Modelling Collaborative Problem-solving Competence with Transparent Learning Analytics: Is Video Data Enough?

Mutlu Cukurova
University College London
m.cukurova@ucl.ac.uk

Qi Zhou
University College London
qtnvqz3@ucl.ac.uk

Daniel Spikol
Malmo University
daniel.spikol@mau.se

Lorenzo Landolfi
Scuola Superiore Sant'Anna
lorenzo.landolfi@santannapisa.it

ABSTRACT

In this study, we describe the results of our research to model collaborative problem-solving (CPS) competence based on analytics generated from video data. We have collected ~500 mins video data from 15 groups of 3 students working to solve design problems collaboratively. Initially, with the help of OpenPose, we automatically generated frequency metrics such as the number of the face-in-the-screen; and distance metrics such as the distance between bodies. Based on these metrics, we built decision trees to predict students' listening, watching, making, and speaking behaviours as well as predicting the students' CPS competence. Our results provide useful decision rules mined from analytics of video data which can be used to inform teacher dashboards. Although, the accuracy and recall values of the models built are inferior to previous machine learning work that utilizes multimodal data, the transparent nature of the decision trees provides opportunities for explainable analytics for teachers and learners. This can lead to more agency of teachers and learners, therefore can lead to easier adoption. We conclude the paper with a discussion on the value and limitations of our approach.

CCS Concepts

• **Applied computing → Computer-assisted instruction** • **Applied computing → Collaborative learning**

Keywords

Multimodal learning analytics; physical learning analytics; collaborative problem-solving; decision trees; video analytics.

ACM Reference format:

Mutlu Cukurova, Qi Zhou, Daniel Spikol and Lorenzo Landolfi. 2020. Modelling Collaborative Problem-solving Competence with Transparent Learning Analytics: Is Video Data Enough?. In Proceedings of ACM Learning Analytics & Knowledge conference (LAK'20). ACM, New York, NY, USA. https://doi.org/10.1145/3375462.3375484

1. INTRODUCTION

Collaborative problem-solving (CPS) is considered an essential skill for learners. However, the measurement and support of CPS are challenging for educators [1]. Nowadays, there is increasing evidence that learning analytics can provide us new means to measure and support students' interactions, collaboration and problem-solving processes [2, 3]. For instance, analytics generated from video data can help us predict students' attendance [4], attention to lectures [5, 6], and learning performance [7]. However, most available research in this area focuses on more monotonous contexts of lectures as it is a challenging task for researchers to detect and interpret complex learner interactions in dynamic classroom contexts. Here, we present our results on modelling students' learning behaviours in dynamic collaborative and their CPS competence, using analytics generated from video data. More specifically, we investigate two research questions. 1) What automated metrics from video data can be used to predict students' speaking, making, listening and watching behaviours during collaborative learning activities? 2) To what extent can video data analytics accurately predict learners' CPS competence? Although, the identification of students' complex interactions with each other, and with other resources around them, from automated metrics generated only from video data is challenging [7]; it can provide opportunities for easy-to-implement learning analytics for the measurement and support of CPS in real-world settings.

2. LITERATURE REVIEW

2.1 Learning Analytics for Co-located Collaborative Learning

Most existing learning analytics (LA) research focuses on investigating computer-based educational environments. However, majority of collaborative learning still occurs in face-to-face or blended settings. Recently, the term "*physical learning analytics*" was coined to refer to research which brings LA methods and innovations into physical learning spaces and attempts to leverage and make sense of physical data to aid teaching practices and learning processes [8]. In this section, we review the previous research that model and analyse collaborative learning and teaching experiences beyond computer-based learning environments. Previous work presented here captures learner and teacher data beyond digital spaces and excludes research studies that leverage only the digital footprints and online logs of interactions [9].

In face-to-face, co-located collaborative learning environments, students communicate and interact with their peers via speech, facial expressions and body gestures while teachers or facilitators monitor these cues and reciprocate accordingly in real-time. As Chua et al. [9] show, research in physical LA can utilise video, audio, or biometric data separately or it can utilise a combination of multiple modalities as in the case of multimodal learning analytics research [10]. There is limited research on physical collaborative learning analytics and it mainly focuses on individual student level analyses rather than group or classroom. This research is mainly at the early maturity level of generating automated metrics rather than creating effective visualisations or providing feedback and reflection opportunites to teachers and students. For instance, Grover, Bienkowski, and Tamrakar [11] collected data of video, audio, clickstream, and screen capture from the activities of pair programming. The authors have asked expert to judge the level of collaboration of each group into three levels: low, medium, and high. Then, built a model with the collected multimodal data to predict the human level judgment. The model published can predict the results better than a baseline with the accuracy of 44%. As a pilot study, this research showed the possibility of using multimodal data to predict the expert judgement of collaboration levels. Similarly, [2] conducted a study collecting multimodal data in students' physical proximity during their collaborative learning activities and analyzed the relationship between these data and

