JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 11(2), 73-90. https://doi.org/10.18608/jla.2024.8253

The Impact of Attribute Noise on the Automated Estimation of Collaboration Quality Using Multimodal Learning Analytics in Authentic Classrooms

Pankaj Chejara[1], Luis P. Prieto[2], Yannis Dimitriadis[3], María Jesús Rodríguez-Triana[4], Adolfo Ruiz-Calleja[5], Reet Kasepalu[6], Shashi Kant Shankar[7]

Abstract

Multimodal learning analytics (MMLA) research has shown the feasibility of building automated models of collaboration quality using artificial intelligence (AI) techniques (e.g., supervised machine learning (ML)), thus enabling the development of monitoring and guiding tools for computer-supported collaborative learning (CSCL). However, the practical applicability and performance of these automated models in authentic settings remains largely an under-researched area. In such settings, the quality of data features or attributes is often affected by noise, which is referred to as attribute noise. This paper undertakes a systematic exploration of the impact of attribute noise on the performance of different collaboration-quality estimation models. Moreover, we also perform a comparative analysis of different ML algorithms in terms of their capability of dealing with attribute noise. We employ four ML algorithms that have often been used for collaboration-quality estimation tasks due to their high performance: random forest, naive Bayes, decision tree, and AdaBoost. Our results show that random forest and decision tree outperformed other algorithms for collaboration-quality estimation tasks in the presence of attribute noise. The study contributes to the MMLA (and learning analytics (LA) in general) and CSCL fields by illustrating how attribute noise impacts collaboration-quality model performance and which ML algorithms seem to be more robust to noise and thus more likely to perform well in authentic settings. Our research outcomes offer guidance to fellow researchers and developers of (MM)LA systems employing AI techniques with multimodal data to model collaboration-related constructs in authentic classroom settings.

Notes for Practice

- Multimodal learning analytics offers a holistic approach to estimate collaboration levels and thus can potentially help teachers to facilitate computer-supported collaborative learning.
- The random forest algorithm is found to be more robust to attribute noise for collaboration-quality estimation tasks.
- The decision tree algorithm could offer a better option for modelling collaboration quality with attribute noise by providing a potential explanation for its results.
- Audio-based collaboration-quality estimation models suffer from 50% degradation in performance as a result of attribute noise in audio features.

Keywords

Multimodal learning analytics, MMLA, collaboration, collaborative learning, computer-supported collaborative learning, CSCL, machine learning, ML

Submitted: 27/09/2023 — **Accepted:** 06/05/2024 — **Published:** 20/06/2024

Corresponding author[1] *Email: pankajch@tlu.ee Address: Center of Educational Technology, room A-317, Narva mnt. 29, Tallinn University, Tallinn, Estonia. ORCID iD: https://orcid.org/0000-0002-7630-5789*

[2] *Email: luispablo.prieto@uva.es Address: GSIC/EMIC Research Group, Paseo de Belén, 15 (Room 2L019), Universidad De Valladolid, Valladolid, Spain. ORCID iD: https://orcid.org/0000-0002-0057-0682*

[3] *Email: yannis@tel.uva.es Address: GSIC/EMIC Research Group, Paseo de Belén, 15 (Room 2L019), Universidad De Valladolid, Valladolid, Spain. ORCID iD: https://orcid.org/0000-0001-7275-2242*

[4] *Email: mjrodtri@uva.es Address: GSIC/EMIC Research Group, Paseo de Belén, 15 (Room 2L019), Universidad De Valladolid, Valladolid, Spain. ORCID iD: https://orcid.org/0000-0001-8639-1257*

[5] Email: adolfo@tlu.ee Address: Center of Educational Technology, room A-317, Narva mnt. 29, Tallinn University, Tallinn, Estonia. ORCID iD: https://orcid.org/0000-0003-1717-6304
[6] Email: reetkase@tlu.ee Address: Center of Educational Technology, room A-317, Narva mnt. 29, Tallinn University, Tallinn, Estonia. ORCID iD: https://orcid.org/0000-0003-3389-8673
[7] Email: shashikant.shankar@ammachilabs.org Address: AMMACHI Labs, Amrita Vishwa Vidyapeetham, Kollam, Kerala, India. ORCID iD: https://orcid.org/0000-0001-8266-3681

1. Introduction

Collaboration is a key 21st-century skill (Voogt & Roblin, 2012), and researchers have found evidence of its positive impact on learning (Laal & Ghodsi, 2012). To develop this skill among students, merely putting them into collaborative learning activities is not sufficient. Teachers, apart from designing the collaborative activity, often need to support students during collaborative learning, which requires monitoring the group's activities to estimate the quality of collaboration in each group and intervening when collaboration problems are detected. However, the estimation of collaboration quality in authentic learning contexts is extremely difficult (Schwarz & Asterhan, 2011) given the multiplicity of learner groups working in the classroom and collaboration's inherently multidimensional nature (e.g., argumentation, mutual understanding) (Rummel et al., 2011).

Automated estimation of collaboration aspects has gained significant traction from researchers due to its potential to support teachers in authentic classroom settings (Schneider et al., 2021). Researchers have demonstrated the feasibility of building models capable of making predictions of high-level collaboration measures, e.g., collaboration quality (Martínez-Maldonado, 2011), confusion (Ma et al., 2023), and rapport (Lubold & Pon-Barry, 2014). To adequately estimate collaboration quality, the rich and multimodal nature of group interactions (e.g., verbal communication, gestures, digital interactions) needs to be taken into account. Consequently, researchers have utilized a variety of data (e.g., audio, video) to build automated estimation models of collaboration quality (Praharaj et al., 2021; Schneider et al., 2021). This field of research, utilizing multiple data sources to understand collaboration (and learning in general), is known as multimodal learning analytics (MMLA) (Ochoa & Worsley, 2016). Researchers have suggested the use of multimodal data to understand complex learning processes in a more holistic way (Drachsler & Schneider, 2018; Worsley & Blikstein, 2018). Researchers have employed simple sensor-based quantitative features (e.g., speaking time, the distance between hands), as well as complex ones with a focus on discourse (e.g., cohesion of speech text, Reilly & Schneider, 2019). These features, along with others, have enabled researchers to build dashboards to support teachers in monitoring group behaviour in their classrooms.

MMLA researchers often collect data using a wide range of data collection devices, e.g., microphones, cameras, heart-rate monitors (Di Mitri et al., 2018; Schneider et al., 2021). The use of these devices often involves a setting-up phase prior to actual data collection. This phase is where researchers configure the device and make sure that it is ready for data collection (e.g., calibrating eye-gaze trackers on the spot). During the data collection phase, these devices need to be used properly (e.g., limited movement while wearing skin-conductance sensors) to minimize errors in the collected data, which is likely to affect the performance of developed machine learning (ML) models (Nettleton et al., 2010). In laboratory settings, where researchers have greater control than in real-world settings, complying with the setup and usage guidelines of multimodal devices can be ensured with higher success. Furthermore, laboratory settings allow the collection of multimodal data with only a few devices (e.g., two eye-trackers) by running the study in multiple batches of participants (e.g., one dyad at a time). This approach permits researchers to use fewer higher-end data collection devices, which are often expensive, to ensure minimal noise.

In contrast, the classroom environment may impede the use of higher-end multimodal devices due to reasons including higher financial cost—due to the need for multiple sensor units (Yan et al., 2022); increased complexity—associated with the setup of multiple data collection devices; room conditions (e.g., student distribution); and intrusiveness of sensors, which may hamper the learning process (Chua et al., 2019; Darvishi et al., 2022). Furthermore, achieving full compliance from participants in terms of device use, such as minimizing movement around a microphone array, may not always be feasible in classroom settings. As a consequence, data collected in authentic classroom settings often suffers from lower quality due to various sources of noise. This decline in data quality is also reflected in the features extracted from different data sources, which are used to construct ML models, such as those for collaboration-quality estimation. Furthermore, the limitations of feature extraction technology (e.g., computer vision–based emotion detection, speech-to-text conversion) add one more source of noise to the extracted features. Consequently, all of these factors, including students' behaviour, hardware limitations, and feature computation errors, affect the quality of features (or attributes) used by automated models to estimate collaboration quality. The aggregate of noise from different sources has been termed "attribute noise" in the ML literature (Nettleton et al., 2010). As a result of attribute noise, the performance of ML models (especially in authentic situations) is often affected (Nettleton et al., 2010).

To bring the benefits of automating collaboration-quality estimation (in the form of monitoring/guiding tools) to the classroom and to improve the reliability of such models, there is a need to investigate how collaboration-quality estimation

models perform in the presence of attribute noise. Additionally, understanding which ML algorithms excel under these conditions is essential. This line of research is of high relevance to the MMLA (and learning analytics (LA) more generally) research fields in order to bridge the gap between research and practice. Furthermore, this also expands our understanding of the noise-handling capability of artificial intelligence (AI) in the educational context. To systematically study the impact of attribute noise in authentic settings, this paper investigates how the performance of collaboration-quality estimation models developed using multimodal data (concretely, audio and log data) is affected in the presence of different types of attribute noise (e.g., noise in all attributes from a specific data source) in different proportions. Moreover, we also analyze the performance of collaboration-quality estimation models developed with ML algorithms that are widely used for similar estimation tasks (Smith et al., 2016; Spikol et al., 2017; Martínez-Maldonado, 2011) to ascertain their noise-handling abilities. For our investigation, we used multimodal datasets collected from 11 different learning contexts during face-to-face collaborative learning activities in Estonian vocational school classrooms. For our research study, we considered learning context as consisting of multiple aspects (e.g., students, teacher, learning activity, subject); thus, change in any of these aspects has been taken into account as a criterion for differences among learning contexts, e.g., differences in tasks as a basis of criteria for task contexts as in Pugh and colleagues (2022).

The rest of the paper is structured in seven sections. In Section 2, we provide background information on collaboration quality, data quality, and attribute noise. Section 3 offers ML research on how attribute noise impacts models' performance and on the state of the art of collaboration estimation using MMLA. Section 4 presents our research questions for the study, and Section 5 explains the research methods and datasets used in the paper. In Section 6, we present the analysis results. Section 7 discusses the results, their implications, and the limitations of the present study. Finally, Section 8 concludes the paper with our plan for future studies.

