Using Facial Expressions to Predict Process Mining Task Performance

Lital Shalev[1], Irit Hadar[1][0000-0002-4267-0235], Rotem Dror[1][0000-0002-9433-8410], Adir Solomon[1][0000-0001-7955-1048], Elizaveta Sorokina[1], Michal Weisman Raymond[2], Pnina Soffer[1][0000-0003-4659-883X]

[1] University of Haifa, Haifa, Israel
[2] Ben Gurion University of the Negev Beer Sheva, Israel
hadari@is.haifa.ac.il

Abstract. Process mining analysis is a complex task that presents significant challenges to human analysts. To aid along this process, it is essential to identify difficulties as they occur. This study takes an initial step in this direction, by predicting the quality of task performance based on analysts' facial expressions while they are engaged in a process mining task. Data were collected using participants' webcams and the iMotions™ cloud application while they performed a process mining task. The data were then utilized to train and evaluate several machine learning classifiers, which classified participants based on the grade given to their task outcome. Our results show the high performance of these classifiers in predicting participants' success based on facial expressions. We further showed that the chosen outcome classifier could accurately classify additional participants, demonstrating its generalizability. Notably, the classifier was able to predict participants' success within a very short time frame. These findings could pave the way for developing a near-real-time support system to detect when analysts engaged in process mining may benefit from assistance.

Keywords: Process of Process Mining, Machine Learning, Facial Expressions.

1 Introduction

It is commonly recognized that process mining (PM) analysts face significant challenges during their work. For example, handling different types of process variations [1] handling multiple perspectives of the same process [2] and coping with missing information in the event logs [3], to name a few. Zimmermann et al. identified as many as 23 such challenges [2].

Even though process mining analysts face these complex difficulties, research efforts in the area of process mining have so far focused primarily on the development of algorithms and approaches for specific process mining tasks, addressing each one separately from a technical perspective [4], [5]. Less attention has been given to supporting PM practitioners along the entire process of process mining (PPM). Wongsuphasawat et al. further indicate a need for analysis guidance, e.g., by augmenting tools with ready-to-use recommendations and templates [6]. To improve the understandability and usability of PM tools for providing better support for process miners, we first need to gain an in-

depth understanding of the cognitive processes underlying the PPM. Several recent works have taken steps in this direction, showing that when analysts explore event logs, they follow different behavioral patterns and strategies to gain insights from the data, and that these may help predicting their chances of succeeding in the PM task [7], [8].

In this study we are setting a first steppingstone towards specialized online support for process mining analysts. To assist analysts, a first step is to detect when they are facing obstacles and could benefit from support. In other words, we need to predict when they are on a path that might set them up to failure. Therefore, our research question for this study was: **How can multimodal data be used for predicting process mining task success?** While attempting to make this prediction as early in the PPM as possible.

By using facial expressions data, collected from process analysts in a simple remote setting, using only the participant's webcam, we envision that a similar near-real-time analysis could be integrated into a supporting system to detect when the process miner could benefit from receiving assistance that would lead them to a more promising path.

We report on the outcomes of this study as well as share our unique dataset of mean facial expressions intensity values, collected as part of this research, with the research community to encourage other research groups in the field of process mining to use multimodal data, such as facial expressions, in their studies.

The rest of the paper is organized as follows. Section 2 presents a summary of related work. Section 3 details the research method and Section 4 its findings. We discuss the findings in Section 5 and conclude in Section 6.

2 Related Work

2.1 The Process of Process Mining (PPM)

Studies into the individual process of process mining (PPM) have just recently started to emerge [7], [8]. To the best of our knowledge, this research is a first attempt of applying machine learning (ML) for predicting the outcome of this process. The most closely related work we found is in the field of process modeling, where a classifier was trained to predict whether the modeler is a novice or an expert process modeler based on layout features of the model under development [9].

