IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Explainable Student Agency Analytics

MIRKA SAARELA[1], VILLE HEILALA[2], PÄIVIKKI JÄÄSKELÄ[2], ANNE RANTAKAULIO[3], and TOMMI KÄRKKÄINEN.[1]

[1]Faculty of Information Technology, P.O. Box 35, FI-40014, University of Jyvaskyla, Finland
[2]Finnish Institute for Educational Research, University of Jyväskylä, Jyväskylä, Finland
[3]Jyväskylä University of Applied Sciences, Jyväskylä, Finland

Corresponding author: Mirka Saarela (e-mail: mirka.saarela@ jyu.fi).

This research was supported by the Academy of Finland (grant no. 311877) and is related to the thematic research area DEMO (Decision Analytics Utilizing Causal Models and Multiobjective Optimization, jyu.fi/demo) of the University of Jyväskylä, Finland.

ABSTRACT Several studies have shown that complex nonlinear learning analytics (LA) techniques outperform the traditional ones. However, the actual integration of these techniques in automatic LA systems remains rare because they are generally presumed to be opaque. At the same time, the current reviews on LA in higher education point out that LA should be more grounded to the learning science with actual linkage to teachers and pedagogical planning. In this study, we aim to address these two challenges. First, we discuss different techniques that open up the decision-making process of complex techniques and how they can be integrated in LA tools. More precisely, we present various global and local explainable techniques with an example of an automatic LA process that provides information about different resources that can support student agency in higher education institutes. Second, we exemplify these techniques and the LA process through recently collected student agency data in four courses of the same content taught by four different teachers. Altogether, we demonstrate how this process—which we call explainable student agency analytics—can contribute to teachers' pedagogical planning through the LA cycle.

INDEX TERMS Explainable Artificial Intelligence, Decision Making, Higher Education, Student Agency

I. INTRODUCTION

THE global COVID-19 and the related closures of educational institutions showed how significant it is for students to be able to rely on their own resources. In particular, to continue learning, the educational institutions' closures placed greater demands on students' autonomy and their capacity for independent learning, executive functioning, and self-monitoring [1]. It also showed that those students who lacked the resilience and engagement to learn on their own, in particular, were at risk of falling behind [1], [2]. In summary, COVID-19 and its consequences for students revealed the importance of being self-determined in learning and being able to adapt to situations involving rapid change.

Student agency equips students to manage such situations. It refers to students' holistic judgement of how they can affect and direct their learning in instructive settings, work effectively, and utilize the assets that are accessible in the learning environment [3], [4]. The importance of agency in education has been emphasized by policy-making informers, especially by the Organisation for Economic Co-operation and Development [5]. Agency is a basic need in any goal-oriented work, particularly in jobs that call for creativity and continuous development in work practices [6]. This means that graduates of higher education institutes, in particular, should be prepared to act as developers and change agents in their field. However, despite this need—especially in the COVID-19 context but also in general—and the particular emphasis on student agency by policy-making informers, student agency has received little explicit attention in educational practice in higher education so far.

Learning analytics (LA) refers to a research field that harnesses data on learners to understand, improve, and optimize learning [7]. The use of LA can, for example, predict academic success, improve quality assurance, and identify at-risk students [8]. Moreover, dashboards are often utilized to visualize learning processes and study pathways—not only to increase awareness but also to give personalized feedback to the learners. This kind of personalized feedback and consideration of the personal traits of learners can positively influence the learning process and outcomes. Since it is usually unfeasible for teachers to manually provide such individualized feedback to all students—especially for teachers in higher education settings who often have to instruct hundreds of students with different backgrounds—such automated feedback can offer significant support.

Jääskelä et al. [9] examined student agency as the theoreti-

cal framework for assessing and enhancing digital education at universities by making use of LA. Based on a factor and robust cluster analysis process, which is conducted to measure students' responses to a validated scale [3], [9], the students receive automated feedback on their individual agency profile. In addition, the teacher of a higher education course gets an aggregated overview of the different student agency profiles. The essence of this automated agency-based process—which is called student agency analytics (SAA)—is to provide actionable information for students on their learning efforts in relation to their perceived affordances in the course and for teachers on students' judgements of their situational agency to increase pedagogical knowledge.

In a recent review, Deeva et al. [10] classified automated feedback systems by their applied educational settings, the properties of their delivered automated feedback, and their design and evaluation approaches. They concluded that applied learning theories or educational frameworks had not been reported in most cases. Moreover, they urged the developers to use more data-based solutions and to be able to explain the reasons behind the automated system. Therefore, the purpose of the present article is to show how the integration of explainable artificial intelligence (XAI) techniques with the SAA process (see Figure 1) can support the transparency and data-based development of automated feedback systems in education. More precisely, we aim to integrate XAI techniques into the SAA process in the context of higher education. This procedure improves awareness of different stakeholders from such organizations on the learning arrangements, considers the complexity of the students' capacities and various contextual resources, and supports reflection.

Another reason why we aim to integrate XAI techniques within SAA is that explainability became a key issue in LA [11]. Relationships in educational data are often complex [8], [12], and several theoretic LA studies have shown that these relationship can be modeled better by complex models than by simple linear ones (e.g., [13]–[16]). However, in practice, these complex models are rarely used because they are reckoned to be inexplicable. XAI is an emerging research direction that can help the user or developer of complex models understand the model's behavior and provide human-understandable justifications for it [17], [18]. Thus, the integration of XAI techniques allow us to also use the better performing complex LA models in SAA and to explain them in such a way that even practitioners with no background in data analysis can easily understand them.

To demonstrate our explainable SAA process (XSAA), we provide the results from a study of four concurrently implemented courses on mathematics in an engineering education degree program. The content and curriculum of these mathematics courses are identical but they are taught independently by four different teachers. This means we built and explained our models not only by using the student-specific agency data but could also link them to the particular teaching approaches of the instructors. Such a setting is new and might help teachers to increase their awareness of the effects of their pedagogical planning and interventions.

The main contributions of this paper are twofold:

- We use XAI to produce explainability and actionability through dashboards. These dashboards not only show summaries of the raw student data (e.g., how active they were with the tasks or how long it took to solve a problem) but also—through nonlinear and universal machine learning models—explain the reasons for the students' actions, linking them to a well-defined body of pedagogical planning by the teacher.
- We discuss the usability of the results gained through XSAA at the teaching practice level; that is, how they may help teachers in reflecting and designing their curriculum and in developing agency-supportive practices in their teaching implementations.

The rest of the paper is organized as follows. Section II outlines the background at the basis of our contribution. First, we locate our research among the previous studies in the field of LA and XAI in higher education. Second, we summarize previous student agency LA studies. Section III provides a discussion of the need for explainable models, especially in LA. It also provides an overview of the different XAI techniques that we are using for our SAA dashboards. Section IV presents an example of an application of our explainable SAA in higher education (i.e., the data and our XSAA results from the four groups of students studying the same mathematics course taught by different teachers at a university of applied science). Finally, Section V presents the main findings and implications of our study.

II. BACKGROUND

A. LA AND XAI STUDIES IN HIGHER EDUCATION

Hundreds of primary studies depicting and analyzing the use of LA to improve educational actions in higher education institutes (HEI) have been published, and their impacts and outcomes have been summarized in many recent reviews (e.g., [19]–[21]). Their overall conclusions suggest that LA should be better grounded to learning science, its effectiveness should be assessed, and actual linkages to teachers and pedagogical planning should be emphasized.

For example, the review by Aldowah et al. [19], which included 402 articles from 2000 to 2017, presented many student-oriented characteristics such as "engagement," "achievement," "participation," "reflection," "motivation," and "satisfaction" to be approachable by using LA techniques. However, no linkage to the actual teaching activities was presented. In the combined review-meta-review by Du et al. [22] from 901 identified research papers from 2011 to 2017, the authors mentioned that instructors need to connect LA with learning science and use dashboards for student monitoring. Similarly, the knowledge gap between the theoretical frameworks of educational domain knowledge and the LA models was emphasized in the review by Cui et al. [23].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

