Impact of window size on the generalizability of collaboration quality estimation models developed 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

Mohammad Khalil
mohammad.khalil@uib.no
University of Bergen
Bergen, Norway

ABSTRACT

Multimodal Learning Analytics (MMLA) has been applied to collaborative learning, often to estimate collaboration quality with the use of multimodal data, which often have uneven time scales. The difference in time scales is usually handled by dividing and aggregating data using a fixed-size time window. So far, the current MMLA research lacks a systematic exploration of whether and how much window size affects the generalizability of collaboration quality estimation models. In this paper, we investigate the impact of different window sizes (e.g., 30 seconds, 60s, 90s, 120s, 180s, 240s) on the generalizability of classification models for collaboration quality and its underlying dimensions (e.g., argumentation). Our results from an MMLA study involving the use of audio and log data showed that a 60 seconds window size enabled the development of more generalizable models for collaboration quality (AUC 61%) and argumentation (AUC 64%). In contrast, for modeling dimensions focusing on coordination, interpersonal relationship, and joint information processing, a window size of 180 seconds led to better performance in terms of across-context generalizability (on average from 56% AUC to 63% AUC). These findings have implications for the eventual application of MMLA in authentic practice.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in collaborative and social computing*; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

MultiModal Learning Analytics, Machine Learning, Collaboration Quality, Generalizability, Temporal Window

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

ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576143

ACM Reference Format:
Pankaj Chejara, Luis P. Prieto, María Jesús Rodríguez-Triana, Adolfo Ruiz-Calleja, and Mohammad Khalil. 2023. Impact of window size on the generalizability of collaboration quality estimation models developed 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, 7 pages. https://doi.org/10.1145/3576050.3576143

1 INTRODUCTION

Collaboration is an essential skill in the 21st Century [6]. To develop this skill among students, collaborative activities are often combined with other pedagogical approaches (e.g., project-based learning [23]) in teaching practices. In such practices, teachers are by default expected to orchestrate and monitor group activities which are extremely difficult [4]. In this direction, the automation of collaboration estimation holds the potential for supporting teachers with the development of monitoring tools [3–5, 9, 17].

There has been a growing interest in automated estimation of collaboration [3, 17]. For example, a tool that identifies low collaboration quality can help the teacher identify the group that needs support in the classroom. The development of such tools demands capturing data from the physical space in addition to the digital space which traditional (log-based) Learning Analytics (LA) fulfills only partially. To address this limitation of capturing physical space, researchers have started employing other data sources (e.g., audio [25], video [23]) in addition to logs, to capture collaboration more holistically. This research field that involves the use of multiple data sources is known as MultiModal Learning Analytics (MMLA) [2].

Earlier MMLA research works have provided preliminary evidence on the feasibility of automating the estimation of collaboration quality (or other aspects of collaboration) using multimodal data (audio and logs) in face-to-face (FtoF) settings [11, 14]. This research has been advanced by MMLA researchers, exploring a variety of modeling techniques (e.g., Random forest [25], Adaboost [22]) with different types of data (e.g., audio [25], eye-gaze [18], video [23]). Furthermore, MMLA work from Olsen et al. [15] on collaboration detection reported performance gains with the use of multimodal data models over models built with unimodal data. These research works' findings suggest the use of MMLA in automating collaboration detection for FtoF settings.

The challenges involved with multimodal data processing and analysis, however, hinder the use of MMLA despite the aforementioned benefits. It asks researchers to make several decisions while building automated models for collaboration behavior. These decisions include a selection of data sources, data features, the type of feature merging, the time scale (window size) on which to merge features, modeling techniques, and model evaluation strategies.

MMLA researchers have taken a closer look at the aforementioned steps (see above) of model building to simplify the use of MMLA. In this direction, Schneider et al. [21] and Di Mitri et al. [7] have offered an in-depth analysis of the use of various data sources in MMLA; Praharaj et al. [17] have provided an analysis of multimodal features used for collaboration modeling; Mu et al. [13] have looked into multimodal feature merging techniques; Chejara et al. [3] have analyzed different model development and evaluation techniques in MMLA. These research works, in addition to others from the literature, provide the current state of the art in their respective domain and thus guide MMLA researchers to make decisions related to modeling. However, in the aforementioned steps (see above), the decision on which window size offers higher-performing collaboration estimation models is an under-explored area of research.