2. Background

This section provides a theoretical background of collaboration quality and its underlying aspects. Additionally, the section also offers an introduction to attribute noise.

2.1 Collaboration Quality

Collaboration is defined as "a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem" (Roschelle & Teasley, 1995). In general terms, collaboration happens when two or more individuals work together toward a common goal. There are multiple underlying aspects to collaboration (e.g., communication or coordination). To understand these aspects of collaboration in depth, researchers from the learning sciences have proposed several frameworks (Cukurova et al., 2018; Rummel et al., 2011; OECD, 2017). For instance, Cukurova and colleagues (2018) proposed a framework to analyze non-verbal behaviour through observations to understand different aspects of collaborative problem solving (e.g., mutual understanding). One of the more widely used frameworks in MMLA to characterize these underlying aspects was initially proposed by Meier and colleagues (2007) and later extended by Rummel and colleagues (2011). This framework specifies five aspects of successful collaboration: communication, joint information processing, coordination, interpersonal relationships, and motivation. The communication aspect emphasizes the importance of how participants communicate their ideas and maintain a unified understanding of the problem/solution. This aspect has been divided into two dimensions, namely, sustaining mutual understanding and collaboration flow. These dimensions focus on participants' having a common understanding and having a "coherent sequence" in their exchanges. The second aspect, joint information processing, focuses on sharing ideas, questioning, and answering during the group activity. This aspect has been divided into two dimensions: knowledge exchange and argumentation. The first dimension focuses on the epistemic side of a group's conversation, while the second dimension captures the process of question answering of the group's participants to reach a consensus. The third aspect, coordination, focuses on the execution of the given task in a timely and orderly manner. This aspect has a single dimension of structuring the problem-solving process, which emphasizes the group's strategy to solve the given problem (e.g., planning of group activities). The fourth aspect, interpersonal relationships, refers to the social plane of group interactions involving relationships between group members. This aspect has one dimension: cooperative orientation, which focuses on group dynamics (e.g., how welcoming the group is for others' ideas/suggestions). The fifth aspect, motivation, focuses on an individual's motivation for working toward completing the given task. This aspect has one dimension: individual task orientation, which, in contrast to the aforementioned dimensions, focuses on the individual rather than the group. These seven dimensions enable a quantitative assessment of collaboration quality, which further opens the door for building support systems for teachers to automatically estimate collaboration quality. While other frameworks have been proposed in the literature to model collaboration (e.g., the NISPI framework from Cukurova et al. (2018) and the PISA framework from OECD (2017)), the way Rummel and colleagues (2011) break down collaboration quality into seven dimensions has enabled the use of intervention strategies from the learning sciences (Kasepalu et al., 2022), which makes the chosen framework

suitable for developing automated support systems for practitioners. This paper models collaboration quality by analyzing the aforementioned seven dimensions.

2.2 Attribute Noise

A widely used definition of quality data is "data fit for use" (Wang & Strong, 1996) that is "free of defects and possess[es] desired features" (Redman, 2001). Accordingly, the quality of attributes (i.e., features extracted from various data) is determined by "how well the attributes characterize instances for modelling purposes (e.g., classification)" (Zhu & Wu, 2004). This quality, however, can be compromised in the presence of noise[1] during the phases of data acquisition and data pre-processing (Nettleton et al., 2010). The effects of these different sources of noise on feature-/attribute-level data have been termed attribute noise (Zhu & Wu, 2004). In a more general sense, attribute noise is defined as "anything which obscures the relationship between attributes and class" (or target variables, e.g., level of collaboration quality) (Hickey, 1996) (as cited in Nettleton et al., 2010).

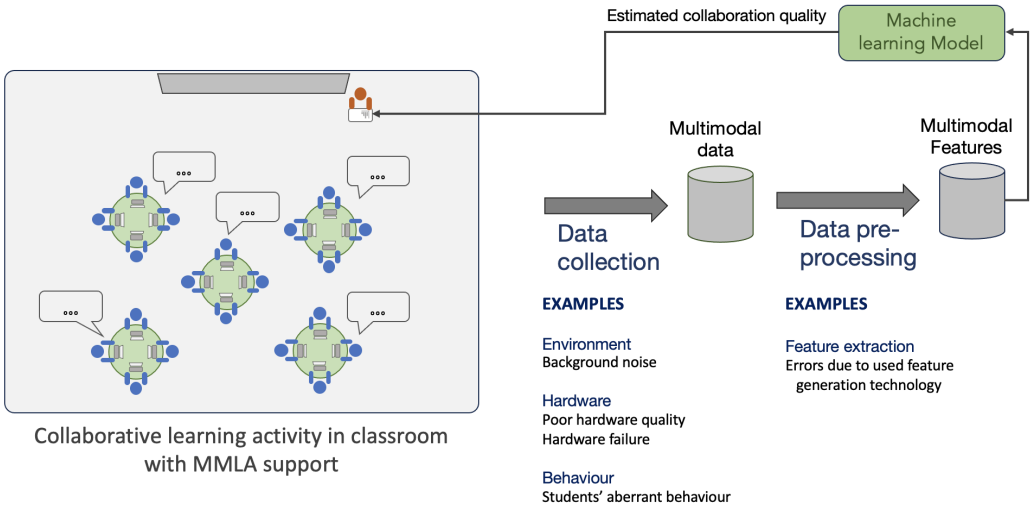

Figure 1. Attribute Noise Factors in Authentic Classroom Settings

We use a hypothetical generic example of the use of ML models in classrooms to explain the notion of attribute noise in the context of collaboration-quality estimation. Figure 1 illustrates the functioning of an ML model that is deployed in a classroom to estimate collaboration quality during authentic learning situations. This process contains data collection and pre-processing of features, which are then fed to an ML model for estimation. During these two phases, i.e., data collection and pre-processing, various factors can give rise to attribute noise. Drawing from MMLA, the ML literature, and our prior engagements in MMLA studies conducted within authentic contexts (Chejara et al., 2024; Chejara, Prieto, et al., 2023; Chejara, Kasepalu, Prieto, Rodríguez-Triana, Ruiz-Calleja, & Shankar, 2023), we have identified four distinct categories of factors. These factors pertain to the environment, hardware, participants' behaviour, and feature computation technology. We have categorized these factors in alignment with the phases of data collection and data pre-processing, which are the stages where attribute noise is known to occur (Nettleton et al., 2010).

- **Environmental factors:** These include causes of attribute noise originating from the surrounding atmosphere. For example, background noise degrading audio data quality (Smith et al., 2016) and improper lighting in the classroom cause low-quality video recording.
- **Hardware-related factors:** These are mainly related to data-recording devices and include malfunctioning equipment, improper configuration of equipment (Kalapanidas et al., 2003), and poor-quality hardware.
- **Behavioural factors:** These are associated with participants' behaviour. An example is when students show exceptional or aberrant behaviour (Sonnenberg & Bannert, 2015) (e.g., an unconscious tapping on the table, which muddles the audio from the mike) causing the model—which learned a relationship between speaking time based on audio energy level and collaboration quality—to perform worse. Another example is when students do not follow specific setup instructions during multimodal data collection (e.g., students moving around a microphone array, causing wrong detection of speaking behaviour). Previous research has also provided preliminary evidence on biased results of LA when there is noise in data due to behavioural factors (Alexandron et al., 2019).

[1] a random error or unexplained variance in a measured variable (Han & Kamber, 2000)

- **Feature computational factors:** These concern the technology used for feature extraction purposes and originate from the technology's limitations. For example, automated facial feature detection technology, while highly sophisticated, is unlikely to yield features with absolute precision. Consequently, the use of such technology introduces an additional error component to the attribute noise.

3. Related Work

This section provides an overview of ML research studies that have investigated attribute noise's impact on model performance. The section next describes state-of-the-art research employing ML to build automated models of collaboration quality (and various aspects of collaboration, e.g., task performance) using multimodal data along with research studies investigating the impact of noise on automated models for collaboration.

3.1 ML Research on How Attribute Noise Impacts Model Performance

Several ML studies have investigated the impact of attribute noise on model performance. Kalapanidas and colleagues (2003) investigated how attribute noise affects regression and classification models developed in the context of air quality modelling (e.g., for toxicity detection). Their results from five-fold cross-validation (CV) showed that a linear regression model performed comparably better in dealing with noisy data. Melville and colleagues (2004) investigated the noise sensitivity of ensemble classifiers (e.g., bagging and boosting). Their results found that the bagging algorithm performed better in the presence of noise. Zhu and Wu (2004) did an extensive study by exploring various noise proportions in training and test data. Their work looked into the accuracy measures of a single ML algorithm on open-source datasets. Nettleton and colleagues (2010) compared the performance of four ML algorithms in dealing with attribute noise. The authors generated a synthetic dataset to study the impact of attribute noise on ML models developed using naive Bayes, decision tree, support vector machine (SVM), and instance-based learning. The noise in their study was manually added in different proportions to input data as well as class labels to create different cases for the study. Their study found naive Bayes to be the most robust algorithm in the case of attribute noise. The aforementioned studies have shown that attribute noise has a considerable impact on the model's performance, although the impact was not uniform across different ML algorithms. Moreover, these studies have also indicated that the robustness of a particular ML algorithm to attribute noise depends on the specific nature of the dataset being processed.

The study of impact has also been conducted in domains, e.g., audio-video speech recognition (Papandreou et al., 2009) and medical diagnosis (Reyes-García et al., 2018), as cited in Ma and colleagues (2023). For example, Reyes-García and colleagues (2018) investigated the impact of noise in terms of missing values (or null values) on the performance of the model for patient physiological health deterioration. They manually added null values to the data in different proportions (i.e., 25%, 50%, 75%). Their results reported that model performance was above 0.90 for 25% missing values.