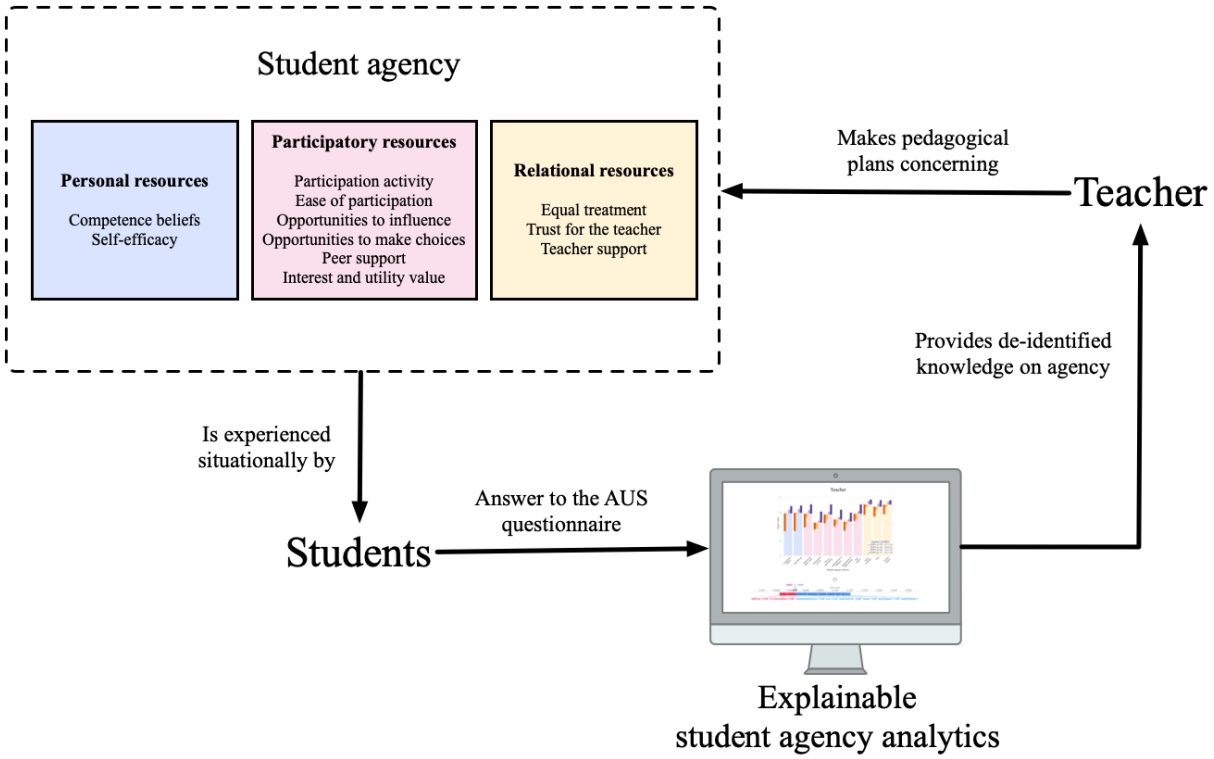

FIGURE 1. The XSAA process can be depicted as a loop, which starts when the teacher makes the initial pedagogical plans. At some point in the learning and teaching process, the students complete the AUS questionnaire, and the agency analytics is executed automatically. The teacher receives results, and can then adjust the pedagogical plans according to the students' experienced resources of agency.

After multistage screening, the review by Larrabee Sonderlund et al. [24] ended up with only 11 studies out of 689 that were found to evaluate the effectiveness of LA interventions, concluding that the lack of intervention studies where the educational institution (in practice, the instructor of a course in HEI) performs and evaluates systematic changes of its actions. Moreover, based on analyzing 252 papers published during 2012 to 2018, Viberg et al. [21] concluded that "the overall potential of LA is so far higher than the actual evidence, which poses a question of how we can facilitate the transfer of this potential into learning and teaching practice." Likewise, Ifenthaler and Yau [20] addressed the study success of HEI students through 46 primary studies, concluding that the lack of "rigorous, large-scale evidence of the effectiveness of LA in supporting study success." To this end, the review by Leitner et al. [25], which was based on 101 articles during 2011–2016, nominated teachers solely as a "side-product" of the research field.

Contrary to the huge amount of LA in HEI studies, studies dealing with XAI in HEI are extremely scarce. A Google Scholar and Scopus search in May 2021 identified only three studies of XAI in HEI [26]–[28]. Putnam and Conati [26] conducted experiments with nine university students testing whether the students would like to receive explanations for hints given in an intelligent tutoring system (ITS). They concluded that the majority of students would like explanations in the ITS, but the actual implementation of XAI was presented as future work. Likewise, Conati et al. [27] discussed only theoretically necessary considerations to make an ITS explainable for the benefit of learning. Alonso and Casalino [28] used XAI for a distance learning set. However, they did not provide any description of XAI techniques and solely used existing software (WEKA) to gather explanations for their prediction models. In sum, all three articles emphasized the need for XAI in automated feedback systems in HEIs, but none implemented and explained the underlying XAI techniques.

B. STUDENT AGENCY ANALYTICS IN A NUTSHELL

a: Student agency in higher education

Agency has been under consideration in several disciplines and has been highlighted in various areas of life. In general, agency is one's capacity to act and cause change. However, different disciplines have their own and more detailed perspective on the meaning of agency. For example, in social cognitive theory, agency is understood as an individual's capability to engage in intentional, self-defined, and meaningful action [29]. Similarly, in social sciences, the concept of agency concerns an individual's capability to take intentional and self-defined (i.e., autonomous) action and is focused on the circumstances and structural factors that constitute frames for action (e.g., [30]). Contemporary educational discourse has emphasized the meaning of agency in lifelong learning [31] and in student-centered learning [32]. Within educa-

tional sciences, agency is seen as an integral part of learning, which manifests itself both as individuals' active action in knowledge construction (e.g., [33]) and a sense of being empowered in learning situations [34].

Our stance on student agency is based on the conceptualization made by Jääskelä et al. [3], who synthesized the previous literature on agency and defined student agency in higher education as "a student's experience of having access to or being empowered to act through personal, relational, and participatory resources, which allow him/her to engage in purposeful, intentional, and meaningful action and learning in study contexts." Student agency consists of three resource areas (see Figure 2). *Personal* agency resources consist of the dimensions of competence beliefs and self-efficacy. *Relational* resources refer to power relations in different educational settings, which include the experiences of equality among the students, trust for the teacher, and support from the teacher. *Participatory* resources of student agency involve dimensions relating to engaged and active participation in learning. Altogether, student agency is composed of 11 dimensions, and it is measured using a validated psychometric Agency of University Student (AUS) scale [3], [35].

b: Student agency analytics

Discerning different study experiences can be demanding in heterogeneous educational settings with a multitude of students. To address this challenge, we apply a LA process called student agency analytics, which utilizes robust statistics and psychometric information obtained using the AUS scale [9]. First, the students in a particular study group or course complete the AUS questionnaire. Second, the individual factor values of agency are calculated for each student using the factor pattern matrix, which enables the determination of the general agency profile of the whole study group. Third, unsupervised learning, specifically robust clustering, is used to provide prototypical agency profiles with four distinct groups based on cluster validation indices, as described in more detail in [9]. Kruskal-Wallis H and Mann-Whitney U tests can then be used for explaining the clustering results through the agency dimensions. Moreover, if the information on the quality of learning outcomes or course grades is available, it can be linked to the prototypical agency profiles using supervised learning.

The main representations obtained using SAA are the students' individual agency profiles (IAPs), the general agency profile (GAP) of a group (e.g., study group, course), and four distinct prototypical agency profiles (PAPs) within a group. IAP (Figure 2) represent the values of individual student's agency dimensions, which can be compared with the GAP. IAP is a personal depiction, and it is aimed only at the student accompanied with general information about student agency. For the teacher, student agency analytics provide a general overview of the agentic resources of the students. To preserve students' privacy, teachers do not receive individual student profiles. Instead, their report consists of de-identified information about the GAP and PAPs. Both the GAP and PAPs are presented in the teacher report as a special combined bar graph (Figure 4).

c: Teacher's perspective

Teachers' actions and their pedagogical choices influence students' learning experiences (e.g., [36]–[40]). In terms of pedagogical planning, teachers would benefit from the analysis results concerning all their students. For instance, peer support can help students in higher education to develop self-regulation skills, decreasing or allowing better management of study-related exhaustion [41]. Thus, it would be worthwhile for the teacher to identify the different experiences of peer support to provide means and opportunities for students to actualized supportive collaboration. Students' prior knowledge can significantly influence student achievement [42]. Failing to consider students' prior knowledge might be manifested as a lack of competence beliefs and self-efficacy. In summary, becoming aware of students' agentic experiences could help teachers make better pedagogical plans and decisions.

From the teacher's perspective, SAA summarizes the inter-individual differences of learning experiences in a visually interpretable form. As a result, students' general assessment of their agency and four distinct student agency profiles are presented to the teacher. The process can be depicted as a loop (see Figure 1), which starts when the teacher makes the initial pedagogical plans. At some point in the learning and teaching process, the students complete the AUS questionnaire, and the agency analytics is automatically executed. The teacher receives results, which visually describe the GAP and the PAPs. The teacher can then adjust the pedagogical plans according to the students' experienced agency resources. In the following sections, we develop the SAA process toward explainable LA.

d: Ethical considerations

A general prerequisite in LA should be the responsible use of educational data [43]. It is worth emphasizing that SAA aims not to evaluate or grade the students or their learning. Instead, the purpose is to identify and make visible different personal learning experiences through the concept of agency. Thus, it is essential to ensure the privacy of the students and teachers. The individual agency profile received by a student is personal and only for the student's use. Teachers or anyone else do not see the student's IAP unless they want to disclose the results, for example, to help study counseling. Generating aggregated results (GAP and PAPs) provide a means to present detailed but de-identified information for the teacher. Similarly, the teacher report depicting the aggregate results of a course is meant only for the teacher to use in personal pedagogical planning. The results should not be used to evaluate the individual teachers or their teaching.

