimensional array to describe the values of the neurons. In computer science, such high-dimensional arrays are called ‘tensors’ these days. One may notice that such representations are not unique. The same layer can be viewed and represented in many different ways. For instance, a 2D layer (an $n \times m$ matrix) can be viewed as a 1D vector of length nm , and be represented as a 1D tensor. Such an operation is called ‘reshaping’, and it always maintains the information without altering the contents of the tensor.

2.3.3. *Network modeling*

We trained a deep-learning model using MATLAB through an iterative algorithm [17]. All the parameters used in our simulation, such as the number of layers of neurons and the shape of weight tensors, were determined based on the TensorFlow tutorial available online, MNIST For ML Beginners [29], and the MATLAB Deep Learning Toolbox tutorial [17].

First, we setup a convolutional network consisting of multiple convolutional and fully connected layers. The model was designed to receive six input variables, one for each student characteristic: gender, pre-test service engagement, experienced elevation, perceived excellence, perceived difficulty for emulation, moral growth mindset. This model predicted one, binomial, outcome variable: post-test service engagement. When the input variables were entered to the

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

model, they were reshaped into a vector. The reshaped vector was convolved with a weight tensor with a dimension of (2, 32) with a same padding. In addition, a bias vector with a component for each output channel was set and added to the convolved result. The output tensor was then applied to the ReLU function:

$$f(x) = \max(0, x)$$

Then, a max pooling layer was added. The max pooling layer consisted of a pool size (1, 1) and a stride (2, 2).

The same procedure was repeated at the second convolutional layer. The resultant tensor from the first convolutional layer was entered to the second layer. At this layer, we used a weight tensor with a dimension of (2, 128) with a same padding; a max pooling layer with a pool size (1, 1) and a stride (2, 2) was also added. Moreover, we added the third convolutional layer. A weight tensor with a dimension of (2, 512) with a same padding was used at the third layer. The output tensor from the third convolutional layer was fed into a fully-connected layer with two neurons. The outputs from the fully-connected layer were forwarded to a Softmax layer for classification. The Softmax layer was used to calculate the likelihood of each category in terms of an exponential probability distribution [35]. In our study, there were two possible categories, participating in any service activity at the post-test (1) and not doing so (0). Thus, the Softmax layer in our study had two output neurons; each representing the probability of whether a student was likely (1) or not likely (0) to participate in any service activity at the post-test.

Second, the predicted outcomes were compared with the actual outcomes in the experimental data, by calculating the cross entropy between them [36]. A predicted probability vector was used for the calculation of the cross entropy. Minimizing the cross-entropy term

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

between the predicted probability and the actual labels is equivalent to maximizing the likelihood of the observed labels, given the predicted probabilities.

In order to minimize the cross entropy, a specific optimization algorithm needs to be selected. However, it is not clear what the best optimizer for deep neural networks is, and finding the best optimizer is one of the important open questions in the deep learning research. There are several popular algorithms that are theoretically and empirically shown to perform well in many deep learning scenarios, and these include SGD with momentum, RMSProp and ADAM, which are briefly described below.

First, SGD, an acronym for Stochastic Gradient Descent, with momentum (SGDM) is an optimization algorithm that estimates the current objective value from a randomly sampled data point (or a mini-batch of data points) and then updates its optimization variables based on the gradient computed with respect to the current sample(s) [37]. SGD with momentum improves the stability of SGD by incorporating the exponential average of past gradients when updating optimization variables. Second, RMSProp makes use of the exponential moving average of past squared gradients to better normalize gradient updates across different parameters [38]. Third, ADAM makes use of both the exponential moving average of past gradients and the exponential moving average of past squared gradients to combine benefits of SGD with momentum and RMSProp [39]. The ADAM optimizer is the most widely used optimizer in most of the deep learning applications, and it was shown to achieve good results in many different cases. Following MATLAB default settings, we used the initial learning rate of .001 for ADAM and RMSProp optimizers and .01 for SGD optimizer.

We evaluated the model accuracy. Because the calculated outcome variables were in the form of probability of two states, yes or no, stored in two output neurons in the Softmax layer,

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

we transformed these into one final output variable. If the value stored in the output neuron representing the probability of service participation was greater than that in the other output neuron, representing the probability of non-participation, then the final output value became 1 (participated). If not, then the final output value became 0 (not participated). Once the final output value was calculated, we compared this predicted value with the actual post-test engagement value in the survey data.

