Deep Learning applied to Learning Analytics and Educational Data Mining: A Systematic Literature Review

Orlando Bisacchi Coelho, Ismar Frango Silveira

Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie (UPM)
Rua da Consolação, 930 – São Paulo – SP – Brazil

{orlandoc,ismar}@mackenzie.br

Abstract. This work presents, to the extent of the authors' knowledge, the first systematic literature review of the application of Deep Learning to Educational Data Mining and Learning Analytics. Previous literature reviews have documented several works in the areas of Educational Data Mining and Learning Analytics that used classical Artificial Neural Networks techniques. But none of them mentioned the new and much more powerful paradigm in Artificial Neural Networks: Deep Learning. This work surveys this new technique and identifies recent works in Learning Analytics and Educational Data Mining that have applied Deep Learning techniques.

1. Introduction

Educational Data Mining (EDM) [Romero and Ventura 2010, Baker and Inventado 2014] aims at developing methods and techniques to deal with the sort of large-scale data that arises in educational settings in order to better understand the educational questions that arise in these environments. The very close discipline of Learning Analytics [Ferguson 2012] also relies on data that is generated in educational situations but is mostly focused on extracting knowledge from the data in order to directly foster the learning process. Irrespective of the difference between the two approaches [Ferguson 2012, Baker and Inventado 2014, Chatti et al. 2014], both use techniques developed by Data Mining, Machine Learning and Statistics research communities in order to study and foster teaching and learning processes. Previous surveys of the main analytical techniques used in Educational Data Mining and Learning Analytics [Romero and Ventura 2010, Baker and Inventado 2014, Avella et al 2016] show that quite a large collection of techniques have been used by the practitioners. In a survey of the state of the art in Educational Data Mining, covering publications up to 2009, Romero and Ventura (2010) state that Artificial Neural Networks was already one of the most used techniques in EDM research.

An Artificial Neural Network (ANN) [Haykin 2008] is a very simplified computational model of the Central Nervous System. It is a labelled oriented graph where the vertices are arranged in layers. The vertices are quite simple computational units that correspond to simplified biological neurons. They integrate, in a nonlinear fashion, the signals they receive from other units. According to the signal that they receive they compute their own activation level. The graph edges correspond to the synapses between units. They encode the influence of a given unit over any other unit that receives a connection from the former. Each given connection can either increase or decrease the

DOI: 10.5753/cbie.sbie.2017.143

receiving unit's activation and this influence is represented by the weight associated to the edge. Just like the Central Nervous System, an ANN is a learning system; it learns to perform a task from data that exemplifies the task being performed. An ANN learning algorithm is a procedure to adapt the network's weights in a way that the network starts to perform a computational task, as expressed by the association of the activation of units in the input layer to the corresponding activation of units in the output layer. ANN learning algorithms comprise the full range of paradigms: unsupervised, supervised and reinforcement learning.

ANN is also a widely used technique in Data Mining [Lu, Setiono and Liu 1996, Stahl and Jordanov 2012]. It has been applied to the tasks of clustering, classification, regression, time series forecasting and visualisation. Despite ANN's quite widespread usage in data mining, it should be noted that part of the data mining community is somewhat critical of the lack of interpretability of the models developed according to this paradigm [Craven and Shavlik 1997]. They do produce accurate predictions from the data, but it is quite hard to extract human-interpretable rules that summarize their predictions.

In the late 90's, ANN research started to loose strength [Le Cun, Bengio and Hinton 2015] for several technical reasons. Those difficulties were only solved in the middle of the last decade. Developments in terms of algorithms, network architectures and hardware then allowed training networks comprising even tens of layers and several millions of connections on datasets of millions of examples. This new paradigm of ANN was called Deep Learning (DL) [Schmidhuber 2014]. These deep networks are trained on GPUs instead of CPUs and benefit from custom software environment. The main architectures and algorithms used in DL are Deep Belief Networks, Convolution Nets and Long Short-Term Memory [Goodfellow, Bengio and Courville, 2016]. In the last few years, DL has managed to solve a series of long-standing problems in Artificial Intelligence, like speech recognition, text translation, textual sequence generation, image classification, and text generation from image. Such developments basically rendered the classic (non-deep) ANNs obsolete.

