Unraveling ML Models of Emotion With NOVA: Multi-Level Explainable AI for Non-Experts

Alexander Heimerl, Katharina Weitz, Tobias Baur, and Elisabeth André

Abstract—In this article, we introduce a next-generation annotation tool called NOVA for emotional behaviour analysis, which implements a workflow that interactively incorporates the 'human in the loop'. A main aspect of NOVA is the possibility of applying semi-supervised active learning where Machine Learning techniques are used already during the annotation process by giving the possibility to pre-label data automatically. Furthermore, NOVA implements recent eXplainable AI (XAI) techniques to provide users with both, a confidence value of the automatically predicted annotations, as well as visual explanations. We investigate how such techniques can assist non-experts in terms of trust, perceived self-efficacy, cognitive workload as well as creating correct mental models about the system by conducting a user study with 53 participants. The results show that NOVA can easily be used by non-experts and lead to a high computer self-efficacy. Furthermore, the results indicate that XAI visualisations help users to create more correct mental models about the machine learning system compared to the baseline condition. Nevertheless, we suggest that explanations in the field of AI have to be more focused on user-needs as well as on the classification task and the model they want to explain.

Index Terms—Tools and methods of annotation for provision of emotional corpora, interactive machine learning, explainable AI, trust, mental model, computer self-efficacy, human-computer interaction, annotation tools

1 Introduction

In this article we propose a framework that allows non-Machine Learning experts to employ AI techniques to their problem domain. More precisely we introduce a tool named NOVA that supports interdisciplinary researchers and end-users during the annotation process of continuous multi-modal data by incorporating Machine Learning techniques that are applied already during the annotation process. This way, users are enabled to interactively enhance their Machine Learning model by incrementally adding new data to the training set, while at the same time they get a better understanding of the capabilities of their model. This happens on multiple levels. First, they get a pure intuition of how well their model performs, by investigating false predicted labels. They might even learn specific cases in the data when their model "always fails" or when they can be sure they can 'trust' their model. Second, besides intuition, we provide so called eXplainable AI (XAI) algorithms within the workflow that allow users to generate local post-hoc explanations on instances their model predicted. This way we combine interactive machine learning techniques and explainable AI algorithms to involve the human in the machine learning process, while at the same time giving back control and transparency to users. Following our previous work [1] we performed a study with 53 participants to investigate how non-expert users can benefit from such a workflow. With this study we want to examine the following research questions:

1) How do people with little or no machine learning experience rate the interaction with the NOVA software?
2) What is the impact of the XAI information presented (confidence values, LIME visualisations, both or none) to non-experts in order to develop a correct mental model about a neural network model for emotion expression recognition?
3) How do non-experts rate the presented information (confidence values, LIME visualisations) in terms of simplicity of understanding and support for explaining the machine learning model?
4) How does the relevant image information of the XAI method LIME for emotion expression classification differ from humans?

We investigate the first research question by descriptively evaluating the feedback of the non-experts. For the second and third question, we calculated comparisons between different groups. To answer the fourth question we contrast LIME visualisations with non-expert drawings of relevant areas in face images.

This article contributes to investigate the impression of Machine Learning on non-experts during a Cooperative Machine Learning (CML) task. It also provides insights into whether non-experts benefit from XAI information.

- *The authors are with the Lab for Human-Centered AI, Augsburg University, 86159 Augsburg, Germany.*
  E-mail: {heimerl, weitz, baur, andre}@hcm-lab.de.

Manuscript received 14 Feb. 2020; revised 16 Nov. 2020; accepted 23 Nov. 2020. Date of publication 9 Dec. 2020; date of current version 6 Sept. 2022.
(Corresponding author: Alexander Heimerl.)
Recommended for acceptance by A. Kleinsmith.
Digital Object Identifier no. 10.1109/TAFFC.2020.3043603

2 Nova Tool

In order to answer the previously introduced research questions, we first give an overview on our machine-supported annotation and explanation tool NOVA. The NOVA tool aims to enhance the standard annotation process with the latest developments from contemporary research fields such as

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

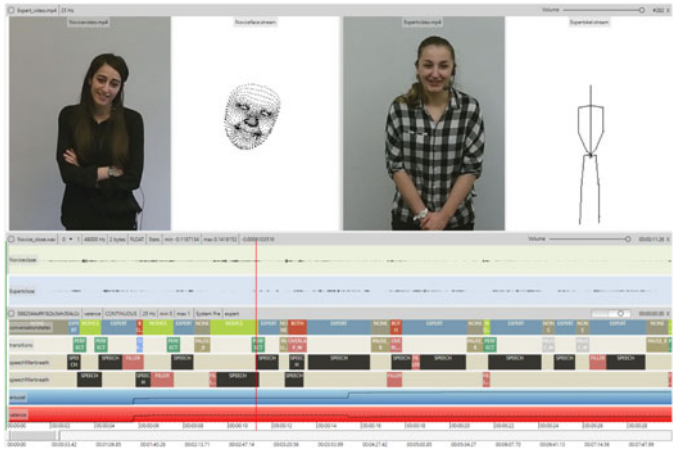

Fig. 1. NOVA allows to visualise various media and signal types and supports different annotation schemes. From top downwards: full-body videos along with skeleton and face tracking, and audio streams of two persons during an interaction. In the lower part, several discrete and continuous annotation tiers are displayed.

Cooperative Machine Learning and eXplainable Artificial Intelligence by giving annotators easy access to automated model training and prediction functionalities, as well as sophisticated explanation algorithms via its user interface.

