Human-crafted Features in Machine Learning Increase Trust but Risk Over-reliance

Liang Tang
University of Illinois Urbana–Champaign
Champaign, IL, USA
ltang29@illinois.edu

Nigel Bosch
University of Illinois Urbana–Champaign
Champaign, IL, USA
pnb@illinois.edu

ABSTRACT

Feature engineering plays a critical role in the development of machine learning systems for educational contexts, yet its impact on student trust remains understudied. Traditional approaches have focused primarily on optimizing model performance through expert-crafted features, while the emergence of AutoML offers automated alternatives for feature extraction. Through an experimental design comparing expert-crafted features with two AutoML approaches (Featuretools and TSFRESH), we investigated the relationship between feature types and student trust in educational systems. Analysis of student interactions with these systems revealed significant variations in trust formation, reliance, and decision-making behaviors based on feature type. We measured trust through multiple metrics including compliance behavior, overreliance tendencies, and decision-making patterns such as response time and decision switching. Our results demonstrate that expert-crafted features led to significantly higher trust and compliance compared to AutoML-generated features, but also resulted in concerning levels of over-reliance when system recommendations were incorrect, whereas computationally complex TSFRESH features encountered persistent undercompliance. Expert-created features were also initially more trusted, and the stability of trust perceptions across all conditions suggests that early impressions remain relatively unchanged. We also provide implications for the design of educational machine learning systems, suggesting that while expert-crafted features may better align with students' mental models, careful attention must be paid to preventing over-reliance. These insights contribute to the development of more effective and trustworthy learning analytics tools that better serve student needs.

Keywords

Feature engineering, Trust, Machine learning

1. INTRODUCTION

The integration of machine learning in education has ushered in a new era of learning, prompting a critical examination of human–machine interactions and their impact on educational outcomes [59, 1, 50, 39, 3]. As machine learning systems become increasingly prevalent in classrooms and online learning environments, so too is it increasingly important to understand how students perceive machine learning systems. Trust forms the cornerstone of effective human–machine collaboration [72]. This is especially true in the educational setting [48, 49], where trust has been recognized as an essential foundation for learning and forming cooperation between different stakeholders in education system, e.g., teachers, students, graders [68, 6]. However, establishing trust in machine learning (ML) systems presents unique challenges [72], primarily due to the often inscrutable nature of their decision-making processes.

Recent years have witnessed ML models achieving human-like performance in various tasks [62]. In fully automated systems, these models' predictions are utilized without human intervention, necessitating a high degree of student trust in algorithmic decisions [31]. This shift raises important questions about how we build and maintain trust in machine systems, especially within educational contexts where stakes are high and participation is often compulsory. Questions of trust in machine learning models have become crucial issues [62, 83, 75]. Studies have revealed that trust in these models is influenced by various factors throughout the ML lifecycle; notably, the stated accuracy, observed accuracy and interpretability of ML model have become a focus point in contemporary trust research [11]. Central to this trust discussion is the concept of trust calibration—the alignment between a user's trust in a system and the system's actual capabilities or reliability [37, 46]. Proper calibration occurs when users trust a system in proportion to its actual performance, while miscalibration manifests as either excessive trust (overcalibration) or insufficient trust (undercalibration) relative to system reliability. In educational settings, where students increasingly rely on ML-driven recommendations, achieving appropriate trust calibration becomes essential for effective learning outcomes. Students who overtrust may uncritically accept erroneous recommendations, while those who undertrust may disregard valuable insights, both scenarios potentially impeding learning progress. To build appropriately calibrated student trust in ML systems, we must first understand and optimize the fundamental processes that shape how these

models function.

Feature engineering is the process of selecting, transforming, and creating relevant input variables for machine learning models. Students expect learning analytics to support their planning and organization of learning processes, provide self-assessments, deliver adaptive recommendations, and produce personalized analyses of their learning activities [63]. However, these expectations require feature engineering in creating analytics-driven assessments that harness formative data from learners to facilitate learning processes. Feature engineering establishes the parameters for a model that ultimately determine how certain input data may lead to decisions. The performance and efficacy of machine systems hinges critically on the quality of feature engineering [57, 26], a process that has demonstrated significant potential in enhancing both fairness [61] and interpretability [20].

Despite the advancements in feature engineering for enhancing fairness and interpretability, little attention has been paid to how the extracted features themselves influence trust. In the educational domain, this issue takes on added significance when we consider students as active users of ML-driven tools. Student-facing dashboards, for instance, allow learners to engage with predictions about their academic outcomes [70]. The type and presentation of features in these dashboards can significantly influence how students interpret and trust the information provided. This democratization of data access transforms students from passive subjects of analysis into empowered stakeholders [73], capable of leveraging insights to refine their learning strategies and educational trajectories. The introduction of dashboard learning analytics systems has had a transformative impact on student engagement [56], yet understanding the factors that drive adoption of these systems remains a critical challenge for their successful implementation.

Educational research has also explored factors that increase individual decisions to trust and adopt learning analytics tools. Klein et al. [34] found that organizational context and commitment play a significant role, while Sjöblom et al. [65] emphasized the importance of integration with existing organizational systems. Moreover, West et al. [76] highlighted the value of communication channels in facilitating the adoption of learning analytics while some researchers cited student interface as an important factor [47]. In this context, features engineered for learning analytics tools serve a dual purpose: they bridge communication between stakeholders and form a crucial part of the student interface. As such, the impact of feature engineering on trust and adoption in learning analytics deserves deeper research. This is particularly important given the potential of well-designed features to enhance student engagement and learning outcomes, as demonstrated by the transformative impact of dashboard systems [56].

Given these considerations, our research investigates how students collaborate with and develop trust in educational ML systems, specifically focusing on their interactions with machine-generated versus expert-designed features. By investigating how students engage with these different types of features, we seek to understand the impact on collaborative learning behaviors and the decision to trust. Our findings will provide implications for the development of more effective, trustworthy, and student-centric feature engineering methodologies in educational technology.

2. BACKGROUND

2.1 Machine Learning in Education

Advanced learning analytics tools with ML embedded offer unprecedented opportunities to dive into complex educational datasets, uncovering hidden patterns and relationships that were previously inaccessible [58]. With data mining, educators can now refine their teaching methodologies and tailor interventions to better meet the diverse needs of their students [41]. For instance, learning systems can group similar materials or students based on their learning and interaction patterns and further reveal which types of learning materials (e.g., videos, interactive simulations, or text-based resources) are most effective for different topics or student groups [4].