Although the difficulty of understanding the solutions generated by Deep Learning in a form that is easily comprehensible to humans persists, the current perception is that the power of DL comes exactly from the networks' ability to learn to develop internal, non human engineered representations that make the computational task learnable [Le Cun, Bengio and Hinton 2015]. Deep Learning is currently seen as the state of the art solution for performing several tasks in Data Mining related to classification, regression and time series forecasting. Najafabadi and colleagues (2015) show applications of DL for extracting complex patterns from massive volumes of data, performing semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. DL seems particularly powerful for making sense of mixed modalities, such as audio, image and video.

According to Romero and Ventura (2010), (classic) Artificial Neural Networks have been used for quite diverse applications such as predicting student performance, student modelling, grouping students according to their personal characteristics, making personalized recommendations, providing adaptive and personalized learning support to students and planning courseware. Nevertheless, in a briefer survey of research developed by both EDM and LA communities, Baker and Inventado (2014) pointed out that ANNs

are not a typical method of choice in EDM (arguably the same could be said for LA). They speculate that this is due to the community's preference for more conservative algorithms. In a survey, Jindal and Borah (2013) identified three articles using ANNs for EDM in the period 2010-2013 [Dutta et al., 2011, Wang and Liao 2011, Yu et al. 2010, on enrolment decision, student retention and language learning. In their review on EDM, Mohamad and Tasir (2013) reported just one paper using ANN for predicting the pattern of behaviour of students [Blagojevic and Micic 2013]. In a systematic literature review of empirical work in LA and EDM, Papamitsiou and Economides (2014) only managed to find one paper using ANN techniques – for prediction of student performance, retention and satisfaction –, only one of them published in the current decade [Guo 2010]. Shahiri, Husain and Rashid (2015) mention four papers using ANNs to predict student performance (according to them the best method for that application), but just one of them from the current decade [Arsad, Buniyamin, Manan 2013]. Avella et al., (2016), in a recent SLR of LA methods in Higher Education, also managed to find just one paper using ANNs [Vahdat et al. 2015]. That begs asking if there has been a sharp decrease in the use of ANNs in EDM and LA recently, and why. Apart from that, a striking feature in all major literature reviews in EDM and LA is the total absence of works that benefit from the current generation of ANNs, Deep Learning.

The current work is a systematic literature review (SLR) of the recent scientific literature in Educational Data Mining and Learning Analytics that uses Deep Learning techniques. It is not a survey, but a systematic literature review, as per the methodological guidelines discussed in Okoli and Schabram (2010) and Moher et al. (2009). The timeframe chosen for this SLR is the period from 2010 to the present for two reasons: (i) this dovetails with the period covered by Romero and Ventura (2010)'s survey and (ii) this is the year the first open access general purpose GPU-based framework for Deep Learning programming was released [Bergstra et al 2010]. This work identifies the very first applications of DL to EDM and LA. Finally, it identifies the computational tasks that support EDM and LA research that can benefit most from DL techniques and the subareas of EDM and LA they can most successfully be applied to. In the next section a detailed account of the methodology used for the SLR is presented and the research questions addressed by the SLR is formulated. Then the results of the research are presented. The final section is a discussion of the main results.

2. Methods

What characterises a literature review as systematic is its adherence to a sound and explicit methodology. Fink (2005) stresses that a research literature review should be systematic, comprehensive and reproducible. Systematic in the sense of abiding to a precise, explicit methodology; comprehensive in terms of breadth of scope, which should be clearly delimited; based on that the review would also be reproducible by any independent researcher. PRISMA [Moher et al. 2009] is a coherent set of guidelines widely used for systematic literature views and meta-analyses. It includes a very useful checklist of items that should be included when reporting a SLR. Nevertheless PRISMA is somehow biased to the health sciences area where it was generated. Okoli and Scahabram (2010) is a quite recent guide for performing SLRs in the area of Information Systems. They propose an eight-step method for conducting a SLR that is used in the current work.

The objective of the current SLR is to survey the research in EDM and LA carried on since 2010 that uses Deep Learning techniques, identifying the educational tasks and EDM/LA research questions DL techniques can help to address. Regarding that period of time and context of application, the research questions this review intends to answer are:

RQ1. Have Deep Learning techniques started to be used? If that is the case, what is the balance between Artificial Neural Network and Deep Learning-based research?
RQ2. What are the main educational tasks or research questions in Learning Analytics and Educational Data Mining that Deep Learning has addressed?