The NOVA user interface has been designed with a special focus on the annotation of long and continuous recordings involving multiple modalities and subjects. A screenshot of a loaded recording session is shown in Fig. 1. On the top, several media tracks are visualised and ready for playback. Note that the number of tracks that can be displayed at the same time is not limited and various types of signals (video, audio, facial features, skeleton, depth images, etc.) are supported. In the lower part, we see multiple annotation tracks of different types (discrete, continuous, and transcriptions) describing the visualised content.

To support a collaborative annotation process, NOVA maintains a database back-end, which allows users to load and save annotations from and to a MongoDB database running on a central server. This gives annotators the possibility to immediately commit changes and follow the annotation progress of others. Beside human annotators, a database may also be visited by one or more "machine users". Just like a human operator, they can create and access annotations. Hence, the database also functions as a mediator between human and machine. NOVA provides instruments to create and populate a database from scratch. At any time new annotators, schemes and additional sessions can be added. NOVA provides several functions to process the annotations created by multiple human or machine annotators. For instance, statistical measures such as Cronbach's α, Pearson's correlation coefficient, Spearman's correlation coefficient or Cohen's κ can be applied to identify inter-rater agreement. Thus the foundations have been laid to fine-tune the number of labelers based on the inter-rater agreement in order to further reduce workload by allocating human resources to instances that are difficult to label (see [2]).

Tasks related to machine learning (ML) are handed over and executed by our open-source Social Signal Interpretation (SSI) framework [3]. Since SSI is primarily designed to build online recognition systems, a trained model can be directly used to detect social cues in real-time [4]. A typical ML pipeline starts by prepossessing data to input data for the learning algorithm, a step known as *feature extraction*. An XML template structure is used to define extraction chains from individual SSI components. A dialogue helps users to extract features by selecting an input stream and a number of sessions. The result of the operation is stored as a new signal in the database. This way, feature streams can be reviewed in NOVA and accessed by all users. Based on the extracted features, a classifier, which may also be added using XML templates, can be trained. Alternatively, NOVA supports Deep and Transfer Learning by providing Python interfaces to Tensorflow and Keras. This way convolutional networks may be trained, respectively retrained, based on annotations saved in NOVA's annotation database on raw video data. Such models may then be used to generate explanations as described in more detail in Section 4.

3 Cooperative Machine Learning

The next aspect of our work is related to the question: How we can make use of machine learning already in the process of labeling data? A common approach to reduce human labelling effort is the selection of instances for manual annotation based on active learning techniques. The basic idea is to forward only instances with low prediction certainty or high expected error reduction to human annotators [5]. Estimation of most informative instances is an art of its own right. A whole range of options to choose from exist, such as calculation of 'meaningful' confidence measures, detecting novelty (e.g.by training auto-encoders and seeing for the deviation of input and output when new data runs through the auto-encoder), estimating the degree of model change the data instance would cause (e.g., seeing whether knowing the label of a data point would make a change to the model at all), or trying to track 'scarce' instances, e.g., trying to find those data instances that are rare in terms of the expected label.

Further, more sophisticated approaches aggregate the results of machine learning and crowd-sourcing processes to increase the efficiency of the labelling process. Kamar *et al.* [6] make use of learned probabilistic models to fuse results from computational agents and human labelers. They show how to allocate tasks to coders in order to optimise crowdsourcing processes based on expected utility. Active learning has shown great potential in a large variety of areas including document mining [7], multimedia retrieval [8], activity recognition [9] and emotion recognition [10].

Most studies in this area focus on the gain obtained by the application of specific active learning techniques. However, little emphasis is given to the question of how to assist users in the application of these techniques for the creation of their own corpora. While the benefits of integrating active learning with annotation tasks has been demonstrated in a variety of experiments, annotation tools that provide users with access to active learning techniques are rare. Recent developments for audio, image and video annotation that make use of active learning include CAMOMILE [11] and iHEARu-PLAY [12]. However, systematic studies focusing on the potential benefits of the active learning approach within the annotation environment from a user's point of view have been performed only rarely [13], [14].

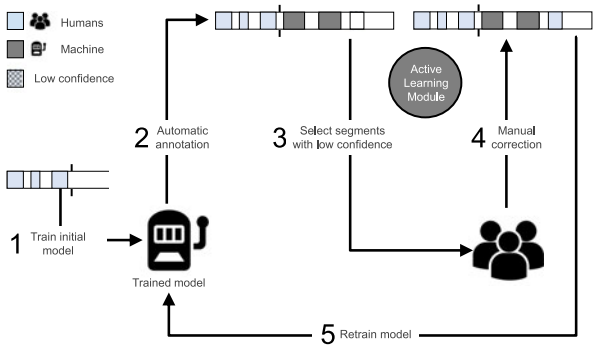

Fig. 2. The scheme depicts the general idea behind Cooperative Machine Learning (CML): (1) An initial model is trained on partially labelled data. (2) The initial model is used to automatically predict unseen data. (3) Labels with a low confidence are selected and (4) manually revised. (5) The initial model is retrained with the revised data.

In this article, we subsume machine learning approaches that efficiently combine human intelligence with the machine's ability of rapid computation under the term *Cooperative Machine Learning* (CML).