During the learning process, we used the 3-fold cross-validation method. We randomly assigned about one-third ($n = 35$) of the entire dataset to a validation dataset, which was not used for learning. This portion of dataset was only used for evaluation, to prevent overfitting, which is associated with whether the trained model is excessively fitted to the learning data, and can predict outcomes within the boundary of the provided learning dataset, but cannot reliably predict real-world cases beyond the boundary [23]. Thus, we did not include the randomly selected validation cases in the training dataset ($n = 72$) and tested to what extent the trained model can predict the validation cases. We sought for the optimal model, which was most robust against overfitting, by applying the early stopping method. For each iterated epoch, we evaluated the accuracy and loss function to determine when an early stopping should occur. In MATLAB, when there are n_y categorical outputs, the default loss function is defined in the form of:

$$V(\theta) = \frac{1}{N} e^T(t, \theta) W(\theta) e(t, \theta)$$

When N is the size of the dataset, $e(t, \theta)$ is n_y -by-1 error vector at a specific time point t parameterized by the parameter vector θ , and $W(\theta)$ is the weighting matrix in the learned neural network model [40].

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

The source code for the simulation written in MATLAB is available via the GitHub: https://github.com/xxelloss/TensorFlow_Intervention [41]. In addition, the data file is available in csv format at https://github.com/xxelloss/TensorFlow_Intervention/blob/master/oxtest1.csv [41]. All source codes contain line-to-line comments to facilitate their modification for practice purposes and for applicability to other datasets.

2.3.4. *Prevention of overfitting and early stopping*

The prediction performance usually decreases after a certain number of iterations. This happens because the prediction model starts capturing noise in the training data set, leading to a decrease in prediction performance, beyond a certain point of model-fitting. This phenomenon is called *overfitting*, meaning that the model is overly fit to the given training data set, including noise. One simple way to avoid overfitting is to use *early stopping*, stopping training the model as soon as overfitting is observed when the validation dataset is entered. To implement this method, we examined the change in loss function for each epoch [42]. When the loss instances on the validation set occurred more than three epochs, we stopped the ongoing training process. Once each run was completed, we calculated prediction accuracy by comparing predicted and actual outcomes for both the learning ($n = 72$) and randomly selected validation dataset ($n = 35$).

2.3.5. *Accuracy evaluation*

We compared the prediction accuracy of the deep learning model with that of logistic regression, structural equation modeling (SEM), and mixed-effects logistic regression, statistical models typically used in educational research. The three methods were selected based on the previous studies that originally reported on the dataset that used in the present study. SEM was performed in [25] and logistic regression analyses were performed in [24].

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

The statistical analyses for the accuracy evaluation were performed with R. All R scripts are also publicly available in the project GitHub repository. In the case of simple logistic regression, we used the post-test service engagement coded in a binary dependent variable, and the pre-test service engagement (in hours), moral growth mindset, gender, intervention condition, moral elevation, perceived moral excellence, and perceived difficulty for emulation as independent variables. Only the main effects were entered to the model, following the models that were used in the studies that originally reported on the dataset used in the current study [24, 25]. R's *glm* package was used for the logistic regression analysis [43].

For mixed-effects logistic regression, we used the same variables used for the simple logistic regression; gender was set as a random effect while the other independent variables were set as fixed effects or covariates. We used *glmer* package for the mixed-effects logistic regression analysis [44]. Similar to the case of the simple logistic regression analysis, we only used the main effects of the independent variables, following the prior studies.

For the SEM, we set a hypothetical model presented in Figure 2. We examined whether the model showed good model fit indicators (i.e., χ^2 statistics, root mean squared error of approximation (RMSEA), comparative fit index (CFI)); if the indicators were not satisfactory, we modified the hypothetical model by referring to modification indices provided by R. We used the *lavaan* package for SEM implemented in R [45]. Likewise, we used only the main effects of the independent variables paralleling the previous study (Han et al., 2017).

Furthermore, we employed different types of deep learning methods to find the optimal deep learning method that can produce the highest prediction accuracy. First, we employed three different optimizers (SGDM, RMSProp, and ADAM) to examine which optimizer would produce the best performance. Second, we compared the prediction accuracy between when

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

convolutional layers were included in the network model and when only the Softmax layer was used, without the convolutional layers, to examine whether the deep learning structure contributes to the improvement of prediction accuracy.

We created the aforementioned three statistical models predicting the post-test service engagement and compared the prediction accuracy of the deep learning model with that of the three traditional statistical models. First, we compared prediction accuracy when the learning dataset ($n = 72$) was used for the deep learning and model estimation. Second, we compared prediction accuracy when the validation dataset ($n = 35$), which was not used for model learning, was entered to the learned model. This analysis process with the validation dataset was used to examine whether the overfitting issue occurred during the model learning process. We performed the learning and validation processes thousand times for each method to collect sufficient results for statistical analysis of performance.

