Proceedings of Machine Learning Research vol 257:29–38, 2024 AI for Education Workshop at AAAI 2024

Exploring the Relationship Between Feature Attribution Methods and Model Performance

Priscylla Silva PRISCYLLA.SILVA@USP.BR
Universidade de São Paulo

Claudio Silva CSILVA@NYU.EDU
New York University

Luis Gustavo Nonato GNONATO@ICMC.USP.BR
Universidade de São Paulo

Editor: Editor's name

Abstract

Machine learning and deep learning models are pivotal in educational contexts, particularly in predicting student success. Despite their widespread application, a significant gap persists in comprehending the factors influencing these models' predictions, especially in explainability within education. This work addresses this gap by employing nine distinct explanation methods and conducting a comprehensive analysis to explore the correlation between the agreement among these methods in generating explanations and the predictive model's performance. Applying Spearman's correlation, our findings reveal a very strong correlation between the model's performance and the agreement level observed among the explanation methods.

Keywords: Explainable Artificial Intelligence, Educational Predictions, Student Success, Explanation Methods, Model Performance, Feature Importance, Correlation Analysis

1. Introduction

Extensive research has been conducted on applying machine and deep learning methods in education. These methods encompass a wide range of automated processes, from grading assignments to generating tailored feedback (Süzen et al., 2020; Bernius et al., 2022). Predicting student success and early course dropout is particularly crucial (Alhothali et al., 2022). Models that address these issues try to identify which students are at higher risk of failing or dropping out (Realinho et al., 2022; Niyogisubizo et al., 2022; Ahmed A. Mubarak and Zhang, 2022). By utilizing these models, educators can proactively intervene and provide tailored support to help students succeed in their coursework. These models analyze student information, such as academic records, engagement, and demographic data, to find patterns predicting future academic outcomes.

Considerable effort has gone into refining the accuracy of predictive models. However, there remains a knowledge gap regarding the inner workings of these models. It is insufficient to identify a potential student failure only; it is necessary to identify the factors analyzed by the model to generate the predictions. Hasib et al. (2022) present a predictive model for student success in secondary education using various classification algorithms; the study emphasizes the importance of interpretability and transparency in model predictions, employing LIME (Local Interpretable Model-agnostic Explanations) to enhance

understanding of the model predictions. Baranyi et al. (2020) conducted a study that aimed to predict the risk of college students dropping out at the Budapest University of Technology and Economics. They employed advanced machine learning models, including deep neural networks and gradient-boosted trees, and focused on interpreting the models by using two techniques - permutation importance and SHAP values. The study sheds light on the importance of model interpretation in predicting student dropout risk.

Predicting student success is challenging, and many models used for this purpose are difficult to interpret because of their black-box nature. This lack of transparency makes it hard to understand how decisions are made and what factors contribute to making predictions. As a result, it is not easy to gain meaningful insights into the factors that impact student success. However, the explainable machine learning community has made significant progress in developing different methods to elucidate the inner workings of models. Some of these methods focus on local explanation techniques that delve into the intricacies of model predictions at an individual instance level. One prominent avenue within local explanation methods involves elucidating feature importance. By employing these techniques, practitioners can gain insights into the importance of each input feature in influencing model predictions. Significant efforts have been made to use explanation methods to understand how a model predicts student success. However, according to Swamy et al. (2022), there is a considerable gap in the literature when it comes to explaining the results in the field of education.

Krishna et al. (2022) have highlighted a significant concern associated with feature attribution explanation methods known as the *disagreement problem*. This issue arises from the notable disparity in identifying the most important features among various explanation methods. The critical nature of the *disagreement problem* becomes evident when considering the implications: if distinct methods yield divergent explanations, the question of trustworthiness arises. In the context of education, specifically within student success prediction, Swamy et al. (2022) present compelling evidence that different explanation methods applied to the same model and course yield markedly distinct feature importance distributions. This underscores the gravity of the *disagreement problem* in educational scenarios, raising crucial questions about the reliability and consistency of explanatory insights derived from these methods. The *disagreement problem* remains unresolved in the existing literature. Our primary goal is to address the following research question: Is there a correlation between the model's performance and the disagreement level observed among explanation methods? To achieve this, we used nine popular instance-based explanation techniques to predict student success in two distinct real-world datasets.

2. Methodology

In this section, we will define the task of predicting student success, the *disagreement problem* that arises when using different explanation methods, and the metrics used to measure the (dis)agreement level between these methods. We will then introduce this study's datasets, model training, and explanation methods. Finally, we will describe our experiment setup in detail.