The main idea is that we train an initially "weak" model on a small labeled dataset, and use that model for predicting the remaining unlabeled dataset. While we probably can not expect our model to produce reliable results in the beginning, the human annotator who interactively gets involved in the training and prediction process gets an idea of in which cases the model succeeds and fails. Additionally, our model provides confidence values based on what it learned in the dataset so far and provides instances with particular low confidence to the annotator. The annotator then corrects or confirms a batch of said instances and the model is retrained with all previously labeled data (manual and corrected). While at the beginning, it might make sense to have a look at instances the model also is confident of (we don't know if we can trust our model at first), in later iterations, once we are aware of the strengths and weaknesses of our model, we only need to look at instances the model itself is not confident enough. In Fig. 2, we illustrate our approach to CML, which creates a loop between a machine learned model and human annotators: an initial model is trained (1) and used to predict unseen data (2). An active learning module then decides which parts of the prediction are subject to manual revision by human annotators (3+4). Afterwards, the initial model is retrained using the revised data (5). Now the procedure is repeated until all data is annotated. By actively incorporating the user into the loop it becomes possible to interactively guide and improve the automatic predictions while simultaneously obtaining an intuition for the functionality of the classifier.

However, the approach not only bears the potential to considerably cut down manual efforts but also to come up with a better understanding of the capabilities of the classification system. For instance, the system may quickly learn to label some simple behaviours, which already facilitates the workload for human annotators at an early stage. Then, over time, it could learn to cope with more complex social signals as well, until at some point it is able to finish the task in a completely automatic manner.

To automatically finish an annotation, the user either selects a previously trained model or temporarily builds one using the

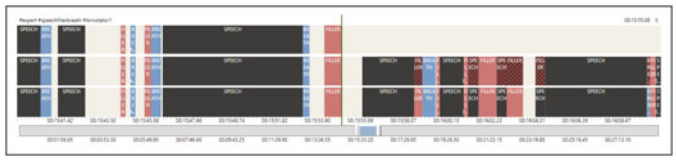

Fig. 3. The upper tier shows a partly finished annotation. ML is now used to predict the remaining part of the tier (middle), where segments with a low confidence are highlighted with a red pattern. The lower tier shows the final annotation after manual revision.

labels on the current tier. An example before and after the completion is shown in Fig. 3. Note that labels with a low confidence are highlighted with a pattern. This way, the annotator can immediately see how well the prediction worked.

To evaluate the efficiency of the integrated CML strategy, in our earlier work [4] we performed a simulation study on an audio-related labeling task. Following this approach, we were able to reduce the initial annotation labour of 9.4h to 5.9h, which is a reduction of 37.23 percent.

While we argue that confidence scores provide information that can help users to understand in which cases the model is or is not confident about its prediction, we aim to provide users with a more sophisticated comprehensible and transparent interpretation of their model. Therefore, we extended the CML workflow with techniques from the explainabale AI research area. In the next section, we give an overview on XAI methods and how we made use of them in the NOVA tool.

4 Explainable AI

Over the last decades, great advances in the field of affective computing and affect recognition have been made. Computational models constantly improved to provide more accurate approximations for highly complex human behaviours. However, with their increasing accuracy they gained ever growing attention from companies and non-research facilities. The AI Now Institute New York recently published their latest report that amongst other topics covers the current developments in the field of facial/affect recognition [15]. They mentioned various applications of computational models in different domains. Those range from call center programs that incorporate voice-analysis algorithms to detect distressed customers to systems that are used in criminal justice to detect potential deception by investigating eye movement and changes in pupil size. Overall many of the mentioned applications are highly safety-critical and make assumptions about sensitive information of the user. That is why they strongly emphasize the fact that those computational models and the application of such systems have to be carefully revised and scrutinized. Moreover, we argue that when classification results may even lead to harmful events for individuals it is important to fully understand the underlying process that leads to a classification. Making complex machine learning models more transparent and comprehensible for the user is the research focus of XAI. In general, Machine Learning models can be distinguished between inherently interpretable models and black-box models [16]. Examples for inherently interpretable ones are Bayesian classifiers or decision trees, whereas neural networks are a typical representative for black-box models. To

make the latter ones interpretable, additional effort has to be made. XAI approaches can be further distinguished between model-agnostic and model-specific techniques. Model-agnostic interpretation methods are able to provide explanations independent of the underlying model type [17]. In contrast to that model-specific approaches exploit the underlying inherent structures of the model and its learning mechanism, which in return bounds them to one specific type of model [16][17]. It is important to note that even though model-agnostic approaches are widely applicable, those techniques often rely on approximation methods, which in return may lead to less accurate explanations, whereas model-specific approaches, due to being specialized on a certain class of machine learning model, usually provide more accurate explanations [18].

Ribeiro *et al.* [19] present LIME, a model-agnostic approach. Their method is based on the idea to approximate an interpretable model around the original model. This way they are capable of creating explanations for various problem domains like text and image classification. Depending on the underlying model their information come in the form of textual or visual feedback and can be used to generate explanations about the model. For an image classification task, LIME is highlighting the sections that have been crucial for the prediction of a specific class. They showed that following their method it has been easier for users to determine from a set of classifiers which one performs best for a given problem domain. This is especially useful when test-accuracy scores themselves are misleading. Moreover, they argue that LIME not only is useful for gaining additional insight about a model but also users have been able to improve performance of classifiers by identifying unnecessary features and removing them based on the information for explanations generated by LIME. In contrast to that Alber *et al.* [20] introduced iNNvestigate a library that provides implementations of common analysis methods for neural networks, e.g., DeepTaylor and LRP. Those interpretation methods are model-specific. The supported approaches have in common that they, similar to LIME, highlight regions in the image, that have been important for the classification.

