Learning to Represent Healthcare Providers Knowledge of Neonatal Emergency Care

Findings From A Smartphone-Based Learning Intervention Targeting Clinicians From LMICs

Timothy Tuti†
Learning and New Technologies,
Oxford University/
KEMRI-Wellcome Trust
timothy.tuti@kellogg.ox.ac.uk

Chris Paton
Center for Global Health & Tropical Medicine,
Oxford University
chris.paton@ndm.ox.ac.uk

Niall Winters,
Learning and New Technologies,
Department of Education,
Oxford University
niall.winters@education.ox.ac.uk

ABSTRACT

Modelling healthcare providers' knowledge while they are gaining new concepts is an important step towards supporting self-regulated personalised learning at scale. This is especially important if we are to address health workforce skills development and enhance the subsequent quality of care patients receive in the Global South, where a huge skills gap exists. Rich data about healthcare providers' learning can be captured by their responses to close-ended problems within conjunctive solution space -such as clinical training scenarios for emergency care delivery- on smartphone-based learning interventions which are being proposed as a solution for reducing the healthcare skills gap in this context. Together with sequential data detailing a learner's progress while they are solving a learning task, this provides useful insights into their learning behaviour. Predicting learning or forgetting curves from representations of healthcare providers knowledge is a difficult task, but recent promising machine learning advances have produced techniques capable of learning knowledge representations and overcoming this challenge. In this study, we train a Long Short-Term Memory neural network for predicting learners' future performance and forgetting curves by feeding it sequence embeddings of learning task attempts from healthcare providers from Global South. From this training, the model captures nuanced representations of a healthcare provider's clinical knowledge and their patterns of learning behaviours, predicting their future performance with high accuracy. More significantly, by differentiating reduced performance based on spaced learning, the model can help provide timely warning that helps support healthcare providers to reinforce their self-regulated learning while providing a basis for personalised instructional support to aid improved clinical outcomes from their professional practices.

CCS CONCEPTS

• Learning latent representations • Knowledge representation and reasoning • E-learning • Social and professional topics

KEYWORDS

Global Health, Clinical training, Smartphones, Neonatal care, Emergency care, Deep Knowledge Tracing, Forgetting Curves

1. Background

Low- and Middle- Income Countries (LMICs) like those in the Global South produce more than 20% of the global disease burden but only has 3% of the global health workforce [1-3]. This severe trained workforce shortage, coupled with skill imbalance, maldistribution, and lack of training opportunities are the key contributors to almost half of avoidable deaths globally, especially when it comes to critical care provided to children under the age of five [3-5]. Sub-Saharan Africa (SSA) has the highest overall risk of death within the first 24 hours of life, reporting 38% of global neonatal deaths [4, 5]. Costs of face-to-face refresher training in the SSA region remain prohibitively high and significantly constrained by the socio-economic, political and institutional landscape [6-8]. This presents a significant challenge of how workers can be (re)training and upskilled economically in these contexts where the need is most urgent and the impact most felt.

Smartphone-based digital learning solutions have shown potential to address this challenge in these contexts, given their increasing uptake rate and pattern of usage [9, 10]. This provides a platform for interventions that are scalable and easily accessible in these regions. Additionally, they provide an avenue for the introduction of adaptive instructional support, which has been shown to significantly outperform teacher-led large-group instruction, non-adaptive computer-based instruction, and paper-based instruction, in enhancing learning [11]. However, these studies tend to be typically in high-resource settings and outside clinical care [11]. Such learning adaptations are common in the Intelligent Tutoring Systems (ITSs) literature, where learner interactions with the digital learning platform tend to be tracked as a sequence of student-driven steps [12]. When a student attempts a learning task (step), the ITS records whether the learner was successful, and whether any system-initiated assistance was provided, and may provide instructional support by restructuring the learning content path, feedback provided, or content presentation, based on the learner's evaluated proficiency [13]. The learning tasks represent unique Knowledge Components (KC) which tend to be some generalisation of learning "…concepts, principles, facts, or skills, and cognitive science terms like schema, production rule, misconception…" [14].

Emergency neonatal care training courses for SSA contexts are typically scenario-based where the components being taught emphasise the steps in critical care needed for early recognition and treatment of new-born babies who need immediate care and hospitalization. Consequently, a specific order of steps of a clinical algorithm is followed, where the probability of new-born surviving is dependent on delivering the key steps in the correct sequence, with each step being timed. We have implemented such a scenario-based training intervention on a smartphone.

1.2 The Intervention

Life-Saving Instruction for Emergencies (LIFE) [15] is a gamified platform for use with low-cost smartphones to provide training in the care of very sick new-born babies, particularly in low-resource settings, with the aim of expanding it to include other clinical care scenarios. It is based on face-to-face scenario-based teaching where the components being assessed emphasise the tenets of neonatal critical care with early recognition of new-borns who need immediate care being key. This is achieved by using game-like training techniques to reinforce the key steps that need to be performed, an approach commonly referred to as serious gaming [16, 17]. Consequently, it follows a specific ordering of clinical care-giving algorithms with each learning task being timed. The learner starts a scenario which provides some background information to the learning task, and, on each learning task, must provide input either through multiple choice questions, selection of items necessary for the learning task, or performing on-screen interactive tasks (e.g. navigating to equipment, switching on machines, etc.) (Figure 1).

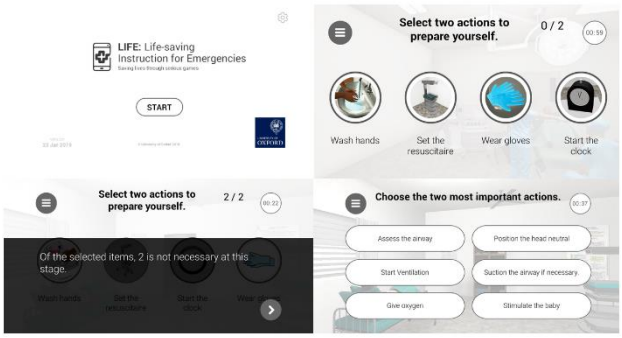

Figure 1: Sample screenshots from LIFE application

For each incorrect attempt by the learner, feedback is provided with the option of more information and the learner must successfully respond to the question before being allowed by the smartphone application to proceed. The end of the scenario is signalled by a crying baby indicating that the baby is now breathing, with a breakdown of the score provided. The scenario model that is used is one that replicates Emergency Triage, Assessment and Treatment plus admission care (ETAT+), a face-to-face training approach training that has already been validated [18, 19]. ETAT+ content has already been used to train over 5,000 healthcare workers and 2,000 medical students across Eastern and Southern Africa, and now East Asia [18, 19]. LIFE is meant to be accessible at scale by healthcare providers and able to function off-line on low-end smartphone devices and provide self-regulated training opportunities akin to continuous professional development at low marginal cost.

