Improving Your Pitch with Facial Action Units… Is It Possible?

A study examining the extent to which pitchers’ facial action units can predict their rankings

Caitlin van Mil
Student number: u1271284

Thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Science in Data Science & Society
Department of Cognitive Science & Artificial Intelligence
School of Humanities and Digital Sciences
Tilburg University

Thesis committee:

Supervisor: Dr. Merel Jung
Second Reader: Prof. Dr. Max Louwerse

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2020

1. ABSTRACT

Entrepreneurs often pitch their idea to potential investors from whom they try to obtain critical resources. Since nonverbal behavior, and facial expressions in particular, are known to influence our daily interaction, but knowledge about their role in specific social contexts is limited, this study investigates the extent to which pitchers' rankings can be predicted from their facial action units (AUs), the extent to which this differs between the pitch presentation and the Q&A session, and which algorithm(s) do this best. Pitchers' facial AUs were expected to carry predictive capability for their rankings, and this capability was expected to be stronger for data from the Q&A session than from the pitch presentation. Data from 25 pitchers from the Jheronimus Academy of Data Science were used to train and test five different algorithms: K-Nearest Neighbor, multinomial logistic regression, Support Vector Machine, decision tree and random forest. Results indicate that pitchers' facial AUs do not carry sufficient predictive capability for their rankings, and that facial AUs during the pitch presentation are slightly better able to predict pitchers' rankings than those during the Q&A session, but only when a decision tree is used. More research is needed to uncover the different features that play a role in investors' evaluations of pitchers and the extent to which a confirmation bias is present among investors. Practical implications are discussed.

Keywords: facial action units, entrepreneurs, pitching, learning algorithm, facial expressions, nonverbal communication.

2. INTRODUCTION

Facial expressions consciously and unconsciously affect our daily interactions (Schmidt & Cohn, 2001). 55 Percent of received information emanates from the speaker's nonverbal communication and only about 7 percent originates from the content of the message (Scott, 1990). Yet, surprisingly little is known about the role of facial expressions in achieving certain results in different social contexts. This study investigates the extent to which pitchers' rankings can be predicted from their facial expressions, as measured by facial action units (AUs) (Cohn, Ambadar, & Ekman, 2007).

Based on the Brunswikian lens model and the regulatory-fit theory, pitchers' facial AUs are expected to affect their rankings. The Brunswikian lens model assumes that a sender's trait or state is communicated through nonverbal cues (Burgoon, Birk, & Pfau, 1990). The regulatory-fit theory suggests that a message's effectiveness is enhanced when the sender uses nonverbal cues that are in line with the sender's orientation towards the message (Cesario & Higgins, 2008). Additionally, since the pitches in this study consisted of two separate parts, a pitch presentation and a subsequent Q&A session, this study explores whether there is a difference in the predictive capability of the facial AUs during the pitch presentation and those during the Q&A session for pitchers' rankings. Based on literature about facial expressions (Frith, 2009), mimicry (Cardon, 2008; Chartrand & Van Baaren, 2009; Duffy & Chartrand, 2015; Van Baaren, Holland, Kawakami, & Van Knippenberg, 2004; Fischer-Lokou, Gueguen, Lamy, Martin, & Bullock, 2014) and entrepreneurial passion (Baron, 2008; Cardon, 2008; Chen, Yao, & Kotha, 2009; Cardon, Sudek, & Mitteness, 2009; Cardon, Wincent, Singh, & Drnovsek, 2009; Elsbach, 2003; Huy and Zott, 2007), pitchers' facial AUs during the Q&A session are expected to be more predictive of their rankings than their facial AUs during the pitch presentation.

Although facial expressions' impact on customer service rating (Kulesza, Dolinski, Huisman, & Maciejewski, 2014), negotiator agreement (Fischer-Lokou et al., 2014) and liking (Duffy & Chartrand, 2015) have been widely researched in the field of mimicry, the broader role of facial expressions in investors' judgements remains hitherto largely unknown (Clarke, Cornelissen, & Healey, 2019), even though they may be important drivers of investors' evaluations of or intuitions about entrepreneurs (Huang & Pearce, 2015). Additionally, the functions of different facial expressions have been extensively studied (cf. Blair, 2003; Schmidt & Cohn, 2001), but their consequences in different social contexts have been largely overlooked. Since individuals' decision-making is strongly affected by nonverbal behavior (Liebregts, Darnihamedani, Postma, & Atzmueller, 2019) and Zebrowitz and Montepare (1992) have found that individuals make assessments about whether a person is competent through nonverbal cues such as the perception of facial features (Huang & Pearce, 2015), there is a need for research investigating the effects of different facial expressions in different social contexts. By exploring the predictability of pitchers' rankings from their facial AUs, this study contributes to filling this gap in the literature.

If the results shows that ranking can indeed be predicted from pitcher facial AUs, investors could incorporate this knowledge into their decision-making. Investors should be aware of this effect and try to not let facial expressions distract from the quality of a pitcher's idea. Vice versa, those pitching their idea could make use of the knowledge that their facial expression can affect an investor's judgement of them and their pitch. More specifically, features' importance in well-performing models could provide insight into which facial AUs are important predictors of pitchers' rankings, thereby providing evidence-based recommendations for entrepreneurs in practice about how to use certain facial AUs, given that, in case of limited funding, investors are assumed to only fund the best ranked pitches.

The above illustrates a need for research investigating the effects of different facial expressions in different social contexts. The research question of this study is therefore as follows: to what extent can pitchers' rankings be predicted from their facial AUs, does this differ between the pitch presentation and the Q&A session, and which learning algorithm(s) perform(s) best? This translates into two sub questions:

1. To what extent can pitchers' rankings be predicted from their facial AUs, and which learning algorithm performs best?
2. Do the predictive capabilities of the facial AUs during the pitch presentation and the facial AUs during the Q&A session differ for pitchers' rankings, and which learning algorithm(s) perform(s) best on these data?

The questions will be answered by comparing the accuracies and macro F1 scores of the classification learning algorithms K-Nearest Neighbor (KNN), multinomial logistic regression, Support Vector Machine (SVM), decision tree and random forest against a majority baseline. The data will be split into a train-, validation- and test set, and nested cross-validation will be used to train the model, find the optimal hyperparameters, and evaluate its performance. For the second sub question in particular, these steps will be taken twice: once based on data from the pitch presentation, and once based on data from the Q&A session.

3. RELATED WORK

Entrepreneurs rely on stakeholders for the provision of critical resources (Nagy, Pollack, Rutherford, & Lohrke, 2012). In order to obtain such resources entrepreneurs often have to pitch their idea to potential investors. The success of such an entrepreneurial pitch largely depends on the pitcher's ability to convince these potential investors that the business model is worth investing in and to persuade them to invest. Likewise, Nagy et al. (2012) conclude that impression management by a pitcher plays a very important role in the provision of critical resources by stakeholders. Moreover, previous research suggests that nonverbal behavior plays an important role in the speaker's persuasiveness (Chidambaram, Chiang, & Mutlu, 2012), credibility (Burgoon et al., 1990) and ability to effectively negotiate (Lincoln, 2000; Scott, 1990). Literature by Scott (1990) contains relevant information if the interplay between a pitcher and a potential investor is viewed as a negotiation about how much to invest in a business model, which states that about 55 percent of received information comes from the nonverbal communication of a speaker and only about 7 percent originates from the content of the message. More broadly, facial expressions can contribute to the structure and meaning of what is being said (Bavelas & Chovil, 1997; Ekman, 1979) and be used to send signals with regards to the overall meaning of the message (Ekman, 1979).

Based on the Facial Action Coding System (FACS), facial expressions can be represented by AUs (Ekman & Friesen, 1977). Facial AUs are widely used and able to objectively describe facial muscle activations (Baltrusaitis, Zadeh, Lim, & Morency, 2018). In other words, facial AUs are the muscular actions that comprise a facial expression (Bartlett et al., 1996). They are the smallest visually distinguishable facial movements (Cohn et al., 2007). In total, 46 facial AUs have been defined (Bartlett et al., 1996). Combinations of facial AUs that occur can be either additive or non-additive. Additive combinations refer to facial AUs that appear independently of each other, while non-additive combinations modify facial AUs' appearances (Cohn et al., 2007).

Two models and theories are applied to illustrate the role nonverbal behavior plays in persuasiveness. First, the modified Brunswikian lens model specifies that "a particular trait or state of the sender is externalized or expressed in distal indicator cues" (Burgoon et al., 1990, p. 143). Simply put, it suggests that people's characteristics are communicated through nonverbal behavior. In line with this model, Burgoon et al. (1990) found that nonverbal behaviors were associated with attributions of credibility and persuasiveness. More specifically, they found that greater facial expressiveness contributed to competence perceptions. Second, the regulatory-fit theory highlights the role of someone's orientation to or concerns about an activity and the way in which this person engages with that activity. In other words, people will experience a feeling of fit when they send a message in a way that is aligned with their attitude towards this message, which makes them "feel right" and enhances their message effectiveness (Cesario & Higgins, 2008).

Prior research suggests investors assess pitchers on several criteria in order to determine whether or not they are going to invest in the idea that is being pitched. One of these criteria, which has received quite some attention in the literature, is entrepreneurial passion (Cardon et al., 2009a). Entrepreneurial passion is defined as "an entrepreneur's intense affective state accompanied by cognitive and behavioral manifestations of high personal value" (Chen et al., 2009, p. 199). It is considered critical for convincing investors to fund a pitcher's start-up (Chen et al., 2009), since it could provide a strong indication of someone's commitment (Vallerand et al., 2003). Notably, research suggests that displayed passion might be just as important as experienced passion (Cardon, 2008), since the extent to which people are expressive in displaying their emotions and the extent to which people display emotions they do not feel differs (Rafaeli & Sutton, 1987; Dasborough & Ashkanasy, 2002; Kring, Smith, & Neale, 1994). Since emotions are conveyed through facial expressions (Schmidt & Cohn, 2001), and entrepreneurial passion is considered an emotion (Cardon et al., 2009b), entrepreneurial passion is also assumed to be exhibited through facial expressions. Although this does not cover the broader topic of facial expressions, which this study aims to do by investigating the extent to which pitchers' rankings can be predicted from their facial AUs, it implies that pitchers can deliberately display passion or other positive emotions to be more persuasive and appear more confident, prompting investors to make more favorable decisions (Baron, 2008; Chen et al., 2009).

Correspondingly, in their quantitative study among entrepreneurs and stakeholders, Huy and Zott (2007) found that entrepreneurs actively manage emotions and that this affects resource mobilization. More specifically, successful entrepreneurs display passion, enthusiasm and low-activation positive emotion to communicate self-control, and that this increases investors' confidence in the business. Similarly, Cardon et al. (2009b) proposed entrepreneurial passion to be associated with creativity as the ability to recognize new patterns of information, perceptually process stimuli and leverage existing knowledge to find creative solutions. This is based on the line of thought assuming that individuals who experience positive affect have an adaptive approach towards environmental stimuli. In turn, creativity is argued to positively affect investment decisions following pitch sessions (Elsbach, 2003). Hereby, the investors particularly look at enthusiasm (affective passion), preparedness (cognitive passion) and commitment (behavioral passion). Remarkably, in their study among U.S. angel investment groups, Cardon et al. (2009a) did not find support for the hypotheses that affective passion and behavioral passion are positively associated with investor funding, although they did find this to be true for cognitive passion. Along the same lines, in their study among participants of a large public U.S. university's annual business plan competition, Chen et al. (2009) found passion to be significantly positively correlated with investment decision, but passion did not have a significant effect on investment decision in their logistic hierarchical regression. The authors of both studies suspect that these unexpected results indicate that entrepreneurial passion is a concept that needs more nuance. This is in line with Baron's (1989) analysis of having to differentiate those who are more sincere and genuine from those who are less sincere and genuine. The former group was found to receive more favorable evaluations than the latter, indicating that genuineness in facial expression might affect investors' evaluations of a pitcher.

When investors do not grasp the value of an entrepreneurial idea, the cause is as much rooted in the pitcher's traits as in the idea's inherent quality. Investors tend to reduce a pitcher to a stereotype and search for visual cues to base this categorization on (Elsbach, 2003). This implies that facial expressions are also used to categorize pitchers. Pitchers categorized as certain types make the investor feel like they are participating in the development of an idea, something investors respond really well to (Elsbach, 2003). Investors should be aware of this effect of categorization though, as it can take away from the amount of attention directed towards the quality of the idea. Moreover, relying too heavily on stereotypes can cause investors to overlook high-potential individuals who possess other qualities like creativity (Elsbach, 2003).