Lundberg *et al.* [21] introduced their own framework SHAP to address the issue that with the broad variety of interpretability methods available it is often not easily comprehensible when an approach suits a given problem domain the best. Their framework focuses on generating explanations by assigning each feature a value, that describes its importance in regard to the prediction.

While such XAI frameworks are of great value in helping to better understand which part of the input data was relevant for a decision, they still require expert knowledge about how to set up the systems and how to incorporate them with one's own model and data. That is why we implemented several of these frameworks into NOVA to provide non-experts with a more comprehensible and transparent machine learning experience. Fig. 4 displays an exemplary instance of explanations with different XAI frameworks in NOVA.

5 Psychological Aspects of Human-Computer Interaction

The goal of our user study is to get an impression of how non-experts perceive the software NOVA and whether the

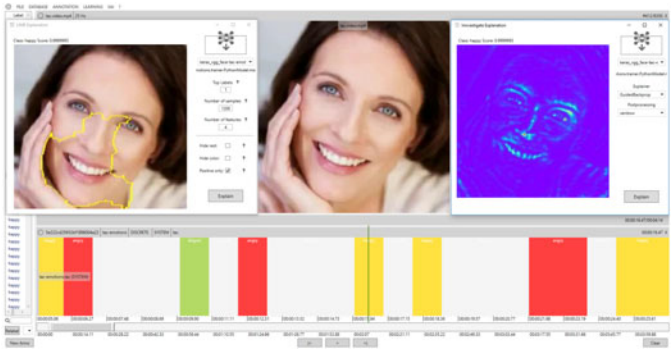

Fig. 4. An instance of the NOVA user interface including XAI visualisations. For each frame multiple explanations can be generated. In this instance the emotion class "happy" was predicted and the explanations show the relevant parts for the decision produced by different approaches (LIME: left window, iNNvestigate: right window) for a particular frame.

XAI information used helps them to better understand the presented Machine Learning model. The non-experts' impression of NOVA and the XAI information can be characterized by different inter- and intraindividual aspects. Therefore, related work about trust, self-efficacy expectation, cognitive workload, and the mental model in the context of interacting with AI will be presented.

5.1 Trust in Technical Systems

One common definition of trust in human-agent interaction sees trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." [22, p. 51].

A variety of studies (e.g., [23], [24], [25]) have shown that numerous factors influence whether people trust AI systems. Theoretical models of trust try to organize these factors. Merritt *et al.* [26], for example, distinguish between dispositional and history-based trust. Dispositional trust depends on past experience of trust, whereas history-based trust is continually changing due to past interactions with a particular system. The approach of Hoff and Bashir [27] has similar components. Hoff and Bashir [27] developed a theoretical approach in which they distinguished between dispositional, situational, and learned trust. Dispositional trust refers to the long-term tendencies that a person has, regardless of the current situation (e.g., age of the person, gender, cultural background, or personality). Situational trust describes external factors, e.g., influences such as the type of system to which the user is exposed, but also characteristics such as the cognitive workload in the situation or the task to be performed. Besides, there are also internal factors, i.e., anchored in the human being, such as the mood or self-confidence of the user. Learned trust, in turn, refers to the trust that someone has already developed based on previous experience (e.g., previous experience with AI).

Besides a positive trust level, authors like Marsh and Dibben [28] point out that the trust is a continuum between positive and negative trust (i.e., distrust).

In our study, we focus mainly on situational and subjective trust (not distrust), i.e., the trust in a deep neural network for emotion expression recognition. The examined factor in our study is the amount of information a non-

expert with no or little knowledge of AI receives about the model and the extent to which this information influences his/her trust in the AI system.

5.2 Perceived Self-Efficacy

Bandura described perceived self-efficacy as the "people's beliefs in their ability to influence events that affect their lives" [29, p.1]. *Computer self-efficacy*, as defined by Compeau and Higgins [30], describes the perceived self-efficacy of people concerning computers and related technologies. To measure computer self-efficacy, they developed the Computer Self-Efficacy Scale (CSE).

Numerous studies have found evidence that computer self-efficacy and user behaviour are related. For example, Hill *et al*. [31] found that there is a connection between perceived self-efficacy and the use of computers.

The information about the perceived self-efficacy of users can also be used to design and adapt AI systems in a more user-centered way. For example, Wiggins *et al*. [32] describe that the information about the perceived self-efficacy of users can be used to adapt intelligent tutoring systems to the abilities and preferences of the user. In addition, they found out in their study that especially people with high and low self-efficacy values benefit from those adaptations.

The computer self-efficacy of users could also be an indicator for peoples attitude towards AI. According to the Eurobarometer report of the European Commission, 75 percent of the European population has a generally positive attitude towards AI [33], when having already heard, seen or read about AI. In comparison, only 49 percent of respondents who have never had interacted with or had received information about AI are positive towards it. In addition to the opportunity to explore AI, the use of XAI could be a valuable support to improve the computer self-efficacy towards AI and thereby change the users' attitude towards AI. In our study, we evaluate the participants' computer-self efficacy towards the software NOVA. Our goal is to gain first insights into whether and how the perceived self-efficacy of users is influenced by the presentation of XAI information.

5.3 Task-Performance & Cognitive Workload

Performance in a psychological view is "any activity or gathering of reactions which leads to an outcome or has an impact on the surroundings" [34]. The performance in a task depends, among other things, on the cognitive workload. The characteristics that describe the cognitive workload of a task are not easily determined objectively. In addition to the requirements of the task, there is always the evaluation of the person performing the task. Therefore, Cognitive workload can be understood as the effort a person puts into fulfilling a task.