1.3 Prior Work On Knowledge Tracing

Given that we are trying to capture aspects of a healthcare provider's knowledge mastery over a sequence of conjunctive ordered tasks, the aim is to map and predict the learner input sequence $(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_T})$ to an output sequence $(\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_T})$. That is, when a student attempts a learning task (step), a record is kept of whether the attempt was successful $(\boldsymbol{y})$ linked to the learning task $(\boldsymbol{x})$. Generally, the knowledge tracing task can be formalized as follows: given a learner's historical interactions X= $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_t}$ up to time t on a Knowledge Component (KC), it predicts some aspects of their next interaction $\boldsymbol{x_{t+1}}$ [20]. Previously, Bayesian Knowledge Tracing (BKT) models have been used to model the knowledge states of KCs using Hidden Markov Models (HMM) [21]. The hidden states in the HMMs represent the student's knowledge state which indicates whether they have mastered the KCs. However, some of the BKT modelling assumptions are impractical: BKT assumes that forgetting does not occur; the KCs are treated as being mutually independent; its typical implementation does not allow for learners to have different learning rates; it assumes that all students have the same probability of knowing a particular skill at their first opportunity [22]; and it suffers from the problem of multiple global maxima when trying to estimate model parameters [23].

To address some of the shortcomings of BKTs, Learning Factors Analysis (LFA) and their different modalities such as Performance Factors Analysis (PFA) [24] have been proposed. They model student knowledge states using logistic regression models in order to deal with the multiple KCs issue while incorporating student ability into the model. They exploit the number of successes or failures of a learner's attempt at a KC to predict whether the learner has acquired understanding about the KC. Although they can handle a learning task that is associated with multiple KCs, they cannot deal with the inherent dependency among KCs [25].

Recently, Recurrent Neural Network (RNN) models have been applied in an approach called Deep Knowledge Tracing (DKT) [26]. These RNNs, have been found to robustly predict student performance on the next learning task, given prior performance [27]. We hypothesis that even in a "vanilla" form, they are a suitable neural network architecture for our analysis, given that they perform well on sequence modelling tasks in other domains [27]. To map the input $(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_T})$ to an output sequence $(\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_T})$, the input vector undergoes a series of transformations via a hidden layer, which captures useful latent information in the form of a sequence of hidden states $(\boldsymbol{h_1}, \boldsymbol{h_2}, \dots, \boldsymbol{h_T})$. More concretely, at time-step $\boldsymbol{t}$, the hidden state $\boldsymbol{h_T}$ is the encoding of the past information. Mathematically, this is represented as:

$$h_t = \tanh(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

$$y_t = \sigma(W_{hy}h_t + b_y) \quad (2)$$

where both the hyperbolic tangent tanh(...) and the sigmoid function $\sigma(...)$ are applied in an element-wise manner. The model is parameterised by a weight matrix W and a bias vector b with appropriate dimensions. To control what information should be stored and used for predicting some aspects of the next interaction $\boldsymbol{x_{t+1}}$, a popular variant of RNN known as Long Short-Term Memory (LSTM) RNN architecture is commonly adopted [28]. LSTM achieves this by incorporating three gates to mimic human memory: forget gate $\boldsymbol{f_t}$, input gate $\boldsymbol{i_t}$, and output gate $\boldsymbol{o_t}$ which control a memory state $\boldsymbol{c_t}$. Mathematically, these are calculated based on the current input of $\boldsymbol{x_t}$ and the previous hidden state $\boldsymbol{h_{t-1}}$ as:

$$f_t = \sigma(W_f[x_t, h_{t-1}] + b_f) \quad (3)$$

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \quad (4)$$

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \quad (5)$$

Where [...] denotates concatenation. $\boldsymbol{f_t}$ decides what information to forget from the previous memory cell state $\boldsymbol{c_{t-1}}$, while $\boldsymbol{i_t}$ decides what new information $\hat{c}_t$ is added to the recent cell state $\boldsymbol{c_t}$. Therefore, $\boldsymbol{c_t}$ depends on the previous cell state after forgetting with new information added from $\boldsymbol{i_t}$. Eventually, the output gate $\boldsymbol{o_t}$ determines what information should be extracted from $\boldsymbol{c_t}$ to form the hidden state $\boldsymbol{h_t}$. These can be expressed mathematically as follows:

$$\hat{c}_t = tanh(W_c[x_t, h_{t-1}] + b_c) \quad (6)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \hat{c}_t \quad (7)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (8)$$

Where $\otimes$ denotes elementwise multiplication. This allows LSTMs to store information that occurred in the distant past, making it more robust in its capability to trace knowledge than a vanilla RNN. The unfolded RNN that we used for LIFE to represent a high-level interpretation of a DKT architecture is illustrated in Figure 2. In the DKT model used in this study, an additional embedding layer was used (Figure 2). This allows for the LSTM-based DTK to capture more information about a learner's trajectory and capture the temporal relationships within the sequence effectively, as demonstrated elsewhere [29].

Arguably, LSTMs mimicking of memory better supports knowledge tracing by accounting for learning 'history', based on the recency and outcome of completed learning tasks. This can be done to ensure that the conjunctive nature of LIFE content (steps to resuscitate a neonate in distress) are factored into the prediction of healthcare workers learning trajectories. Evidence of the successful implementation of student-modelling approaches on digital platforms in clinical training in order to facilitate adaptive learning and improve learning outcomes is scarce[30, 31], and virtually non-existent for emergency care training in low-income settings [32, 33].

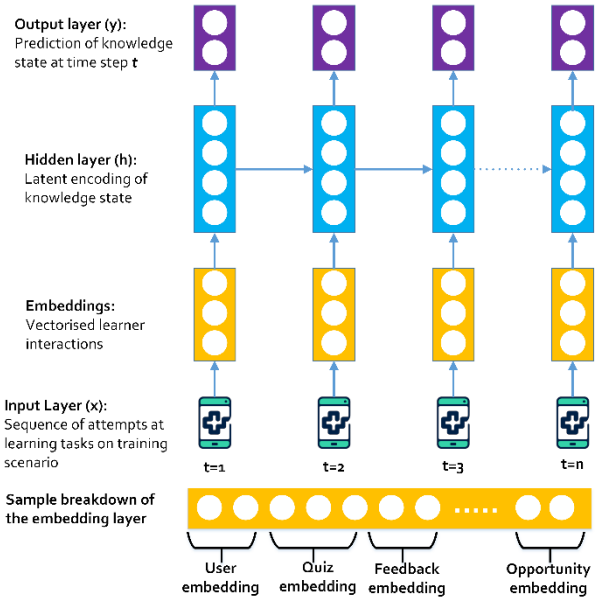

Figure 2: Illustration of LSTM model used

In high-income settings, despite the important role of smartphones in facilitating personalised learning, there is still a lack of research investigating mobile-based ITSs [34]. Thus, we do not have much evidence through empirical examples supporting how learning analytics such as DKT can be implemented in LMICs context to support self-regulated learning at scale for healthcare providers on low-end smartphone devices. Additionally, we do not have evidence of if and how the healthcare providers' length of learning trajectories used in knowledge tracing affects the ability to predict some aspects of their next interaction for this conjunctive solution space of clinical training. Finally, we are yet to find an empirical basis that explores prediction performance while explicitly addressing skill decay over time. The emphasis on skill decay over time rests in the risk implications for patient outcomes in this context. This study represents efforts to start addressing the aforementioned research gaps.