A few MMLA research works have shown in their preliminary analysis that the window size does impact the model's performance [11, 23]. However, those works only assessed the impact in terms of the model's performance in a single learning context[1]. There is still a lack of research on the evaluation of model performance beyond a single context (across-context generalizability) and systematic exploration of how across-context generalizability is affected by window sizes.

In this paper, we systematically investigate the impact of different window sizes (e.g., 30 seconds, 60s, 90s, 120s, 180s, 240s) on the collaboration quality estimation model's performance across learning contexts. This paper is structured into seven sections. We provide related work and point out the contribution and novelty of our work in section 2. Section 3 describes the datasets used in our investigation. In section 4, we offer details on our data preprocessing, machine learning model development, and evaluation strategies. Section 5 presents the results. In section 6, we discuss the main findings and limitations of the presented work. Section 7 concludes our work.

2 RELATED WORK

The current MMLA research has shown a growing interest in building automated models for collaboration behavior [3, 11, 12, 16, 23, 25]. A variety of data sources have been utilized, e.g., audio [12], video [23], eye-gaze [21], etc. (for more details refer to [17, 20]). The use of multiple data sources has provided data with different sampling rates, thus, on a different time granularity level. To bring data from different levels of granularity to a common level, MMLA researchers have used a windowing operation. This operation divides the multimodal data into smaller chunks by a particular size of the time window. Following the windowing operation, the aggregation of features (either individual or group-level) using various statistics (mean, standard deviation) is performed. For example, Martínez-Maldonado et al. [12] used a time window of 30 seconds to segment their dataset of audio and log data, and then aggregated the features using mean and standard deviation.

The majority of the research into modeling collaboration has used a window size smaller or equal to 60 seconds [1, 3, 8, 22, 25]. In particular, a window size of 30 seconds has been used frequently by MMLA researchers [1, 3, 22, 25]. Besides, a smaller window size of 10 seconds has been employed by Prieto et al. [18] for modeling the social plane of orchestrating collaborative learning. On the contrary, few research works [14, 23] have used larger window sizes of 400s and 240s, respectively.

Some authors [11, 23] have investigated the use of several window sizes for building their collaboration estimation models. For example, Martínez-Maldonado et al. [11] in their investigation of three window sizes (30, 60, and 90 seconds) found that the 30 seconds time window enabled the development of high-performing models for classifying collaboration quality. Similarly, Spikol et al. [23] investigated the role of three window sizes (120s, 240s, and 360s) and found that a window size of 240 seconds offered better performing models in classifying group artifacts' quality. The results of these research works suggest that the window size has an impact on the collaboration estimation model's performance. Therefore, choosing an adequate window size is essential.

The current MMLA research on automating collaboration estimation falls short on two fronts. First, there is a lack of research on the systematic exploration of window size for modeling collaboration quality estimation models. Second, an across contexts model evaluation (in general and) with the use of different window sizes is currently absent in MMLA.

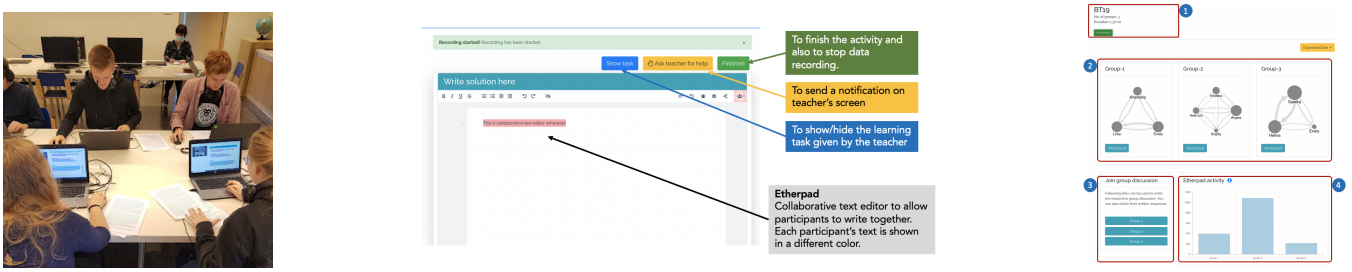
This paper, thus, sets out to analyze the impact of various window sizes (30 seconds, 60s, 90s, 120s, 180s, and 240s) on the performance of machine learning models in classifying collaboration quality. The heterogeneity of the resulting window in prior research also motivated us to get a broader, more systematic understanding of what window size is suitable for what dimension/aspect of collaboration quality from a generalizability perspective. Therefore, we also decided to investigate the impact of window size on the underlying dimensions/aspects of collaboration quality as per the Rummel et al. [19] framework. Moreover, the development of more generalizable models for collaboration quality dimensions could enable the building of automated guiding tools for teachers' support [10].