Hart and Staveland [35] developed the NASA-TLX questionnaire to measure the cognitive worload. In the field of Human-Computer interaction (HCI), Ramkumar *et al* [36] used the NASA-TLX to evaluate the process of interactive segmentation.

In our study, the NASA-TLX will be used to investigate whether the type of XAI information presented has an impact on the cognitive workload of the participants.

5.4 Mental Models

A mental model is a cognitive representation that a user has about a complex model [37], [38]. Through the interaction with a system, the mental model of the user can be formed or changed [39]. In this context, XAI can support users to create correct mental models. Therefore, XAI can be an important part of trust-calibration and technology adoption [39]. Ensuring that XAI can unfold its full potential, Richardson and Rosenfeld (2018) [40] indicate that it is important to evaluate why, what, how, and when an AI system should give explanations to the user.

With our study, we want to investigate which and how XAI information influences the user's mental model about AI.

6 Study

6.1 Study-Setup

NOVA was used in this study to improve a neural network model that recognizes emotional facial expressions based on image data. Accepting image data as input and predicting specific domain classes as output is generally known as image classification [41]. As a neural network architecture, we chose a convolutional neural network (CNN). CNNs set the benchmark on various famous image classification challenges like MNIST and ImageNet Large Scale Visual Recognition Challenge [42]. Moreover, we applied transfer learning to improve the performance of our model. Transfer learning is based on the idea to take already learnt knowledge about one domain and transfer it onto another domain to improve generalization [43]. In our case, we took advantage of the learnt knowledge of the VGG16 [44] CNN which performed exceptionally well on the ImageNet dataset. The assumption is that the network has already learnt meaningful features to classify images. However, the VGG16 CNN was trained to predict the classes of the ImageNet dataset. Therefore we stripped the fully connected layers of the network that are responsible for the mapping onto the domain-specific classes. We then added our own fully connected layers that correspond to our task of recognizing emotional facial expressions. Finally, our network was trained on the AffectNet facial expression corpus [45]. The corpus provides amongst other data annotations for the classes Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain and Non-face. Out of those categories we have chosen a subset (anger, disgust, happiness, sadness, and neutral) to train our neural network model. This subset consists out of four from Ekman's six basic emotions (happiness, sadness, anger, surprise, disgust, fear) [46]. We chose to not consider surprise and fear to reduce the complexity of the classification task. In the user-study, our trained model was used to predict visible emotions in images of facial expressions. Those images have not been part of the training set and therefore unknown to the model.

6.2 Study-Design

We conducted a study to investigate the influence of different types of XAI information (confidence values and LIME visualisations) on task-performance, computer self-efficacy, cognitive workload, and subjective trust of NOVA users with no or little machine learning background. The participants should help to improve the model's performance by identifying as much wrongly classified images as possible given a five minute time

frame, similar to the revision step described in Section 3. Additionally, they had to find as many images as possible in five minutes that the model has already classified well. For this purpose, the participants were presented with 254 images and the corresponding classifications of the neural network model in NOVA. The 254 images were equally distributed between the 5 classes and presented in an unsorted way. They were supposed to navigate freely through these images to get an overview of the model. This task varied the information displayed to the participants: One group (baseline condition) only received the images and the classification labels, another group (confidence values condition) additionally received the confidence values (i.e., how certain was the model?) similar to how they would be displayed with the CML workflow (see Section 3), a third group received LIME visualisations (LIME condition) as an additional stream which was displayed together with the images, and a fourth group (LIME & confidence values condition) received LIME visualisations as well as confidence values as information. After filling out a questionnaire which is described in 6.3, the participants were shown pictures with emotional facial expressions, where they had to classify on their own and report how sure they were with their decision. After this, they had to draw on each image which areas were important to them for their classification.

In order to make a statement about the required study size, we conducted a power analysis.

6.3 Evaluation Methods

After interacting with NOVA, the participants completed a questionnaire, including the following items:

Personal Information. At the beginning of the questionnaire, we asked the participant about personal information. These questions included age, gender, experience with ML in general and NOVA, and their knowledge about AI and XAI.

Impression of NOVA. After finishing the task using NOVA, we asked the participants to indicate their overall impression about NOVA. For this purpose, we used two questions, i.e., "The information NOVA provides are easy to understand", and "The information provided by NOVA helps to understand the model". These questions were rated on a 7-point Likert scale (1= don't agree, 7=totally agree).

Impression of XAI Methods. In addition to the general impression of NOVA, the participants were asked to rate the helpfulness and explainablity of the presented XAI methods. The first question was "The *XAI visualisations* NOVA provides are easy to understand", the second question was "The *XAI visualisations* provided by NOVA help to explain the model". The phrases in italic were changed, depending on the experimental condition. Again, these questions were rated on a 7-point Likert scale (1= don't agree, 7=totally agree).

Mental Model. To gain insight into the mental model of the users, we used the task reflection method, an approach recommended by Hofmann *et al.* [47]. This methods allows the user to describe their reasoning about the AI system. Therefore, after each five-minute interaction with NOVA, we asked them about their assumptions why the model recognized the pictures wrong or well respectively. This free-form feedback was combined with a Likert scale that allows users to evaluate their confidence in their statement.

