How to Build More Generalizable Models for Collaboration Quality? Lessons Learned from Exploring Multi-Context Audio-Log Datasets using Multimodal Learning Analytics

Pankaj Chejara
pankajch@tlu.ee
Tallinn University
Tallinn, Estonia

Luis P. Prieto
luisp@tlu.ee
Tallinn University
Tallinn, Estonia

María Jesús Rodríguez-Triana
mjrt@tlu.ee
Tallinn University
Tallinn, Estonia

Adolfo Ruiz-Calleja
adolfo@tlu.ee
Tallinn University
Tallinn, Estonia

Reet Kasepalu
reetkase@tlu.ee
Tallinn University
Tallinn, Estonia

Shashi Kant Shankar
shashik@tlu.ee
Tallinn University
Tallinn, Estonia

ABSTRACT

Multimodal learning analytics (MMLA) research for building collaboration quality estimation models has shown significant progress. However, the generalizability of such models is seldom addressed. In this paper, we address this gap by systematically evaluating the across-context generalizability of collaboration quality models developed using a typical MMLA pipeline. This paper further presents a methodology to explore modelling pipelines with different configurations to improve the generalizability of the model. We collected 11 multimodal datasets (audio and log data) from face-to-face collaborative learning activities in six different classrooms with five different subject teachers. Our results showed that the models developed using the often-employed MMLA pipeline degraded in terms of Kappa from Fair (.20 < Kappa < .40) to Poor (Kappa < .20) when evaluated across contexts. This degradation in performance was significantly ameliorated with pipelines that emerged as high-performing from our exploration of 32 pipelines. Furthermore, our exploration of pipelines provided statistical evidence that often-overlooked contextual data features improve the generalizability of a collaboration quality model. With these findings, we make recommendations for the modelling pipeline which can potentially help other researchers in achieving better generalizability in their collaboration quality estimation models.

KEYWORDS

MultiModal Learning Analytics, Machine Learning, Collaboration Quality, Generalizability

ACM Reference Format:
Pankaj Chejara, Luis P. Prieto, María Jesús Rodríguez-Triana, Adolfo Ruiz-Calleja, Reet Kasepalu, and Shashi Kant Shankar. 2023. How to Build More Generalizable Models for Collaboration Quality? Lessons Learned from Exploring Multi-Context Audio-Log Datasets using Multimodal Learning Analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576050.3576144

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576144

1 INTRODUCTION

Multimodal learning analytics (MMLA) allows researchers to capture activity traces from the physical space of collaborative learning, unlike traditional log-based learning analytics [2]. By capturing activity traces from both physical and digital spaces, MMLA enables a more holistic understanding of collaboration. In this direction, researchers have explored the use of different data sources (e.g., audio and video) [26] and identified several indicators of collaboration behavior [20]. For example, speaking time has been found to be an indicator of equality of participation [20] which was found as an indicator of the quality of collaboration [3, 15]. Furthermore, MMLA research has also been extended to automatically detect collaboration in the context of collaborative learning. For example, researchers have also built models that can automatically estimate collaboration behavior [3, 22, 30].

Estimation models for collaboration have the potential for building automated tools (e.g., [14]) that can support teachers monitor student activity during collaborative learning. In fact, such tools were also found to be effective in improving teachers' awareness of collaborative activity as illustrated by [9]. These advances may contribute toward propelling MMLA research on automating collaboration estimation [5]. Consequently, there has been a growing interest in building models that can automate the collaboration estimation in MMLA [3, 8, 11, 15, 22, 30].

The collaboration estimation models developed in MMLA have performed well compared to baseline performances (e.g., chance performance [8] or majority performance [12]). In some cases, studies have reported achieving accuracy above 90% for their classification models of collaboration [33]. The majority of research, however, has only used data from a single learning context[1] for collaboration estimation model development and evaluation [11, 12, 33]. This

[1]Here, we consider a learning context composed of multiple aspects, e.g., students, learning activity, teacher, and learning environment. For example, if two learning contexts involve the same students, teacher, and learning environment, but different learning activities, then these two will be perceived as different in terms of an activity task.

means that using data from a single context confines the applicability of developed models to a particular context, thus, limiting the knowledge on the generalizability of developed models [16].

The generalizability of collaboration estimation models to other learning contexts -which the models have never seen- is still in question [31, 33]. A systematic evaluation of the model's generalizability at different levels (how the model performs within and across contexts) is currently absent. Moreover, it is critical for the implementation in authentic contexts because any MMLA system once put into practice will see a new context (by our criteria of contextual differences) every time it is used for estimation [5]. Consequently, there is a lack of knowledge on how to build generalizable estimation models for collaboration.

The evaluation of generalizability demands the collection of additional MMLA datasets [16], preferably from a different learning context to assess the generalizability of the model across contexts. With this demand of collecting multiple MMLA datasets, an additional question arises: would contextual features which contain crucial information about the learning context contribute to the generalizability of the model? Furthermore, an appropriate selection of modelling steps (e.g., outlier handling or data scaling) is crucial to improve the generalizability of the models [28]. However, there is a lack of research on how different modelling steps impact the generalizability of the collaboration estimation model.

These aforementioned gaps in MMLA research led us to pose three research questions: **RQ1**- How do collaboration estimation models which are developed using a standard MMLA pipeline perform across different contexts?; **RQ2** - Which MMLA pipeline offers further improvement in the model's performance across contexts, in other words, its generalizability?; **RQ3** - What is the impact of adding/removing a modelling step and contextual features on the model's performance across contexts?

To answer the aforementioned research questions, we collected 11 multimodal datasets (audio and log) from six Estonian classrooms with five subject teachers during face-to-face collaborative learning activities. We used these datasets to develop collaboration estimation models using a standard MMLA pipeline (e.g., the one used in [15]) and performed a systematic generalizability evaluation, addressing RQ1. To tackle RQ2, we explored multiple pipelines (32 in total) to develop collaboration estimation models and identified high-performing model pipelines in terms of across-context generalizability. We then examined the performance of these pipelines to study the impact of different modelling steps and the use of contextual features on the generalizability of the model using a statistical method, thus addressing RQ3.