students' learning outcomes. They used cameras, wearable equipment, and an integrated development environment to collect the data about students' position, movement, speech, and their interactions with the computer. The results showed that the distance between students' hands and faces can be strong predictors of students' learning outcomes in collaborative learning. Similarly, Worsley [12] collected gesture, speech and electro-dermal activation data from pair collaboration. The gesture data was used to learn a set of canonical clusters and the relation between students' clusters and their behaviors showed that gesture data can be used to predict students' CPS behaviours. Moreover, in some previous studies biometric data was also used to analyse students' collaborative learning. Lubold and Pon-Barry [13] conducted a study related to acoustic-prosodic features with rapport in collaborative learning and found that students' pitch may be similar when they collaborated well with each other. Similarly, Dikker et al. [14] used portable electroencephalogram (EGG) to record students' brain activities and showed that, brain-to-brain synchrony can also be a possible indicator of dynamic social interaction and effective collaboration.

At a relatively more mature research level, aiming to generate visualisations from automatically generated metrics of students' collaborative learning in physical spaces, Martinez-Maldonado et al. [3] used multimodal data to record students' learning activities in healthcare simulations. Similar to group's previous work by Echeverria et al. [15], they tracked and visualized how teams of students occupy the space in healthcare simulations. Although, visualisations of how students occupied space in the learning environment were created, the authors argued that teachers need the additional contextual information to have an interpretation of the students' learning process from these visualisations.

On the other hand, some researchers focused on individual student data and analysis rather than the group data as presented previously. In order to identify the different performances and behaviours of individual students during collaborative learning, Oviatt et al. have conducted a series of studies [16-19]. The group mainly explored the differences between expert and novice students, and compared their collaborative learning behaviours. The participants were asked to solve math problems in groups of three and video, audio, and written data was collected. The authors found that expert students performed more fluently in both writing and speaking during the process of collaboration. They also found that expert students had a higher ratio of using non-linguistic symbolic representations and structured diagrams to elemental marks. Similarly, Schneider et al. [20] used eye-tracking, video, and audio data to analyze individual student's learning motivation in pair collaboration. Their results show that, using eye-tracking data only is not enough to fully present students' different levels of learning dynamics.

As the reviewed research above shows, LA research, particularly with multimodal data, can provide promising results to investigate collaborative learning in physical spaces. However, the collection and analyses of multimodal data from real-world classroom environments are challenging. On the other hand, video recording is a method which is used frequently to collect data from the classrooms to study student or teacher behaviours. Due to their low financial and technical costs, video-based analytics of collaborative learning can provide valuable opportunities for immediate real-world impact. Therefore, researchers also focused on investigating the potential of video data to analyze student behaviors in various learning activities. In the next section, we will review this research.

2.2 Video data to generate relevant metrics on learner behaviours in physical spaces

Different types of information are extracted from the video data to generate physical learning analytics. Nowadays, the most common technology implemented to generate metrics from video data on learning is face recognition. Firstly, it can contribute to identify which students appear in videos. For instance, Mao et al. [4] designed a system which uses face recognition to identify students' attendance in class. In fact, a similar approach was used in most existing learning analytics studies using video data, because in all previous studies, the first essential step to generate analytics from video data was to identify each student in the video. Secondly, head motion is another key metric which was utilised by researchers to measure and support learning. For instance, previously head motion was used to measure students' attention in lectures [5, 6]. Thirdly, gestures are frequently extracted from video data for LA. Won et al. [7] used a Kinect cam to capture teachers' and students' gestures in one-to-one tutorial settings and found that both students' and teachers' gestures can be used to predict the effectiveness of learning and teaching.