2.1. Problem Formulation

In this study, we are considering the student success prediction as a binary classification task. Let X be the feature space, representing the input features of a student. The feature vector for a particular student is denoted as $x \in X$. Let Y be the label space, where $y \in \{0, 1\}$ represents the binary outcome of student success. Here, $y = 1$ may signify success, while $y = 0$ denotes otherwise. A binary classification model is a function $f : X \to [0, 1]$ that assigns a probability to each instance, indicating the likelihood of success.

A local attribution method for the model f is a mapping $g : (f, X) \to E$ that, based on f, takes instances from X to the explanation space E, where $g(x) = (e_1, \ldots, e_K)$ is a point in E, K denotes the number of features, and e_i are the importance of each feature as to f. Consider two distinct local attribution methods, g_1 and g_2. For a given instance x, let $g_1(x)$ and $g_2(x)$ be the explanations generated by g_1 and g_2 in x, respectively. The disagreement problem occurs when $g_1(x) \neq g_2(x)$.

Krishna et al. (2022) introduced a set of metrics to measure the (dis)agreement between two local attribution explanations. The metrics evaluate (dis)agreement in the top-k most important features identified by two explanation methods. Our focus in this study is on the metrics, namely, Feature Agreement (FA), Sign Agreement (SA), Rank Agreement (RA), and Signed Rank Agreement (SRA).

Feature Agreement (FA) determines the proportion of common features between the sets of top-k features in two explanations. Sign Agreement (SA) assesses the proportion of common features with the same sign among the top-k features of two explanations. The positive and negative signs indicate the effect of a feature on the model's prediction. A positive attribution score means a feature contributes positively, while a negative score indicates the opposite. Rank Agreement (RA) calculates the fraction of common features in the same position of the rank of importance among the top-k features of two explanations. Signed Rank Agreement (SRA) combines the previous methods, incorporating both rank and sign. All the metrics listed above are in the interval [0, 1], with zero indicating complete disagreement and one representing total agreement[1].

2.2. Experimental Setup

In our experiment, we utilized two datasets. The first dataset was provided by Amrieh et al. (2015) and consisted of 480 students and 16 predictive features collected from a Kalboard 360 e-learning system. The target of this dataset is a multiclass label that classified student grades into low, medium, and high categories. After the data preprocessing step[2], we were left with 12 features. Since we were working with binary classification, we only used students classified in the low and high categories, where high represented the positive class, leaving us with a dataset of 269 students.

The second dataset was collected from a group of 132 computer science and computer engineering students taking an Introduction to Programming course during their first semester at a university in Brazil. It consists of 16 predictive features, and the target label is binary, indicating whether the student passed or failed the course.

1. Additional information on the metrics can be found in appendix A of Silva et al. (2024).
2. See appendix C of Silva et al. (2024) for more details on preprocessing.

We trained Neural Network models for each dataset. The model for the Amrieh et al. (2015) dataset consists of two hidden layers, with 16 and 8 neurons, respectively. On the other hand, the model for the Introduction to Programming course dataset includes two hidden layers with 32 and 16 neurons, respectively. In the experiment, we used 70% of the data for training, 15% for validation, and 15% for testing. Throughout the training, we systematically saved the models from intermediate epochs, creating a series of snapshots that captured the evolving state of the neural network.

We employed nine state-of-the-art feature attribution techniques to explain the predictions made by the models for the data in the testing set. These methods included six gradient-based methods, namely DeepLift (Shrikumar et al., 2017a), Guided Backprop (Springenberg et al., 2015), Input X Gradient (Shrikumar et al., 2017b), Integrated Gradients (Sundararajan et al., 2017), Smooth Gradient (Smilkov et al., 2017), and Vanilla Gradients (Simonyan et al., 2014), along with three other techniques named LIME (Ribeiro et al., 2016), Occlusion (Zeiler and Fergus, 2014), and KernelShap (Lundberg and Lee, 2017).

2.2.1. Model Performance in Intermediate Epochs

Using the test data, we computed the Area Under the Receiver Operating Characteristic Curve (AUC) metric for each model from the intermediate epochs. The AUC metric is a valuable measure for binary classification models, quantifying the model's ability to distinguish between positive and negative instances across different probability thresholds. Calculating the AUC at each intermediate epoch, we obtained a dynamic model performance profile throughout training.

2.2.2. (Dis)agreement measurement

We used the saved models to predict the test set data at each intermediate epoch. Then, we applied the selected explanation methods to generate explanations for each individual prediction. These methods generated importance scores for the features of each instance in the test set, offering a detailed understanding of the contribution of each input feature to the model's decision-making process.