2 RELATED WORK

MMLA research has grown on estimating collaboration (or its aspects) using a variety of data sources, ranging from logs, audio, and video to eye-gaze trackers [26]. The automation of collaboration estimation can potentially help to develop monitoring tools that can support teachers during a collaborative learning activity [9]. However, given the multidimensional nature of collaboration [25], this estimation would limit the actionability aspect, i.e., the teacher would be clueless about how the current situation of low collaboration quality can be improved. Therefore, some MMLA researchers developed models for underlying dimensions of collaboration quality such as argumentation or knowledge exchange [5, 22]. Through these studies, the collaboration quality construct is examined in greater depth, which further opens the possibility of automated guiding tools to support teachers' interventions. For example, the availability of intervention strategies for collaboration quality dimensions from computer-supported collaborative learning (CSCL) literature together with the automated models could help in developing guiding tools to assist teachers with intervention strategies [9].

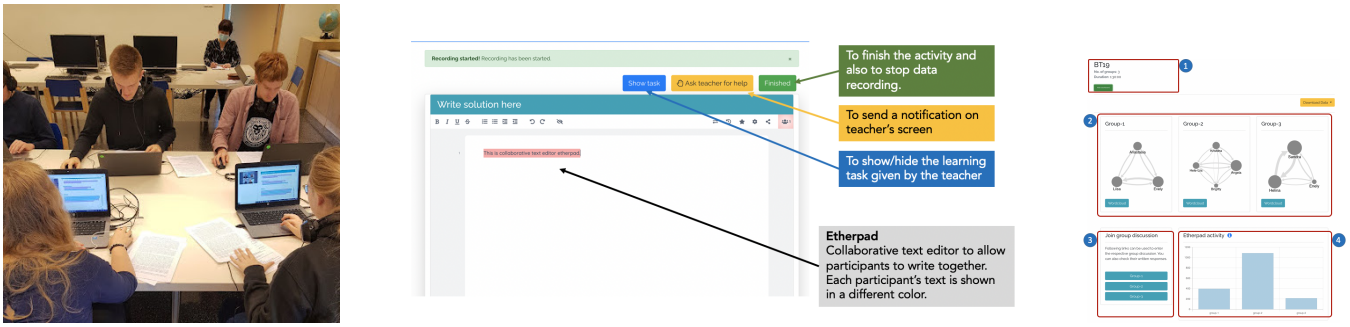
MMLA research on automating the estimation of collaboration quality and its dimensions has reported achieving accuracy from moderate (70% accuracy [19]) to high (90% accuracy [33]). The majority of these studies have used accuracy as a performance metric for their models, it is problematic for cases with a class imbalance problem (e.g., binary labels with 80%/20% ratio). Therefore, researchers have also used other performance metrics for reporting their results, e.g., f1-score [15], Kappa [33], Area under the curve [11], etc. These performance metrics are often reported without any reference point, hindering the assessment of solutions for automatic collaboration detection at community level in MMLA. This need for a reference point has also been argued in a framework [5] proposed as a model evaluation framework for MMLA.

In order to develop their models, researchers have often followed a standard machine learning process, mainly including data collection, feature extraction, model development and evaluation [1, 3, 8, 11, 15, 17, 19, 30, 33]. In this process, the data flow through multiple steps (e.g., scaling and outlier handling) which are altogether called the modeling pipeline. Each of the modeling steps usually has more than one choice to select from. For example, there are multiple techniques to perform data scaling such as Standard or MinMax scaling. Since the selection of a particular choice can impact the performance of the developed model [29] it is essential to understand the impact of the different choices on the generalizability of the model.

With the availability of diverse choices for each of the modelling steps in the pipeline, the total number of alternatives becomes very large. For example, the researcher may wonder which data scaling and which strategy to use for handling class imbalance. This makes the exploration of all the choices for each modelling step a time-exhausting task. Therefore, to expedite the model development process, MMLA researchers have often used a similar pipeline to develop collaboration estimation models [1, 3, 11, 31]. This approach saves time but deprives models from achieving high performance with an adequate selection of choices.

For model evaluation, K-fold cross-validation (CV) with varying values of K has been used (e.g., 10 [12, 15], 5 [23], or 4 [31]). In educational terms, this evaluation assesses the model using data instances never seen by the model in training but taken from the same kind of context (i.e., contexts with the same teacher, learning activity, students, and learning environment). As a consequence, the K-fold CV results approximate the performance of the model in a similar situation that is not realistic in authentic practice (e.g., every new class would have a new activity or different teacher or students).

The majority of MMLA research has developed estimation models with data from a single context [1, 15, 17, 19, 33]. Consequently, their results from model evaluation only indicate how well the



(a) (b) (c)

Figure 1: (a) Students working on the collaborative activity in the classroom (b) Collaborative activity space in CoTrack (c) A real-time multimodal dashboard for teachers to track monitoring students, component-1 showing activity details, component-2 showing speaking dynamics, component-3 showing controls to join group activities and component-4 showing writing activity

model performs within the same context. To gain an understanding of models' across-context generalizability, there is a need for additional datasets from different contexts. In this direction, the information about differences among contexts (from which data is collected) is highly relevant from educational points of view [13]. The importance of this contextual information, however, is yet to be explored for the development of collaboration estimation models. This, therefore, raises the question of whether contextual information in addition to multimodal data features impacts the generalizability of the collaboration quality model.

A few research works assessed the generalizability of their models on datasets from different learning contexts [5, 22]. For example, recent work from Pugh et al. [22] assessed the generalizability of their language-based models in the task of classifying collaborative problem-solving facets (e.g., constructing shared knowledge). Their model achieved a transfer ratio of 93% on performing in a context with a different learning activity. Similarly, Chejara et al. [5] reported their linguistic feature-based models filling 40-60% between chance and human performance at the task of rating collaboration quality dimensions. These studies provide preliminary evidence of the feasibility of building generalizable models. On the other hand, the aforementioned research was limited in terms of the number of evaluated contexts (e.g., datasets from two contexts different only in terms of activity). Therefore, further research is needed to evaluate generalizability in a wider variety of contexts.

3 MATERIALS AND METHODS

This section presents details of the study context, data collection setup, and research methods.

3.1 Context and data collection

The datasets used in this paper were collected as part of a previous study focusing on investigating teachers' perspectives on using multimodal analytics for monitoring and guiding collaborative learning activities [9].

3.1.1 Study context. The study was conducted in an Estonian vocational school during fourteen sessions of collaborative learning activities in the autumn semester of 2021. Out of fourteen, there were three sessions with missing data for all the groups. As a result, we only considered eleven sessions from our study. These eleven sessions were conducted in six different classrooms by five different teachers. The subjects were English language, Mathematics, Chemistry Integrated with Woodwork, Communication, and Estonian language with a total of 105 students. The language of instruction in all sessions was Estonian except for English language as a foreign language (English language). The majority of students in the study were Estonian. Table 1 presents the details of the contexts and learning activities. In all cases, the collaborative activity entailed the use of a collaborative text editor where students were asked to write the responses of their group.