TABLE 1
Demographic Information of the Participants

Characteristic	Conditions				Total
	0	1	2	3	
n	13	13	14	13	53
Age					
M	22.46	22.85	22.36	22.23	22.47
SD	2.47	3.02	2.59	2.89	2.68
Gender					
male	2	7	3	5	17
female	11	6	11	8	36
Experience					
NOVA	0	0	0	0	0
Machine Learning	1	4	1	2	8

0=Baseline condition; 1=Confidence values condition; 2=LIME condition; 3=LIME and confidence values condition.

Trust. For the assessment of the trustworthiness of the AI system, we used the TiA questionnaire [48]. Here, trust is regarded as a trait of the user. The TiA scale is one of the most commonly used trust scales in HCI [47]. With 11 items, the TiA measures 6 subscales of Trust: Fidelity, loyalty, reliability, security, integrity, and familiarity.

Computer Self-Efficacy. To measure the computer self-efficacy of the participants, we used the CSE scale [30]. This scale consists of 10 items that asked the user to estimate his/her perceived self-efficacy when using the NOVA (e.g., "I could complete the job using the software package if I had only the software manuals for reference"). These items were initially answered with "Yes" or "No". If a user answered "Yes", he or she was then asked on a 10-point Likert scale how confident he/she would be with this item (1= not confident at all, 10 = totally confident).

Cognitive Workload. We also collected data about their subjective workload using the NASA-TLX questionnaire [49]. On six scales, (i.e., mental demand, physical demand, temporal demand, performance, effort, and frustration level) participants were asked for their subjective assessment of the previously performed task using NOVA.

6.4 Participants

In total, 53 participants took part in the study (see Table 1 for more detailed demographic information).

All participants stated that they have heard the term *artificial intelligence* before. On average, they rated their impression of AI with 4.77 (SD = 0.91) clearly positive (range from 1 = extremely negative, 7 = extremely positive).

In contrast to this, only two participants stated that they have heard about XAI. After giving the participants the information what the goal of XAI is, in average participants rated XAI as important for politicians (M = 5.17, SD = 1.61), companies (M = 5.40, SD = 1.39), researchers (M = 5.47, SD = 1.35) as well as for non-experts (M = 5.60, SD = 1.45).

Most of the participants had no experience with machine learning and none of the participants used the software NOVA before.

7 Results

In the following, the results of the study will be presented. Starting with the evaluation of the software NOVA, followed by the results of the experimental groups comparisons.

TABLE 2
Rating of Participants, if the Confidence Values and LIME Visualisations are Helpful and Easy to Understand (Conditions: 1=Confidence Values; 2=LIME Visualisations; 3=LIME Visualisations and Confidence Values)

Characteristic	Conditions			Total
	1	2	3	
n	13	14	13	40
Confidence values (easy)				
M	6.77	-	6.23	6.50
SD	0.44	-	0.83	0.71
Confidence values (helpful)				
M	6.00	-	6.31	6.15
SD	0.91	-	0.75	0.83
LIME visualisations (easy)				
M	-	5.43	5.85	5.63
SD	-	1.22	1.41	1.31
LIME visualisations (helpful)				
M	-	5.71	5.62	5.67
SD	-	0.99	1.12	1.04

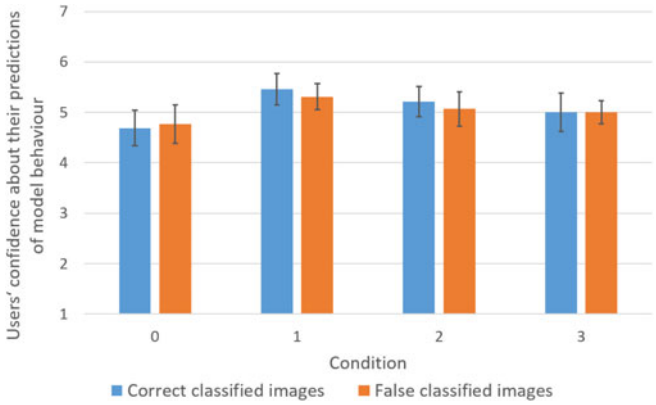

Fig. 5. Rating of the participants to what extent they are confident in their description of the behaviour of the neural network model. The rating was scaled between 1=disagree to 7=fully agree. 0=Baseline condition; 1=Confidence values condition; 2=LIME condition; 3=LIME and confidence values condition. Error bars represent the standard error.

Afterwards, the used LIME visualisations are compared with the human areas of interest.

7.1 Non-Experts Impression of NOVA

All 53 participants of the study interacted with the NOVA software for the first time. The overall impression of NOVA was particularly high (1=disagree to 7=fully agree). With an average of $M = 6.02$ ($SD = 0.84$), participants rated NOVA as easy to understand. They also rated NOVA to be helpful to understand the machine learning model ($M = 5.39$, $SD = 1.06$).

The evaluation of the CSE scale [30] showed that with an overall average of $M = 7.62$ ($SD = 1.18$) (1=not confident at all, 10=totally confident), the participants were confident that they would be able to cope successfully with the given tasks when interacting with NOVA again.

7.2 Subjective Trust, Self-Efficacy, and Cognitive Workload of Non-Experts

A one-way MANOVA was calculated to investigate the differences between the four conditions regarding subjective trust using the TiA questionnaire [48], computer-self efficacy using the CSE questionnaire [30], and cognitive workload using the NASA-TLX questionnaire [49]. The result of the MANOVA was not statistically significant, Wilks' Lambda = 0.80, $F(9, 115) = 1.21$, $p = .293$, which means there were no statistical differences between the conditions regarding the TiA, CSE and NASA-TLX ratings of the participants.

