Combining Cognitive and Machine Learning Models to Mine CPR Training Histories for Personalized Predictions

Florian Sense
InfiniteTactics, LLC
Dayton, OH, USA
florian.sense@infinitetactics.com

Michael Krusmark
L3Harris Technologies
Melbourne, FL, USA
michael.krusmark.ctr@us.af.mil

Joshua Fiechter
Ball Aerospace
Dayton, OH, USA
jfiechte@ball.com

Michael G. Collins
ORISE at AFRL
Dayton, OH, USA
michael.collins.74.ctr@us.af.mil

Lauren Sanderson &
Joshua Onia
RQI Partners, LLC
Dallas, TX, USA
lauren.sanderson@rqipartners.com
josh.onia@rqipartners.com

Tiffany Jastrzembski
Air Force Research Laboratory
Dayton, OH, USA
tiffany.jastrzembski@us.af.mil

ABSTRACT

Cardiopulmonary resuscitation (CPR) is a foundational life-saving skill for which medical personnel are expected to be proficient. Frequent refresher training is needed to prevent the involved skills from decaying. Regular low-dose, high-frequency training for staff at fixed intervals has proven successful at maintaining CPR competence but does not take into account inherent performance variability across learners. Tailoring refreshers to an individual's past performance would minimize personnel being trained too (in)frequently and would ensure faster knowledge acquisition for new learners. To maximize the benefits of individualized schedules, learning needs gleaned from past training history need to be identified. A recent field study conducted among nursing students showed that a cognitive model-based approach was able to successfully trace the knowledge acquisition and decay of learners and prescriptively devise personalized training regimes that outperformed fixed schedules with regards to both training efficiency and learners' performance. Here, we report a post-hoc analysis of the collected data to investigate whether an alternative modeling approach, blending cognitive modeling and machine learning, could have resulted in even higher quality predictions. Our results reveal modest improvements in predictive accuracy for ensemble models, in which machine learning models predict the prediction errors (i.e., residuals) of the standalone cognitive model. These promising findings reveal strong applied utility for future use in domains where sustained proficiency is a requirement.

Keywords

Predictive modeling; Cognitive model; Machine learning; Cardiopulmonary resuscitation; Learning

1. INTRODUCTION

Cardiopulmonary resuscitation (CPR) is a basic life-saving skill but it has been shown that medical professionals are not able to perform it consistently [1]. To remedy this shortfall, several improvements to skill acquisition and maintenance programs have been proposed [6]. One dimension of the shift in educational focus [5] emphasizes increased re(training) efficiency by moving towards personalized, adaptive scheduling. The current work aims to facilitate this development.

Currently, as in many domains, refresher trainings at fixed intervals (i.e., regular and the same for everyone) are required to maintain CPR compliance. A recent effort [14, 18] has shown that a cognitive model representing regularities of memory can be leveraged to devise personalized training schedules that maintain proficiency at lower cost and risk. This effort will be referred to as *the CPR field study* throughout the current text, and the data collected during this experiment (see sections 2.1 and 2.2 for details) will form the bedrock on which the efforts presented here will build.

Specifically, we conducted a post-hoc simulation study of the CPR field study data to explore whether the cognitive model's predictions could be enhanced by combining it with machine learning (ML) models that can leverage additional information to improve predictive accuracy. The combination of the two modeling approaches is achieved by fitting the models sequentially, forming an ensemble model in which the cognitive model's residuals are learned by the ML models. We show that their combined predictions afford a modest improvement over the cognitive model by itself and are preferable to using the ML models by themselves.

Modern CPR training is an interesting educational data mining domain and modeling task because trainings are conducted on advanced manikins equipped with an array of sensors that quantify various aspects of a student's performance against objective performance guidelines [16]. Consequently, large amounts of high-resolution data are readily collected for a given event. The challenge is that events are usually spaced months apart, which provides a sparse sampling space for knowledge tracing. Consequently, it is difficult to make precise predictions. However, given quality CPR's central role in the "chain of survival" [6, 5], even small improvements in predictive accuracy can conceivably have large real-life impacts—especially if predictive models can identify those most in need of more frequent refresher trainings and help them to maintain compliance.