3.1.2 Data collection tool. The study was conducted with the use of a tool: CoTrack [4]. This web-based application allows teachers to create collaborative learning activities with monitoring functionalities. It offers a collaborative writing space for groups, with the use of Etherpad[3], to draft the solution to a given problem together. CoTrack also records every writing activity and student's audio. The audio data is processed by CoTrack allowing extraction of data features in real-time (e.g., speaking time, turn-taking, and speech-to-text). These features are used by CoTrack to generate a real-time dashboard. Figure 1 shows the collaborative learning context, the student learning environment in CoTrack, and the dashboard.

3.1.3 Procedure. All collaborative learning activities were planned prior to the enactment of the study. The subject teacher together with a researcher (also co-author of this paper) from educational sciences created the learning design. As the activity involved the use of a particular web application (CoTrack), the help materials on how to use it were shared with the participating students in advance. Consent was asked from adult students and additionally, consent was also taken from the parents of students who were younger than 18 years old. On the day of the activity, the aforementioned researcher was present in the classroom and briefly explained the purpose of the study to the students before the activity.

The researcher also provided information on what data would be recorded during the activity. The data recording part was optional

[3]https://etherpad.org/

Table 1: 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 will do if they were given a particular sum of money (10,000 euros).	6	2	English Language
2	The activity involved reading a magazine's 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 to 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, organisation and language use.	4	2	English language
4	The task involved preparing a presentation in the group on one of the epics[2] topics (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 summarise 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 the 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	Wood work 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. Firstly, they put down ideas supporting the topic, and secondly, they write 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 in 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
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

and students were free to opt out of being recorded. In the case of opting out from recording, students were simply instructed not to allow camera and microphone access to the application, but could still be involved in the activity. The students were given instructions on how to access the collaborative learning environment and then asked to work on the given activity in the groups formed by the teacher. Each student had a laptop and a microphone for the activity (see the student distribution in Figure 1a). During the activity, the teacher had access to a dashboard provided by the web application (see Figure 1c).

3.2 Dataset processing

We obtained pre-processed data from CoTrack. CoTrack has integrated features for tracking Voice Activity Detection (VAD) and converting audio data to speech data using the Google Speech-to-Text API. This data was provided by CoTrack in the form of a CSV (Comma Separated Values) file. The VAD data included a timestamp, the speaking duration, a group-id, and user-id. The speech-to-text data included a timestamp, a group-id, a user-id, and a speech transcript. We also downloaded writing logs from CoTrack containing a timestamp, a user-id, a group-id, characters written/deleted, the length of text before the operation (e.g., write, delete), the type

Table 2: Extracted Features

Data source	Feature	Description
Audio	speaking_time	Speaking time in seconds
	turn_taking	Number of turns taken by student in every 30 seconds
	freq_I	Frequency of word 'I'
	freq_WE	Frequency of word 'WE'
	freq_WH	Frequency of WH-words (e.g., why, what)
Logs	char_add	Number of characters written
	char_del	Number of characters deleted

of operation, the difference in terms of text length after the operation, and the final text on Etherpad. The identifiers used in the downloaded CSV files (e.g., user-id) were already anonymized by CoTrack.

3.2.1 Data exclusion. We had to discard data from some groups because of missing video data from one or more of the participants in those groups. The video data was used for annotation purposes and having missed the data of even one participant restricted the annotation for the entire group. Thus, we decided to discard the data of the groups where video recordings for all participants were not available. In total, ten groups were excluded.

3.2.2 Feature extraction. We used the CSV files obtained from CoTrack for feature extraction (see Table 2 for extracted features). We extracted speaking time and turn-taking from VAD data. These features are found to be predictors of collaboration quality in MMLA [12, 19, 20]. From speech data, we extracted the frequency of "I" and "We". Our decision for selecting these features was based on past research which found differences in high and low collaborating groups in terms of their usage of the words "I" and "We" [32]. Additionally, we also extracted the frequency of wh-words from speech data. As there were three sessions where Estonian language was used for communication, for those sessions we extracted the frequency of Estonian translations of "I", "We" and wh-words ("I":"Ma", "We":"Me", kes":"who", "kus":"where", "mis":"what", "miks":"why", "kuidas":"how", "milline":"which", "millal":"when"). From the writing logs, we extracted the number of characters written or deleted by the participants of the groups. As these features were collected on an individual level, we further took the average and the standard deviation to compute group-level features for each extracted feature.

3.2.3 Annotation. To obtain the ground truth of collaboration quality and its underlying dimensions, we used a rating scheme from Rummel et al. [25]. This rating scheme assigns scores for seven dimensions of collaboration quality, namely: argumentation, sustaining mutual understanding, cooperative orientation, structuring problem solving and time management, individual task orientation, knowledge exchange, and collaboration flow. Four MA graduate students from the School of Digital Technologies were trained using the adapted rating scheme as suggested by the author [25]. We assigned a score to each of the seven dimensions of collaboration quality every 30 seconds. The scores were on a 5-point scale (i.e., -2,-1,0,1,2). Cohen's Kappa was above .60 for all seven dimensions of collaboration quality, indicating substantial (as per [10] guidelines) inter-rater agreement. We added all the scores of the seven dimensions and took their average to compute a measure of collaboration quality. We further mapped[4] the scores for each dimension and the collaboration quality into binary labels (High, Low) for developing classification models. .

3.3 Methods

This section presents the methods employed to address the research questions set in the study.

3.3.1 Model development with a typical MMLA pipeline and generalizability evaluation (RQ1). We employed the widely used MMLA pipeline (multimodal data collection → feature extraction → model development → model evaluation) for building classification models for collaboration quality and its underlying dimensions. We used the random forest algorithm to build the models for two reasons: firstly, this algorithm has been found to achieve high performance in the field of MMLA [1, 21, 33]; secondly, we also achieved similar results in our previous study exploring machine learning models for estimating collaboration quality and its dimension [5].