We employed established (dis)agreement metrics (FA, SA, RA, and SRA) to systematically quantify the (dis)agreement level between the explanation methods. These metrics operate on a per-instance basis, so we calculated the average (dis)agreement across all instances in the test set, providing a comprehensive assessment of the overall agreement among the selected explanation methods. This process was repeated for every model stemming from the intermediate epochs, enabling us to discern patterns in the evolution of explanation methods disagreements throughout the neural network's training. By averaging the (dis)agreement scores for all instances, we obtained a robust measure of the consensus or divergence among the explanation methods[3].

3. In Silva et al. (2024), appendix E provides an example of the disagreement between the pairs of methods for the two datasets used in our study. Additionally, appendix D provides an example of how the distribution of the (dis)agreement scores can vary.

2.2.3. Correlation Analysis

In order to evaluate the (dis)agreement metrics accurately, it is necessary to vary the value of k between 1 and the total number of features present in each dataset. This is because the size of the top-k features significantly impacts the (dis)agreement metrics.

We compute Spearman's rank correlation to explore the relationship between model performance, as measured by the AUC metric, and the (dis)agreement level. In Figure 1, the results are presented for the Introduction to Programming course dataset, with columns representing the four metrics used to measure (dis)agreement and lines showcasing the variation in k used in the top-k. Each dot on the charts corresponds to models from the intermediate epochs.

3. Results and Discussions

In this section, we analyze the relationship between model performance and (dis)agreement level among the employed explanation methods by examining the Spearman correlation results. A higher AUC value indicates better model performance, while a higher disagreement metric value indicates stronger consensus among methods.

Our findings are visually represented in Figures 1 and 2, with the x-axis denoting the (dis)agreement level and the y-axis representing the model's performance. The charts are organized as follows: columns present charts for each (dis)agreement metric, while rows showcase the variation in the k values used for calculating the metric.

Figure 1 shows a sample of the results for models trained on the dataset of the n Introduction to Programming course. In the figure, k ranges from 1 to 3[4]. In 87.5% of cases for this dataset, the Spearman correlation values surpassed 0.8, indicating a robust and consistent correlation. Correlations below 0.8 were primarily observed for the FA metric as the k value increased, approaching the total number of features in the dataset. The unique behavior of the FA metric explains this phenomenon. Specifically, when k equals the number of features, the FA metric results in 100% agreement between explanations. This characteristic arises from the metric considering the intersection between sets of top-k features from two explanation methods. When k aligns with the total number of features, the two sets become identical, yielding unanimous agreement between explanations.

Figure 2 presents a sample of the study outcomes on Amrieh et al. (2015)'s dataset, where k ranges from 10 to 12. Our analysis revealed that for this dataset, in 79% of cases, the Spearman correlation score surpassed 0.8, indicating a very strong positive correlation between AUC and the level of agreement. In 16% of cases, the Spearman correlation fell between 0.5 and 0.8, signifying a strong correlation. As we previously discussed, the correlation dropped below 0.5 in certain cases related to the FA metric, which was consistent with the behavior observed before. It has been noted before that when k equals the number of features, the (dis)agreement level is always 1, regardless of the model's performance (as shown in the chart in the first column and third line). This highlights the behavior of the FA metric[5].

4. For the complete figure with k ranging from 1 to the total number of features, please refer to appendix F in Silva et al. (2024)
5. Appendix F of Silva et al. (2024) contains the complete figure.