From the perspective of the educational settings, most research that uses metrics from only video data focuses on lectures. In these studies, researchers mainly analyze individual learner behaviors, such as attendance, attention, and engagement. On the other hand, very few studies focused on using video data to generate metrics on students' interaction in collaborative learning settings. For instance, Schneider and Blikstein [21] conducted a study using video data to analyze pair collaboration. Students' gestures were extracted from the 3D video data from Kinect cameras. The authors have explored whether the bimanual coordination, body synchronization, and body distance can be used to predict the learning outcomes in collaboration. The results showed that, although body distance was not strongly associated with students' learning outcomes, students with low scores tended to be further away from their partners. So far, previous studies which use only video data in collaborative learning settings were not able to generate models with high prediction accuracy, particularly compared to those studies focusing on relatively more monotonous contexts such as lectures. It is a challenging task for researchers to identify students' complex interactions with each other, and with other resources around them, from metrics generated only from video data [7]. However, this challenge if addressed to a satisfactory level, can provide opportunities for learning analytics of physical collaborative learning settings that can be more easily adopted in real-world education settings.

3. METHODOLOGY

3.1 Participants and the learning activities

The participants in this study were 18 engineering students (17 male, 1 female). Their average age was 20 years old. Participants were selected from a class of 30 students by their lecturers according to their performance and they had similar levels of domain knowledge. This was intentional to ameliorate the bias of knowledge and skill differences between the students on their CPS performance. The participants were divided into six groups to complete three sequential open-ended design tasks: 1) prototyping an interactive toy, 2) prototyping a colour sorting machine, and 3) building an autonomous car. The group members in the three activities were consistent. No specific instruction on how to allocate tasks or time was given to the learners.

3.2 Generating metrics from video data

The final dataset encompassed 15 videos. Two open source libraries were used to extract metrics from videos. A face recognition database, FaceNet, was used to assign student IDs; and OpenPose, a powerful deep learning-based library, was used to extract students' body poses from the videos. Initially, a random identifier was assigned to each person in the scene and these identifiers were associated with student ID's using FaceNet.

OpenPose was used to detect multi-person human body poses in real-time. For each video frame, OpenPose outputs the location of 18 key points of the bodies found in the scene. Based on these, 8 metrics were calculated, namely face-in-the-screen, right-hand in the screen, left-hand in the screen, both hands in the screen, hand distance of individual students, distance between bodies, distance between faces, and distance between group members' hands. The first four are metrics of frequency, and they refer to the frequency of students' body key points being found in the screen. The rest four are metrics of distance, and they refer to the distance between students' key body points. These metrics were automatically extracted from video data using Python. Next, we will describe methods used to calculate these metrics.

3.2.1 Metrics of frequency

The metrics were used to represent the frequency of a student's body point identified in the scene. To exemplify the approach, we will explain the face in screen (FIS) metric. When a student's face position as generated from OpenPose is not (0,0), it means that the student's face is found in the scene. If K(A) is the total number of windows in video A, X_P (A1, n) would be the x value of student A1's face point P in the window n. Similarly, Y_P (A1, n) would be the y value of student A1's face point P in the window n. FIS, P = 0. FIS(A1) was be defined as:

$$\text{FIS(A1)} = \frac{\sum_{n=1}^{K(A)}(x_P(A1,n)\neq 0 \text{ and } y_P(A1,n)\neq 0)}{K(A)}$$

Similarly, right-hand in screen (RHIS), left-hand in screen (LHIS) and both-hands in screen (BHIS) metrics were defined as follows:

$$\text{RHIS(A1)} = \frac{\sum_{n=1}^{K(A)}(x_{RH}(A1,n)\neq 0 \text{ and } y_{RH}(A1,n)\neq 0)}{K(A)}$$

$$\text{LHIS(A1)} = \frac{\sum_{n=1}^{K(A)}(x_{LH}(A1,n)\neq 0 \text{ and } y_{LH}(A1,n)\neq 0)}{K(A)}$$

$$\text{BHIS(A1)} = \frac{\sum_{n=1}^{K(A)}(x_{RH}(A1,n)\neq 0 \text{ and } y_{RH}(A1,n)\neq 0 \text{ and } x_{LH}(A1,n)\neq 0 \text{ and } y_{LH}(A1,n)\neq 0)}{K(A)}$$

3.2.2 Metrics of distance

Hand distance (HD) was used to represent the mean distance between a student's left hand and right hand. The distance between a student's left hand and right hand was calculated only when both hands were found in the scene. If we define T_{BH}(A1) to represent the set of windows in which student A1's both hands are found in the screen and t_{BH}(A1) as the total number of windows in this dataset. HD(A1) is calculated as below to measure student A1's mean hand distance.

$$\text{HD(A1)} = \frac{\sum_{n\in T_{BH}(A1)}\left(\sqrt{(x_{(A1)}(4,n)-x_{(A1)}(7,n))^2+(y_{(A1)}(4,n)-y_{(A1)}(7,n))^2}\right)}{t_{BH}(A1)}$$