To the best of our knowledge, this investigation of window size impact on the classification models for *multiple collaboration quality dimensions/aspects* using a *multimodal dataset* has not been done yet. Moreover, this impact in terms of *generalizability* of models is also missing from the current research. Thus, this paper offers insights into what window size is adequate for building more generalizable classification models for collaboration quality and its dimensions using multimodal datasets.

3 DATASETS

The datasets used in this study were collected as part of our previous study focusing on investigating teachers' perspectives on

[1]Here, we consider a learning context composed of multiple aspects, e.g., particular (groups of) students, a particular 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 contexts will be different in terms of learning activity.



(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

multimodal analytics for collaborative learning activity monitoring and guiding [10].

3.1 Study contexts

The study took place during collaborative learning activities in three different subject classrooms in an Estonian vocational school. The subjects were Math, Woodwork, and Estonian language. We treated these as three different contexts because of their variation in multiple aspects, i.e., type of activity, teacher, classroom, and subject. Students and teachers used the Estonian language for communication. There were 12 groups of varying group sizes (e.g., 2,3,4) and the activity duration was 45-60 minutes.

3.2 Data collection tool

The study was conducted with a tool called CoTrack[2]. CoTrack is a web-based application that allows teachers to create collaborative learning activities with monitoring functionality. It offers a collaborative writing space for groups to draft the solution to a given problem together. CoTrack also records every writing activity and students' 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, student learning environment in CoTrack, and teacher's dashboard.

3.3 Activity tasks

The study was conducted in Mathematics, Woodwork, and Estonian language classroom sessions. For the Mathematics classroom, the collaborative activity involved solving a set of geometric problems. Each group was given a similar set of problems but with different measurements. For example, one problem for group 3 was to calculate the perimeter and area of a rectangle with a diagonal of 84 cm forming an angle of 25 degrees with a larger side. The task for Woodwork involved a hypothetical situation of a person, Steve, who needed to renovate a particular portion of his house (exterior facade, bathroom, and 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 labor and materials, and prepare the final document with all details for Steve. The task for the Estonian language learning involved preparing a presentation in the group on one of the epic[3] 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 summarize the central story of the epic. At the end of the session, the groups were asked to present in front of the class.

3.4 Procedure

We designed all three collaborative learning activities beforehand for the research study. A researcher from educational science and the concerned teacher were involved in the learning design. The same researcher was present during the enactment of the study in the classroom and briefed the students about the purpose of the study. The consents were taken from the students (consents were taken before from the parents in case of students younger than 18 years). Students were grouped by the teacher and then asked to complete the given activity. Each student had a laptop and a microphone for the activity (see Figure 1a).

3.5 Features

We extracted speaking time and turn-taking from audio using the Voice Activity Detection algorithm which detected voice activity every 200 ms. These features (e.g., speaking time) are found to be predictors of collaboration quality in MMLA [11, 16, 17]. From speech data, we extracted the frequency of "I" and "we" (as in [24]) and the wh-words, all in Estonian. From writing logs, we extracted the number of characters written or deleted by group participants. We further took the average and standard deviation for each extracted feature to compute group-level features.