The aim of this study was to determine: (a) given a healthcare provider's sequence thus far in the emergency care tasks they have prescribed, whether we can accurately predict the healthcare provider will successfully complete the next caregiving task, (b) how this prediction is affected by the healthcare provider's spacing of their own learning, and (c) how this prediction varies by learner's performance. This research's intended practical implication for the future is to advance evidence on how learning analytics can integrate a wholistic approach by undertaking research that aims to understand and optimise the learning process on platforms that increase both scale and access of the learning

interventions for *all* students across *all* contexts, especially those from under-represented contexts in learning analytics research e.g. Sub-Saharan Africa [35].

2 Methods

2.1 Study Design, Setting And Participants

This study was a retrospective observational study [36] of healthcare providers from both public and private hospitals in Low- and Middle-Income Countries (LMICs), in clinical cadres such as nurses, clinical officers and medical doctors, with experience levels varying from students to consultants. Participants were enrolled into the study through a combination of snowball and convenience sampling strategy. Recruitment occurred through use of peer referrals among clinicians, publicised through (a) private professional social media and social network accounts, (b) regional clinical meetings, (c) clinical conferences, (d) medical training institutions, (e) local hospitals and (f) international clinical professional forums focused on health systems in LMICs. In total, 697 participants were recruited. The eligibility criteria for inclusion were that the participants had to be healthcare providers from LMICs either in training for, and/or actively providing bedside clinical care.

2.2 Study Variables, And Data Management

The learning scenario used in this study provides simulation training on the contextualised management of newborn resuscitation through a series of sixteen learning interactions that elicit responses from learners in the form of multiple-choice answers or performing interactive tasks. At the end of a successful completion of simulation tasks, the platform provides performance score feedback based on the outcome of the learner's first attempt at each learning interaction. Data collection was through the Android-based LIFE smartphone application, which would securely transmit a copy of anonymised student-step data to a Google Firebase distributed database. For the purposes of the proposed analysis, the outcome of interest was specified as getting the next try correct given previous attempts at the learning scenario. The variables of interest were time spent on learning task, number of previous tries (i.e. opportunities) per learning task, and whether feedback had been provided for each unique try per learning task. Demographic data from study participants was optional and collected using an 'opt-in' mechanism since ethical review process required an extra informed consent process for any other type of data that was not student-step data. It included years of experience, clinical cadre (i.e. role) and the age-bracket.

2.3 Model Training

A student's trajectory consists of *k* learning task submissions, which represent how far they got in providing the emergency care needed (i.e. the number of attempts they made at learning tasks) within the LIFE game. Each task represents a question that they had to respond to. Additionally, time taken to complete the task, level of feedback provided on incorrect try, cumulative count of opportunities at the specific learning task, and time since last attempt, were captured with each submission. These data points were converted into embeddings (meaningful vector representations of feature combinations from an individual learner's submissions) for use within the LSTM model, similar to the one described in [37] and illustrated in Figure 2. Embeddings consisting of user identifier, quiz component identifier, time to complete task (recoded as an incremental counter for every 3 seconds passed), level of feedback provided, hours passed after last attempt and cumulative opportunity at the current task were created and concatenated for each submitted attempt at learning task. This model utilised previous temporal information; e.g. at timestep *t*, it utilised all embeddings from timestep in the last *t-1* timesteps.

We used a two-layer deep LSTM. To make the prediction at the end of the sequence i.e. the last timestep, we pass the hidden state at the last timestep through a fully connected layer and then a sigmoid layer. The output y from the prediction layer – which is a sigmoid layer - is an estimated probability distribution over two binary classes, indicating whether the healthcare provider would successfully solve the current task *t* given their past performance. For objectives B and C, predictions from the LSTM model output were evaluated after being grouped by the learning spacing categories and learner performance respectively.

From the learning data, 20% was held-out as test dataset. Of the remaining 80%, in each of the 1000 forward and backward passes over the training dataset (i.e. epochs), the model would hold-out 50% of randomly selected samples from the training dataset as a validation dataset. The use of multiple epochs aids in finding optimal model parameters which minimise the training and validation losses. We used Adam [38] as the model's stochastic gradient optimizer together with a Tanh activation function [39] for the deep layers of the model, with Adam's learning rate set to 0.0005. To minimise overfitting, we used early stopping for the number of epochs, and set at 50% the fraction of the hidden layer units to drop in the model during the transformation of both the hidden layer and the recurrent state [40, 41]. In addition, L2 weight regularization was employed to smooth oscillations over training loss, in order to avoid overfitting and reduce generalisation error [42], with the weight decay rate set to 0.001.

2.4 Model Evaluation

We employed different evaluation metrics including the Accuracy metric for calculating how often predictions matches labels in the validation dataset as an optimising strategy. The models' performance was evaluated on the held-out test dataset (20% of the samples generated from sliding window approach which is explained in the next section), with area under a curve (AUCROC), F1 score, and accuracy being used as evaluation metrics. The model set to evaluate learner prediction performance based on temporal sequence of learners' input responses in LIFE. In efforts to train a model that uses the same dense function applied at every time (in our case positional) step, a time-distributed layer was included, which allowed to make a prediction at every timestep ***t*** based on the hidden state at ***t***. The

performance would be evaluated using the already described metrics at every temporal slice provided in the data samples.

2.5 Statistical Methods, Missing Data, And Sensitivity Analyses

Data manipulation and statistical analyses were performed using *Python* 3.7.4 and *Tensorflow* 1.13.1 [43, 44]. In order to expand DKT for understanding learners as they produce valuable responses over time within learning scenarios, using variables of interest, we created meaningful embeddings of the data generated from their interaction with the learning tasks. This follows common practice of using LSTMs to create embeddings of leaner responses [37, 45].

Sliding window approach was used to determine the samples to use for modelling. In this approach, samples with varying history lengths, where history length was in the set [2+1, 5+1, 8+1, 11+1, 14+1], were generated with each sample constrained to belong to a single learner. The +1, indicates the step whose outcome was being predicted. This is illustrated in Figure 3 where the index notation in the bottom left of the boxes illustrate sample counter and the number of circles illustrating single ***x*** input at timestep ***t***.

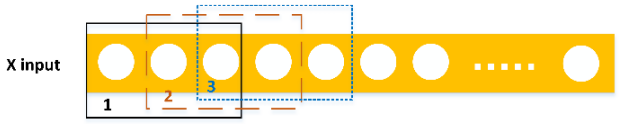

Figure 3: Sliding window sample generation based on varying timesteps. In this illustration, timestep length (i.e. history) is three i.e. using two most recent previous steps to predict the current step which is the third step.