The methods for calculating the distance between bodies (DBB) and distance between faces (DBF) were similar. For example, DBF metric was used to present the mean distance between one student's face and the other two students' faces during the CPS activity. The student who stood in the middle always had a smaller distance to the other two students even if this student was not engaged with the CPS activity. Therefore, a method of triangle visualisation was used. As figure 2 shows, points A1, A2 and A3 were used to represent the students' faces. These three points make up a digital triangle and the closer the students stood to each other, the smaller area the triangle had. The height values of A1, A2, and A3 points represent how these three points affect the area of the triangle. Therefore, the height value of these points, HS1, HS2, and HS3, were used to define and calculate the DBF values.

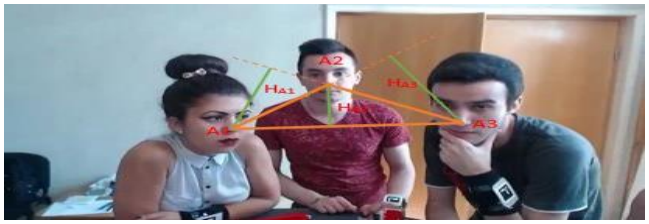

Figure 2. Distance between faces metric

This method contributed to differentiate distance between students' faces in a triangle relation. DBF(A1) was defined as:

$$\text{DBF(A1)} = \text{Average of } D_f\text{(A1,n)}$$

D_f(A1,n) was used to represent the distance between student A1's face and the other students' faces in the window *n*. If student A1's face was not found in the scene or there were no other faces in the scene, DBF value was not calculated. If only two students' faces were found in the scene (including student A1), D_f(A1,n) was calculated as the distance between A1 and midpoint of the other two students' faces.

$$D_f\text{(A1,n)}=D_f\text{(A2,n)}=\sqrt{(x_P(A1,n)-(\frac{x_P(A1,n)+x_P(A2,n)}{2}))^2+(y_P(A1,n)-(\frac{y_P(A1,n)+y_P(A2,n)}{2}))^2}$$

If there were three students' faces found in the scene, D_f(A1,n) was calculated as the height value of point A1.

$$D_f\text{(A1,n)}=\frac{|(x_P(A2,n)-x_P(A1,n))*(x_P(A3,n)-x_P(A1,n))-(y_P(A3,n)-y_P(A1,n))*(y_P(A2,n)-y_P(A1,n))|}{\sqrt{(x_P(A2,n)-x_P(A3,n))^2+(y_P(A2,n)-y_P(A3,n))^2}}$$

Similarly, the distance between group members' hands (DBH) was used to show the distance between students' hands.

$$\text{DBH(A1)} = \text{Average of } D_h\text{(A1,n)}$$

The first step in calculating D_h(A1,n) was defining students' hand position. If student's hands were both in the scene, the midpoint was used to represent the position of the students' hand. If only a student's right hand or left hand was found in the scene, the position of this hand was used to represent the hand position. If no hands were found in the scene for a student, this students' DBH value was not calculated. (x_{HA1},y_{HA1}) was used to represent student A1's hand position. Then, a similar method to calculating distance between faces and distance between bodies was used to define D_h(A1,n). If student A1's hand position was not found in the scene or there were no other hand positions defined in the scene, this value was not calculated. If there were two students' hand positions found in the scene (including student A1), D_h(A1,n) was calculated as the distance between A1 and midpoint of the two students' hand positions.

$$D_h\text{(A1,n)}=D_h\text{(A2,n)}=\sqrt{(x_{HA1}(n)-(\frac{x_{HA1}(n)+x_{HA2}(n)}{2}))^2+(y_{HA1}(n)-(\frac{y_{HA1}(n)+y_{HA2}(n)}{2}))^2}$$

If there were three students' hand positions found in the scene, D_h(A1,n) was calculated as the height value of point A1.

$$D_h(A1,n)=\frac{\left|(x_{HA2}(n)-x_{HA1}(n))*(x_{HA3}(n)-x_{HA1}(n))-(y_{HA3}(n)-y_{HA1}(n))*(y_{HA2}(n)-y_{HA1}(n))\right|}{\sqrt{(x_{HA2}(n)-x_{HA3}(n))^2+(y_{HA2}(n)-y_{HA3}(n))^2}}$$

