Analyzing Student Success and Mistakes in Virtual Microscope Structure Search Tasks

Benjamin Paaßen*
Insitute of Infomatics
Humboldt-University of Berlin
Berlin, Germany
benjamin.paassen@hu-berlin.de

Andreas Bertsch
Educational Technology Lab
German Research Center for Artificial Intelligence
Berlin, Germany

Katharina Langer-Fischer
Institute of Molecular and Cellular Anatomy
University of Ulm
Ulm, Germany

Sylvio Rüdian
Humboldt-University of Berlin & Weizenbaum Institute
Berlin, Germany

Xia Wang
Educational Technology Lab
German Research Center for Artificial Intelligence
Berlin, Germany

Rupali Sinha
Educational Technology Lab
German Research Center for Artificial Intelligence
Berlin, Germany

Jakub Kuzilek
Insitute of Infomatics
Humboldt-University of Berlin
Berlin, Germany

Stefan Britsch†
Institute of Molecular and Cellular Anatomy
University of Ulm
Ulm, Germany

Niels Pinkwart†
Insitute of Infomatics
Humboldt-University of Berlin
Berlin, Germany

ABSTRACT

Many modern anatomy curricula teach histology using virtual microscopes, where students inspect tissue slices in a computer program (e.g. a web browser). However, the educational data mining (EDM) potential of these virtual microscopes remains under-utilized. In this paper, we use EDM techniques to investigate three research questions on a virtual microscope dataset of $N = 1,460$ students. First, which factors predict the success of students locating structures in a virtual microscope? We answer this question with a generalized item response theory model (with 77% test accuracy and 0.82 test AUC in 10-fold cross-validation) and find that task difficulty is the most predictive parameter, whereas student ability is less predictive, prior success on the same task and exposure to an explanatory slide are moderately predictive, and task duration as well as prior mistakes are not predictive. Second, what are typical locations of student mistakes? And third, what are possible misconceptions explaining these locations? A clustering analysis revealed that student mistakes for a difficult task are mostly located in plausible positions ('near misses') whereas mistakes in an easy task are more indicative of deeper misconceptions.

*Corresponding author
†Shared senior authors

Benjamin Paaßen, Andreas Bertsch, Katharina Langer-Fischer, Sylvio Rüdian, Xia Wang, Rupali Sinha, Jakub Kuzilek, Stefan Britsch and Niels Pinkwart "Analyzing Student Success and Mistakes in Virtual Microscope Structure Search Tasks". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 559-565. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

Keywords

anatomy education, clustering, item response theory, performance modeling, virtual microscopes

1. INTRODUCTION

Histology is a core subject that all medicine students have to pass in their studies. An important part of classic histology training is the microscopy course where students examine a large number of slides of human or animal tissue with an optical microscope in order to identify cellular structures with the aim of establishing structure-function relationships [21]. In recent years, more and more virtual microscopes (VMs) have been developed and integrated into teaching [21]. Such VMs reduce the need for resources (students only require a computer and a software), offer the opportunity to annotate slides with teacher notes, and enhance the student learning experience [21]. Prior work has provided numerous case studies of VMs being successfully integrated into anatomy education around the globe, e.g. [5, 6, 10, 13, 21, 22]. Moreover, several evaluation studies have shown that students using VMs perform at least as well as students using optical microscopes [11, 15].

To the best of our knowledge, no study to date has considered the educational data mining potential of VMs. For example, VMs enable us to record which slides students have seen, which areas on the slides they have focused on, etc. In this work, we consider the MyMi.mobile VM that is used in anatomy courses at two German universities [10]. In this VM, students can view a slide with expert annotations (exploration), and they can test their knowledge by either locating a structure in a slide (structure search; refer to Figure 1), or identifying the tissue sample and staining (diagnosis).

We analyze the performance of $N = 1,460$ students in structure search tasks with respect to three research questions:

RQ1: Which features predict student success?

RQ2: What are typical locations of student mistakes?

RQ3: What are possible misconceptions explaining these locations?