III. TOWARD EXPLAINABLE LEARNING ANALYTICS

From a technical point of view, LA is about modeling students and learning. Its methods have roots in several different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

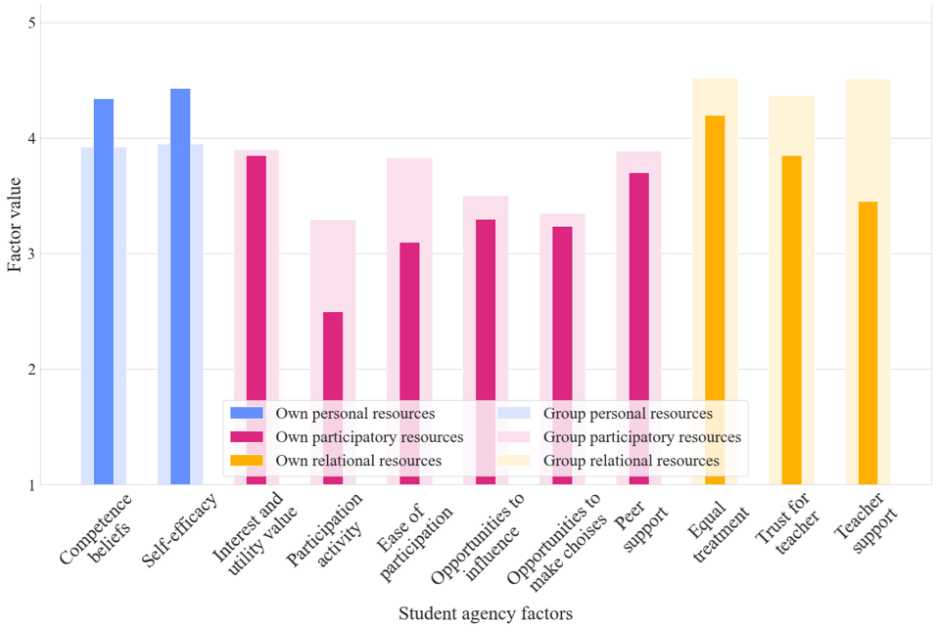

FIGURE 2. Student agency analytics provides information about the inter-individual differences relating to resources of student agency. This figure shows a student's personal report that consists of his/her individual agency profile in comparison with the general agency profile of the group. A teacher's report consists of a general agency profile of the group combined with four prototypical agency profiles (as visualized in Figure 3).

disciplines, such as statistics, education, psychology, and machine learning [44], [45]. While traditionally, statistical models were mainly used in LA to scaffold students and help teachers, the machine learning models have gained in importance in recent years [46]. This is mainly due to the challenge of modeling the increasingly rich, varied, and multimodal (such as eye tracking, physical movement, and face recognition for emotion detection) LA data [47], [48].

Often a trade-off occurs between the performance of a specific machine learning model and its explainability. For example, in supervised learning, the performance (i.e., the difference between the real outputs and the outputs of the model) is usually better for complex models with nonlinear combinations of inputs, but such models are harder or even impossible to understand. These kinds of models are also called "black boxes." On the contrary, simple linear methods are prone to perform worse, but they are easier to interpret and understand. One example of the latter is a linear regression model, where the coefficient of an input can be directly interpreted as the importance of that input.

Although they usually perform better, black boxes have several problems. One problem relates to assuring that such a model works as intended. If not even the designer of the model can explain the model's underlying logic and how it arrived at a result, it is impossible to verify that the model uses the right justifications for its decisions. In the worst case scenario, such black-box models may use questionable reasons for their decisions without anyone noticing them. This usually happens if they adopt bias in the training data. Bolukbasi et al. [49], for instance, showed that a model that was trained on a corpus of Google News text, learned the correct word embedding "man is to woman as king is to queen," but at the same time also learned the worrisome embedding "man is to woman as computer programmer is to homemaker."

Another example, discussed by Freitas [50], comes from the military: The military trained a classifier to distinguish pictures of enemy tanks from pictures of friendly tanks. This classifier was performing well on the training set but showed poor performance when it was used in the field. Later it was discovered that the pictures of enemy tanks in the training set were taken mostly on overcast days, while the pictures of other tanks were taken on fair weather days. It turned out that the classifier had learned this pattern from the training set and consequently mostly used background features to classify the tanks. Such examples prevent users from trusting a black box model. In fact, some studies have shown that even if they are proven to be more accurate than human forecasters, most people exhibit an inherent distrust of automated predictive models [51]. If the users do not trust a model or a prediction, they will not use or deploy it. Thus, the explainability of models is important, not only for developers but also for the end users, and all other parties involved.

XAI is a new research field. It refers to approaches attempting to make machine learning models more explainable and to address the above-mentioned issues. Several XAI review papers were recently published, indicating its importance and topicality [18], [52]–[56]. Generally, the explainability of a model refers to any approach that helps the user or developer understand the model behavior and its reasoning process [17]. While no definition of XAI is uniformly accepted, it can be conceptualized as the ability to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

provide human-understandable justifications explaining the way in which a model works so that observers can understand how and why it has delivered particular outcomes. For example, in the military classifier case discussed above [50], an explanation would have shown that the classifier used the background instead of the features of the tanks for classifying the photos. Thus, XAI can help to identify potential bias in the training data, ensure algorithmic fairness, and verify that the algorithms perform as intended [53].

As pointed out by Baker [11], explainability is also one of the biggest challenges in LA nowadays. Several LA studies have shown that complex models outperform the simpler ones. However, if an instructor does not understand such a complex LA model and if a development team cannot explain it, the LA model will probably never be employed in practice (ibid). Instead, only simple linear models that have been around for years continue to be used. This is a problem, because as argued for example in [12], relationships in educational data are often complex and cannot be modeled well enough with the simple models. If the better performing complex models could also be explained in such a way that even practitioners with no background in data analysis could easily understand them, they would probably be employed more often.

Conati et al. [27] argued that the explainability of models is also important for learners: For instance, if learners cannot comprehend the logic of an intelligent tutoring system, they are not motivated to follow the systems instructions and their trust in the system as a whole will decrease. Another reason the explainability of LA models has become increasingly important is that the new General Data Protection Regulation (GDPR) now includes a right to explanation and information [57], [58]. This means that if automatic profiling "(e.g., in student analytics) is used, it is not only a desiderata but actually a requirement to be able to explain to a student why he/she was assigned to a particular profile.

In general, one can distinguish XAI methods that are intrinsic, meaning interpretable due to their simple structure, and post-hoc XAI methods, meaning methods applied after model training to explain the model's logic in retrospect. Moreover, one distinguishes between local and global explanations [59], [60]. While modular global explanations provide interpretation for the model as a whole, approaching it holistically, a local explanation provides interpretation for a specific observation (such as one particular student). Finally, explanation techniques can be model specific, meaning the explanation technique is specific to its model, or model agnostic, meaning the explanation technique can be applied to any model.

In this work, we use both intrinsic model-specific and post-hoc model-agnostic explanations as well as global and local explanations. Moreover, we want to explain not only the most important characteristics of the different agency profiles (global explanations) but also explain, for specific observations, why they were assigned to a particular group (local explanations). The latter are especially interesting for instructors who receive a report about their students' agency and can then see why a particular student was assigned to a particular agency group. Finally, as pointed out above, students have a right to information about individual decisions made by agency algorithms, and the local XAI techniques enable us to provide such information.

a: Multinomial logistic regression

Logistic regression is an example of a machine learning method that because of its linear structure is intrinsically explainable and offers model-specific modular global explanations. It is probably the most traditional technique to predict a categorical response variable (i.e., the class). If the class is dichotomous, a simple logistic regression can be used that employs a logistic function to measure the relationship between the class and the explanatory variables through estimating probabilities. If the class has more than two categories, multinomial logistic regression should be used. Multinomial logistic regression uses the softmax function (i.e., a generalization of the logistic function to multiple dimensions) to calculate the probabilities of each class category over all possible class categories. These calculated probabilities are then used for determining the class (i.e., the response variable category) for the given inputs.

Logistic regression is intrinsically explainable through its coefficients. The coefficient of a continuous explanatory variable can be explained as the estimated change in the natural log of the odds for the reference event for each unit increase in the predictor [61]. In general, the larger the absolute magnitude of a coefficient is, the more relevant the corresponding explanatory variable is for the classification. Moreover, the sign of the coefficient indicates whether the explanatory variable increases or decreases the probability of belonging to a certain class. Furthermore, if the logistic regression model is penalized with the l_1 norm, some of the feature coefficients shrink to exactly zero, which makes the model simpler and easier to explain [62]. However, although (multinomial) logistic regression generally meets the characteristics of an explainable model, Arrieta et al. [63] point out that it may also demand post-hoc explainability techniques, such as visualizations, particularly if the model is to be explained to non-expert audiences.

b: Multilayer perceptron

A multilayer perceptron (MLP) is an example of a machine learning technique that is also able to find and model complex nonlinear interactions in data and, thus, often outperforms linear techniques, such as the previous discussed logistic regression. It consists of an input layer, at least one hidden layer, and an output layer. Each layer consists of nodes, and except for the input nodes, all nodes are neurons with nonlinear activation functions. MLPs are fully connected, meaning that each node in one layer connects with a certain weight w_{ij} to every node in the succeeding layer. These weights on the nodes are automatically adjusted to construct the mathematical model that most accurately maps the input

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

features (such as the agency dimensions of the students and the information in which course he/she was studying) to the output labels.