A different relevant line of work, related to the analysis and prediction of participants' engagement and emotions based on facial expressions, was conducted in classroom [10] and in online settings with 10 seconds clips [11]. Similarly, research focusing on different ML approaches aimed to predict participants' performance based on static predictive variables [12]–[15]. Research focusing on learning and emotions [16]–[18] may also inform our work, since we hypothesize that the PPM entails learning activities, i.e., learning about the process. Positive emotions affect learning by increasing students' attention and motivation [17]. Surprise, in particular, has been explored as a phenomenon that affects learning in both child development and education settings [19], [20]. Surprise has an effect on learning by requiring the individual to explain unexpected outcomes; the more unexpected the outcome is and requires more explanation, the more memorable the learning will be [16].

As PM is characterized as a knowledge-intensive and unstructured process that often yields unpredictable outcomes, it may entail varying levels of cognitive load. Cognitive load has been vastly investigated in relation to biometric sensors and multimodal data which are similar to some extent to the tools used in this study. Two ML predictive models have been trained to detect cognitive load during driving assignment based on

eye measurements only [21]. In an e-learning setup, a cognitive load predictive models were developed based on eye-movements, heart rate and skin-based measures [22]. The combination of blink rate and galvanic skin response (GSR) measures were also shown to be highly distinctive features in different ML classifiers [23].

2.2 Prediction Error Minimization for the Process of Process Mining (PEM4PPM)

Results analysis in this paper will be considered in light of the PEM4PPM model, since it provides conceptualization of the cognitive process that process miners might follow. The PEM4PPM model is an adaptation of Prediction Error Minimization (PEM) to the PPM [7]. PEM is a principal within the cognitive theory of Predictive Processing (PP), viewing the brain as a sophisticated machine that attempts to predict what it will sense and to create a model of what might be causing the sensory inputs it receives. According to the PEM principal, the brain then aims to minimize the difference between its predictions and the real input as much as possible. If its predictions are already quite accurate, it might not need to change its models; if they are not, it will continue adjusting the prediction models until they are accurate enough [24].

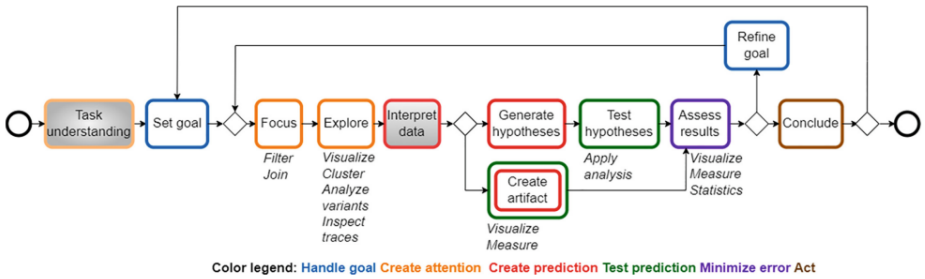

Fig. 1. The PEM4PPM model.

Fig. 1 illustrates the PEM4PPM model, highlighting the sequence of steps and their corresponding cognitive operations. The PPM begins with high-level business goals, which serve as the starting point for any process mining endeavor. These goals can be decomposed or refined into more specific objectives as needed. The refinement process continues until the goals are concrete enough to be achieved through available mining operations. For example, a high-level goal can be to find how cycle times can be reduced. This goal can be refined to detection of bottlenecks in the process. To focus attention on relevant aspects of the input data, a relevant subset of the data is filtered and organized. This step enables subsequent exploration of the data to identify behavioral patterns that are of interest considering the identified goals. Based on the exploration results, concrete hypotheses are formed as predictions to be tested. For example, hypotheses can be formed about specific activities which may act as a bottleneck in the process. Predictions are tested through the creation of specific artifacts, such as discovered process models. Available PM techniques are applied to validate the hypotheses or create artifacts that support the predictions. The obtained results are assessed against the original goal or hypothesis to evaluate prediction errors and take actions for their minimization. This

assessment serves as a basis for determining whether the goal has been achieved or if further refinement is needed [7]. This process is iterative in nature, and may involve additional filtering, focusing, and exploration to form new hypotheses and to test them.