4 METHODS

4.1 Annotation

We used a rating scheme from [19] to obtain the ground truth of collaboration quality and its underlying dimensions. This rating

[2] https://www.cotrack.website

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

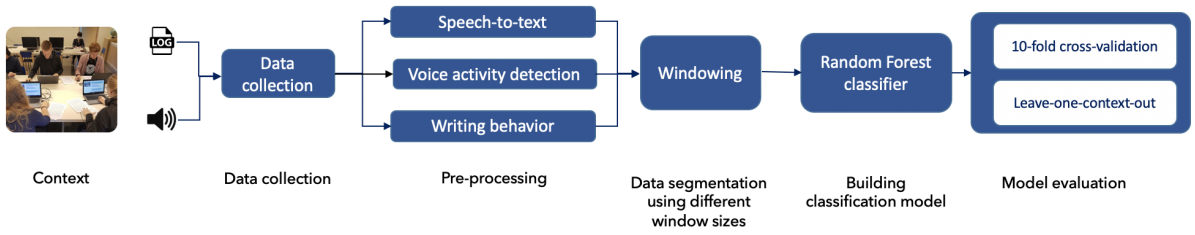


Figure 2: Model development process

scheme assigns scores on a 5-point scale (i.e., -2,-1,0,1,2) for seven dimensions of collaboration quality. These dimensions are argumentation, sustaining mutual understanding, cooperative orientation, structuring problem solving and time management, individual task orientation, knowledge exchange, and collaboration flow. The inter-rater reliability score (Cohen's Kappa >.60) was at a substantial level for all seven dimensions of collaboration quality. We added all seven dimensions' scores and averaged them to compute a measure of collaboration quality following prior research work in MMLA (e.g.,[12]).

4.2 Dataset generation for different window sizes

We used a rolling window operation, following [12], to generate datasets with window sizes of the 30s, 60s, 90s, 120s, 180s, and 240s. As the annotation was done for every 30s, we merged labels in the following manner: we assigned a final label of 'High' if 50% or more labels of the same were observed in windows under rolling operation. For example, consider three consecutive 30s windows with labels[4] High, Low, Low; for generating a dataset for 60s window size, the rolling operation took the first two labels and merged them into a label of High; while for the 90s dataset, the final label was Low.

4.3 Model development

Figure 2 shows our model-building process involving data collection, feature extraction, segmentation by different window sizes, training classification models, and then evaluating the developed models. We employed the random forest algorithm to build[5] classification models for collaboration quality and its dimensions. We decided to use the random forest algorithm for two reasons: first, this algorithm has been found to achieve high performance in the field of MMLA [1, 18, 25]; second, we also found the random forest to achieve comparatively better performance in estimating collaboration quality and its dimension [3].

For model evaluation, we used 10-fold cross-validation (CV) and leave-one-context-out [3]. The first evaluation scheme divided the datasets into 10 equal portions, using 9 portions for training and 1 for testing. This process is iterated 10 times with the selection of a different portion for testing. In educational terms, the 10-fold CV assesses how well a model predicts a situation from the same set of contexts it has been trained on which is unrealistic in practice (i.e., in practice every classroom activity will be in a different context as per our definition). Thus, it offers a measure for within-context generalizability. On the contrary, in leave-one-context-out, datasets from two contexts were used for training and other for testing. This was iterated three times. This evaluates the models for their generalizability to a different learning context which has never been seen by the model before. Thus, it provides a measure of across-context generalizability that is more relevant to MMLA, i.e., after the model is put into practice, it will see a completely different context each time it is used in the classroom by teachers.

We report the results of models' performance in terms of Area Under the ROC Curve (AUC) which takes into account true positive rate and false positive rate. The AUC scores range from 0 to 1 which we scaled between 0 and 100% with 0 representing a model making all wrong predictions and 100% when the model makes all correct predictions. An AUC score of 50% represents chance performance.

5 RESULTS

Table 1 presents the AUC score of random forest classifiers for collaboration quality and its dimensions, evaluated with a 10-fold CV and leave-one-context-out. On 10-fold CV, the classification models on average improved their performance from 64% AUC to 89% AUC when developed with the 30s and 240s window sizes, respectively. All the developed models achieved their highest performance or within-context generalizability with the use of a 240s window size. In particular, models of collaboration flow, collaboration quality, and sustaining mutual understanding performed comparably higher (AUC score 91%) than other models with a 240s window size. On contrary, the structuring problem solving[6] model achieved the lowest AUC score (87%) with a 240s window.