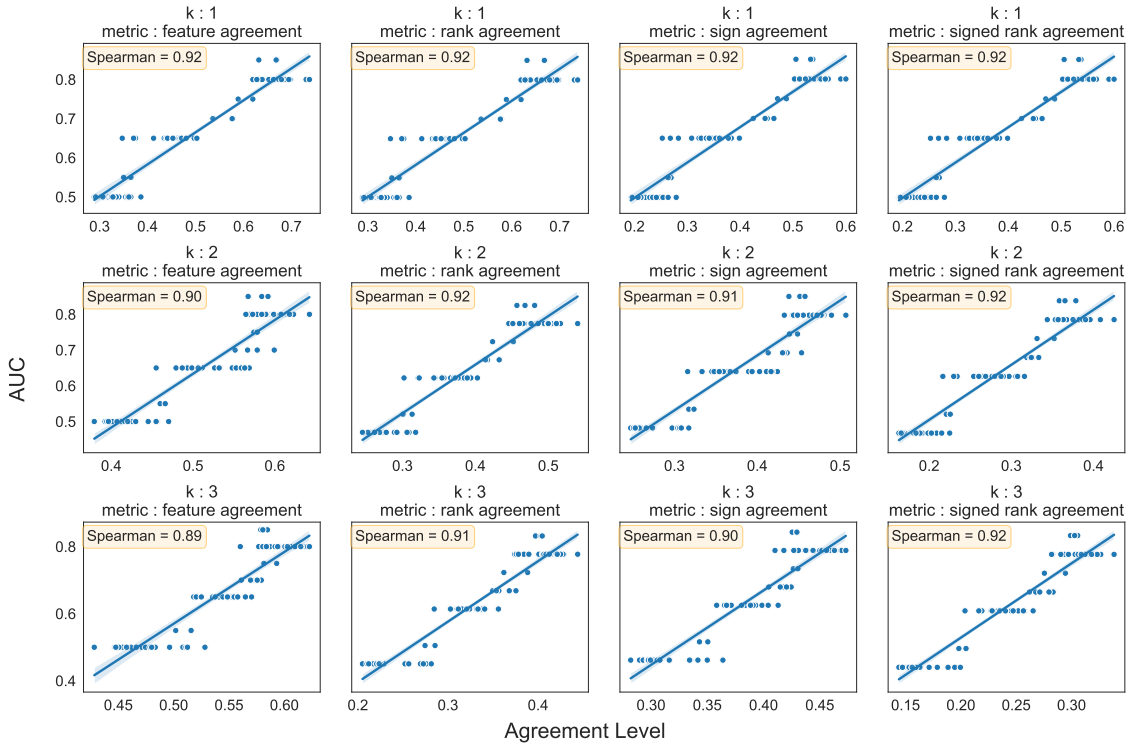

Figure 1: Correlation between Model Performance (AUC) and (Dis)agreement Metrics for Models Trained on the Introductory Programming Course Dataset.

4. Discussion and Conclusion

In the results section, we showed that for both datasets analyzed in the student success prediction task, we were able to observe that there is a strong correlation between the model's performance, measured using AUC, and the (dis)agreement level between the methods, measured using the FA, SA, RA, and SRA metrics. The strong correlation we identified implies that the agreement among explanation methods becomes more evident as the model's performance improves. A higher-performing model tends to yield explanations that exhibit more substantial consensus across various explanation techniques. This finding underscores the intrinsic connection between model quality and the interpretability of its predictions.

Our results have significant implications for practitioners and experts using explanation methods. Notably, we advocate for thoughtful consideration of the model's performance before employing any explanation method. Figures 1 and 2 depict this relationship, illustrating that models with an AUC greater than or equal to 0.8 consistently exhibit the highest levels of agreement among explanation methods. In conclusion, our study emphasizes the intertwined nature of model performance and explainability, reinforcing the importance of a robust model before delving into the realm of explanation methods. By prioritizing mod-

els with AUC values above the 0.8 threshold, practitioners can enhance the reliability and coherence of explanations generated by various methods.