Varying the lengths of the samples was to evaluate whether model performance in predicting ***y*** at timestep **t** is affected by how far back the history goes. Hence, the sensitivity of the predictions was evaluated against the length of the history of the performance used to predict next step. As part of post-hoc analysis, we used a popular method for exploring high-dimensional data known as t-SNE, to explore how well the model embeddings had been captured [46]. The t-SNE parameters used were: learning rate=10, iterations = 5000, metric = cosine, and perplexity =100. The use of these metrics is explained in detail elsewhere [46, 47]. In the modelling of the samples generated from learning data using the sliding window approach, learners without enough data points were omitted from analysis. Table 1 describes the numbers omitted for each length of history used in LSTM models.

Table 1: **Learners excluded due to insufficient data based on timesteps used in LSTM model**

History* (Step being predicted)	Learners excluded (N)	% of learners excluded
2 (3)	54	7.78
5 (6)	117	16.86
8 (9)	140	20.17
11 (12)	165	23.78
14 (15)	202	29.11

Note: **Length of previous learning tasks submission used in predicting outcome in the current learning step*

No evaluation was conducted of whether excluding observations with insufficient data would bias the results, and the limitations of this approach will be addressed in a later discussion.

3. Results

Results from the running of the multiple LSTM models demonstrated that using a longer history had higher accuracy, and more predictive and better discriminatory power that using shorter chains (Table 2). Although in general, the models showed considerably better prediction performance compared to alternative modelling approaches [48] where history length was ≥ 5 learning task submissions, which were not limited to a single learning session or single KC.

Table 2: **Model performance results from using outcome from n-1 steps to predict step outcome on step n**

History length* (No. analysis samples**)	AUC Score	Accuracy	F1 Score
2 (20,419)	0.8116	0.7480	0.7450
5 (18,578)	0.8921	0.8172	0.8163
8 (16,872)	0.9300	0.8613	0.8607
11 (15,228)	0.9415	0.8775	0.8771
14 (13,679)	0.9474	0.8832	0.8834

Note: **Length of previous learning tasks submission used in predicting outcome in the current learning step. **Number of samples generated from the 'sliding window' process.*

When considering how spacing of learning sessions would affect model performance given the varying length of previous timesteps informing current step's outcome prediction, the models using six timesteps and above showed consistently decent predictive performance with an AUC above 0.85 except for predicting spacing longer than one month, while models with nine timesteps and above were better (Figure 4).

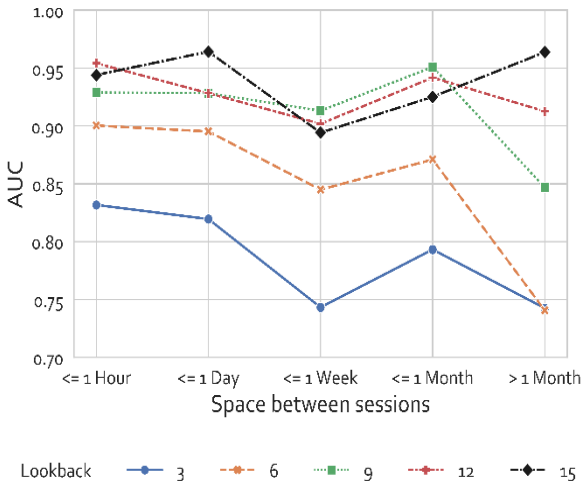

Figure 4: Prediction accuracy of the different models by time between learning sessions

Intuitively, this can be attributable to longer chains representing providing more latent information for knowledge tracing of healthcare providers despite their varying learning spacing behaviour. It is noteworthy that the learning spacing was not balanced across the categories: from 1712 learning sessions among the 697 learners, 694 (40.54%) were repeated immediately after the learning session ended, 577 (33.70%) were repeated within the hour, 204 (11.92%) were learning sessions repeated within the day but not within the hour, 126 (7.36%) were learning sessions repeated within the week but not in the same day, 80 (4.67%) were learning sessions repeated within the month but not within the same week, and 31 (1.81%) were learning sessions that were repeated after a month.

This finding from LIFE's learning data showed tightly spaced learning sessions when healthcare providers were given autonomy to decide when to (re)use the learning intervention. While typical studies in spaced learning focus on weeks or months as the temporal unit for analysis, LMICs healthcare providers' learning behaviours using LIFE is indicative that when they are allowed to self-regulate on digital learning interventions, they prefer to reinforce their own learning on more frequent, tightly spaced learning cycles. In general, the wider the learning spacing, the less accurate the prediction of learning performance on next learning step, but this is mitigated by using longer learning sequences for predictions (Figure. 4). For emergency care scenarios in typical LMICs hospital settings from which LIFE learning content mimics, because of prevalence in comorbidities or cyclic nature of clinical care, they tend to have longer sequences and would thus produce longer learning scenarios. We highlight that the lack of DKT model performance degeneration in prediction with longer histories can be leveraged for such scenarios.

Figure 5 illustrates results from exploring whether the best performing model's spacing embedding layer was sensitive to the healthcare providers' pace of learning. The plot dimensions of the Figure 5 are un-interpretable because t-SNE does not preserve distances, rather visualises clustering of weights from the DKT model.

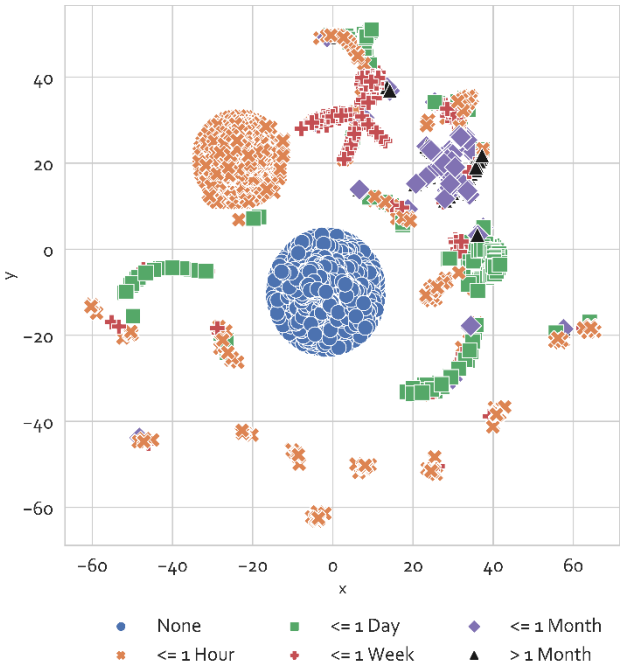

Figure 5: Example t-SNE results from the learning spacing embedding layer.

From the DKT learning spacing embedding layer visualisations illustrated in Figure 5, while both tail end of spacing are easily distinguishable by the distinctive clustering, where the healthcare providers chose to go through LIFE again within a span of one hour, their learning behaviour produces relatively more outliers that are difficult to distinguish from the alternative spacing options.

Several competing plausible explanations for this include (1) that the learning data was strongly skewed towards those who had shorter spaced learning intervals, thereby making it challenging for the model(s) to learn distinguishing behaviours from longer spaced learning intervals; or (2) Given that these data is only for 697 healthcare providers, more data would be needed from more healthcare providers to be able to generate better embeddings of their spaced learning behaviours.