In across contexts evaluation, on average, models achieved their lowest performance at 30s window size (AUC 56%) and highest at 180s window size (AUC 63%). These results showed a steep degradation in the model's across-context generalizability compared to their within-context performance. Besides, the use of different window sizes led to higher performance for models of collaboration quality and its dimensions. For example, the argumentation model and collaboration quality model achieved their highest across contexts performance of 64% and 61% AUC scores, respectively, with a 60s window size. While classification models for collaboration flow,

[4]For 30s window size, we mapped scores below or equal to zero as 'Low' otherwise 'High'.

[5]Source code: https://github.com/pankajchejara23/Time-window-size-impact-on-model-performance

[6]Referring to structuring problem solving and time management

Table 1: Random forest model's performance within and across contexts (AUC scaled in the range of 0-100%)

Target	Within-context						Across-context					
	30s	60s	90s	120s	180s	240s	30s	60s	90s	120s	180s	240s
ARG	61 (4)	69 (5)	71 (3)	76 (5)	85 (4)	**88 (3)**	57 (6)	**64 (6)**	62 (5)	60 (5)	61 (7)	59 (7)
CF	65 (5)	70 (4)	74 (6)	77 (3)	86 (4)	**91 (2)**	56 (5)	60 (4)	61 (4)	63 (7)	**64 (6)**	62 (7)
CO	66 (5)	70 (5)	74 (6)	78 (3)	86 (4)	**91 (3)**	56 (5)	62 (4)	60 (4)	62 (10)	**65 (7)**	62 (6)
CQ	65 (6)	70 (5)	71 (7)	76 (5)	81 (6)	**88 (4)**	55 (4)	**61 (5)**	57 (3)	59 (6)	60 (6)	58 (6)
ITO	64 (6)	69 (5)	73 (4)	76 (4)	85 (3)	**90 (2)**	61 (2)	60 (4)	**62 (5)**	61 (6)	**62 (6)**	62 (6)
KE	64 (5)	69 (6)	74 (5)	75 (4)	84 (4)	**90 (3)**	60 (4)	59 (5)	64 (9)	63 (9)	**65 (7)**	64 (6)
SPST	65 (7)	65 (6)	69 (6)	73 (2)	83 (4)	**87 (4)**	55 (3)	64 (7)	62 (7)	60 (6)	**66 (7)**	61 (3)
SMU	64 (5)	70 (4)	74 (4)	80 (3)	86 (4)	**91 (3)**	55 (4)	60 (4)	60 (6)	62 (8)	65 (8)	**67 (10)**
	64	69	72	76	84	89	56	61	61	61	63	61

ARG: Argumentation, CF: Collaboration flow, CO: Cooperative orientation, CQ: Collaboration quality, ITO: Individual task orientation, KE: Knowledge exchange, SPST: Structuring problem solving and time management, SMU: Sustaining mutual understanding

cooperative orientation, knowledge exchange, and structuring problem solving dimensions, achieved a higher AUC score (64%, 65%, 65%, 66%, respectively) for a window size of 180s. For sustaining mutual understanding, the model showed a higher performance (AUC 67%) at a 240s window size, but with comparably high variation. The classification model for individual task orientation showed the most stable performance (60% to 62%) across different window sizes in comparison with other dimension models. It achieved an AUC score of 62% with minimal variation on the 90s window.

6 DISCUSSION

We present here the main findings from our study and their implications for the MMLA research community.

6.1 Main findings

A larger window size of 240 seconds seems better for building classification models for collaboration quality and its dimensions when the goal is to use the models on contexts very similar to the one's model was trained on.

On 10-fold CV, our results showed that having a larger window size of 240s for data segmentation helps the model to achieve higher performance. In educational terms, however, it only indicates that the models will achieve similar results if performed on data coming from the same learning contexts. This is highly unlikely because each time a model is applied, it will see a different context (as per our aforementioned criteria of contextual differences). Nevertheless, the results from the 10-fold CV allow the researcher to see how well a model fits the data and also help in understanding the predictive power of used features. Therefore, our results suggest the use of the 240s window size if the goal of developing the model is to understand the feature importance for a particular context towards collaboration prediction.