However, MLP models are generally regarded as black boxes and opaque. For example, even when techniques are used to identify the features that a particular MLP model assigned significant weights to, the relationships between those features and the classification can be weak because a small permutation in a seemingly unrelated aspect of the data can result in a significantly different weighting of features [64]. Moreover, different initial settings can result in the construction of different models [65].

c: Random forest

Random forests, as well as other tree-based techniques, are one of the most popular nonlinear supervised machine learning methods nowadays [66]. They are ensemble learners based on decision trees, which are on the one hand, explainable and able to model nonlinear relationship in data, but on the other hand, generally low performing because they tend to overfit the training data. Through growing each tree in the ensemble (i.e., the forest) only on a bootstrap sample from the original data and by randomly using only a subset of the features for each node in each tree, random forest keep the main advantages of decision trees while at the same time overcoming their disadvantage. In other words, random forest are also explainable and able to model nonlinear relationship in data, but—through the bagging of many uncorrelated decision trees—surmount the overfitting and low-performance issue of decision trees. In fact, they perform so well that they are often the winner in machine learning competitions [66], [67]. Nevertheless, although the importance of a global model-specific feature is generally provided with the random forest implementation (for example, in Python, Gini measures the global importance of the input features), less attention has been paid o far to local explanations for random forest predictions [66].

d: Local interpretable model-agnostic explanations

Local interpretable model-agnostic explanations (LIME) are a XAI tool developed by Ribeiro et al. [68]. LIME provides explanations, such as features and rules of features, that were important for predicting a specific observation (i.e., local explanations). It can be used for any prediction model, meaning it is model agnostic, because it does not even need know the actual "black box" prediction model f; it just uses its predictions. More specifically, it changes the model's inputs and then uses the model's outputs to make conclusions about the model. The main idea is that if the model prediction does significantly changes after the value of a feature is slightly adjusted, that feature may be an important predictor. Vice versa, if the prediction does not change, the changed feature may not be important at all.

It accomplishes this by taking the observation x for which the prediction should be explained and permuting its feature values. All of these permuted fake observations are weighted by their distance to x. Then, the black box model f is used to predict the permuted observations, and a new surrogate/explanation model (can be any explainable model, such as a linear model or decision tree) g is trained that reflects the original predictions as accurately as possible, while the complexity of this surrogate model is kept as low as possible. Then the explanations of the simple surrogate model (for example, the weights if g is a linear model) are used to explain the local behavior of $f(x)$.

Mathematically, this can be expressed as follows:

$$\xi = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

where π_x is the proximity measure to define locality around x, and $\Omega(g)$ is the complexity of g that should be kept low (for example, by minimizing the number of non-zero weights if g is a linear model).

The advantages of LIME are that it is relatively easy to use and understand. However, certain drawbacks are associated with it. One of these is the potential inconsistency between the surrogate model prediction $g(x)$ and the real model prediction $f(x)$. Anther drawback is the lack of comparative values for the LIME values. SHAP, which will be discussed below, overcomes these drawbacks.

e: SHapley Additive exPlanations

Shapley values, introduced by Shapley [69], originate from cooperative game theory. They measure the fair payout that each player should receive based on his/her contribution to the total payout of the game. The payout for each player is proportional to his/her marginal contribution to the total payout. Similarly, when used as an explanation for a prediction, a Shapley value measures the contribution of an individual feature to the total prediction. This means a Shapley value is the average marginal contribution of a feature value across all possible coalitions of the features.

The fair contribution of feature i is obtained by taking the average of the contribution over the possible different permutations in which the coalition can be formed. Mathematically, this can be expressed as follows:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S)),$$

where N is the number of all features, S a subset of the N features, and $v(S)$ the prediction of the S features. When feature i joins the S features, its marginal contribution is $v(S \cup \{i\}) - v(S)$.

Shapley values come with four desirable properties: (i) efficiency, meaning that the sum of the Shapley values of all features equals the value of the total coalition; (ii) symmetry, meaning that all features have a fair chance to join the prediction; (iii) dummy, meaning if a feature contributes nothing to any coalition S, then the contribution of that feature is zero; and (iv) additivity, meaning that for any pair of predictions v, w: $\phi(v + w) = \phi(v) + \phi(w)$, where $(v + w)(S) = v(S) + w(S)$ for all S.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

SHapley Additive exPlanations (SHAP) are a XAI tool developed by Lundberg and Lee [70] that uses these Shapley values to explain machine learning models. It includes the model-agnostic SHAP `KernelExplainer` that works universally for any prediction model. The `KernelExplainer` builds a weighted linear regression by using the given data, the predictions, and the function/model that predicts the predictions. It computes the feature importance values based on the Shapley values and the coefficients from a local linear regression. Besides the `KernelExplainer`, the SHAP tool also includes other explainers that have been optimized for specific models. One example is the `TreeExplainer`, which was optimized for tree-based prediction models [66]. According to Lundberg et al. [66], it is the only tool that enables the exact computation of optimal local explanations for tree-based models. The `TreeExplainer` can also be used as a global explanation method by averaging local explanations. For example, if this is done over all instances in a dataset, it results in a global measure of feature importance.

IV. APPLICATION OF EXPLAINABLE STUDENT AGENCY ANALYTICS

In this section, we present the results from an application of XSAA in higher education. All the analytics were performed in Python 3.8.2, using `LIME` and `SHAP` toolboxes.

a: Sample and study context

Four courses on mathematics (A1-A4) of first-year engineering students ($n = 141$) in a Finnish higher education institution (university of applied sciences, ISCED Level 6) were studied. Each course had a different responsible teacher but the same basic contents and learning goals. The teaching arrangements as a whole were mostly traditional: lectures and guided exercises in a classroom and additional homework. The courses consisted of instructional videos, automatic tests that guided the student depending on the answers, and a final test. In addition to class hours, teachers sent emails to the whole student group using the virtual learning environment. Personal messages between teachers and students were exchanged by email. In all the courses, mid-term feedback was collected, and depending on the results, some small modifications were made (for example, more time was allocated to topics the students found challenging). All the courses also had voluntary support classes guided by the teacher.

Different practices were also used between the courses. Attendance affected the evaluation in one course (A2). Two courses (A1 and A4) made continuous self-assessments; one based on homework and their model solutions (A1) and the other based on the results of automatic tests in the learning environment (A4). One course (A3) had extra support hours guided by a student assistant. In one course (A4), the students had the opportunity to get a small amount of personal guidance from the teacher if necessary. Moreover, this course (A4) made weekly applications on the topics practiced and had small teams.

b: Analysis between prototypes

Prototypical student agency profiles were created using clustering. The different prototypical agency profiles (PAP1–PAP4) and the general agency profile (GAP) are presented in Figure 3. GAP is the profile of all the analyzed students. All the agency dimensions maintain the order from the lowest profile PAP1 to the highest profile PAP4. In general, the relational resources of student agency (equal treatment, trust for the teacher, and teacher support) were experienced as the highest resource domain and > 4 in all profiles except in PAP1. Three of the participatory resources (participation activity, opportunities to make choices, and opportunities to influence) were generally experienced as lower than other resources in all the profiles. The rest of the participatory resources and the personal resources were experienced close to the factor value of 4 at the GAP level. PAP1 was particularly characterized by low personal resources.

c: Analysis between courses

The analysis between courses revealed differences in student agency between the four different course instances (A1–A4). Figure 4 presents the box plots of each student agency dimension in each of the course instances. There were statistically significant differences in all the dimensions based on the pairwise comparison using the Mann-Whitney U statistics. In particular, the student agency dimensions of trust for the teacher, teacher support, and opportunities to influence were experienced as lower in the A3 course instance comparing to other courses, and the difference was statistically significant.

We also examined if there were any dominant prototypical profiles present in each of the courses (Table 1). Based on the chi-square test of the contingency table, statistically significant differences were observed; $\chi^2(9, n = 141) = 30.1, p < .001$. More students were assigned to the higher agency profiles PAP3–PAP4 in the courses A1 and A4. In course A4, no students were observed in the low agency profile PAP1. In course A3, the majority of the students were in the profiles PAP1–PAP3, and only 5% were in the high agency profile. In A2, a somewhat equal quantity of students were assigned to each PAP.

TABLE 1. Students representing the different prototypical profiles PAP1–PAP4 in each course instance A1–A4, with row-wise percentages; $\chi^2(9, n = 141) = 30.1, p < .001$.

	PAP1	PAP2	PAP3	PAP4
A1	4 (13%)	4 (13%)	16 (53%)	6 (20%)
A2	10 (30%)	8 (24%)	7 (21%)	8 (24%)
A3	16 (28%)	17 (29%)	22 (38%)	3 (5%)
A4	0 (0%)	11 (37%)	9 (30%)	10 (33%)

d: Prediction results

In comparison to earlier work, we not only created the student agency profiles here but also built models predicting these profiles. Using these models, their global model-specific explanations, and local model-agnostic LIME and SHAP explanations on top of them allows us to identify the most