3.2 MMLA Research on Collaboration-Quality Estimation

MMLA researchers have explored the use of various types of data sources from the physical space for modelling collaboration, e.g., video (Spikol et al., 2018), audio (Ponce-Lopez et al., 2013), and eye-gaze (Schneider & Pea, 2013). Certain research work has even investigated the feasibility of real-time prediction of collaboration processes in terms of group collaboration quality (Martinez-Maldonado et al., 2015). Researchers have employed various ML algorithms to build automated models for collaboration quality, group task performance, and group artifact quality, e.g., naive Bayes, decision tree, best first tree, AdaBoost, SVM, neural network, random forest, logistic regression, K-means, and Res-NET (Martinez-Maldonado et al., 2013; Spikol et al., 2018; Reilly & Schneider, 2019; Chejara et al., 2021). Random forest was frequently reported to outperform other ML algorithms (e.g., SVM, naive Bayes) for estimating collaboration quality (Reilly & Schneider, 2019; Viswanathan & Vanlehn, 2018). For evaluation, researchers have mainly employed k-fold CV (Spikol et al., 2018; Martinez-Maldonado et al., 2013) and rarely evaluated their models across different contexts (Pugh et al., 2022). In educational terms, k-fold CV evaluation only estimates the performance in authentic learning contexts similar to the one used during model training. However, if we aim for a wider adoption of ML solutions in real practice, where the dataset will most likely come from a different context than the one used for training, then we need to go beyond single-context evaluations. This issue has also been raised in a generalizability evaluation framework of MMLA (EFAR-MMLA), suggesting the evaluation of a model's performance on higher levels of generalizability (e.g., group, context) (Chejara et al., 2021).

MMLA research works on collaboration estimation have shown the feasibility of building automated collaboration models. The current research work has achieved moderate (70% accuracy) to high (90% accuracy) performance (Ponce-Lopez et al., 2013; Viswanathan & Vanlehn, 2018). However, the majority of these MMLA studies on collaboration have been conducted in laboratory settings (Chua et al., 2019). A recent MMLA review reported an increasing number of MMLA studies being conducted in authentic settings (approximately 50%), though this is not the case when considering MMLA studies on group collaboration (only four out of 100 reviewed studies, Yan et al., 2022).

The noise issues inherent to the use of multiple data sources (e.g., technical issues, hardware failure) and authentic classroom settings (e.g., audio from other students) make it of high relevance for research fields (e.g., MMLA, LA, computer-supported collaborative learning (CSCL)) that use AI in educational contexts to develop models that can deal with noise (Crescenzi-Lanna, 2020; Emerson et al., 2020). However, there is a lack of research in this direction, which can potentially limit the transition of research to practice (Crescenzi-Lanna, 2020). In their study, Ma and colleagues (2023) showed the negative impact of noise on the prediction of confusion and conflict moments during collaborative learning. Their study developed ML models using features extracted from audio and video data (e.g., speech and facial features) and investigated the models' performance. Their results showed that the noise in audio led to a substantial decrease in the performance accuracy from 0.79 to 0.61. This study offers preliminary evidence of the negative impact of noise on the performance of ML models using multimodal data for collaborative learning activities. While predicting collaborative problem-solving aspects with large language models, Pugh and colleagues (2021) reported a performance decrease from 0.83 to 0.78 when models were applied in authentic noisy classroom settings.

Furthermore, the findings from ML research on noise's impact on performance may not be directly applicable to the aforementioned research fields (e.g., MMLA, CSCL) because the datasets used are from non-educational contexts. Thus, their findings may not hold, given that noise impact is conditioned by the datasets used (Nettleton et al., 2010). Besides, in educational settings, the developed ML models are likely to perform on data coming from a different learning context. Therefore, the knowledge of attribute noise's impact on performance across context is also needed in addition to impact over within-context performance. However, there is still a research gap on how attribute noise will impact the model's performance in authentic educational contexts and, more specifically, how noise affects performance *across* different learning contexts.

4. Research Questions

Based on the aforementioned state of the art in ML and MMLA, this study explores the following research questions in the context of collaboration-quality estimation:

RQ1: How do different types and different proportions of attribute noise in test data impact the collaboration-quality estimation models' performance within and across context? This question aims to investigate how the performance of the collaboration-quality model is affected when it is used for estimation tasks with attribute noise emulated by manually introducing it in the test data set. This case reflects an education situation where an ML model is trained using high-quality data, which is collected from the classroom and has gone through a data-cleaning phase (e.g., discarding data with missing values, handling outliers) and upon deployment is unlikely to have the same quality of data. Thus, the investigation assesses the potential of existing models in authentic scenarios (where the aforementioned factors are likely to affect the quality of attributes).

RQ2: Which ML algorithms show more robust performance in the presence of attribute noise in test data for the task of estimating collaboration quality? This question entails gaining insights into the performance of ML algorithms for collaboration-quality estimation in the presence of noise. Recognizing that performance is not the sole measure of success and that there is often a trade-off between performance and explainability, we also explore the explainability aspect while investigating the noise-handling capabilities of the ML algorithms used.

5. Materials and Methods

In this section, we present the multimodal datasets used in the study, and methods for generating attribute noise, building collaboration-quality estimation models, and assessing the performance of the developed models. The datasets were collected as part of our previous work, where we developed collaboration-quality estimation models and investigated teachers' perception of the use of collaboration monitoring systems (Chejara, Prieto, et al., 2023; Kasepalu et al., 2023). We decided to use these datasets because of the unavailability of a benchmark multimodal dataset of collaboration quality in the field due to ethical concerns.

5.1 Study Context

The datasets for the study were collected from an Estonian vocational school in a total of 11 classroom sessions (Please refer to Table 8 in Appendix A for more details.) These sessions involved five different subject teachers, six classrooms, and 105 students. These subjects were English language, Estonian language, woodwork with integrated chemistry, mathematics, and communication. The participants were older than 18 years of age and had prior experience working in group activities. The participants were of Estonian background and used the Estonian language for communication during the learning activities. Each session was 30 to 60 minutes long and involved face-to-face interaction as well as using a collaborative text editor (i.e., Etherpad). Before the activity, the teacher formed groups of two to four based on the available number of students. During

the activities, the teacher was present in the classroom and carried out routine tasks involving monitoring groups' activities and offering help when needed. Students were in the same classrooms and instructed to use headphones primarily capturing high-quality audio through microphones. Teachers seated students based on the classroom's logistics. Participants from the same group sat close to each other. Figure 2 shows a group of students during the activity.

Figure 2. Typical Classroom Setup Used for Data Collection in the Present Study

5.2 Datasets

We collected 11 multimodal datasets (audio and log) corresponding to the 11 sessions in an Estonian vocational school where the collaborative learning activities took place. The data were collected using CoTrack (Chejara et al., 2024), which is a web-based multimodal data collection tool. CoTrack provided data features through its integrated pre-processing functionality, including voice activity detection, speech recognition using Google Speech-to-Text, and writing analytics. Out of this pre-processed data, we extracted speaking time; turn-taking; frequency of "I," "You," and "We"; frequency of wh words (e.g., what, why); and the number of characters added or deleted as the key features (i.e., attributes) for model development. Such simple quantitative measures, which do not require sophisticated discourse analysis, have been used by prior researchers for collaboration-quality estimation (Martínez-Maldonado, 2011; Ponce-Lopez et al., 2013). Due to the simplicity of these features, they may be easily applied and adopted by practitioners. Moreover, these features, despite being simple, have been demonstrated to achieve across-context generalizability for collaboration-quality estimation in current research (Chejara, Kasepalu, Prieto, Rodríguez-Triana, Ruiz Calleja, & Schneider, 2023). Our decision to use these features is thus based on prior research in MMLA for CSCL. Table 1 provides further details on related research studies with features chosen in the presented study. Please refer to Appendix B for further details.

Regarding audio data collection, each participant was wearing headphones to allow the audio data to be as noise-free as possible. We further excluded 10 groups whose audio data was either missing or corrupted due to hardware issues. For ground truth, four graduate masters students were first trained using the rating scheme from Rummel and colleagues (2011) and then assigned videos for annotation. The rating scheme involves assigning scores on a 5-point scale to seven dimensions of collaboration quality: argumentation, sustaining mutual understanding, knowledge exchange, collaboration flow, cooperative orientation, structuring problem solving and time management, and individual task orientation. The inter-rater reliability score for all the aforementioned dimensions using the rating scheme was at a substantial level (Kappa ≥ 0.65). The scores of the dimensions were added and the average was taken to get a measure of overall collaboration quality, following similar work in MMLA (Martínez-Maldonado, 2011).

5.3 Methods

We present methods for emulating different types of attribute noise in the dataset, and for building and evaluating ML algorithms for collaboration-quality estimation.

5.3.1 Generating Attribute Noise

To emulate attribute noise, we adopted a systematic approach proposed by Nettleton and colleagues (2010) (Algorithm 1). This approach involves a two-step process: First, a specific number ($x\%$) of cases is randomly selected using a uniform distribution. Then, for each selected case, the value of each attribute is replaced with an error value. This error value is initially drawn from a

Table 1. Data Features

Data type	Feature name	Description	Related studies
Audio	speaking_time	Speaking time in seconds	MMLA research studies have found speaking time and turn-taking as indicators of collaboration (Martinez-Maldonado et al., 2013; Ponce-Lopez et al., 2013).
	turn_taking	Number of speaking turns taken by the participant	
	freq_I, freq_you and freq_we	Frequency of "I," "You," and "We"	Storch (2001) found differences between high- and low-collaborating groups in terms of their use of personal pronouns.
	freq_wh	Frequency of wh-words (e.g., what, why, who, where, how)	Question-asking behaviour indicates argumentative nature of collaboration as per rating scheme of Rummel and colleagues (2011).
Log	chars_add chars_del	Number of characters added Number of characters deleted	These are indicators of individual participation, which have been found as one of the key quantitative metrics for collaborative learning (Weinberger & Fischer, 2006).

standard Gaussian distribution and then scaled to match the range of the corresponding attribute. Our choice to use a Gaussian distribution for introducing attribute noise was informed by the central limit theorem, which posits that "the normalized sum of random variables follows a Gaussian distribution" (Peebles, 1987). In our study, we conceptualized attribute noise as the result of the accumulation of multiple error components introduced during the phases of data collection and data pre-processing, influenced by various factors (e.g., environmental and hardware-related factors). For example, consider a scenario where audio data was collected with background noise and later analyzed using a feature computation technique, which added another error component given the limitation of the technique. The attribute noise in this case can be successfully modelled using our conceptualization and the property of the central limit theorem. This conceptualization, however, is limited in terms of modelling attribute noise to the cases where multiple factors contribute to the noise. For example, this conceptualization is inapplicable to the scenarios where a single factor influences the data quality.