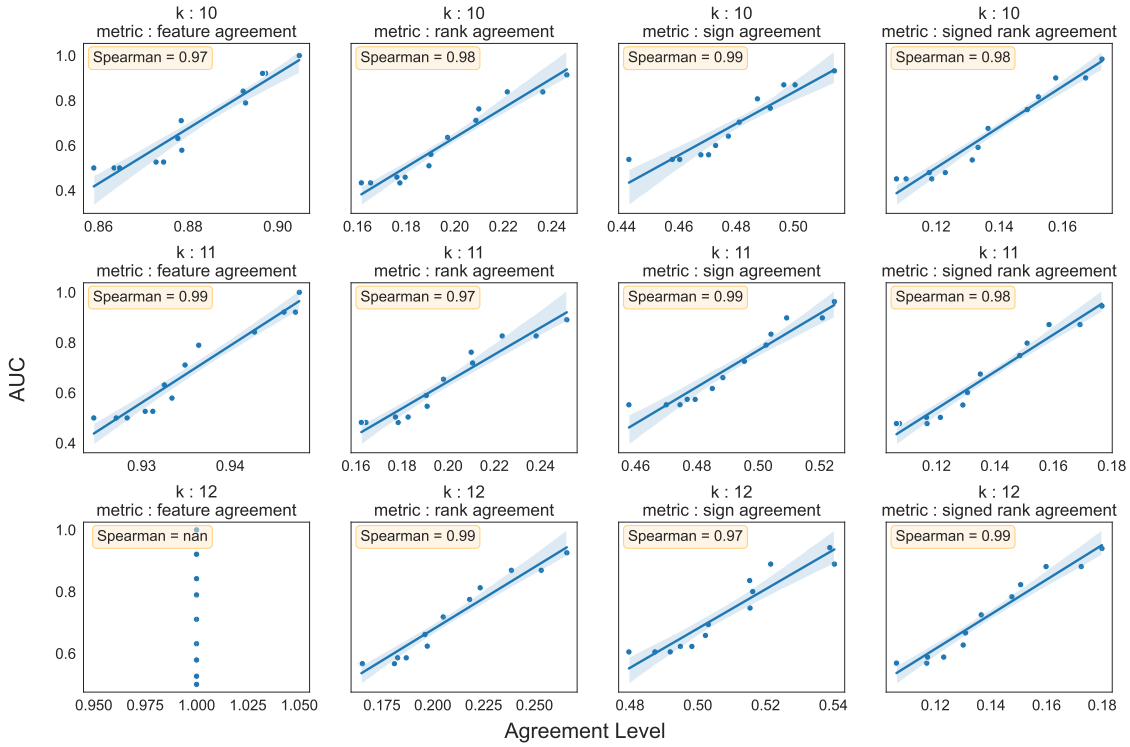

Figure 2: Correlation between Model Performance (AUC) and (Dis)agreement Metrics for Models Trained on Amrieh et al. (2015)'s Dataset.

Acknowledgments

This work was supported by FAPESP grants 2022/09091-8, 2023/05783-5, 2022/03941-0 and CNPq grant 307184/2021-8. The opinions, hypotheses, and conclusions or recommendations expressed in this material are those of responsibility of the author(s) and do not necessarily reflect FAPESP's view.

References

Han Cao Ahmed A. Mubarak and Weizhen Zhang. Prediction of students' early dropout based on their interaction logs in online learning environment. *Interactive Learning Environments*, 30(8):1414–1433, 2022. doi: 10.1080/10494820.2020.1727529.

Areej Alhothali, Maram Albsisi, Hussein Assalahi, and Tahani Aldosemani. Predicting Student Outcomes in Online Courses Using Machine Learning Techniques: A Review.

Sustainability, 14(10), 2022. ISSN 2071-1050. doi: 10.3390/su14106199. URL https://www.mdpi.com/2071-1050/14/10/6199.

Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. Preprocessing and analyzing educational data set using X-API for improving student's performance. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5, 2015. doi: 10.1109/AEECT.2015.7360581.

Máté Baranyi, Marcell Nagy, and Roland Molontay. Interpretable Deep Learning for University Dropout Prediction. In *Proceedings of the 21st Annual Conference on Information Technology Education*, SIGITE '20, page 13–19, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370455. doi: 10.1145/3368308.3415382. URL https://doi.org/10.1145/3368308.3415382.

Jan Philip Bernius, Stephan Krusche, and Bernd Bruegge. Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, 3:100081, 2022. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2022.100081. URL https://www.sciencedirect.com/science/article/pii/S2666920X22000364.

Khan Md. Hasib, Farhana Rahman, Rashik Hasnat, and Md. Golam Rabiul Alam. A machine learning and explainable ai approach for predicting secondary school student performance. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0399–0405, 2022. doi: 10.1109/CCWC54503.2022.9720806.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jovial Niyogisubizo, Lyuchao Liao, Eric Nziyumva, Evariste Murwanashyaka, and Pierre Claver Nshimyumukiza. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3:100066, 2022. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2022.100066. URL https://www.sciencedirect.com/science/article/pii/S2666920X22000212.

Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins. Predicting Student Dropout and Academic Success. *Data*, 7(11), 2022. ISSN 2306-5729. doi: 10.3390/data7110146. URL https://www.mdpi.com/2306-5729/7/11/146.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page

1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017a.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences, 2017b.

Priscylla Silva, Claudio Silva, and Luis Gustavo Nonato. Exploring the relationship between feature attribution methods and model performance. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*, 2024. URL https://openreview.net/forum?id=HkJwIypfus.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6806.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

Vinitra Swamy, Bahar Radmehr, Natasa Krco, Mirko Marras, and Tanja Käser. Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In Antonija Mitrovic and Nigel Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 98–109, Durham, United Kingdom, July 2022. International Educational Data Mining Society. ISBN 978-1-7336736-3-1. doi: 10.5281/zenodo.6852964.

Neslihan Süzen, Alexander N. Gorban, Jeremy Levesley, and Evgeny M. Mirkes. Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, 169:726–743, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.02.171. URL https://www.sciencedirect.com/science/article/pii/S1877050920302945. Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society), held August 15-19, 2019 in Seattle, Washington, USA.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.