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

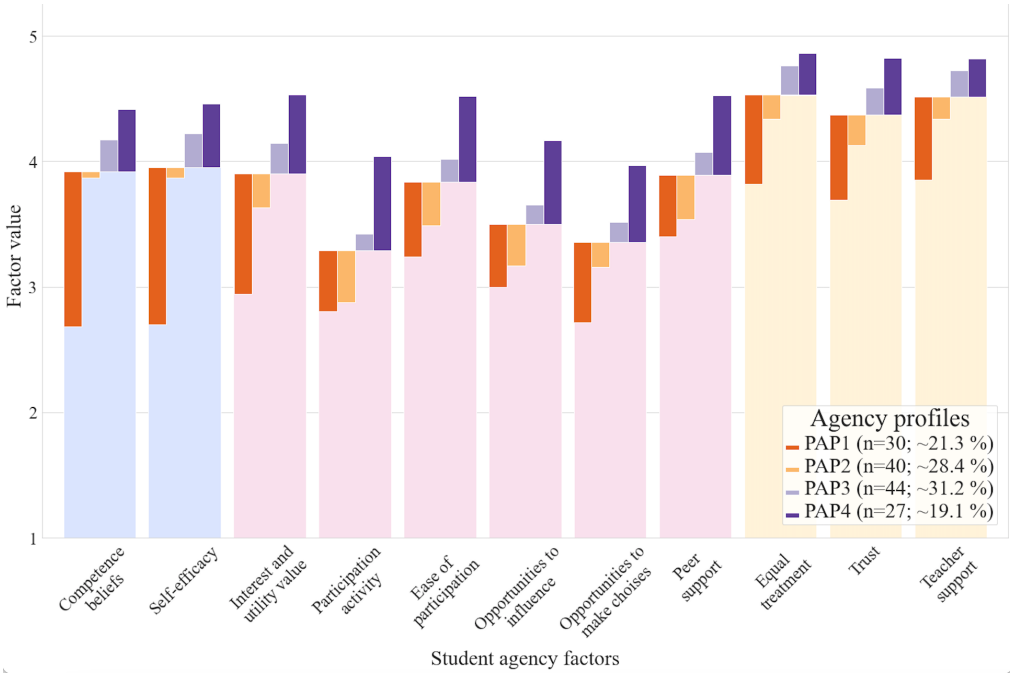

FIGURE 3. Student agency prototype profiles (PAP1–PAP4) and the general average profile of students (n = 141) studying in an engineering education program in a higher education institution.

important characteristics explaining why certain students are assigned to certain profiles. To predict the multinomial class label (i.e., the agency profile), we used all 15 features: the 11 agency dimensions and the four course variables that were one-hot encoded into binary features.

To estimate and compare the models for the supervised task (i.e, predicting the student profile), we divided the data with a stratified split into a training (80%) and an independent test set (20%). Then, we used stratified fivefold cross-validation on the training set to estimate the best hyperparameters for the classifiers. We compared the multinomial logistic regression (MLR) with l_1, l_2, and elasticnet penalization, random forest, and MLP classification models to predict the agency profile. Table 2 summarizes the best model for each classifier as determined through the fivefold cross-validation on the training set and its performance on the independent test set. As shown in the table, the two nonlinear classifiers (random forest and MLP) outperformed the three linear classifiers. Overall, random forest was the best performing classifier when comparing all classifiers, and multinomial logistic regression with l_1 penalization was the best linear classifier.

TABLE 2. Accuracy of the supervised models predicting the agency profile PAP1–PAP4 of the student.

Classifier	test set accuracy	train set mean (std)
MLR l_2	0.724	0.767 ($\pm$ 0.098)
MLR l_1	0.897	0.839 ($\pm$ 0.061)
MLR elastincnet	0.793	0.829 ($\pm$ 0.088)
Random forest	0.966	0.863 ($\pm$ 0.069)
MLP	0.897	0.875 ($\pm$ 0.103)

e: Global explanations

Since random forest was the best classifier overall and the multinomial logistic regression with l_1 penalization the best linear classifier, we focused on these two models to explain the prediction results. Figure 8 shows the coefficients of the multinomial logistic regression with l_1 penalization predicting the highest agency profile PAP4. Figure 9 shows the coefficients of the multinomial logistic regression with l_1 penalization for all four agency profiles. The figures illustrate that overall, the agency dimensions seem more important for the prediction model than the course variables. However, being in a certain course can also increase or decrease the probability of belonging to a particular agency profile. For example, being in course A1 decreases the probability of belonging to agency profile PAP2 and increases the probability of belonging to agency profile PAP3 (see Figure 9).

Figure 10 shows the importance of the features of the random forest model predicting the agency profile. In comparison to the coefficients from the multinomial logistic regression, the feature importance levels of the random forest are always positive and do not encode which class a feature is indicative of. The random forest feature importance levels can tell us that a certain feature is important, but not whether it is indicative of a student having agency profile PAP1, PAP2, PAP3, or PAP4. Moreover, they provide no information in regard to whether a high feature value increases or decreases the probability for a certain class. They just summarize the importance of each feature for the whole model.

If we combine all the local SHAP values (the results of the individual local explanations are provided in the next

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

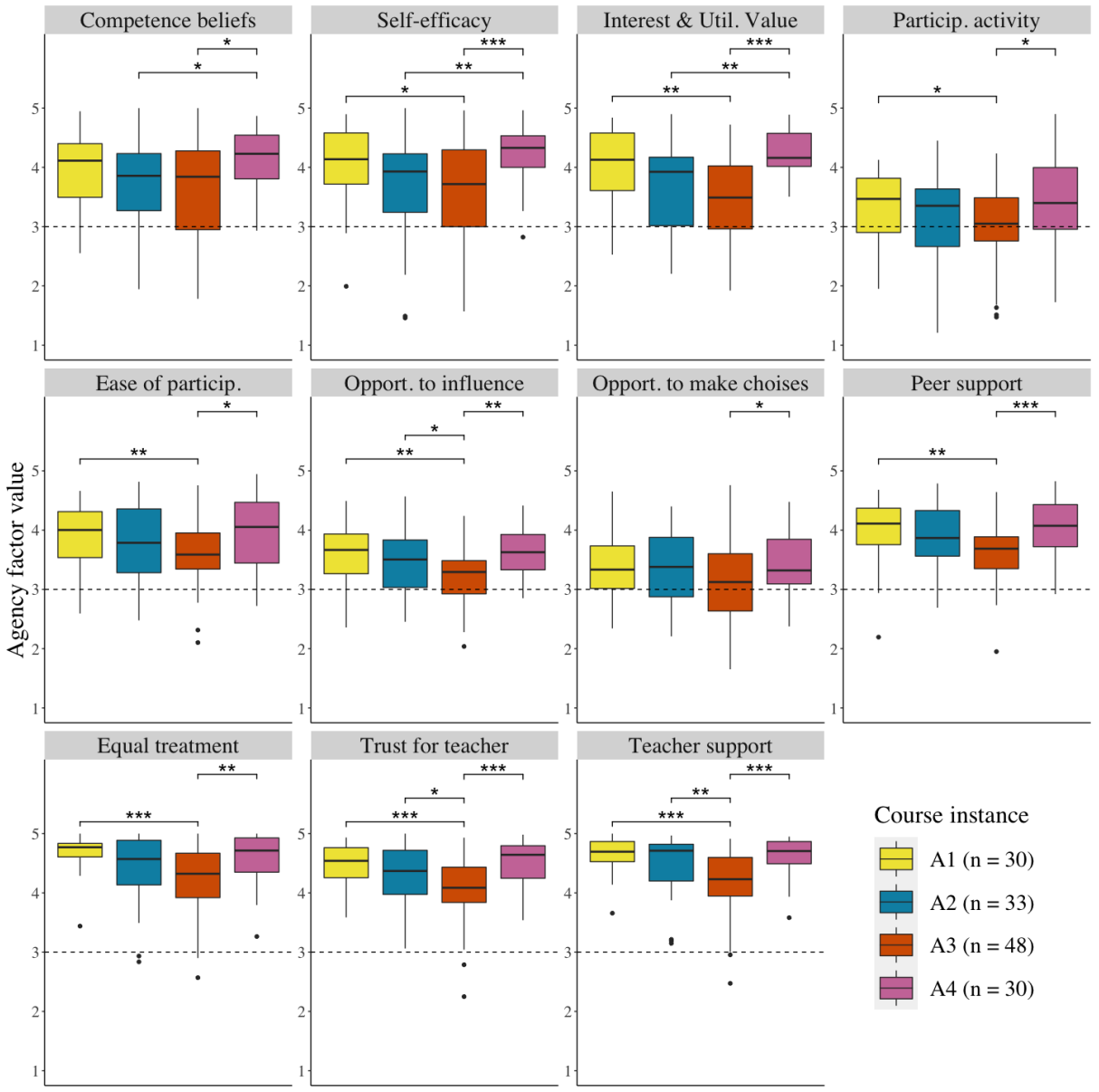

FIGURE 4. Student agency dimension in each course instance and pairwise statistical significance using Mann-Whitney U statistics. As usual, $\star$ corresponds to $p < 0.05$, $\star\star$ to $p < 0.01$, and $\star\star\star$ to $p < 0.001$.

section) for all the students, we can also get the global SHAP explanations for a model. This is shown in Figure 5 for the random forest classification model. As the figure shows, a student's competence belief was the most important feature for the model, especially when determining if he/she belongs to the lowest (PAP1) agency profile. This model-agnostic explanation is the same as that from the model-specific feature importance levels (see Figure 10, here the competence belief was also the most important feature) but more informative as it also shows which features are important for each profiles.

f: Local explanations

As explained in Section III, local explanations enable us to explain why a certain student received his/her prediction and the contributions of the individual predictors. Global feature importance, as discussed above, only shows the results across the entire population, but not on each individual student. The local explanations, in contrast, enable us to pinpoint and contrast the impacts of the factors for particular students.