Four process mining cognitive strategies have been identified and validated based on the PEM4PPM model, and their effect on the PM task performance was analyzed. The strategies are: (1) NNN - No data interpretation - No indicated hypothesis - No testing, where the conclusions of the participants were based on the data exploration stage only; (2) WNN - With data interpretation - No indicated hypothesis - No testing, where participants based their conclusions mostly on data exploration and interpretation; (3) WWN - With data interpretation - With hypothesis - No testing, where participants formulated hypotheses but did not follow a trial-and-error approach; (4) WWW - With data interpretation - With hypothesis - With testing, where participants performed all the activities of the PEM4PPM model. Analysts who followed this full WWW strategy demonstrated significantly better performance than analysts who applied other strategies [7].

In relation to learning, we posit that the PEM4PPM steps of Task Understanding, Explore, Interpret Data and Assess Results, require some extent of learning the mined process. We further speculate that the phases of Task Understanding, Focus, Generate Hypothesis, Set Goal and Refine Goal would require high cognitive load. We will analyze our findings in light of these assumptions.

3 Methods

3.1 Study Settings

The data for this research were collected from 16 B.Sc. and M.Sc. students in the Department of Information Systems at the University of Haifa, taking an advanced course in Process Mining. As facial expression recognition may involve bias in terms of gender and ethnicity [25] we note that all participants were Caucasian / Middle eastern, 13 females and 3 males. For the participation in the study, students received bonus points to their final course grade. Students who chose not to participate in the study were offered an alternative non-experimental assignemnt with the same bonus points. This study setup was approved by the IRB.

The students who chose to participate in the study were presented with the following question about the Road Traffic Fine Management (RTFM) event log, “Based on the data in the log, if the offender wants to pay as little as possible, how should they act? Suggest at least two alternative actions, show why they are fulfilling the requirement of paying as little as possible, and compare between them.” Students were asked to use the Disco (Fluxicon Disco™) application and think-aloud, namely, verbally describe their thinking process when performing the task.

3.2 Data Collection

The data were collected with both iMotions™ cloud platform and Zoom application. The iMotions™ cloud platform was used to collect eye tracking data, screen recordings, and participants’ face recordings. The Zoom application was used to record the think-aloud data. The data collection was done remotely, with the participants using their own

computers and sharing their screens and web camera through the iMotions™ cloud platform. Figure 2 shows an example of a processed video exported from iMotions™. The video includes voice recording, participant's face recording (anonymized in the figure), screen recording with gaze path, and two graphs of selected facial expressions (further explained below).

Facial Expressions. The participants' face recordings were analyzed using AFFDEX 3.0 SDK [26], [27], a toolkit for analyzing facial expressions in real life setups. AFFDEX was used through its integration in iMotions[1]. Every 30 milliseconds the state of the participant was recorded by iMotions and analyzed by AFFDEX, which provides detections of the following emotional expressions: Anger, Contempt, Disgust, Fear, Joy, Sadness, Surprise, Engagement, Valence, Sentimentality, and Confusion. The detection of facial Action Units is translated to facial expressions such as Chick Raise, Blink Rate, Smile, and Smirk. For each facial expression, an intensity score is provided, between 0 and 100. The higher the score, the higher the likelihood the participant is presenting that emotional expression. From iMotions™ it is possible to export all the facial expression data into a CSV file for further analysis. Fig. 2 shows a screenshot taken from iMotions™, where the main part of the screen shows the participant's screen recording while using Disco. The orange circle on the Disco screen recording represents the location the participant's eyes were fixated on. The bottom part of the figures shows a graph of the AFFDEX facial expressions analyzes for Confusion and Contempt.