One of the most promising applications of ML in education is its ability to predict student performance and identify those who may be at risk of academic difficulties [78]. This predictive capability allows for timely interventions and personalized support, evolving the way educational institutions approach student success [22]. Moreover, the integration of ML technologies in education extends beyond predictive functions; it also has the potential to empower students directly by providing them with access to insights derived from their own learning data [23]. For instance, personalized dashboards can display a student's progress across different subjects, highlighting areas of strength and those needing improvement. Some systems even provide comparative analytics, letting students anonymously benchmark their performance against peers, which can motivate self-directed learning [60]. When students can access and review data about their own academic work—such as their assignment completion patterns, quiz scores over time, and topic-specific performance—they can better identify which concepts they have mastered and where they need additional practice[53, 9, 32]. ML in educational contexts has therefore attracted significant scholarly attention. Researchers have explored various aspects of ML in education, from data collection methodologies to student behavior analysis techniques [81, 14, 9, 2, 29]. These studies have highlighted the potential of ML to optimize curriculum delivery and enhance overall learning outcome. For example, Wu et al. used machine learning predictions to automatically decide when to provide a quiz to students, resulting in greater learning versus a control group who received quizzes randomly, and Ball et al. employed machine learning techniques to analyze institutional data, identifying key factors influencing graduation rate then implementing targeted curriculum changes to address specific student needs and enhance graduation outcomes[5, 77]. In the Eye-Mind Reader project, researchers employed a support vector machine to predict mind wandering episodes during reading tasks. Using features from eye-gaze patterns, including fixation durations and pupil diameters, the model identified moments when learners' attention drifted from the instructional text[45].

As the volume and complexity of educational data grow exponentially, researchers and practitioners face the challenge of extracting meaningful insights that can inform pedagogi-

cal strategies and enhance student outcomes [13]. A crucial step in developing effective machine learning models is feature engineering: the process of selecting and transforming relevant attributes from raw data for ML models to utilize. The importance of this step cannot be overstated, as the performance and efficacy of ML models are linked to the quality and comprehensibility of their input features [84]. However, designing features that are both statistically meaningful and interpretable to educators and students poses a significant challenge [33].

One set of approaches to feature engineering that attempts to reduce the manual labor involved is automatic machine learning (AutoML) [12], with a particular focus on automating the feature engineering process. Tools such as Featuretools (for relational data) [30] and TSFRESH (Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests) [19] have revolutionized the way we approach feature extraction, offering unprecedented efficiency and adaptability [7, 19]. These AutoML methods employ mathematical and statistical methods to explore data, potentially uncovering patterns that might be hidden to human experts and saving time.

However, the advent of AutoML has introduced a new set of considerations. While AutoML approaches offer the promise of systematic data exploration, questions remain about students' confidence in using the resulting features. Research indicates that even domain experts often struggle to make sense of AutoML-generated features [7], raising concerns about their practical applicability in educational contexts where transparency and understanding are paramount. In contrast, traditional expert-driven feature engineering relies heavily on domain knowledge and human intuition. This approach often yields features that are more contextually relevant and easier to interpret. However, it is not without its limitations, potentially constrained by cognitive biases [18] and the considerable time and effort required for manual feature selection. For instance, consider these examples in the context of predicting student academic performance:

Expert: score_higher_than_mean

Translation: Numbers of test scores students received that were higher than average of class

Featuretools: MAX(assessmentsmerged.CUM_COUNT (assessment_type))

Translation: Largest value of the count of how many assignments were submitted by the student so far

TSFRESH: mouse_click_left_agg_linear_trend_attr_"rvalue"_ _chunk_len_5__f_agg_"mean"

Translation: Correlation of a line drawn through the sequence of values that consist of the average number of mouse clicks in the last 5 actions done by the student

These three features demonstrate different levels of complexity in their calculations. The Expert feature is relatively straightforward, using a basic statistical comparison to count above-average test scores—a calculation that directly relates to familiar educational metrics. The Featuretools example introduces temporal aggregation by tracking cumulative assignment submissions over time, requiring more complex data aggregation across multiple database tables but still maintaining a connection to recognizable educational concepts. The TSFRESH feature is the most complex, incorporating sequence analysis of mouse behavior with both temporal chunking and correlation calculations. Each shows a distinct approach to capturing student behavior: direct performance metrics (Expert), engagement through submission patterns (Featuretools), and detailed interaction behavior through mouse movements (TSFRESH).

2.2 Trust

As a fundamental aspect of human–machine collaboration, trust requires close inspection with regard to its preconditions, implications, and consequences across multiple disciplines. In ML systems specifically, this manifests through dimensions of perceived competence, benevolence, and integrity [21]. Trust in ML involves a willingness to rely on the system's outputs and recommendations, based on positive expectations of its performance and reliability [42]. While high levels of trust can promote the adoption and effective use of ML technologies, excessive trust without critical evaluation may lead to overlooking potential biases or limitations inherent in these systems [83, 24, 37]. As ML continues to evolve, including the evolution of its strengths *and* weaknesses, striking the right balance of trust between humans and ML becomes crucial for ensuring both the efficacy and ethical implementation of these technologies. This is particularly important because students, especially younger ones, are often more susceptible to trusting technology due to their age and less experience evaluating automated recommendations [64, 15]. Their heightened trust makes them potentially more vulnerable to accepting ML-driven insights.

Trust formation is a complex process influenced by multiple factors. To address this complexity, researchers have proposed various frameworks to analyze and explain trust in automated systems. McKnight et al.'s [43] trust in technology model focuses primarily on institutional, calculative, and knowledge-based dimensions, which may not fully capture the experiential aspects of student interactions with ML features. Lee & See's [37] framework emphasizes performance, process, and purpose, providing valuable insights into the functional aspects of trust. Similarly, Mayer et al.'s [42] organizational trust model, while robust in addressing ability, benevolence, and integrity, was originally developed for human-human trust rather than human-machine interactions. The Hoff & Bashir [24] model emphasizes the significance of factors such as the machine's ability to engage socially, its overall appeal to users, and its communication approach [51]. By focusing on these aspects, the framework provides insights into how the design of machine learning systems can influence trust and interaction. Among these layers of trust, learned trust plays a pivotal role in shaping attitudes toward ML systems, developing through repeated interactions and accumulated experiences aligns particularly well with our focus on how different feature types influence trust formation and decision-making in educational ML systems. Research indicates that prior knowledge, expectations, observed performance, and perceived reliability

of ML models can significantly influence the trust formation process [67, 82].

Distinguishing between different aspects of trust is also crucial. Two key concepts are automation complacency and the propensity to trust technology. Automation complacency refers to suboptimal monitoring of model performance, potentially leading to overreliance on technology and failure to recognize situations requiring human intervention [44, 37]. It manifests as a trust behavior and emerges from prolonged exposure to consistently reliable ML systems [44, 52], potentially compromising users' critical thinking and decision-making abilities. In contrast, the propensity to trust technology is an individual's general tendency to rely on or believe in technological systems. It represents a more dispositional trait, reflecting an individual's general inclination to trust machines. Therefore, we employ a multifaceted approach from previous trust research to measure trust in ML-generated features within educational contexts [75]. First, *Appropriate Compliance* captures instances where students correctly accept or reject ML recommendations. Second, *Overcompliance* identifies cases where students accept incorrect ML recommendations, and serves as a behavioral indicator of overreliance, directly linking to the concept of automation complacency discussed earlier. Third, *Undercompliance* tracks situations where students reject correct ML recommendations, possibly suggesting under-trust. Additionally, we analyze *Decision Time*, which measures the duration students take to accept or reject ML recommendations. Furthermore, the *Switch Ratio*, which quantifies the frequency with which students change their mind from disagreement with the ML system at first to eventual agreement. Lastly, we adopt the Propensity to Trust[27] scale as a measure of non-behavioral level trust, providing insight into students' general inclination to trust ML systems.