3.3 Human coding of the video data for ground truth

3.3.1 Learner Behaviours in CPS

Learners present various behaviours while they are engaged with CPS activities. As has been shown in previous work [22], some of these relate to speaking such as vocalizing knowledge to establish shared understanding; some relate to listening such as adopting an improved version of a hypothesis suggested; some others to watching such as observing an agreed action being undertaken to achieve shared understanding; or making such as taking appropriate actions to solve the problem at hand. Therefore, we categorized the learner behaviours in four main categories of making (M), watching (W), speaking (S) and listening (L). In previous research, learners were marked with three different states according to the extent they were physically engaged in CPS [23]. However, such focus on physical activity emphasises on students' making and watching behaviours, while omitting communication behaviours of listening and speaking. Here, by involving codes on listening and speaking, we aim to cover learner behaviours in CPS more holistically. Making refers to the situations that learners interact with objects around them. Watching represents situations when learners are looking at the resources around them (including human resources). Speaking refers to situations when learners were talking to other learners or teachers. Listening refers to situations when a learner's head was facing towards another learner who is speaking.

This coding scheme was implemented for 15 videos. 30-second windows were used to code each learner's behaviours. The researchers observed the video for three seconds in every 30 seconds to judge learners' behaviours. Five digits were used to represent learner states. In each window, the codes (1), (2), (3), and (4) were used to represent making(M), watching(W), speaking(S), and listening(L) respectively. The code (0) was used to code windows when a learner was not in any of the states mentioned above, or was not found in the scene. The coding scheme was implemented by two human coders. Whenever there was disagreement, the coders observed the video for five seconds to revise their coding. In the end, there was %98 agreement with the ordinal k alpha value 0.912. M, W, S, and L values were then calculated as the relative frequency of these behaviours. For instance, M was defined as:

$$M = \frac{Number\ of\ code\ '1'}{Number\ of\ windows}$$

3.3.2 Collaborative Problem-solving Quality

Similar to the previous studies [11, 23], the group videos were watched by two researchers and manually coded using the frameworks introduced in [22, 23]. Based on these scores student groups were categorized as high, medium, and low competence CPS groups. In final scores, there was high agreement between researchers, with the ordinal k alpha = 0.876.

3.4 Decision trees

We use decision tree algorithm in RapidMiner v.9.3 to explore the relation between automated metrics and behaviours, as well as the relation between learner behaviours and CPS competence. Before implementing the operator, the numerical data was transferred into nominal data. For the values of video metrics and the values of behaviours, each value was compared with the mean score of the same learning activity (an interactive toy, a colour sorting machine, and an autonomous car). So, if the M value was higher than the mean making value of an interactive toy activity, it was labelled with 'high'. Otherwise, it is labelled with 'low'. For the CPS competence, groups were divided into three groups of high, medium, and low competence CPS groups. Given the scale of the sample, pruning and pre-pruning were used to avoid overfitting if the models. The confidence of pruning was 0.1. The minimal gain of pre-pruning was 0.01, the minimal leaf size was chosen as 2 and the maximum depth was set to 5. All learning behavior prediction models were evaluated with eight-fold cross-validation and CPS competence prediction model was evaluated with a three-fold cross validation due to smaller number of data points at the group level.

4. RESULTS

4.1 Using video metrics to predict learner behaviours of CPS

4.1.1 Making

The metrics of distance played a significant role in predicting the frequency of making behaviours. The DBB value was the root of the decision tree built which shows that when student bodies were close to each other, they were likely to have a high frequency of making behaviours.

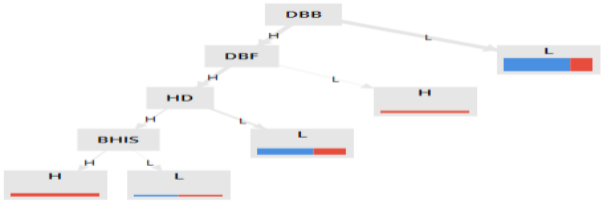

Figure 4. Decision tree for predicting the M value

The overall accuracy of this decision tree was at 63.75%±20.43%. Predicting both high making (66.67%) and low making (63.89%) behaviours in groups more accurately than the baseline measures.

Table 1. Confusion matrix for the M value

	True H	True L	Class Precision
Pred. H	6	3	66.67%
Pred. L	13	23	63.89%
Class Recall	31.58%	88.46%	

4.1.2 Watching

The decision tree which was generated to predict the W value is shown in figure 5. Both, the metrics of frequency and the metrics of distance, contributed significantly to the prediction of watching behaviours. The frequency of student's both hands being found in the screen was the most significant metric to predict their watching behaviours. When student's both hands were found in the screen frequently, their watching behaviours tended to be low.