Think Aloud. The Zoom voice recordings of the participants describing their thought process were combined with the screen recordings from iMotions™ to generate a single video and audio recording for each participant. This combined video was then used to determine the grade for the task performance. Grades were on the scale of 0-100, considering both the provided answer, and the evidence that supported the answer. E.g. a student who provided the expected answer (the offender should wait for 90 days or pay the fine immediately), but did not analyze the data in Disco to support this answer, received a grade of 50. The recordings were viewed carefully as part of the participants' task performance evaluation in order to decide on the appropriate grade.

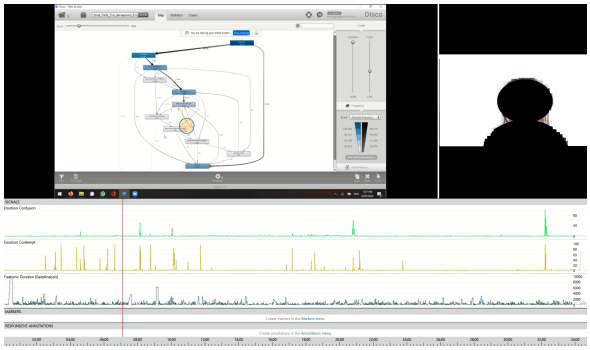

Fig. 2. Snapshot from a processed video exported from iMotions™.

[1] https://imotions.com/

3.3 Data Preprocessing

We preprocessed the collected data in the following manner. First, the facial expressions data were exported from iMotions™ for each participant. Second, we set two grade groups, High Performers, participants whose grades are above 55, and Low Performers, participants whose grades are below 55. This threshold was set since grades below 55 indicate poor performance in the provided task, and since it marked a clear break between two balanced groups of 8 participants (see grades distribution in Fig. 3).

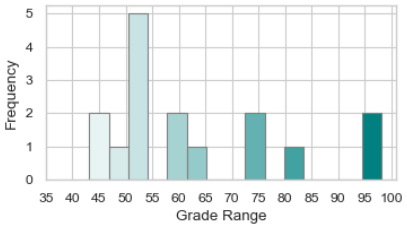

Fig. 3. Grades Distribution.

We then used a Wilcoxon statistical test for the hypothesis that there is a significant difference between the High Performers and Low Performers groups in average facial expression values. The data we used for the test were the first 10,000 samples from each participant's data, representing the first 5 minutes of the session. To prepare the data for the ML classifiers training, the mean value of each facial expression intensity was calculated for each participant for all the valid data points in the session.

3.4 ML Model Training

Multiple ML classifiers, including Decision Tree Classifier, Random Forest Classifier, and Naïve Bayes Classifier, were trained to predict if the participant belonged to the High or to the Low Performer groups. We chose to train a Naïve Bayes Classifier due to its ability to handle small data sets. Further, the Decision Tree Classifier was selected for its proficiency in handling non-linear relationships effectively, and its inherent ability to accommodate correlations among features, making it particularly suited for complex but small datasets. Lastly, the Random Forest Classifier was chosen due to its ensemble approach, which integrates multiple decision trees, and mitigates overfitting associated with single decision models to provide a reliable assessment of feature relevance. We used Python Sklearn package implementations of these models. Since our dataset was relatively small, with mean values for 16 participants for each facial expression, we chose to use Leave One Out Cross Validation (LOOCV) approach for the evaluation of the classifiers' performance. We then calculated the mean accuracy of each iteration, where in each, 1 participant was used as a test set and the rest of the 15 served as a train set. This approach guarantees that each participant will be in the test set only once. This approach is appropriate for small datasets such as the one we had.