While the learning spacing was skewed towards less than 24 hours, it is encouraging to see that in general, the best model was typically not associated with substantive degradation of its predictive performance when prediction accuracy was grouped by learner's performance, even as the space between learning sessions growing wider (Figure. 6).

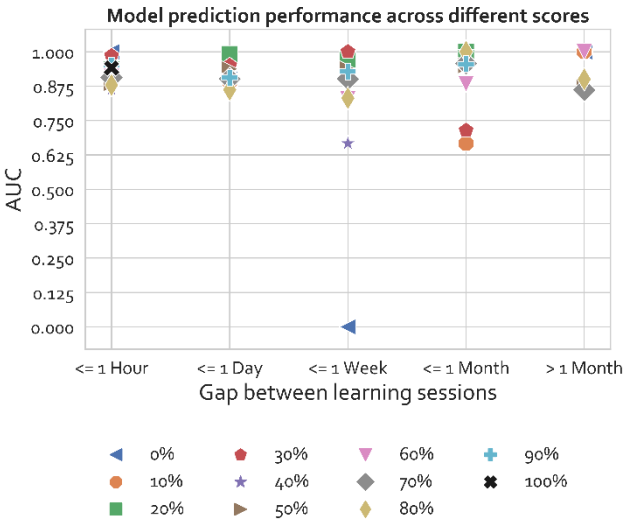

Figure 6: LSTM predictive performance by learner performance grouped by learning spacing

Learning sessions with less than 50% knowledge retention were in the minority and tended to be more spread across the learning spacing spectrum (Table 3). The low numbers in the wider spacing options might arguably have negatively impacted knowledge tracing for healthcare providers in those performance groups by reducing ability predictions precision (Figure 6).

Table 3: Spaced Learning by performance

Learning Space	Score ≥ 50 % (N= 876)	Score < 50 % (N = 142)
<= 1 Hour	506 (57.76%)	71 (50.0%)
<= 1 Day	177 (20.21%)	27 (19.01%)
<= 1 Week	110 (12.56%)	16 (11.27%)
<= 1 Month	63 (7.19%)	17 (11.97%)
> 1 Month	20 (2.28%)	11 (7.75%)

Exploration of the impact of the varying sequence lengths of LSTM time slices indicated that longer sequence lengths had better model prediction performance at the later timepoints than the earlier ones (Figure 7), in addition to the evidence that models with longer sequence lengths being better in prediction performance across different spacing of learning sessions (Figure 4).

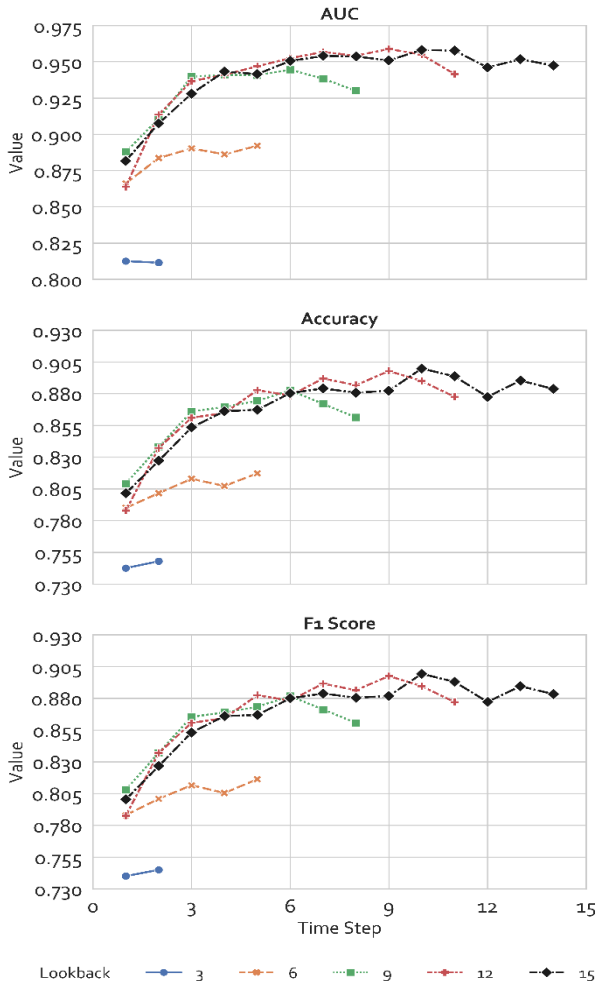

Figure 7: LSTM model sensitivity to length of history

From the healthcare providers learning behaviour, forgetting curves (Figure 8) would indicate that for LIFE, spacing learning to weekly basis had a reasonable effect on knowledge retention of about ≥ 70% in terms of learner performance. The learning spacing embedding layer of the LSTM model had demonstrated it could capture this spacing reasonably well (Figure 5).

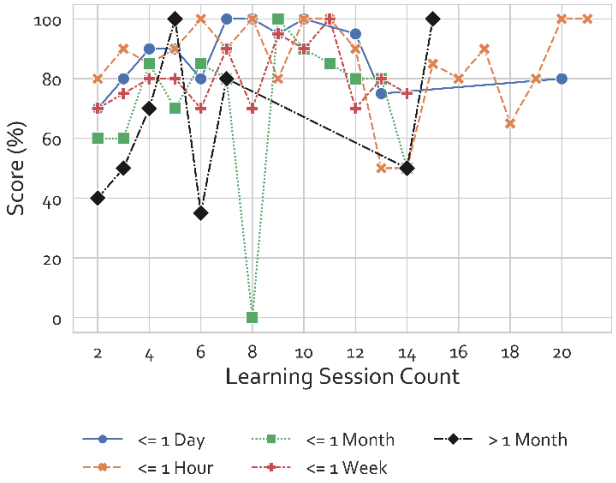

Figure 8: Healthcare providers' forgetting curves by spaced learning categories. Performance summarised using median statistic

However, with this spaced learning option, the prediction accuracy across different scores was moderate to poor (Figure 6), which might partly be explained by the very few numbers in this spacing spectrum (Table 3). One way to overcome this challenge might be using a DKT model with shorter lengths of the previous learning steps in predicting next step like length 9 (Figure 4), which had a reasonably good AUC, accuracy and F1 score (Figure 7, Table 2).

Only 164 (23.53%) of the study participants consented to their demographic data being collected for analysis. Admittedly, with such huge missingness, including these variables in the DKT model would have likely led to misleading findings. Instead, we provide exploratory analysis from model predictions linked to demographic data where it was available. However, this ought to be interpreted cautiously. Learner performance prediction was more varied in healthcare providers for healthcare providers with 9-12 years of experience. In general, there was no easily discernible pattern of model prediction performance across the clinical cadres with increase in years of experience. This is illustrated in Figure 9.

