British Journal of Educational Technology *Vol 50 No 6 2019* 3032–3046
doi:10.1111/bjet.12829

Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring

Mutlu Cukurova , Carmel Kent and Rosemary Luckin

Mutlu Cukurova is a lecturer in Digital Technologies in Education at UCL. Dr Cukurova investigates the potential of emerging technology such as Artificial Intelligence and Learning Analytics to continuously evaluate and support human development. Carmel Kent is a senior researcher at UCL. Dr Kent has 20 years of industry and academic experience, having worked for IBM research, EdTech and healthcare providers' companies and start-ups, as a software engineer, data scientist, entrepreneur, teacher and a researcher. Rosemary Luckin is a professor of Learner Centred Design at UCL and is the director of the EDUCATE programme. Dr Luckin has been developing and writing about the Learning Sciences, Educational technology and Artificial Intelligence in Education for over 20 years. Address for correspondence: Mutlu Cukurova, University College London, 23-29 Emerald Street, London, WC1N 3QS. Email: m.cukurova@ucl.ac.uk

Abstract
The question: "What is an appropriate role for AI?" is the subject of much discussion and interest. Arguments about whether AI should be a *human replacing* technology or a *human assisting* technology frequently take centre stage. Education is no exception when it comes to questions about the role that AI should play, and as with many other professional areas, the exact role of AI in education is not easy to predict. Here, we argue that one potential role for AI in education is to provide opportunities for human intelligence augmentation, with AI supporting us in decision-making processes, rather than replacing us through automation. To provide empirical evidence to support our argument, we present a case study in the context of debate tutoring, in which we use prediction and classification models to increase the transparency of the intuitive decision-making processes of expert tutors for advanced reflections and feedback. Furthermore, we compare the accuracy of unimodal and multimodal classification models of expert human tutors' decisions about the social and emotional aspects of tutoring while evaluating trainees. Our results show that multimodal data leads to more accurate classification models in the context we studied.

Introduction

There is little doubt about the great potential, and the disruptive nature of AI. However, in many professional areas, there are still debates about when and where AI technologies are appropriate for use, or indeed, whether they are appropriate at all (Floridi *et al*., 2018). Education is one of these areas, where AI technologies and their impact are currently not fully investigated. For instance, although there are AI technologies currently available and used to automate evaluation tools in education systems (Moser, 2015), the value of AI used only for automation in education systems is questioned (Baker, 2016). There is an emerging need for investigations into the potential use and impact of AI technologies in education. We need to explore the ability of these systems to cope with the complex social contexts in education, and to serve learners equally and equitably as appropriate. We also need to identify the potential unintended consequences of these systems. Clearly, whether we will ever attain the level of AI maturity that will enable

Practitioner Notes

What is already known about this topic

- There is little doubt about the great potential, and the disruptive nature, of AI in Education.
- There are key differences in the positioning and purpose of multimodal machine learning/AI and multimodal learning analytics.
- Many educational constructs are complex and ill-defined so human-decisions on them are often intuitive, rather than analytical.

What this paper adds

- A case study in which transparent classification models with multimodal data are created to support the decision-making process of educators.
- The results show that explicit and traceable models can support the complex and intuitive decisions of expert tutors in the context of tutoring.
- Multimodal data in models of expert human tutors' intuitive decisions might lead to more accurate classifications.

Implications for practice and/or policy

- One potential role of AI in Education is to support advanced reflections and feedback on human decision-making processes, rather than automating them.
- Transparent models can give educators the opportunity to reflect upon their complex decisions and provide learners with more advanced feedback, particularly on ill-defined constructs of education.
- Multimodal data collection and analysis is suggested in investigations of complex educational constructs.

fully automated systems to be part of everyday activities in education systems is an interesting research question in and of its self.

In this regard, there has been significant progress made in multimodal machine learning to potentially achieve automation of complex tasks in social contexts (Baltrušaitis, Ahuja, & Morency, 2019). More specifically in learning contexts, there are recent attempts that aim to interpret various modalities of data to automate the evaluation of complex learning processes (Di Mitri *et al.*, 2017), or to automatically predict learning performance from multimodal data (Giannakos, Sharma, Pappas, Kostakos, & Velloso, 2019). Such automation is particularly useful for the provision of adaptive support to learners. However, AI systems in education can also be used to support human decision-making processes, rather than automating them. AI systems in Education should be considered a continuum with regard to the extent they are decoupled from humans, rather than only an approach to provide full automation (Cukurova, 2019). As argued by Malone (2018), AI systems have the potential to augment human intelligence in machine-human collaborations as "superminds." The main question we are interested in this study is, *what might the augmentation of human intelligence to create "super educator minds" look like in the context of decision-making in education?*

Multimodal machine learning and multimodal learning analytics

Modality can be defined as the type of communication channel used by two agents to convey and acquire information that defines the data exchange (Kress, 2009). In this paper, we use two modalities. First one is the tabular information collected through psychometrics and surveys,

the other one is the audio data. Multimodality has been studied for around three decades in the context of social semiotics and its potential to help us understand the world around us led AI researchers to try and build models that can process information from multiple modalities through machine learning and social signal processing (Vinciarelli, Pantic, & Bourlard, 2009). The literature on multimodal prediction is rich with examples of audio-visual speech recognition (Zhou & De la Torre, 2012); multimedia content indexing and retrieval (Atrey, Hossain, El Saddik, & Kankanhalli, 2010), and multimodal affect recognition (D'mello & Kory, 2015; Grawemeyer *et al.*, 2017). Learning from multimodal data provides opportunities to gain an in-depth understanding of complex processes and, for AI research to make progress, it should focus on multimodal AI models that can process and relate information from multiple modalities (Baltrušaitis, Ahuja, & Morency, 2019). In multimodal machine learning research, the ultimate aim is to automate the decision-making process, rather than making it transparent to humans for reflection, feedback and learning opportunities.