To compare the performance of the different statistical methods, first, we conducted an omnibus ANOVA to examine whether the type of modeling method significantly contributed to the differences in prediction accuracy. Two omnibus ANOVAs were performed for training and validation datasets. Because the variable indicating the type of specific prediction method was nested, we used a nested ANOVA with the *nlme* package in R [46]. Under the category of the deep learning methods, there were six specific types of prediction methods that were differentiated in terms of an optimizer (SGDM, RMSProp, and ADAM) and whether or not the convolutional layers were used (with convolutional layers vs. only with Softmax). In the case of the classical non-deep learning regression methods, there were three methods, logistic regression, SEM, and mixed-effects logistic regression. Thus, we set the main effect of the prediction

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

method (one of the aforementioned nine methods) as the independent variable and the calculated prediction accuracy as the dependent variable while performing ANOVA.

Second, we performed planned t -tests to test whether the deep learning method showed better prediction accuracy compared with the three classical regression methods. These t -tests were performed for both the learning and validation datasets. In addition to the t -tests, we also examined the effect sizes in each comparison in terms of Cohen's D . $D \geq .2$ refers to a small effect, $D \geq .5$ a medium effect, and $D \geq .8$ a large effect [47]. In addition to the conventional t -tests, we performed additional Bayesian t -tests with BayesFactor package in R [48].

In addition to the evaluation of prediction accuracy, we also evaluated sensitivity and selectivity of the tested methods with receiver operating characteristic (ROC) analysis. For quantitative comparison, we compared the area under the ROC curve (AUC) between the deep learning and traditional regression analysis methods. In the case of the evaluation of the deep learning method, we focused on one deep learning condition that showed the optimal prediction accuracy outcome. Thus, for the AUC comparison, four methods (i.e., the best deep learning method, logistic regression, mixed-effect logistic regression, SEM) were compared. We compared the mean AUCs between the aforementioned four methods.

3. Results

3.1.Descriptive statistics and correlation analysis

Descriptive statistics, mean, standard deviation, skewness, and kurtosis values, are demonstrated in Table 1. The result of the correlation analysis is presented in the same table.

3.2.Learning dataset prediction accuracy

<Place Table 3 and Figure 3 about here>

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

We examined prediction accuracy when the learning dataset ($n = 72$) was entered. The omnibus nested ANOVA model reported a significant main effect of prediction method (three deep learning, three only with Softmax, three traditional regression methods), $F(8, 8,987) = 1,576.50, p < .001$. The results from the t -tests also reported significant differences. The best prediction accuracy was achieved with the deep learning method with the ADAM optimizer when the learning dataset was used (See Table 3 and Figure 3). Although other two optimizers demonstrated a slightly worse accuracy compared with the ADAM optimizer, overall, the deep learning method significantly outperformed all traditional statistical methods. However, when only the Softmax layer was used without any convolutional neural network, the outcome performance of our learning method became significantly worse than the three traditional statistical methods.

3.3. Validation dataset prediction accuracy and robustness against overfitting

We also examined prediction accuracy when the validation dataset was entered, to test whether the model was robust against overfitting (see Table 3). The performed omnibus nested ANOVA reported a significant main effect of prediction method when the nine prediction methods (three deep learning, three only with Softmax, three traditional regression methods), $F(8, 8,987) = 126.50, p < .001$. When the planned t -tests were performed, the deep learning methods showed accuracy that was not significantly different from that of the traditional regression methods when convolutional layers and ADAM optimizer were used. The result suggests that when the deep learning method was used with ADAM optimizer, its prediction accuracy did not become worse than that of traditional regression analysis methods.

3.4. ROC and AUC comparison

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

We compared AUCs between the deep learning method with ADAM optimizer that showed the best prediction accuracy and three traditional regression methods. The plotted ROC curves are presented in Figure 4. The deep learning method showed the highest mean AUC, .91 ($SD = .03$). The mean AUC of logistic regression was .84 ($SD = .01$), that of mixed-effects logistic regression was .84 ($SD = .02$), and that of SEM was .80 ($SD = .00$). Both classical and Bayesian t -tests between the deep learning and three traditional regression methods resulted in $p < .001$ and $2\log BF > 10$. These result indicated that the deep learning method showed the significantly higher mean AUC compared with the three other methods.