Building on the work outlined above, but focusing on the second sub question, facial AUs of the pitcher during the Q&A session are expected to have more predictive capability for their rankings than those during the pitch presentation. This because a Q&A session is more interactive than a pitch presentation, thus adding a more communicative purpose to the functions of facial expressions (Frith, 2009). For example, an important function that presents itself in a more interactive context is mimicry. Mimicry happens predominantly subconsciously and occurs in almost every social interaction (Duffy & Chartrand, 2015). More specifically, facial mimicry takes place when someone mimics the facial expression(s) of someone else (Duffy & Chartrand, 2015). Based on the fact that mimicry encourages affiliation between interaction partners and generates a prosocial orientation, mimicry was expected and found to make the interaction partners like, trust and help each other more (Chartrand & Van Baaren, 2009; Duffy & Chartrand, 2015). It also induces generosity, according to a study by Van Baaren et al. (2004) among undergraduate students from the University of Nijmegen. Furthermore, based on the knowledge that imitation has the ability to bind individuals to each other, Fischer-Lokou et al. (2014) expected that mimicry could increase the chance of agreement during negotiations. The findings of their study among students and mediators supported this expectation, but only when imitation was repeated and conducted over a longer period of time. Additionally, display of emotions through facial expressions and mimicry is proposed to invoke passion within employees through the contagion process of emotional mimicry which assumes that emotions are contagious (Cardon, 2008). The same is expected for investors. This passion is in turn argued to generate more positive investor decisions as explained above (Baron, 2008; Chen et al., 2009; Cardon et al. 2009a, 2009b; Elsbach, 2003; Huy and Zott, 2007). Therefore, data on the facial AUs stemming from this part of the pitch are expected to be more predictive of pitchers' rankings than data on the facial AUs stemming from the pitch presentation part of the pitch.

4. EXPERIMENTAL SETUP

4.1. DATA

Videos and survey data (Liebregts, Urbig, & Jung, 2018-2020) from multiple pitch sessions held at the Jheronimus Academy of Data Science (JADS) were provided for this study. In total there were 4 pitch sessions: one from November 2018 – January 2019, conducted for the third year of the joint bachelor data science (7 pitches); one from September 2019 – December 2019, conducted for the second year of the master data science & entrepreneurship (5 pitches); two from November 2019 – January 2020, conducted for the third year of the joint bachelor data science (15 pitches in total, 13 of which provided informed consent for using the data for research purposes). This resulted in a total of 25 pitches that could be analyzed. For each pitch, a video of the pitcher, videos of each of the three investors, and a video of all four persons combined, was provided. All pitcher videos were fed into OpenFace 2.2.0, a validated tool for facial behavior analysis which is also based on FACS (Baltrusaitis et al., 2018). OpenFace 2.2.0 outputted a csv file with columns on seventeen different AUs (both presence, represented as classification scores, and intensity, represented as regression scores), eye gaze, landmarks, head position and rigid and non-rigid face shape. This study focused on the facial AUs in particular, which are often used to represent facial expression, and objectively describe facial muscle activations (Ekman & Friesen, 1977). More specifically, only the regression scores on facial AUs were used in order to take the intensity of the facial AUs into account. OpenFace uses linear kernel SVM for AU recognition and the model demonstrates good performance with noisy data (Baltrusaitis et al., 2018). Each row of the csv output data represented 0.04 seconds in the video (= 25 FPS (Frames Per Second)), so the number of rows in the datasets was dependent on the length of the video. The number of rows had an average of 19.97 (≈ 13 minutes, 19 seconds), a minimum of 15.21 (≈ 10 minutes, 8 seconds) and a maximum of 32.67 (≈ 21 minutes, 47 seconds).

4.1.1. Data cleaning

First, the data sets from the different pitchers were loaded. In Python 3, missing values were replaced with the previous value. Low confidence rows and unnecessary columns were deleted. Analyses in R showed there were no missing values after removing these aforementioned rows and columns. Next, the different data sets of the pitchers were merged in R to generate the descriptives in the next section. Moreover, the pitchers' rankings from the three different judges were averaged. This average was then used to create the classes 'low', 'middle' and 'high' relative to the number of pitchers in that particular session, i.e. pitchers with the top 33% scores were assigned a 'high' ranking, pitchers with the next 33% scores were assigned a 'middle' ranking, and pitchers with the bottom 33% scores were assigned a 'low' ranking. This approach, rather than working with the original, averaged rankings, was chosen as the latter would have required regression, for which this study did not have enough samples. Additionally, lowering the number of classes to three ensured more generic classes with a higher, more reliable number of samples. A link to all code can be found in Appendix 1.

4.1.2. Descriptives

After data were cleaned and the data frame was prepared, some descriptives were generated in R. First, the descriptive statistics of the facial AUs were examined (see Table 1).

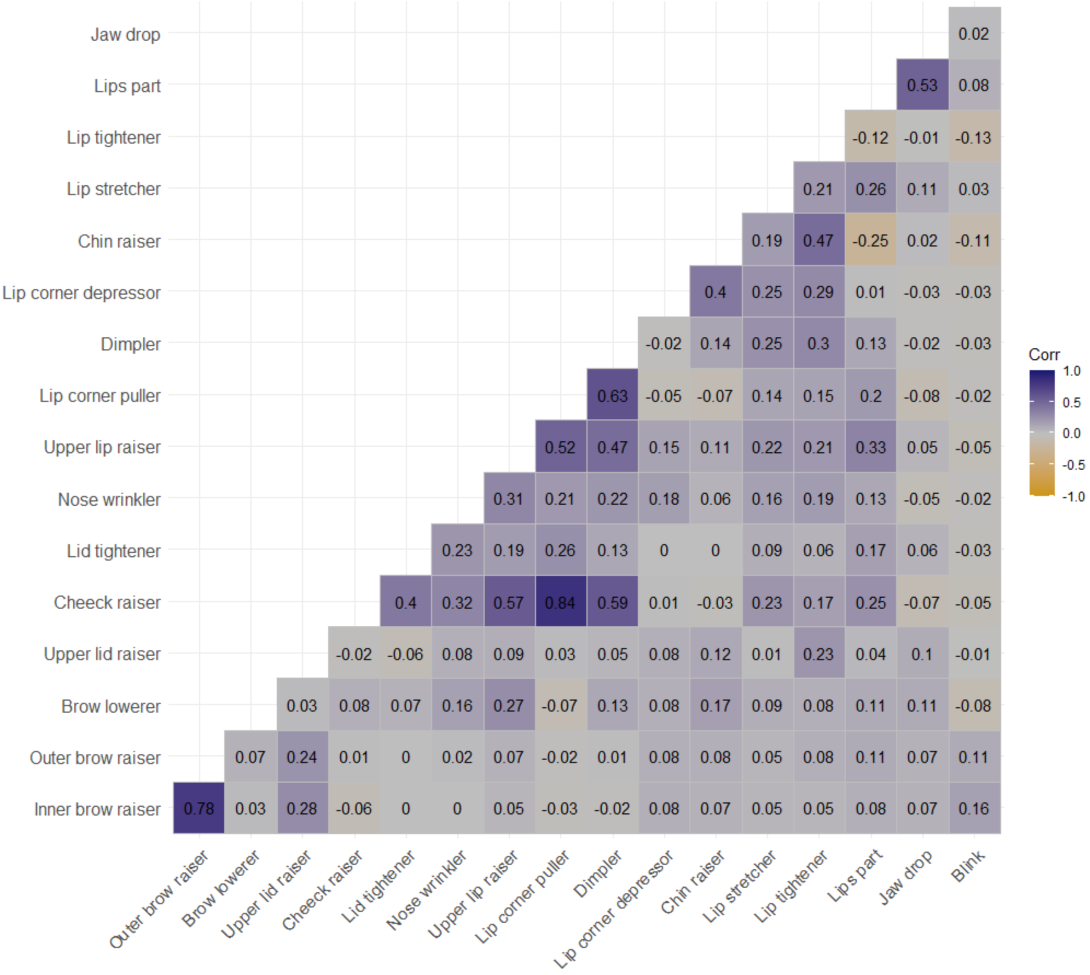
Table 1

Descriptive statistics of facial AUs

Action unit		*M*	*SD*	*Maximum*
Inner brow raiser	AU01	0.26	0.50	4.86
Outer brow raiser	AU02	0.13	0.34	5.00
Brow lowerer	AU04	0.37	0.42	4.19
Upper lid raiser	AU05	0.08	0.22	3.69
Cheek raiser	AU06	0.18	0.41	4.51
Lid tightener	AU07	0.42	0.54	4.78
Nose wrinkler	AU09	0.07	0.20	4.07
Upper lip raiser	AU10	0.36	0.52	3.81
Lip corner puller	AU12	0.23	0.52	3.76
Dimpler	AU14	0.26	0.43	4.33
Lip corner depressor	AU15	0.18	0.34	5.00
Chin raiser	AU17	0.50	0.49	5.00
Lip stretcher	AU20	0.13	0.25	5.00
Lip tightener	AU23	0.16	0.33	5.00
Lips part	AU25	0.68	0.68	5.00
Jaw drop	AU26	0.55	0.59	5.00
Blink	AU45	0.21	0.39	3.73

Note. All minima were 0.000(0).

Second, the Pearson correlations between the facial AUs and the Pearson correlations between the facial AUs and the timestamp were examined to see how certain facial AUs correlate and whether the use of certain facial AUs changes over time (see Appendix 2). Remarkable is that almost all correlations are statistically significant at the 0.001 level (2-tailed), including many Pearson correlations close to zero. Since these correlations are based on pairs of datapoints that represent just one point in time (one pair per value on the variable *timestamp*), the data set used for these correlations was very large. As confirmed by Sari et al. (2017), large data sets like these cause coefficients of low magnitude to have a high confidence interval amplitude, thus showing statistical significance. Moreover, Tian, Kanade, and Cohn (2001) confirm that facial AUs are often correlated, and these correlations somewhat seem in line with the study by Yüce, Gao, Cuendet, and Thiran (2017), that found correlations higher than 0.25 between roughly half the facial AUs included. Figure 1 displays the correlation matrix heatmap, which helps with the interpretation of the Pearson correlations between the AUs. Based on the guidelines by Cohen (1988) the Pearson correlations between inner brow raiser and outer brow raiser (*r* = 0.776), cheek raiser and lid tightener (*r* = 0.402), cheek raiser and nose wrinkler (*r* = 0.322), nose wrinkler and upper lip raiser (*r* = 0.312), cheek raiser and lip corner puller (*r* = 0.842), upper lip raiser and lip corner puller (*r* = 0.519), cheek raiser and dimpler (*r* = 0.593), upper lip raiser and dimpler (*r* = 0.475), lip corner puller and dimpler (*r* = 0.635), lip corner depressor and chin raiser (*r* = 0.403), dimpler and lip tightener (*r* = 0.301), chin raiser and lip tightener (*r* = 0.465), upper lip raiser and lips part (*r* = 0.328), and lips part and jaw drop (*r* = 0.532) were considered relatively strong. These stronger correlations are somewhat similar as those found by Yüce et al. (2017).



	Outer brow raiser	Brow lowerer	Upper lid raiser	Cheeck raiser	Lid tightener	Nose wrinkler	Upper lip raiser	Lip corner puller	Dimpler	Lip corner depressor	Chin raiser	Lip stretcher	Lip tightener	Lips part	Jaw drop	Blink
Jaw drop																0.02
Lips part															0.53	0.08
Lip tightener														-0.12	-0.01	-0.13
Lip stretcher													0.21	0.26	0.11	0.03
Chin raiser												0.19	0.47	-0.25	0.02	-0.11
Lip corner depressor											0.4	0.25	0.29	0.01	-0.03	-0.03
Dimpler										-0.02	0.14	0.25	0.3	0.13	-0.02	-0.03
Lip corner puller									0.63	-0.05	-0.07	0.14	0.15	0.2	-0.08	-0.02
Upper lip raiser								0.52	0.47	0.15	0.11	0.22	0.21	0.33	0.05	-0.05
Nose wrinkler							0.31	0.21	0.22	0.18	0.06	0.16	0.19	0.13	-0.05	-0.02
Lid tightener						0.23	0.19	0.26	0.13	0	0	0.09	0.06	0.17	0.06	-0.03
Cheeck raiser					0.4	0.32	0.57	0.84	0.59	0.01	-0.03	0.23	0.17	0.25	-0.07	-0.05
Upper lid raiser				-0.02	-0.06	0.08	0.09	0.03	0.05	0.08	0.12	0.01	0.23	0.04	0.1	-0.01
Brow lowerer			0.03	0.08	0.07	0.16	0.27	-0.07	0.13	0.08	0.17	0.09	0.08	0.11	0.11	-0.08
Outer brow raiser		0.07	0.24	0.01	0	0.02	0.07	-0.02	0.01	0.08	0.08	0.05	0.08	0.11	0.07	0.11
Inner brow raiser	0.78	0.03	0.28	-0.06	0	0	0.05	-0.03	-0.02	0.08	0.07	0.05	0.05	0.08	0.07	0.16

Figure 1. Correlation matrix heatmap of the facial AUs.