Figure 5. Decision tree for the W value

The accuracy of the decision tree was at 64.58%±12.34%, and it performed better in predicting the low watching behavior of learners compared to high watching behaviours.

Table 2. Confusion matrix for the W value

	True H	True L	Class Precision
Pred. H	5	5	50.00%

Pred. L	11	24	68.57%
Class Recall	31.25%	82.76%	

4.1.3 Speaking

Distance between students' faces was the most significant metric to predict students' speaking behaviours. Moreover, other metrics of distance were related to the prediction of the S value. Whereas, only one metrics of frequency, LHIS value, was found significantly relevant to learners' speaking behaviours.

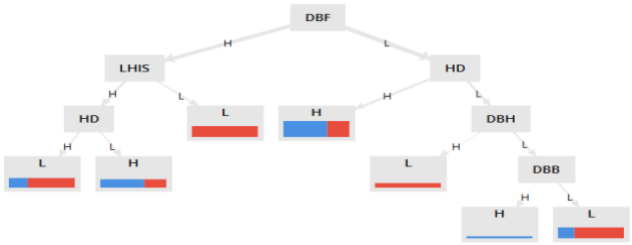

Figure 6. Decision tree for the S value

The overall accuracy of the model was at 63.75%±26.27% with the low speaking behaviour prediction (75.00%) being much higher than the high speaking behaviour prediction (52.38%).

Table 3. Confusion matrix for the S value

	True H	True L	Class Precision
Pred. H	11	10	52.38%
Pred. L	6	18	75.00%
Class Recall	64.71%	64.29%	

4.1.4 Listening

Although the root of the listening decision tree was RHIS metric, both the metrics of distance and the metrics of frequency were significant in predicting the students' listening behaviours in CPS.

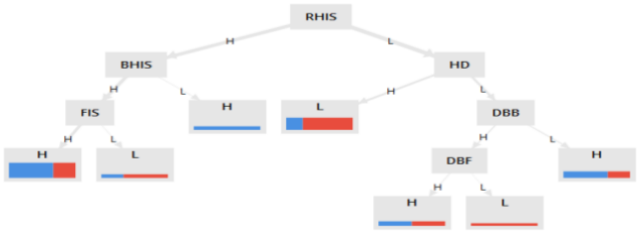

Figure 7. Decision tree for the L value

The overall accuracy of this decision tree was at 60.00%±14.36% with similar class precision for high and low listening behaviours.

Table 4. Confusion matrix for the L value

	True H	True L	Class Precision
Pred. H	17	12	58.62%
Pred. L	6	10	62.50%
Class Recall	73.91%	45.45%	

4.2 Predicting CPS from students' Making, Listening, Speaking, and Watching behaviours

In order to predict human judgements of CPS competence, we categorized learners' behaviour values in three groups. For each behaviour, groups' total values are ranked and every five group from top down is labelled as labelled as "Low", "Med", and "High", respectively. We chose categorizing behaviours cross-activity, rather than within activity, because the CPS competence judgements of humans were also made at the cross-activity level.

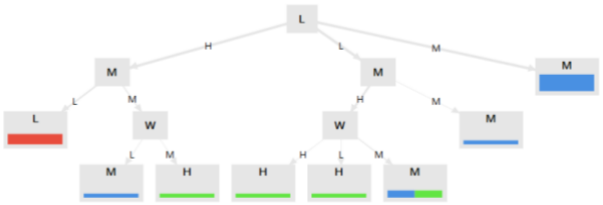

Figure 8. Decision tree for the CPS competence

As table 5 shows, the overall accuracy of this decision tree is 62.50%±44.32% which is better than the baseline and the previous work in the field [11]. The L collaboration class prediction precision was the highest (75.00%) which shows the approach's efficiency to predict low competence CPS groups from video data only. Whereas the prediction performance of high competence CPS groups was relatively low.