On the contrary, in educational contexts, multimodal learning analytics (Blikstein, 2013) approaches are emerging and providing promising opportunities. For instance, Ochoa *et al.* (2018) recently proposed a multimodal feedback approach for learner reflections based on posture, gaze, volume and performance data. Within the context of collaborative learning, Martinez-Maldonado, Dimitriadis, Martinez-Monés, Kay, and Yacef (2013) collected data from verbal and physical interactions of students to provide insights into their collaborative actions around tabletop computers. Similarly, Di Mitri *et al.* (2017) investigated the potential of multimodal data from Fitbit wristband and computer logs to predict learners' self-regulation performance. In the context of project-based learning, Cukurova, Luckin, Millán, and Mavrikis (2018), and Spikol, Ruffaldi, Dabisias, and Cukurova (2018) collected data from learners' hand movements and head direction to predict their success in open-ended design tasks. In the area of professional development, Echeverria, Martinez-Maldonado, Power, Hayes, and Shum (2018) used sensor data to capture trainee nurses' interactions during healthcare training and created visualisations of their interaction and movements for effective reflections. From the educators' point of view, Prieto, Sharma, Kidzinski, Rodríguez-Triana, and Dillenbourg (2018) recently collected eye-tracking, audio-visual and accelerometer data of educators to help them support the management of classroom activities. Although certain tasks and activities might be automated to support learners and educators, in essence, most multimodal learning analytics approaches aim to provide explicit and comprehensible ways of presenting information to learners and teachers to make them more informed in their decisions. This positioning differs significantly from the multimodal machine learning approaches that aim to automate the decision-making process itself. Both multimodal machine learning and learning analytics approaches are valuable for the advancement of the educational research and practice, and they are not commensurable due to different epistemological values driving them (Cukurova, 2018, 2019).

The contribution of this paper is twofold. First, we present the results of our investigation into the decision-making process of expert tutors while they evaluate trainees. We illustrate that multimodal classification models can support these tutors by making some aspects of their decision-making processes more explicit and traceable, therefore, potentially, subjecting them to a more analytical decision-making process. We demonstrate our approach to utilizing a combination of non-transparent predictive models and transparent classification models in order to exemplify a hybrid AI system in the service of complex human decision-making in education. Second, we compare the accuracy of unimodal and multimodal models of expert human tutors' intuitive decision-making processes on the social and emotional aspects of tutoring while they evaluate tutor candidates. We show that multimodality in data leads to more accurate predictions in these conditions.

Decision-making process for humans and transparent systems to potentially augment it

Different mechanisms are suggested for the process of decision-making (Evans, 2008), there is however, a consensus about the process nature of it. Broadly, there are two categories of decision-making processes argued for in the literature. One is mainly associated with heuristic processes that are argued to be operating autonomously and automatically (Kahneman & Frederick, 2002). These are processes that function without conscious control and cannot easily be accessed for inspection. Furthermore, they can process multiple pieces of information in parallel (Betsch & Glöckner, 2010). The other one is mainly associated with analytic processes, that are, in contrast, performed step-by-step. The sequence and direction of these processes can be deliberatively controlled, and the individual is consciously aware of performing these processes and thus able to reflect on them (Betsch, 2008). Intuitive decision-making is faster and less effortful, but as a result it is subjected to decision biases, which makes the outcomes suboptimal or simply wrong (Evans, 2003; Kahneman, 2003). In contrast, the analytical decision-making is more slow, deliberative, reflective, has more normative outcomes (Evans, 2003) and is able to guard us against biases in intuitive decision-making (Kahneman & Frederick, 2002).

Based on these descriptions of two categories of decision-making processes, the decision-making process of the so-called "good old-fashioned" AI technologies that require explicit representation of knowledge and well-defined goals, and that use transparent modelling approaches (ie, Bayesian Knowledge Tracing (Corbett & Anderson, 1994) or more traditional statistical modelling approaches such as logistic regression), may be considered as analytical processes. Compared to more modern approaches, these approaches are limited in terms of what types of knowledge and distributions they can model (Khajah, Lindsey, & Mozer, 2016). However, arguably, their most significant advantage is that they are understandable by humans (Russell & Norvig, 1995). These are transparent approaches that are often implemented step-by-step. When they are investigated, it is relatively easy to trace the origin of each output and to understand what is the weight given to each input and whether it has been contributing to the decision-making process of the model. As these models require a good level of understanding of the process that is being modelled, they also have the potential to provide opportunities for humans to better understand the processes themselves, therefore can be utilized to support the decision-making processes of educators.