Generally speaking, the fields of cognitive science and machine learning have approached the computational modeling of a task like CPR skill acquisition and maintenance with different mindsets [24]. Specifically, cognitive models primarily focus on explaining the mechanisms that drive individual differences; machine learning models primarily focus on out-of-sample prediction. Aiming to combine the best of both approaches, we engineered a pipeline of predictive models: First, a cognitive model that was specifically developed as a prescriptive tool [13] is fit to the training data, which results in residuals that indicate which instances are fit poorly by the cognitive model. Next, a ML model is fit to those residuals, effectively learning to fine-tune the cognitive model's predictions. We show that such an ensemble approach can provide improvements in predictive accuracy without sacrificing interpretability, which is important to retain so that the model's personalized prescriptions are fully explainable. With an eye on advancing predictive tools in the domain of CPR training, our core motivation is to assess whether alternative predictive approaches could have yielded better result in the CPR field study, so that insights gained can be leveraged in future applied settings.

1.1 The current study

Here, we will use the exact version of the cognitive model that was used in the CPR field study as a yardstick to determine: **(I)** How well would alternative instantiations of the cognitive model have performed?, **(II)** Would a number of off-the-shelf ML approaches have yielded superior predictions than the cognitive model?, and **(III)** Could ML models be used to learn the prediction error of the cognitive model?

We believe the last question is the most pertinent. However, the combined approach should be compared to the approaches that only use either of the two modeling approaches in isolation to ascertain whether it has any benefit.

2. METHODS

This section will outline the data we used for our exploration, the input features that are available in the data, the predictive model we employed, and the setup of the simulation study we conducted to address our research questions. Figure 1 provides a schematic overview of the approach and its parts and connections are going to be explained in the following.

2.1 Data

Data were from a multi-phased, longitudinal study conducted at 10 nursing schools across the United States [18]. A total of 475 nursing students started the study. Participants were randomly assigned to 4 initial acquisition conditions where they completed 4 consecutive CPR training sessions that were spaced by 1 day, 1 week, 1 month, or 3 months. Students were additionally randomized to 3 maintenance training conditions where they refreshed their skills for 1 year at intervals of 3 months, 6 months, or at personalized intervals prescribed by the cognitive model. During each session, students completed a series of CPR events using the Resuscitation Quality Improvement (RQI®) system with Laerdal's Resusci Anne® adult QCPR manikin.

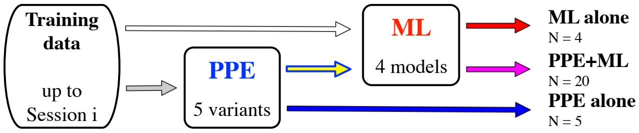

Figure 1: Schematic of the various predictive models.

First, students completed a pre-test consisting of 60 compressions or 12 bag-mask ventilations with no feedback from the manikin, followed by trainings where students received real-time, dynamic feedback to guide, and then post-tests with no feedback. RQI provides composite scores for the quality of compressions and ventilations on scales of 0 to 100, with higher scores corresponding to better performance. The compression score is based on depth, rate, release, and hand placement. The ventilation score is based on volume, rate, and compliance with inspiration time.

Prior to the onset of the study, participants completed a demographic questionnaire. Of the 475 participants that began the study, we included in our study the 393 that completed the initial acquisition phase. Due to the variations in retraining schedules across the maintenance phase, not all students completed the same number of sessions. We focus here on data from the first through the eighth session since the majority of students completed 8 sessions.

2.2 Input features

This section describes all information available in the "training data" box of Figure 1. As sessions progress, the number of available training instances grows but the number of input features is constant. Using the color-coded arrows in Figure 1 to categorize them, these features are detailed in the following and the `labels` correspond to those in Figure 4A.

Gray arrow in Figure 1: `time/lag`: An event's timestamp expressed as the number of seconds since the first/previous recorded event; `score`: The composite performance score recorded for each event.