3. THE CURRENT STUDY

The literature reviewed thus far underscores the complex interplay between student trust, system design, and the nature of features in ML models. It highlights a critical gap in understanding of how different feature engineering approaches—specifically, expert-crafted versus AutoML-generated features—might influence student trust and behavior in educational contexts. We focus on the context of predicting student outcomes, a critical application of ML in education. Our research is guided by the following research question and hypotheses:

> *Research question (RQ): In the context of predicting student outcomes, how do expert-crafted features compare to two AutoML-generated feature approaches (Featuretools and TSFRESH) in terms of student trust and propensity to use?*

H1: Expert-crafted features will elicit higher levels of initial trust compared to AutoML-generated features.

Reasoning: We expect that expert-crafted features are likely to be more aligned with domain knowledge and interpretable understanding of educational processes because they arise from human experts' knowledge. This alignment may lead to higher initial trust and more accurate decision-making by students [33]. Simultaneously, algorithm aversion—a tendency to distrust algorithmic decision-making after observing algorithmic errors—might lead to lower trust in AutoML-generated features [10], given that students may have experienced algorithmic errors in the past. This aversion could potentially result in relatively higher trust or appropriate compliance for expert-crafted features, as students may perceive human expertise as more reliable.

H2: AutoML-generated features will result in higher overcompliance rates compared to expert-crafted features.

Reasoning: The complexity and potential opacity of AutoML-generated features may lead to a form of automation bias, in which people often exhibit a tendency to trust computer-generated outputs more than human-generated ones due to their perceived sophistication [51, 69]. The automation bias stems from the perception that machine-generated information is more objective, comprehensive, and free from human error [66, 66]. This phenomenon might result in higher overcompliance rates for AutoML-generated features, despite possible initial lower trust hypothesized in H1. While initial trust might be lower for AutoML features (H1), overcompliance represents a different behavioral pattern where students may defer to suggestions even when they have doubts. This distinction between trust and potentially harmful overcompliance is crucial—trust, in general, reflects a considered judgment, while overcompliance indicates an excessive reliance that may override critical thinking and professional judgment.

H3: Expert-crafted features will lead to higher appropriate compliance compared to AutoML-generated features.

Reasoning: When features align with domain knowledge, students are more likely to make decisions that accurately reflect the system's reliability. This alignment in expert-crafted features may facilitate better judgment about when to trust or question the system's suggestions [33]. For example, research shows that when students rely purely on their personal experiences of learning (like expectations about workload patterns or forum participation), they sometimes interpret data very differently from statistical approaches and from each other [28]. The interpretability of expert features could also enable students to better calibrate their trust, leading to more appropriate compliance decisions.

H4: AutoML-generated features will result in higher undercompliance rates compared to expert-crafted features.

Reasoning: The complexity and unfamiliarity of AutoML-generated features may trigger skepticism and resistance, leading students to reject system suggestions even when they are correct. This cognitive barrier could stem from algorithm aversion [10], where the opacity of machine-generated features makes students more likely to dismiss valid system recommendations.

H5: Decision time will be shorter for expert-crafted features compared to AutoML-generated features.

Reasoning: The lexical familiarity and interpretability of

expert-crafted features may facilitate quicker decision-making processes. In contrast, the potentially novel or complex nature of AutoML-generated features might require more cognitive processing time [84].

H6: Students will demonstrate a higher propensity to trust expert-crafted features compared to AutoML-generated features, with this difference remaining consistent.

Reasoning: The familiarity and interpretability of expert-crafted features may sustain this trust advantage over time, as students might find these features consistently more relatable and understandable compared to potentially complex or opaque AutoML-generated features [71].

H7: The switch ratio will be higher for expert-crafted features compared to AutoML-generated features.

Reasoning: Consistent with our earlier hypotheses about the propensity to trust and decision time, students are likely to maintain higher trust in expert-crafted features throughout their interactions. This sustained trust in expert features may lead to more instances where students initially disagree with the system but then switch to agreement after further consideration [24].

In general, we hypothesize that features that are more interpretable or align closely with students' understanding of learning processes may foster greater trust. While highly abstract or complex features may demonstrate strong predictive power, their reduced interpretability may make them less effective at convincing students to thoughtfully reconsider their initial judgments when disagreements arise[35]. This is distinct from overcompliance patterns that develop over repeated interactions; here, we focus on those critical single instances where students reconsider their initial stance. By analyzing students' responses to these different feature types, we seek to uncover patterns in trust formation. Do students place more trust in models that use familiar, easily understood features? Or do they have more confidence in systems that employ sophisticated, machine-generated features that might be less interpretable? Understanding these dynamics is crucial for bridging the gap between the technical aspects of ML and the practical needs of students. Our findings inform the design of more trustworthy and effective educational ML systems, potentially leading to increased student engagement and improved learning outcomes.

4. METHOD

This research protocol was pre-registered through OSF(`https://doi.org/10.17605/OSF.IO/3H2GM`) and conducted under approval from the Institutional Review Board of the University of Illinois.

4.1 Data Preparation

We utilized two publicly available, de-identified educational datasets: the Open University Learning Analytics Dataset (OULAD) [36] and Educational Process Mining (EPM) data [74]. From these datasets, we extracted and designed variables related to student interactions and behavior within learning systems. While study participants only saw the plain English translations to ensure naming conventions didn't influence their judgments, the calculation difference vary substantially between these AutoML and expert-created features.

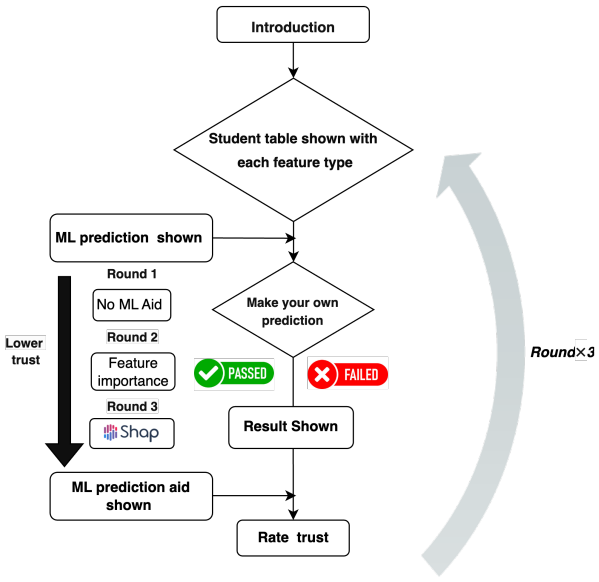

Figure 1: Research Procedure

.