The developed models were assessed using 10-fold cross-validation (CV) and a leave-one-context-out evaluation schemes, both of a nested nature (two levels of cross-validation, one for hyperparameter tuning and another for evaluation purposes). The 10-fold CV was performed separately on the dataset from each context to obtain a measure of generalizability within the context (or generalizability to the same context in terms of activity, teacher, students, etc.). In this evaluation, the dataset was divided into ten approximately equal portions, using nine portions for training and one for testing. This process is iterated ten times with the selection of a different portion for testing. The second evaluation scheme (leave-one-context-out) divided the datasets based on learning contexts. For example, our datasets were collected from eleven different learning contexts[5]. Thus, the datasets were split into eleven portions. Similarly, ten portions were used for training and one for testing. This was iterated eleven times. This evaluation allowed us to approximate how well the developed models perform in a learning context that could be different from the ones used for model development. Thus, it enabled us to investigate the generalizability aspect of the models developed with a typical MMLA pipeline. In other words, these two evaluation strategies assess generalizability within (at the instance level) and across contexts (at the context level), respectively (as per [5]). We report the model's performance using accuracy and kappa metrics.

[4]We mapped scores below or equal to zero as 'Low' otherwise 'High'.
[5]Here, we consider a learning context different if there are changes in any of the following: teacher, students, subject, education level, learning activity, and type of activity.

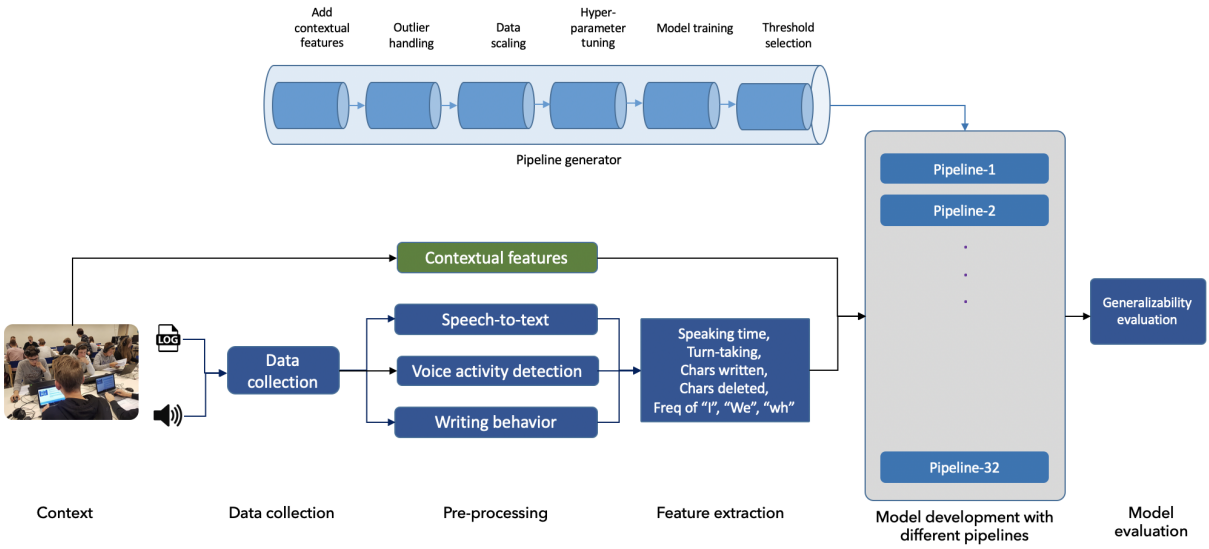


Figure 2: Exploration of pipelines with different configurations

3.3.2 Exploring modeling pipeline with different configurations (RQ2). Figure 2 shows our exploration[6] of various pipelines having different configurations for the following steps: data scaling, outlier handling, hyperparameter optimization, and classification threshold selection. In MMLA, it is common for the extracted features to have a different range, which could affect the process of building the model. For this reason, we included data scaling in our exploration. This data scaling step is likely to be influenced by the existence of outliers (anomaly values). Therefore, we additionally included the outlier handling step. The hyperparameter tuning step was also explored as it is often a part of the model development involving machine learning techniques that require manual configuration of parameters (refer to [6] for further details). For example, in our case of a random forest, the number of trees had to be decided before training. Finally, we decided to explore the classification threshold selection step, which is often used in machine learning and found to be effective in handling class imbalance problems [7].

For data scaling, we used the widely adopted Standard scaling which transforms data distribution with a mean zero and standard deviation of one. Additionally, we also employed the MaxAbs scaling technique which instead of shifting the mean of data like Standard scaling, scales the data while preserving its sparsity. The use of the data scaling step, however, also depends on the used machine learning algorithms. For example, in our case using data scaling was irrelevant due to the use of a tree-based algorithm (i.e., random forest). Despite we kept the step in the methodology so that it could be useful with other types of machine learning algorithms, in particular distance-based algorithms (e.g., support vector classifier).

For outlier handling, we decided to employ the clipping method to handle outliers instead of removing them due to the size of the dataset. The clipping method replaces the outlier values with the border value of the clipping interval (e.g., interval .05 - .95 replacing the large value with 95%tile value). For hyperparameter tuning, we used GridSearch CV from python's SciLearn library [18]. For the classification threshold, we chose a criterion that maximizes the kappa based on past research [7].

With all these steps except data scaling, we generated modelling pipelines either including the step or skipping it. For example, for the classification threshold, there were two different versions of pipelines generated, one with it and another one without it. For data scaling, we used Standard scaling in one group and MaxAbs in the second group.

All these configurations generated 16 pipelines in total. These pipelines were then used with two groups of features. The first group only contained the extracted multimodal features (Table 2). The second group contained multimodal features with contextual features (Table 3). This step resulted in a total of 32 pipelines. These pipelines were used to develop estimation models. For evaluation, we followed the same strategy as above.

Table 3: Details of contextual features

Feature	Description
activity_type	Type of collaborative learning activity, e.g., collaborative writing, problem-based.
time_of_day	When the activity took place, e.g., in the morning or in the evening.
students	Number of total students present in the classroom during the activity.
teacher_id	Teacher present in the activity.
classroom_id	Classroom identifier.
language	Language used for instruction and communication, e.g., English, Estonian.