To explain the model predictions for particular students, we used the true positives with the highest probability for each agency profile; that is, those four students from the test set that the model correctly predicted to be PAP1, PAP2, PAP3, and PAP4, respectively, with the highest probability. Table 3 summarizes these local explanations for the random forest model. As we saw already in the global model-specific explanations (Figure 10), the opportunities to influence was one of the most important variables for the random forest

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

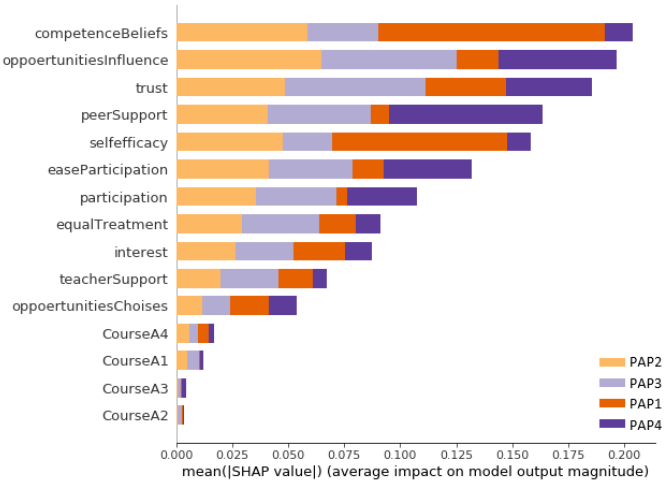

FIGURE 5. Global SHAP explanations for the random forest model. For competence beliefs, the mean absolute SHAP values are 0.1 for PAP1, 0.06 for PAP2, 0.03 for PAP3, and 0.01 for PAP4, making it altogether the most important global predictor for this model.

model. However, from Table 3, we can also see for which profiles this variable was especially important (namely, agency profile PAP2, PAP4, and especially PAP3).

The LIME rules can also be presented visually. Figure 11 shows the LIME rule visualization for the PAP2 student who was predicted to be a PAP2 profile with the highest probability with the random forest model. For comparison, Figure 6 shows the SHAP local explanations for the same model and student. This plot provides a more comprehensive explanation overview of the prediction than the LIME rules.

More specifically, as Figure 6 shows, the model predicted an 88 percent chance that this student was a PAP2 student, whereas the base value (i.e., the prediction if nothing would be known about this student) for PAP2 was a 29 percent chance. The feature values causing increased predictions are in red, and their visual size shows the magnitude of the feature's effect. The biggest impact comes from the opportunities to influence, which is 3.16 for this student. The feature values decreasing the prediction are in blue. As can be seen in Figure 6, the fact that this student is in course A1 had a meaningful effect, decreasing the prediction. The model predicted some tiny probabilities that this student was a PAP1 or PAP3 student, but his/her competence beliefs are lower than for PAP3 and higher than for PAP1 students. If one subtracts the length of the blue bars from the length of the red bars, it equals the distance from the base value to the output. This means that the baseline plus the sum of individual effects add up to the prediction as discussed in Section III.

g: Local explanations for the student needing the most support

The local explanations also enable us to locate the students needing support the most and to receive the explanations describing which factors could affect a change toward higher agency. Based on Table 1 and Figure 9, we can conclude that the students in course A3 needed the most support. Since profile PAP1 represents the lowest agency profile, we chose the student from the test set who was in course A3 and was predicted to have the lowest agency profile PAP1 with the highest probability for the local explanations. Figure 7 shows the SHAP values explaining why this student was assigned to profile PAP1 with the highest probability. As Figure 7 illustrates, the base value of the prediction in the absence of any information on the independent variables is 0.2138. Knowing that the competence beliefs of this student are only 1.907 increased the prediction that this student is PAP1 by 0.222, and knowing that the self-efficacy value of this student is 1.878 increased the prediction for profile PAP1 by another 0.176 (see Table 4).

A. SUMMARY AND DISCUSSION OF RESULTS

Our results can be summarized from the application level and the methodological level. From the application level, we can conclude that the level of student agency was higher in the two courses, A1 and A4, where continuous task-driven self-assessment took place. No students were in the lowest agency profile PAP1 in the course A4, and the majority of the students in A1 and A4 were in the higher agency profiles PAP3 and PAP4. One reason for the students' generally high sense of agency in course A4 might be the personal guidance that the teacher offered in the course. Furthermore, a joint analysis of Figure 8, Figure 5, and Table 3 suggests that if the students found support from their peers and experienced opportunities to influence and participate in the course, they tended to have higher agency profiles.

From the teacher's perspective, the XSAA results could provide insight for pedagogical planning. For example, the students in course A1 seem to have received the proper amount of teacher's support and attention, as relational resources were scored high and those resources represented some of the most important resource areas for the second highest agency profile PAP3 (Figure 3, Figure 5, and Table 3). To foster student agency of the PAP2 and PAP3 students in A1, the teacher could provide low-threshold ways for participation because the participatory resources were considered important in the highest profile PAP4. In addition, suggestions to improve pedagogical planning could be made by analyzing the characteristics of the students in the lowest agency profile PAP1. The findings suggest that low self-efficacy and competence beliefs are important common nominators for students in PAP1 (Figure 3, Figure 5, and Figure 7). As there were many PAP1 students in course A3, these students might benefit from more extensive encouragement as well as more attention and support in understanding the course contents (cf., [71]).

From the methodological level, our results showed that the complex nonlinear methods, especially the random forest, improved the accuracy of the predictive models. The traditional linear techniques performed worse but came with more informative global model-specific explanations. For

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

TABLE 3. LIME rules explaining the true positive students for each profile from the test set with the highest probability with the random forest model. For each student, the rules are ordered by importance with the most important rule first.

Profile PAP1		**Profile PAP2**	
rule	importance	rule	importance
competenceBeliefs <= 3.44	0.241	participation <= 2.81	0.095
selfefficacy <= 3.41	0.169	3.03 < oppoertunitiesInfluence <= 3.43	0.08
trust <= 3.99	0.066	peerSupport <= 3.56	0.072
interest <= 3.33	0.042	3.99 < trust <= 4.44	0.065
equalTreatment <= 4.17	0.03	3.38 < easeParticipation <= 3.76	0.062
oppoertunitiesChoises <= 2.86	0.029	3.44 < competenceBeliefs <= 3.96	0.039
teacherSupport <= 4.17	0.029	3.41 < selfefficacy <= 4.03	0.032
oppoertunitiesInfluence <= 3.03	0.028	3.33 < interest <= 3.94	0.029
easeParticipation <= 3.38	0.021	4.17 < equalTreatment <= 4.61	0.026
CourseA4 <= 0.00	0.02	4.17 < teacherSupport <= 4.61	0.021
peerSupport <= 3.56	0.009	CourseA4 <= 0.00	-0.02
CourseA1 <= 0.00	-0.006	CourseA1 > 0.00	-0.018
CourseA2 <= 0.00	-0.005	2.86 < oppoertunitiesChoises <= 3.26	0.016
participation <= 2.81	0.001	CourseA2 <= 0.00	-0.007
0.00 < CourseA3 <= 1.00	0.0	CourseA3 <= 0.00	0.001
Profile PAP3		**Profile PAP4**	
rule	importance	rule	importance
3.43 < oppoertunitiesInfluence <= 3.79	0.078	peerSupport > 4.27	0.138
competenceBeliefs > 4.34	0.054	oppoertunitiesInfluence > 3.79	0.113
trust > 4.71	0.054	easeParticipation > 4.28	0.081
4.61 < equalTreatment <= 4.83	0.053	trust > 4.71	0.075
teacherSupport > 4.81	0.052	participation > 3.72	0.064
3.88 < peerSupport <= 4.27	0.045	competenceBeliefs > 4.34	0.03
interest > 4.35	0.044	interest > 4.35	0.028
3.76 < easeParticipation <= 4.28	0.031	oppoertunitiesChoises > 3.70	0.024
4.03 < selfefficacy <= 4.40	0.021	selfefficacy > 4.40	0.018
CourseA1 <= 0.00	-0.015	equalTreatment > 4.83	0.013
CourseA4 > 0.00	-0.014	CourseA4 > 0.00	0.011
2.81 < participation <= 3.34	0.012	teacherSupport > 4.81	0.005
oppoertunitiesChoises > 3.70	0.011	CourseA3 <= 0.00	0.005
CourseA3 <= 0.00	-0.01	CourseA1 <= 0.00	-0.001
CourseA2 <= 0.00	0.009	CourseA2 <= 0.00	-0.001

example, while the global model-specific explanations from the random forest simply provided a ranking of the input features, the global model-specific explanations of the logistic regression with l_1 penalization also showed which feature was important for which class and which direction (i.e., whether it increased or decreased the probability for this class). Moreover, several features were dropped from the model, making it sparser and more interpretable.

Through recently developed model-agnostic XAI tools, we were able to also explain the better performing classifiers. LIME and SHAP can be used on top of any (complex) classifier to explain predictions for particular students (local explanations). These local explanations are very important, mainly for two reason. First, the GDPR now includes a right for explanation [57]. This means that if an automatic profiling is used in an LA tool, the student has a right to receive an explanation about his/her particular profiling.

Second, the local and global explanations can be different, and it is thus not enough to use the global explanations to explain why a particular student was assigned to a certain profile. For example, according to Figure 5, the most important agency dimensions for PAP2 (visual consideration of the lengths of the orange bars) are opportunities to influence, competence beliefs, and then trust for the teacher and self-efficacy. However, according to the LIME rules for that student in the test set who was assigned to PAP2 with the highest probability (Table 3), the order of importance concerning agency dimensions was participation activity, opportunities to influence, peer support, and then trust for the teacher.

In other words, the LIME rules (also those for the students that are representative for their PAP-profiles) do not always resemble the global explanations (Figure 5). For example, for the particular PAP2 student analyzed in Table 3, the value of participation activity was extremely low (2.69, see Figure 11), and the local surrogate model built by LIME to explain this prediction relied on this feature to a significant degree. This exemplifies the "local fidelity" of LIME: LIME explanations can be trusted only locally around the specific instance being explained. In contrast, the local SHAP explanations can—because of their additivity—be combined so that they can also be used to explain the global behavior of the model (Figure 5), being therefore more in line with the global model-specific explanations (Figure 10).