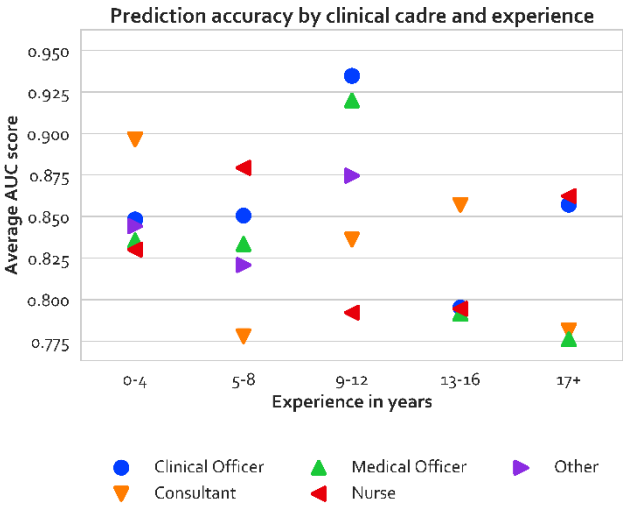

Figure 9: DKT prediction performance by clinical role and level of experience

4. Discussion

4.1 Summary Of Findings

From healthcare providers' learning data captured through the LIFE intervention, Deep Knowledge Tracing Models can predict learner's performance with an accuracy of 0.748 – 0.883. With healthcare providers allowed to self-regulate their learning behaviour, a weekly spacing of learning sessions demonstrated reasonable impact on knowledge retention of about ≥ 70% in terms of learner performance. From the learning spacing embedding layer of the LSTM model, it demonstrated it could capture this spacing reasonably well with few outliers observed for learning spaced hourly or daily. However, with the weekly spaced learning option, the prediction accuracy across different scores was moderate to poor, which might partly be explained by the very few numbers of learning sessions in this spacing spectrum. Using DKT model with shorter trajectory histories in predicting next steps, like length 9, helped overcome this challenge and produced reasonably good AUC, accuracy and F1 score.

4.2 Comparison To Other Studies

There are very limited cases of use of DKTs together with forgetting curves (i.e. learning decay) in healthcare provider training. An example of learning decay in a clinical care setting has been in radiography skills where it was mitigated against by exposure to multiple scenarios (irrespective of feedback) every two months [49]. This differs from our approach where only a single care scenario was used for this pilot work, but where we found that a shorter spacing of a week might be better in reducing learning decay. In the health domain, LSTMs have been used in clinical settings to predict the occurrence of clinical events and clinical diagnoses [50, 51], but in our use-case, we extend this to healthcare providers application of knowledge to simulated clinical scenarios. This study's emphasis on neonatal emergency care in LMICs is due to both the need for scalable access to training for emergency care training and the subsequent impact it promises on better global patient outcomes by reducing avoidable deaths [6, 7]. How different clinical training interventions use metacognitive scaffolds to improve healthcare providers' knowledge gain is not new [52-54], but this evidence has been fairly opaque as to how knowledge tracing was achieved, and hardly presents research evidence from low resource settings [32] making its impact limited to the Global North. Such evidence does not improve understanding of situated learning from diverse contexts such as LMICs [35].

Moreover, while adaptive learning produces significantly higher knowledge gains than alternatives [52-54], the current training models used in LMICs -which are not adaptive to individual learner needs- are typically face-to-face, and offered at a very high cost [8, 55, 56]. The findings from this study help to start exploring how the use of smartphone devices to deliver personalised clinical training for short simulation-based learning activities can begin to accommodate self-regulated learning over time, which arguably optimises learning outcomes in the health domain [57].

4.3 Implications Of Findings

For LMICs-based neonatal emergency care training delivered through smartphone-based learning interventions, DKT models' representation of spaced learning -which is useful for producing forgetting curves- might be at best, moderately accurate. The associated practical implication of this is that, where there is a risk associated with applying errant knowledge which exacerbates negative outcomes (e.g. for patient outcomes), encouraging up-skilling/retraining from inferred knowledge gain decay based on personalised

learning trajectories is not as precise as we would like it to be. In a context where the healthcare providers tend to be significantly overworked, underpaid, and under-resourced, interventions such as LIFE ought to optimise when and the way they encourage them to carve out time to refresh their clinical training knowledge. Additionally, as a knowledge tracing intervention that is linked to skill performance mastery, while this approach provides a necessary starting point for bridging knowledge gaps in healthcare providers, adding a layer of multi-modal learning using *in situ* high fidelity simulation training provided by platforms such as Virtual Reality (VR), or Mixed Reality (MR) might arguably enhance hands-on experiential learning by transforming the learning experience into a more meaningful one aimed at building performance skills, confidence and self-efficacy of the healthcare provider in providing necessary life-saving care [58, 59]. The study reported here sets the stage for progression to such type of healthcare training experiences research in LMICs. Future work into this type of multimodal learning needs to cater for scaling up learning interventions geographically (globally), while including design thinking for LMICs context, and situated learner behaviours in these contexts.

There is a potential risk of errant knowledge creation within a completely self-regulated, gamified platform such as LIFE. This undesirable learning effect is further compounded by 'dark play' where healthcare providers may purposefully make wrong choices to obtain feedback that elucidates on the consequences of their choices as a way of learning [60]. Where the status-quo in LMIC is delivery of care by a healthcare provider who typically has challenges accessing training, the errant knowledge risk becomes as equally detrimental to patient outcomes as having no knowledge. To minimise this risk in such gamified platforms, using formative evaluation linked to immediate feedback to constrain learner progression until they achieve a certain knowledge threshold can be useful. Learning paths can be constrained to follow the clinical algorithms, minimising the potential risk of errant knowledge creation from a completely self-regulated, "gamified" system. For the type of dark play being described here, which is a learning disposition, it can be harnessed through accommodating various learning modes within the design of gamified platforms that account for such anomalous learning behaviour. Such gamification learning modes can be linked to a human instructor, professional colleagues, or even chatbots to help guide this elaborative way of learning. Future research should consider exploring the motivations of LMICs healthcare providers and how their personalised goals informs their learning strategy. This would also make it possibility to explore how to integrate such learner behaviours into DKT modelling approach for use in adaptive learning systems.

Theoretically, in the same way there exist common architectures for deep learning models such as ResNets, LeNet, AlexNet etc. for computer vision application areas [61], for future growth in the area of Deep Learning application in education, there ought to be more concerted work on DKT models that seeks to produce curated architectures that maximise on educational concepts such as guessing, slipping, learning opportunities, spacing and/or knowledge components. This would greatly aid in external model validation, easier model transfer learning, and a better-shared understanding of how such DKT models can be more rapidly implemented and scaled up in other/newer digital learning interventions.