To answer RQ1, we analyzed the collected learning data with a generalized item response theory [2, 12] model, which consists of a difficulty parameter for each task, an ability parameter for each student, and four weights for additional features (see section 3.2). To answer RQ2 and RQ3, we employed a Gaussian mixture model [7] on the locations of mistakes and interpreted the resulting clusters with the help of domain experts. Our results can contribute to enhanced teaching quality in VM courses as well as establish interpretable models to analyze data from such courses.

In the remainder of this paper, we cover related work, our experimental setup, the results, and a conclusion.

2. RELATED WORK

Prior work on machine learning on virtual microscope data has focused on applications outside education. For example, major prior work has been done in training convolutional neural networks to solve classification tasks on microscope images such as detecting fluorescence on images [4]. Successful applications can assist anatomy experts in predicting carcinogens in human cells [23]. Due to the high accuracy of these models [1], they are helpful in cancer diagnostics.

Related to education, prior work of virtual microscopes can be roughly distributed into two categories. First, there are case studies describing how virtual microscopes were integrated into anatomy curricula and the requirements for successful integration, e.g. [5, 6, 10, 13, 21, 22]. Second, several studies have investigated whether students with optical microscopes have higher learning gain compared to students with a virtual microscope and found that this is not the case, e.g. [11, 15].

One of our research questions in this paper is to identify factors that are related to success in locating structures in a virtual microscope. Models that predict student success are a common topic of educational data mining research [3]. For example, Dietz-Uhler et al. [8] summarized which kind of data is often used to predict students success, classified into data gathered from the Learning Management System (e.g. clicks on resources) and performance data (e.g. feedback or grades, created by the instructor or respectively the system). Other papers use demographic data and prior success to predict success rates, e.g. [16]. Prior work has shown that, depending on the knowledge domain, different features have high importance to predict students' success. For example, Ramos et al. [20] found that hits in a discussion forum have high importance to predict students success. Yukselturk et al. [24] used a correlational research design and concluded that self-regulation variables have a highly statistically significant relation to learning success using interpretable methods. To our best knowledge, there is no prior work on success prediction in virtual microscopes. We want to close this research gap.

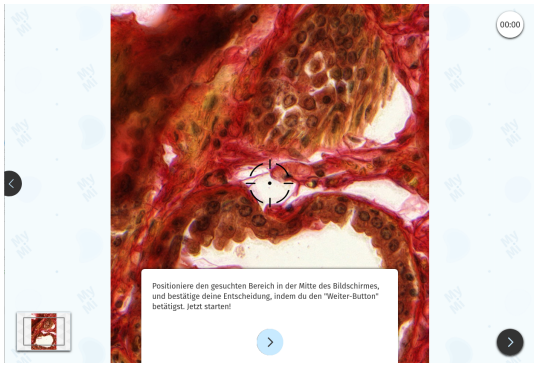

Figure 1: Screenshot of the MyMi.mobile structure search mode.

To do so, we turn to item response theory. Item response theory is concerned with modeling the probability of success of a student i at a task j via a logistic distribution over the difference between a student's ability parameter θ_i and a task's difficulty parameter b_j [2, 12]. Generalizations of this model include more parameters and other distributions [2, 12]. In this paper, we use the standard logistic distribution but include auxiliary parameters for features that capture student behavior.

To analyze the locations of typical mistakes, we perform a clustering analysis using Gaussian mixture models [7]. Clustering is a well-established technique in educational data mining [3], e.g. to identify groups of student solutions that may warrant similar feedback [9]. Our reasoning is similar: We wish to identify typical locations of mistakes in structure searches such that we have a reasonably sized set of representative locations that a teacher can inspect and for which feedback may be developed.