[6]Source code: https://github.com/pankajchejara23/collaboration-quality-classification-modeling-using-MMLA

Table 4: Random forest classification performance using base MMLA pipeline

Target	Class balance	Instance generalizability (10-fold cv)		Context generalizability (leave-one-context-out)	
		Accuracy	Kappa	Accuracy	Kappa
Collaboration quality	68%	72% (3)	.23 (.06)	69% (16)	.15 (.11)
Argumentation	53%	63% (4)	.25 (.07)	57% (11)	.17 (.11)
Sustaining mutual understanding	67%	72% (3)	.23 (.05)	70% (14)	.16 (.10)
Collaboration flow	71%	74% (2)	.22 (.04)	71% (15)	.14 (.11)
Knowledge exchange	71%	75% (3)	.26 (.05)	74% (11)	.18 (.11)
Cooperative orientation	70%	73% (2)	.22 (.04)	70% (16)	.14 (.10)
Individual task orientation	65%	69% (2)	.17 (.09)	64% (14)	.10 (.10)
Structuring problem solving	69%	72% (3)	.15 (.07)	65% (24)	.07 (.09)

3.3.3 Statistical analysis on the impact of contextual features and modelling steps on the generalizability of the model (RQ3). To study the impact of a particular modelling step as well as the contextual features on the generalizability of the model, we divided the total number of pipelines into two sets. One with all the pipelines using that step (or the contextual feature, or MaxAbs scaling) and the other with the same but without that step (or no use of the contextual feature or use of Standard scaling). These two sets of pipelines were then employed to develop models for collaboration quality and its dimension, resulting in two sets of performance measures. The performance measures were not following a normal distribution. Therefore, we used a non-parametric Wilcoxon signed-rank test [34] to assess the statistical significance of differences in the model's generalizability as a result of adding a modeling step.

4 RESULTS

We present the results from our systematic evaluation of developed classification models using a typical modelling pipeline (RQ1), the exploration of 32 different pipelines with different configurations (RQ2), and the statistical analysis of its performance to study the impact on the models' generalizability across contexts (RQ3).

4.1 Performance of classification models developed using typical MMLA pipeline (RQ1)

Table 4 presents the classification models' performance in terms of kappa and accuracy using a widely used modeling pipeline (multimodal data collection → feature extraction → model development → model evaluation). Models in general achieved better performance than the baseline of chance performance (Kappa = 0). As expected, the performance of models degraded across different contexts in terms of kappa. However, when we look at the accuracy metric of performance, models seem to perform stably with a slight decrease in performance. For example, the collaboration quality model achieved 72% accuracy at 10-fold CV and 69% accuracy across contexts evaluation. This, however, may only indicate the model's performance in terms of classifying the positive class correctly (e.g., High collaboration quality) given class imbalance issue. In terms of classifying both, positive and negative classes, the models' predicted labels were in fair[7] agreement with ground truth (.20 < Kappa < .40) for instance generalizability except for individual task orientation and structuring problem-solving[8] dimensions. For these two dimensions, the Kappa was < .20, indicating poor agreement. On across contexts evaluation, all models' kappa exhibited poor agreement with ground truth (< .20).

4.2 Exploring pipelines to develop more generalizable models (RQ2)

We present the results of modeling pipelines that enabled the improvement over the base pipeline in terms of the generalizability of developed models (Table 5). The reported pipelines improved model's performance at both levels of generalizability, instance and context. At instance level (with 10-fold CV), these pipelines improved Kappa from fair (with base pipeline) to moderate (.40 < Kappa < .60) for all target labels (as per [10]). However, we only report the performance across contexts in table 5 to offer a comparison in terms of generalizability across contexts. We can notice that the new pipelines improved the performance for each target label in terms of kappa, bringing it from poor (Kappa < .20) to fair (.20 < Kappa < .40).

For individual task orientation and structuring problem solving the performance enhanced the most with the improvement of +.20 and +.30 in kappa, respectively. It can be seen that the most stable (standard deviation = .09) model for performing across contexts is the collaboration quality classification model developed with the pipeline (CON→MAX→HP→TH). In terms of accuracy, there was not much improvement, indicating models making an approximately equal number of correct classifications. However, the improvement in kappa suggests that the new models improved on classifying not just the positive class but also the negative class.

We also provide the performance measure for the base and new pipelines in terms of how much gap they were able to fill from chance to human performance (Figure 3). The new modeling pipelines enabled the models to fill more than 25% gap from chance to human performance for collaboration quality and most of its dimensions. All the best-performing pipelines have utilized contextual data features, suggesting a positive impact of contextual

[7]Kappa interpretation as per [10] guidelines
[8]Referring to structuring problem solving and time management.

Table 5: Random forest classification performance across contexts with best performing MMLA pipeline

Target	Base pipeline	More generalizable model building pipeline		
		Pipeline	Performance	
			Accuracy	Kappa
Collaboration quality	.15 (.11)	CON-Max-HP-TH	71% (14)	.27 (.09)
Argumentation	.17 (.11)	CON-Max-HP-TH	63% (18)	.26(.19)
Sustaining mutual understanding	.16 (.10)	CON-OH-Sta-HP-TH	70% (13)	.28 (.17)
Collaboration flow	.14 (.11)	CON-OH-Sta	73% (14)	.25 (.18)
Knowledge exchange	.18 (.11)	CON-OH-Sta	75% (11)	.29 (.12)
Cooperative orientation	.14 (.10)	CON-OH-Max-HP-TH	68% (18)	.28 (.19)
Individual task orientation	.10 (.10)	CON-Max-TH	71% (13)	.30 (.18)
Structuring problem solving	.07 (.09)	CON-Sta	75% (21)	.37 (.26)

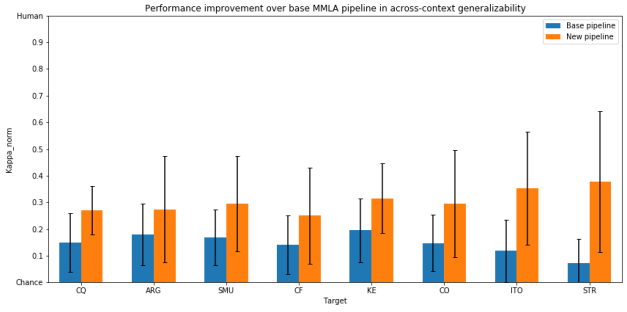


Figure 3: Improvement in model's generalizabilty across contexts (Kappa normalized with respect to human-performance); CQ: Collaboration quality, ARG: Argumentation, SMU: Sustaining mutual understanding, CF: Collaboration flow, KE: Knowledge exchange, ITO: Individual task orientation, STR: Structuring problem solving

features. Moreover, pipelines involving the use of Standard scaling had outlier handling in 3 out of 4 cases.

4.3 Statistical analysis of impact of contextual features and different modeling steps on model's generalizability (RQ3)

Table 6 reports the results of the Wilcoxon signed-rank test on the impact of contextual features and used modeling steps on the generalizability of developed models.