The Open University Learning Analytics Dataset (OULAD) [36], collected between 2013 and 2014, captures student behaviors, interactions with course materials, and their final outcomes. OULAD includes information on 32,592 students, with over 10 million data points describing their interactions with an online learning system. The dataset contains student demographics, assessment results, and daily online activity logs. For our study, we simplified the original four-category student outcomes ("fail", "withdrawn", "pass", and "distinction") into two categories: "fail" and "pass". This binary classification makes it easier to interpret our machine learning models' results—especially for participants, who make judgments about pass/fail with the help of these machine learning models—and ensures consistency across different datasets.

The EPM dataset offers more granular student interface interaction data from 115 first-year undergraduate engineering students [74]. It contains 230,318 rows of data about their interactions with electronic tutoring software and assessment scores for different topics. Student outcomes in EPM range from 0 to 5 and were first converted into pass and fail for use in prediction.

4.2 Machine Learning Model Development

To ensure that the feature sets presented to participants were both fair and statistically valid, we selected feature sets that balanced predictive power with a degree of uncertainty. We built machine learning models using combinations of variables to determine their utility in predicting students' future academic outcomes. These models formed the basis for the predictions presented to participants in the subsequent survey experiments. The survey displayed predicted

outcomes from machine learning models to student participants. To create these predictions, we trained and applied random forest classification and regression models using the AutoML-generated features and the hand-crafted features separately, with data split into testing and training sets using a 30/70 ratio. For each of the three feature generation methods (Featuretools, TSFRESH, and expert-crafted), we selected 4 distinct sets, each comprising 5 variable combinations. These feature sets were selected to achieve a predictive accuracy within the range of 0.7 to 0.75. This specific range was chosen based on multiple studies that suggest a threshold of about 70% accuracy is appropriate for examining trust in decision-support systems incorporating AI [80, 83]. After that, we utilized SHAP (SHapley Additive exPlanations) [40] values to identify and visualize the importance of each feature within these sets, providing participants with graphical representations of how each variable contributes to the model's predictions. The purpose of including SHAP values was to provide participants with additional machine learning-supported information about feature importance. This inclusion is based on the premise that having more information to support decision-making can actually lead to less trust in the machine [75].

In our study, the decision to train a random forest model was driven by two main considerations. First, random forest models allow for fast and efficient SHAP value calculations [79, 40], which is crucial for our study design that incorporates feature importance visualization. This computational efficiency is not available with other popular models such as neural networks or support vector machines. Second, Random forest's robustness against overfitting makes it an appropriate choice for prediction tasks with many features [8], such as in this study. Moreover, our research specifically examines how different feature types influence student trust and decision-making, with the model architecture serving only as a means for this investigation. Thus, comparing different model types would extend beyond the scope of our core research questions about student trust. The focus was not on comparing model types but rather on highlighting the utility of the features within typical educational data mining contexts. By prioritizing feature selection, we aim to ensure that our study's findings are grounded in the quality and relevance of features, rather than the capabilities of any particular modeling technique.

4.3 Prediction Task and Trust Evaluation

Participants engaged in a multi-stage study to explore their use of different types of ML features for making decisions, and especially what their usage patterns might tell us about trust and reliance on features (Figure 1):

Initial Prediction: Participants were presented with a set of features (Figure 2), and then made an initial prediction of outcomes (Pass or Not Pass).

Model Prediction and Re-evaluation: Participants were shown the ML model's prediction and asked to make other predictions. Regardless of their decision on the first level, all participants progressed through the other two levels of prediction tasks.

Variables	Value
Average score student received from assessment	98
Clicks on the different sections in the course pages	16
Clicks on the online video discussions	6
Numbers of scores student received that were higher than average of class	1
Clicks on the online tutorial sessions	0

Figure 2: Screenshot of an example table of expert features shown to participants.

1. Level 1: Participants received basic information about the ML model and made predictions using a subset of variables.
2. Level 2: Additional insights into the ML model's predictions (variable importance) were provided, and participants made another set of predictions.
3. Level 3: Detailed graphical explanations (SHAP) of variables were given, and participants made final predictions.

Trust Assessment: Following each prediction task, participants rated their trust in the model's most recent prediction on a 5-point Likert-type question, ranging from 1 (not trustworthy at all) to 5 (completely trustworthy).

Participants were assigned to interact with three feature creation methods: Expert, TSFRESH, or Featuretools. They were presented with a series of recommendations and asked to make decisions based on these recommendations. Their interactions were recorded, including their decisions, decision times, and any changes in their decisions.

4.4 Trust Metrics

Our study employed trust-related behavioral measures to assess participants' trust in three different feature creation methods: Expert, TSFRESH, and Featuretools. To assess trust, we employed a set of behavioral measures derived from participants' interactions with system recommendations. Specifically, we focused on four key indicators that reflect different aspects of trust and decision-making behavior. Moreover, we measured trust compliance as the number of times participants followed the system's recommendations. This was further broken down into three specific metrics:

- **Appropriate Compliance**: The sum of correct recommendations accepted and incorrect recommendations rejected. This measure indicates the overall alignment between the participant's decisions and the system's accurate recommendations. A high value implies that participants are effectively discerning between accurate and inaccurate system recommendations, indicating a balanced level of trust. A low value suggests either indiscriminate acceptance or rejection of recommendations, pointing to potential issues in trust calibration [75].

- **Overcompliance**: The number of incorrect recommendations accepted. A high overcompliance rate suggests excessive trust in the system, potentially leading to errors and poor decision-making [75].

- **Undercompliance**: The number of correct recommendations rejected. High undercompliance might indicate a lack of trust in the system, even when it is providing accurate information. This could lead to missed opportunities for improved decision-making and underutilization of the system's capabilities [75].

- **Decision Time**

  We recorded the time taken by participants to accept or reject a recommendation, providing insight into the cognitive processing involved in trust-based decision-making. A shorter decision time might indicate higher trust in the system or a more interpretable understanding of the features, while longer decision times could suggest more careful consideration or uncertainty about the recommendation's validity.

- **Switch Ratio**

  The switch ratio was calculated to measure the frequency with which participants changed their initial disagreement with the system to eventual agreement. This ratio is expressed as a percentage and calculated using the following formula:

$$\text{Switch Ratio (\%)} = \frac{\text{Num. switches to agreement}}{\text{Num. disagree initially}} \times 100 \quad (1)$$

  A high switch ratio could imply that participants are willing to reconsider their initial judgments based on the system's recommendations, potentially indicating growing trust in the system over time. Conversely, a low switch ratio might suggest that participants are more confident in their initial judgments or less willing to rely on the system's recommendations, possibly indicating lower trust toward the system.

- **Initial Trust and Trust Difference** In addition to behavioral measures, we assessed participants' self-reported trust at the beginning of the interaction (Initial Trust) and the change in trust over the course of the interaction (Trust Difference).