Since a single person carried out the review, no formal written protocol was developed. In order to perform a comprehensive selection of the target studies, the following scientific bases were used: ACM Digital Library, Google Scholar, IEEE Xplore and the non preview-only content of Sprinker Link. Only works in English were considered. In order to also capture the recent works using Artificial Neural Networks, the following search string was used (from a logical point of view, adaptations had to be made for each search engine) was: *("neural network" OR "deep learning") AND ("educational data mining" OR "learning analytics")*. The string was searched for in the full text of the article, since most bases did not allow searching only in the title, abstract and keywords. The search was performed in June of 2017.

The search resulted in a huge amount of results; the number of total distinct results is not reported since searching different bases could return the same paper more than once. Also the four different combinations of substrings that sometimes had to be used in searching a given base (e.g. *"deep learning " AND "learning analytics"*) returned many repeated results across all searches. Results that were unrelated to the focus of the survey were discarded. This practical screening [Okoli and Schabram 2010] was based on reading the title, abstract and keywords of each work and, if necessary, skimming through the text. A practical step taken was that from each of the four different searches carried on Google Scholar only the first 25 results returned were taken into consideration. As it happened, the practical screening ended up by excluding all the non preview-only content of Sprinker Link. That process led to the identification of 46 works relevant for this review. On the other hand, 3 articles cited in the surveys referred to in the previous section but not found in the databases searched were added to the set, since the surveys were found in Google Scholar and the three articles would also found if the 25-result limited was not imposed for practical reasons. That took the total to 49.

After consolidating the works obtained, a quality appraisal step excluded just a few of the works, on the basis of three criteria. First, only primary research was included, not surveys or reviews of the EDM & LA area. Second, works that did not report any implemented and tested system nor presented a well-developed theoretical discussion were also excluded. A total of 3 works were so excluded. Lack of quality was the last exclusion criterion; only 1 paper was discarded for that reason. Therefore, at the end, 45 articles remained. The works were then read in order to extract from them the information to answer the two research questions listed before. The results of the process are presented in the next session.

3. Results

The search using the string: *("neural network" OR "deep learning") AND ("educational data mining" OR "learning analytics")*, followed by the quality appraisal step afforded 45 results, as shown in Table 1.

Table 1. Number of relevant results obtained from each scientific base using the search string after the quality appraisal step.

Scientific base	ACM Digital Library	IEEE Xplore	Google Scholar	From older reviews	Grand total
ANN-related	3	20	13	3	39
DL-related	4	2	0	0	6
Total	7	22	14	3	45

Figure 1 helps to answer RQ1. EDM and LA research based on ANN techniques has been published every year. But the number of papers is quite low: an average of 5 papers per year. Publications using DL techniques first appeared in 2015. Every year saw the publication of only 2 papers using DL, one third of the amount of ANN-based publications in the same time. Half the works sprang from American institutions and the other ones from three different Asian countries. No group published twice. Adding up both techniques, the average is 8 papers per year for the last three years.

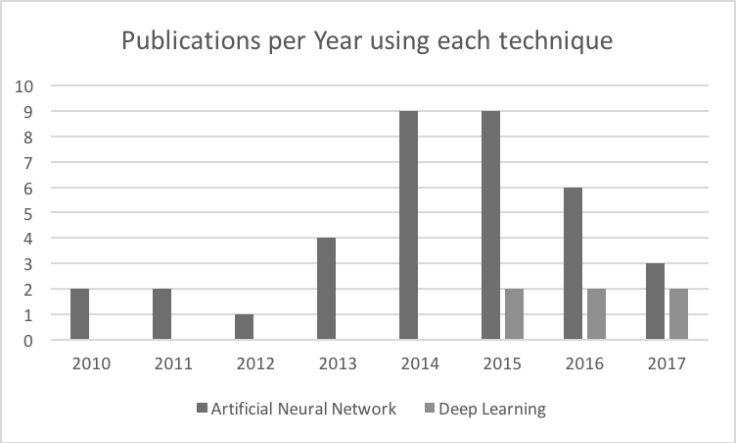

Figure 1. Number of publications in EDM and LA per year using either ANN or DL.