Table 5. Confusion matrix for CSP competence

	Ture High	True Medium	True Low	Class Precision
Pred. High	2	3	0	40.00%
Pred. Med	2	4	0	66.67%
Pred. Low	0	1	3	75.00%
Class Recall	50.00%	50.00%	100.00%	

5. Discussion and Conclusions

In this study, students' speaking, listening, making, and watching behaviours were interpreted via eight metrics generated from video data, which were than used to predict students' CPS competences. Our ultimate goal is to generate transparent and explainable models that predict learners' CPS competences from video data. These models can be used to support teachers' co-orchestration of collaborative learning activities in classrooms. We chose decision trees due to their easy to interpret nature for teachers to be able to interrogate the analytics suggestions of learners' CPS competences. We avoided the use of "black-box" modelling approaches, even though they are more frequently used in video analysis due to their high precision to generate insights from complex data [24]. As "black-box" approaches, we refer to the use of machine learning approaches (i.e deep neural networks) in learning analytics that are non-transparent and unexplainable. Not being able to explain a machine learning model doesn't matter in all cases. However, in educational contexts and learning analytics, where usually the main goal of a model is not only to get good predictions but also to understand which factors influence a learning outcome measure and to what extent, opaque approaches might have limited value [30]. However, as also argued in the literature [24] transparent modelling approaches such as decision trees can be more valuable as they allow teachers and learners to scrutinise analytics suggestions generated and allow opportunities for feedback and reflection [30]. This increased human agency in transparent models has the potential to lead to a better adoption in practice [25].

Our first research question was: What automated metrics from video data can be used to predict students' speaking, making, listening and watching behaviours during collaborative learning activities? The results show that, all-eight automated metrics we calculated can contribute to the prediction of student behaviours. However, the distance between students' faces and the distance between student's hands were the most significant metrics as they were closer to the roots of four behaviours' decision trees. These results are aligned with previous research that showed hands distance, and the students' body location can be used to predict

various learning outcomes [26, 27]. More specifically for CPS, the distance between students' hands, and the distance between students' faces were also found to be strong features for predicting CPS performance [2]. However, in contrast to previous research the number of faces looking at the screen did not perform well in predicting students' CPS performance [2]. The reason for this difference may be that the ground truths used in those studies were different. For instance, [2] evaluated outcomes of students' CPS activity, whereas we evaluated students' CPS competence regardless of their activity outcomes to categorise groups. While in education the goal of learning domain-specific knowledge is often accompanied with the goal of learning how to collaborate, both may require different guidance and support [28,31]. Our second research question was: To what extent can video data analytics accurately predict learners' CPS competence? Although the decision trees illustrate that the automated metrics from video data can be used to predict CPS competence better than baseline, the accuracy of the decision trees need to be improved. The accuracy of the transparent models we built are inferior to previous LA work in the field that involves multimodal data [2, 3, 20]. It might be the case that different modalities of data can be used to extract different types of metrics to build more accurate models [i.e audio data [24, 29]]. However, transparent nature of the models built here helps us generate insights of learners' CPS competence from video data. For instance, learners' listening, making, and watching behaviours were all strong predictors of their CPS competence. Considering the hands-on nature of the learning activities we studied, perhaps these results are not surprising. Yet, it is still interesting to observe that high competence CPS learners are frequently engaged in making and/or watching their peers' making behaviours during the learning activities. On the other hand, low competence CPS students were spending most of their time listening to others while not making or watching other's making behaviours. These insights generated from the models can serve to LA tools that provide suggestions for interventions to teachers and learners. Learners and teachers need to know their performance through analytics, but they also need to know the reasons behind the analytics' predictions on their performance. Our approach of building layered models from analytics metrics to learner behaviours (distance and frequency metrics → to predict S, M, L, and W) and from learner behaviours to CPS (S, M, L, and W → to predict high, medium, and low CPS competence groups) provides valuable contributions towards the design and use of non-disruptive LA technologies that can be used to improve real-world practice of CPS. It should be noted that since we used video data from classroom activities, there were multiple issues such as the cameras being too close/too far, not being able record all interactions between students, or being moved away by students during the activities. This all led to missing data and lower accuracy in our predictions. Therefore, a similar research with more reliable data sources can help improve the results presented here.

6. REFERENCES

[1] Cukurova, M., Bennett, J. and Abrahams, I. Students' knowledge acquisition and ability to apply knowledge into different science contexts in two different independent learning settings. *Research in Science & Technological Education*, 36, 1 (2018), 17-34.

[2] Spikol, D., Ruffaldi, E., Dabisias, G. and Cukurova, M. Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, 34, 4 (2018), 366.

[3] Martinez-Maldonado, R., Power, T., Hayes, C., Abdiprano, A., Vo, T., Axisa, C. and Buckingham Shum, S. *Analytics meet patient manikins: challenges in an authentic small-group healthcare simulation classroom*. ACM, 2017.

[4] Mao, Y., Li, H. and Yin, Z. *Who missed the class?—Unifying multi-face detection, tracking and recognition in videos*,ACM, 2014.