5. RESULTS

5.1 Participants

Participants were recruited through Prolific, an online platform for research studies with exceptional data quality [55]. The inclusion criteria specified currently enrolled U.S. college students, at least 18 years old. A total of 179 participants completed the study. Participants received $5 base compensation for their participation. Additionally, to incentivize engagement and performance, participants were scored based on their prediction accuracy rate throughout the study. The accuracy rate was calculated as the percentage of correct predictions made by each participant across all tasks. Participants whose accuracy rates ranked in the top 10% of all participants received a bonus of $2.

5.2 Compliance Metrics

We employed linear mixed-effects models to analyze the data, with the feature creation method (Expert, TSFRESH, Featuretools) as a fixed effect and participant ID as a random effect. This approach allowed us to account for the effect of individual differences on multiple observations per participant, while examining the effect of the feature creation method on our trust-related measures. Figure 3 illustrates the compliance metrics across the three groups.

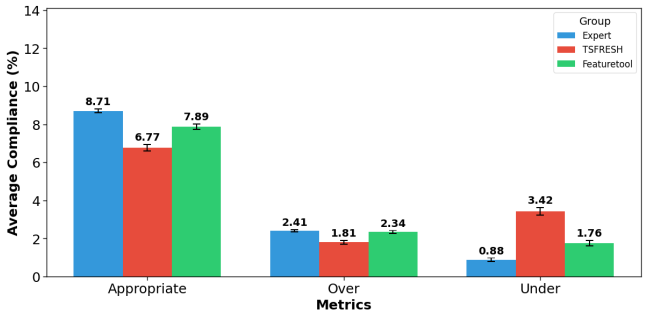

Figure 3: Compliance Metrics by Condition

5.2.1 *Appropriate Compliance*

Figure 3 shows that the Expert condition had the highest average appropriate compliance made by participants at 8.71 out of 36 rounds, followed by Featuretools at 7.89 and TSFRESH at 6.77. A linear mixed model analysis revealed significant differences among the conditions. The Expert condition showed higher appropriate compliance compared to both TSFRESH ($\beta = -0.615, p < .001$) and Featuretools ($\beta = -0.240, p = .002$). These results support H3, as the Expert condition showed significantly higher appropriate compliance.

5.2.2 *Overcompliance*

For overcompliance (Figure 3), participants averaged 2.41 incorrect acceptances with Expert features, 2.34 with Featuretools, and 1.81 with TSFRESH, out of the 9 rounds where the ML system made incorrect recommendations. The linear mixed effects model, with TSFRESH as the reference, showed significant differences. The Expert condition demonstrated significantly higher overcompliance compared to TSFRESH ($\beta = 0.598, p < .001$). Featuretools also showed higher overcompliance than TSFRESH ($\beta = 0.531, p < .001$). These results contradict H2, which predicted higher overcompliance for AutoML-generated features.

5.2.3 *Undercompliance*

Undercompliance was highest for TSFRESH at 3.42 correct recommendations rejected out of the 28 rounds where the ML system made correct predictions, followed by Featuretools at 1.76 and Expert at 0.88 (Figure 3). Significant differences were observed among conditions. Compared to the Expert condition, Featuretools showed higher undercompliance ($\beta = 0.201, p < .001$), as did TSFRESH ($\beta = 0.732, p < .001$). These results support H4, showing significantly higher undercompliance rates in both AutoML conditions compared to the Expert condition.

5.3 Trust Metrics

5.3.1 *Initial Trust*

Figure 4 shows the distribution of initial trust across conditions.

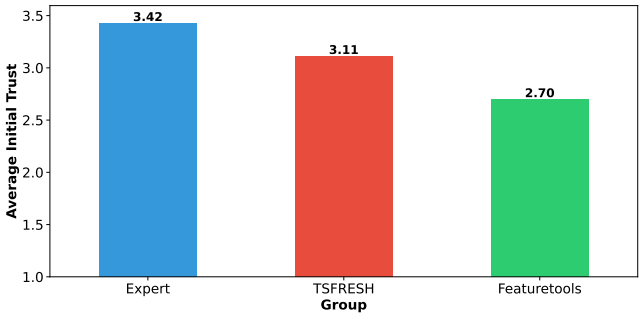

Figure 4: Distribution of Initial Trust by Condition

The Expert condition exhibited a higher median trust (3.52 out of 5) and a smaller interquartile range compared to TSFRESH and Featuretools, which both had median trust around 3.0 and larger interquartile ranges. The linear mixed model revealed significant differences in initial trust among conditions. Compared to the Expert condition, both Featuretools ($\beta = -0.315, p < .001$) and TSFRESH($\beta = -0.729, p < .001$) showed significantly lower initial trust. This finding supports H1, suggesting that expert-crafted features indeed elicit higher levels of initial trust and appropriate compliance.

5.3.2 *Trust Difference*

Analysis of trust differences between initial trust and post-trust did not reveal statistically significant changes across conditions. Neither Featuretools ($\beta = -0.017, p = .276$) nor TSFRESH ($\beta = 0.007, p = .705$) showed significant differences in trust change compared to the Expert condition. This finding partially contradicts H6, which predicted differences in trust levels between expert-crafted and AutoML-generated features over time.

5.4 Switch Ratio

Figure 5 presents the switch ratio by condition.

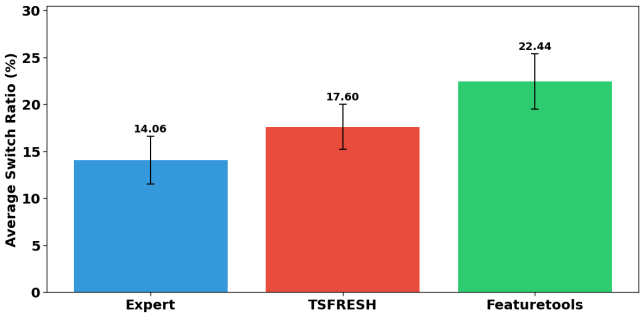

Figure 5: Switch Ratio by Condition

Featuretools had the highest switch ratio at 22.44%, followed by TSFRESH at 17.60%, and Expert at 14.06%. The linear mixed-effects model showed that Featuretools had a significantly higher switch ratio compared to Expert ($\beta = 8.380, p = .008$). TSFRESH, however, did not show a statistically significant increase compared to Expert ($\beta = 3.538, p = .064$), although the p-value suggests future work to confirm a trend in the same direction. This partially supports H7, with Featuretools exhibiting a higher switch ratio as predicted for AutoML-generated features, but no statistically significant difference for TSFRESH features.

5.5 Decision Time

Figure 6 illustrates the time differences by feature type.