White arrow in Figure 1: `compvent`: Does the recorded event correspond to performing compressions or ventilations?; `acqint` and `maintint`: What acquisition-maintenance interval combination was this user assigned to? Together, these define the 12 experimental conditions (see previous section for detail) that the condition PPE variant is based on; `session`: Session counter 1 through 8; `pretrnpst`: Was this a pre-test, training, or post-test?; `site`, `age`, `gender`, `height`, and `weight`: Demographic information associated with each user (time-invariant). There were ten sites/locations and other information was coded in years, male/female, inches, and pounds, respectively; `profile`: We reduced the unique user IDs to a low-dimensional set of performance profiles. This approach was inspired by earlier work [2, 23] that showed a small number of descriptive profiles could be obtained by performing k-means clustering [9]. Here, we used $k = 4$, and re-partitioned the training data on each iteration of the simulation. Specifically, only pre-test scores across both skills were taken into account.

Yellow arrow in Figure 1: `decay rate` and `model time`: The two PPE terms computed when fitting PPE (see section 2.3.1) are passed to the ML models; `residual`: The difference between PPE's fit and `score`.

The three rightmost arrows indicate the predictions that are made, emphasizing that the PPE+ML ensemble models (purple) uses both the cognitive (blue) and machine learning (red) models. Next, we outline which ML models were used and how the five PPE variants were fit to the data.

2.3 Predictive models

As noted in Figure 1, there are a total of 29 models. Here, we describe the PPE and ML models that make up the PPE alone/ML alone predictions. The remaining majority of models are based on combining the PPE+ML models by first fitting the PPE as described below and subsequently computing the PPE's residuals in the training data and training the ML model to predict those. The PPE+ML predictions can thus be thought of PPE predictions that were fine-tuned by a given ML model.

2.3.1 Predictive performance equation (PPE)

The PPE is a set of nested mathematical equations that capture findings in the cognitive science literature associated with the temporal dynamics of human learning and forgetting [26]. These include the power law of learning, the power law of forgetting, the spacing effect, and effects of relearning. Two essential components of PPE are sub-equations that model time and the rate of knowledge/skill decay. The *model time* equation captures the idea that the age of items in memory should be skewed toward the most recent presentations, but the full study history should be represented. Hence, model time for each instance i (across n instances) is $w_i \times t_i$, where $w_i = \frac{t_i^{-0.6}}{\sum_{j=1}^{n} t_j^{-0.6}}$ and t_i is the time, in seconds, relative to the first instance. The *decay rate* equation captures the idea that spacing practice across time produces more stable knowledge that decays at a slower rate, while massing practice produces less stable knowledge that decays at a faster rate. Since model time and the decay rate are essential to how PPE captures learning and decay, we include them separately as features in the machine learning models. For a more extensive description of the mechanics of the PPE, please see [25, 26].

In the CPR field study, PPE was fit separately to each participant's history of compression and ventilation scores. We refer to this variant as the *original* PPE throughout the paper. The rationale for individual fits was from prior research suggesting that each individual would have unique learning and forgetting trajectories across sessions due to psychometric differences, the trajectories would vary for compressions and ventilations, and thus predictive accuracy would be maximized by fitting to each student on each skill.

Here, we conduct post-hoc simulations to explore these assumptions by comparing the methodology used in the field study to less granular PPE variants in which free parameters are fit to: experimental condition (acquisition and maintenance intervals), skill (compressions and ventilations), user, or user's performance profile. By exploring these different groupings for which a set of unique parameters are estimated, we evaluate the trade-space between model flexibility and predictive accuracy, and how this interacts with the amount of data available for fitting the models.

2.3.2 Machine learning models

We used four different machine learning models. Depending on the approach, these were either trained to predict the score (red arrow in Figure 1) or PPE's residuals (purple arrow in Figure 1). For either task, all models had access to all input features outlined in section 2.2 (also see x-axis of Figure 4A). All models were run with the default settings of the cited `R` packages [21].

As the simplest model, we fit a single **decision tree** to the scores/residuals. Each tree was pruned through 10-fold cross validation—as implemented in [22]—to avoid overfitting. In most cases, this resulted in very shallow trees and sometimes even single node "stumps." Hence, the decision trees can be thought of as baseline models.