3. METHOD

3.1 MyMi.mobile VM and Dataset

The MyMi.mobile VM provides three modes: *exploration*, which shows expert annotations, *structure search*, where students need to locate a structure in a slide, and *diagnosis*, where students need to identify the slide and the stain. The structure search mode is shown in Figure 1. Students see a tissue slice and are supposed to move the field of view (by panning and zooming) until the crosshair is located over the correct structure. Then, they confirm their choice by clicking the arrow on the bottom right. As additional interface elements, students see an explanatory text at the bottom of the screen ("Position the area to be searched in the center of the screen and confirm your decision by pressing the 'continue-button'. Start now!"), a 'minimap' of the slide on the bottom left, and a timer on the top right. Students can select structure searches in any order from a list sorted alphabetically according to the slides (e.g. armpit, eye, colon)[1]. Students can attempt the same search as many

[1]The alphabetical ordering probably introduces an ordering bias. In particular, we observe that the two most attempted

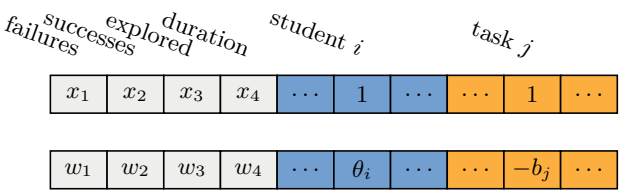

Figure 2: Illustration of the feature vector $\vec{x}$ (top) for an attempt of student i on task j, and the parameter vector $\vec{w}$ (bottom) of the item response theory model.

times as they want.

We consider a dataset of 19,525 structure search attempts by 1,460 students recorded in the summer term 2020 at two German universities. Most students were second semester undergraduate students of medicine, with some students from fourth semester dentistry (45) and molecular medicine in the second or fourth semester (39). Most students (817) did not attempt any structure search. Of the 643 who did, most attempted few structure searches (median 7) with some 'heavy users' making hundreds of attempts (mean 30.37, maximum 649). 68.19% of attempts were successful.

For the purpose of validating our model, we also asked four anatomical teachers using the VM to rate the difficulty of the 30 most attempted structure search tasks on the platform. The teachers received the following instruction: any structure search that at least 65% of students are expected to solve on their first try should be rated as 'easy'; any structure search between $40-65\%$ should be rated 'moderate'; and any structure search below 40% should be rated as 'difficult'. These boundaries were chosen based on the actual success rates of students: 10 of the tasks had an actual success rate over 65%, 10 had an actual success rate between $40\%-65\%$, and 10 had an actual success rate below 40%.

3.2 Item Response Theory

In order to investigate RQ1, we trained a generalized item response theory model implemented via logistic regression. In particular, we pre-processed each structure search attempt to be represented as a 1,859 dimensional, highly sparse feature vector (see Figure 2). The first four dimensions (gray) contain auxiliary features, namely: 1) How often has the student failed on the same structure search? (*failures*) 2) How often has the student succeeded on the same structure search? (*successes*) 3) Has the student already seen the same slide in the exploratory mode? (*explored*), and 4) How many minutes has the student spend on the current structure search? (*duration*). The next 1,460 dimensions (blue) indicate which student made the attempt, i.e. feature $x_{4+i} = 1$ if the current attempt was made by student $i \in \{1, \ldots, 1460\}$ and 0 otherwise. The remaining 395 features (orange) indicate which task the attempt was made on, i.e. feature $x_{1464+j} = 1$ if the current attempt was made on task $j \in \{1, \ldots, 395\}$ and 0 otherwise.

structure searches are both on the alphabetically first slide.

Our model, then, has the form

$$P(1|\vec{x}) = \frac{1}{1+\exp(-\vec{w}^T \cdot \vec{x})} \tag{1}$$
$$= \frac{1}{1+\exp(b_j - \theta_i - w_1 \cdot x_1 \ldots - w_4 \cdot x_4)}$$

where $\vec{x}$ is the sparse feature vector of an attempt and $\vec{w}$ is the parameter vector (see Figure 2). Note that we obtain a classic IRT model if the first four features x_1, x_2, x_3, and x_4 are 0. We used the implementation of logistic regression from the scikit-learn library [19].