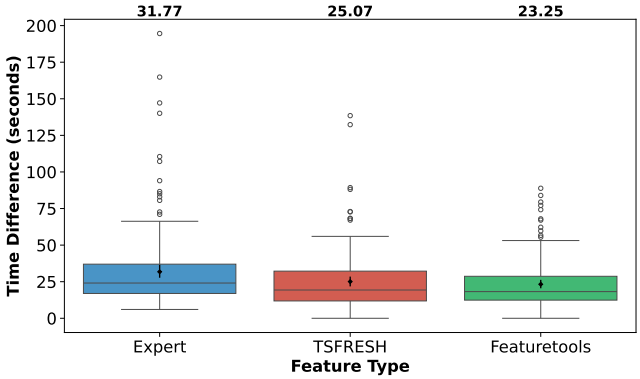

Figure 6: Decision Time by Feature Type

The Expert condition yielded the highest median decision time indicating that participants spent more time making decisions when presented with expert-crafted features. A linear mixed-effects model analysis revealed significant differences in time differences among the conditions. The Expert condition showed the highest time difference ($\beta = 38.980$, $p < .001$). Compared to the Expert condition, both Featuretools ($\beta = -12.839, p = .007$) and TSFRESH ($\beta = -10.000, p = .035$) demonstrated significantly lower decision time. This result contradicts H5, which predicted shorter decision times for expert-crafted features.

In summary, our results indicate significant differences in compliance metrics and initial trust across the three conditions, with the Expert condition generally showing higher appropriate compliance and initial trust. However, trust changes over time were not significantly different across conditions. The Featuretools condition demonstrated a significantly higher switch ratio compared to the Expert condition, while decision time differences showed some variation but with considerable overlap between conditions.

6. DISCUSSION

6.1 Compliance Patterns and Feature Interpretation

The observed differences in compliance across conditions provide insight into how students interpret and act upon different types of features. The Expert condition's higher appropriate compliance suggests that expert-crafted features may align more closely with students' mental models of what

predicts educational outcomes. This alignment could facilitate more accurate interpretations and decisions [34]. When features are grounded in educational domain knowledge and presented in familiar terms, students appear better equipped to make well-calibrated decisions about when to trust the system's suggestions (H3). This finding suggests that interpretability and domain alignment may serve as crucial mechanisms for fostering appropriate trust calibration.

Interestingly, the Expert condition also showed the highest overcompliance, followed closely by Featuretools and then TSFRESH. This pattern suggests that students may be more inclined to accept recommendations based on expert-created features, even when incorrect (H2). This concurrent finding might present an intriguing paradox. One possible explanation is that the familiarity and interpretability of expert features might create an "interpretability bias" where students become overconfident in their understanding of the system's decision-making process and become less likely to check for errors [17], leading to excessive trust. While this pattern might lead to better performance in cases where the machine learning model is quite accurate, it also highlights a potential risk of over-reliance due to a specific type of feature, which could propagate biases or errors present in expert-created features—particularly given that we matched accuracy across feature types to control for the effect of accuracy, indicating that overcompliance is not explained by higher accuracy. The Featuretools condition's close second in overcompliance suggests that its machine-generated features, while not matching expert features in perceived authority, still inspire a considerable degree of student confidence. This could be due to the features' apparent simplicity or relevance compared to TSFRESH, even if not fully understood by students.

Unlike overcompliance which reflects excessive trust, undercompliance indicates an ongoing trust deficit—where students continue to reject valid system recommendations due to their inability to build confidence in the system's decision-making process. The patterns of undercompliance, where students reject correct recommendations, also merit careful consideration. TSFRESH features in particular resulted in substantially more undercompliance. This could reflect a form of algorithm aversion where students are less willing to accept correct recommendations from machines compared to human experts [16], suggesting—along with appropriate compliance rates—that TSFRESH features are the least trusted and perhaps the most clearly "machine-like" in nature from students' perspectives (H4). This persistent skepticism likely originated from the features' computational complexity and lack of interpretable meaning, creating a barrier to trust development that persists throughout repeated interactions.

The consistently lower trust levels for machine-generated features might suggest that the opacity of AutoML features may create a "trust ceiling"—where students' trust remains consistently below optimal levels despite exposure and experience with the machine-created features. This finding is particularly important since it indicates that mere exposure to AutoML systems may not be sufficient to overcome trust barriers; instead, students may need additional support to develop appropriate levels of trust in these more complex, machine-generated features.

6.2 Trust Formation and Evolution

The significant differences in initial trust, with Expert features receiving higher trust ratings, reflect a persistent human preference for expert knowledge over machine-generated insights (H1). However, the lack of significant differences in trust change over time across conditions suggests that initial trust impressions are relatively stable. This stability in trust perceptions, despite exposure to features' performance in machine learning models, raises questions about the malleability of trust in machine systems and the potential need for more explicit trust calibration mechanisms [25].

6.3 Decision-making Processes

The time difference analysis reveals that participants spent significantly more time on Expert-created features compared to Featuretools and TSFRESH (H5). This increased deliberation time could indicate deeper cognitive processing of expert-crafted features, possibly due to their perceived complexity or relevance [54, 38]. The switch ratio analysis further complicates this picture, with Featuretools showing a higher propensity for students to change their initial disagreements (H7). This finding suggests that while machine-generated features may initially be viewed with skepticism, they have the potential to alter student judgments upon further consideration and interaction.

Overall, expert-created features consistently inspired higher levels of trust and compliance, highlighting the enduring value of domain expertise in machine system design. However, this also raised concerns about potential over-reliance on expert-created features. Machine-generated features, particularly those from Featuretools, showed promise in their ability to influence student decisions over time, as evidenced by higher switch ratios.

6.4 Implications for Student–AI Interaction Design

The results of the experiment in this paper have significant implications for the design of automated, machine learning-powered systems, especially in educational contexts where stakeholders (e.g., students, teachers, administrators, others) may rely on predictions of learning outcomes from a machine learning algorithm to help them make decisions exactly like those explored in this paper. The persistence of higher trust in expert-crafted features underscores the importance of incorporating domain expertise in feature engineering processes. However, the malleability of judgments regarding machine-generated features, as evidenced by the Switch Ratio, suggests potential for improving student acceptance of machine-generated insights. Future machine learning-powered systems might benefit from hybrid approaches that combine expert knowledge with machine learning capabilities, potentially leveraging the strengths of both to enhance student trust and system performance. Additionally, the development of more transparent and interpretable machine learning models could help bridge the trust gap between expert-crafted and machine-generated features, though only if interpretability extends beyond explanations in terms of feature values—which will only be as useful (and trusted) as the features are.

6.5 Limitations and Future Directions

While our findings demonstrate how students develop trust in AI systems, educational decision-making often involves more complex real-world contexts. The controlled nature of the experiment may not fully capture the complexities of real-world educational decision-making contexts. Future research could explore these dynamics in more naturalistic settings, potentially through longitudinal studies that track trust evolution over extended periods of interaction with AI-assisted decision-making. Furthermore, investigating the impact of different explanation strategies for machine-generated features could yield valuable insights into how to foster appropriate trust in machine systems. Exploring individual differences in trust propensity and machine literacy could also provide a more nuanced understanding of trust formation in student-AI collaborations.