Algorithm 1: Add *x*% noise in dataset D

```
Input: Multimodal dataset D
Output: Multimodal dataset with attribute noise
cases ← select x% cases from D using uniform distribution
foreach case in cases do
    foreach attribute in case do
        Compute the range of attribute
        Draw a random value from standard Gaussian distribution
        Scale the random value to the attribute range
        Replace the attribute's value with the scaled value
    end
end
```

We considered three different types of scenarios in our study:

- **Single-source attribute noise:** This scenario represents a situation where the noise is present in attributes that are extracted from a specific data source. Several factors may introduce noise in audio or log data. For example, attributes from audio data are likely to have a higher degree of noise given the poor quality of classroom audio data. While the noise[2] in audio data is likely to occur in classrooms, the log data is less prone to becoming noisy. However, there is still a possibility due to factors related to the classroom environment, hardware, and students' unexpected behaviour. For instance, in classroom settings, sometimes there are fewer devices (e.g., laptops, tablets) than students, and the sharing of devices may obscure collaboration dimensions, like individual task orientation, as captured by the log data. Students' unexpected behaviour, e.g., extensive use of copy/paste, may also obscure the measurement of participation from log data

[2]Here, we'd like to remind the reader that our criterion for noise is "... anything which obscures the relationship between features and target label" (Hickey, 1996).

Table 2. Example of Noisy Audio Attributes at 20% and 100% Levels (erroneous values are presented in red)

Cases	Original dataset		20% noisy dataset		100% noisy dataset	
	speaking_mean	turn-taking_mean	speaking_mean	turn-taking_mean	speaking_mean	turn-taking_mean
1	3.9	2	3.9	2	26.1	7.4
2	2.5	1	12.6	2.4	9.3	4.4
3	5.4	2	5.4	2	1.7	.9
4	2.5	1	2.5	1	6.2	10.5
5	3.1	1	3.1	1	0.7	6.5

(e.g., a single keystroke transforms into a spike on log features like "number of characters written"). Another example of log noise is when the laptop fails to connect to the Etherpad or the server fails to record log data.

To emulate this noise scenario, we added noise first in features extracted from audio data only, and then in features extracted from log data only. This step resulted in two[3] sets of datasets: one with audio noise and another with log noise. We added different proportions of noise (0.10, 0.20, 0.30, 0.40, 0.50) following studies on ML in similar investigations (Zhu & Wu, 2004). We emphasize here that the noise levels were *relative*, not absolute. For example, an audio noise level of 0.10 represented a dataset in which 10% of its cases had their original (cleaned) audio attributes' values replaced with erroneous values. To clarify this concept further, consider Table 2 as an illustrative example. The dataset presented therein comprises five cases and two attributes extracted from audio data, specifically the mean of speaking time and the mean of turns taken. In a 20% noisy audio dataset, one out of five cases has erroneous values for both attributes (i.e., 12.6 and 2.4).

- **Single-source missing attributes:** This scenario represents the situation when the complete data from a specific source is not available due to a hardware malfunction. For this scenario, we first replaced all audio attributes' values with erroneous values (i.e., 100% noise), and then we repeated this process for log attributes' values. The last two columns of Table 2 illustrate how the dataset appears after the introduction of 100% noise. Importantly, we chose to completely randomize attribute values by introducing 100% noise rather than discarding these features. This choice was made to maintain uniformity in the number of input features, ensuring consistency between the training and testing phases of the developed models.

- **Multiple-source attribute noise:** This scenario represents a situation where different factors simultaneously cause noise in attributes of multiple data sources. We emulated this by simply using the aforementioned algorithm to add noise to all the features of the datasets. We added different proportions of noise in all features, as in the single-source attribute noise scenario.

5.3.2 Building Collaboration-Quality Classification Models Using Different ML Algorithms

We employed four ML algorithms, namely, random forest, AdaBoost, naive Bayes, and decision tree, for two reasons. First, these algorithms have been used frequently and found to achieve high performance at the task of estimating collaboration quality (or its aspects) with multimodal data (Martinez-Maldonado et al., 2013; Spikol et al., 2017; Smith et al., 2016; Chejara et al., 2021). Second, ML research on noise handling has found that these algorithms offer the best performance with noisy data in contexts other than education (Melville et al., 2004; Nettleton et al., 2010).

Our model development pipeline included five steps: data scaling, outlier handling, hyper-parameter optimization, threshold selection, and use of contextual features (e.g., the total number of students in the classroom). Our decision to use this particular configuration was based on our previous study where we explored various configurations of pipelines toward building generalizable collaboration-quality estimation models (Chejara, Prieto, et al., 2023). We used Kappa as a performance metric of our models, which takes into account how well the model is performing for positive (e.g., high) and negative (e.g., low) labels of collaboration quality. The metric has been used by other researchers for similar modelling problems (Viswanathan & Vanlehn, 2018).

[3] Our datasets also had video data but that was meant for annotation purposes only.

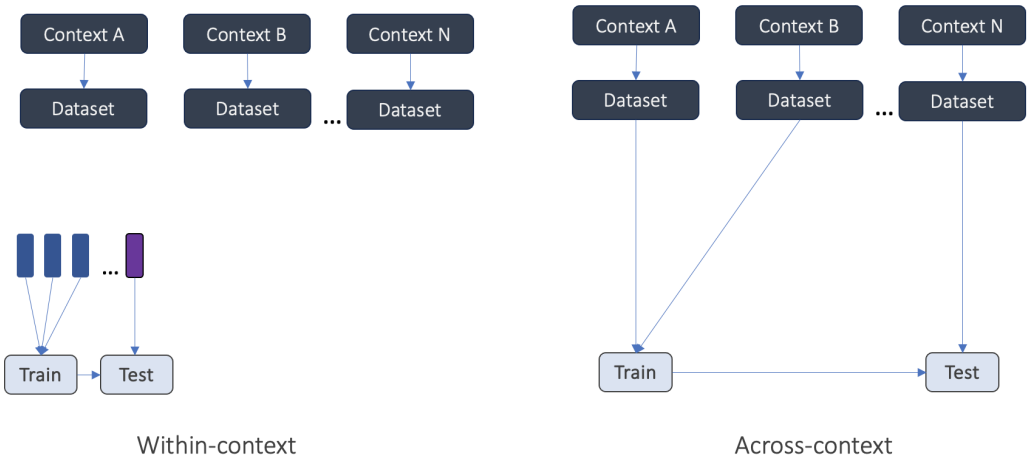

Figure 3. Model Evaluation (in within-context evaluation, smaller boxes represent partitions of the dataset after dividing it under 10-fold CV)

5.3.3 Evaluating Model Performance

For model evaluation, we used 10-fold CV and leave-one-context-out evaluation schemes (Figure 3). The first evaluation method, i.e., 10-fold CV, estimates how well a model performs on data coming from a similar data distribution or, in educational terms, from a similar learning context. We used the performance measure from 10-fold CV as a measure of within-context performance. The leave-one-context-out evaluation method assesses the model's performance on a dataset that is coming from a different learning context than the one used during the training phase. Thus, it provides a measure of across-context performance, which is a more realistic evaluation, because once a model is put into practice, the model will perform on data from a new context each time it is used. In both evaluations, the test set was induced with the attribute noise with different proportions using the aforementioned noise-generation method.

5.3.4 ML Algorithms' Effectiveness at Handling Attribute Noise

To study which ML algorithm is better at dealing with attribute noise in all three aforementioned noise scenarios, we followed the same approach as suggested by Nettleton and colleagues (2010). We trained models using all four algorithms and assigned a rank score as per their across-context performance on noisy data (e.g., the lower the rank, the higher the performance). For example, random forest performed comparatively better than other algorithms on noisy audio data (single-source attribute noise scenario) at leave-one-context-out evaluation; therefore, we assigned it rank 1. Finally, we took the average and standard deviation of the model's ranks across different noise configurations (as in Nettleton et al., 2010).

6. Results

The results are structured according to the research questions posed in Section 4.

6.1 Model Performance in the Presence of Attribute Noise (RQ1)

We present here performance measures of collaboration-quality estimation models with attribute noise in the aforementioned three different scenarios, namely, single-source attribute noise, single-source missing attributes, and multiple-source attribute noise. The performance measures are reported at two levels of generalizability, denoted as "within" and "across," signifying within-context and across-context evaluations, respectively. Additionally, the performance metrics are reported across various levels of attribute noise. Here, we would like to remind the reader again that the noise levels in this study are not absolute. For example, a noise level of 0% does not imply that the data is completely noise-free; rather it indicates that the quality of attributes remains consistent between the training and test sets. Conversely, a noise level of 10% signifies that the test dataset contains 10% more noise than the training dataset.