From this study findings, in smartphone-based clinical training, self-regulated learning will arguably tend to produce tighter spacing of learning sessions where risk from errant knowledge is costly. Subsequently, any measure of learning intervention effectiveness of the 'pre-post' form that does not account for how learners spaced their learning sessions might be, at best, slightly biased, and at worst, grossly misleading, when it comes to the interpretation of learning performance. Considering healthcare providers diverse and agentic spaced use of digital learning interventions like LIFE, it raises the question of whether decay rate of knowledge gain is arguably more informative about intervention effectiveness over '*immediate*' knowledge gain, especially where the bulk of repeated learning sessions are within the few hours that follow. The challenge here is in designing DKT models for (clinical) digital training interventions that not only concur with existing learning theories [62] but address the 'last mile' challenge: in the real world and for a multitude of interdisciplinary global learning challenges, if and how the modelling approach is connected to achieving the intended learning effectiveness of knowledge gain and for how long this effect is sustained. This would arguably improve healthcare providers conscientisation and agentic action in the adoption of digital interventions based on the intervention merits when it comes to the balance between initial knowledge gain and subsequent decay rate. One way forward on this might be including evidence-based metrics for decay rates based on learning spacing options as a means of advancing debate on how the effectiveness of such interventions should be reported.

4.4 Study Limitations

While DKT models have demonstrated reasonably moderate performance on LIFE data, the relatively average weighted average of F1 score -which consistently lagged behind AUC scores- is disconcerting. This might be due to the low numbers of observations analysed in the whole study in general, making it challenging to provide more accurate estimates. However, given that the data collected is from a pilot -arguably unique- study looking at the utility of digital learning metrics in knowledge tracing prediction for clinical training in low-income settings, it sheds light into a previously underexplored topic. This limitation can be revisited at a later stage as we continue to generate data to support the evidence base of these kinds of interventions for atypical application domains such as healthcare. While this study's sample is hardly generalisable, its inclusive constitution (from students to consultants, in all clinical cadres) makes it highly informative as a realistic data source on developing cognitive models for adaptive emergency care

training on smartphone platforms delivered to health workers in low-income settings. We are yet to find a comparable student-step data source (and studies) for this subject in this context.

5. Conclusions

Our work focuses on multi-step clinical scenario exercises with bounded solution spaces. Unlike open-ended exercises where flexible problem solving is encouraged, the conjunctive nature of the required solution and the inferred accumulating risk to patient care with each errant submission makes understanding the learner's progression essential, especially when considering their own spacing of self-regulated learning.

Given that digital learning platforms such as gamified smartphone-based learning applications more easily capture the temporal dimension of learner performance, we proposed an approach not commonly used in digital-based clinical training for learning representations of healthcare providers knowledge by using embeddings of learner submissions from LIFE over time from typical student-step features. We showed that based on the length of trajectories of these representations, they produce varied model performance for different student learning spacing behaviours after accounting for performance. We also showed that these representations can predict future healthcare provider's performance with high accuracy when we account for longer accounts i.e. (histories) of their previous student-step learning data.

We envisage this type of work being used in implementing automated hint systems for platforms such as LIFE, where deep knowledge tracing has the potential to identify learner weaknesses and provide personalised feedback – a use-case that we are working towards exploring [63]. By being able to anticipate particularity of learners struggles on a self-regulating digital learning platform such as LIFE, we hope to provide instructional support at scale to healthcare providers with varied knowledge levels in an unsupervised fashion in order to minimise learner dropout and encourage learner retraining. These applications could help improve and personalise the learning experience of healthcare providers in supporting access to knowledge that is crucial to saving lives at birth especially in settings such as LMICs.

ACKNOWLEDGEMENTS

Funds from Economic and Social Research Council (ESRC) awarded to TT through Oxford University, with additional funds from *Global Challenges Research Fund* (*GCRF*) Incubator Award Number: 161/105 grant awarded to NW, ME and CP supported aspects of this work. The authors would like to thank Mike English, Hilary Edgcombe, Jakob Rossner, Conrad Wanyama, and Naomi Muinga who have supported the implementation of LIFE project. The funders had no role in drafting or submitting this manuscript.

REFERENCES

1. Anyangwe, S. and C. Mtonga, *Inequities in the Global Health Workforce: The Greatest Impediment to Health in Sub-Saharan Africa.* International Journal of Environmental Research and Public Health, 2007. **4**(2): p. 93.
2. Sousa A. and Flores G., *Transforming and Scaling up Health Professional Education and Training*, in *Policy Brief on Financing Education of Health Professionals*. 2013, WHO: Geneva, Switzerland
3. M. Roser and H. Ritchie. *Burden of Disease*. 2019 [cited 2019 27th September]; Available from: https://ourworldindata.org/burden-of-disease.
4. UNICEF, *Levels & Trends in Child Mortality. Report 2018. New York, USA, 2018*. 2018, United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME).
5. WHO. *Children: reducing mortality*. 2019 [cited 2019 27th September]; Available from: https://www.who.int/news-room/fact-sheets/detail/children-reducing-mortality.
6. Couper, I., et al., *Curriculum and training needs of mid-level health workers in Africa: a situational review from Kenya, Nigeria, South Africa and Uganda.* BMC Health Services Research, 2018. **18**(1): p. 553.
7. Barteit, S., et al., *E-Learning for Medical Education in Sub-Saharan Africa and Low-Resource Settings.* Journal of medical Internet research, 2019. **21**(1).
8. Chaudhury, S., et al., *Cost analysis of large-scale implementation of the 'Helping Babies Breathe'newborn resuscitation-training program in Tanzania.* BMC health services research, 2016. **16**(1): p. 681.
9. Edgcombe, H., C. Paton, and M. English, *Enhancing emergency care in low-income countries using mobile technology-based training tools.* Arch Dis Child, 2016.
10. Silver, L. and C. Johnson. *Internet Connectivity Seen as Having Positive Impact on Life in Sub-Saharan Africa*. Pew Research Center - Global Attitudes and Trends 2018 [cited 2018 18th Dec]; Available from: http://www.pewglobal.org/2018/10/09/majorities-in-sub-saharan-africa-own-mobile-phones-but-smartphone-adoption-is-modest/.
11. Ma, W., *Intelligent Tutoring Systems and Learning Outcomes: Two Systematic Reviews*. 2017, Education: Faculty of Education.
12. Vanlehn, K., *The behavior of tutoring systems.* International journal of artificial intelligence in education, 2006. **16**(3): p. 227-265.
13. Chi, M., et al., *Instructional factors analysis: A cognitive model for multiple instructional interventions.* 2011.
14. VanLehn, K., P. Jordan, and D. Litman. *Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed*. in *Workshop on Speech and Language Technology in Education*. 2007.
15. University of Oxford. *Life-Saving Instructions for Emergency (LIFE)*. 2016 [cited 2018 18th Dec]; Available from: https://oxlifeproject.org/.
16. Bergeron, B., *Developing Serious Games. Game Development Series. Charles River Media.* Inc., Massachucetts, 2006.
17. Wang, R., et al., *A systematic review of serious games in training health care professionals.* Simulation in Healthcare, 2016. **11**(1): p. 41-51.
18. Ayieko, P., et al., *A Multifaceted Intervention to Implement Guidelines and Improve Admission Paediatric Care in Kenyan District Hospitals: A Cluster Randomised Trial.* PLOS Medicine, 2011. **8**(4): p. e1001018.
19. Irimu, G., et al., *Developing and Introducing Evidence Based Clinical Practice Guidelines for Serious Illness in Kenya.* Archives of disease in childhood, 2008. **93**(9): p. 799-804.
20. Corbett, A.T. and J.R. Anderson, *Knowledge tracing: Modeling the acquisition of procedural knowledge.* User modeling and user-adapted interaction, 1994. **4**(4): p. 253-278.
21. Yudelson, M.V., K.R. Koedinger, and G.J. Gordon. *Individualized bayesian knowledge tracing models.* in *International conference on artificial intelligence in education*. 2013. Springer.
22. Pardos, Z.A. and N.T. Heffernan. *Modeling individualization in a bayesian networks implementation of knowledge tracing*. in *International Conference on User Modeling, Adaptation, and Personalization*. 2010. Springer.
23. Gong, Y., J.E. Beck, and N.T. Heffernan, *How to construct more accurate student models: Comparing and optimizing knowledge tracing and performance factor analysis.* International Journal of Artificial Intelligence in Education, 2011. **21**(1-2): p. 27-46.
24. Cen, H., K. Koedinger, and B. Junker. *Learning factors analysis–a general method for cognitive model evaluation and improvement*. in *International Conference on Intelligent Tutoring Systems*. 2006. Springer.
25. Pavlik Jr, P.I., H. Cen, and K.R. Koedinger, *Performance Factors Analysis--A New Alternative to Knowledge Tracing.* Online Submission, 2009.
26. Piech, C., et al. *Deep knowledge tracing*. in *Advances in neural information processing systems*. 2015.
27. Khajah, M., R.V. Lindsey, and M.C. Mozer, *How deep is knowledge tracing?* arXiv preprint arXiv:1604.02416, 2016.
28. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.
29. Wang, L., et al. *Learning to Represent Student Knowledge on Programming Exercises Using Deep Learning*. in *EDM*. 2017.