A window size of 60 seconds seems better for building more generalizable collaboration quality and argumentation classification models.

Our results showed that a window size of 60 seconds is better than other window sizes for building more generalizable models for collaboration quality (AUC 64%) and argumentation dimension (AUC 61%). These results are consistent with Chounta and Avouris [4] work from Learning Analytics on developing classification models for collaboration quality and its dimensions using log-based features (e.g., number of chat messages). Their work found a 60 seconds window size better for modeling collaboration quality. Our finding on the collaboration quality model, however, is discordant with Martínez-Maldonado et al. [12] research work in which a 30s window size was found as a better window size (compared to 60s and 90s) for building a collaboration quality model on the 10-fold CV. Their work, however, did not assess models for across-context generalizability. The differences could be further explained by the type and time duration of collaborative learning activity. Martínez-Maldonado et al. [12] investigated groups, performing a job scheduling task using a specific software requiring them to interact using a mouse. The activity duration in their case was comparatively shorter than ours (17 minutes < 45-60 minutes).

A window size of 180s/240 seconds seems better for modeling collaboration flow, knowledge exchange, cooperative orientation, structuring problem solving, and sustaining mutual understanding dimensions.

Our results showed that the classification models for collaboration flow, knowledge exchange, cooperative orientation, and structuring problem solving dimensions performed comparably better across contexts when using a window size of 180s. These four dimensions cover three aspects of collaboration quality: joint-information processing, interpersonal relationship, and coordination [19]. These aspects are complex to understand and also require the qualitative aspect of students' interaction [4]. However, our features were mainly quantitative and included very simple speech features (e.g., frequency of 'Ma'). For the sustaining mutual understanding (SMU) dimension, our models with 240s window size datasets achieved high performance compared to other window sizes. This performance gain with a larger window size (180s/240s) could be explained by the sparsity of our features (e.g., turn-taking) for smaller window sizes which might have been addressed with the use of a larger window size, allowing the model to learn effectively. Our finding on cooperative orientation and structuring problem solving dimensions (i.e., 180s window size) is consistent with Chounta and Avouris [4] research.

A window size of the 90s or 180s seems better for modeling the motivation aspect of collaboration quality. In the case of individual task orientation, the models achieved the most stabilized performance across all window sizes, achieving higher performance on the 90s window size and 180s window size (AUC score 62%). This stabilized performance indicates that the changes in window size do not have a significant impact on the performance of the classification model for individual task orientation. Thus, a smaller or larger window size could be appropriate for modeling the motivation aspect of collaboration quality.

Need for more across contexts model evaluation in MMLA. Our results showed that models in general performed very well, achieving an average score of 89% AUC in 10 fold-CV (with 240s window size). However, this performance deteriorated when the same models were evaluated across contexts. This highlights the shortcomings of the frequently employed K-fold CV in MMLA [3] and raises the need to perform a stricter evaluation across contexts. This relates to the evaluation framework proposed in MMLA, suggesting multi-level (e.g., across groups, contexts) generalizability evaluation [3]. There is a need to assess models for across-context generalizability, in MMLA as well as in LA, to enable the community to gain an understanding of how far are we from production-ready models.

6.2 Limitations and future work

The present study has five main limitations. The first limitation is the small dataset size. Even though the datasets were collected from three different contexts, they were limited in activity types, teachers, education level, and activity duration. This limits the generalizability of our findings due to a very narrow scope of explored contexts. The second limitation is related to the data used. We mainly utilized datasets of audio and log files for modeling purposes. Other types of data (particularly high-frequency sensors) may result in different window sizes. The third limitation is the use of a particular machine learning technique. We used a random forest classifier for modeling collaboration quality and its seven dimensions. There is a possibility that different window sizes than the ones we recommended might be optimal for building high-performing models with the use of other machine learning techniques. The fourth limitation is with our use of a majority voting approach for merging labels for window size larger than 30s. This approach may affect class balance, e.g., a rare occurrence of 'Low' when aggregated for a larger window size become even rarer. The fifth limitation is with our criteria for defining learning context. It is debatable what defines learning context, thus, further exploration is needed.