Model performance in single-source attribute noise Table 3 shows the models' performance on data with noise in log attributes. At 0% noise level, random forest and decision tree performed better than other algorithms within and across context. Random forest achieved the highest performance of Kappa = 0.44 and Kappa = 0.27 (with 0% attribute noise in test data) when evaluated within the context and across different contexts, respectively. Decision tree also achieved a similar performance with a slight difference, e.g., 0.43 and 0.24 at within-context and across-context generalizability levels. Overall, noise in the log data seems to only slightly affect the performance of the random forest, decision tree, and naive Bayes models' performance (i.e.,

Table 3. Collaboration-Quality Model Performance (mean±standard deviation) in Single-Source (log) Attribute Noise Scenario (higher performance measures are represented in bold text) DT: Decision Tree, NB: Naive Bayes, RF: Random Forest, AB: AdaBoost

Model	Attribute noise (all log features at a time)											
	0%		**10%**		**20%**		**30%**		**40%**		**50 %**	
	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across
DT	.43±.07	.24±.20	**.43±.07**	.23±.21	.43±.07	.23±.21	.42±.06	.24±.19	.43±.06	.23±.20	.43±.06	.23±.21
NB	.18±.06	.09±.10	.17±.07	.08±.10	.17±.07	.08±.10	.17±.06	.07±.10	.15±.06	.08±.10	.18±.09	.10±.12
RF	**.44±.08**	**.27±.20**	**.44±.07**	**.26±.1**	**.46±.08**	**.28±.19**	**.46±.08**	**.28±.20**	**.44±.07**	**.27±.19**	**.44±.08**	**.26±.21**
AB	.40±.05	.21±.19	.39±.06	.20±.19	.36±.07	.19±.18	.36±.09	.21±.16	.34±.08	.18±.16	.30±.09	.18±.15

reduction in Kappa in the range of 0.01 to 0.03) at both levels of generalizability. Among these three, naive Bayes showed a slightly larger decline in the within-context performance for 40% noise level, going from Kappa = 0.18 (at 0% noise) to 0.15 (at 40% noise). Comparatively, the AdaBoost models experienced more degradation (i.e., reduction in Kappa from 0.40 to 0.30) in their within-context performance when performed on test data with 50% attribute noise.

Table 4 presents the models' performance with test data where only audio-based attributes were induced with noise. Decision tree and random forest showed a slight decline in their within-context performance, i.e., a reduction of 0.04 to 0.05 in Kappa at 50% attribute noise in comparison with Kappa at 0% noise. Even with 50% audio attribute noise, random forest and decision tree achieved a performance of 0.40 and 0.38 Kappa within context, respectively. Naive Bayes performed poorly in comparison with others but the performance was only slightly decreased within and across context with an increase in audio attribute noise. AdaBoost suffered the most in terms of performance at both levels of generalizability, e.g., at 50% attribute noise, there was a degradation of 0.16 and 0.07 in Kappa from the initial within-context (Kappa = 0.40) and across-context (Kappa = 0.21) performance, respectively. Random forest achieved the highest across-context performance for different proportions of audio attribute noise.

Table 4. Collaboration-Quality Model Performance in Single-Source (audio) Attribute Noise Scenario

Model	Attribute noise (all audio features at a time)											
	0%		**10%**		**20%**		**30%**		**40%**		**50 %**	
	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across
DT	.43±.07	.24±.20	.42±.07	.22±.21	.40±.07	.23±.18	.40±.07	.21±.20	.38±.07	.19±.18	.38±.08	.21±.18
NB	.18±.06	.09±.10	.18±.06	.09±.10	.17±.04	.09±.09	.14±.07	.07±.09	.15±.07	.09±.08	.13±.09	.07±.09
RF	**.44±.08**	**.27±.20**	**.43±.09**	**.26±.19**	**.41±.09**	**.24±.18**	**.40±.07**	**.24±.18**	**.39±.08**	**.23±.18**	**.40±.08**	**.22±.19**
AB	.40±.05	.21±.19	.38±.04	.22±.17	.32±.05	.17±.15	.31±.06	.16±.16	.27±.06	.14±.13	.24±.04	.14±.12

Model performance in single-source missing attributes scenario Table 5 shows the performance of models for different cases of missing complete data from a specific source (audio, log, or both). Our decision tree model of collaboration quality emerged as a high-performing model for dealing with missing audio data (emulated by adding 100% attribute noise in all audio attributes), achieving 0.33 and 0.16 Kappa at within-context and across-context generalizability, respectively. For the same case of missing audio, naive Bayes and AdaBoost degraded to chance model performance (Kappa = 0 or close to zero) on both generalizability levels. For missing log data, random forest outperformed other algorithms, achieving 0.46 and 0.26 Kappa at the within-context and across-context generalization, respectively. Decision tree also performed close to random forest for missing log data, achieving Kappas of 0.41 and 0.23, respectively. In the case of missing audio and log data, all the models became chance models, reaching a Kappa of zero (or close to zero). This is due to the loss of all information contained in the attributes used by models.

Model performance in multiple-source attribute noise scenario Table 6 shows the performance of collaboration-quality models on test data having different proportions of noise added to both audio and log features. As the noise increased in the data, all models experienced a decline in their performance (as expected). The pattern of performance degradation was similar among random forest, AdaBoost, and decision tree. Naive Bayes, on the contrary, showed relatively lower degradation. However, its performance was the lowest for all cases in both noisy and non-noisy data. Overall, random forest performed better than others for attribute noise $\leq 30\%$, while, for higher noise levels ($> 30\%$), AdaBoost outperformed (or achieved the same performance as) random forest on the levels of within context and across context. At the 50% noise level, all high-performing models showed close to 50% loss in their performance (as measured by Kappa).

Table 5. Collaboration-Quality Model Performance in Single-Source Missing Attributes

Model	Attribute noise (missing data emulated using 100% noise)					
	Audio missing		**Log missing**		**Both missing**	
	Within	Across	Within	Across	Within	Across
DT	**.33±.09**	**.16±.18**	.41±.06	.23±.20	.01±.05	.02±.07
NB	0±.00	0±.05	.06±.05	.04±.06	0±0	0±0
RF	.29±.11	.15±.16	**.46±.07**	**.26±.21**	.04±.08	.02±.06
AB	.03±.02	.01±.03	.17±.11	.10±.11	.03±.04	.02±.05

Table 6. Collaboration-Quality Model Performance in Multiple-Source Attribute Noise Scenario (noise in both audio and log)

Model	Attribute noise (all feature)											
	0%		**10%**		**20%**		**30%**		**40%**		**50 %**	
	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across
DT	.43±.07	.24±.20	.36±.05	.20±.17	.31±.07	.18±.15	.27±.08	.16±.14	.22±.07	.12±.12	.19±.08	.11±.10
NB	.18±.06	.09±.10	.17±.07	.09±.09	.14±.06	.06±.09	.17±.05	.07±.10	.12±.06	.06±.09	.11±.07	.06±.08
RF	**.44±.08**	**.27±.20**	**.40±.08**	**.26±.19**	.33±.05	**.20±.15**	**.31±.06**	**.20±.13**	.24±.07	**.14±.13**	.20±.07	.12±.12
AB	.40±.05	.21±.19	.38±.04	.19±.18	**.34±.06**	.18±.16	.30±.04	.17±.14	**.27±.05**	**.14±.14**	**.22±.08**	**.13±.12**

6.2 ML Algorithms' Capabilities in Dealing with Attribute Noise (RQ2)

Table 7 shows the average ranking (and standard deviation) of algorithms on the task of estimating collaboration quality across different types of noisy scenarios (single-source attribute noise, single-source missing attributes, and multiple-source attribute noise). Random forest emerged as the most noise resistant algorithm among the four algorithms used. The rank was in the range of 1 to 1.3 (with smaller variation), which can be interpreted as random forest outperforming other algorithms in terms of working with attribute noise, in almost every noise scenario considered in this study. Decision tree, despite being a comparatively simple algorithm, achieved a rank in the range of 2 to 2.3, making it a strong candidate for modelling with attribute noise. Naive Bayes was found to achieve the lowest performance, reflected in its rank of 4 across different proportions of attribute noise.

Table 7. Ranking of Algorithms (average rank ± standard deviation)

Model	Attribute noise (all feature)											
	0%		**10%**		**20%**		**30%**		**40%**		**50 %**	
	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across
DT	2±0	2±0	2.3±.4	2.1±.3	2.3±.4	2.1±0.3	2±0.8	2.1±0.6	2.3±0.4	2.3±0.4	2.3±0.4	2.3±0.4
NB	4±0	4±0	4±0	4±0	4±0	4±0	4±0	4±0	4±0	4±0	4±0	4±0
RF	**1±0**	**1±0**	**1±0**	**1±0**	**1.3± .4**	**1.1±0**	**1±0**	**1±0**	**1.3±0.4**	**1.1±0.3**	**1.3±0.4**	**1.1±0.4**
AB	3±0	2.8±0	2.6±.4	2.6±.4	2.3±.9	2.5±0.7	2.6±0.4	2.6±0.4	2.3±0.9	2.3±0.9	2.3±0.9	2.3±0.9

7. Discussion

This section discusses the implications of the results presented above for the research questions we set for the study.

7.1 RQ1: How Do Different Proportions of Attribute Noise in Test Data Impact the Collaboration-Quality Estimation Models' Performance Within and Across Context?

Finding 1: The attribute noise in log data seems to have minimal impact on the performance of collaboration-quality estimation models at the within-context and across-context levels of generalizability. Our results showed that the noise in the log data features in general had a minimal impact on the models' performance. The reason for the minimal impact of noise on the models' performance could be explained by the importance of the log-based features in the different collaboration estimation tasks. We performed an impurity-based feature importance analysis and found that our random forest model had speaking time–based features (average and standard deviation) with the highest feature importance. In contrast, all other features, including log-based features, had comparatively lower (50% or less) feature importance. This could be due to the face-to-face nature of collaborative learning activities: the verbal interaction between the participants was higher than their interaction with the collaborative writing editor (i.e., Etherpad). It was also found in the analysis of audio and log features that

almost 46% of log features were sparse, while sparsity for audio features was 11%. This suggests more face-to-face interaction than writing. The face-to-face nature, sparsity of log features, and comparatively low feature importance might have led models (except AdaBoost) to learn more from audio features. This may have contributed to the minimal impact of log noise on the within-context and across-context performance. Consequently, the noise in log data was handled with graceful degradation by the models we developed.

Finding 2: Noise in audio data seems to cause a higher degradation in the performance of the AdaBoost model. We expected a severe degradation in all models' performance due to audio attribute noise. However, only AdaBoost and naive Bayes showed that degradation. On the contrary, random forest and decision tree performed with only a small degradation in their performance. Moreover, the random forest model was affected the least by noise at the within-context level, while decision tree was the least affected model at the across-context level.