7.3 Non-Experts' Impression of XAI Methods

After the participants interacted with NOVA and described their impression about NOVA, participants in the three XAI information conditions were asked about the simplicity and helpfulness of this information, using two items (for a detailed description see Section 6.3). Overall, the results show that confidence values as well as LIME visualisations both reached values beyond 5 (1 = disagree, 7 = fully agree), which means they tend to be helpful and easy to understand (see Table 2).

To evaluate the two items, we conducted two one-way MANOVAs. The first MANOVA compared the impressions (simplicity and helpfulness) of the two conditions which saw the LIME visualisations. Here we found no significant differences between the conditions, Wilk's Lambda = 0.82, $F(2, 24) = 2.57$, $p = .097$.

The second MANOVA compared the impressions (simplicity and helpfulness) of the two conditions which saw the confidence values. Here we found a significant difference between the conditions, Wilk's Lambda = 0.71, $F(2, 23) = 4.80$, $p = .018$. The followed ANOVA revealed that there was a significant difference in the variable "easy to understand", $F(1,24) = 4.26$, $p = 0.05$, where participants of condition 1, who only saw the confidence values rated the confidence values as easier to understand compared to participants of condition 3, who saw confidence values combined with the LIME visualisation.

7.4 Non-Experts' Mental Model About the Neural Network

In order to determine the non-experts' mental model about the neural network model for emotion expression recognition, the participants were given the task of finding correctly and incorrectly classified images. Subsequently, they had to explain what aspects were relevant for the classification by the model. In addition, participants should state how confident they were in their explanation. Overall, the participants were as confident in their explanations about the relevant aspects for the neural network for correctly classified images ($M = 5.09$) as for incorrectly classified images ($M = 5.03$). When considering the confidence of the statements in the four conditions, a fairly equal rating between the conditions can be seen (see Fig. 5). To evaluate the ratings between the four conditions statistically, we conducted a one-way MANOVA. Here we found no statistical difference for all four groups, Wilks' Lambda = 0.94, $F(6, 96) = 0.51$, $p = .806$.

The similar quite good ratings between the conditions, even in the baseline condition without objective information in the form of XAI, make the evaluation of the open questions on the participants' reasons even more interesting. The lack of XAI information did not disturb the participants of the baseline condition to generate explanations about the models' behaviour. They simply justified the behaviour of the

TABLE 3
Explanations Given by the Participants About the Behaviour of the Neural Network

Condition	Example feedback of the participants
Baseline (no XAI information)	• *The emotions were clearly recognizable. the facial features were clear* • *Happy was especially recognized by a laughing mouth; also corresponded most often to my own opinion* • Images and emotions have been well matched. Only neutral faces did not always fit perfectly • The Indian woman's eyes were so full of make-up that the system predicted she would be happy, but for me she looked neutral to disgusted
Confidence values	• *For the pictures classified as Happy: on the smile, teeth often shown; Neutral: Few facial expressions"* • *Often happy, because of teeth & smiling* • Large eyes are classified as aggressive in pictures • Sad often did not correspond to a sad expression. Apart from that nothing special noticed
LIME visualisations	• *I think the model recognized the pictures correctly, especially because it looked at the mouth and/or eyes* • *The model had focused the XAI visualization on relevant areas of the face. eye area, mouth area* • I believe that the model has misclassified the images because it has often focused the mouth - and just because the mouth is open does not mean that the image shows someone "Happy" or "Angry" • In some cases, the XAI visualization did not refer to the face at all, but marked the background or clothing
Confidence values & LIME visualisations	• *On certain parts of the face, the model was able to easily identify the appropriate emotions* • *The model focuses on the eye and mouth area* • Sometimes the eyes are not taken into account, e.g. if the teeth are seen, the person can still be sad • Unnecessary areas such as the background are taken into account, mouth and eyes are hardly or not at all considered. Why the program does not concentrate on these areas is not understandable

Sentences in italic refer to the networks behaviour when classifying images correctly, non-italic statements to incorrect classifications.

neural network with the arguments they themselves use for emotion classification (see Table 3 for examples of participants' feedback). In the baseline condition, most of the participants described their assumptions about the models' behaviour for the emotion happiness, followed by descriptions for the emotion sadness. Here, prototypical facial expressions (e.g., for happiness: pull up of the corners of the mouth, show teeth) were used as explanations. Furthermore, participants often used their own assumptions as a reference for the behaviour of the model (e.g., "corresponded to my own opinion" or "for me she looked disgusted").

In contrast to this, in the two conditions with the LIME visualisations, it can be seen that the participants described less their own strategies for emotion recognition, but used the XAI information instead. They refer to superpixel areas presented to them by LIME.

Interestingly, in the two conditions where confidence values were displayed, the information about the uncertainty of the model was not used by the participants to explain its behavior. The decisive factor was whether people were additionally shown LIME visualisations or whether they only saw confidence values. If they saw LIME visualisations, the answers were similar to the condition that only saw LIME visualisations. If they only saw confidence values, the responses were very similar to the baseline condition who assumed their own assumptions were those of the model.

In Fig. 6, two images which are presented in the study using NOVA are shown. The superpixels generated by LIME for the classification of happiness are displayed. On the left image, the neural network model focuses on the mouth for the classification of happiness. On the right image, the model focuses on the background to classify happiness. This faulty learning of the neural network with simultaneous correct prediction was only recognized and mentioned as a problem by participants in the two LIME visualisation conditions.