The full list of the thirty-nine papers found in the search that used ANN techniques – which space restriction makes it impossible to discuss in this work – can be found in the following address: https://goo.gl/8BRMR3 .

Gross and colleagues (2015) present the handwriting recognition component of the Xerox Ignite Educator Support System, a product designed to be used for K-12 educators, a commercial system that lifts and identifies student responses from paper or tablet, autoscores the answers and allows the teacher to validate the grades. Student

answers can be either numerical, a combination of number and letters or simple sentences consisting of up to five words. The recognition occurs in a pipeline that begins by image pre-processing followed by character segmentation, character recognition and finally word recognition. The character recognition step is based on Deep Learning. The recognition process depends on a majority voting by Convolutional Neural Networks and Stacked De-noising Auto-encoders [Goodfellow, Bengio and Courville, 2016]. The training is done on top of a combination of public datasets and private datasets collected from students. The authors report 92.8% accuracy in the task.

Guo and collaborators (2015) develop a student performance prediction system that can be used to act as an early-warning system by detecting students at risk of failing a course. Based on data collected from 12,000 9-year students from 100 Chinese schools, comprising background, personal, school, past study, and current study data, they train a six-layer feedforward network to predict student grade in the high school Entrance Examination, modelled as a one-in-five class result. The hidden networks of the network are first trained in an unsupervised way, layer by layer, using auto-encoders. After all the weights are so initialized, backpropagation [Goodfellow, Bengio and Courville, 2016] is then used to train the whole network in the grade prediction task. The authors show that their approach compares favourably (on the same dataset) to other methods used in the literature, achieving overall accuracy of 77.2%.

The work by Li, Wong and Kankanhal (2016) is the first automated presentation assessment system based on Deep Learning. Each presentation is scored by humans in terms of posture, gesture, eye contact, fluency, liveliness, pronunciation, speech rate, and presenter-audience interaction. Using Kinect sensors, Google Glass and audio, information about body language, eye contact and speech generated by each presenter is captured. Then one set of bidirectional Long-Short Term Memory (LSTM) networks [Graves, Mohamed and Hinton 2013] processes the audio input while another set of bidirectional LSTMs (gated by a further layer of bidirectional LSTMs acting as an attention focusing mechanism) processes the video frames and skeleton representations. The networks that deal with the different modalities are connected together and the resulting deep framework is trained end-to-end. The authors show that the proposed framework achieves better results when different modalities are combined and outperforms alternative models.

Okubo and colleagues (2017) develop a predictor for the final grade at a course taken by university students. The prediction is made on the basis of the grading the students received in several learning activities performed each week. Compared to multiple regression analysis, the Long-Short Term Memory achieves far better results, with an accuracy of more than 90% from the point where 40% of the course had been completed and 100% when 2/3 of the course had been completed.

Tang, Peterson and Pardos (2016) outline two possible applications of Deep Learning in Intelligent Tutoring Systems. One is word suggestion when a student gets stuck while writing an essay. The other is suggesting a learning resource for a student on the basis of the action stream she generates while interacting with a MOOC. Both are geared for university students and are modelled using Long-Short Term Memory. Both experiments are quite preliminary but positive, entailing further exploration.

Finally, Wang and collaborators (2017) develop an LSTM-based model to predict a given student's performance in a future programming exercise on the basis of the

sequence of solution attempts she submitted to the system in the current exercise. But their main goal is not to be able to predict future student performance but to understand the learning process of the student. The LSTM model obtained an increase of 5% in accuracy compared to baseline.

Table 2 summarises the educational and data mining tasks addressed by the works that use Deep Learning.

Table 2. Educational and Data Mining tasks addressed using Deep Learning.

Article	Educational Task	Task Description	Data Mining Task
[Gross et al. 2015]	Computer Assisted Instruction	K-12 student handwriting recognition	Classification
[Guo et al. 2015]	Student Performance Prediction	9-year performance prediction in high school entrance examination	Classification
[Li, Wong and Kankanhal 2016]	Student Assessment	Automated presentation assessment	Classification & Regression
[Okubo et al. 2017]	Student Performance Prediction	University students final grade on course based on weekly learning activities	Classification & Time series forecasting
[Tang, Peterson and Pardos 2016]	Intelligent Tutoring System	Word suggestion (for essay writing) & Learning resource recommendation system for university students	Time series forecasting & Classification
[Wang et al. 2017]	Student Performance Prediction & Student Modelling	Prediction of student performance in programming exercises	Classification

4. Discussion

As expected, the Data Mining taks addressed by the models are the ones that are amenable to supervised learning (classification, regression and time series forecasting), reflecting what can be performed by the most used learning algorithm in Deep Learning: backpropagation.