3.5 Independent Temporal Data Evaluation Setting

For the temporal evaluation of the trained Random Forest classifier, we used additional data collected from 4 participants, as part of a previous data collection study, which

used the same event log and asked the participants similar questions [7]. The independent data used for the evaluation were collected with Zoom application only. The Zoom video recordings were concatenated into a new video file which included only the participants' face recording. The new video file was then imported into iMotions™ and post-processed using the AFFDEX 3.0 toolkit to provide the same facial expressions data format used to train the ML classifiers. The data were then aggregated in the same way as the training data; the mean value was calculated for each relevant facial expression throughout the whole session. The trained classifier was then used to predict the grade group of the participant in different time points during the process mining task.

4 Findings

Three ML classifiers for predicting participants' task performance were trained during this study. Their accuracy was evaluated by both using the LOOCV method and additional real-world evaluation data. Fig. 4 presents the mean accuracy results for the three ML classifiers with the LOOCV evaluation method. Fig. 5 presents the F1-score for each classifier. The classifier that had the best performance in both the mean accuracy and F1-score metrics was the Random Forest Classifier with mean accuracy of 87.5% and F1-score micro of 88% for both classes. The Decision Tree classifier and the Naïve Bayes classifier had the same mean accuracy results of 62.5%.

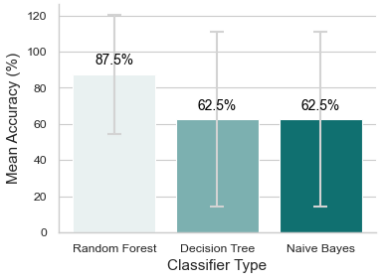

Fig. 4. Classifiers Comparison Mean Accuracy.

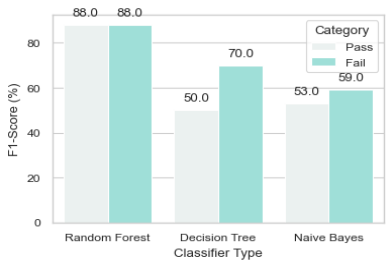

Fig. 5. Classifiers Comparison F1-score micro.
Legend: Pass = High Performers group, Fail = Low performers group

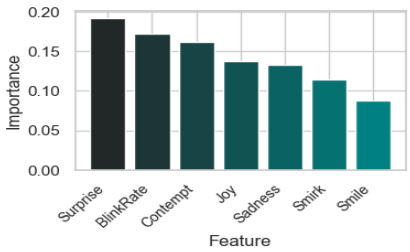

Fig. 6. Feature Importance of the Random Forest Classifier.

Fig. 6 presents the features used for the Random Forest classifier and their importance. The highest importance ranks were for Surprise, Blink Rate, and Contempt, and then Joy, Sadness, Smirk, and Smile. The parameters used for the Random Forest were minimal samples split: 5, number of estimators: 200, and balanced class weight.

To explore the generalization capabilities and temporal responsiveness of the Random Forest classifier, independent data evaluation was performed for different time points during the process. Data of 4 participants from a previous study were used [7]. The focus of the previous study was not the analysis of facial expressions, hence video had to be processed for facial expressions. As a result, most of the data were partial, since the participants' faces were not clearly visible part of the time. Results are presented in Table 1. For 3 out of 4 of the participants, the classifier was able to detect correctly the class of the participant (High Performers/Low Performers) after the first minute. After 7.5 minutes, the classifier had accuracy of 100% for the 4 participants. For the participant with ID 2, the classification was correct for all time points except for the 5 minutes window. We assume that is due to the low visibility of the participant's face which resulted in lower quality of the facial expressions analysis.

Table 1. Temporal Independent Data Evaluation

ID	*1 Min.*	*2.5 Min.*	*5 Min.*	*7.5 Min.*	*10 Min.*	*15 Min.*	*Full Session*
1							
2							
3							
4							
Accur.	75%	75%	75%	100%	100%	100%	100%

Fig. 7 shows the significant differences (p-value < 0.001 in a Wilcoxon statistical test) in the facial expression mean values between the High Performers and Low Performers groups during the first 5 minutes of the process mining task. Positive values in Fig.7 show that the High Performers group had a higher mean value, and negative values indicate that the Low Performers group had higher values. It is evident that Blink Rate, Smile, and Joy had higher mean values for the High Performers group, where Contempt, Smirk and Surprise had higher mean values for the Low Performers group.