In education and training settings, some constructs, particularly those that relate to so-called 21st-century skills, and/or social and emotional learning, are neither well-defined nor are they widely understood. Concepts such as creativity, empathy, self-awareness, social awareness or ethical responsibility can all be considered as ill-defined constructs within education. Even some fairly well-defined, studied and understood constructs such as collaborative problem solving (CPS), when considered as generic skills, are extremely complex skill sets (Scoular, Care, & Hesse, 2017). They should be considered as "a bundle of skills, knowledge and abilities that are required to deal effectively with complex and dynamic non-routine situations in different domains" (Funke, Fischer, & Holt, 2018, p. 42). Therefore, it is safe to assume that educators sometimes find themselves in situations where they make intuitive decisions while evaluating complex constructs. Based on this assumption, *our first hypothesis is that the analytic nature of transparent models can support educators in their intuitive decision-making processes about some of the ill-defined constructs of learning*.

Additionally, we also know from the existing literature that, that multiple modalities of information, improve the decision-making process by influencing both the speed and the accuracy of decisions in humans (Ratcliff, Smith, Brown, & McKoon, 2016) and animals (Kulahci, Dornhaus, & Papaj, 2008). Therefore, *our second hypothesis is that multimodal data in decision-making processes*

might increase the accuracy of classification decisions. These two hypotheses generated from the literature on decision-making processes are the core of our two research questions.

1. How can analytic classification models support the intuitive decision-making processes of expert tutors while they are evaluating tutor trainees?
2. What are the relative accuracies of unimodal and multimodal models in classifying tutor trainees?

The context of the study and the intuitive decision-making process investigated

The decision-making process we studied is in the context of evaluating the performance of people who have applied to become a tutor for an education organisation that offers tutoring in debating skills to school students. The evaluation is based on the performance of candidates whose academic achievements are reflected in their CV. Initially, we elicited knowledge from the expert debate tutors who currently conduct the evaluation of applicants using observations and interviews over a three-month period. We identified that there were predominantly two essential domains for the evaluation of tutors: candidates' social and emotional skills, and their tutoring skills. Frequently emerging themes within the domain of *social and emotional* skills for the "expected candidates," were social interactivity, engagement, emotional intelligence and appropriate encouragement/praise of others. Skills such as the use of contingent tutoring techniques, appropriate pitching and management of the students were considered as essential for *tutoring* skills. In addition, the expert tutors identified the "style" (eg, persuading and engaging), professionalism (use of appropriate vocabulary and manner of debating), content (specific debating techniques) and strategy (eg, relevance and clarity of the argumentation) as essential elements of debating.

During the evaluation process, three different expert tutors independently used these criteria to give each candidate a score ranging from 1 to 5; where 1 and 2 designate an excellent debate tutor, who can generally be placed at any school, 3 is used for those who are acceptable, but might need some further training, and 4 and 5 are not desirable debate tutors who need in-depth training. Following the independent scoring, in cases of a discrepancy among experts, the evaluators negotiated the scores and reached a consensus. Some of the criteria mentioned by the expert tutors were evaluated through an analytical decision-making process with rubrics such as the content and strategy of candidates' debates. However, the discussions conducted by expert tutors revealed that many other skills were judged through an intuitive decision-making process, where expert tutors relied upon their experience and observational clues. When we questioned expert tutors on what they think makes a good tutor with respect to these social and emotional aspects, they all struggled to define these constructs or break what they were looking for into smaller observable constructs. As one expert tutor put it, "when you see one, you know they are a great debate tutor." In order to get more insights into such intuitive decision-making processes, we also analysed the evaluation notes of expert tutors about the candidates. As can be seen below, the concepts used to justify scores given are neither well-defined, nor well-understood.

Notes taken by expert tutors to justify their scores on social and emotional skills of the new candidates observed.	Energetic, engaging, Good sense of humour, Good audience interaction, Not inspiring, Very confident, academic, Friendly, nice and fun, Teaching oriented, Cool vibes, Needs more interactivity, Kids will like her, Seems socially awkward, Controlled class well, Very interactive, Good commitment and effort, A bit shy, High energy, Authoritative, Sweet

It is very hard for humans to untangle this knot of knowledge, skills, emotions and perceptions on the ill-defined constructs of education contexts due to their intuitive nature. On the contrary, in transparent systems, the decision-making processes are analytical types by definition. In these models of decision-making, it is possible to trace the modality origin of each trace and cue, whatever the source of information input and to understand what is the weight given to this modality in the model, and to what extent it has been contributing to the decision-making processes.

Methods and sample

We initially undertook a review of the learning sciences literature on debate tutoring skills, with the purpose of identifying potentially relevant data sources for a classifying model of the intuitive decision-making processes of expert tutors, Based on the literature review, we identified relevant psychometric measures of effective debate tutoring including the following personal characteristics: temperament, emotional intelligence, charisma and several personality traits (extraversion/introversion, neuroticism/emotional stability, openness, conscientiousness). In particular, temperament was represented by two dimensions—social closeness and social anger, that are assessed by the Adult Temperament Questionnaire (ATQ; Evans & Rothbart, 2007). We used empathy as a representative of the emotional intelligence, the measurement of which was adapted from the Trait Emotional Intelligence Questionnaire (TEIQue- SF; Petrides, 2009). The charismatic abilities are examined by the General Charisma Inventory (GCI; Tskhay, Zhu, Zou, & Rule, 2018). Finally, the reflection of the tutors on their abilities to follow plans and commitments were assessed by items utilized from the Big Five Inventory (BFI; John & Srivastava, 1999). These personal characteristics aim to reflect the candidates' social and emotional capabilities, specifically, testing their ability to develop social connections, be orientated to communicate and readiness to be exposed to a large volume of social interaction, along with abilities to empathy and emotion control that are all crucial in the context of debate tutoring. Furthermore, we added two items about candidates' previous experience in debating and tutoring. The questionnaire was filled in by 127 candidates.