3.3 Clustering

To investigate RQ2 and RQ3, we applied clustering on the locations of mistakes. More specifically, we used a Gaussian mixture model with K components, which approximates the probability density over locations (x, y) of mistakes in an image as

$$p(x,y) = \sum_{k=1}^{K} \mathcal{N}\big((x,y)\big|\vec{\mu}_k, \mathbf{\Sigma}_k\big) \cdot \pi_k, \tag{2}$$

where $\mathcal{N}((x,y)|\vec{\mu}_k, \mathbf{\Sigma}_k)$ denotes the 2D Gaussian density with mean $\vec{\mu}_k \in \mathbb{R}^2$ and covariance matrix $\mathbf{\Sigma}_k \in \mathbb{R}^{2\times 2}$; and where $\pi_k \in [0,1]$ is the prior for the kth Gaussian component. Compared to other clustering algorithms, Gaussian mixtures have at least two advantages. First, they can deal with non-spherical clusters by adjusting the covariance matrix accordingly. Second, they provide a probability density of the data. Moreover, they remain fast to train with an expectation maximization scheme [7]. We use the scikit-learn implementation of Gaussian mixtures [19]. To select the optimal number of components K, we use the Bayesian information criterion [18].

4. RESULTS AND DISCUSSION

In this section, we present the results of our experiments. We begin with the teacher difficulty ratings, then continue with the item response theory model (regarding RQ1), and conclude with the clustering analysis (regarding RQ2 and RQ3).

4.1 Teacher difficulty ratings

As the result of our teacher survey, we obtained difficulty ratings ('easy', 'moderate', or 'difficult') for the 30 most attempted structure search tasks. We observe that the teachers agreed moderately. On average, the Kendall τ for pairwise agreement is 0.4 and the overall Krippendorff's α is 0.44. To enhance reliability, we consider the average rating of each task in the subsequent analysis. On average, teachers ranked most tasks as 'easy' (about 55%), fewer as 'moderate' (just below 35%), and very few as 'difficult' (about 10%; refer to blue bars in Figure 3). Recall that, according to actual success rate, all blue bars would have height 1/3. This indicates that teachers tended to underestimate the actual difficulty, which may be an instance of the 'expert blind spot', i.e. the phenomenon that experts may fail to imagine the difficulties of novices [17]. We will use the teacher ratings as reference to further validate our item response theory model below.

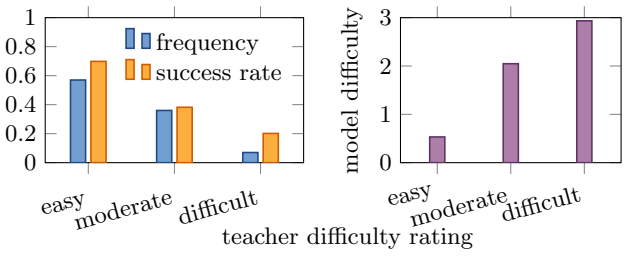

Figure 3: Left: The frequency (blue) and the mean actual success rate (orange) of tasks rated as easy, moderate, or difficult by teachers. Right: The average difficulty parameter assigned by the model to tasks rated as easy, moderate, or difficult by teachers.

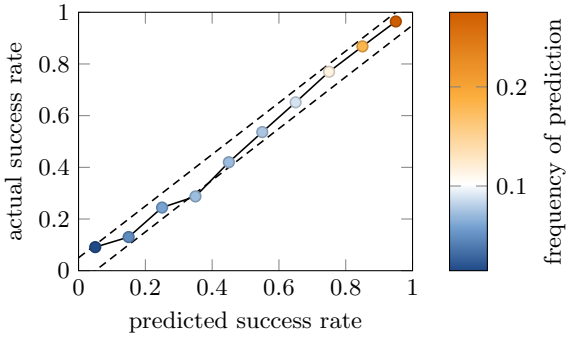

Figure 4: Calibration plot of the IRT model. Dashed lines indicate bin width. The color indicates how full each bin is.