Naturally, our results are limited to the relatively small amount of data. Further data collection is required to increase the reliability of the observed connections between student agency and course implementations in higher education. In this paper, we have established the foundations for the use of XAI techniques in analyzing students' agency. Further work is required to examine, for example, the causal relationships of teaching practices and student agency.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

FIGURE 6. SHAP values explaining why the random forest model predicted an agency profile 2 student from the test set to be profile PAP2 and not profile 1, 3, or 4 (the bars are ordered by the profile number; i.e, the first bar predicts PAP1, the second PAP2, and so on). For each bar, the values explain how to get from the *base value* that would be predicted if no feature would be known to the current output for this particular profile 2 student. Feature values causing increased predictions are in red, and feature values decreasing the prediction are in blue. Their visual size shows the magnitude of the feature's effect.

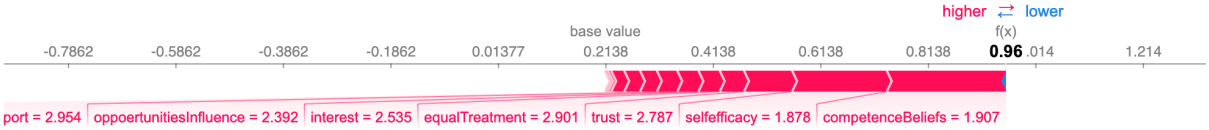

FIGURE 7. SHAP values explaining why the random forest model predicted an agency profile PAP1 student studying in course A3 from the test set to be profile PAP1 (true positive). The most important explanations are the low competence beliefs and self-efficacy values of this student.

V. CONCLUSION

Student agency is a key construct in the contemporary discourse about student-centered learning in higher education [3]–[5]. Jääskelä et al. [9] developed an LA process called student agency analytics (SAA), which utilizes a psychometric questionnaire instrument [35] and machine learning to provide information about the different resources of student agency. The recent literature on LA has highlighted the importance of explainability when utilizing complex models in education (e.g., [11], [14], [72]). In this study, we employed XAI techniques to derive more detailed information from student agency data. The purpose was to illustrate how the SAA process, combined with XAI techniques, could advance teachers' pedagogical awareness and reflection.

The purpose of the XAI techniques is to help to gain an understanding of how and why a model works. We used the multinomial logistic regression coefficients, feature importance levels of the random forest model, and combined SHAP values to explain the essential characteristics of the different agency profiles (global explanation). The prediction of the student profiles showed that the nonlinear techniques (especially random forest) modeled the data the best. The finding indicates that the relationships between the prototypical profiles of student agency and the teaching practices in higher education are relatively complex. Local explanations gave insight into why a student was assigned to a particular agency profile. Altogether, the XSAA results could be used to derive tentative explanations of the different experiences of student agency and to suggest ideas for pedagogical planning, as summarized in Section IV-A.

Educators at all levels of education need to take steps toward supporting student agency. To promote the educators' efforts, Moses et al. [4] called for connecting theory and practice and suggested increasing the research and practitioner-focused work about how teachers could support student agency. They emphasize that student agency "is a practice-embedded construct that shapes the daily work of educators" by involving them in reflecting the ways to create agentic spaces for students and making pedagogical decisions based on that reflection [4]. We see that this kind of teacher reflecting, pedagogical planning, and sharing of experiences of the agency-supporting practices among the colleagues could be facilitated using research-based tools and explainable SAA. These tools could help teachers to detect and understand the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3116664, IEEE Access

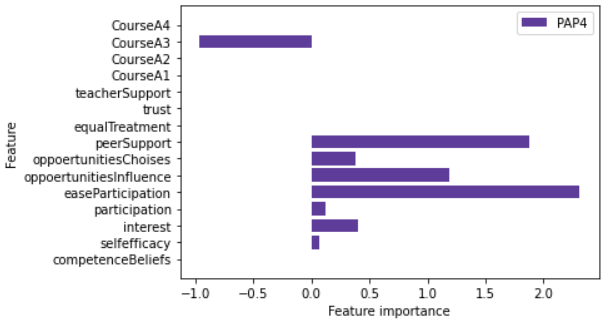

FIGURE 8. Coefficients of the multinomial logistic regression with l_1 penalization predicting the highest agency profile (PAP4). For seven features, the coefficient is zero, meaning they were irrelevant for this prediction model. A high value in all the picked features (except CourseA3) increases the probability that a student will be assigned to PAP4. However, if the student is in course A3, the probability that he/she will be assigned to PAP4 decreases.

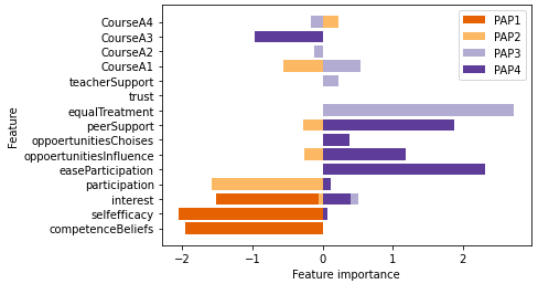

FIGURE 9. Coefficients of the multinomial logistic regression with l_1 penalization predicting agency profiles PAP1-PAP4. As a whole, the course features seem not as important as the agency dimensions but they are contributing. For example, if a student is in course A1, the probability that he/she will have the second highest agency (PAP3) increases.

different experiences of student agency in their courses.

In summary, explainable models can provide more detailed and meaningful information about the different dimensions of student agency. By getting an overview of the different experiences of student agency in their courses, teachers could better meet the practical challenges of supporting student agency. Furthermore, higher education institutions could better adapt their capabilities to different learners' needs now and in the future. Thus, XSAA has the potential to contribute to teachers' pedagogical planning through the LA cycle.

APPENDIX

REFERENCES

[1] A. Schleicher, "The impact of covid-19 on education insights from education at a glance 2020," 2020.

[2] M. Silverman, R. Sibbald, and S. Stranges, "Ethics of covid-19-related school closures," *Canadian Journal of Public Health*, vol. 111, no. 4, pp. 462–465, 2020.

[3] P. Jääskelä, A.-M. Poikkeus, P. Häkkinen, K. Vasalampi, H. Rasku-Puttonen, and A. Tolvanen, "Students' agency profiles in relation to student-perceived teaching practices in university courses," *International Journal of Educational Research*, vol. 103, p. 101604, 2020.

[4] L. Moses, D. Rylak, T. Reader, C. Hertz, and M. Ogden, "Educators' perspectives on supporting student agency," *Theory Into Practice*, vol. 59, no. 2, pp. 213–222, 2020.

[5] A. Schleicher, "Concept note: OECD learning compass 2030," OECD, Tech. Rep., 2019.

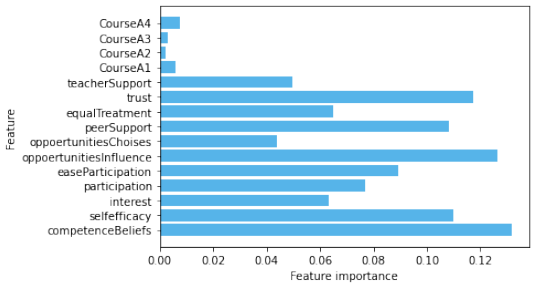

FIGURE 10. Feature importances of the random forest model predicting the agency profile. For the random forest, the agency dimensions are more important than the course features.

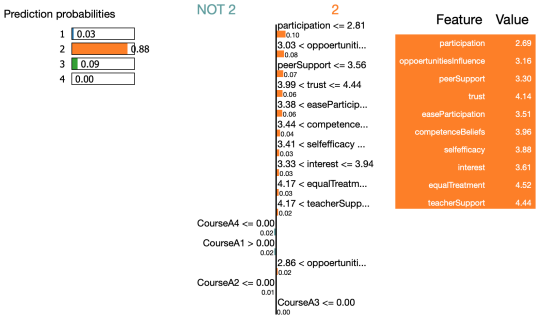

FIGURE 11. LIME rules explaining why the random forest model predicted an agency profile PAP2 student from the test set to be profile PAP2 (i.e., a true positive) with the highest probability. The most important local explanation why this student was assigned to this profile are his/her participation values.

TABLE 4. SHAP values (rounded to three decimals) for that student from the test set, who is in course A3 and was assigned to PAP1 with the highest probability by the random forest classifier. The table shows that the competence belief was the most important variable for the prediction.

variable	PAP1	PAP2	PAP3	PAP4
competenceBeliefs	0.222	-0.163	-0.042	-0.016
selfefficacy	0.176	-0.132	-0.03	-0.014
interest	0.04	-0.008	-0.024	-0.008
participation	0.006	0.034	-0.022	-0.017
easeParticipation	0.031	0.017	-0.025	-0.023
oppoertunitiesInfluence	0.039	0.036	-0.043	-0.032
oppoertunitiesChoises	0.029	-0.016	-0.008	-0.005
peerSupport	0.024	0.021	-0.013	-0.033
equalTreatment	0.045	-0.008	-0.031	-0.006
trust	0.089	-0.015	-0.049	-0.026
teacherSupport	0.038	-0.009	-0.025	-0.004
CourseA1	0.0	0.001	-0.002	0.0
CourseA2	-0.001	0.0	0.0	-0.0
CourseA3	0.0	0.0	0.001	-0.002
CourseA4	0.005	-0.005	0.001	-0.001

[6] K. Collin, S. Lemmetty, S. Herranen, S. Paloniemi, T. Auvinen, and E. Riivari, "Professional agency and creativity in information technology work," in *Agency at Work*. Springer, 2017, pp. 249–270.