In order to collect data on the emotional traits of tutor candidates, we collected audio recordings of 47 candidates. To analyse the audio data, we used OpenSMILE's prediction models. OpenSMILE is an open-source software for automatic extraction of features from audio signals and for classification of speech and music signals. It extracts audio features such as shimmers (auditory spectra), loudness, formant frequencies and bandwidths. It then uses audio analysis algorithms to infer on the speaker's interest level, emotional and affective states and traits. For example, anger is typically characterized by fast speech rate/tempo, high voice intensity/sound level, much voice intensity/sound level variability, much high-frequency energy, high pitch level and variability, rising pitch contour, fast voice onsets/tone attacks and microstructural irregularity (Juslin & Laukka, 2003). OpenSMILE's models are trained on voice recordings and annotated for emotions, affections and more by humans. Specifically, in this paper, we have used 16 voice-based emotion features. These include seven classes of basic emotions (Anger, Fear, Happiness, Disgust, Boredom, Sadness, Neutral) according to the Berlin Speech Emotion Database (EMO-DB) (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005), the Airplane Behavior Corpus (ABC) (Schuller, Arsic, Rigoll, Wimmer, & Radig, 2007) with six affections (Aggressive, Cheerful, Intoxicated, Nervous, Neutral, Tired) and the Audio Visual Interest Corpus (AVIC) with three levels of interest (passive, neutral and strong interest levels) (Schuller, Steidl, & Batliner, 2009). To clean the audio data, we omitted windows that are smaller than 1600 ms, computed *SD* for each variable/candidate, omitted samples outliers, and computed mean values without outliers.

Data preparation

In order to reduce the large set of variables into a smaller set of components, that account for most of the variance in the tailored questionnaire variables, a principal components analysis (PCA) was run on the data from the 22-question questionnaire. Inspection of the correlation matrix showed that two variables: social closeness and rare social anger, had no correlation coefficient greater than 0.3, thus they were removed from the PCA. The 20 variables remaining had at least one correlation coefficient greater than 0.3 (KMO = .791, Barlett's sphericity was significant $p < .0005$). PCA revealed five components that had eigenvalues greater than one and which explained 29.728%, 10.847%, 9.159%, 7.458% and 5.683% of the total variance, respectively. Visual inspection of the scree plot indicated that four components should be retained (Cattell, 1966). In addition, a four-component solution met the interpretability criterion. So, four components were retained. The four-component solution explained 57.193% of the total variance. A Varimax orthogonal rotation was employed to aid interpretability. The rotated solution exhibited "simple structure" (Thurstone, 1947) with strong loadings of

1. Extraversion, outgoingness, and leadership items on component 1, which we will call the extrovert leader factor.
2. Charisma, enthusiasm and the tendency to make people comfortable items on component 2, the charismatic factor.
3. Assertiveness, organization and the tendency of being influential items on component 3, the assertive organized factor, and
4. Neuroticism, non-assertiveness items on component 4, the neurotic factor.

Figure 1 shows a radar chart of the median value of each of the four factors, for the three score categories.

There was insufficient data to implement a similar PCA approach for the audio data. However, we omitted highly correlated variables from the data. Our data sets were generally not normally distributed, therefore, we have used a Spearman ranking correlation tests, based on which variables were omitted:

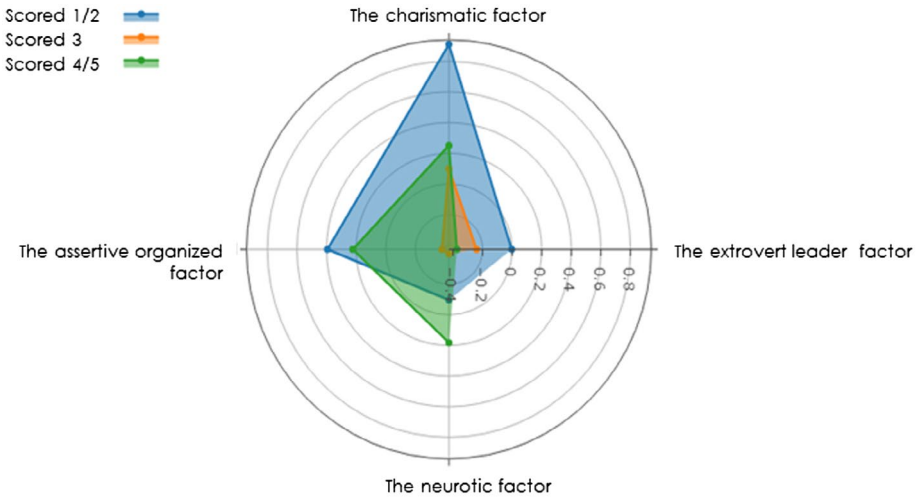

Figure 1: Significant components accounting for most of the variance of the designed questionnaire [Colour figure can be viewed at wileyonlinelibrary.com]

neutral interest, since it is negatively correlated with strong interest $r = -0.958$ $p < .005$, and happiness since it is highly correlated with anger 0.892 $p < .005$. In the literature, there is evidence that affective dimensions are interrelated in a highly systematic fashion (Russell, 1980) and, here, it is not possible to assert from openSMILE's predictive models whether the score a candidate is given is because they are "angry" or "happy." Therefore, we used the common denominator of two emotions, and labelled this merged factor as "arousal" (Russell, 1980). In order to be able to make comparisons on the accuracy of unimodal and multimodal models and answer our second research question, we used only the set of 41 candidates who had a clear human decision-making output (their final score), audio sample, and a full survey.