The findings on the higher performance of random forest for estimating collaboration quality with the presence of attribute noise are aligned with ML research on the model's performance with attribute noise (Melville et al., 2004). Notably, our findings diverge from previous studies that found AdaBoost as efficient as bagging-based methods (e.g., random forest) for handling attribute noise. The discrepancy could be attributed to the distinct nature of our educational datasets compared to those (i.e., automobiles, disease) in prior studies, potentially influencing the impact of attribute noise (Nettleton et al., 2010).

Finding 3: Noise affecting all features from audio and log led to severe degradation in the performance of all models except naive Bayes. The degradation in the performance of developed models was expected as a result of attribute noise in audio and log data features. All models showed degradation in their within-context as well as across-context performance except naive Bayes. This result could be explained by the probabilistic nature of naive Bayes, which makes it "relatively insensitive to noise" (Webb, 2010). This is also reported by Nettleton and colleagues (2010). However, the initial performance (i.e., at zero noise level) of the naive Bayes model was significantly lower than that of other models for collaboration-quality estimation tasks, which makes it less suitable for modelling collaboration quality with simple audio and log features.

A 50% attribute noise level led to around 50% reduction in Kappa of our high-performing models (i.e., random forest and decision tree). The reduction was in both within-context and across-context performance. This reduction in performance due to attribute noise may provide an estimate of how much a model's performance could be degraded in classroom settings. For example, a recent work from Pugh and colleagues (2022) reported that their language-based model achieved a performance of 80% AUCROC (area under the ROC curve) for classifying collaborative problem-solving aspects (i.e., constructing shared knowledge). It would be interesting to see whether their model suffers a similar performance loss when exposed to noisy datasets in classroom settings. Furthermore, the reduction in performance raises the need for extensive exploration of a model's performance in the presence of noise before putting that model into practice. It will equip users with prior information about the expected decline in the model's performance with noisy data, such as in noisy environments.

7.2 *RQ2: Which ML Algorithms Show More Robust Performance in the Presence of Attribute Noise for the Task of Estimating Collaboration Quality?*

Finding 4: Random forest and decision tree were found to be most effective in dealing with noisy data for collaboration-quality estimation tasks. Our results showed that random forest outperformed other models developed using decision tree, naive Bayes, and AdaBoost algorithms for better estimating collaboration quality within and across context. This finding on random forest is consistent with work from Melville and colleagues (2004), who found ensemble models better at handling attribute noise. The findings on the decision tree are also consistent with Nettleton and colleagues' (2010) research, where naive Bayes and decision tree both performed closely when noise was present in test data. We also expected the naive Bayes model to perform well with noisy data (Nettleton et al., 2010), and our results showed that the naive Bayes model was affected the least by attribute noise. However, due to poor initial performance (e.g., with 0% noise), naive Bayes got the last rank.

This finding is more in line with MMLA research (Reilly & Schneider, 2019; Viswanathan & Vanlehn, 2018) on collaborative learning and less with that of Nettleton and colleagues (2010). The results on random forest are aligned with the results of MMLA studies (Reilly & Schneider, 2019; Viswanathan & Vanlehn, 2018) building collaboration-quality models, which found that random forest outperformed other ML algorithms (e.g., naive Bayes). Our results take their findings further by showing that the same algorithm also performs best with attribute noise. Thus, we suggest the use of random forest with a multimodal dataset involving audio data for building collaboration-quality estimation models for authentic settings where the quality of attributes is likely to be compromised.

The deviation in findings from Nettleton and colleagues (2010) is most likely due to differences in the dataset used for analysis (i.e., Nettleton's datasets were synthetically generated to mimic different real-world datasets, while ours were from the particular domain of educational settings). Nettleton and colleagues (2010) indeed suggest that the robustness of ML algorithms with attribute noise is conditioned by the datasets used.

This finding, however, should be considered in the context of the set of four ML algorithms (namely decision tree, random forest, naive Bayes, and AdaBoost) that we used for our study. Though we selected these because they frequently achieve high performance for collaboration modelling tasks in LA and MMLA, they only represent a small subset of all possible ML algorithms. For example, researchers have also illustrated the potential of deep learning for estimating collaboration aspects (Spikol et al., 2018). Furthermore, prior research has also shown the potential of deep learning models to utilize histogram-based feature representation of students' participation to assess collaboration quality (Som et al., 2021). Ma and colleagues (2023) in their recent study also showcased the ability of deep learning models to perform gracefully in the presence of noise. However, deep learning requires large amounts of training data to avoid overfitting and it works as a "black box"—which makes the inspection and/or explanation of their outputs difficult, which we believe is critical in educational applications. The fact that our datasets were limited in terms of size and the degree to which model estimations could be explained also influenced our choice.

This finding also addresses the concerns raised by Alexandron and colleagues (2019) in their study about the need for more robust and generalizable analysis techniques that can deal with data noise. Their study provided preliminary evidence that introducing noise by adding simulated data of fake learners (i.e., "learners who do not rely on learning to achieve high performance"; Alexandron et al., 2019) biased results of LA significantly. Thus, there is a need to employ more robust techniques that can also be generalized across courses. Our results on the robust performance of random forest and decision tree, not only within a learning context but across different learning contexts, have implications for (MM)LA communities.

Finding 5: Decision tree may provide a robust, generalizable, and potentially explainable model for collaboration-quality estimation in noisy settings. Our results showed that though random forest outperformed all other models, the performance gain over decision tree was only 4% to 5% on within-context and across-context levels when noise was present in all of the features. This helps us to see that often overlooked easy to understand models (i.e., decision tree), also referred to as white-box models (Loyola-González, 2019), can perform better across context and even with the presence of attribute noise. Decision tree in comparison with random forest is easier to understand and is self-explanatory (Loyola-González, 2019). This finding provides preliminary results on the effectiveness of explainable (or white-box) models for the task of collaboration estimation with attribute noise. Furthermore, in educational fields, the use of a decision tree could enable practitioners to easily understand the estimation, which may promote greater adoption of automated tools in practice. This can potentially help in addressing the concerns related to trust with the use of AI in an educational context (Chounta et al., 2022). The use of such explainable models could also help in reducing AI bias (particularly, algorithmic bias) in education (Baker & Hawn, 2022).

7.3 Limitations

The presented study comes with a set of limitations. The first limitation is that the accuracy of the presented findings relies on the assumption of simulated noise being similar to the real environment of the study. In particular, our conceptualization of attribute noise limits the applicability of our findings to scenarios where multiple factors cause attribute noise. Further research work is needed to explore different noise generation processes, for example, recording noisy audio separately and fusing it with clean audio data. The second limitation comes from the use of our datasets, which were collected from different classrooms with different subject teachers but from a single school. Thus, the narrow context variation limits the generalizability of our results beyond an Estonian school. Furthermore, our datasets only had features from audio and log data; thus, the findings are not necessarily applicable to other types of data (e.g., video data). The fourth limitation is our use of simple quantitative features. This study makes use of literature-based simple quantitative features that do not take sophisticated discourse analysis (e.g., students having on-task or off-task discussion) of collaboration into account for modelling. This use of only quantitative features is likely to affect the robustness of developed models for collaboration estimation. The fifth limitation is with our use of Rummel and colleagues' (2011) framework for collaboration-quality modelling. This use of a specific framework also narrows down the scope of applicability of the presented findings. Other frameworks are also available to be explored, e.g., the PISA framework (OECD, 2017) and NISPI (Cukurova et al., 2018). Finally, we explored only a small subset of ML algorithms, even if we chose the most common ones used in MMLA for collaborative learning.

8. Conclusion and Future Work

This paper addresses a research gap regarding the impact of attribute noise on collaboration-quality estimation models developed with supervised ML using multimodal data. In this paper, we provided a systematic analysis of how different proportions of attribute noise affect the performance of collaboration-quality estimation models. Our results showed that noise in log data does not have a significant impact on the performance of the ML algorithms used. However, the selection of ML algorithms greatly affects performance when noise is present in audio data. Our results suggest that random forest and decision tree algorithms could be a good choice for building automated collaboration estimation models for authentic classroom settings where attribute quality is likely to be compromised.

In our future work, we plan to explore other noise generation processes from related ML research (i.e., speech recognition and image processing) to study the impact of noise. We aim to extend our presented research by modelling/characterizing noise and then using those models to study noise's impact on model performance for the real-world environment. In this sense, future research work should consider recording and analyzing audio data from a variety of classrooms so that the distribution of noise in audio data can be better understood. We will collect data from authentic learning contexts with a larger range of variation (e.g., different kinds/lengths of collaborative learning activities, more variation in the type of schools, varied educational levels, and studies conducted in several countries) to use for model development and assess whether differences among students or communication styles associated with different cultures also affect the performance of ML models. Moreover, we also plan to explore other non-verbal audio features (e.g., pitch, energy) and video features (e.g., emotions, facial action units, head movement), along with more features from Etherpad log data (e.g., higher-level operations, edit, paste). In addition, we also aim to utilize the speech text for extracting discourse-related features (e.g., argumentation level, agreement, rapport) to model collaboration quality. The use of such discourse-related complex features, in addition to adding more robustness to the resultant collaboration-quality models, is likely to be more convincing to the practitioners and, therefore, may contribute to adoption. In this paper, we have explored the impact of noise on collaboration quality, as a whole. To go deeper, we also plan to investigate the impact of attribute noise on the estimation models of underlying dimensions of collaboration quality. This investigation would differ from the present study on the modelling stage, which will build ML models for each dimension separately. This investigation could offer more insights into which dimensions can be estimated robustly even in the case of attribute noise and which ones are likely to be significantly affected by attribute noise.

Our findings and methodological approach may help other researchers (e.g., from MMLA and CSCL communities) to understand and evaluate the attribute noise effect in building automated collaboration-quality models and the potential of white-box models (like decision tree) to build reasonably robust and generalizable AI solutions that are also explainable. Furthermore, this paper also opens a new door for analyzing the practice-readiness of the current state of the art in collaboration estimation using ML (i.e., how much performance loss could be expected due to attribute noise). We hope the community will further explore this research direction, addressing attribute noise issues and advancing the realization of reliable automated models for practical applications.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Funding

The presented work has been partially funded by the Estonian Research Council's Personal Research Grant (PRG) under grant number PRG1634; grants PID2020-112584RB-C32, RYC2021-032273-I and RYC2022-037806-I, financed by MCIN/AEI/10.13039/501100011033, the European Union's "NextGenerationEU/PRTR", and ESF+; and grant VA176P23, funded by the Junta de Castilla y León and FEDER.