Finally, the number of pitchers in each of the classes for the rankings were plotted in order to see their proportions (see Figure 2).

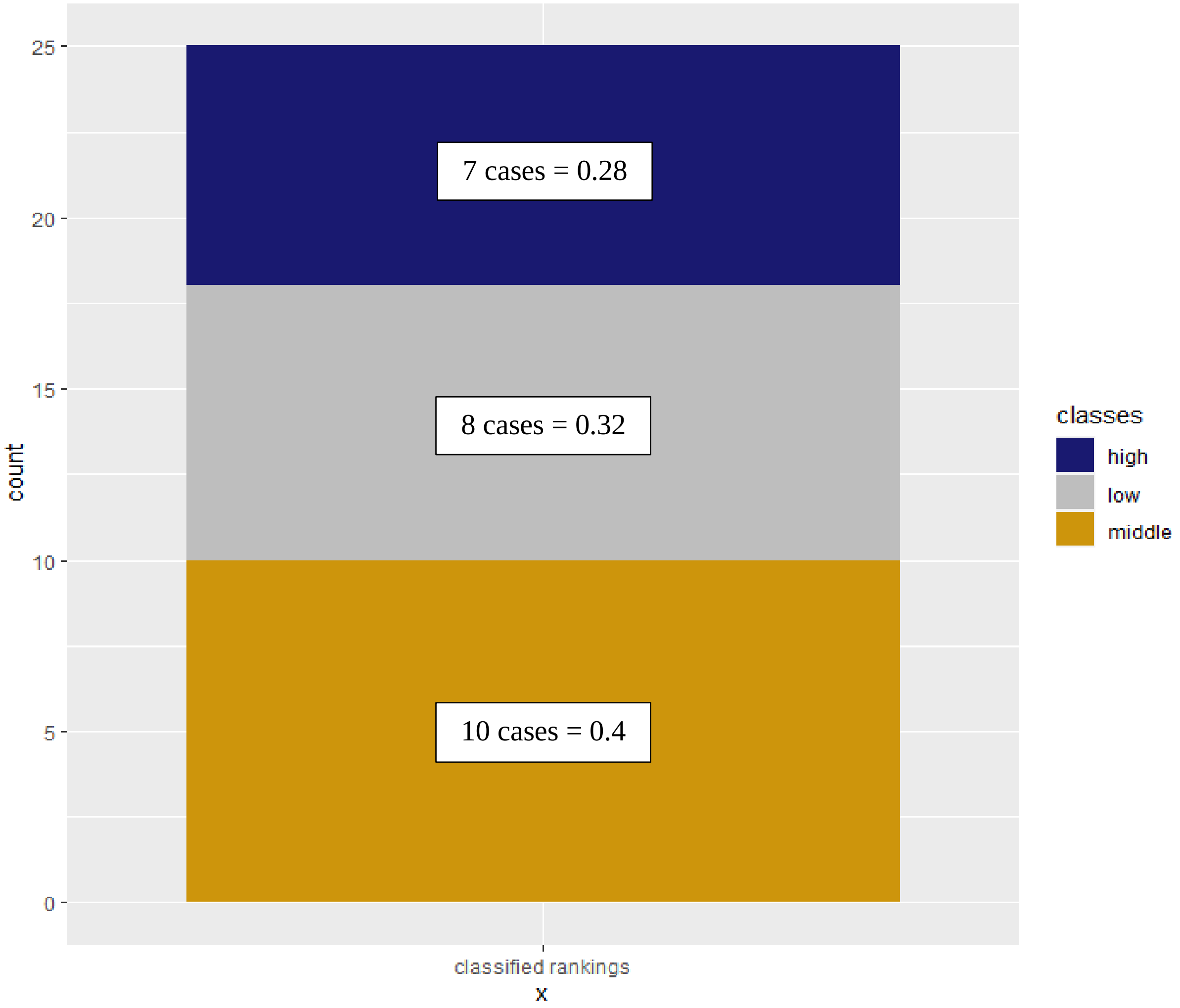


Figure 2. Counts of the classes for the rankings.

4.1.3. Pre-processing

All pre-processing was done in Python 3. First, the data were split to create separate data sets for the pitch presentation and the Q&A session, for the purpose of answering the second sub question, which aims at investigating whether there is a difference between the predictive capability of facial AUs during the pitch presentation and during the Q&A session. Next, features were created based on the mean, maximum, standard deviation, skewness and kurtosis of the different facial AUs. The minimum was disregarded since all minima were 0.00, therefore not carrying predictive capability. This resulted in seventeen (the number of facial AUs) times five (statistical features per facial AU) so 85 different features. In order to improve interpretability of the models and to make the data more suitable for the KNN algorithm, which does not deal very well with high dimensionality, features were selected based on their Pearson correlation with the classified ranking. First, the feature data were split into five folds. Three folds were set apart for training, one for validation and one for testing. These folds subsequently rotated resulting in five train-, validation- and test sets. Next, each train set, rather than all the data, was used to determine the features to be kept in order to prevent overfitting by selecting features based on data that are also included in the validation and test set. Only the features that correlated ≥ |0.2| with the classified ranking were selected. This feature selection was thereupon applied to the corresponding validation and test sets as well. The number of features in each of the five combinations of train-, validation- and test sets in the entire data set were 7, 18, 18, 25 and 21, in the pitch presentation data set 12, 23, 32, 24 and 24, and in the Q&A data set 15, 24, 28, 27 and 24. The features included in these folds can be found in Appendix 3. Additionally, the classes of the target were encoded as numbers to make them suitable for the algorithms that were used.

4.2. METHODS & MODELS

In this section, the packages used in R and Python 3, the optimization approach, the different classification algorithms and the evaluation metrics used to assess the performance of the algorithms are discussed.

4.2.4. Packages

All the packages used for data cleaning, preprocessing and processing can be found in Table 2 and 3 below. Table 2 represents the packages used in R 3.6.1, mainly used for cleaning and generating descriptives. Table 3 represents the packages used in Python 3, which were used for cleaning, pre-processing and processing.

Table 2

R packages

Package	Application	Source
dplyr (version 0.8.3)	Selecting only relevant columns (cleaning)	Kroeg et al. (2019)
reshape2 (version 1.4.3)	Creating correlation matrix heatmap (pre-processing)	Wickham (2017)
ggplot2 (version 3.3.0)	Creating rankings bar graph (pre-processing)	Wickham et al. (2020)
readr (version 1.3.1)	Reading csv files (loading data)	Wickham, Hester, Francois, & RStudio (2018)
ggcorrplot (version 0.1.3)	Correlation matrix heatmap (pre-processing)	Kassambara (2019)

Table 3

Python packages

Package	Application	Source
pandas (version 0.25.1)	Reading data, creating data frames (cleaning, pre-processing & processing)	Reback, Van den Bossche, MeeseeksMachine, Augspurger, Mendel (2019)
numpy (version 1.16.5)	Creating matrix with features, concatenating rows, averaging scores, setting seed (pre-processing & processing)	Shadchin et al. (2019)
scikit-learn (sklearn) (version 0.21.3)	Label encoding for target, scaling, classification algorithms, evaluation metrics (pre-processing & processing)	Pedregosa et al. (2011)

4.2.5. Nested cross-validation

Nested cross-validation was used to account for the small sample size and get an idea of the models' performances. Cross-validation in general is a good approach when the sample is small, since it varies which data to train and test on, thereby generating a more reliable insight into model performance (Varma & Simon, 2006). Nested cross-validation uses a pair of nested loops, whereby the inner loop sets the parameters to optimize training like sklearn's grid search, which deploys an exhaustive search over the specified parameter values to find those that maximize model performance. The outer loop applies these optimal parameters to a test set and estimates model performance (Cawley & Talbot, 2010; Varma & Simon, 2006). As can be concluded by the data splits described earlier, the number of folds in the outer loop was five, and the number of folds in the inner loop was four, three of which were used for training, one of which was used for validation. The advantage of this approach as opposed to (non-nested) cross-validation, is that it reduces the bias in the model performance estimate because tuning of the parameters is repeated in each cross-validation loop, while in (non-nested) cross-validation, parameter tuning happens on the data not left out in each cross-validation loop, therefore resulting in biased estimates of model performance (Varma & Simon, 2006).

4.2.6. Classification algorithms

In order to examine to what extent pitchers' rankings can be predicted from their facial AUs, to what extent this differs between the pitch presentation and the Q&A session, and which algorithm(s) do this best, five different algorithms were fitted, tuned and tested: KNN, multinomial logistic regression, SVM, decision tree and random forest.

4.2.6.1. K-Nearest Neighbor

K-Nearest Neighbor (KNN), is an algorithm predominantly used for classification tasks, but occasionally regression tasks as well. It projects the learning data into a large dimension space, whereby each feature of the data is represented by a dimension (Lubis, Lubis, & Al-Khowarizmi, 2020). When classifying a novel case, the algorithm places this case onto the large dimension space as well, and determines its distance from neighbors. It subsequently uses the targets of these neighbors to assign a target to the new case. In other words, it classifies the case based on data on which the algorithm was trained by comparing this previous data with the current data, i.e. the data belonging to the new case (Lubis et al., 2020). KNN is often called a lazy learning algorithm as it is not so much learning from the data as storing it (Mulak & Talhar, 2013). Additionally, it should be noted that KNN suffers from the curse of dimensionality which causes the distances between the nearest and furthest datapoints to become almost equal, thereby decreasing the algorithm's ability to provide valuable predictions (Kouiroukidis & Evangelidis, 2011).

The main hyperparameters of KNN are the number of neighbors (*K*), which represents a trade-off between overfitting and underfitting (Daumé, 2017), the weight function used in prediction, and the distance metric. The weights can be set to either uniform or distance. Uniform weights imply that every neighbor is treated equally important, while distance weights imply that closer neighbors are assigned a heavier weight than those further away from the new case, and therefore have more influence (Scikit-learn, 2007-2019). The distance metric calculates the distance i.e. similarity between the new case and the nearest neighbor (Lubis et al., 2020). The most common distance measure is Euclidean distance, which measures the distance as the length of the straight line segment joining the data points (O'Neil, 2008). This measure was also used in this study. It is calculated as follows (Daumé, 2017):

$$d(\boldsymbol{a}, \boldsymbol{b}) = \left[\sum_{d=1}^{D} (a_d - b_d)^2\right]^{\frac{1}{2}}$$

The variables *a* and *b* represent the vectors in a *D*-dimensional space between which the Euclidean distance *d* is to be calculated. So, the label of a new example is predicted by finding the training example that is most similar to the new example, i.e. the example for which the Euclidean distance *d* to the new example is minimized.

As mentioned, the hyperparameters of the algorithms were optimized through the inner loop of the nested cross-validation. For *K*, the number of neighbors, one up to and including three were included. For the weights, the options 'uniform' and 'distance' were included. The Euclidean distance was used for as the measure for distance.

4.2.6.2. Multinomial Logistic Regression

Logistic regression is a mathematical model used to describe the relationship between a dichotomous dependent variable, and several independent variables, in this case the pitchers' facial AUs (Kleinbaum & Klein, 2002). Despite its name, logistic regression is not used for regression problems, but for classification problems. It first calculates the logistic function *f*(*z*) (Kleinbaum & Klein, 2002):

$$f(z) = \frac{1}{1 + e^{-z}}$$

The variable z is an index that combines the independent variables and ranges from -∞ to +∞. It is calculated as follows (Kleinbaum & Klein, 2002):

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Both α and β_i are constant and represent unknown parameters and the X variables represent the independent variables in the model. The logistic function represents the probability of the outcome being one for a certain individual and thus ranges between zero and one. The function *f(z)* always takes an S-shape indicating that the outcome is likely to be zero for low values of *z*, until some threshold is reached, after which high values of *z* occur and the outcome is likely to be one (Kleinbaum & Klein, 2002).