Results

Initially, to ensure that the different variables originating from different modalities of data investigated in this study actually do explain different constructs, we created a correlation matrix of all the variables investigated. Figure 2 shows that there are many and significant in-modal correlations, however, very few and non-significant inter-modal correlations. This information suggests that different modalities of data used in this research bring in different input to the classification model built.

Predicting the intuitive decision outputs of expert tutors from various data inputs

We used multinomial logistic regression to classify the scores of the candidates. Table 1 shows the results of the model built with just the two experience variables. The model's goodness of fit shows that the model fits the data well $\chi^2(\text{df} = 14) = 10.73$, $p = .707$. However, the model's fitting

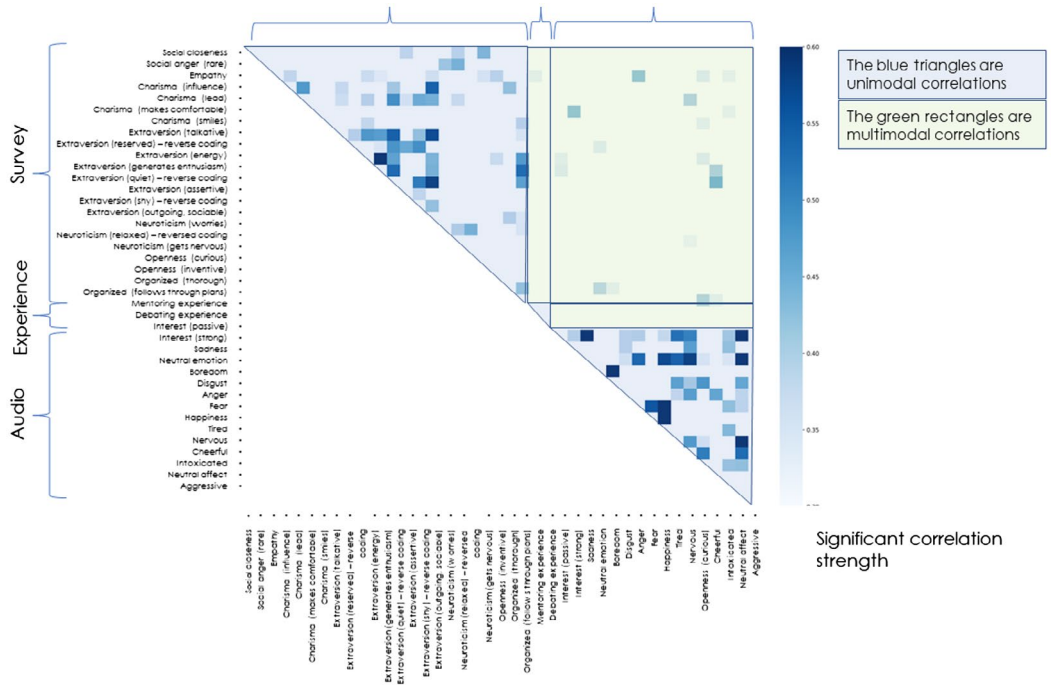

Figure 2: Correlation matrix of the variables from the survey, audio data and the previous experience on debating and tutoring [Colour figure can be viewed at wileyonlinelibrary.com]

information shows that the full model does not significantly predict the score, not better than the intercept-only model.

Table 2 shows the results of just audio variables. The model's goodness of fit shows that the model fits the data well $\chi^2(\text{df} = 56) = 61.63$, $p = .282$. Moreover, the model's fitting information shows that the full model significantly predicts the score, better than the intercept-only model alone $\chi^2(\text{df} = 24) = 41.72$, $p = .014$. In the likelihood ratio tests, Interest – passive $\chi^2(\text{df} = 2) = 15.33$, $p = .001$, Emotion – arousal $\chi^2(\text{df} = 2) = 9.06$, $p = .011$, Affect – nervous $\chi^2(\text{df} = 2) = 20.19$, $p = .001$, Affect – aggressive $\chi^2(\text{df} = 2) = 8.40$, $p = .015$ were found to be significant in the classification model.

Table 3 shows the results of the model with survey + experience. The model's goodness of fit shows that the model fits the data well $\chi^2(\text{df} = 56) = 61.13$, $p = .297$. However, the model's fitting information shows that the full model does not significantly predict the score.