Table 6: Results of Wilcoxon signed-rank test

Pipeline step	Mean_diff_kappa	Z-statistic	P-value
Outlier	.002	2.24	.02 *
Scaling	.002	1.5	.13
Tuning	-0.04	-6.43	.000 *
Thresholding	.03	5.81	.000 *
Context	.05	9.48	.000 *

*:p-value < .05

All the investigated modeling steps (and use of contextual features) except data scaling were found to have a statistically significant (p-value < 0.05) impact on the model's generalizability (or performance across contexts). The use of contextual features, outlier handling, and threshold selection showed a positive impact, while hyper-parameter tuning was found to have a negative impact on the model's generalizability. We also tested these effects for their statistical significance using one-tailed test. Our results from one-tailed (diff > 0) had a p-value < .05, supporting our inference over the positive impact of data scaling, threshold selection, and use of contextual features on the model's generalizability. In the one-tailed (diff < 0), only the hyper-parameter step was found to have a statistically significant impact, supporting the claim over its negative impact on the model's generalizability.

5 DISCUSSION

This section discusses the results for each research question and also outlines the limitations of the study.

5.1 RQ1:How do collaboration estimation models which are developed using a standard MMLA pipeline perform across different contexts?

The base MMLA modelling pipeline produced models that performed better in terms of accuracy achieved for instance generalizability with respect to a baseline of chance performance. The accuracy of classification models (for collaboration quality and its dimensions) was in the range of 63%-83%. The kappa measure, however, was in the range of .15-.25, mostly having a fair level (.20 < Kappa, according to [10]) of agreement between the ground truth and predicted labels. The models were able to retain their performance across contexts in terms of accuracy metric with a slight decline of 1%-6% in the performance. However, the overall performance in classifying positive and negative classes degraded badly (Kappa < .20, slight agreement).

The difference between the accuracy and the kappa measures of model performance can be explained by the uneven distribution of classes in our dataset. For collaboration quality and its dimensions, the distribution was approximately 70% to 30% on positive v/s negative class. If we then take a base model which always predicts the majority class, we will end up achieving 70% accuracy. That is what the models developed with the MMLA base pipeline

achieved, which suggests that the developed models did not do much better than the base model. These differences, however, become more visible with kappa. Thus, our results highlight why the use of accuracy is inadequate in MMLA where class labels are not necessarily evenly distributed.

One possible explanation for the model's poor performance in terms of kappa < .20 at the context level could be due to the datasets from contexts varying in multiple aspects (e.g., classroom level, teacher, subject, etc.). This makes the classification task difficult, even for human experts as shown in our inter-rater agreement scores for all the dimensions of collaboration quality (.60 < kappa < .80).

5.2 RQ2:Which pipeline offers further improvement in the model's performance across contexts, in other words, its generalizability?

Our exploration of different modelling pipelines enabled us to identify high-performing modelling pipelines which improved the generalizability of the models. Major improvements were seen for individual task orientation and structuring problem solving models. For individual task orientation, performance across contexts improved from kappa of .10 to .30 using a pipeline with the following steps: use of contextual data → MaxAbs scaling → hyper-parameter tuning → threshold selection. For structuring problem solving the use of contextual data with standard scaling resulted in an improvement of kappa from .07 to .37. The variation in the performance (across contexts) of structuring problem-solving model was the highest with .26. Simultaneously, the collaboration quality model was the most stable in terms of generalizing across contexts (.09).

The structuring problem solving dimension is comparably difficult to learn as also shown in our inter-rater agreement scores, achieving the lowest (kappa = .65) for the same dimension. Moreover, given the use of simple features, the model learned from contextual features which were observed in the feature importance exploration (e.g., there were 4 contextual features among the top 10 important features). This might have caused instability in the generalizability of the model. Notwithstanding, for collaboration quality and individual task orientation models, there were respectively zero and only one contextual feature in the top 10 important features. It might suggest that the use of contextual features in a complementary way with multimodal features can lead to a stable generalizable model. For example, the collaboration quality model was mainly harnessing multimodal data features (10 most important features were multimodal features) and additionally leveraging contextual features with relatively less importance in prediction.

We would like to point out that the models themselves are more of an illustration than a contribution. The features used in the paper were selected based on prior research and also because of their simplicity. This can make the developed models easily interpretable which is of significant importance in educational settings. We deem a more important contribution of our presented methodology which despite using simple features led to significant model performance improvement when evaluated across contexts.

5.3 RQ3:What is the impact of adding/removing a modelling step and contextual features on the model's performance across contexts?

Our results showed that outlier handling, use of contextual features, and threshold selection steps positively impact a model's generalizability. One of the possible explanations for improvement in the model's performance across contexts because of outlier handling is that our features (e.g., number of characters added or deleted) had some extreme values. For example, when the student did a copy-paste operation, it resulted in tracking a high number of characters added in a very short time. Having those extreme values in the datasets and not handling them before performing scaling is most likely to affect the data. Therefore, having an outlier handling step might avoid this issue.

The use of context data positively impacts the generalizability of the model. The potential of contextual information has already been advocated in the field of Learning Analytics (LA) [24] and MMLA [27] for understanding learning behavior (e.g., performance). With our results, we further provide evidence on the positive impact of using contextual information in model building and its benefit in improving the generalizability of the model across contexts.

For threshold selection, we were anticipating an improvement because the class distribution in our datasets was uneven. Thus, the use of 0.5 for decision-making was not necessarily an adequate option. The threshold selection step tries to mitigate the issue of uneven class distribution by learning the threshold from training data.

There was no statistically significant difference found in the model's performance across contexts as a result of adding MaxAbs scaling. This is because of the insensitivity of tree-based machine learning algorithms, random forest in our case, towards data scaling.

The results for hyperparameter tuning were surprising. We were expecting it to improve the model's performance. However, this step had a negative impact on the model's generalizability. This could be due to the introduction of an overfitting effect when the hyperparameter tuning step selected a model only performing well for a particular set of datasets. Moreover, the small size of the datasets may have increased the likelihood of overfitting the model in the training phase.