However, in this study, the dependent variable, namely the classified rankings of pitchers, is not dichotomous, but polytomous, since there are more than two classes. For such multiclass problems, multinomial logistic regression is used. In this approach, a linear predictor function is constructed as follows:

$$score(X_i, k) = \beta_k \cdot X_i$$

The variable X_i is the vector of the independent variables that describes observation *i.* The variable β_k is a vector of weights that corresponds to the outcome *k*. Consequently, $score(X_i, k)$ is the score that is associated with observation *i* when assigned to outcome *k*. It can also be written similar to the function of *z*, depicted above:

$$f(k, i) = \beta_{0,k} + \beta_{1,k} X_{1,i} + \beta_{2,k} X_{2,i} + \cdots + \beta_{M,k} X_{M,i}$$

Hence, $\beta_{M,k}$ is the regression coefficient associated with the independent variable *m* and outcome *k*, and $X_{M,i}$ is the independent variable *m* associated with observation *i.*

The main hyperparameters of this algorithm are the penalty, the inverse of regularization strength *C*, and the solver. The penalty is the norm used in the penalization. The main penalties are ‘l1’ and ‘l2’. The ‘l1’ penalty, also referred to as lasso regression, shrinks the less important feature’s coefficient to zero, while the ‘l2’ penalty, also known as the ridge regression adds the squared magnitude of coefficients as penalty term to the loss function, preventing over-fitting (Nagpal, 2017; Scikit-learn, 2007-2019). Alternatively, no penalty can be applied. Finally, the solver is the algorithm to use in the optimization problem. The following solvers are supported by the Python library sklearn: newton-cg, lbfgs, liblinear, sag and saga (Scikit-learn, 2007-2019).

Given the fact that the target in this study included 3 classes, the multiclass hyperparameter was set to multinomial. Since only the ‘newton-cg’, ‘sag’, ‘saga’ and ‘lbfgs’ solvers handle multinomial loss, only these solvers were included. Since these solvers only handle ‘l2’ or no penalty, only these two penalties (‘l2’ and ‘none’) were included. For the inverse of regularization strength, C, the numbers 0.001, 0.01, 0.1, 1, 10 and 100 were included.

4.2.6.3. Support Vector Machine

The Support Vector Machine (SVM) is an approach to an optimization problem that tries to find a separating hyperplane with as large a margin as possible. The constrained optimization problem is written as follows (Daumé, 2017):

$$\underset{w,b}{min} \frac{1}{\gamma(w,b)}, \text{ subject to } y_n(w \cdot x_n + b) \geq 1$$

The variables w and b represent the weight and bias, respectively, and γ represents the margin. The constraint represents the requirement that the classification y of training example x, taking into account the weight and bias, has to be greater than one. The objective is to find the parameters that maximize this margin (Daumé, 2017).

The main hyperparameters of this algorithm are the kernel type to be used in the algorithm, gamma, and the decision function shape. The following kernel types are supported by the Python library sklearn: linear, poly, rbf and sigmoid. Gamma, which is the kernel coefficient for the kernel types 'rbf', 'poly' and 'sigmoid', can be set to either 'scale', in which case the algorithm uses 1/(n_features * X_var()) as the value of gamma, or 'auto', in which case it uses one divided by the number of features. Finally, the decision function shape can be set to either 'ovr' (one-vs-rest) of shape (n_features, n_classes) or 'ovo' (one-vs-one) of shape (n_samples, n_classes * (n_classes – 1) / 2). The latter is always used for multiclass problems like the one in this study, and is implemented in Support Vector Classification (SVC), which was used in this study (Scikit-learn, 2007-2019).

The kernel types 'rbf', 'linear', 'sigmoid' and 'poly' were included. For the hyperparameter gamma, the options 'scale' and 'auto' were included.

4.2.6.4. Decision Tree

A decision tree is a set of questions and guesses, written in a tree format, where the internal tree nodes represent the questions, and the leaves represent the guesses (Daumé, 2017). The first question is always the most useful question, i.e. the question based on which the most cases would be predicted correctly. This can be done by minimizing the impurity of the split. Two common measures of impurity are Gini impurity and entropy. Gini impurity represents how often a random element would be labeled incorrectly if the labels were assigned at random from a given distribution. It is calculated as follows (Xia, Zhang, Li, & Yang, 2008):

$$I_{gini}(T) = \sum_{i=1}^{k} p_i(1 - p_i)$$

The variable p_i is the probability of that an example belongs to class i, and k the number of classes. Entropy is a measure of uncertainty, where more uniform distributions have more uncertainty. It is calculated as follows:

$$I_H(P) = -\sum_i P_i \log_2(P_i)$$

The two measures are statistically indistinguishable (Xia et al., 2008).

The main hyperparameters of this algorithm are the function to measure the quality of a split, which can be either Gini impurity or entropy, the maximum depth of the tree, the maximum number of leaf nodes, and the minimum numbers of samples required to split internal nodes.

For the function to measure the quality of the splits, ‘gini’ and ‘entropy’ were included. For the maximum depth of the tree, the numbers four, six, eight and twelve were included. For the maximum number of leaf nodes, the numbers ranging from three to ten were included. Finally, for the minimum number of samples required to split an internal node, two, three and four were considered.

4.2.6.5. Random Forest

The random forest is an ensemble method that entails a collection of decision trees, in which the trees have fixed structures and random features. These decision trees are generated independently by the algorithm, and the features used at the branches of these trees are selected randomly. The leaves, which represent the predictions, are based on the training data. The resulting class of the random forest is decided based on a voting of the decision trees (Daumé, 2017).

The main hyperparameters of this algorithm are the same as for the decision tree, with the addition of the number of trees in the forest.

The inputs for the hyperparameters of the random forest were the same as those for the decision tree, with an addition for the number of trees in the forest including the numbers 2, 4, 6, 8, 10, 30, 50, 70 and 100.

4.2.7. Evaluation Metrics

To assess the performance of the algorithms, the evaluation metrics accuracy and macro F1 were chosen. Macro F1 scores give equal weights to each class, resulting in rare classes having the same impact as frequent classes (Van Asch, 2013). This is more suitable for these data than micro F1, because the proportions of the different classes in the target are somewhat unbalanced, although this is relatively challenging to determine with such few cases.

Accuracy is calculated as follows:

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ positives + false\ negatives}$$

Macro F1 is calculated as depicted below (Shmueli, 2019):

$$macro\ F1 = \frac{F1\ class\ 1 + F1\ class\ 2 \ldots + F1\ class\ k}{k}$$

Here, *k* represents the total number of classes in the data.

The evaluation metric F1 is calculated as follows:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

The calculation of the evaluation metrics precision and recall are depicted in the formulas below:

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

These evaluation metrics were computed for each of the five learning algorithms corresponding to the data of each of the three data sets, and compared to the majority baseline of 0.4 (see Figure 2).

5. RESULTS

In this section, classification performance of the five different learning algorithms on the three data sets are presented, thereby uncovering the extent to which pitchers' facial AUs can predict their rankings, whether this differs for the pitch presentation and the Q&A session, and which learning algorithm(s) perform best. The optimal hyperparameter settings can be found in Appendix 4.

5.1. K-NEAREST NEIGHBOR

Based on the entire data set, the five outer loops produced the following accuracy scores: 0.4, 0.2, 0.0, 0.4 and 0.0. This resulted in an average accuracy of 0.2, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.49, 0.11, 0.17, 0.22 and 0.0. This resulted in an average macro F1 score of 0.2, which does not outperform the majority baseline of 0.4.

Based on the pitch presentation data, the five outer loops produced the following accuracy scores: 0.8, 0.4, 0.0, 0.4 and 0.2. This resulted in an average accuracy of 0.36, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.82, 0.36, 0.0, 0.36 and 0.22. This resulted in an average macro F1 score of 0.35, which does not outperform the majority baseline of 0.4.

Based on the Q&A session data, the five outer loops produced the following accuracy scores: 0.4, 0.4, 0.2, 0.2 and 0.6. This resulted in an average accuracy of 0.36, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.33, 0.22, 0.17, 0.13 and 0.61. This resulted in an average macro F1 score of 0.29, which does not outperform the majority baseline of 0.4.

5.2. Multinomial Logistic Regression

Based on the entire data set, the five outer loops produced the following accuracy scores: 0.4, 0.2, 0.0, 0.0 and 0.4. This resulted in an average accuracy of 0.2, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.43, 0.17, 0.0, 0.0 and 0.36. This resulted in an average macro F1 score of 0.19, which does not outperform the majority baseline of 0.4.

Based on the pitch presentation data, the five outer loops produced the following accuracy scores: 0.6, 0.6, 0.2, 0.0 and 0.2. This resulted in an average accuracy of 0.32, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.67, 0.61, 0.17, 0.0 and 0.22. This resulted in an average macro F1 score of 0.33, which does not outperform the majority baseline of 0.4.

Based on the Q&A session data, the five outer loops produced the following accuracy scores: 0.4, 0.2, 0.4, 0.0 and 0.4. This resulted in an average accuracy of 0.28, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.5, 0.13, 0.19, 0.0 and 0.33. This resulted in an average macro F1 score of 0.23, which does not outperform the majority baseline of 0.4.

5.3. Support Vector Machine

Based on the entire data set, the five outer loops produced the following accuracy scores: 0.6, 0.2, 0.0, 0.0 and 0.2. This resulted in an average accuracy of 0.2, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.49, 0.11, 0.0, 0.0 and 0.11. This resulted in an average macro F1 score of 0.14, which does not outperform the majority baseline of 0.4.

Based on the pitch presentation data, the five outer loops produced the following accuracy scores: 0.8, 0.6, 0.0, 0.0 and 0.2. This resulted in an average accuracy of 0.32, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.82, 0.61, 0.0, 0.0 and 0.22. This resulted in an average macro F1 score of 0.33, which does not outperform the majority baseline of 0.4.

Based on the Q&A session data, the five outer loops produced the following accuracy scores: 0.4, 0.4, 0.2, 0.0 and 0.6. This resulted in an average accuracy of 0.32, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.5, 0.19, 0.17, 0.0 and 0.61. This resulted in an average macro F1 score of 0.29, which does not outperform the majority baseline of 0.4.

5.4. Decision Tree

Based on the entire data set, the five outer loops produced the following accuracy scores: 0.6, 0.4, 0.0, 0.6 and 0.4. This resulted in an average accuracy of 0.4, which is the same as the majority baseline. Moreover, the five outer loops produced the following macro F1 scores: 0.82, 0.22, 0.0, 0.29 and 0.3. This resulted in an average macro F1 score of 0.33, which does not outperform the majority baseline of 0.4.

Based on the pitch presentation data, the five outer loops produced the following accuracy scores: 0.8, 0.6, 0.0, 0.6 and 0.4. This resulted in an average accuracy of 0.48, which slightly outperforms the majority baseline of 0.4. Because this score was above the majority baseline, the features' importance in this model was examined (see Appendix 5). This suggested the mean, maximum and kurtosis of lid tightener, the mean of outer brow raiser and the kurtosis of upper lid raiser to be relatively important features. Moreover, the five outer loops produced the following macro F1 scores: 0.6, 0.56, 0.0, 0.44 and 0.36. This resulted in an average macro F1 score of 0.39, which does not outperform the majority baseline of 0.4.

Based on the Q&A session data, the five outer loops produced the following accuracy scores: 0.2, 0.4, 0.4, 0.6 and 0.4. This resulted in an average accuracy of 0.4, which is the same as the majority baseline. Moreover, the five outer loops produced the following macro F1 scores: 0.0, 0.22, 0.19, 0.29 and 0.22. This resulted in an average macro F1 score of 0.18, which does not outperform the majority baseline of 0.4.

5.5. RANDOM FOREST

Based on the entire data set, the five outer loops produced the following accuracy scores: 0.4, 0.2, 0.4, 0.0 and 0.2. This resulted in an average accuracy of 0.24, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.33, 0.17, 0.36, 0.13 and 0.11. This resulted in an average macro F1 score of 0.22, which does not outperform the majority baseline of 0.4.

Based on the pitch presentation data, the five outer loops produced the following accuracy scores: 0.4, 0.2, 0.2, 0.0 and 0.4. This resulted in an average accuracy of 0.24, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.27, 0.13, 0.13, 0.0 and 0.13. This resulted in an average macro F1 score of 0.13, which does not outperform the majority baseline of 0.4.