As can be seen in Table 4, when classifying the candidates data including the survey variables, the two experience variables and two of the significant audio variables, affect nervous, and emotion arousal variables, the model's goodness of fit shows that the model fits the data well $\chi^2(\text{df} = 52) = 48.56$, $p = .610$. Moreover, the model's fitting information shows that the full model significantly predicts the score, better than the intercept-only model alone $\chi^2(\text{df} = 28) = 47.05$, $p = .014$. Considering our small sample size, we could not add more audio variables to the model, since it would cause it to overfit. In the likelihood ratio tests, the extrovert leader factor $\chi^2(\text{df} = 2) = 11.08$, $p = .004$, the assertive organized factor $\chi^2(\text{df} = 2) = 13.03$, $p = .001$, the neurotic factor $\chi^2(\text{df} = 2) = 8.2$, $p = .017$, the social closeness survey item $\chi^2(\text{df} = 2) = 11.50$, $p = .003$, the social anger rare survey item $\chi^2(\text{df} = 2) = 7.35$, $p = .025$, tutoring experience $\chi^2(\text{df} = 6) = 14.38$, $p = .026$, debating experience $\chi^2(\text{df} = 6) = 21.92$, $p = .001$, the nervous affect $\chi^2(\text{df} = 2) = 19.92$, $p = .001$ and the emotion arousal $\chi^2(\text{df} = 2) = 19.92$, $p = .001$ variables were found to be significant in the classification model. As can be seen, when two audio variables (affect nervous and emotion arousal) were added to the model, the value of other variables has become significant in the classification outputs. Specifically, it was found that candidates who were given 3 by expert tutors are more likely to show higher level of arousal relatively to those who scored 1 and 2 ($p = .042$, $B = 78.34$) and are less likely to show signs of being nervous ($p = .014$, $B = -426.697$), where these differences were not significant between those who scored 4 and 5 and those who scored 1 and 2.

Discussion

The results reported here illustrate that transparent classification models of expert tutors' intuitive decision-making about trainee tutors' social and emotional skills can potentially be built with the help of relevant multimodal data. Transparent decision-making models can be utilized

Table 1: Classification based on two experience variables

	Estimated score			
Observed score	*1/2*	*3*	*4/5*	*Per cent correct*
1/2	0	9	0	0.0
3	0	22	0	100.0
4/5	0	10	0	0.0
Overall	0.0%	100.0%	0.0%	53.7

Table 2: Classification based on audio variables

	Estimated score			
Observed score	*1/2*	*3*	*4/5*	*Per cent correct*
1/2	8	1	0	88.9
3	2	15	5	68.2
4/5	0	3	7	70.0
Overall	24.4%	46.3%	29.3%	73.2

Table 3: Classification based on two experience variables and the survey variables

	Estimated score			
Observed score	*1/2*	*3*	*4/5*	*Per cent correct*
1/2	4	5	0	44.4
3	3	17	2	77.3
4/5	1	8	1	10.0
Overall	19.5%	73.2%	7.3%	53.7

Table 4: Multimodal classification, based on the experience, survey and two of the stronger predicting variables

	Estimated score			
Observed score	*1/2*	*3*	*4/5*	*Per cent correct*
1/2	8	1	0	88.9
3	1	19	2	86.4
4/5	0	4	6	60.0
Overall	22.0%	58.5%	19.5%	80.5

by expert tutors to reflect upon their own decision-making processes enabling them to provide better feedback to trainees' about their evaluations to trainees. In order to exemplify the value of transparent models to support the expert tutors' decision-making process, we will now detail a specific candidate's evaluation. The trainee Jane (pseudonym) was initially scored 3 by the expert tutors, which is a score that represents a trainee who is acceptable, but might need further training to become a tutor. First, she was predicted by the model presented in Table 3 (ie, a model not taking into account the audio analysis) as a 4/5 tutor, which means, that according only to her previous experience and her personality data, without taking into account the voice modality, she would have been found not to be suitable to serve as a tutor. However, using the multimodal model at Table 4, which takes the voice modality into account, she was accurately predicted to be a score 3 level trainee.

In the case of Jane, the qualitative notes taken by the expert tutors to justify their decision about the social and emotional aspects of tutoring were very intuitive and were not detailed enough to provide the candidate with an appropriate feedback. They were neither explicit enough for expert tutors to reflect upon their own decisions. For example, one tutor evaluator noted about

the candidate that "She has good vibes and kids will like her in the classroom." In contrast, the multimodal model in Table 4 was able to more transparently open up this intuitive decision of "good vibe," and propose possible explanations about what it is exactly about Jane's collected data that contributes to Jane's score. The explicability embedded in the analytical nature of the classification models built, provides educators with the opportunity to track and monitor each factor's contribution to the decisions made. For example, Jane's level of arousal as spotted by the audio analysis can be utilized as a feedback tool for Jane to reflect upon (Figure 3). Similar to the arousal emotion, other factors that were found to be significant in the likelihood ratio tests of the multimodal data model (see the last paragraph of the results section). All of these features can be utilized for effective feedback and reflection opportunities. Additional information through the analysis of complex audio data with openSMILE's prediction models and their combination with the transparency of the multimodal regression model have the potential to make expert tutors more informed in their decisions.

Therefore, returning to our first research question, it is possible to argue that classification models can potentially support the intuitive decision-making processes of expert tutors while evaluating tutors, through making the processes more transparent yielding better educator reflection and advanced feedback for learners. Here, the value of the multimodal classification models built is not necessarily in their ability to accurately classify trainee tutors. In this sense, the area of multimodal machine learning is making significant progress in building better models that can automatically predict and classify human behaviours (Baltrušaitis, Ahuja, & Morency, 2019). However, with regard to supporting the human decision-making processes in complex educational contexts, there might also be value in approaches that can combine the power of modern AI approaches with more limited yet transparent approaches. This is not to criticise modern "black-box" modelling approaches such as deep learning or machine learning as they provide excellent opportunities to generate insights from complex data (Khajah, Lindsey, & Mozer, 2016). Yet, our intention is to argue that there is value in the careful match of the modelling approaches with the objective of the model. In this study, we used non-transparent prediction models of openSMILE to predict the emotional traits of tutor candidates based on their audio data, and