5.4 Limitations

Our work has six main limitations. The first limitation is our selection of contextual aspects to determine the difference in contexts. This selection, along with our definition of context, is debatable. The second limitation is that most of the datasets we collected were from English language classrooms restricting the variability of contexts. Other contextual aspects could be regarded for differentiation of contexts. Thus, a more systematic evaluation is needed in a wider range of contexts to address these two limitations. The third limitation is that the reported results should be interpreted in the context of a specific type of collaborative learning activity with a specific tool. This does not allow our findings (or our resulting models) to be necessarily applicable to other types of collaborative learning activities. To validate our findings, we will include other types of group activities in our future studies. The fourth limitation is with

our use of simple features for modeling collaboration. We acknowledge that the used features are unlikely to generalize to contexts different on a higher level (e.g., different schools) and would require the use of additional features (e.g., language features). Furthermore, our use of the contextual features, in particular teacher_id and classroom_id, limits the applicability scope of developed models to contexts having the same set of teachers and classrooms. The fifth limitation is the use of a specific machine learning algorithm (random forest). There is a possibility that with the use of a different machine learning algorithm, a different modelling pipeline could work better in developing generalizable models in other particular contexts. Hence, the exploration of used pipelines with a different set of machine learning algorithms would be required. Additionally, we did not explore other pipeline steps (e.g., feature merging techniques, different window-size merging, etc.) in our current investigation of pipelines to keep the focus on the investigated modelling steps. We plan to include those steps in our pipeline exploration toward collaboration quality modelling in our future studies. Our longer-term goal is to add more generalizable guiding functionalities to CoTrack with the purpose of supporting teachers in the classroom during collaborative learning activities.

6 CONCLUSION

This paper addressed the gap on the lack of MMLA research in building generalizable models for classifying collaboration quality and its dimensions. Our findings offer insights into where the current model development pipeline in MMLA falls short in building a generalizable model. We also shed light on how it can be addressed with the inclusion of contextual features along with additional steps in the modelling pipeline. The findings presented in this paper along with the used methodology take a step towards building generalizable collaboration quality models and understanding the generalizability of current (state-of-the-art) models, which is critical for real-world implementation of MMLA. The development of such models could further help the community to achieve the goal of building automated guiding tools to help teachers better understand and guide students during collaborative learning activities. We hope that other MMLA researchers will take up the exhaustive exploration of pipelines and results presented in this paper so that we can better understand how far we are from human performance.

ACKNOWLEDGMENTS

The research presented in the paper has received partial funding from Estonian Research Council's Personal Research Grant (PRG) project PRG1634.

REFERENCES

[1] Nikoletta Bassiou, Andreas Tsiartas, Jennifer Smith, Harry Bratt, Colleen Richey, Elizabeth Shriberg, Cynthia D'Angelo, and Nonye Alozie. 2016. Privacy-Preserving Speech Analytics for Automatic Assessment of Student Collaboration. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, Nelson Morgan (Ed.). ISCA, San Francisco, CA, USA, 888–892. https://doi.org/10.21437/Interspeech.2016-1569

[2] Paulo Blikstein and Marcelo Worsley. 2016. Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics* 3, 2 (Sep. 2016), 220–238. https://doi.org/10.18608/jla.2016.32.11

[3] Pankaj Chejara, Luis P. Prieto, Adolfo Ruiz-Calleja, María Jesús Rodríguez-Triana, Shashi Kant Shankar, and Reet Kasepalu. 2020. Quantifying Collaboration Quality in Face-to-Face Classroom Settings Using MMLA. In *Collaboration Technologies and Social Computing - Proceedings of the 26th International Conference, CollabTech 2020 (Lecture Notes in Computer Science, Vol. 12324)*, Alexander Nolte, Claudio Alvarez, Reiko Hishiyama, Irene-Angelica Chounta, María Jesús Rodríguez-Triana, and Tomoo Inoue (Eds.). Springer, Tartu, Estonia, 159–166. https://doi.org/10.1007/978-3-030-58157-2_11

[4] Pankaj Chejara, Luis Pablo Prieto, Adolfo Ruiz-Calleja, María Jesús Rodríguez-Triana, Shashi Kant Shankar, and Reet Kasepalu. 2021. CoTrack2: A Tool to Track Collaboration Across Physical and Digital Spaces with Real Time Activity Visualization. In *Companion Proceedings of 11th International Conference on Learning Analytics & Knowledge LAK21*. 406 – 406. https://www.solaresearch.org/wp-content/uploads/2021/04/LAK21_CompanionProceedings.pdf

[5] Pankaj Chejara, Luis P. Prieto, Adolfo Ruiz-Calleja, María Jesús Rodríguez-Triana, Shashi Kant Shankar, and Reet Kasepalu. 2021. EFAR-MMLA: An Evaluation Framework to Assess and Report Generalizability of Machine Learning Models in MMLA. *Sensors* 21, 8 (Apr 2021), 2863. https://doi.org/10.3390/s21082863

[6] Matthias Feurer and Frank Hutter. 2019. Hyperparameter optimization. In *Automated machine learning*. Springer, Cham, 3–33.

[7] Elizabeth A. Freeman and Gretchen G. Moisen. 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling* 217, 1-2 (2008), 48–58. https://doi.org/10.1016/j.ecolmodel.2008.05.015

[8] Shuchi Grover, Marie A. Bienkowski, Amir Tamrakar, Behjat Siddiquie, David A. Salter, and Ajay Divakaran. 2016. Multimodal analytics to study collaborative problem solving in pair programming. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge, LAK 2016*, Dragan Gasevic, Grace Lynch, Shane Dawson, Hendrik Drachsler, and Carolyn Penstein Rosé (Eds.). ACM, Edinburgh, United Kingdom, 516–517. https://doi.org/10.1145/2883851.2883877

[9] Reet Kasepalu, Luis P. Prieto, Tobias Ley, and Pankaj Chejara. 2022. Teacher Artificial Intelligence-Supported Pedagogical Actions in Collaborative Learning Coregulation: A Wizard-of-Oz Study. *Frontiers in Education* 7 (2022). https://doi.org/10.3389/feduc.2022.736194

[10] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977).

[11] Yang Liu, Tingting Wang, Kun Wang, and Yu Zhang. 2021. Collaborative Learning Quality Classification Through Physiological Synchrony Recorded by Wearable Biosensors. *Frontiers in Psychology* 12, April (2021), 1–12. https://doi.org/10.3389/fpsyg.2021.674369

[12] Roberto Martínez Maldonado, Yannis A. Dimitriadis, Alejandra Martínez-Monés, Judy Kay, and Kalina Yacef. 2013. Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *Int. J. Comput. Support. Collab. Learn.* 8, 4 (2013), 455–485. http://dblp.uni-trier.de/db/journals/cscl/cscl8.html#MaldonadoDMKY13

[13] Katerina Mangaroska, Kshitij Sharma, Dragan Gaševic, and Michalis Giannakos. 2020. Multimodal Learning Analytics to Inform Learning Design: Lessons Learned from Computing Education. *Journal of Learning Analytics* 7, 3 (2020), 79–97.