<Place Figure 4 about here>

4. Discussion

In the present study, we examined deep learning as a possible method to model and predict potential outcomes of educational interventions based on relatively small-scale data. We showed that simulated results from deep learning models can outperform the result from traditional statistical methods, logistic regression, SEM, and mixed-effects logistic regression, in terms of prediction accuracy. Moreover, we found that the early stopping method addressed the overfitting issue successfully in our study. The results from the performance comparisons using validation data demonstrated that deep learning was able to show the prediction accuracy that was not significantly different compared with the traditional regression analysis methods when ADAM optimizer was used (see results for deep learning with ADAM optimizer in Table 3).

Interestingly, although the dataset used in the present study was small ($N = 107$), our prediction model showed the enhanced performance when the deep learning methods were used. In general, researchers in the field have considered that the deep learning methods are suitable for analyzing large datasets and the majority of studies have been conducted with bigdata [49].

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

Even if this might be the case in general, several researchers, for example those who are interested in applying the deep learning methods in medical and clinical science, have been trying to test whether the methods can contribute to improving automatized diagnosis procedures with relatively small datasets [50]. In their studies, they have found that even with small datasets, the deep learning methods can more accurately predict clinical outcomes compared with traditional analysis methods and they can be well used in the aforementioned context dealing with small datasets [51–53]. Hence, consistent with the prior research in clinical and medical science, our study that showed the enhanced performance of the deep learning method with a small dataset may suggest that the deep learning method can contribute to simulating outcomes of educational interventions, which is frequently conducted with small datasets collected in lab or classroom settings.

Machine learning algorithms have the ability to learn from and make predictions about a dataset, without the need to explicitly organize or structure the internal dynamics of a neural network model. Any quantifiable data can be used as input, and after a training period, the algorithm can predict outcomes (dependent variables) within a specific margin of error. The goal of the training period is to decrease the margin of error. Cross-validating the algorithm requires a new set of data, one the model was not trained with, to test if the model can predict the dependent variables or outcomes within a desired threshold. During the training, when the network output does not match the expected values, the algorithm changes the weights among the neurons in the network by implementing a backpropagation, supervision, or reinforcement approach. Once trained, the weights among the neurons are fixed, and a different sub-set of the data is used to test the success of the network in predicting the values for the dependent variables. This approach not only allows finding patterns in complex data, and predicting

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

outcomes given a set of independent measures, but also shifts the goal of analysis from finding significant effects to developing a model that can flexibly explain the gradual contribution of different factors to the outcomes of interest. We have demonstrated that the deep learning method can more accurately predict longitudinal behavioral outcomes of interventions compared to traditional statistical methods. Thus, we propose a more rigorous and feasible way for longitudinal predictive modeling in behavioral research in general, including in psychology, education, and policy-making.

Interdisciplinary efforts in educational research incorporate a wide-range of indicators, from psychophysiological measures (e.g., cortisol level, neural measures) to socio-cultural and economic indicators (e.g., levels of education, socio-economic status) to understand the effects of and to predict the outcomes for educational interventions [54]. Connecting data across multiple levels (e.g., genetic, neural, behavioral, socio-cultural) and developing a model that not only shows whether a specific intervention leads to successful outcomes, but also predicts the outcomes of the same intervention under different conditions is highly valuable [54]. Implementing an intervention at a large scale is costly and requires convincing stakeholders (e.g., teachers, parents, administrators, and policy makers) about the intervention's effectiveness. Therefore, models that can predict outcomes in a specific context based on a wide-range of indicators can be crucial in decision making.

It is important to think about where we apply the deep learning method in the wider context of an intervention study. Here, we propose a model to help consider what would precede the application of a deep learning model, to make sense of complex patterns and predict outcomes of an intervention, and how the insights acquired from the deep learning study can be used to inform decision making.

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

<Place Figure 5 about here>

Even though here we illustrate a limited use of the deep learning model, we conceive use of deep learning models as part of a larger project that includes multiple steps, each step involving a different scale of design and implementation (see Figure 5). The first step of the model involves a design-based study, where ideas for the design of the intervention is tested out on small groups of participants [55]. This step involves iterative cycles of design and implementation. The insights gathered from successive trials of design and implementation would help refine the intervention, and prepare it for the next step. In the next pilot step, the intervention is implemented in the authentic context with the target population (e.g., small number of classrooms). Finally, the large-scale intervention study includes, for example, multiple classrooms across different schools, which would inform the large-scale use and outcomes of the intervention in a wide range of authentic contexts.