30. Fontaine, G., et al., *Effectiveness of adaptive e-learning environments on knowledge, competence, and behavior in health professionals and students: protocol for a systematic review and meta-analysis.* JMIR research protocols, 2017. **6**(7).
31. Car, L.T., et al., *Health professions digital education on clinical practice guidelines: a systematic review by Digital Health Education collaboration.* BMC medicine, 2019. **17**(1): p. 139.
32. Opiyo, N. and M. English, *In-service training for health professionals to improve care of seriously ill newborns and children in low-income countries.* The Cochrane database of systematic reviews, 2015(5): p. 1.
33. Gentry, S.V., et al., *Serious Gaming and Gamification Education in Health Professions: Systematic Review.* Journal of medical Internet research, 2019. **21**(3): p. e12994-e12994.
34. Mousavinasab, E., et al., *Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods.* Interactive Learning Environments, 2018: p. 1-22.
35. Dawson, S., et al. *Increasing the Impact of Learning Analytics*. in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. 2019. ACM.
36. Mann, C., *Observational research methods. Research design II: cohort, cross sectional, and case-control studies.* Emergency medicine journal, 2003. **20**(1): p. 54-60.
37. Piech, C., et al., *Learning program embeddings to propagate feedback on student code.* arXiv preprint arXiv:1505.05969, 2015.
38. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.
39. Kalman, B.L. and S.C. Kwasny. *Why tanh: choosing a sigmoidal function*. in *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*. 1992. IEEE.
40. Gal, Y. and Z. Ghahramani. *A theoretically grounded application of dropout in recurrent neural networks*. in *Advances in neural information processing systems*. 2016.
41. Caruana, R., S. Lawrence, and C.L. Giles. *Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping*. in *Advances in neural information processing systems*. 2001.
42. Xu, Z., et al., *L 1/2 regularization.* Science China Information Sciences, 2010. **53**(6): p. 1159-1169.
43. Van Rossum, G. and F.L. Drake, *The python language reference manual*. 2011: Network Theory Ltd.
44. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
45. Socher, R., et al. *Recursive deep models for semantic compositionality over a sentiment treebank*. in *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013.
46. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.
47. Wattenberg, et al. *How to Use t-SNE Effectively*. 2016 [cited 2019 19th July]; Available from: http://distill.pub/2016/misread-tsne.
48. Yudelson, M. *Elo, I Love You Won't You Tell Me Your K*. in *European Conference on Technology Enhanced Learning*. 2019. Springer.
49. Boutis, K., et al., *The effect of testing and feedback on the forgetting curves for radiograph interpretation skills.* Med Teach, 2019. **41**(7): p. 756-764.
50. Lipton, Z.C., et al., *Learning to diagnose with LSTM recurrent neural networks.* arXiv preprint arXiv:1511.03677, 2015.
51. Kaji, D.A., et al., *An attention based deep learning model of clinical events in the intensive care unit.* PloS one, 2019. **14**(2): p. e0211057.
52. Feyzi-Behnagh, R., et al., *Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system.* Instructional science, 2014. **42**(2): p. 159-181.
53. Veredas, F.J., et al., *A web-based e-learning application for wound diagnosis and treatment.* Computer methods and programs in biomedicine, 2014. **116**(3): p. 236-248.
54. Wong, V., et al., *Adaptive tutorials versus web-based resources in radiology: A mixed methods comparison of efficacy and student engagement.* Academic radiology, 2015. **22**(10): p. 1299-1307.
55. Atukunda, I.T. and G.A. Conecker, *Effect of a low-dose, high-frequency training approach on stillbirths and early neonatal deaths: a before-and-after study in 12 districts of Uganda.* The Lancet Global Health, 2017. **5**: p. S12.
56. Willcox, M., et al., *Incremental cost and cost-effectiveness of low-dose, high-frequency training in basic emergency obstetric and newborn care as compared to status quo: part of a cluster-randomized training intervention evaluation in Ghana.* Globalization and health, 2017. **13**(1): p. 88.
57. van Houten-Schat, M.A., et al., *Self-regulated learning in the clinical context: a systematic review.* Medical education, 2018. **52**(10): p. 1008-1015.
58. Sheik-Ali, S., H. Edgcombe, and C. Paton, *Next-generation Virtual and Augmented Reality in Surgical Education: A Narrative Review.* Surgical technology international, 2019. **35**.
59. Winters, N., et al., *Using mobile technologies to support the training of community health workers in low-income and middle-income countries: mapping the evidence.* BMJ Global Health, 2019. **4**(4): p. e001421.
60. Andrade, F.R., R. Mizoguchi, and S. Isotani. *The bright and dark sides of gamification*. in *International conference on intelligent tutoring systems*. 2016. Springer.
61. Das Siddharth. *CNN Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more ....* 2017 [cited 2019 29th July]; Available from: https://medium.com/@sidereal/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5.
62. Smolen, P., Y. Zhang, and J.H. Byrne, *The right time to learn: mechanisms and optimization of spaced learning.* Nat Rev Neurosci, 2016. **17**(2): p. 77-88.
63. Tuti, T., et al., *Evaluation of Adaptive Feedback in a Smartphone-Based Serious Game on Health Care Providers' Knowledge Gain in Neonatal Emergency Care: Protocol for a Randomized Controlled Trial.* JMIR Res Protoc, 2019. **8**(7): p. e13034.