Based on the Q&A session data, the five outer loops produced the following accuracy scores: 0.6, 0.0, 0.4, 0.2 and 0.4. This resulted in an average accuracy of 0.32, which does not outperform the majority baseline of 0.4. Moreover, the five outer loops produced the following macro F1 scores: 0.5, 0.0, 0.22, 0.11 and 0.27. This resulted in an average macro F1 score of 0.22, which does not outperform the majority baseline of 0.4.

On the entire data set, the decision tree performed the best, and KNN, the multinomial logistic regression and the SVM performed the worst when accuracy was chosen as the evaluation metric, while when macro F1 was chosen as the evaluation metric, the SVM performed the worst. Table 4 below sums up the different algorithms, their average accuracy and macro F1 over the five outer loops, and their performance in comparison with the majority baseline of 0.4.

Table 4

Algorithm performances on entire data set

Algorithm	Average accuracy	Δ	Average macro F1	Δ
KNN	.20	-50%	.20	-50%
Multinomial logistic regression	.20	-50%	.19	-52.5%
SVM	.20	-50%	.14	-65%
Decision tree	.40	0%	.33	-17.5%
Random forest	.24	-40%	.22	-45%

On the pitch presentation data, the decision tree performed the best, and the random forest performed the worst. Table 5 below sums up the different algorithms, their average accuracy and macro F1 over the five outer loops, and their performance in comparison with the majority baseline of 0.4.

Table 5

Algorithm performances on pitch presentation data

Algorithm	Average accuracy	Δ	Average macro F1	Δ
KNN	.36	-10%	.35	-12.5%
Multinomial logistic regression	.32	-20%	.33	-17.5%
SVM	.32	-20%	.33	-17.5%
Decision tree	.48	+20%	.39	-2.5%
Random forest	.24	-40%	.13	-67.5%

On the Q&A session data, when accuracy was chosen as the evaluation metric, the decision tree performed the best and the multinomial logistic regression performed the worst. When macro F1 was chosen as the evaluation metric, KNN and the SVM performed the best and the decision tree performed the worst. Table 6 below sums up the different algorithms, their average accuracy and macro F1 over the five outer loops, and their performance in comparison with the majority baseline of 0.4.

Table 6

Algorithm performances on Q&A session data

Algorithm	Average accuracy	Δ	Average macro F1	Δ
KNN	.36	-10%	.29	-27.5%
Multinomial logistic regression	.28	-30%	.23	-42.5%
SVM	.32	-20%	.29	-27.5%
Decision tree	.40	0%	.18	-55%
Random forest	.32	-20%	.22	-45%

6. DISCUSSION

This study examined the following: to what extent can pitchers' rankings be predicted from their facial AUs, does this differ between the pitch presentation and the Q&A session, and which learning algorithm(s) perform(s) best? Data from 25 students from the JADS were analysed. Five different algorithms were used: KNN, multinomial logistic regression, SVM, decision tree and random forest. The research question was tackled by addressing two sub questions.

The first sub question was "To what extent can pitchers' rankings be predicted from their facial AUs, and which learning algorithm performs best?" The highest accuracy and macro F1 scores on the entire data set were 0.40, which is the same as the majority baseline and 0.33, which is 17.5% lower than the baseline, respectively. Both were produced by the decision tree. Since this accuracy score is merely the same as the majority baseline, and this macro F1 score is even lower than the majority baseline of 0.4, these results suggest that the algorithm is not learning anything. This implies that the features used in this study to measure pitchers' facial AUs do not carry sufficient predictive capability for their rankings.

The second sub question was "Do the predictive capabilities of the facial AUs during the pitch presentation and the facial AUs during the Q&A session differ for pitchers' rankings, and which learning algorithm(s) perform(s) best on these data?". This question was answered by evaluating the performances of the same five algorithms based on either the data representing the pitch presentation or the data representing the Q&A session. Based on the pitch presentation data as well as the Q&A session data, the highest accuracy scores were produced by the decision tree. These accuracy scores were 0.48, which outperforms the majority baseline of 0.4 by 20%, and 0.4, which is the same as the majority baseline of 0.4, respectively. Judging from these accuracy scores, pitchers' facial AUs during the pitch presentation seem to carry some predictive capability for their rankings if a decision tree is used, and the Q&A session data seem to carry less predictive capability for pitchers' rankings. However, averaged over the different algorithms, the accuracy based on the pitch presentation data and the Q&A session data were very similar, although the macro F1 was somewhat lower for the Q&A session data. Anyhow, this study did not find support for the expectation that the data from the Q&A session would be better able to predict pitchers' rankings than the data from the pitch presentation. This expectation was based on the idea that facial AUs are more functional in a more interactive setting like the Q&A session, for example through mimicry. The fact that this study did not find support for this expectation could be explained by the mitigating impact of first impressions and confirmation bias. According to Gaffey (2014), pitchers have roughly five minutes to generate the right first impression. After this, the judges' decisions are less likely to be influenced by the pitchers' facial AUs because they will have formed their opinion about the pitcher and will have a tendency to look for, interpret and use evidence in a way that confirms their existing beliefs about the pitcher, called conformation bias (Charness & Dave, 2017), meaning that investors form their opinion about the pitcher during the pitch presentation, and look for confirmation of that opinion during the Q&A session that follows.

The fact that most of the times, the decision tree performed the best could be partly due to the fact that decision trees are less prone to overfitting when the sample size is small (Domingos, 1998; Morgan, Daugherty, Hilchie, & Carey, 2003; Oates & Jensen, 1997). Furthermore, it should be noted that, contrary to what one might expect, the random forests did not outperform the decision trees. This is in line with Ali, Khan, Ahmad, and Maqsood (2012), who found that the decision tree deals better with small data sets. Moreover, the performance estimates based on the random forests were more instable than those of the decision trees. This strokes with Wang, Yang, and Luo (2016) who found that random forests lack stability and that high dimensional and small sample data sets increase randomness.

Finally, it is remarkable that the best performance was yielded by a model based on the pitch presentation data set, rather than the entire data set. Although the limitations that will be discussed below should give rise to great caution, this could mean that decision trees are better able to build well-performing models based on data solely from the pitch presentation rather than the entire pitch.

6.1. Limitations & Future Research

Although this study has contributed to the literature addressing the broader role of facial expressions in investors' judgements and the consequences of facial expressions in different social contexts, some methodological limitations and theoretically critical notes have to be discussed.

First, data from only 25 pitchers were available. Although this issue was dealt with by using cross-validation, more data would have most likely resulted in higher accuracy scores (Stockwell & Peterson, 2002). Additionally, it is also argued that cross-validation performed on small samples results in excessive variance and unreliable individual performance estimates (Braga-Neto & Dougherty, 2004). More data thus would have resulted in more reliable model performance estimates. Additionally, more data would have created the opportunity to use a convolutional neural network (CCN), which handles 3D data. Since the original format of the data was 3D, this model would have been preferable and would have retained more information than the models ultimately used in this study. However, the large number of parameters in CNNs require a lot of training samples (Keshari, Vatsa, Singh, & Noore, 2018), which were not available in this study. So for more solid research, more data is necessary, and future studies that have access to a larger amount of data should explore the predictability of pitchers' rankings with facial AUs with a CNN.

Second, the generalizability of the results is limited as the sample in this study consisted of students from the JADS in the Netherlands. Future research should conduct similar research among non-students and in different countries, controlling for cultural influences, given the proven relationship between culture and entrepreneurship (cf. Doepke & Zilibotti, 2014; Huggins & Thompson, 2014; Liñán & Fernandez-Serrano, 2014). Additionally, the fact that the sample was represented by students rather than actual entrepreneurs, and that the rankings were provided by hypothetical rather than actual investors, could limit the generalizability of the results of this study to the context of entrepreneurship.

Third, because the results suggest that when a decision tree is used, the predictive capability of pitchers' facial AUs during the pitch presentation for their rankings is higher than those during the Q&A sessions, future studies should investigate whether a confirmation bias is present with potential investors.

Fourth, the effects of pitchers' characteristics and the individual judges on pitchers' rankings was not taken into account due to lack of sufficient data. Perhaps certain characteristics of the pitchers and judges might carry predictive capability for pitchers' rankings. Future research could develop a more comprehensive overview of the different features playing a role in investors' judgements of pitchers and uncover effects like the relationship between an entrepreneur's gender and funding decisions, which has yielded mixed conclusions so far (cf. Johnson, Stevenson, & Letwin, 2018; Thébaud & Sharkey, 2016).

Furthermore, much related work was rooted in the line of thought that entrepreneurial passion positively affects investors' judgements since it conveys how committed the pitcher is. Nevertheless, the question remains whether this is caused because entrepreneurial passion triggers a logical process within the investor that pitchers' passion will make them succeed in their endeavors as a result of devotion, or because it triggers an unconscious process in which entrepreneurial passion as a positive emotion is transferred onto the investors. Although the first is the main idea of the literature discussing this construct, other literature does confirm this transfer of positive emotion (Parkinson, 2011). Future research should investigate the extent to which these different processes play a role since this could be something investors should be aware of when evaluating a pitch, especially when this affective transfer is of great influence.

6.2. Practical Implications

The results of this study suggest that facial AUs during a pitch, have none to very limited predictive capability for pitchers' rankings, and that those during the Q&A session are not better at predicting pitchers' rankings than those during the pitch presentation. Rather, the opposite was found to be true when a decision tree was used. Although replication studies that address this study's limitations are warranted, this finding could provide investors with the recommendation to be aware of the effect of facial expressions during the pitch presentation, to prevent their evaluation from being based on the pitchers' facial expressions rather than the quality of the entrepreneurial idea. Vice versa, those pitching could keep in mind that their facial expressions during the pitch presentation can affect investors' judgements of them and their pitch, especially their use of lid tightener, outer brow raiser and upper lid raiser. However, the magnitude of this effect seems limited and facial expressions could be something that investors consciously want to evaluate if the success of the business will be dependent on the entrepreneur's ability to influence others. If the facial expressions of a pitcher are considered unimportant to the success of the pitcher's entrepreneurial idea, investors could consider doing a blind pitch, which puts all the attention on the inherent quality of the entrepreneurial idea. And although more research is needed, this result may recommend investors to be aware of the presence of a potential confirmation bias.

7. CONCLUSION

This study investigated the extent to which pitchers' rankings could be predicted from their facial AUs, the extent to which this differed between the pitch presentation and the Q&A session and which learning algorithm(s) performed best on the data. The results of this study suggested that facial AUs have none to very limited predictive capability for pitchers' rankings. Only if a suitable algorithm is used, do pitchers' facial AUs during the pitch presentation seem to carry some predictive capability for their rankings. Although this study contributed to the literature on the role of facial expressions in investors' judgements of pitchers, further research pertaining to the different features that play a role in investors' evaluations of pitchers, the presence of a confirmation bias with investors, and the performance of a CNN on this kind of data is warranted. Pitchers and investors should be aware of a potential confirmation bias and the potential influence of the pitchers' facial expressions on funding decisions and consider doing a blind pitch.

8. ACKNOWLEDGEMENTS

There are several people to whom I owe my gratitude and whose contributions should be acknowledged. First of all, Dr. Werner Liebregts, Assistant Professor at JADS and Tilburg University, who provided me with the videos and survey data for this study. Secondly, my supervisor Dr. Merel Jung, who provided me with feedback and ideas for further improvement. Third, the DSS master students in the same thesis circle and my dear friend Maartje Verhoeven, for acting as valuable sparring partners. Fifth, my twin brother Sean van Mil and Roan van Sheppingen for providing me with feedback. Finally, my partner Jeroen Crompvoets, for helping me improve my skills in programming with matrices.

9. REFERENCES

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues, 9*(5), 272-278. Retrieved from: http://www.uetpeshawar.edu.pk/TRP-G/Dr.Nasir-Ahmad-TRP/Journals/2012/Random%20Forests%20and%20Decision%20Trees.pdf

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018, May). Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018*, 59-66. doi: 10.1109/FG.2018.00019

Baron, R. A. (1989). Impression management by applicants during employment interviews: The “too much of a good thing” effect. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 204-215). Thousand Oaks, CA: SAGE Publications, Inc.

Baron, R. A. (2008). The role of affect in the entrepreneurial process. *Academy of Management Review, 33*(2), 328-340. doi: 10.5465/amr.2008.31193166

Bartlett. M. S., Viola, P. A., Sejnowski, T. J., Golomb, B. A., Larsen, J., Hager, J. C., & Ekman, P. (1996). Classifying facial action. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems* (pp. 823-829). Cambridge, MA: MIT Press.