[7] G. Siemens, "Learning analytics: The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, 2013.

[8] J. Jovanović, M. Saqr, S. Joksimović, and D. Gašević, "Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success," *Computers & Education*, p. 104251, 2021.

[9] P. Jääskelä, V. Heilala, T. Kärkkäinen, and P. Häkkinen, "Student agency analytics: learning analytics as a tool for analysing student agency in higher education," *Behav. Inf. Technol.*, pp. 1–19, Feb. 2020.

[10] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. De Weerdt,

"A review of automated feedback systems for learners: Classification framework, challenges and opportunities," *Computers & Education*, vol. 162, p. 104094, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S036013152030292X

[11] R. S. Baker, "Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes," *Journal of Educational Data Mining*, vol. 11, no. 1, pp. 1–17, 2019.

[12] V. Heilala, M. Saarela, P. Jääskelä, and T. Kärkkäinen, "Course satisfaction in engineering education through the lens of student agency analytics," in *2020 IEEE Frontiers in Education Conference (FIE)*, 2020, pp. 1–9.

[13] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, and N. Radi, "Machine learning approaches to predict learning outcomes in massive open online courses," in *2017 International Joint Conference on Neural Networks*. IEEE, 2017, pp. 713–720.

[14] A. Dutt and M. A. Ismail, "Can we predict student learning performance from lms data? a classification approach," in *3rd International Conference on Current Issues in Education*. Atlantis Press, 2019, pp. 24–29.

[15] A. P. Patil, K. Ganesan, and A. Kanavalli, "Effective deep learning model to predict student grade point averages," in *2017 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2017, pp. 1–6.

[16] M. Stapel, Z. Zheng, and N. Pinkwart, "An ensemble method to predict student performance in an online math learning environment." *International Educational Data Mining Society*, pp. 231–238, 2016.

[17] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[18] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2020.

[19] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, Apr. 2019.

[20] D. Ifenthaler and J. Y.-K. Yau, "Utilising learning analytics to support study success in higher education: a systematic review," *ETR&D Educational technology research and development*, vol. 68, no. 4, pp. 1961–1990, Aug. 2020.

[21] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Computers in Human Behavior*, vol. 89, pp. 98–110, Dec. 2018.

[22] X. Du, J. Yang, B. E. Shelton, J.-L. Hung, and M. Zhang, "A systematic meta-review and analysis of learning analytics research," *Behaviour & information technology*, pp. 1–14, Sep. 2019.

[23] Y. Cui, F. Chen, A. Shiri, and Y. Fan, "Predictive analytic models of student success in higher education: A review of methodology," *Information and Learning Sciences*, vol. 120, no. 3/4, pp. 208–227, Jan. 2019.

[24] A. Larrabee Sønderlund, E. Hughes, and J. Smith, "The efficacy of learning analytics interventions in higher education: A systematic review," *British journal of educational technology*, vol. 31, p. 209, Nov. 2018.

[25] P. Leitner, M. Khalil, and M. Ebner, "Learning analytics in higher Education—A literature review," in *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning*, A. Peña-Ayala, Ed. Cham: Springer International Publishing, 2017, pp. 1–23.

[26] V. Putnam and C. Conati, "Exploring the need for explainable artificial intelligence (xai) in intelligent tutoring systems (its)." in *IUI Workshops*, vol. 19, 2019, pp. 1–7.

[27] C. Conati, K. Porayska-Pomsta, and M. Mavrikis, "AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling," *CoRR*, vol. abs/1807.00154, 2018.

[28] J. M. Alonso and G. Casalino, "Explainable artificial intelligence for human-centric data analysis in virtual learning environments," in *International workshop on higher education learning methodologies and technologies online*. Springer, 2019, pp. 125–138.

[29] A. Bandura, "Toward a psychology of human agency," *Perspectives on Psychological Science*, vol. 1, no. 2, pp. 164–180, Jun. 2006.

[30] M. S. Archer and M. S. Archer, *Structure, agency and the internal conversation*. Cambridge, England: Cambridge University Press, 2003.

[31] Y.-H. Su, "The constitution of agency in developing lifelong learning ability: The 'being'mode," *Higher Education*, vol. 62, no. 4, pp. 399–412, 2011.

[32] S. Hoidn and K. Reusser, "Foundations of Student-Centered learning and teaching," in *The Routledge International Handbook of Student-Centered Learning and Teaching in Higher Education*, S. Hoidn and M. Klemenčič, Eds. Abingdon-on-Thames: Routledge, 2020, pp. 17–46.

[33] J. Martin, "Self-Regulated learning, social cognitive theory, and agency," *Educational psychologist*, vol. 39, no. 2, pp. 135–145, Jun. 2004.

[34] L. Starkey, "Three dimensions of student-centred education: a framework for policy and practice," *Critical Studies in Education*, pp. 1–16, Jan. 2017.

[35] P. Jääskelä, A.-M. Poikkeus, K. Vasalampi, U. M. Valleala, and H. Rasku-Puttonen, "Assessing agency of university students: validation of the AUS scale," *Studies in Higher Education*, vol. 42, no. 11, pp. 1–19, 2017.

[36] J. Santos, A. S. Figueiredo, and M. Vieira, "Innovative pedagogical practices in higher education: An integrative literature review," *Nurse education today*, vol. 72, pp. 12–17, 2019.

[37] L.-M. Hero and E. Lindfors, "Students' learning experience in a multidisciplinary innovation project," *Education and training*, vol. 61, no. 4, pp. 500–522, 2019.

[38] P. Garnjost and L. Lawter, "Undergraduates' satisfaction and perceptions of learning outcomes across teacher-and learner-focused pedagogies," *The International Journal of Management Education*, vol. 17, no. 2, pp. 267–275, 2019.

[39] P. Koskinen, J. Lämsä, J. Maunuksela, R. Hämäläinen, and J. Viiri, "Prime-time learning: collaborative and technology-enhanced studying with genuine teacher presence," *International Journal of STEM Education*, vol. 5, no. 1, p. 20, 2018.

[40] M. J. Leenknecht, L. Wijnia, S. Loyens, and R. Rikers, "Need-supportive teaching in higher education: Configurations of autonomy support, structure, and involvement," *Teaching and Teacher Education*, vol. 68, pp. 134–142, 2017.

[41] M. Räisänen, L. Postareff, and S. Lindblom-Ylänne, "Students' experiences of study-related exhaustion, regulation of learning, peer learning and peer support during university studies," *European Journal of Psychology of Education*, pp. 1–23, 2020.

[42] T. Hailikari, N. Katajavuori, and S. Lindblom-Ylanne, "The relevance of prior knowledge in learning and instructional design," *American journal of pharmaceutical education*, vol. 72, no. 5, 2008.

[43] Y.-S. Tsai, D. Rates, P. M. Moreno-Marcos, P. J. Muñoz-Merino, I. Jivet, M. Scheffel, H. Drachsler, C. D. Kloos, and D. Gašević, "Learning analytics in european higher education—trends and barriers," *Computers & Education*, vol. 155, p. 103933, 2020.

[44] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, "A reference model for learning analytics," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5-6, pp. 318–331, 2012.

[45] C. P. Rosé, "Learning analytics in the learning sciences," in *International handbook of the learning sciences*. Routledge, 2018, pp. 511–519.

[46] M. Saarela, *Automatic knowledge discovery from sparse and large-scale educational data: case Finland*. University of Jyväskylä, 2017, no. 262.

[47] P. Blikstein, "Multimodal learning analytics," in *Proceedings of the third international conference on learning analytics and knowledge*, 2013, pp. 102–106.

[48] R. Luckin and M. Cukurova, "Designing educational technologies in the age of AI: A learning sciences-driven approach," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 2824–2838, 2019.

[49] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[50] A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD explorations newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[51] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114, 2015.

[52] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[53] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[54] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[55] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[56] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, "Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai," *arXiv preprint arXiv:1902.01876*, 2019.

[57] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[58] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accountable AI for Robotics," *Science Robotics*, vol. 2, no. 6, 2017.

[59] C. Molnar, *Interpretable Machine Learning*. Lean Publishing, 2019.

[60] M. Saarela and T. Kärkkäinen, "Can we automate expert-based journal rankings? Analysis of the Finnish publication indicator," *Journal of Informetrics*, vol. 14, no. 2, p. 101008, 2020.

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

[62] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[63] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[64] A. J. London, "Artificial intelligence and black-box medical decisions: accuracy versus explainability," *Hastings Center Report*, vol. 49, no. 1, pp. 15–21, 2019.

[65] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[66] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[67] M. Saarela, O.-P. Ryynänen, and S. Äyrämö, "Predicting hospital associated disability from imbalanced data using supervised learning," *Artificial intelligence in medicine*, vol. 95, pp. 88–95, 2019.

[68] M. T. Ribeiro, S. Singh, and C. 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*, 2016, pp. 1135–1144.

[69] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[70] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[71] V. Heilala, P. Jääskelä, T. Kärkkäinen, and M. Saarela, "Understanding the study experiences of students in low agency profile: Towards a smart education approach," in *International conference on smart Information & communication Technologies*. Springer, 2019, pp. 498–508.

[72] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020.