In our future work, we plan to run the study with a larger dataset from a wider range of learning contexts (e.g., different tasks, education level, time duration, teacher, etc.). We will also utilize other machine learning techniques (e.g., AdaBoost) to develop models and investigate how their performance will be affected by the change in window size. We will also explore other label aggregation techniques for merging labels. We then envision the development of an automated guiding system to support teachers in interventions with the help of developed classification models for collaboration quality and its dimensions.

7 CONCLUSION

Our paper fills a gap in MMLA research about the window size that enables the development of more generalizable classification models for collaboration quality. We analyzed three audio-log multimodal datasets and developed random forest classifiers for collaboration quality and each of its dimensions using different window sizes (30 seconds, 60s, 90s, 120s, 180s, 240s). The results showed that a 60s window is better suited to modeling collaboration quality and argumentation dimension. For dimensions including collaboration flow, knowledge exchange, sustaining mutual understanding, cooperative orientation, and structuring problem solving, we suggest a large size window of 180s/240s. This paper contributes to the community's knowledge by offering guidelines that can help in developing not just high-performing models but also models generalizable across contexts. We also highlight the need for more across-context evaluation which we hope will help the MMLA community to bridge the gap between MMLA research and practice.

ACKNOWLEDGMENTS

The research presented in the paper has received partial funding from Estonian Research Council's Personal Research Grant (PRG) project PRG1634, European Union's Horizon 2020 research and innovation programme under grant agreement No 856954.

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. 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

[4] Irene-Angelica Chounta and Nikolaos M. Avouris. 2015. Towards a Time Series Approach for the Classification and Evaluation of Collaborative Activities. *Comput. Informatics* 34, 3 (2015), 588–614. http://www.cai.sk/ojs/index.php/cai/article/view/3222

[5] Yi Han Victoria Chua, Justin Dauwels, and Seng Chee Tan. 2019. Technologies for Automated Analysis of Co-Located, Real-Life, Physical Learning Spaces: Where Are We Now?. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (Tempe, AZ, USA) *(LAK19)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/3303772.3303811

[6] Chris Dede. 2010. Comparing frameworks for 21st century skills. *21st century skills: Rethinking how students learn* 20, 2010 (2010), 51–76.

[7] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. 2018. From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning* 34, 4 (2018), 338–349.

[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 Sixth 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, Pankaj Chejara, Luis Pablo Prieto, and Tobias Ley. 2022. Do Teachers Find Dashboards Trustworthy, Actionable and Useful? A Vignette Study Using a Logs and Audio Dashboard. *Technol. Knowl. Learn.* 27, 3 (2022), 971–989. https://doi.org/10.1007/s10758-021-09522-5

[10] 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

[11] 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

[12] 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

[13] Su Mu, Meng Cui, and Xiaodi Huang. 2020. Multimodal data fusion in learning analytics: A systematic review. *Sensors* 20, 23 (2020), 6856.

[14] 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

[15] Jennifer K. Olsen, Kshitij Sharma, Nikol Rummel, and Vincent Aleven. 2020. Temporal analysis of multimodal data to predict collaborative learning outcomes. *Br. J. Educ. Technol.* 51, 5 (2020), 1527–1547. https://doi.org/10.1111/bjet.12982

[16] 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

[17] 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.

[18] 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

[19] 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.

[20] 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.

[21] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2015. Augmenting the senses: a review on sensor-based learning support. *Sensors* 15, 2 (2015), 4097–4133.

[22] J Smith, H Bratt, C Richey, N Bassiou, E Shriberg, A Tsiartas, C D'Angelo, and N Alozie. 2016. Spoken interaction modeling for automatic assessment of collaborative learning. *Proceedings of the International Conference on Speech Prosody* 2016-Janua (2016), 277–281. https://doi.org/10.21437/SpeechProsody.2016-57

[23] 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.

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

[25] 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