The most complex model was a **random forest** [4], which is an ensemble of decision trees. Using the default settings in [15], we used both bagging and random feature sub-setting to grow a forest of 500 trees. A recent comparison of gradient boosting algorithms included random forests as a comparison and nicely showed that they perform very well on a range of ML tasks and have the added benefit of not requiring hyperparameter tuning [3]. The disadvantage of random forests—as with many ML approaches—is that the internal mechanics that result in a prediction are not easily inspected (see our discussion around Figure 4 below).

The two other models were **ridge** regression and the **lasso** [10], which apply slightly different shrinkage terms when coefficients are estimated. The key difference between the two methods is that ridge regression will retain coefficients for all input features, while the lasso effectively performs feature selection (see section 6.2 in [12] for an introduction). This generally makes lasso models more interpretable. Both models have a single hyperparameter, λ, that was tuned for each iterative prediction using the cross-validation procedure implemented in [8] and all numerical features were standardized.

2.4 Simulation study and analysis

The approach to our simulation study can be summarized as follows: For each session $s = 1, 2, \ldots, 7$, train the models on data up to session s and issue predictions for the pre-test of session $s + 1$. We focused on pre-test scores because we were interested in predicting students' readiness to perform CPR compressions and ventilations, prior to additional training. The procedure was run for all 29 (combinations of) models described above and generated iterative predictions for sessions 2 through 8. The quality of predictions across the models will be compared by computing the mean absolute error (MAE), which summarizes how accurately, on average, each model predicts the scores recorded in the subsequent session. This yields $7 \times (4 + 20 + 5) = 203$ (i.e., number of predicted sessions times number of models) errors. For the sake of easier presentation of these results, we subsequently

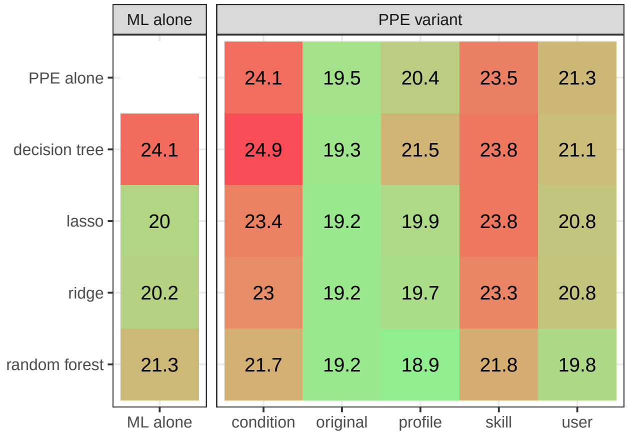

Figure 2: Average MAE across sessions for all models.

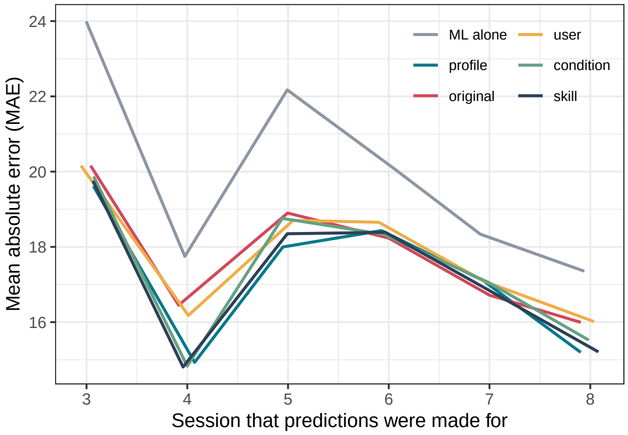

Figure 3: Comparison of all models in the "random forest" row of Figure 2, showing prediction errors for each session.

summarize the errors by (i) computing the average MAE for each model (Figure 2), and (ii) ranking the models based on their errors (Table 2). These overall results are elaborated on with additional figures and tables that highlight relevant details.

3. RESULTS

Figure 2 presents the average MAE for all 29 models and speaks to all three research questions posed in section 1.1. As detailed in section 2.4, the 203 prediction errors were aggregated across sessions and the average MAE for each combination of models is presented as a heatmap. The color-coding corresponds to the magnitude of the errors; lower values are better. By averaging across sessions, variations in predictive accuracy as a function of session (see Figure 3) is lost but it becomes easier to assess the model's relative performance in one glance.