[5] Qin, J., Zhou, Y., Lu, H. and Ya, H. *Teaching video analytics based on student spatial and temporal behavior mining*. ACM, 2015.

[6] Raca, M., Kidzinski, L. and Dillenbourg, P. *Translating head motion into attention-towards processing of student's body-language*. City, 2015.

[7] Won, A. S., Bailenson, J. N. and Janssen, J. H. Automatic detection of nonverbal behavior predicts learning in dyadic interactions. *IEEE Transactions on Affective Computing*, 5, 2 (2014).

[8] Martinez-Maldonado, R., Echeverria, V., Santos, O. C., Santos, A. D. P. D. and Yacef, K. *Physical learning analytics: A multimodal perspective*. ACM, City, 2018.

[9] Chua, Y. H. V., Dauwels, J. and Tan, S. C. *Technologies for automated analysis of co-located, real-life, physical learning spaces: Where are we now?* ACM, City, 2019.

[10] Blikstein, P. *Multimodal learning analytics*. ACM, 2013.

[11] Grover, S., Bienkowski, M., Tamrakar, A., Siddiquie, B., Salter, D. and Divakaran, A. *Multimodal analytics to study collaborative problem solving in programming*. ACM,, 2016.

[12] Worsley, M. *(Dis) engagement matters: Identifying efficacious learning practices with multimodal learning analytics*. ACM, 2018.

[13] Lubold, N. and Pon-Barry, H. *Acoustic-prosodic entrainment and rapport in collaborative learning dialogues*. ACM, City, 2014.

[14] Dikker, S., Wan, L., Davidesco, I., Kaggen, L., Oostrik, M., McClintock, J., Rowland, J., Michalareas, G., Van Bavel, J. J. and Ding, M. Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27, 9 (2017), 1375-80.

[15] Echeverria, V., Martinez-Maldonado, R., Power, T., Hayes, C. and Shum, S. B. *Where is the nurse? towards automatically visualising meaningful team movement in healthcare education*. Springer, City, 2018.

[16] Oviatt, S. *Problem solving, domain expertise and learning: Ground-truth performance results for math data corpus.* ACM, 2013.

[17] Oviatt, S. and Cohen, A. *Written and multimodal representations as predictors of expertise and problem-solving success in mathematics*. ACM, City, 2013.

[18] Oviatt, S. and Cohen, A. *Written activity, representations and fluency as predictors of domain expertise in mathematics*. ACM,2014.

[19] Oviatt, S., Hang, K., Zhou, J. and Chen, F. *Spoken interruptions signal productive problem solving and domain expertise in mathematics*. ACM, City, 2015.

[20] Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P. and Pea, R. *Detecting collaborative dynamics using mobile eye-trackers.* Singapore: International Society of the Learning Sciences, City, 2016.

[21] Schneider, B. and Blikstein, P. *Unraveling Students' Interaction Around a Tangible Interface Using Gesture Recognition*. Citesee2014.

[22] Cukurova, M., Avramides, K., Spikol, D., Luckin, R. and Mavrikis, M. *An analysis framework for collaborative problem solving in practice-based learning activities: a mixed-method approach*. ACM, City, 2016.

[23] Cukurova, M., Luckin, R., Millán, E. and Mavrikis, M. The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education*, 116, (2018).

[24] Khajah, M., Lindsey, R. V. and Mozer, M. C. How deep is knowledge tracing? *arXiv, arXiv:1604.02416* (2016).

[25] Baker, R. S. Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26, 2 (2016).

[26] Blikstein, P. *Using learning analytics to assess students' behavior in programming tasks*. ACM, 2011.

[27] Ochoa, X., Chiluiza, K., Méndez, G., Luzardo, G., Guamán, B. and Castells, J. *Expertise estimation based on simple multimodal features*. ACM, City, 2013.

[28] Kirschner, P. A., Sweller, J., Kirschner, F. and Zambrano, J. From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning*, 13, 2 (2018),213.

[29] Owens, M. T., Seidel, S. B., Wong, M., Bejines, T. E., Lietz, S., Perez, J. R., Sit, S., Subedar, Z.-S., Acker, G. N. and Akana, S. F. Classroom sound can be used to classify teaching practices in college science courses. *Proceedings of the National Academy of Sciences*, 114, 12 (2017), 3085.

[30] Cukurova, M., Kent, C., and Luckin, R. Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *British Journal of Educational Technology*, 50*(6)*, (2019).

[31] Kent, C. and Cukurova, M. Investigating Collaboration as a Process with Theory-driven Learning Analytics. *Journal of Learning Analytics*, (2020).