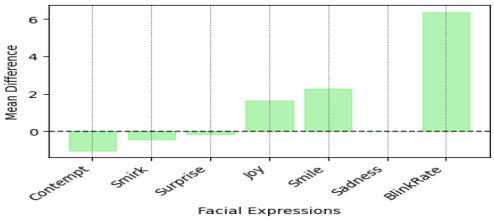

Fig. 7. Significant Differences Between the High Performers and Low Performers Groups.

5 Discussion

The presented findings show that it is possible to predict if a participant will be successful in a process mining task, solely based on facial expressions. We show that this

ability remains also when the dataset is partial and generated from a regular Zoom recording with no special setup. Furthermore, the classification had 75% accuracy after only 1 minute of partial data, and 100% accuracy after 7.5 minutes of partial data.

We hypothesize that it is possible to predict the success in a process mining task early in the process, due to the cognitive strategies the High Performers and Low Performers follow. Participants whose strategy included generating and testing hypotheses had higher rates of success [7]. We believe there are facial cues presented during the Generate Hypothesis step of the PEM4PPM model, and that they occur at an early stage of the session for most of the participants. For other participants it might take a few minutes longer to generate a hypothesis, but if it takes longer than that, they might not generate hypotheses at any time during the analysis. If this is indeed the case, it might be an explanation for the finding that after 7.5 minutes, the classifier was able to predict the outcome correctly for the 4 independent participants. Future research combining the settings of the two studies will enable us to test our hypothesis.

Another possible explanation could be that other PEM4PPM steps could also be reliable predictors of the process mining task outcome. These steps might be Task Understanding, Focus, Set Goal and Refine Goal which we anticipate will require higher cognitive load. Blink Rate has been indicated as a measure of cognitive load [28], [29], and, as shown in Fig. 7, it has the highest difference in the mean values between the two grade groups. The High Performers group had significantly higher mean values than the Lower performers group, which might indicate a higher cognitive load for the High Performers. Further investigation is required to validate this explanation.

The feature importance of the trained classifier resonates well with the conceptual arguments we presented, as Surprise was the feature of the highest importance. Surprise has previously been reported as being related to learning processes [16]. Since we view the PPM as partially a learning process, we expected Surprise to be one of the high importance features for a trained outcome classifier.

The additional selected features are Contempt, Joy, Sadness, Smirk and Smile. We note that the differences in the mean values showed that the Low Performers had higher negative facial expressions (Contempt and Smirk), and High Performers had higher values of positive facial expressions (Joy and Smile). It has been established that positive emotions improve learning, while negative emotions reduce it [17]. We therefore speculate that the Low Performers had less phases of learning than the High Performers.

We suggest that the random forest model outperformed other classification algorithms in predicting the success or failure of participants completing the task, because it leverages various combinations and interactions among features — in our case, the different facial expressions displayed by the participants — resulting in a more comprehensive and accurate model. As the random forest constructs an ensemble of decision trees, capturing the complex patterns and nuances in facial expressions that a single decision tree might miss. Based on these results, we recommend that researchers intending to explore the use of facial expressions in PM tasks use ensemble classifiers, as we believe that the relationships between different expressions can explain various phenomena in unique ways, making ensembles more adequate for this kind of task in this domain. We are publicly sharing the dataset and implementation of the classifiers

presented in this paper, to facilitate the use of multimodal data and ML models in the research community of process mining[2].