First, we can look at the "PPE alone" row in Figure 2 to compare the five PPE variants that were tested. Overall, the original instantiation of the cognitive model as used in the CPR field study—if used alone—does indeed outperform the more constrained variants explored here. This is somewhat surprising since the original model exhibited signs of overfitting (i.e., fit MAE lower than prediction MAE) that were ameliorated for the constrained variants of PPE. However, it appears that despite overfitting the training data, the original variant of PPE did produce the best predictions after all.

Second, whether a number of off-the-shelf ML approaches would have yielded superior predictions than the cognitive model can be assessed by comparing the cell original PPE alone (MAE = 19.5) with the prediction errors in the "ML alone" column in Figure 2. All ML approaches yield average errors larger than 19.5 when applied alone, which suggests that the ML models tested here—if used by themselves—would not have resulted in better predictions overall. However, the differences in prediction errors are not large and the ridge and lasso regression in particular perform well on average.

And third, and most importantly, we turn to the PPE+ML combinations. These correspond to the larger facet labeled "PPE variants" in Figure 2. A number of notable patterns emerge: the average MAE for the original PPE is hardly affected by adding any of the ML models to predict its residuals. This might be because this variant of PPE is very flexible, which restricts the residuals in the training data that the ML can actually fit to. For all other PPE variants, we see a gradient from top to bottom, with average MAEs decreasing with ML models relative to PPE alone. The decision trees are an exception to this pattern and seem to worsen the performance more often than not. Otherwise, we generally see the lasso and ridge regressions improving on PPE alone and the PPE+random forest resulting in the best performance for all PPE variants.

Zooming in on the models using random forests, Figure 3 shows the session-by-session prediction errors of the random forest alone (ML alone) and the PPE+random forest combinations. We omitted predictions for the second session because they are quite poor for the random forests combined with the condition and skill PPE variants, which distorts the y-axis and obscures differences between models in the later sessions[1]. The figure highlights that the random forest alone consistently performs worse than all other combinations of models, in which the random forest is used specifically to learn PPE's residuals rather than observed performance. This suggests that the most promising approach is an ensemble of a PPE variant that captures the overall temporal dynamics to issue predictions that are subsequently fine-tuned by a random forest that can leverage all other available input features.

Another way to summarize these results is by ranking all 29 models' MAE within each session and computing the average rank for each PPE variant. These average ranks are shown

[1] Predictions for session 2 were generally much worse than for all subsequent sessions. We ran all analyses reported here without session 2 predictions to confirm that our conclusions do not depend on differences between models on session 2. If session 2 is omitted, the skill PPE variant performs a little better overall but results are not otherwise affected drastically.

Table 1: Comparison of PPE variants across all ML models.

PPE variant	average rank	average MAE
profile	11.9	20.1
skill	12.7	23.2
original	15.1	19.3
user	17.2	20.8
condition	18.4	23.4

in Table 1, and reveal that although the original PPE yields the lowest overall average MAE, both the profile and skill variants achieve better average ranks. This suggests that if ML models are leveraged to predict PPE's residuals, more constrained variants of PPE tend to perform better. However, even the lowest average ranks listed in Table 1 is relatively high, suggesting that no model consistently outperforms the others. This observation is confirmed by inspecting the models' MAEs in detail (not shown here), which reveals that for some sessions, most models perform effectively identically.

Lastly, we present the top 10 models in terms of their ranking in Table 2. Here, the ranks are computed as an average across the seven sessions each model made predictions for. The best-performing model is the PPE with parameters for each performance profile whose residuals are predicted by a random forest. Figure 2 corroborates this observations, showing that this combination of models obtains the lowest average MAE overall. Notably, all five instances of the original PPE and five out of six instances of random forests are represented in the top 10, confirming that these models perform very well in various combinations.

3.1 Peeking into the random forest

Space constraints limit the amount of model interrogation we can report here. However, we want to at least showcase one prominent example. Table 2 and Figure 2 show that the best model overall is the combination of the PPE variant with unique parameters for each performance `profile` and a random forest that learns its residuals. (This model is the blue line in Figure 3, which highlights that other models perform very similarly.) Figure 4A shows the normalized feature importance computed for each input feature (white and yellow arrows in Figure 1) for each iterative session that predictions are made for. Superimposed are the average importance and the spread (in black) and features are sorted from least to most important based on average importance. Notably, most time-invariant features (gender, age, etc.) are equally important across the seven iterations. The `session` counter, `stability`, and `model time`, on the other hand, become gradually more important as more sessions were included in the training data, while the opposite pattern is evident for `compvent` and users' performance `profile`.