References

Alexandron, G., Yoo, L. Y., Ruipérez-Valiente, J. A., Lee, S., & Pritchard, D. E. (2019). Are MOOC learning analytics results trustworthy? With fake learners, they might not be! *International Journal of Artificial Intelligence in Education*, *29*, 484–506. https://doi.org/10.1007/s40593-019-00183-1

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Chejara, P., Kasepalu, R., Prieto, L., Rodríguez-Triana, M. J., & Ruiz-Calleja, A. (2024). Bringing collaborative analytics using multimodal data to masses: Evaluation and design guidelines for developing a MMLA system for research and teaching practices in CSCL. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Kyoto, Japan (pp. 800–806). ACM. https://doi.org/10.1145/3636555.3636877

Chejara, P., Kasepalu, R., Prieto, L. P., Rodríguez-Triana, M. J., Ruiz Calleja, A., & Schneider, B. (2023). How well do collaboration quality estimation models generalize across authentic school contexts? *British Journal of Educational Technology*, *55*(4), 1602–1624. https://doi.org/https://doi.org/10.1111/bjet.13402

Chejara, P., Kasepalu, R., Prieto, L. P., Rodríguez-Triana, M. J., Ruiz-Calleja, A., & Shankar, S. K. (2023). Multimodal learning analytics research in the wild: Challenges and their potential solutions. In *CrossMMLA'23: Leveraging Multimodal Data for Generating Meaningful Feedback*, 14 March 2023, Arlington, Texas, USA (pp. 1–5). https://ceur-ws.org/Vol-3439/paper5.pdf

Chejara, P., Prieto, L. P., Rodríguez-Triana, M. J., Kasepalu, R., Ruiz-Calleja, A., & Shankar, S. K. (2023). How to build more generalizable models for collaboration quality? Lessons learned from exploring multi-context audio-log datasets using multimodal learning analytics. In *Proceedings of the 13th International Conference on Learning Analytics and Knowledge* (LAK 2023), 13–17 March 2023, Arlington, Texas, USA (pp. 111–121). ACM. https://doi.org/10.1145/3576050.3576144

Chejara, P., Prieto, L. P., Ruiz-Calleja, A., Rodríguez-Triana, M. J., Shankar, S. K., & Kasepalu, R. (2021). EFAR-MMLA: An evaluation framework to assess and report generalizability of machine learning models in MMLA. *Sensors*, *21*(8), 1–27. https://doi.org/10.3390/s21082863

Chounta, I. A., Bardone, E., Raudsep, A., & Pedaste, M. (2022). Exploring teachers' perceptions of artificial intelligence as a tool to support their practice in Estonian K-12 education. *International Journal of Artificial Intelligence in Education*, *32*(3), 725–755. https://doi.org/10.1007/s40593-021-00243-5

Chua, Y. H. V., Dauwels, J., & Tan, S. C. (2019). Technologies for automated analysis of co-located, real-life, physical learning spaces. In *Proceedings of the Ninth International Conference on Learning Analytics and Knowledge* (LAK 2019), 4–8 March 2019, Tempe, Arizona, USA (pp. 11–20). ACM. https://doi.org/10.1145/3303772.3303811

Crescenzi-Lanna, L. (2020). Multimodal learning analytics research with young children: A systematic review. *British Journal of Educational Technology*, *51*(5), 1485–1504. https://doi.org/10.1111/bjet.12959

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, *116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007

Darvishi, A., Khosravi, H., Sadiq, S., & Weber, B. (2022). Neurophysiological measurements in higher education: A systematic literature review. *International Journal of Artificial Intelligence in Education*, *32*(2), 413–453. https://doi.org/10.1007/s40593-021-00256-0

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, *34*(4), 338–349. https://doi.org/10.1111/jcal.12288

Drachsler, H., & Schneider, J. (2018). JCAL special issue on multimodal learning analytics. *Journal of Computer Assisted Learning*, *34*(4), 335–337. https://doi.org/10.1111/jcal.12291

Emerson, A., Henderson, N., Rowe, J., Min, W., Lee, S., Minogue, J., & Lester, J. (2020). Early prediction of visitor engagement in science museums with multimodal learning analytics. *Proceedings of the 2020 International Conference on Multimodal Interaction* (ICMI 2020), 25–29 October 2020, online, 107–116. https://doi.org/10.1145/3382507.3418890

Han, J., & Kamber, M. (2000). Data mining: Concepts and techniques. In J. Gray (Ed.), *The Morgan Kaufmann series in data management systems*. Morgan Kaufmann Publishers. https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

Hickey, R. J. (1996). Noise modelling and evaluating learning from examples. *Artificial Intelligence*, *82*(1-2), 157–179. https://doi.org/10.1016/0004-3702(94)00094-8

Kalapanidas, E., Avouris, N., Craciun, M., & Neagu, D. (2003). Machine learning algorithms: A study on noise sensitivity. In Y. Manolopoulos & P. Spirakis (Eds.), *Proceedings of the First Balkan Conference in Informatics*, 21–23 November 2003, Thessaloniki, Greece (pp. 356–365). http://delab.csd.auth.gr/bci1/Balkan/0prefaceBalkan.pdf

Kasepalu, R., Chejara, P., Prieto, L. P., & Ley, T. (2023). Studying teacher withitness in the wild: Comparing a mirroring and an alerting & guiding dashboard for collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 1–32. https://doi.org/10.1007/s11412-023-09414-z

Kasepalu, R., Prieto, L. P., Ley, T., & Chejara, P. (2022). Teacher artificial intelligence-supported pedagogical actions in collaborative learning coregulation: A Wizard-of-Oz study. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.736194

Laal, M., & Ghodsi, S. M. (2012). Benefits of collaborative learning. *Procedia—Social and Behavioral Sciences*, *31*, 486–490. https://doi.org/10.1016/j.sbspro.2011.12.091

Loyola-González, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, *7*, 154096–154113. https://doi.org/10.1109/ACCESS.2019.2949286

Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA 2014), 12 November 2014, Istanbul, Türkiye (pp. 5–12). ACM. https://doi.org/10.1145/2666633.2666635

Ma, Y., Celepkolu, M., Boyer, K. E., Lynch, C. F., Wiebe, E., & Israel, M. (2023). How noisy is too noisy? The impact of data noise on multimodal recognition of confusion and conflict during collaborative learning. In E. André, M. Chetouani, D. Vaufreydaz, G. Lucas, T. Schultz, L.-P. Morency, & A. Vinciarelli (Eds.), *Proceedings of the 25th International Conference on Multimodal Interaction* (ICMI 2023), 9–13 October 2023, Paris, France (pp. 326–335). ACM. https://doi.org/10.1145/3577190.3614127

Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2015). MTFeedback: Providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Transactions on Learning Technologies*, *8*(2), 187–200. https://doi.org/10.1109/TLT.2014.2365027

Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., & Yacef, K. (2013). Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 455–485. https://doi.org/10.1007/s11412-013-9184-1

Martínez-Maldonado, R. (2011). Modelling symmetry of activity as an indicator of collocated group collaboration. In J. Konstan, R. Conejo, J. Marzo, & N. Oliver (Eds.), *Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and Lecture notes in bioinformatics)* (pp. 207–218, Vol. 6787). https://doi.org/10.1007/978-3-642-22362-4_18

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *International Journal of Computer-Supported Collaborative Learning*, *2*, 63–86. https://doi.org/10.1007/s11412-006-9005-x

Melville, P., Shah, N., Mihalkova, L., & Mooney, R. J. (2004). Experiments on ensembles with missing and noisy data. In F. Roli, J. Kittler, & T. Windeatt (Eds.), *Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and Lecture notes in bioinformatics)* (pp. 293–302, Vol. 3077). https://doi.org/10.1007/978-3-540-25966-4_29

Nettleton, D. F., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, *33*(4), 275–306. https://doi.org/10.1007/s10462-010-9156-z

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, *3*(2), 213–219. https://doi.org/10.18608/jla.2016.32.10

OECD. (2017). *PISA 2015 collaborative problem-solving framework*. https://doi.org/10.1787/9789264281820-8-en

Papandreou, G., Katsamanis, A., Pitsikalis, V., & Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(3), 423–435. https://doi.org/10.1109/TASL.2008.2011515

Peebles, J. (1987). *Probability, random variables, and random signal principles*. McGraw-Hill.

Ponce-Lopez, V., Escalera, S., & Baro, X. (2013). Multi-modal social signal analysis for predicting agreement in conversation settings. *Proceedings of the 2013 ACM International Conference on Multimodal Interaction* (ICMI 2013), 9–13 December 2013, Sydney, Australia, 495–502. https://doi.org/10.1145/2522848.2532594

Praharaj, S., Scheffel, M., Drachsler, H., & Specht, M. (2021). Co-located collaboration modelling using multimodal learning analytics—Can we go the whole nine yards? *IEEE Transactions on Learning Technologies*, *14*(3), 367–385. https://doi.org/10.1109/TLT.2021.3097766

Pugh, S. L., Rao, A., Stewart, A. E., & D'Mello, S. K. (2022). Do speech-based collaboration analytics generalize across task contexts? In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 208–218). ACM. https://doi.org/10.1145/3506860.3506894

Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E. B., Andrews-Todd, J., & D'Mello, S. K. (2021). Say what? Automatic modeling of collaborative problem solving skills from student speech in the wild. In *Proceedings of the 14th International Conference on Educational Data Mining* (EDM 2021), 29 June 2021–2 July 2021, Paris, France, and online (pp. 55–67). International Educational Data Mining Society. https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_141.pdf

Redman, T. C. (2001). *Data quality: The field guide*. Digital Press.