7. CONCLUSION

As machine learning systems become increasingly important in educational contexts, understanding and addressing these dynamics will be crucial for developing systems that are not only accurate but also trustworthy and effectively utilized by human decision-makers. This experiment provides crucial insights into the dynamics of student trust and behavior in educational machine systems, specifically comparing expert-crafted features with those generated by AutoML tools. We reveal an important consideration: the origin of features significantly influences student trust, compliance, and decision-making processes.

8. REFERENCES

[1] O. Ali, P. A. Murray, M. Momin, Y. K. Dwivedi, and T. Malik. The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change*, 199:123076, Feb. 2024.

[2] Z. Alin, R. Sharma, D. Constantinescu, M. Pana, and C. Toma. Using learning analytics for analyzing students' behavior in online learning. *Studies in Informatics and Control*, 31:63–74, Sept. 2022.

[3] R. S. Baker and A. Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4):1052–1092, Dec. 2022.

[4] R. S. Baker and P. S. Inventado. Educational data mining and learning analytics. In J. A. Larusson and B. White, editors, *Learning Analytics: From Research to Practice*, pages 61–75. Springer, New York, NY, 2014.

[5] R. Ball, L. DuHadway, K. Feuz, J. Jensen, B. Rague, and D. Weidman. Applying machine learning to improve curriculum design. In *SIGCSE '19: Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, pages 787–793, Feb. 2019.

[6] F. Borgonovi and T. Burns. The educational roots of trust. Technical report, OECD, Paris, May 2015.

[7] N. Bosch. AutoML feature engineering for student modeling yields high accuracy, but limited interpretability. *Journal of Educational Data Mining*, 13(2):55–79, Aug. 2021.

[8] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[9] S. Caspari-Sadeghi. Applying learning analytics in online environments: measuring learners' engagement unobtrusively. *Frontiers in Education*, 7, Jan. 2022. Publisher: Frontiers.

[10] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.

[11] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, Mar. 2017.

[12] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Muller, L. Ju, and H. Su. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 297–307, Mar. 2020.

[13] B. Drăgulescu and M. Bucos. Hyperparameter ttuning using automated methods to improve models for predicting student success. In A. Lopata, R. Butkienė, D. Gudonienė, and V. Sukackė, editors, *Information and Software Technologies*, Communications in Computer and Information Science, pages 309–320, Cham, 2020. Springer International Publishing.

[14] A. Elsayed, M. Caeiro Rodriguez, F. Mikic Fonte, and M. Llamas Nistal. Research in learning analytics and educational data mining to measure self-regulated learning: A systematic review. In *Proceedings of the 18th World Conference on Mobile and Contextual Learning (mLearn 2019)*, pages 46–53, Sept. 2019.

[15] N. Ezer, A. D. Fisk, and W. A. Rogers. Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human Factors*, 50(6):853–863, Dec. 2008.

[16] I. Filiz, J. R. Judek, M. Lorenz, and M. Spiwoks. The extent of algorithm aversion in decision-making situations with varying gravity. *PLOS ONE*, 18(2):e0278751, Feb. 2023.

[17] C. J. Fitzsimmons, C. A. Thompson, and P. G. Sidney. Confident or familiar? The role of familiarity ratings in adults' confidence judgments when estimating fraction magnitudes. *Metacognition and Learning*, 15(2):215–231, Aug. 2020.

[18] C. G. Harris. Mitigating cognitive biases in machine learning algorithms for decision making. In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 775–781, New York, NY, USA, 2020. Association for Computing Machinery.

[19] T. Gervet, K. Koedinger, J. Schneider, and T. Mitchell. When is deep learning the best approach to knowledge tracing? *Journal of Educational Data Mining*, 12(3):31–54, Oct. 2020.

[20] A. Gosiewska, A. Kozak, and P. Biecek. Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, 150:113556, Nov. 2021.

[21] P. S. Greenberg, R. H. Greenberg, and Y. L. Antonucci. Creating and sustaining trust in virtual teams. *Business Horizons*, 50(4):325–333, July 2007.

[22] Z. Han, J. Wu, C. Huang, Q. Huang, and M. Zhao. A review on sentiment discovery and analysis of educational big-data. *WIREs Data Mining and*

Knowledge Discovery, 10(1):e1328, 2020.

[23] A. B. Hernández-Lara, A. Perera-Lluna, and E. Serradell-López. Applying learning analytics to students' interaction in business simulation games. the usefulness of learning analytics to know what students really learn. *Computers in Human Behavior*, 92:600–612, Mar. 2019.

[24] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, May 2015.

[25] C. Hoffner and M. Buchanan. Young adults' wishful identification with television characters: the role of perceived similarity and character attributes. *Media Psychology*, 7(4):325–351, Nov. 2005.

[26] H.-N. Huang, H.-M. Chen, W.-W. Lin, C.-J. Huang, Y.-C. Chen, Y.-H. Wang, and C.-T. Yang. Employing feature engineering strategies to improve the performance of machine learning algorithms on echocardiogram dataset. *Digital Health*, 9, Oct. 2023.

[27] S. A. Jessup, T. R. Schneider, G. M. Alarcon, T. J. Ryan, and A. Capiola. The Measurement of the Propensity to Trust Automation. In J. Y. Chen and G. Fragomeni, editors, *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, pages 476–489, 2019.

[28] L. Jiang and N. Bosch. Mining and assessing anomalies in students' online learning activities with self-supervised machine learning. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, pages 549–554, 2022.

[29] N. A. Johar, S. N. Kew, Z. Tasir, and E. Koh. Learning analytics on student engagement to enhance students' learning performance: a systematic review. *Sustainability*, 15(10):7849, Jan. 2023.

[30] J. M. Kanter and K. Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, Oct. 2015.

[31] C. Kern, F. Gerdon, R. L. Bach, F. Keusch, and F. Kreuter. Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, 3(10):100591, Oct. 2022.

[32] S. N. Kew and Z. Tasir. Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data. *Technology, Knowledge and Learning*, 27(2):405–427, June 2022.

[33] B. Kim, J. Shah, and F. Doshi-Velez. Mind the Gap: a generative approach to interpretable feature selection and extraction. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2260–2268, Cambridge, MA, USA, 2015.

[34] C. Klein, J. Lester, H. Rangwala, and A. Johri. Learning analytics tools in higher education: Adoption at the intersection of institutional commitment and individual action. *The Review of Higher Education*, 42(2):565–593, 2019.

[35] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, 12, 2021.

[36] J. Kuzilek, M. Hlosta, and Z. Zdrahal. Open University Learning Analytics dataset. *Scientific Data*, 4:170171, Nov. 2017.

[37] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, Mar. 2004.

[38] M. L. Lewis and M. C. Frank. The length of words reflects their conceptual complexity. *Cognition*, 153:182–195, Aug. 2016.