Feature importance plots as shown here can be informative but should be interpreted with caution since they do not capture and visualize the potential intricate non-linear relationships between the various input features [17]. Furthermore, feature importance and their impact on predictions are not necessarily the same—more advanced approaches exist [19] but are beyond the scope of the current paper.

Table 2: The top 10 models overall sorted by average rank across the seven predictions made by each model.

	PPE variant	ML model	average rank
1	profile	random forest	5.3
2	skill	random forest	6.7
3	condition	random forest	7.9
4	original	lasso	8.1
5	original	random forest	8.3
6	original	decision tree	8.4
7	original	ridge	8.6
8	user	random forest	10.1
9	original	PPE alone	10.7
10	condition	ridge	11.9

Figure 4B zooms in on two important features and shows the predictions made for the profile PPE model for the fourth session against the residuals the random forest predicts for each instance. We generally see the most differentiation between models on Session 4, which is why we chose it—however, this figure is broadly representative of the profile PPE+RF dynamics for other sessions. Figure 4B suggests that ventilations are more often down-adjusted than compressions (i.e., more triangles below the equality line) unless PPE predicts near-ceiling performance. The fact that model time is consistently identified as the most important feature (see Figure 4A) but no clear relationship between the magnitude of the adjustment (i.e., distance from equality line) emerges in Figure 4B highlights the disadvantage of applying ML models—such as a random forest—that are challenging to interrogate.

4. DISCUSSION

The post-hoc simulation study reported here suggests that the original cognitive model used for prescriptive, adaptive scheduling in the CPR field study performed very well overall. In fact, in the aggregate, it resulted in lower average prediction errors than both the more constrained variants of PPE and the machine learning models included in the current comparison. Thus, it is unlikely that the tested off-the-shelf ML models would have performed better than the original PPE, although the regularized regression models (ridge and lasso) in particular achieved prediction errors similar to the original PPE. We expected the ML models to outperform the cognitive model because the latter's main "insights" (the estimated *model time* and *decay rate*) were included as input features to the ML models (yellow arrow in Figure 1. This suggests that the PPE, using much less information, was slightly better at extrapolating performance to the next session.

The current explorations also showed, however, that an ensemble cognitive and ML model has the potential to perform slightly better than either alone. Notably, the more constrained variants of PPE performed particularly well in this ensemble arrangement. One possible explanation is that the less flexible cognitive model operates as a smoothing function on the temporal information, which leaves the ML to learn under which conditions (i.e., [combinations of] input features) the general temporal dynamics should be adjusted to further improve predictions. This framing of the procedure is