Threats to Validity. Several threats to the validity of the study need to be considered. First, to date, there is no agreement in the literature whether facial expressions are indications of emotions or not. This could be relevant to the conclusions we draw from the results, however, it does not present a threat to the validity of our classifier, since we record facial expressions and predict based directly on them. Our interpretations of the meaning of the facial expressions are what could be compromised. Second, the datasets we used in this study are relatively small. As this is a clear limitation of our study, it is also one of its strengths, since we were able to show that it is possible to gain reliable results and train an accurate ML classifier also with small datasets. In future research, a larger dataset would also enable to refine the binary classification of high and low performers into finer-grained performance groups. Third, the data collected in this study relate to a single PM tool and a single PM task. while this adequately addresses our research questions regarding the prediction of a PM task outcome, further studies are necessary to establish broader conclusions. In particular, different tasks (e.g., process discovery, conformance checking) should be investigated.

6 Conclusion

This research has several novel methods and findings. Firstly, we show that collecting facial expressions data in a simple remote setup, using only a webcam, can provide valuable insights for better understanding the PPM, and potentially other related tasks, e.g. process modeling and data mining. Our findings show that the intensity scores of Surprise and Blink Rate are associated with PM task performance, and that based on this association, task performance can be predicted early in the process. We hypothesize this is due to differences in cognitive load and learning-related phases during the PPM. We intend to further investigate this relationship in future research.

Secondly, we present that it is possible to train ML classifiers for process mining task performance based on a small dataset, containing data of only 16 participants. In addition, we present that it is possible to make relatively accurate predictions (75%) of students' task performance by using only 1 minute of partial facial expressions data gathered from a simple Zoom recording that was post-processed by using iMotions™. After 7.5 minutes of data, our classifiers accuracy increased to 100%. We plan to increase our dataset to further validate these results as well as to generalize our chosen classifier for similar tasks, such as business process modeling.

Lastly, our current classifier uses only facial expressions intensity data. We view this as an advantage for remote setups and the usage of it in a future online support system for process miners. However, we believe it would be interesting and possibly beneficial to enhance this classifier with additionally collected data, such as eye tracking measurements, therefore our future research plans include that as well. We believe that once the presented classifier is appropriately generalized, it would be possible to use it for

[2] https://github.com/litalshl/Facial-Expressions-Classifiers

near-real-time predictions as part of a specialized support system, thereby identifying when an analyst requires support to avoid their predicted failure.

Acknowledgement
The research was funded by the Israel Science Foundation, grant no. 2005/21.