The development of the deep learning model starts early on, with prototypes using the data from the pilot study. The generalizability and predictive power is low at this stage but these early pilots can help with designing the initial architecture of the model. The final training and testing of the model takes place after the completion of the large-scale implementation study. Once completed, the predictive model can be used to inform scaled interventions across different contexts. The usefulness of the predictive model relies on how much the stakeholders and decision makers (e.g., researchers, superintendents, principals, teachers) are informed about what the model has to offer them, and how much they incorporate insights acquired from the model in their decision making. Efforts in bringing machine learning techniques in analyzing the data, and developing predictive models can be thought as part of a wider trend in incorporating “big data” in decision making [56].

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

Further studies are required to address several issues that might limit the application of the prediction method proposed in our study, in diverse contexts. First, we only tested the deep learning model with a specific dataset, the moral story intervention dataset, as an illustrative example. Future studies should test the model with other complex datasets. Second, related to the first point, we tested the deep learning prediction method with a dataset demonstrating longitudinal changes in prosocial behavior twelve weeks after the last intervention session, which might not be a sufficiently long-period to examine long-term outcomes. Although the present study showed that the deep learning method is suited to predicting long-term outcomes, it is necessary to replicate the present study with other long-term intervention datasets. Third, only one binominal variable was used for the dependent variable for the prediction. The simulation model should be modified and upgraded to predict various forms outcome variables, such as multiple continuous variables. Fourth, because programming skills are required to customize the current simulation program for other purposes, educators and policy makers without coding skills will not be able to utilize the simulation program without outside help. Thus, a graphical user interface should be developed to provide such potential users with user-friendly access to the simulation model. Fourth, because the deep learning method is suitable for the prediction of categorical outcome variables, such as pattern recognition and classification [10], such as the binary variable that was used in the present study, it would be difficult to predict continuous outcome variables with the deep learning method. Future research should focus on use of other machine learning methods (e.g., ensemble learning and dimensionality reduction algorithms), with datasets that involve more complex outcome variables.

5. Conclusion

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

In the present study, we examined whether the deep learning methods can improve the accuracy while predicting outcomes of educational interventions. We compared prediction accuracy between the deep learning and traditional regression methods with a small-scale educational intervention dataset. We found that the deep learning methods reported better prediction accuracy compared with the traditional regression methods. Although there were several methodological limitations that should be addressed by future studies, in the present study, we were able to show that the deep learning methods can be applied to research on educational psychology and potentially contribute to educational program designing and policy-making by allowing educators and educational researchers to simulate outcomes of educational interventions more accurately.

Funding: This study was not funded by any funding sources.

Conflict of Interest: The authors declare that they have no conflict of interest.

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

References

1. Yeager DS, Trzesniewski KH, Dweck CS (2013) An implicit theories of personality intervention reduces adolescent aggression in response to victimization and exclusion. *Child Dev* 84:970–988. doi: 10.1111/cdev.12003
2. Yeager DS, Trzesniewski KH, Tirri K, et al (2011) Adolescents' implicit theories predict desire for vengeance after peer conflicts: Correlational and experimental evidence. *Dev Psychol* 47:1090–1107.
3. Cohen GL, Garcia J, Purdie-Vaughns V, et al (2009) Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science* (80-) 324:400–403. doi: 10.1126/science.1170769
4. Walton GM, Cohen GL, Cwir D, Spencer SJ (2012) Mere belonging: The power of social connections. *J Pers Soc Psychol* 102:513–532. doi: 10.1037/a0025731
5. Brannon T, Walton G (2013) Improving Attitudes by Enacting Interests: How Intergroup Contact Can Spark Interest in an Outgroup's Culture and Reduce Prejudice. *Psychol Sci* 24:1947–1957.
6. Yeager DS, Walton GM (2011) Social-psychological interventions in education: They're not magic. *Rev Educ Res* 81:267–301. doi: 10.3102/0034654311405999
7. Yeager DS, Romero C, Paunesku D, et al (2016) Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *J Educ Psychol* 108:374–391. doi: 10.1037/edu0000098
8. Yeager DS, Fong CJ, Lee HY, Espelage DL (2015) Declines in efficacy of anti-bullying programs among older adolescents: Theory and a three-level meta-analysis. *J Appl Dev Psychol* 37:36–51. doi: 10.1016/j.appdev.2014.11.005