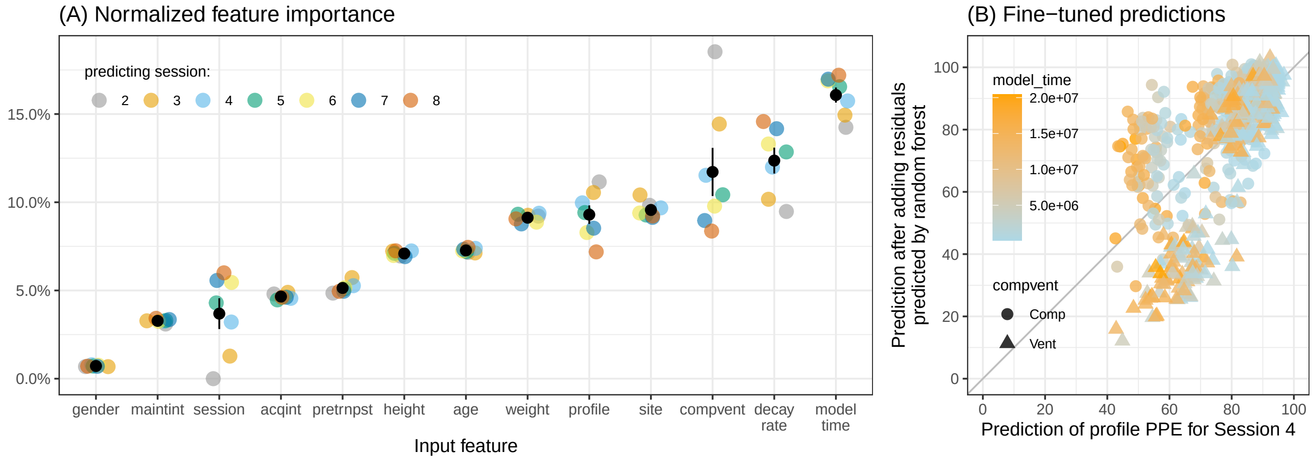

Figure 4: Details on the best-performing model. (A): The normalized feature importance for each iteration of the model with superimposed averages (black dots). (B): Initial PPE predictions for Session 3 (x-axis) plotted against fine-tuned predictions (y-axis); color-coding indicates the `model time` for each predicted instance and shape differentiates `compvent`.

conceptually akin to a two-step boosting algorithm [7] and the “fine-tuning” of predictions induced by the second step (the ML model in this case) is nicely illustrated in Figure 4B, which shows the results of the first step (the constrained cognitive model’s predictions on the x-axis), against the results after the second boosting-like step (the PPE+ML predictions on the y-axis). In this process, the initial PPE prediction’s quite restricted range is expanded by the random forest, which can—and does (cf. Figure 4A)—draw on various input features.

The ensemble approach outlined here has the added benefit of a modular structure. Thus, it is easy to make design decisions, particularly regarding (i) which constraints should be built into the parameter fitting procedure for PPE, and (ii) what type of ML model is most informative. The latter will determine where on the continuum of interpretability the ensemble will fall. For example, the dynamics of the random forest that predicts the profile PPE’s residuals (highlighted in Figure 4), does not lend itself to straightforward model interrogation but the PPE+lasso and PPE+ridge combinations would not reduce the interpretability of the ensemble, while slightly reducing prediction errors relative to PPE alone (see Figure 2).

It should be pointed out, however, that the improvement in prediction errors relative to the original PPE alone is minor. Nevertheless, we consider these findings significant for two reasons: First, the small improvement vindicates that PPE’s time-based mechanisms capture the majority of variance in this task domain. Second, the PPE+ML ensemble approach used here serves as a proof-of-concept that illustrates how the core mechanism of PPE can be preserved while incorporating an arbitrary number of additional input features. For example, some of the input features used here were specific to the field study’s design (notably `acqint` and `maintint`) and would not be present in the hospital setting RQI systems are primarily deployed in. In such settings, however, other input features would be available (e.g., job title or department) and samples would be larger and more heterogeneous, which would conceivably introduce more variance that is not a function of time-based features alone. We expect that under these conditions, the ensemble approach’s advantage over PPE alone would be more pronounced.

In the current effort, we choose to assess the models’ ability to make session-by-session predictions. This approach meant that events did not line up chronologically (a student in the weekly acquisition condition will have completed the first four session before a student in the 3-month condition returned for their second session) but the amount of training data available for each student is equalized—only the lag between events varies. This reveals, for example, that predictions improve up to session 4 (the end of the acquisition phase; see Figure 3) and then get worse for session 5, which is when students switch to the maintenance phase. This suggests that the models get better at forecasting performance as more data from a consistent schedule becomes available, and that one should expect a dip in predictive accuracy as the temporal dynamics are altered.

Future work in this domain should validate the approach presented here in more naturalistic data that more closely resemble how medical professionals train and maintain CPR proficiency. We believe that cognitive models in particular—and a cognitive-machine learning ensemble specifically—hold great promise in moving towards a predictive framework that affords personalized, adaptive refresher training schedules that are tailored towards individual learning needs—either of an individual or groups of learners that exhibit similar performance profiles. Furthermore, the outlined predictive pipeline’s potential value in adaptive, educational learning system outside of the medical domain should be explored.