A good percentage of the works (in a very limited sample, what shall be taken into consideration) fits in what has ben been called Academic Analytics [Campbell and Oblinger 2007], [Guo et al. 2015] and [Okubo et al. 2017]. As expected the results achieved using Deep Learning techniques outperform the ones obtained earlier with ANN techniques.

One research area where Deep Learning can surely make a major impact is multimodal learning analytics [Blikstein 2013]. In fact, DL has become the major

technique for learning from image, video and speech [Le Cun, Bengio and Hinton 2015]. It has also shown itself to be very able to combine multiple modalities [Ngiam et al. 2011]. In this sense, Li and al. (2016)'s paper can be seen as the beginning of a very fruitful line of research.

It is still early days to judge the impact that Deep Learning will have on research in Learning Analytics and Educational Data Mining. The difficulty of interpreting the results produced by both ANN and DL was also an issue for Data Mining researchers that spun from a background in Statistics, in the beginning (but it has never been a big deal for the Machine Learning practitioners). Nowadays, face the astounding results achieved by Deep Learning, that basically turned into a non-issue. It is open to future assessment if that will also happen in the Learning Analytics and Educational Data Mining communities. Nevertheless, there is no denying the power achieved by letting the learning algorithm to automatically develop the internal representations that happen to be the most useful for learning the task at hand – the very idea of Deep Learning.

5. References

Arsad, P.M., Buniyamin, N. and Manan, J.A. (2013) *A Neural Network Students' Performance Prediction Model (NNSPPM)*. Proc. IEEE Intl. Conf. Smart Instrumentation, Measurement and Applications (ICSIMA).

Avella, J., Kebritchi, M., Nunn, S. and Kanai, T. (2016) Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning Journal*, *20*(2), p. 1-17.

Baker, R.S.J.D. and Inventado, P.S. (2014) *Educational Data Mining and Learning Analytics*. In J.A. Larusson, & B. White (Eds.), Learning Analytics: from Research to Practice, p. 61-75. NY, USA: Springer.

Bergstra, J. et al. (2010) *THEANO: A CPU and GPU Math Expression Compiler*. Python for Scientific Computing Conference.

Blagojevic, M. and Micic, Z. (2013) A web-based intelligent report e-learning system using data mining techniques. *Computers & Electrical Engineering*, *39*(2), p. 465-474.

Blikstein, P. (2013) *Multimodal Learning Analytics*. 3rd Intl. Conf. Learning Analytics and Knowledge, p. 102-106.

Campbell, J. P. and Oblinger, D.G. (2007) *Academic Analytics*. EDUCAUSE.

Chatti, M.A., Lukarov, V., Thüs, H. and Schoeder, U. (2014) Learning Analytics: Challenges and Future Research Directions. e-learning and education: eleed, Dec.

Craven, M. W. and Shavlik, J.W. (1997) Using Neural Networks for Data Mining. *Future Generation Computer Systems*, *13* (2-3), p. 211-229.

Dutta Borah, M., Jindal, R., Gupta, D. and Deka, G.C (2011) *Application of knowledge based decision technique to Predict student's enrolment decision*. Proc. Intl. Conf. on Recent Trends in Information Systems, p.180-184.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6) p. 304–317.

Finks, A. (2005) *Conducting Research Literature Reviews: From the Internet to Paper*. 2 ed. Thousand Oaks: Sage Publications.

Graves, A., Mohamed, A. and Hinton, G.E. (2013) *Speech recognition with deep recurrent neural networks*. IEEE Intl. Conf. Acoustic, Speech and Signal Processing, p. 6645-6649.

Goodfellow, I., Bengio, Y. and Courville, A. *Deep Learning*. MIT Press, 2016.

Gross, E., Wshah, S., Simmons, I. and Skinner, H. (2015) *A Handwriting Recognition System for the Classroom*. 5th. Intl. Conf. Learning Analytics and Knowledge, p. 218-222.