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

9. Brown AL (1992) Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *J Learn Sci* 2:141–178. doi: 10.1207/s15327809jls0202_2
10. LeCun Y, Bengio Y, Hinton G, et al (2015) Deep Learning. *Nature* 521:436–444. doi: 10.1038/nature14539
11. Yarkoni T, Westfall J (2017) Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci* 53:174569161769339. doi: 10.1177/1745691617693393
12. Shmueli G (2010) To Explain or to Predict? *Stat Sci* 25:289–310. doi: 10.1214/10-STS330
13. Simmons JP, Nelson LD, Simonsohn U (2011) False-Positive Psychology. *Psychol Sci* 22:1359–1366. doi: 10.1177/0956797611417632
14. Loken E, Gelman A (2017) Measurement error and the replication crisis. *Science* (80-) 355:584–585. doi: 10.1126/science.aal3618
15. Han H, Park J, Thoma SJ (2018) Why do we need to employ Bayesian statistics and how can we employ it in studies of moral education?: With practical guidelines to use JASP for educators and researchers. *J Moral Educ* 47:519–537. doi: 10.1080/03057240.2018.1463204
16. Han H, Park J (2019) Bayesian meta-analysis of fMRI image data. *Cogn Neurosci* 10:66–76. doi: 10.1080/17588928.2019.1570103
17. MathWorks (2018) Deep Learning Toolbox.
18. Black JE, Reynolds WM (2016) Development, reliability, and validity of the Moral Identity Questionnaire. *Pers Individ Dif* 97:120–129. doi: 10.1016/j.paid.2016.03.041
19. Han H, Lee K, Soyulu F (2016) Predicting long-term outcomes of educational interventions

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

- using the Evolutionary Causal Matrices and Markov Chain based on educational neuroscience. *Trends Neurosci Educ* 5:157–165. doi: 10.1016/j.tine.2016.11.003
20. Han H, Lee K, Soylu F (2018) Simulating outcomes of interventions using a multipurpose simulation program based on the evolutionary causal matrices and Markov chain. *Knowl Inf Syst*. doi: 10.1007/s10115-017-1151-0
 21. Bengio Y (2009) Learning Deep Architectures for AI. *Found Trends® Mach Learn* 2:1–127. doi: 10.1561/22000000006
 22. Srivastava N, Hinton G, Krizhevsky A, et al (2014) Dropout: prevent NN from overfitting. *J Mach Learn Res* 15:1929–1958. doi: 10.1214/12-AOS1000
 23. Srivastava N, Hinton GE, Krizhevsky A, et al (2014) Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* 15:1929–1958. doi: 10.1214/12-AOS1000
 24. Han H, Kim J, Jeong C, Cohen GL (2017) Attainable and Relevant Moral Exemplars Are More Effective than Extraordinary Exemplars in Promoting Voluntary Service Engagement. *Front Psychol* 8:283. doi: 10.3389/fpsyg.2017.00283
 25. Han H, Choi Y-J, Dawson KJ, Jeong C (2018) Moral growth mindset is associated with change in voluntary service engagement. *PLoS One* 13:e0202327. doi: 10.1371/journal.pone.0202327
 26. Crocetti E, Jahromi P, Meeus W (2012) Identity and civic engagement in adolescence. *J Adolesc* 35:521–532.
 27. Porter TJ (2013) Moral and political identity and civic involvement in adolescents. *J Moral Educ* 42:239–255. doi: 10.1080/03057240.2012.761133
 28. Malin H, Han H, Liauw I (2017) Civic purpose in late adolescence: Factors that prevent

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

- decline in civic engagement after high school. *Dev Psychol.* doi: 10.1037/dev0000322
29. Google (2017) MNIST For ML Beginners.
 30. Wagenmakers E-J (2007) A practical solution to the pervasive problems of p values. *Psychon Bull Rev* 14:779–804. doi: 10.3758/BF03194105
 31. Love J, Selker R, Marsman M, et al (2017) JASP (Version 0.8.2).
 32. Benjamin DJ, Berger JO, Johannesson M, et al (2018) Redefine statistical significance. *Nat Hum Behav* 2:6–10. doi: 10.1038/s41562-017-0189-z
 33. Kass RE, Raftery AE (1995) Bayes Factors. *J Am Stat Assoc* 90:773–795. doi: 10.2307/2291091
 34. Deng L, Yu D (2014) Deep learning: Methods and applications. *Found Trends® Signal Process* 7:197–387. doi: 10.1561/20000000039
 35. Tang Y (2013) Deep Learning using Linear Support Vector Machines.
 36. Huang P-S, He X, Gao J, et al (2013) Learning deep structured semantic models for web search using clickthrough data. In: *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM '13*. ACM Press, New York, New York, USA, pp 2333–2338
 37. Qian N (1999) On the momentum term in gradient descent learning algorithms. *Neural Networks* 12:145–151. doi: 10.1016/S0893-6080(98)00116-6
 38. Hinton G, Srivastava N, Swersky K (2012) Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.
 39. Kingma D, Ba J (2014) Adam: A Method for Stochastic Optimization. <http://arxiv.org/abs/1412.6980>.
 40. MathWorks (2019) Loss function and model quality metrics.
 41. Han H, Lee K, Soylu F (2017) TensorFlow_Intervention.