[39] H. Luan, P. Geczy, H. Lai, J. Gobert, S. J. H. Yang, H. Ogata, J. Baltes, R. Guerra, P. Li, and C.-C. Tsai. Challenges and Future Directions of Big Data and Artificial Intelligence in Education. *Frontiers in Psychology*, 11, Oct. 2020.

[40] S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions, Nov. 2017.

[41] F. Martin and D. Bolliger. Engagement matters: student perceptions on the importance of engagement strategies in the online learning environment. *Online Learning*, 22(1):205–222, Mar. 2018.

[42] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3):709–734, 1995.

[43] D. H. McKnight, L. L. Cummings, and N. L. Chervany. Initial trust formation in new organizational relationships. *The Academy of Management Review*, 23(3):473–490, 1998.

[44] S. M. Merritt, A. Ako-Brew, W. J. Bryant, A. Staley, M. McKenna, A. Leone, and L. Shirase. Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology*, 10, 2019.

[45] C. Mills, J. Gregg, R. Bixler, and S. K. D'Mello. Eye-mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction*, 36(4):306–332, July 2021.

[46] B. M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5):527–539, Nov. 1987.

[47] L. Márquez, V. Henríquez, H. Chevreux, E. Scheihing, and J. Guerra. Adoption of learning analytics in higher education institutions: A systematic literature review. *British Journal of Educational Technology*, 55(2):439–459, 2024.

[48] T. Nazaretsky, M. Ariely, M. Cukurova, and G. Alexandron. Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 53(4):914–931, 2022.

[49] S. Niedlich, A. Kallfaß, S. Pohle, and I. Bormann. A comprehensive view of trust in education: Conclusions from a systematic literature review. *Review of Education*, 9:124–158, Nov. 2020.

[50] E. M. Onyema, K. K. Almuzaini, F. U. Onu, D. Verma, U. S. Gregory, M. Puttaramaiah, and R. K. Afriyie. Prospects and challenges of using machine learning for academic forecasting. *Computational Intelligence and Neuroscience*, 2022:5624475, June

2022.

[51] R. Parasuraman and C. A. Miller. Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47(4):51–55, 2004.

[52] R. Parasuraman and V. Riley. Humans and Automation: Use, Misuse, Disuse, Abuse - Raja Parasuraman, Victor Riley, 1997.

[53] A. Pardo, J. Jovanovic, S. Dawson, D. Gašević, and N. Mirriahi. Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, 50(1):128–138, 2019.

[54] S. Paul and D. L. Nazareth. Input information complexity, perceived time pressure, and information processing in GSS-based work groups: An experimental investigation using a decision schema to alleviate information overload conditions. *Decision Support Systems*, 49(1):31–40, Apr. 2010.

[55] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4):1643–1662, Aug. 2022.

[56] G. Ramaswami, T. Susnjak, and A. Mathrani. Effectiveness of a learning analytics dashboard for increasing student engagement levels. *Journal of Learning Analytics*, 10(3):115–134, Dec. 2023. Number: 3.

[57] F. M. Rohrhofer, S. Saha, S. D. Cataldo, B. C. Geiger, W. v. d. Linden, and L. Boeri. Importance of feature engineering and database selection in a machine learning model: a case study on carbon crystal structures, Jan. 2021.

[58] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura. Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146, 2013.

[59] C. Romero and S. Ventura. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, Nov. 2010.

[60] T. Rotsaert, E. Panadero, and T. Schellens. Anonymity as an instructional scaffold in peer assessment: its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education*, 33:75–99, Jan. 2018.

[61] R. Salazar, F. Neutatz, and Z. Abedjan. Automated feature engineering for algorithmic fairness. *Proc. VLDB Endow.*, 14(9):1694–1702, 2021.

[62] P. Schmidt and F. Biessmann. Quantifying Interpretability and Trust in Machine Learning Systems, Jan. 2019.

[63] C. Schumacher and D. Ifenthaler. Features students really expect from learning analytics. *Computers in Human Behavior*, 78:397–407, Jan. 2018.

[64] K. L. Seaman, A. P. Christensen, K. D. Senn, J. A. Cooper, and B. S. Cassidy. Age-related Differences in the Social Associative Learning of Trust Information. *Neurobiology of aging*, 125:32, Feb. 2023.

[65] A. Sjöblom, A. Silvola, and J. Lallimo. How deployment processes affect the adoption of learning analytics in higher education institutions: improving potential for impact with better deployment practices, Aug. 2021.

[66] L. J. Skitka, K. L. Mosier, and M. Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, Nov. 1999.

[67] R. Spain. The Effects of Automation Expertise, System Confidence, and Image Quality on Trust, Compliance, and Performance. *Psychology Theses & Dissertations*, July 2009.

[68] M. Strier and H. Katz. Trust and parents' involvement in schools of choice. *Educational Management Administration & Leadership*, 44(3):363–379, May 2016.

[69] S. S. Sundar and J. Kim. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–9, New York, NY, USA, 2019. Association for Computing Machinery.

[70] T. Susnjak, G. S. Ramaswami, and A. Mathrani. Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19(1):12, Feb. 2022.

[71] L. Tang and N. Bosch. Can students understand AI decisions based on variables extracted via AutoML? In *Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3342–3349, Piscataway, NJ, 2024. IEEE.

[72] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 272–283, Barcelona Spain, Jan. 2020. ACM.

[73] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos. Implementing AutoML in educational data mining for prediction tasks. *Applied Sciences*, 10(1):90, Jan. 2020.

[74] M. Vahdat, L. Oneto, D. Anguita, M. Funk, and M. Rauterberg. A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In G. Conole, T. Klobučar, C. Rensing, J. Konert, and E. Lavoué, editors, *Design for Teaching and Learning in a Networked World*, Lecture Notes in Computer Science, pages 352–366, Cham, 2015. Springer International Publishing.

[75] O. Vereschak, G. Bailly, and B. Caramiaux. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):327:1–327:39, 2021.

[76] D. West, H. Huijser, and D. Heath. Putting an ethical lens on learning analytics. *Educational Technology Research and Development*, 64(5):903–922, 2016.

[77] Z. Wu, T. He, C. Mao, and C. Huang. Exam paper generation based on performance prediction of student group. *Information Sciences*, 532:72–90, Sept. 2020.

[78] Yasmin. Application of the classification tree model in predicting learner dropout behaviour in open and

distance learning. *Distance Education*, 34(2):218–231, Aug. 2013.

[79] M. Yağcı. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1):11, Mar. 2022.

[80] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[81] M. R. Young. The art and science of fostering engaged learning. *The Academy of Educational Leadership Journal*, Nov. 2010.

[82] N. Yuviler-Gavish and D. Gopher. Effect of descriptive information and experience on automation reliance. *Human Factors*, 53(3):230–244, June 2011.

[83] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, New York, NY, USA, 2020. Association for Computing Machinery.

[84] A. Zytek, I. Arnaldo, D. Liu, L. Berti-Equille, and K. Veeramachaneni. The need for interpretable features: motivation and taxonomy. *ACM SIGKDD Explorations Newsletter*, 24(1):1–13, 2022.