Bavelas, J. B., & Chovil, N. (1997). Faces in dialogue. In J. A. Russel, & J. M. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 334-346). Cambridge, England: Cambridge University Press.

Blair, R. J. R. (2003). Facial expressions, their communicatory functions and neuro-cognitive substrates. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 561-572. doi: 10.1098/rstb.2002.1220

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics, 20*(3), 374-380. doi: 10.1093/bioinformatics/btg419

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal Behaviors, Persuasion, and Credibility. *Human Communication Research, 17*(1), 140-169. doi: 10.1111/j.1468-2958.1990.tb00229.x

Cardon, M. S. (2008). Is passion contagious? The transference of entrepreneurial emotion to employees. *Human Resource Management Review, 18*(2), 77-86. doi: 10.1016/j.hrmr.2008.04.001

Cardon, M. S. Sudek, R., & Mitteness, C. (2009a). The impact of perceived entrepreneurial passion on angel investing. *Frontiers of Entrepreneurship Research, 29*(2), 1-15. Retrieved from: https://www.researchgate.net/profile/Melissa_Cardon/publication/228466857_The_impact _of_perceived_entrepreneurial_passion_on_angel_investing/links/0deec51701e6c7eec8000000.pdf

Cardon, M. S., Wincent, J., Singh, J., & Drnovsek, M. (2009b). The Nature and Experience of Entrepreneurial Passion. *Academy of Management Review, 34*(3), 511-532. doi: 10.5465/amr.2009.40633190

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research, 11*(70), 2079-2107. Retrieved from: http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Cesario, J., & Higgins, E. T. (2008). Making Message Recipients "Feel Right": How Nonverbal Cues Can Increase Persuasion. *Psychological Science, 19*(5), 415-420. doi: 10.1111/j.1467-9280.2008.02102.x

Charness, G., & Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior, 104*, 1-23. doi: 10.1016/j.geb.2017.02.015

Chartrand, T. L., & Van Baaren, R. (2009). Human mimicry. *Advances in experimental social psychology, 41,* 219-274. doi: 10.1016/S0065-2601(08)00405-X

Chen, X., Yao, X., & Kotha, S. B. (2009). Entrepreneur passion and preparedness in entrepreneurs' business plan presentations: A persuasion analysis of venture capitalists' funding decisions. *Academy of Management Journal, 52*(1), 199-214. doi: 10.5465/amj.2009.36462018.

Chidambaram, V., Chiang, Y-H., & Mutlu, B. (2012). Designing Persuasive Robots: How Robots Might Persuade People Using Vocal and Nonverbal Cues. *HRI'12: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction,* 293-300. doi: 10.1145/2157689.2157798

Clarke, J. S., Cornelissen, J. P., & Healey, M. P. (2019). Actions speak louder than words: How figurative language and gesturing in entrepreneurial pitches influences investment judgments. *Academy of Management Journal*, *62*(2), 335-360. doi: 10.5465/amj.2016.1008

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. In J. A. Coan, & J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment* (pp. 203-221). New York, NY: Oxford University Press.

Dasborough, M. T., & Ashkanasy, N. M. (2002). Emotion and attribution of intentionality in leader-member relationships. *Leadership Quarterly, 13*(5): 615-634. doi: 10.1016/S1048-9843(02)00147-9

Daumé, H. (2017). *A Course in Machine Learning.* Retrieved from: http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf

Doepke, M., & Zilibotti, F. (2014). Culture, Entrepreneurship, and Growth. In P. Aghion, & S. N. Durlauf (Eds.), *Handbook of Economic Growth* (2nd ed.) (pp. 1-48). Oxford, England: Elsevier B.V.

Domingos, P. (1998). Occam's two razors: The sharp and the blunt. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 37-43. Retrieved from: https://www.aaai.org/Papers/KDD/1998/KDD98-006.pdf

Duffy, K. A., & Chartrand, T. L. (2015). Mimicry: causes and consequences. *Current Opinion in Behavioral Sciences, 3,* 112-116. doi: 10.1016/j.cobeha.2015.03.002

Ekman, P. (1979). About brows: emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, & D. Ploog (Eds.), *Human ethology: claims and limits of a new discipline* (pp. 169-222). New York, NY: Cambridge University Press.

Ekman, P., & Friesen, W. V. (1977). *Manual for the Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press.

Elsbach, K. D. (2003). How to pitch a brilliant idea. *Harvard Business Review, 81*(9), 117-123. Retrieved from: https://escholarship.org/uc/item/1w82h22c

Fischer-Lokou, J., Guéguen, N., Lamy, L., Martin, A., & Bullock, A. (2014). Imitation in mediation: Effects of the duration of mimicry on reaching agreement. *Social Behavior and Personality: an international journal, 42*(2), 189-195. doi: 10.2224/sbp.2014.42.2.189

Frith, C. (2009). Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B, 364*(1535), 3453-3458. doi: 10.1098/rstb.2009.0142

Gaffey, A. (2014). *The elevator pitch.* Retrieved from: https://www.apa.org/SCIENCE/ABOUT/PSA/2014/06/ELEVATOR-PITCH

Huang, L., & Pearce, J. (2015). Managing the Unknowable: The Effectiveness of Early-stage Investor Gut Feel in Entrepreneurial Investment Decisions. *Administrative Science Quarterly, 60*(4), 634-670. doi: 10.1177/0001839215597270

Huggins, R., & Thompson, P. (2014). Culture, entrepreneurship and uneven development: a spatial analysis. *Entrepreneurship & Regional Development, 26*(9-10), 726-752. doi: 10.1080/08985626.2014.985740

Huy, Q., & Zott, C. (2007). How entrepreneurs manage stakeholders' emotions to build new organizations. *Academy of Management Proceedings, 2007*(1), 1-6. doi: 10.5465/ambpp.2007.26518309

Johnson, M. A., Stevenson, R. M., & Letwin, C. R. (2018). A woman's place is the… startup! Crowdfunder judgements, implicit bias, and the stereotype content model. *Journal of Business Venturing, 33*(6), 813-831. doi: 10.1016/j.jbusvent.2018.04.003

Kassambara, A. (2019). *Package 'ggcorrplot'.* Retrieved from: https://cran.r-project.org/web/packages/ggcorrplot/ggcorrplot.pdf

Keshari, R., Vatsa, M., Singh, R., & Noore, A. (2018). Learning structure and strength of CNN filters for small sample size training. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9349-9358. Retrieved from: http://openaccess.thecvf.com/content_cvpr_ 2018/papers/Keshari_Learning_Structure_and_CVPR_2018_paper.pdf

Kleinbaum, D. G., & Klein, M. (2002). *Logistic Regression: A Self-Learning Text* (2nd ed.). New York, NY: Springer-Verlag.

Kouiroukidis, N., & Evangelidis, G. (2011). The Effects of Dimensionality Curse in High Dimensional kNN Search. *2011 15th Panhellenic Conference on Informatics*, 41-45. doi: 10.1109/PCI.2011.45

Kring, A. M., Smith, D. A., & Neale, H. M. (1994). Individual differences in dispositional expressiveness: Development and validation of the emotional expressivity scale. *Journal of Personality and Social Psychology, 66*(5): 934-949. doi: 10.1037/0022-3514.66.5.934

Kroeg, A. J., Scheidr, B., Vaughan, D., Hicks, D., Farns, G. V., Patil, I., … Kercheval, S. (2019). *dplyr 0.8.3*. Retrieved from: https://www.tidyverse.org/blog/2019/07/dplyr-0-8-3/

Kulesza, W., Dolinski, D., Huisman, A., & Majewski, R. (2014). The echo effect: The power of verbal mimicry to influence prosocial behavior. *Journal of Language and Social Psychology, 33*(2), 183-201. doi: 10.1177/0261927X13506906

Liebregts, W., Darhihamedani, P., Postma, E., Atzmueller, M. (2019). The promise of social signal processing for research on decision-making in entrepreneurial contexts. *Small Business Economics,* 1-17. doi: 10.1007/s11187-019-00205-1

Liebregts, W. J., Urbig, D., & Jung, M. M. (2018-2020). [Survey and video data regarding entrepreneurial pitches and investment decisions]. Unpublished raw data.

Liñán, F., & Fernandez-Serrano, J. (2014). National culture, entrepreneurship and economic development: different patterns across the European Union. *Small Business Economics, 42,* 685-701. doi: 10.1007/s11187-013-9520-x

Lincoln, M. G. (2000). Negotiation: The opposing sides of verbal and nonverbal communication. *Journal of Collective Negotiations, 29*(4), 297-306. Retrieved from: https://triggered.edina.clockss.org/ServeContent?url=http://baywood.stanford.clockss. org%2FBWCN%2FBAWOOD_BWCN_29_4%2FL278ETET3BLGE9QP.pdf

Lubis, A. R., Lubis, M., & Al-Khowarizmi, A. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics, 9*(1), 326-338. doi: 10.11591/eei.v9i1.1464

Morgan, J., Daugherty, R., Hilchie, A., & Carey, B. (2003). Sample size and modeling accuracy of decision tree based on data mining tools. *Academy of Information and Management Sciences Journal, 6*(2), 77-91. Retrieved from: https://pdfs.semanticscholar.org/52af/82237a3da053481d1afc17c4c853b429e9bf.pdf

Mulak, P., & Talhar, N. (2013). Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset. *International Journal of Science and Research, 4*(7), 2101-2104. Retrieved from: https://www.semanticscholar.org/paper/Analysis-of-Distance-Measures-Using-K-Nearest-on-Mulak-Talhar/63f91b934f3dadec75c12a786f0e99d6df45ff37

Nagpal, A. (2017). *L1 and L2 Regularization Methods*. Retrieved from: https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

Nagy, B. G., Pollack, J. M., Rutherford, M. W., & Lohrke, F. T. (2012). The Influence of Entrepreneurs' Credentials and Impression Management Behaviors on Perceptions of New Venture Legitimacy. *Entrepreneurship Theory & Practice, 36*(5), 941-965. doi: 10.1111/j.1540-6520.2012.00539.x

Oates, T., & Jensen, D. (1997). The effects of training set size on decision tree complexity. *Proceedings of the 14th International Conference on Machine Learning.* Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.7365&rep=rep1&type=pdf

O'Neil, D. J. (2008). Nearest Neighbors Problem. In S. Shekhar, H. Xiong, & X. Zhou (Eds.), *Encyclopedia of GIS* (pp. 783-787). Cham, Switzerland: Springer.

Parkinson, B. (2011). Interpersonal Emotion Transfer: Contagion and Social Appraisal. *Social and Personality Psychology Compass, 5*(7), 428-439. doi: 10.1111/j.1751-9004.2011.00365.x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12,* 2825-2830. Retrieved from: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Rafaeli, A., & Sutton, R. I. (1987). Expression of emotion as part of the work role. *Academy of Management Review, 12*(1), 23-37. doi: 10.5465/amr.1987.4306444

Reback, J., Van den Bossche, J., MeeseeksMachine, Augspurger, T., & Mendel, B. (2019). *What's new in 0.25.1 (August 21, 2019).* Retrieved from: https://pandas.pydata.org/pandas-docs/version/0.25.1/whatsnew/v0.25.1.html#

Sari, B. G., Lúcio, A. D. C., Santana, C. S., Krysczun, D. K., Tischler, A. L., & Drebes, L. (2017). Sample size for estimation of the Pearson correlation coefficient in cherry tomato tests. *Ciência Rural, 47*(10), 1-6. doi: 10.1590/0103-8478cr20170116

Schmidt, K. L., & Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology, 116*(S33), 3-24. doi: 10.1002/ajpa.20001

Scikit-learn. (2007-2019). *sklearn.linear_model.LogisticRegression.* Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Scikit-learn. (2007-2019). *sklearn.neighbors.KNeighborsClassifier.* Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Scikit-learn. (2007-2019). *sklearn.svm.SVC.* Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Scott, G. G. (1990). *Resolving Conflict with Others and Within Yourself*. Oakland, CA: New Harbinger Publications Inc.

Shadchin, A., Haldane, A., Merry, B., Harris, C., Snyder, C., Allan, D., … Hoyer, S. (2019). *NumPy 1.16.5 Release Notes.* Retrieved from: https://numpy.org/devdocs/release/1.16.5-notes.html

Shmueli, B. (2019). *A Tale of Two Macro-F1's*. Retrieved from: https://towardsdatascience.com/a-tale-of-two-macro-f1s-8811ddcf8f04.