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

- https://github.com/xxelloss/TensorFlow_Intervention.
42. Ruder S (2016) An overview of gradient descent optimization algorithms.
 43. R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
 44. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. doi: 10.18637/jss.v067.i01
 45. Rosseel Y (2012) lavaan : An R Package for Structural Equation Modeling. J Stat Softw. doi: 10.18637/jss.v048.i02
 46. McDonald JH (2014) Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, MD
 47. Cohen J (1992) A power primer. Psychol Bull 112:155–159. doi: 10.1037/0033-2909.112.1.155
 48. Morey RD, Rouder JN, Jamil T, et al (2018) Package ‘BayesFactor.’ <https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf>.
 49. Chou C-N, Shie C-K, Chang F-C, et al (2019) Representation learning on large and small data. In: Vronchidis S, Huet B, Chang EY, Kompatsiaris I (eds) Big Data Anal. Large-Scale Multimed. Search. Wiley, Hoboken, NJ, pp 3–30
 50. Hekler EB, Klasnja P, Chevance G, et al (2019) Why we need a small data paradigm. BMC Med 17:133. doi: 10.1186/s12916-019-1366-x
 51. Ahmed K Ben, Hall LO, Liu R, et al (2019) Neuroimaging Based Survival Time Prediction of GBM Patients Using CNNs from Small Data. In: 2019 IEEE Int. Conf. Syst. Man Cybern. IEEE, pp 1331–1335
 52. He G (2019) Lung CT Imaging Sign Classification through Deep Learning on Small Data.

COMPUTER SIMULATION FOR EDUCATIONAL RESEARCH

ArXiv

53. Aydin F, Zhang M, Ananda-Rajah M, Haffari G (2019) Medical Multimodal Classifiers Under Scarce Data Condition. ArXiv
54. Han H, Soylu F, Anchan DM (2019) Connecting Levels of Analysis in Educational Neuroscience: A Review of Multi-level Structure of Educational Neuroscience with Concrete Examples. Trends Neurosci Educ 100113. doi: 10.1016/j.tine.2019.100113
55. Barab S, Squire K (2004) Design-Based Research: Putting a Stake in the Ground. J Learn Sci 13:1–14. doi: 10.1207/s15327809jls1301_1
56. Liu M-C, Huang Y-M (2017) The use of data science for education: The case of social-emotional learning. Smart Learn Environ 4:1. doi: 10.1186/s40561-016-0040-4

Tables

Table 1

Classic and Bayesian correlation analysis of dependent and independent variables

| | <i>M</i> | <i>SD</i> | Skewness | Kurtosis | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---------------------------------|----------|-----------|----------|----------|----------|----------|----------|----------|---------|---------|------|
| 1. Pre-test service engagement | 1.40 | .44 | 1.33 | 1.54 | — | | | | | | |
| 2. Post-test service engagement | .64 | .48 | -.57 | -1.71 | .39***‡ | — | | | | | |
| 3. Gender | 1.53 | .50 | -.13 | -2.02 | -.02 | .00 | — | | | | |
| 4. Group assignment | .49 | .50 | -.06 | -2.04 | .07 | .31**† | -.01 | — | | | |
| 5. Elevation | 3.07 | .77 | -.62 | .27 | .12 | 0.22* | -.21* | .06 | — | | |
| 6. Perceived excellence | 3.56 | .68 | -1.82 | 3.95 | .08 | -.09 | -.19* | -.08 | .55***‡ | — | |
| 7. Perceived difficulty | 2.38 | .88 | -.15 | -.79 | -.33***† | -.34***† | .09 | -.52***‡ | -.16 | -.06 | — |
| 8. Moral growth mindset | 4.77 | 1.30 | -.44 | -.38 | .17 | .13 | -.35***† | .03 | .39***‡ | .36***‡ | -.13 |

Note. Post-test service engagement: 0: not-engaged, 1: engaged. Gender: 1: female, 2: male. Group assignment: 0: historic

figure condition, 1: peer exemplar condition. * $p < .05$, ** $p < .01$, *** $p < .001$. † $5 \leq 2\log\text{BF} < 10$, ‡ $2\log\text{BF} \geq 10$.