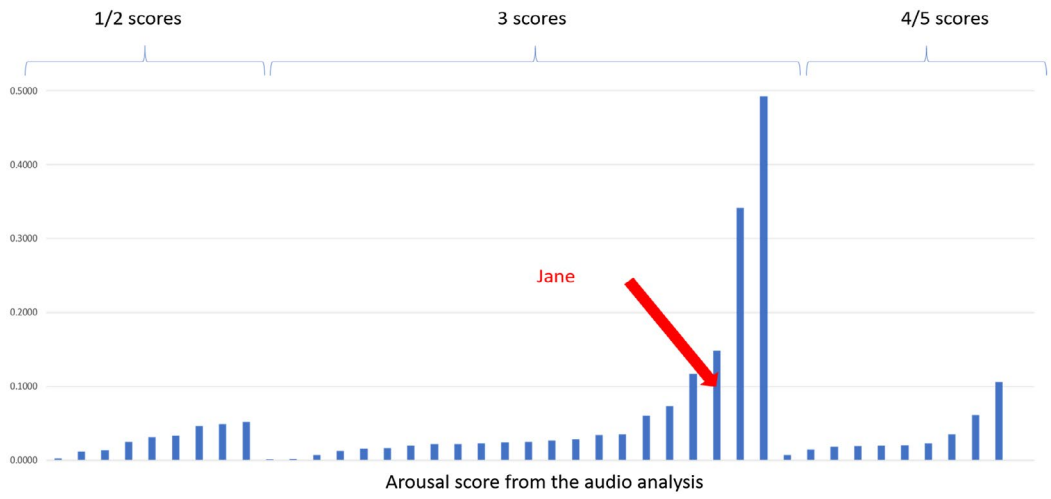

Figure 3: Levels of the component operationalising arousal in OpenSMILE's audio analysis [Colour figure can be viewed at wileyonlinelibrary.com]

we used transparent logistic regression models to identify exactly what personality, emotion and experience traits lead to effective debate tutoring skills. Predictive models were very powerful to make sense of complex and non-linear audio data, whereas the transparent regression models were valuable to identify key aspects for tutors to reflect upon their own decisions and provide tutor candidates with feedback on their performance.

Our second research question was to investigate the relative accuracy of unimodal and multimodal models to classify trainee tutors. The results show that the accuracy of the models built increases with multimodal data. There are a number of possible mechanisms to explain the increased accuracy of the classification models. As suggested in our correlation matrix (Figure 2), it might be the case that different data modalities are effective at providing information on different dimensions of the multidimensional constructs evaluated by expert tutors, so multiple modalities would lead to an increase in overall performance. Based on these results, we argue that the richness and complexity of human decision-making can potentially be better interpreted, evidenced, and supported with rich, multimodal data. Similar results are also echoed in multimodal machine learning research in educational contexts (Giannakos *et al.*, 2019). Furthermore, relatively easy access to large volumes of data produced through the use of digital technologies that are large in volume, even for single learners, provide us many opportunities to collect multimodal data to support human decision-making processes. We argue that such large-volume, heterogeneous and complex data, when combined with different modelling approaches, can potentially help us to shed light on and support complex human decisions in a holistic manner.

Limitations and future studies

The research reported here has been conducted in a particular context with a relatively small sample size of 127 candidates, from which 41 were subjected to the multimodal classification. Multimodal data would naturally result in a large number of features for use in modelling, which ideally require a respectively large number of observations to avoid overfitting. In particular, we had limited data from trainees who scored 4 or 5, which might have affected the accuracy of our classification models. Therefore, any conclusions we draw about effective debate tutor features should be approached with caution. Moreover, the prediction ability of openSMILE models used isn't perfect. As detailed in the data preparation section, "happiness" feature is highly correlated to "anger"; potentially due to shared high arousal and opposite valence (Russell, 1980). openSMILE models employed here were not able to detect such differences. Thus, the outputs created from the predictive models of openSMILE should also be interpreted with caution. Furthermore, while our results show the value of multimodal models in classifying the outputs of intuitive decision-making processes in the context of debate tutoring, carefully designed future studies are needed to distinguish the value of different modality's value for different decision contexts. We chose debate tutoring context as a case study, as it frequently required tutors to take intuitive decisions and it lacked clear standards for evaluating "success." A similar approach can be used in other relevant educational settings. In the future, we also aim to study the value of such multimodal approaches to support the decision-making processes of educators with qualitative investigations to get better insights into their potential value to educators and learners. More specifically, we are interested in studying the value of different decision support tools in educational settings. We aim to set up a series of studies looking at the adoption of such decision support approaches by tutors. Future studies should focus on what decisions do educators expect to be supported with learning analytics and AI in different educational contexts? To what extent they are expecting the decisions on these tasks to be automated by modelling approaches? as well as investigations on how data from such systems should be visualized and presented back to educators and students for advanced reflections and feedback?

Conclusion

In this paper, we present the results of a case study in which we have combined predictive and transparent models to support the human decision-making processes involved in tutor trainee evaluations. Our results showed that models with multimodal data can accurately classify tutors and have the potential to support the intuitive decision-making of expert tutors in the context of evaluating trainee applicants. Through the analysis of multimodal data, collected *in situ* to investigate the social and emotional aspects of debate tutoring, our aim was to exemplify a potential approach to utilising AI in the service of complex human decision-making. This "subservient" role of AI that feeds into more transparent analytics interpreted by humans might be more valuable for educational contexts than the current discourse on AI as a human-replacing technology.