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling, 148*(1), 1-13. doi: 10.1016/S0304-3800(01)00388-X

Thébaud, S., & Sharkey, A. J. (2016). Unequal Hard Times: The Influence of the Great Recession on Gender Bias in Entrepreneurial Financing. *Sociological Science, 3*, 1-31. doi: 10.15195/v3.a1

Tian, Y-l., Kanade, T., Cohn, J. F. (2001). Recognizing Action Units for Facial Expressions Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(2), 97-115. doi: 10.1142/9789812778543_0002

Vallerand, R. J., Mageau, G. A., Ratelle, C., Leonard, M., Blanchard, C., Koestner, R., & Gagne, M. (2003). Les Passions de l'Ame: On obsessive and harmonious passion. *Journal of Personality and Social Psychology, 85*(4), 756-767. doi: 10.1037/0022-3514.85.4.756

Van Asch, V. (2013). Macro- and micro-averaged evaluation measures [[BASIC DRAFT]]. *Belgium: CLiPS, 49*, 1-27. Retrieved from: https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf

Van Baaren, R. B., Holland, R. W., Kawakami, K., & Van Knippenberg, A. (2004). Mimicry and Prosocial Behavior. *Psychological Science, 15*(1), 71-74. doi: 10.1111/j.0963-7214.2004.01501012.x

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*(91), 1-8. doi: 10.1186/1471-2105-7-91

Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics, 17*(60), 1-18. doi: 10.1186/s12859-016-0900-5

Wickham, H. (2017). *reshape2 v1.4.3*. Retrieved from: https://www.rdocumentation.org/packages/reshape2/versions/1.4.3

Wickham, H., Chang, W., Henry, L., Pederson, T. L., Takahashi, K., Wilke, C., … Dunnington, D. (2020). *ggplot2 3.3.0.* Retrieved from: https://ggplot2.tidyverse.org/index.html

Wickham, H., Hester, J., Francois, R., & RStudio. (2018). *readr 1.3.1.* Retrieved from: https://readr.tidyverse.org/index.html

Xia, F., Zhang, W., Li, F., & Yang, Y. (2008). Ranking with decision tree. *Knowledge and Information Systems, 17,* 381-395. doi: 10.1007/s10115-007-0118-y

Yüce, A., Gao, H. Cuendet, G. L., & Thiran, J-P. (2017). Action Units and Their Cross-Correlations for Prediction of Cognitive Load during Driving. *IEEE Transactions on Affective Computing, 8*(2), 161-175. doi: 10.1109/TAFFC.2016.2584042

Zebrowitz, L. A., & Montepare, J. M. (1992). Impressions of Babyfaced Individuals Across the Lifespan. *Developmental Psychology, 28*(6), 1143-1152. doi: 10.1037/0012-1649.28.6.1143

10. APPENDIX 1: CODE

All code used for this master’s thesis can be found here (https://github.com/caitlinvanmil/masterthesis2020).

11. Appendix 2: Correlation table facial AUs

Table 7 below shows the correlations between the different action units and the timestamp.

Table 7

Pearson correlations between the facial AUs & timestamp

	Inner brow raiser	Outer brow raiser	Brow lowerer	Upper lid raiser	Cheek raiser	Lid tightener	Nose wrinkler	Upper lip raiser	Lip corner puller	Dimpler	Lip corner depressor	Chin raiser	Lip stretcher	Lip tightener	Lips part	Jaw drop	Blink
Inner brow raiser	-																
Outer brow raiser	0.776***	-															
Brow lowerer	0.034***	0.073***	-														
Upper lid raiser	0.282***	0.240***	0.033***	-													
Cheek raiser	-0.056***	0.006***	0.084***	-0.019***	-												
Lid tightener	-0.004**	0.001	0.071***	-0.062***	0.402***	-											
Nose wrinkler	-0.000	0.021***	0.164***	0.080***	0.322***	0.230***	-										
Upper lip raiser	0.047***	0.065***	0.273***	0.093***	0.567***	0.190***	0.312***	-									
Lip corner puller	-0.029***	-0.018***	-0.071***	0.026***	0.842***	0.260***	0.208***	0.519***	-								
Dimpler	-0.016***	0.013***	0.134***	0.049***	0.593***	0.129***	0.224***	0.475***	0.635***	-							
Lip corner depresssor	0.080***	0.080***	0.080***	0.076***	0.014***	0.002	0.178***	0.155***	-0.052***	-0.016***	-						
Chin raiser	0.073***	0.083***	0.171***	0.120***	-0.033***	-0.001	0.062***	0.105***	0.073***	0.143***	0.403***	-					
Lip stretcher	0.047***	0.051***	0.092***	0.013***	0.225***	0.092***	0.159***	0.217***	0.139***	0.253***	0.252***	0.185***	-				
Lip tightener	0.047***	0.081***	0.077***	0.231***	0.173***	0.065***	0.193***	0.207***	0.148***	0.301***	0.285***	0.465***	0.207***	-			
Lips part	0.085***	0.106***	0.115***	0.037***	0.247***	0.173***	0.126***	0.328***	0.198***	0.128***	0.008***	-0.251***	0.263***	-0.123***	-		
Jaw drop	0.069***	0.072***	0.112***	0.095***	-0.074***	0.055***	-0.047***	0.049***	-0.081***	-0.022***	-0.027***	0.023***	0.109***	-0.015***	0.532***	-	
Blink	0.157***	0.112***	-0.081***	-0.009***	-0.050***	-0.034***	-0.021***	-0.054***	-0.024***	-0.035***	-0.031***	-0.112***	0.033***	-0.129***	0.082***	0.017***	-
Timestamp	-0.114***	-0.123***	-0.028***	-0.070***	-0.006***	-0.072***	-0.031***	-0.048***	-0.030***	0.043***	-0.002	0.026***	-0.040***	0.023***	-0.160***	-0.140***	-0.112***

*** Correlation is significant at the 0.001 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

12. APPENDIX 3: FEATURE SELECTION

Table 8 below shows which features were included in the five folds of the three different data sets.

Table 8

Feature selection

Dataset	Fold	Features included
Entire data set	1	kurtosis cheek raiser mean lid tightener mean lip corner puller mean dimpler mean chin raiser mean lips part maximum blink
	2	maximum, skewness and kurtosis inner brow raiser maximum brow lowerer skewness and kurtosis cheek raiser skewness and kurtosis lid tightener skewness and kurtosis nose wrinkler skewness and kurtosis lip corner puller maximum dimpler mean, maximum and standard deviation lip corner depressor skewness and kurtosis lip stretcher
	3	maximum, skewness and kurtosis inner brow raiser maximum brow lowerer maximum and standard deviation cheek raiser mean, maximum, standard deviation and kurtosis lid tightener skewness and kurtosis upper lip raiser maximum lip corner puller mean, maximum and standard deviation lip corner depressor mean and standard deviation lips part
	4	mean and standard deviation inner brow raiser mean and standard deviation outer brow raiser maximum and standard deviation brow lowerer mean and standard deviation upper lid raiser maximum and standard deviation cheek raiser mean, maximum and standard deviation lid tightener mean nose wrinkle skewness and kurtosis upper lip raiser mean, maximum and standard deviation lip corner puller mean and standard deviation dimpler mean chin raiser mean lip stretcher mean and standard deviation lips part

Table 8 Continued

Dataset	Fold	Features included
Entire data set	5	mean and standard deviation inner brow raiser mean and standard deviation outer brow raiser mean, maximum and standard deviation brow lowerer kurtosis upper lid raiser mean, maximum and standard deviation cheek raiser mean and standard deviation lid tightener mean, maximum and standard deviation lip corner puller mean and standard deviation dimpler mean and standard deviation lips part maximum blink
Pitch presentation data set	1	mean and kurtosis lid tightener skewness and kurtosis nose wrinkler mean lip corner puller mean and standard deviation dimpler mean chin raiser kurtosis jaw drop mean, maximum and standard deviation blink
	2	Maximum, skewness and kurtosis inner brow raiser skewness and kurtosis outer brow raiser skewness and kurtosis upper lid raiser maximum, skewness and kurtosis lid tightener skewness and kurtosis lip corner puller mean, maximum and standard deviation dimpler mean, maximum and standard deviation lip corner depressor skewness and kurtosis lip stretcher skewness and kurtosis jaw drop kurtosis blink
	3	maximum, skewness and kurtosis inner brow raiser maximum, skewness and kurtosis outer brow raiser skewness and kurtosis brow lowerer skewness and kurtosis upper lid raiser maximum and kurtosis cheek raiser maximum, skewness and kurtosis lid tightener skewness and kurtosis upper lip raiser maximum, skewness and kurtosis lip corner puller maximum and standard deviation dimpler mean, maximum and standard deviation lip corner depressor maximum, skewness and kurtosis lip stretcher maximum, skewness and kurtosis lips part kurtosis jaw drop
	4	mean, maximum and standard deviation inner brow raise mean, max and standard deviation outer brow raise maximum, skewness and kurtosis brow lowerer skewness and kurtosis upper lid raiser maximum cheek raiser maximum and kurtosis lid tightener skewness and kurtosis upper lip raiser maximum, skewness and kurtosis lip corner puller mean and standard deviation dimpler mean lip stretcher mean and standard deviation lips part

Table 8 Continued

Dataset	Fold	Features included
Pitch presentation data set	5	mean and standard deviation inner brow raiser
		mean, maximum and standard deviation outer brow raiser
		mean, maximum, standard deviation and kurtosis brow lowerer
		mean, maximum, standard deviation and kurtosis cheek raiser
		mean, maximum and kurtosis lid tightener
		skewness and kurtosis nose wrinkler
		kurtosis upper lip raiser
		mean, maximum and standard deviation lip corner puller
		mean dimpler
		maximum blink
Q&A session data set	1	mean and standard deviation inner brow raiser
		mean and standard deviation outer brow raiser
		maximum and standard deviation brow lowerer
		mean and standard deviation lid tightener
		mean lip corner puller
		mean dimpler
		mean lip corner depressor
		mean chin raiser
		mean and standard deviation lips part
		maximum blink
	2	mean, maximum and standard deviation inner brow raiser
		mean and standard deviation outer brow raiser
		maximum and standard deviation brow lowerer
		mean and standard deviation upper lid raiser
		skewness and kurtosis cheek raiser
		mean, skewness and kurtosis lid tightener
		standard deviation, skewness and kurtosis nose wrinkler
		skewness and kurtosis lip corner puller
		maximum dimpler
		mean and standard deviation lip corner depressor
		mean and standard deviation lips part
	3	mean, maximum and standard deviation inner brow raiser
		mean and standard deviation outer brow raiser
		maximum brow lowerer
		mean and standard deviation upper lid raiser
		maximum and standard deviation cheek raiser
		mean, maximum, standard deviation and kurtosis lid tightener
		mean and standard deviation nose wrinkler
		standard deviation lip corner puller
		standard deviation dimpler
		mean and standard deviation lip corner depressor
		mean chin raiser
		mean and standard deviation lip stretcher
		mean and standard deviation lip tightener
		mean and standard deviation lips part
		mean jaw drop

`Table 8 Continued

Dataset	Fold	Features included
Q&A session data set	4	mean and standard deviation inner brow raiser maximum and standard deviation brow lowerer mean, maximum and standard deviation upper lid raiser maximum and standard deviation cheek raiser mean, maximum and standard deviation lid tightener mean and standard deviation nose wrinkler skewness and kurtosis upper lip raiser mean, maximum and standard deviation lip corner puller mean and standard deviation dimpler mean lip corner depressor mean and standard deviation chin raiser mean lip stretcher mean and standard deviation lips part
	5	mean and standard deviation inner brow raiser kurtosis outer brow raiser mean, maximum and standard deviation brow lowerer mean, maximum and standard deviation cheek raiser mean and standard deviation lid tightener mean nose wrinkler mean, maximum and standard deviation lip corner puller mean and standard deviation dimpler mean lip corner depressor standard deviation chin raiser mean lip stretcher mean and standard deviation lips part skewness and kurtosis blink

13. Appendix 4: Optimal hyperparameters

Table 9 below sums up the hyperparameter settings applied to each outer loop corresponding to the five folds of the algorithms in the three different data sets.