Table 2

Voluntary service participation statuses at the pre- and post-test

| | Pre-test | Post-test | | Total |
|--------------------------|----------|------------------|-------------------------|-------|
| | | Participated (1) | Did not participate (0) | |
| Participated (1) | | 56 | 15 | 71 |
| Did not participated (0) | | 12 | 24 | 36 |
| Total | | 68 | 39 | 107 |

Table 3

Comparisons of prediction accuracy among deep learning and logistics regression, SEM, and mixed-effects logistic regression

| | | Deep-learning Accuracy | | Accuracy of classical methods and comparisons with deep-learning | | | | | | | | | | | | | | |
|----------------|---------|------------------------|-----------|--|-----------|-----------|---------------|----------|----------------------------------|-----------|-----------|---------------|----------|----------|-----------|----------|---------------|----------|
| | | | | Logistic regression | | | | | Mixed-effect logistic regression | | | | | SEM | | | | |
| | | <i>M</i> | <i>SD</i> | <i>M</i> | <i>SD</i> | <i>t</i> | <i>2logBF</i> | <i>D</i> | <i>M</i> | <i>SD</i> | <i>t</i> | <i>2logBF</i> | <i>D</i> | <i>M</i> | <i>SD</i> | <i>t</i> | <i>2logBF</i> | <i>D</i> |
| Learning set | | | | | | | | | | | | | | | | | | |
| Softmax | ADAM | 75.85% | 8.12% | | | -10.02*** | 91.30*** | -.45 | | | -12.73*** | 148.95*** | -.57 | | | 27.53*** | 634.05*** | 1.23 |
| | SGDM | 76.91% | 8.32% | | | -6.10*** | 30.61*** | -.27 | | | -8.95*** | 72.05*** | -.40 | | | 29.89*** | 729.71*** | 1.34 |
| | RMSPROP | 75.72% | 8.28% | 78.65% | 7.35% | -10.31*** | 96.93*** | -.46 | 79.59% | 4.52% | -12.98*** | 154.79*** | -.58 | 64.87% | 9.65% | 26.99*** | 612.71*** | 1.21 |
| Deep learning | ADAM | 94.79% | 2.16% | | | 76.59*** | Inf*** | 3.43 | | | 66.40*** | Inf*** | 2.97 | | | 84.61*** | Inf*** | 3.78 |
| | SGDM | 89.26% | 8.20% | | | 37.57*** | 1058.40*** | 1.68 | | | 32.65*** | 845.31*** | 1.46 | | | 60.90*** | Inf*** | 2.72 |
| | RMSPROP | 89.99% | 4.37% | | | 43.97*** | 1342.72*** | 1.97 | | | 38.11*** | 1082.09*** | 1.70 | | | 65.49*** | Inf*** | 2.93 |
| Validation set | | | | | | | | | | | | | | | | | | |
| Softmax | ADAM | 69.01% | 7.32% | | | -10.49*** | 100.53*** | -.47 | | | -5.52*** | 24.01*** | -.25 | | | 15.15*** | 210.11*** | .68 |
| | SGDM | 69.52% | 8.26% | | | -8.24*** | 60.39*** | -.37 | | | -3.64*** | 7.16** | -.16 | | | 15.71*** | 225.57*** | .70 |
| | RMSPROP | 69.32% | 7.70% | 72.25% | 6.44% | -9.22*** | 76.68*** | -.41 | 70.74% | 6.71% | -4.41*** | 13.20*** | -.20 | 62.95% | 10.31% | 15.64*** | 223.61*** | .70 |
| Deep learning | ADAM | 72.22% | 7.23% | | | -.08 | -5.97* | -.00 | | | 4.75*** | 16.26*** | .21 | | | 23.28*** | 471.33*** | 1.05 |
| | SGDM | 69.37% | 7.81% | | | -9.00*** | 72.93*** | -.40 | | | 4.23*** | 11.72*** | -.19 | | | 15.68*** | 224.62*** | .70 |
| | RMSPROP | 70.48% | 7.19% | | | -5.80*** | 27.13*** | -.26 | | | -.85 | -5.26* | -.04 | | | 18.93*** | 322.05*** | .85 |

Note. *D*: Cohen's *D* for an effect size from *t*-tests. For *P*-values, *** $p < .001$. For *2logBF* values, *: $2\log BF > 2$; ***: $2\log BF$

≥ 10 .

Figure Captions

Figure 1. The structure of neural network for deep learning

Figure 2. Hypothetical SEM model. Error terms were excluded from the diagram.

Figure 3. Actual SEM used for the performance comparisons model only with significant paths ($p < .05$). Error terms were excluded from the diagram.

Figure 4. ROC curves of the four compared methods.

Figure 5. A big picture view of an intervention study, from early stages of design to scaled implementation