4.2 Factors of student success

In order to investigate which factors contribute to student success (RQ1), we trained an item response theory model (refer to Section 3.2) on our data.

Model validation To validate the model, we performed three analyses. First, we performed a 10-fold cross-validation over attempts, yielding 80.19% ± 0.13% training accuracy and 77.73% ± 1.83% test accuracy on average ± standard deviation. Because our data is imbalanced (with less failures than successes), we also considered AUC (0.86 ± 0.001 in training and 0.82 ± 0.02 in test), and F1 score (0.69 ± 0.003 in training and 0.66 ± 0.024 in test with a test precision of 0.73 ± 0.06 and a test recall of 0.60 ± 0.03). All measures indicate good generalization from training to test set. For the remainder of this section, we consider a model trained on all data.

Second, we assessed model calibration. *Calibration* means that the predicted success probability of a student corresponds to the actual success rate [14]. To analyze this, we aggregated data into bins according to the predicted success probability (each bin had a width of 10%) and then computed the actual success rate within each bin. Figure 4 shows the corresponding calibration curve, where the dashed lines indicate the width of each bin in the analysis. Given that the curve remains within the dashed zone, we conclude that our model was well-calibrated. Most predictions

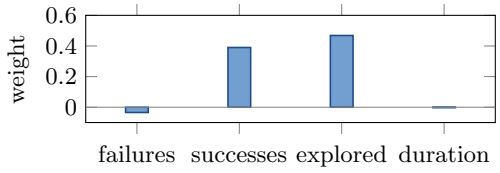

Figure 5: The scaled weights of auxiliary features.

(27.5%) were in the 90% – 100% bin (orange dot), i.e. our model predicted successful attempts with high confidence.

Third, we compared the difficulty parameters of our model with the human ratings from Section 4.1. Figure 3 (right) displays the average difficulty parameter assigned by the IRT model for each difficulty class. We observe that tasks rated as more difficult by teachers were also rated as more difficult by the model. Tasks rated as 'easy' by the teachers have a mean difficulty parameter of 0.5, tasks rated as 'moderate' have a mean difficulty parameter of 2, and tasks rated as 'difficult' a mean parameter of 3.

Overall, we note that the model is reasonably accurate, well-calibrated, and agrees with teacher ratings of difficulty.

Factors to Success Next, we inspect the weights of our model to infer which features are predictive of student success. To make the weights comparable, we normalized the auxiliary features to the same scaling as the binary features.

Regarding auxiliary features (Figure 5), we observe that the number of prior failures had a low negative weight, i.e. it is not predictive of student success. This is likely explained by the design of the MyMi.mobile VM. On a failure, students only learned that they were wrong but not where the right answer might be. This ensures that students can not get the right answer by trial and error. Attempt duration also had a low negative weight. This may be because duration is an ambiguous feature. Students may take longer both for productive reasons – e.g. inspecting the slide in more detail to validate the image against the definition of the structure – and unproductive reasons – e.g. being distracted. Accordingly, duration may not provide predictive information either way.

By contrast, we obtained positive scaled weights for the *successes* (0.39) and *explored* (0.47) features. The explanation for the former is obvious: If you have found the correct solution for the task once, chances are you memorized the location and can find it again. An explanation for the latter is that having seen an annotated example of the structure helps to find another instance of it in a structure search. That being said: We can not make causal inferences in this model. It is also possible that students who are more likely to succeed for other reasons are also more likely to consult the exploratory slides. On the other hand, we account for a general underlying student ability via the student ability parameter (Figure 6).

We observe that student ability parameters vary in the range from −1.97 to 1.55 and most parameters are clumped around

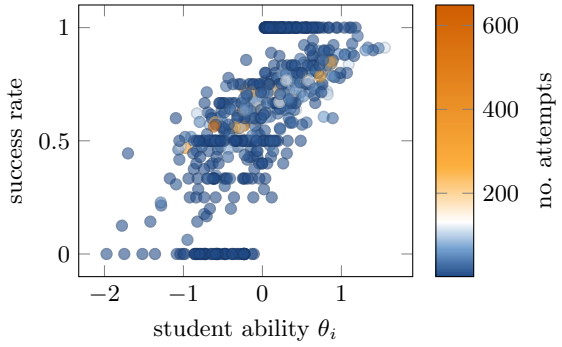

Figure 6: The success rate vs. the student ability of structure searches. Each dot represents a student. Color indicates the number of attempted structure searches by a student.