Table 9

Optimal hyperparameter settings

Data set	Algorithm	Fold	Hyperparameter settings accuracy	Hyperparameter settings macro F1

Entire data set	KNN	1	n_neighbors = 1
			weights = uniform
				n_neighbors = 3
				weights = distance

		2	n_neighbors = 2
			weights = uniform
				n_neighbors = 2
				weights = uniform

		3	n_neighbors = 1
			weights = uniform
				n_neighbors = 3
				weights = uniform

		4	n_neighbors = 2
			weights = uniform
				n_neighbors = 2
				weights = uniform

		5	n_neighbors = 2
			weights = uniform
				n_neighbors = 2
				weights = uniform

	Multinomial logistic regression	1	penalty = l2
			C = 0.001
			solver = newton-cg
				penalty = l2
				C = 100
				solver = saga

		2	penalty = none
			C = 0.001
			solver = lbfgs
				penalty = l2
				C = 100
				solver = saga

		3	penalty = none
			C = 0.001
			solver = newton-cg
				penalty = none
				C = 0.001
				solver = newton-cg

		4	penalty = l2
			C = 0.001
			solver = newton-cg
				penalty = l2
				C = 0.001
				solver = newton-cg

		5	penalty = l2
			C = 1
			solver = newton-cg
				penalty = l2
				C = 1
				solver = newton-cg

	SVM	1	kernel = rbf
			gamma = scale
				kernel = rbf
				gamma = scale

		2	kernel = rbf
			gamma = scale
				kernel = rbf
				gamma = scale

		3	kernel = linear
			gamma = scale
				kernel = linear
				gamma = scale

		4	kernel = rbf
			gamma = scale
				kernel = rbf
				gamma = scale

		5	kernel = linear
			gamma = scale
				kernel = linear
				gamma = scale

Table 9 Continued

Data set	Algorithm	Fold	Hyperparameter settings accuracy	Hyperparameter settings macro F1

Entire data set	Decision tree	1	criterion = gini
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 4
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 5
				min_samples_split = 3

		2	criterion = gini
			max_depth = 12
			max_leaf_nodes = 6
			min_samples_split = 2
				criterion = gini
				max_depth = 12
				max_leaf_nodes = 6
				min_samples_split = 2

		3	criterion = gini
			max_depth = 4
			max_leaf_nodes = 3
			min_samples_split = 2
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 3
				min_samples_split = 2

		4	criterion = gini
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 3
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 4
				min_samples_split = 3

		5	criterion = entropy
			max_depth = 4
			max_leaf_nodes = 5
			min_samples_split = 4
				criterion = entropy
				max_depth = 4
				max_leaf_nodes = 6
				min_samples_split = 3

	Random forest	1	criterion = gini
			n_estimators = 2
			max_depth = 8
			max_leaf_nodes = 8
			min_samples_split = 2
				criterion = gini
				n_estimators = 2
				max_depth = 8
				max_leaf_nodes = 8
				min_samples_split = 2

		2	criterion = entropy
			n_estimators = 2
			max_depth = 8
			max_leaf_nodes = 3
			min_samples_split = 2
				criterion = entropy
				n_estimators = 2
				max_depth = 8
				max_leaf_nodes = 3
				min_samples_split = 2

		3	criterion = gini
			n_estimators = 2
			max_depth = 8
			max_leaf_nodes = 5
			min_samples_split = 3
				criterion = gini
				n_estimators = 2
				max_depth = 8
				max_leaf_nodes = 5
				min_samples_split = 3

		4	criterion = entropy
			n_estimators = 2
			max_depth = 6
			max_leaf_nodes = 8
			min_samples_split = 3
				criterion = entropy
				n_estimators = 2
				max_depth = 4
				max_leaf_nodes = 9
				min_samples_split = 2

		5	criterion = gini
			n_estimators = 4
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 2
				criterion = gini
				n_estimators = 4
				max_depth = 4
				max_leaf_nodes = 4
				min_samples_split = 2

Table 9 Continued

Data set	Algorithm	Fold	Hyperparameter settings accuracy	Hyperparameter settings macro F1

Pitch presentation data set	KNN	1	n_neighbors = 3
			weights = uniform
				n_neighbors = 3
				weights = uniform

		2	n_neighbors = 1
			weights = uniform
				n_neighbors = 1
				weights = uniform

		3	n_neighbors = 1
			weights = uniform
				n_neighbors = 1
				weights = uniform

		4	n_neighbors = 3
			weights = uniform
				n_neighbors = 3
				weights = uniform

		5	n_neighbors = 1
			weights = uniform
				n_neighbors = 1
				weights = uniform

	Multinomial logistic regression	1	penalty = l2
			C = 1
			solver = newton-cg
				penalty = l2
				C = 1
				solver = newton-cg

		2	penalty = l2
			C = 1
			solver = newton-cg
				penalty = l2
				C = 1
				solver = newton-cg

		3	penalty = l2
			C = 0.1
			solver = newton-cg
				penalty = l2
				C = 0.1
				solver = newton-cg

		4	penalty = l2
			C = 0.001
			solver = newton-cg
				penalty = l2
				C = 0.001
				solver = newton-cg

		5	penalty = l2
			C = 1
			solver = newton-cg
				penalty = l2
				C = 1
				solver = newton-cg

	SVM	1	kernel = linear
			gamma = scale
				kernel = linear
				gamma = scale

		2	kernel = linear
			gamma = scale
				kernel = linear
				gamma = scale

		3	kernel = linear
			gamma = scale
				kernel = linear
				gamma = scale

		4	kernel = sigmoid
			gamma = scale
				kernel = sigmoid
				gamma = scale

		5	kernel = linear
			gamma = scale
				kernel = linear
				gamma = scale

Table 9 Continued

Data set	Algorithm	Fold	Hyperparameter settings accuracy	Hyperparameter settings macro F1

Pitch presentation data set	Decision tree	1	criterion = entropy
			max_depth = 4
			max_leaf_nodes = 5
			min_samples_split = 4
				criterion = entropy
				max_depth = 4
				max_leaf_nodes = 5
				min_samples_split = 4

		2	criterion = gini
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 4
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 5
				min_samples_split = 3

		3	criterion = gini
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 3
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 4
				min_samples_split = 3

		4	criterion = entropy
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 2
				criterion = entropy
				max_depth = 4
				max_leaf_nodes = 4
				min_samples_split = 2

		5	criterion = gini
			max_depth = 4
			max_leaf_nodes = 6
			min_samples_split = 2
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 6
				min_samples_split = 2

	Random forest	1	criterion = gini
			n_estimators = 2
			max_depth = 12
			max_leaf_nodes = 7
			min_samples_split = 2
				criterion = gini
				n_estimators = 2
				max_depth = 12
				max_leaf_nodes = 7
				min_samples_split = 2

		2	criterion = gini
			n_estimators = 2
			max_depth = 12
			max_leaf_nodes = 3
			min_samples_split = 4
				criterion = gini
				n_estimators = 2
				max_depth = 12
				max_leaf_nodes = 3
				min_samples_split = 4

		3	criterion = gini
			n_estimators = 2
			max_depth = 4
			max_leaf_nodes = 5
			min_samples_split = 3
				criterion = gini
				n_estimators = 2
				max_depth = 4
				max_leaf_nodes = 5
				min_samples_split = 3

		4	criterion = entropy
			n_estimators = 2
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 2
				criterion = entropy
				n_estimators = 6
				max_depth = 4
				max_leaf_nodes = 7
				min_samples_split = 3

		5	criterion = gini
			n_estimators = 10
			max_depth = 8
			max_leaf_nodes = 8
			min_samples_split = 4
				criterion = gini
				n_estimators = 2
				max_depth = 4
				max_leaf_nodes = 6
				min_samples_split = 3

Table 9 Continued

Data set	Algorithm	Fold	Hyperparameter settings accuracy	Hyperparameter settings macro F1

Q&A session data set	KNN	1	n_neighbors = 1
			weights = uniform
				n_neighbors = 1
				weights = uniform

		2	n_neighbors = 1
			weights = uniform
				n_neighbors = 1
				weights = uniform

		3	n_neighbors = 3
			weights = uniform
				n_neighbors = 3
				weights = uniform

		4	n_neighbors = 1
			weights = uniform
				n_neighbors = 1
				weights = uniform

		5	n_neighbors = 2
			weights = uniform
				n_neighbors = 2
				weights = uniform

	Multinomial logistic regression	1	penalty = l2
			C = 0.001
			solver = newton-cg
				penalty = l2
				C = 0.1
				solver = newton-cg

		2	penalty = l2
			C = 0.1
			solver = newton-cg
				penalty = l2
				C = 0.1
				solver = newton-cg

		3	penalty = l2
			C = 0.1
			solver = newton-cg
				penalty = l2
				C = 0.1
				solver = newton-cg

		4	penalty = l2
			C = 0.001
			solver = newton-cg
				penalty = l2
				C = 10
				solver = newton-cg

		5	penalty = l2
			C = 0.1
			solver = newton-cg
				penalty = l2
				C = 0.1
				solver = newton-cg

	SVM	1	kernel = rbf
			gamma = scale
				kernel = linear
				gamma = scale

		2	kernel = sigmoid
			gamma = scale
				kernel = sigmoid
				gamma = scale

		3	kernel = poly
			gamma = scale
				kernel = poly
				gamma = scale

		4	kernel = poly
			gamma = scale
				kernel = poly
				gamma = scale

		5	kernel = linear
			gamma = scale
				kernel = linear
				gamma = scale

Table 9 Continued

Data set	Algorithm	Fold	Hyperparameter settings accuracy	Hyperparameter settings macro F1

Q&A session data set	Decision tree	1	criterion = gini
			max_depth = 4
			max_leaf_nodes = 3
			min_samples_split = 2
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 7
				min_samples_split = 3

		2	criterion = gini
			max_depth = 4
			max_leaf_nodes = 4
			min_samples_split = 2
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 4
				min_samples_split = 2

		3	criterion = gini
			max_depth = 4
			max_leaf_nodes = 3
			min_samples_split = 2
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 3
				min_samples_split = 2

		4	criterion = gini
			max_depth = 8
			max_leaf_nodes = 5
			min_samples_split = 3
				criterion = gini
				max_depth = 8
				max_leaf_nodes = 5
				min_samples_split = 3

		5	criterion = gini
			max_depth = 4
			max_leaf_nodes = 5
			min_samples_split = 2
				criterion = gini
				max_depth = 4
				max_leaf_nodes = 5
				min_samples_split = 2

	Random forest	1	criterion = gini
			n_estimators = 2
			max_depth = 4
			max_leaf_nodes = 8
			min_samples_split = 4
				criterion = gini
				n_estimators = 2
				max_depth = 4
				max_leaf_nodes = 8
				min_samples_split = 4

		2	criterion = gini
			n_estimators = 2
			max_depth = 4
			max_leaf_nodes = 3
			min_samples_split = 3
				criterion = gini
				n_estimators = 2
				max_depth = 4
				max_leaf_nodes = 3
				min_samples_split = 3

		3	criterion = entropy
			n_estimators = 2
			max_depth = 6
			max_leaf_nodes = 4
			min_samples_split = 2
				criterion = entropy
				n_estimators = 2
				max_depth = 6
				max_leaf_nodes = 4
				min_samples_split = 2

		4	criterion = gini
			n_estimators = 2
			max_depth = 6
			max_leaf_nodes = 7
			min_samples_split = 3
				criterion = gini
				n_estimators = 2
				max_depth = 6
				max_leaf_nodes = 7
				min_samples_split = 3

		5	criterion = gini
			n_estimators = 6
			max_depth = 12
			max_leaf_nodes = 5
			min_samples_split = 3
				criterion = gini
				n_estimators = 6
				max_depth = 12
				max_leaf_nodes = 5
				min_samples_split = 3

14. APPENDIX 5: FEATURE IMPORTANCE

Table 10 below shows the features' importance in each fold of the model that performed above the baseline, namely the decision tree based on the pitch presentation data.

Table 10

Feature importance

Fold	Feature	Feature importance
1	Mean blink	0.41
	Mean lid tightener	0.40
	Kurt lid tightener	0.19
2	Max lid tightener	0.44
	Kurt inner brow raiser	0.37
	Kurt upper lid raiser	0.19
3	Max dimpler	0.50
	Kurt upper lid raiser	0.30
	Kurt upper lip raiser	0.20
4	Max lid tightener	0.48
	Mean outer brow raiser	0.29
	Kurt lid tightener	0.24
5	Max brow lowerer	0.31
	Mean lid tightener	0.27
	Mean outer brow raiser	0.26
	Kurt cheek raiser	0.16