Reilly, J. M., & Schneider, B. (2019). Predicting the quality of collaborative problem solving through linguistic analysis of discourse. In *Proceedings of The 12th International Conference on Educational Data Mining* (EDM 2019), 2–5 July 2019, Montréal, Québec, Canada (pp. 149–157). https://educationaldatamining.org/edm2019/proceedings/

Reyes-García, J., Galeana-Zapién, H., Galaviz-Mosqueda, A., & Torres-Huitzil, C. (2018). Evaluation of the impact of data uncertainty on the prediction of physiological patient deterioration. *IEEE Access*, *6*, 38595–38606. https://doi.org/10.1109/ACCESS.2018.2853701

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer supported collaborative learning. NATO ASI Series* (pp. 69–97, Vol. 128). Springer. https://doi.org/10.1007/978-3-642-85098-1_5

Rummel, N., Deiglmayr, A., Spada, H., Kahrimanis, G., & Avouris, N. (2011). Analyzing collaborative interactions across domains and settings: An adaptable rating scheme. In S. Puntambekar, G. Erkens, & C. Hmelo-Silver (Eds.), *Analyzing interactions in CSCL* (pp. 367–390). https://doi.org/10.1007/978-1-4419-7710-6_17

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 375–397. https://doi.org/10.1007/s11412-013-9181-4

Schneider, B., Sung, G., Chng, E., & Yang, S. (2021). How can high-frequency sensors capture collaboration? A review of the empirical links between multimodal metrics and collaborative constructs. *Sensors*, *21*(24), 8185. https://doi.org/10.3390/s21248185

Schwarz, B. B., & Asterhan, C. S. (2011). E-moderation of synchronous discussions in educational settings: A nascent practice. *Journal of the Learning Sciences*, *20*(3), 395–442. https://doi.org/10.1080/10508406.2011.553257

Smith, J., Bratt, H., Richey, C., Bassiou, N., Shriberg, E., Tsiartas, A., D'Angelo, C., & Alozie, N. (2016). Spoken interaction modeling for automatic assessment of collaborative learning. In *Proceedings of the International Conference on Speech Prosody*, 31 May–3 June 2016, Boston, Massachusetts, USA (pp. 277–281). https://doi.org/10.21437/SpeechProsody.2016-57

Som, A., Kim, S., Lopez-Prado, B., Dhamija, S., Alozie, N., & Tamrakar, A. (2021). Automated student group collaboration assessment and recommendation system using individual role and behavioral cues. *Frontiers in Computer Science*, *3*, 728801. https://doi.org/10.3389/fcomp.2021.728801

Sonnenberg, C., & Bannert, M. (2015). Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *Journal of Learning Analytics*, *2*(1), 72–100. https://doi.org/10.18608/jla.2015.21.5

Spikol, D., Cukurova, M., & Ruffaldi, E. (2017). Using multimodal learning analytics to identify aspects of collaboration in project-based learning introduction PELARS system and context. In *Making a Difference: Prioritizing Equity and Access in CSCL, 12th International Conference on Computer Supported Collaborative Learning* (CSCL 2017), 18–22 June 2017, Philadelphia, Pennsylvania, USA (pp. 263–270). https://repository.isls.org/handle/1/240

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, *34*(4), 366–377. https://doi.org/10.1111/jcal.12263

Storch, N. (2001). How collaborative is pair work? ESL tertiary students composing in pairs. *Language Teaching Research*, *5*(1), 29–53. https://doi.org/10.1177/136216880100500103

Viswanathan, S. A., & Vanlehn, K. (2018). Using the tablet gestures and speech of pairs of students to classify their collaboration. *IEEE Transactions on Learning Technologies*, *11*(2), 230–242. https://doi.org/10.1109/TLT.2017.2704099

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299–321. https://doi.org/10.1080/00220272.2012.668938

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Webb, G. I. (2010). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 713–714). Springer US. https://doi.org/10.1007/978-0-387-30164-8_576

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, *46*(1), 71–95. https://doi.org/10.1016/j.compedu.2005.04.003

Worsley, M., & Blikstein, P. (2018). A multimodal analysis of making. *International Journal of Artificial Intelligence in Education*, *28*, 385–419. https://doi.org/10.1007/s40593-017-0160-1

Yan, L., Zhao, L., Gasevic, D., & Martinez-Maldonado, R. (2022). Scalability, sustainability, and ethicality of multimodal learning analytics. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 13–23). https://doi.org/10.1145/3506860.3506862

Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, *22*(3), 177–210. https://doi.org/10.1007/s10462-004-0751-8

Appendix A

Table 8. Characteristics of Learning Activities

id	learning activity task	students	groups	subject
1	The task was to complete the given sentences on past and present tenses. The activity also asked the groups to discuss and write collaboratively a paragraph on what they would do if they were given a particular sum of money (10,000 euros).	6	2	English language
2	The activity involved reading a magazine article containing multiple paragraphs explaining the journey of a girl who became a press reporter from a librarian in a month. The students were asked to first assign headings to each paragraph and then discuss their opinion on the possibility of learning a completely new job in a month.	8	3	English language
3	The task was to write an essay collaboratively on given topics (e.g., "The generation today is less healthy than our parents"). The groups were also asked to assess their essay against a set of checklists focusing on content, communication, organization, and language use.	4	2	English language
4	The task involved preparing a presentation in the group on one of the epics[4] (e.g., *Gilgamesh*, *Song of My Cid*). The groups were given instructions on the content to put in the presentation, e.g., describe the main characters and summarize the central story of the epic. At the end of the session, the groups were asked to present in front of the class.	7	3	Estonian language
5	The groups were given a speech conversation transcript involving people asking questions about their habits. Groups were first asked to gather the same information from their peers and write complete sentences, including the name of their peer.	9	3	English language
6	The task involved a hypothetical situation of a person, Steve, who needed to renovate a particular portion of his house (exterior facade, bathroom, and living room). The groups were given a map of the house with measurements of each wall as well as the floor. The groups were asked to first prepare a list of tools and materials needed to complete the renovation. The groups were also asked to discuss the estimated cost of labour and materials and prepare the final document with all details for Steve.	15	5	Woodwork and chemistry
7	The collaborative activity involved solving a set of geometric problems. Each group was given a similar set of problems with different measurements. For example, one problem for group 3 was to calculate the perimeter and area of a rectangle with a diagonal of 8.5 dm forming an angle of 25 degrees with the longer side.	13	4	Mathematics
8	The groups were given topics to choose from and then write a for-and-against essay, First, they put down ideas supporting the topic, and second, they found arguments opposing it. The groups were given a structure to follow for the essay.	9	3	English language
9	The activity involved dividing student groups into two categories: Employee and Employer. For each category, students were given a set of questions/tasks to discuss and write down in the text editor. For example: one of the tasks for the Employer group was "You are the owners of a construction company. Please think about which personal traits are important for a construction worker. Put down the traits below and also the reason why they are important."	13	4	English language
10	Same as #9.	8	3	English language

[4]Oxford definition: a long poem about the actions of great men and women or a nation's history

Table 8. Characteristics of Learning Activities (continued)

id	learning activity task	students	groups	subject
11	The task was to write a discursive essay on the topic "The Growth of Online Shopping Has Greatly Improved Life for the Consumer" after brainstorming the arguments for and against the topic's idea. The groups were given a similar checklist as #3 for evaluating their essay.	11	3	English language

Appendix B

Figure 4 shows a snapshot of log data collected from the group activity in the presented research. Whenever a participant interacted with the collaborative text editor (i.e., Etherpad), an entry was recorded in the log data. This entry included *timestamp*, *author*, *group*, *char_bank*, *changeset*, *source_length* (i.e., text length before the action), *operation* (i.e., type of operation), and *difference* (i.e., the difference in the text length as a result of the action). The *timestamp* represents the time when the action was performed. The participant id and group id were captured in the *author* and *group* columns. The *char_bank* column recorded what text was added or deleted by the participant. The *changeset* column encoded information related to where changes were made in the document. The editor uses this encoded form to reproduce the change history. The *source_length* column recorded the text length in the editor before the current interaction. The *operation* recorded whether the text was added (>) or deleted (<). The *difference* column tracked the difference in the length as a result of the current interaction. Finally, the *text* column recorded the text present in the editor.

timestamp	author	group	char_bank	changeset	source_length	operation	difference	text
02.04.2022 21:10	a.qELM24Qsmgw22hrh	1	t	Z:1>1*0+1$t	1	>	1	t
02.04.2022 21:10	a.qELM24Qsmgw22hrh	1	his is	Z:2>6=1*0+6$his is	2	>	6	this is
02.04.2022 21:10	a.qELM24Qsmgw22hrh	1	pank	Z:8>5=7*0+5$ pank	8	>	5	this is pank<

Figure 4. Sample of Log Data

Figure 5 shows a snapshot of the VAD (voice activity detection) data file, which recorded participants' speaking activity. This file contained information related to *timestamp*, *user*, *group*, and *speaking_time*. The timestamp is when the participant spoke and the speaking time is the continuous speaking duration in seconds. The information about the participant and the group the participant belongs to was recorded in the *user* and *group* columns.

timestamp	user	group	speaking_time(sec.)
2022-04-02 18:16:25+00:00	a.qELM24Qsmgw22hrh	2	4.013
2022-04-02 18:16:44+00:00	a.qELM24Qsmgw22hrh	2	8.114
2022-04-02 18:16:58+00:00	a.qELM24Qsmgw22hrh	2	2.207

Figure 5. Sample of Voice Activity Detection Data

Figure 6 presents a sample of a speech data file. The information in *timestamp*, *user*, and *group* represents the same to the VAD file. The last column (i.e., *speech*) contains the speech-to-text data.

timestamp	user	group	speech
2022-02-10 06:58:50+00:00	a.R4U2erGWXBpQqJk2	2	even then you can login using your already inside
2022-02-10 06:59:08+00:00	a.PhHnktSyQDJAIM4w	2	the left.
2022-02-10 06:59:09+00:00	a.R4U2erGWXBpQqJk2	2	the last one
2022-02-10 06:59:17+00:00	a.PhHnktSyQDJAIM4w	2	Also you said
2022-02-10 06:59:29+00:00	a.Qmx2guSAy0usXyXh	1	other than a headset

Figure 6. Sample of Speech Data