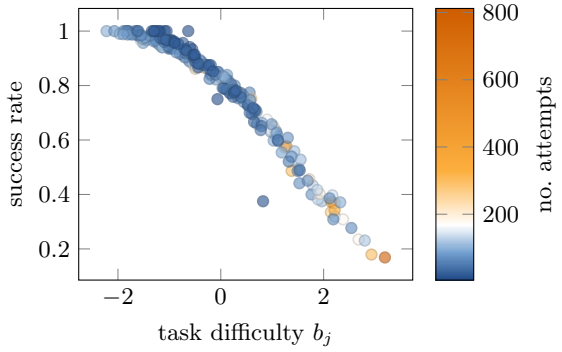

Figure 7: The success rate vs. the task difficulty of structure searches. Each dot represents a task. The heatmap colors indicate the number of attempts of a given task.

0 ± 0.54 (Figure 6). We also observe that the correlation between the ability parameter and actual success rate is relatively weak (Kendall $\tau = 0.52$). To further investigate the role of student ability, we performed another 10-fold cross validation over students instead of attempts, i.e. we tried to generalize to students that the model had never seen before and who thus had an ability parameter of 0. In this setting, we still obtained an average training accuracy of $80.19\% \pm 0.13\%$ and an average test accuracy of $77.93\% \pm 1.83\%$, indicating that the student ability parameters contributed little to an accurate prediction. We have two possible explanations for this finding: First, it may just be that the underlying 'true' student ability is relatively uniform because almost all students were in the same semester at the same two universities. Second, student ability may change during usage of the microscope, such that a single parameter may not be able to capture student ability particularly well.

Finally, we find that the task difficulty had the clearest relation to success compared to the other features. As shown in Figure 7, parameters range from -2.22 to 3.19 (mean 0 ± 1.19) and anti-correlate very well with the actual success rate (Kendall $\tau = -0.91$). This indicates that tasks had a roughly consistent difficulty across students. It also explains how our IRT model generalized well to new students.

In summary, we observe that prior success on the same task, having seen the corresponding exploratory slide, and task difficulty were most predictive of student success, whereas student ability was only moderately predictive and prior failures as well as duration were not predictive.

4.3 Typical mistakes

To investigate RQ2 and RQ3, we consider the two most attempted structure search tasks, namely searching for the nucleus of a myoepithelial cell and searching for an apocrine gland in human armpit tissue (refer to Figure 8 left and right, respectively).

The myoepithelial cell search (Figure 8, left) was the hardest task in the whole dataset with only 16.87% correct guesses (shown as green dots), with a difficulty parameter of 3.19, and unanimous consent of all four experts that it is difficult. Figure 8 (left) illustrates why the task is difficult: The correct regions (in green) are small and hard to spot.

By contrast, the slide for the apocrine gland task (Figure 8, right) exhibits many and large correct regions. Accordingly, 57.72% of guesses were correct (green dots), the model assigned a lower difficulty rating (1.28), and all experts agreed that this task is easy.

To identify typical mistakes, we trained a 10-component[2] Gaussian mixture model to cluster all the mistake locations (shown as blue dots). The cluster means are plotted as orange shapes in Figure 8. Interestingly, most clusters for the myoepithelial cell search task, namely the orange squares in Figure 8 (left) could plausibly be cell cores of myoepithelial cells. The bottom-most orange diamond is also located near a correct region. Only the remaining orange diamonds are clearly wrong because they are not located at cell cores. Generally, many students seemed to have a correct understanding of the structure to be found but failed to spot unambiguously correct locations.