5. ACKNOWLEDGEMENTS

This work was funded through the 711th Human Performance Wing Chief Scientist Seedling award at the Air Force Research Laboratory. Data were wrangled in `R` using [28]; tables were created with [11], and figures with [27] and [20].

6. REFERENCES

[1] B. S. Abella, J. P. Alvarado, H. Myklebust, D. P. Edelson, A. Barry, N. O'Hearn, T. L. V. Hoek, and L. B. Becker. Quality of cardiopulmonary resuscitation during in-hospital cardiac arrest. *Jama*, 293(3):305–310, 2005.

[2] E. Ayers, R. Nugent, and N. Dean. Skill set profile clustering based on student capability vectors computed from online tutoring data. *Educational Data Mining*, 2008.

[3] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, pages 1–31, 2020.

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] A. Cheng, D. J. Magid, M. Auerbach, F. Bhanji, B. L. Bigham, A. L. Blewer, K. N. Dainty, E. Diederich, Y. Lin, M. Leary, et al. Part 6: resuscitation education science: 2020 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 142(16_Suppl_2):S551–S579, 2020.

[6] A. Cheng, V. M. Nadkarni, M. B. Mancini, E. A. Hunt, E. H. Sinz, R. M. Merchant, A. Donoghue, J. P. Duff, W. Eppich, M. Auerbach, et al. Resuscitation education science: educational strategies to improve outcomes from cardiac arrest: a scientific statement from the american heart association. *Circulation*, 138(6):e82–e122, 2018.

[7] Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[8] J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010.

[9] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108, 1979.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009.

[11] M. Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables.* Central European Labour Studies Institute (CELSI), Bratislava, Slovakia, 2018. R package version 5.2.2.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning.* Springer, 2013.

[13] T. S. Jastrzembski, K. A. Gluck, and G. Gunzelmann. Knowledge tracing and prediction of future trainee performance. In *Interservice/Industry Training, Simulation, and Education Conference*, pages 1498–1508. National Training Systems Association, 2006.

[14] T. S. Jastrzembski, M. Walsh, M. Krusmark, S. Kardong-Edgren, M. Oermann, K. Dufour, T. Millwater, K. A. Gluck, G. Gunzelmann, J. Harris, et al. Personalizing training to acquire and sustain competence through use of a cognitive model. In *International conference on augmented cognition*, pages 148–161. Springer, 2017.

[15] A. Liaw and M. Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[16] R. M. Merchant, A. A. Topjian, A. R. Panchal, A. Cheng, K. Aziz, K. M. Berg, E. J. Lavonas, D. J. Magid, A. Basic, P. B. Advanced Life Support, R. E. S. Advanced Life Support, Neonatal Life Support, and S. of Care Writing Groups. Part 1: Executive summary: 2020 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 142(16_Suppl_2):S337–S357, 2020.

[17] C. Molnar. *Interpretable Machine Learning.* 2019. https://christophm.github.io/interpretable-ml-book/.

[18] M. Oermann, M. Krusmark, S. Kardong-Edgren, T. S. Jastrzembski, and K. A. Gluck. Personalized training schedules for retention and sustainment of CPR skills. *Simulation in Healthcare*, 2021.

[19] T. Parr, J. D. Wilson, and J. Hamrick. Nonparametric feature impact and importance. *arXiv preprint arXiv:2006.04750*, 2020.

[20] T. L. Pedersen. *patchwork: The Composer of Plots*, 2019. R package version 1.0.0.

[21] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

[22] B. Ripley. *tree: Classification and Regression Trees*, 2019. R package version 1.0-40.

[23] F. Sense, M. Collins, M. Krusmark, and T. S. Jastrzembski. Using k-means clustering for out-of-sample predictions of memory retention. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020.

[24] G. Shmueli et al. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

[25] M. M. Walsh, K. A. Gluck, G. Gunzelmann, T. Jastrzembski, and M. Krusmark. Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive science*, 42:644–691, 2018.

[26] M. M. Walsh, K. A. Gluck, G. Gunzelmann, T. Jastrzembski, M. Krusmark, J. I. Myung, M. A. Pitt, and R. Zhou. Mechanisms underlying the spacing effect in learning: A comparison of three computational models. *Journal of Experimental Psychology: General*, 147(9):1325, 2018.

[27] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

[28] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.