References

[1] W. M. P. van der Aalst and A. J. M. M. Weijters, "Process mining: a research agenda," *Comput. Ind.*, vol. 53, no. 3, pp. 231–244, Apr. 2004.
[2] L. Zimmermann, F. Zerbato, and B. Weber, "Process Mining Challenges Perceived by Analysts: An Interview Study," in *Enterprise, Business-Process and Information Systems Modeling*, 2022, pp. 3–17.
[3] R. P. J. C. Bose, R. S. Mans, and W. M. P. van der Aalst, "Wanna improve process mining results?," in *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013, pp. 127–134.
[4] A. Augusto *et al.*, "Automated Discovery of Process Models from Event Logs: Review and Benchmark," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 686–705, Apr. 2019.
[5] V. Pasquadibisceglie, A. Appice, G. Castellano, and W. van der Aalst, "PROMISE: Coupling predictive process mining to process discovery," *Inf. Sci.* , vol. 606, pp. 250–271, Aug. 2022.
[6] K. Wongsuphasawat, Y. Liu, and J. Heer, "Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study," *arXiv [cs.HC]*, 01-Nov-2019.
[7] E. Sorokina, P. Soffer, I. Hadar, U. Leron, F. Zerbato, and B. Weber, "PEM4PPM: A Cognitive Perspective on the Process of Process Mining," in *Business Process Management*, 2023, pp. 465–481.
[8] F. Zerbato, P. Soffer, and B. Weber, "Process Mining Practices: Evidence from Interviews," in *Business Process Management*, 2022, pp. 268–285.
[9] A. Burattin *et al.*, "Who Is Behind the Model? Classifying Modelers Based on Pragmatic Model Features," in *Business Process Management*, 2018, pp. 322–338.
[10] G. Tonguç and B. Ozaydın Ozkara, "Automatic recognition of student emotions from facial expressions during a lecture," *Comput. Educ.*, vol. 148, p. 103797, Apr. 2020.
[11] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, Glasgow, UK, 2017, pp. 33–40.
[12] C. G. Nespereira, E. Elhariri, N. El-Bendary, A. F. Vilas, and R. P. D. Redondo, "Machine Learning Based Classification Approach for Predicting Students Performance in Blended Learning," in *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt*, 2016, pp. 47–56.
[13] A. Joshi, P. Saggar, R. Jain, M. Sharma, D. Gupta, and A. Khanna, "CatBoost — an ensemble Machine Learning model for prediction and classification of student academic performance," *Adv. Data Sci. Adapt. Anal.*, vol. 13, no. 03n04, Jul. 2021.
[14] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of Machine Learning: A review," *Appl. Sci.*, Feb. 2020.
[15] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student' performance prediction using Machine Learning techniques," *Neveléstudomány*, Sep. 2021.
[16] M. I. Foster and M. T. Keane, "The Role of Surprise in Learning: Different Surprising Outcomes Affect Memorability Differentially," *Top. Cogn. Sci.*, vol. 11, no. 1, pp. 75–87, Jan. 2019.
[17] R. Pekrun, "Emotions and learning," *Educational practices series*, vol. 24, no. 1, pp. 1–31, 2014.
[18] P. Hökkä, K. Vähäsantanen, and S. Paloniemi, "Emotions in Learning at Work: a Literature Review," *Vocations and Learning*, vol. 13, no. 1, pp. 1–25, Apr. 2020.
[19] N. M. Tsang, "Surprise in Social Work Education," *Soc. Work Educ.*, vol. 32, no. 1, pp. 55–67, Feb. 2013.
[20] M. Ramscar, M. Dye, J. W. Gustafson, and J. Klein, "Dual routes to cognitive flexibility: learning and response-conflict resolution in the dimensional change card sort task," *Child Dev.*, vol. 84, no. 4, pp. 1308–1323, Jan. 2013.
[21] L. Fridman, B. Reimer, B. Mehler, and W. T. Freeman, "Cognitive Load Estimation in the Wild," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, <conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>, 2018, pp. 1–9.

[22] N. Herbig *et al.*, “Investigating Multi-Modal Measures for Cognitive Load Detection in E-Learning,” in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, <conf-loc>, <city>Genoa</city>, <country>Italy</country>, </conf-loc>, 2020, pp. 88–97.
[23] F. Chen *et al.*, *Robust Multimodal Cognitive Load Measurement*. Springer International Publishing.
[24] D. Williams, “Predictive Processing and the Representation Wars,” *Minds Mach.*, vol. 28, no. 1, pp. 141–172, 2018.
[25] R. Singh, P. Majumdar, S. Mittal, and M. Vatsa, “Anatomizing Bias in Facial Analysis,” *Proc. Conf. AAAI Artif. Intell.*, vol. 36, no. 11, pp. 12351–12358, Jun. 2022.
[26] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. el Kaliouby, “AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit,” in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, San Jose, California, USA, 2016, pp. 3723–3726.
[27] N. Jmour, S. Masmoudi, and A. Abdelkrim, “A New Video Based Emotions Analysis System (VEMOS): An Efficient Solution Compared to iMotions Affectiva Analysis Software,” *Adv. Sci. Technol. Eng. Syst*, 2021.
[28] A. Magliacano, S. Fiorenza, A. Estraneo, and L. Trojano, “Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm,” *Neurosci. Lett.*, vol. 736, p. 135293, Sep. 2020.
[29] S. Chen and J. Epps, “Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load,” *Human–Computer Interaction*, vol. 29, no. 4, pp. 390–413, Jul. 2014.