By contrast, the cluster means for the apocrine gland search (Figure 8, right) indicate deeper misconceptions. All cluster centers are clearly wrong. More specifically, the diamond in the bottom right corresponds to an eccrine instead of andocrine gland, and the center diamond corresponds to a broken structure.

In both tasks, we can use cluster centers as a tool to find typical misconceptions that need to be discussed in class.

5. CONCLUSION

In this paper, we investigated three research questions regarding structure search tasks in virtual microscopes, namely 1) Which features predict student success? 2) What are typical locations of student mistakes? 3) What are underlying misconceptions explaining these locations?

[2]We observed that only little improvement in Bayesian information criterion could be achieved for more than 10 components. We also observed that 10 components were sufficient such that some components ended up unused in Figure 8. For other slides, different numbers may be needed.

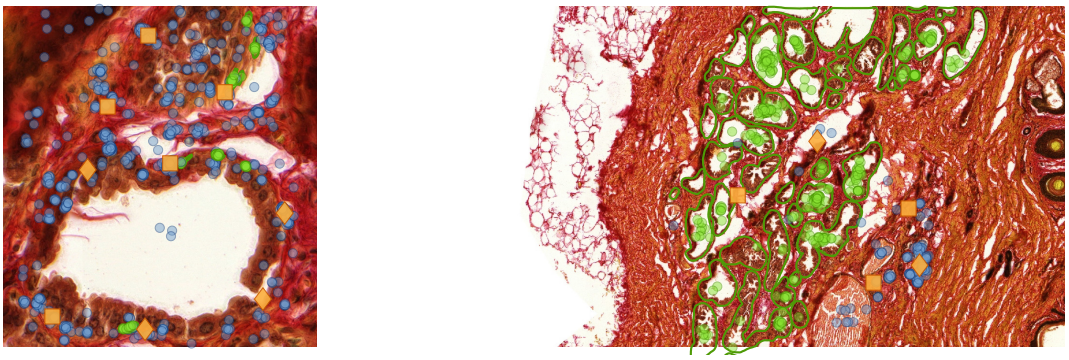

Figure 8: Students' correct (green) and wrong (blue) guesses on structure searches for myoepithelial cell cores (left) and apocrine glands (right). Correct structures are outlined in green. The centers of mistake clusters are orange shapes.

To answer the first question, we trained a generalized item response theory (IRT) model, obtaining 77% accuracy and 0.82 AUC in 10-fold cross-validation as well as solid calibration. Of the features considered, we found that task difficulty was particularly predictive of student success and the obtained difficulty parameters aligned well with actual student success rates and expert ratings. We observed less predictive value of student ability, illustrated by the fact that IRT models could generalize without loss of accuracy to new students. Moreover, prior success on the same task and having seen an annotated version of the same histological slide were predictive of success, whereas prior failures and duration spent on the task were not. This is interesting because it suggests that time stamps could be removed from the data, enhancing the privacy of the system.

Regarding the second and third research question, we applied clustering on mistake locations and interpreted the cluster centers in terms of misconceptions that may have led students to wrongly click at these locations. Such misconceptions can then be discussed in class to improve students' learning, or can be used to provide adaptive feedback in the virtual microscope tool.

Overall, this work represents the first step towards educational data mining on virtual microscope data with results that can be used to improve virtual microscope education, e.g. by ordering structure searches according to difficulty, by discussing typical misconceptions in class, and by enhancing annotations. Further work remains to be done, though. In particular, more features should be included to both enhance accuracy and find educational interventions that support student performance (like the exploratory view). Further, one could include relations between tasks in the model, thus identifying tasks that share an underlying skill, and extend the analysis to more advanced knowledge tracing methods. Finally, convolutional neural networks could be utilized to generalize teacher annotations and to identify regions of images that are easy to confuse with a structure to be searched.

6. ACKNOWLEDGMENTS

BP has been supported by the German Research Foundation (DFG) under grant number PA 3460/2-1, SR has been supported by the German Federal Ministry of Research (BMBF) under grant number 16DII127, and SB has been supported by grants from the Ministerium für Wissenschaft, Forschung und Kunst (MWK) Baden-Wuerttemberg, and by the Medical Faculty of the University of Ulm.