[14] Roberto Martinez-Maldonado. 2019. A handheld classroom dashboard: Teachers' perspectives on the use of real-time collaborative learning analytics. *International Journal of Computer-Supported Collaborative Learning* 14, 3 (2019), 383–411.

[15] Roberto Martínez Maldonado, James R. Wallace, Judy Kay, and Kalina Yacef. 2011. Modelling and Identifying Collaborative Situations in a Collocated Multi-display Groupware Setting. In *Artificial Intelligence in Education - 15th International Conference, AIED 2011 (Lecture Notes in Computer Science, Vol. 6738)*, Gautam Biswas, Susan Bull, Judy Kay, and Antonija Mitrovic (Eds.). Springer, Auckland, New Zealand, 196–204. https://doi.org/10.1007/978-3-642-21869-9_27

[16] Charles I Mosier. 1951. The need and means of cross validation. I. Problems and designs of cross-validation. *Educational and Psychological Measurement* 11 (1951), 5–11.

[17] Yukiko I. Nakano, Sakiko Nihonyanagi, Yutaka Takase, Yuki Hayashi, and Shogo Okada. 2015. Predicting Participation Styles Using Co-Occurrence Patterns of Nonverbal Behaviors in Collaborative Learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 91–98. https://doi.org/10.1145/2818346.2820764

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] Víctor Ponce-López, Sergio Escalera, and Xavier Baró. 2013. Multi-modal social signal analysis for predicting agreement in conversation settings. In *2013 International Conference on Multimodal Interaction, ICMI '13*, Julien Epps, Fang Chen, Sharon L. Oviatt, Kenji Mase, Andrew Sears, Kristiina Jokinen, and Björn W. Schuller (Eds.). ACM, Sydney, NSW, Australia, 495–502. https://doi.org/10.1145/2522848.2532594

[20] Sambit Praharaj, Maren Scheffel, Hendrik Drachsler, and Marcus Specht. 2021. Literature review on co-located collaboration modeling using multimodal learning analytics—can we go the whole nine yards? *IEEE Transactions on Learning Technologies* 14, 3 (2021), 367–385.

[21] Luis Pablo Prieto, Kshitij Sharma, Lukasz Kidzinski, María Jesús Rodríguez-Triana, and Pierre Dillenbourg. 2018. Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *J. Comput. Assist. Learn.* 34, 2 (2018), 193–203. https://doi.org/10.1111/jcal.12232

[22] Samuel L Pugh, Arjun Rao, Angela E.B. Stewart, and Sidney K. D'Mello. 2022. Do Speech-Based Collaboration Analytics Generalize Across Task Contexts?. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, Online, USA, 208–218. https://doi.org/10.1145/3506860.3506894

[23] Joseph M. Reilly and Bertrand Schneider. 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*, Michel C. Desmarais, Collin F. Lynch, Agathe Merceron, and Roger Nkambou (Eds.). International Educational Data Mining Society (IEDMS), Montréal, Canada, 149–157. https://drive.google.com/file/d/1CvvFOlkdJYHWqpPriMbYw09_ZcKUqcmJ

[24] María Jesús Rodríguez-Triana, Alejandra Martínez-Monés, Juan I. Asensio-Pérez, and Yannis A. Dimitriadis. 2015. Scripting and monitoring meet each other: Aligning learning analytics and learning design to support teachers in orchestrating CSCL situations. *Br. J. Educ. Technol.* 46, 2 (2015), 330–343. https://doi.org/10.1111/bjet.12198

[25] Nikol Rummel, Anne Deiglmayr, Hans Spada, George Kahrimanis, and Nikolaos Avouris. 2011. *Analyzing collaborative interactions across domains and settings: An adaptable rating scheme.* Springer US, Boston, MA, 367–390.

[26] Bertrand Schneider, Gahyun Sung, Edwin Chng, and Stephanie Yang. 2021. How Can High-Frequency Sensors Capture Collaboration? A Review of the Empirical Links between Multimodal Metrics and Collaborative Constructs. *Sensors* 21 (2021), 32 pages.

[27] Shashi Kant Shankar, María Jesús Rodríguez-Triana, Adolfo Ruiz-Calleja, Luis P. Prieto, Pankaj Chejara, and Alejandra Martínez-Monés. 2020. Multimodal Data Value Chain (M-DVC): A Conceptual Tool to Support the Development of Multimodal Learning Analytics Solutions. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* 15, 2 (2020), 113–122. https://doi.org/10.1109/RITA.2020.2987887

[28] Kshitij Sharma, Evangelos Niforatos, Michail Giannakos, and Vassilis Kostakos. 2020. Assessing Cognitive Performance Using Physiological and Facial Features. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 41 pages. https://doi.org/10.1145/3411811

[29] Kshitij Sharma, Zacharoula Papamitsiou, and Michail Giannakos. 2019. Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology* 50, 6 (2019), 3004–3031. https://doi.org/10.1111/bjet.12854

[30] Daniel Spikol, Emanuele Ruffaldi, and Mutlu Cukurova. 2017. Using Multimodal Learning Analytics to Identify Aspects of Collaboration in Project-Based Learning. In *12th International Conference on Computer Supported Collaborative Learning, CSCL 2017*, Brian K. Smith, Marcela Borge, Emma Mercier, and Kyu Yon Lim (Eds.). International Society of the Learning Sciences, Philadelphia, Pennsylvania, USA, 263–270. https://repository.isls.org/handle/1/240

[31] Daniel Spikol, Emanuele Ruffaldi, Giacomo Dabisias, and Mutlu Cukurova. 2018. Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning* 34, 4 (2018), 366–377.

[32] Neomy Storch. 2001. How collaborative is pair work? ESL tertiary students composing in pairs. *Language teaching research* 5, 1 (2001), 29–53.

[33] Sree Aurovindh Viswanathan and Kurt VanLehn. 2018. Using the Tablet Gestures and Speech of Pairs of Students to Classify Their Collaboration. *IEEE Trans. Learn. Technol.* 11, 2 (2018), 230–242. https://doi.org/10.1109/TLT.2017.2704099

[34] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer, 196–202.