Guo, W.W. (2010) Incorporating statistical and neural network approaches for student course satisfaction analysis and prediction. *Expert Systems with Applications*, *37*(4), 3358–3365.

Guo, B., Zhang, R., Shi, C. and Yang, L. (2015) *Predicting Students Performance in Educational Data Mining*. 2015 Intl. Symp. Educational Technology, p. 125-128.

Haykin, S. (2008) *Neural Networks and Learning Machines*. 3. ed, Pearson.

Jindal, R. and Borah M. D. (2013) A Survey on Educational Data Mining and Research Trends, *International Journal of Database Management Systems*, *5*(3), p. 53-73.

Längkvista, M., Karlssona, L. and Lout, A. (2014) A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters, 42*, 1, pp. 11–24.

Li, J., Wong, Y. and Kankanhali, M.S. (2016) *Multi-stream Deep Learning Framework for Automated Presentation Assessment*. 2016 IEEE Intl. Symp. Multimedia, p. 222-225.

Le Cun, Y., Bengio, Y. and Hinton, G. E. (2015) Deep Learning. *Nature, 521*,p. 436-444.

Lu, H., Setiono, R. and Liu, H. (1996) Effective Data Mining Using Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6).

Mohamad, S.K. and Tasir, Z. Educational data mining: A review. (2013) Proc. 9th. Intl. Conf. Cognitive Science.

Moher D., Liberati A., Tetzlaff J. and Altman D.G., The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses*:* The PRISMA Statement. *PLoS Med 6*(7).

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E. (2015) Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*:1.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y. (2011) *Multimodal Deep Learning*. 28th Intl. Conf. Machine Learning.

Ninrutsirikun, U., Watanapa, B., Arpnikanondt, C. and Phothikit, N. (2016) *Effect of the Multiple Intelligences in Multiclass Predictive Model of Computer Programming Course Achievement*. 2016 IEEE Region 10 Conference.

Okoli, C. and Schabram, K. (2010). *A Guide to Conducting a Systematic Literature Review of Information Systems Research*. Sprouts: Working Papers on Information Systems, 10(26).

Okubo, F., Yamashita, T., Shimada, A. and Ogata, H. (2017) *A Neural Network Approach for Students' Performance Prediction*. 7th Intl. Conf. Learning Analytics and Knowledge.

Papamitsiou, Z. and Economides, A.A. (2014) Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, *17*(4).

Romero, C., and Ventura, S. (2010). Educational data mining: a review of the state-of-the-art. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, 40*(6), p. 601–618.

Sarker, F., Tiropanis, F. and Davis, H.C. (2014) Linked Data, Data Mining and External Open Data for Better Prediction of at-risk Students. 2014 International Conference on Control, Decision and Information Technologies, p. 652-657.

Shahiri, A.M., Husain, W. and Rashid, N.A. (2015) A Review on Predicting Student's Performance using Data Mining Techniques. Proc. 3rd Information Systems Intl. Conf., p. 414-422.

Srivastava, N. and Salakhutdinov, R. (2012) *Multimodal Learning with Deep Boltzmann Machines*. Proc. Advances in Neural Information Processing Systems 25 (NIPS 2012).

Stahl, F. and Jordanov, I. (2012) An Overview of the Use of Neural Networks for Data Mining Tasks. *WIREs Data Mining and Knowledge Discovery*, *2*, p. 193–208.

Tang, S., Peterson, J.C. and Pardos, Z.A. (2016) *Deep Neural Networks and How They Apply to Sequential Education Data*. 3rd ACM Conf. Learning @ Scale, p. 321-324.

Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M. and Rauterberg, M. (2015) *Advances in learning analytics and educational data mining*. 2015 European Symp. Artificial Neural Networks, Computational Intelligence and Machine Learning.

Wang, Y.-H. and Liao, H.-C. (2011) Data mining for adaptive learning in a TESL-based e-learning system. *Expert Systems with Applications 38*, p. 6480–6485

Wang. L., Sy, A., Liu, L. and Piech, C. (2017) Deep Knowledge Tracing On Programming Exercises. 4th ACM Conf. Learning @ Scale, p. 201-204.

Yu, C.H., DiGangi, S., Jannasch-Pennell, A. and Kaprolet, C. (2010) A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, *8*, p.307-325.