Acknowledgements

We would like to thank DebateMate (www.debatemate.com), Scarlett McCabe, Margaret McCabe, Ekaterina Cooper, and Allen Atwell for their help with the literature review, data collection, and data analysis.

Statements on open data, ethics and conflict of interest

Data can be accessed for research purposes upon request.

The ethical approval of the project was granted by the UCL Institute of Education Ethics Committee (REC 1056).

There are no conflicts of interest in the communications of this research.

References

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, *16*(6), 345–379.

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443.

Betsch, T. (2008). The nature of intuition and its neglect in research on judgment and decision making. In H. Plessner, C. Betsch, & T. Betsch (Eds.), *Intuition in judgment and decision making* (pp. 3–22). New York, NY: Erlbaum.

Betsch, T., & Glöckner, A. (2010). Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry*, *21*(4), 279–294.

Blikstein, P. (2013). Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 102–106). New York, NY: ACM.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Ninth European Conference on Speech Communication and Technology* (pp. 1517–1520). Lisbon, Portugal.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Cukurova, M. (2018). A syllogism for designing collaborative learning technologies in the age of AI and multimodal data. In *European Conference on Technology Enhanced Learning* (pp. 291–296). Cham: Springer.

Cukurova, M. (2019). Learning analytics as AI extenders in education: Multimodal machine learning versus multimodal learning analytics. In *Proceedings of the Artificial Intelligence and Adaptive Education Conference* (pp. 1–3).

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

Di Mitri, D., Scheffel, M., Drachsler, H., Börner, D., Ternier, S., & Specht, M. (2017). Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, (pp. 188–197).

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*(3), 43.

Echeverria, V., Martinez-Maldonado, R., Power, T., Hayes, C., & Shum, S. B. (2018). Where is the nurse? Towards automatically visualising meaningful team movement in healthcare education. *International Conference on Artificial Intelligence in Education*, 74–78.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends Cognition Sciences*, *7*, 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Evans, D. E., & Rothbart, M. K. (2007). Developing a model for adult temperament. *Journal of Research in Personality*, *41*(4), 868–888.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*, 1–34. https://doi.org/10.1093/hmg/ddy137/4972370

Funke, J., Fischer, A., & Holt, D. V. (2018). Competencies for complexity: Problem solving in the twenty-first century. In E. Care, P. Griffin, & M. Wilson (Eds), *Assessment and teaching of 21st century skills. Educational Assessment in an Information Age* (pp. 41–53). Cham: Springer.

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, *48*, 108–119.

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., & Rummel, N. (2017). Affective learning: Improving engagement and enhancing learning with affect-aware feedback. *User Modeling and User-Adapted Interaction*, *27*(1), 119–158.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research* (Vol. *2*, pp. 102–138).

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, *49*, 81.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). *How deep is knowledge tracing?* arXiv preprint arXiv:1604.02416.

Kress, G. (2009). *Multimodality: A social semiotic approach to contemporary communication*. Abingdon: Routledge.

Kulahci, I. G., Dornhaus, A., & Papaj, D. R. (2008). Multimodal signals enhance decision making in foraging bumble-bees. *Proceedings of the Royal Society of London B: Biological Sciences*, *275*(1636), 797–802.

Malone, T. W. (2018). How human-computer 'superminds' are redefining the future of work. *MIT Sloan Management Review*, *59*(4), 34–41.

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

Moser, L. (2015). A controversial teacher-evaluation method is heading to court. Here's why that's a huge deal. *Slate*. Retrieved from http://www.slate.com/blogs/schooled/2015/08/11/vam_lawsuit_in_new_york_state_here_s_why_the_entire_education_reform_movement.html

Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). The rap system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 360–364) . Sydney, Australia.

Petrides, K. V. (2009). Psychometric properties of the Trait Emotional Intelligence Questionnaire (TEIQue). In J. Parker, D. Saklofske, C. Stough (Eds), *Assessing emotional intelligence. The Springer Series on Human Exceptionality* (pp. 85–101). Boston, MA: Springer.

Prieto, L. P., Sharma, K., Kidzinski, Ł., Rodríguez-Triana, M. J., & Dillenbourg, P. (2018). Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of Computer Assisted Learning*, *34*(2), 193–203.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161. https://doi.org/10.1037/h0077714

Russell, S., & Norvig, P. (1995). Artificial Intelligence: A modern approach. *Prentice-Hall, Englewood Cliffs*, *25*(27), 79–80.

Schuller, B., Arsic, D., Rigoll, G., Wimmer, M., & Radig, B. (2007). Audiovisual behavior modeling by combined feature spaces. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Vol. 2, pp. II-733). Honolulu, HI: IEEE.

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association.* (pp. 312-315). Brighton, UK.

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *Journal of Educational Measurement*, *54*(1), 12–35.

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

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Tskhay, K. O., Zhu, R., Zou, C., & Rule, N. O. (2018). Charisma in everyday life: Conceptualization and validation of the General Charisma Inventory. *Journal of Personality and Social Psychology*, *114*(1), 131.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743–1759.

Zhou, F., & De la Torre, F. (2012). Generalized time warping for multi-modal alignment of human motion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1282–1289). Washington, DC: IEEE.