7. REFERENCES

[1] I. Anagnostopoulos and I. Maglogiannis. Neural network-based diagnostic and prognostic estimations in breast cancer microscopic instances. *Medical and Biological Engineering and Computing*, 44:773–784, 2006.

[2] F. Baker. *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD, USA, 2001.

[3] R. S. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, 2009.

[4] R. Brent and L. Boucheron. Deep learning to predict microscope images. *Nature methods*, 15:868–870, 2018.

[5] L. David, I. Martins, M. Ismail, ..., and C. Carrilho. Interactive digital microscopy at the center for a cross-continent undergraduate pathology course in Mozambique. *Journal of Pathology Informatics*, 9(1):42, 2018.

[6] F. R. Dee. Virtual microscopy in pathology education. *Human Pathology*, 40(8):1112 – 1121, 2009.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[8] B. Dietz-Uhler and J. E. Hurn. Using learning analytics to predict (and improve) student success: A faculty perspective. *Journal of interactive online learning*, 12(1):17–26, 2013.

[9] S. Gross, B. Mokbel, B. Paaßen, B. Hammer, and N. Pinkwart. Example-based feedback provision using structured solution spaces. *International Journal of Learning Technology*, 9(3):248–280, 2014.

[10] K. Langer-Fischer, D. Brandt, C. Braun, ..., and S. Britsch. MyMi.mobile - adaptives individualisiertes lernen in der mikroskopischen anatomie. In *Joint Annual Conference of the Medical Education Society (GMA), AKWLZ, and CAL*, 2019. German.

[11] B.-C. Lee, S.-T. Hsieh, Y.-L. Chang, . . ., and S.-C. Chang. A web-based virtual microscopy platform for improving academic performance in histology and pathology laboratory courses: A pilot study. *Anatomical Sciences Education*, 13(6):743–758, 2020.

[12] G. J. Mellenbergh. Generalized linear item response theory. *Psychological Bulletin*, 115(2):300–307, 1994.

[13] M. Merk, R. Knuechel, and A. Perez-Bouza. Web-based virtual microscopy at the rwth aachen university: Didactic concept, methods and analysis of acceptance by the students. *Annals of Anatomy - Anatomischer Anzeiger*, 192(6):383 – 387, 2010.

[14] M. E. Miller, S. L. Hui, and W. M. Tierney. Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8):1213–1226, 1991.

[15] S. Mione, M. Valcke, and M. Cornelissen. Evaluation of virtual microscopy in medical histology teaching. *Anatomical Sciences Education*, 6(5):307–315, 2013.

[16] B. Naik and S. Ragothaman. Using neural networks to predict mba student success. *College Student Journal*, 38(1):143–150, 2004.

[17] M. J. Nathan and A. Petrosino. Expert blind spot among preservice teachers. *American Educational Research Journal*, 40(4):905–928, 2003.

[18] A. A. Neath and J. E. Cavanaugh. The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics*, 4(2):199–203, 2012.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, . . ., and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] C. Ramos and E. Yudko. "hits" (not "discussion posts") predict student success in online courses: A double cross-validation study. *Computers & Education*, 50(4):1174–1182, 2008.

[21] C. Schmidt, M. Reinehr, O. Leucht, N. Behrendt, S. Geiler, and S. Britsch. MyMiCROscope—intelligent virtual microscopy in a blended learning model at Ulm university. *Annals of Anatomy - Anatomischer Anzeiger*, 193(5):395–402, 2011.

[22] M. M. Triola and W. J. Holloway. Enhanced virtual microscopy for collaborative education. *BMC medical education*, 11(1):4, 2011.

[23] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7:12474, 2016.

[24] E. Yukselturk and S. Bulut. Predictors for student success in an online course. *Journal of Educational Technology & Society*, 10(2):71–83, 2007.