7.5 Areas of interests for LIME and Humans

As the final task of the study, the participants were asked to highlight areas of relevance for classifying emotions in images. They were explicitly told that they should mark areas, which they think have been important for their recognition of a specific emotion. In the following, we are going to compare heatmaps generated from participants' reported areas with XAI visualisations generated by LIME. Fig. 7 displays heatmaps and LIME visualisations for five different faces (A to E) which all correspond to a specific emotion. Following emotions are present: A: anger, B: neutral, C: disgust, D: sadness, E: happiness. The top row covers the heatmaps.

Fig. 6. XAI visualisation generated by LIME for two images classified as happy by a neural network model. While in the left image the network focused on the mouth region, in the right picture the background seems to have had an impact on the model's decision.

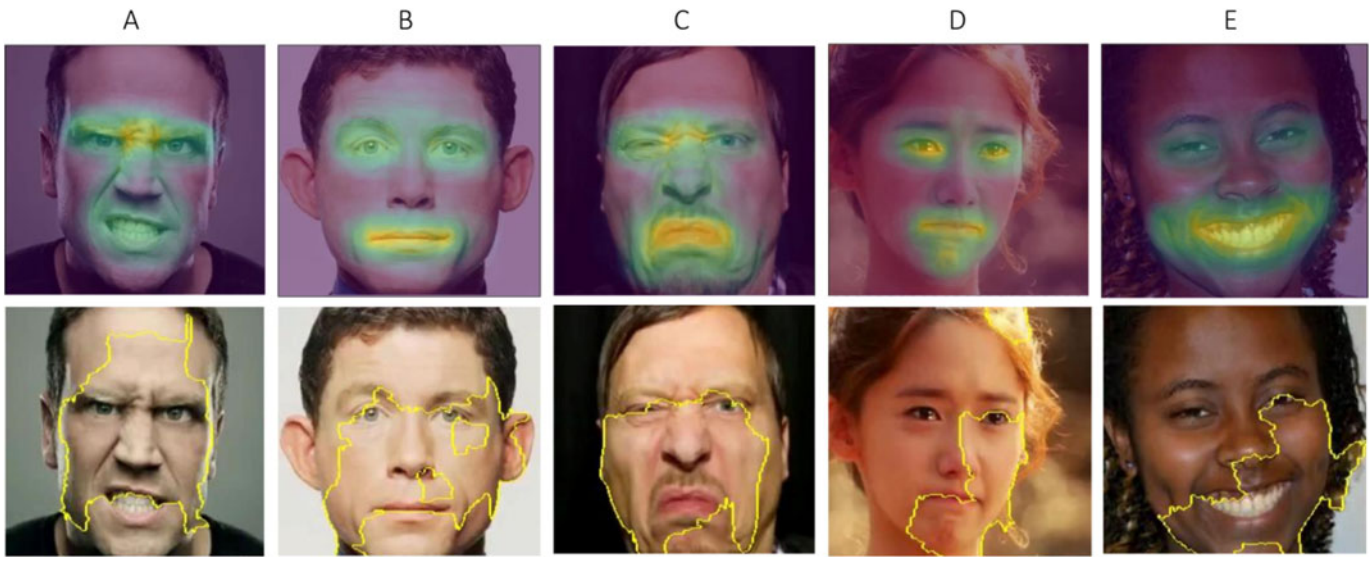

Fig. 7. Comparison between the average areas of interest according to the study participants and model agnostic explanations generated with LIME. The different faces show varying emotions. A: anger, B: Neutral, C: Disgust, D: Sadness, E: Happiness.

The brightness of the coloring describes the importance of the facial areas, as marked by the participants. The bottom row covers the XAI visualisation from LIME. The spaces defined by the yellow bounds describe the areas of the face that have been important for classifying a specific emotion expression. When analyzing the heatmaps, it is conspicuous that the participants identified for all faces the eye and mouth area to be most important. Little to no attention has been paid to other facial regions. For the angry face (A) most emphases were put on the region between the eyes and the eyes itself. This is most likely due to the presence of wrinkles. The mouth region played a subordinate role. In the neutral face (B) especially the area around the mouth has been considered important. In addition to that, the eyes have been given attention. The disgusted face (C), similar to the angry face, displays wrinkles in between the eyes, which have been identified by the participants as a relevant area. Moreover, the specific shape of the mouth, with the corners of the mouth facing downwards were marked as very important. For the sad facial expression (D) the mouth and the wrinkled chin have been recognized as valuable information. It is noteworthy that for this facial expression the eyes themselves have been considered exceptionally important. That is most likely due to the fact that for this image tears have been present in the corners of the woman's eyes. In the happy looking face (D) the region around the mouth, displaying a big smile and the corresponding wrinkles around the cheeks have been identified as most important. Additional little attention was given to the area around the eyes. In contrast to that, the automatically generated LIME visualisations cover larger areas of importance. This difference is especially evident for the angry, neutral and disgusted face. In general, it seems that humans focus more on specific facial features, whereas the trained model takes a rather holistic approach by putting more emphasis on larger areas of the face.

Following the question about important areas for emotion recognition, we asked the participants after each shown image how certain they were with their decision. The rating was scaled between 1=unsure to 8=fully sure. The corresponding results are displayed in Fig. 8. Overall, the participants have been very confident in their own decisions. None of the different emotions resulted in a score below 6. However, there have been differences between the emotions. The participants have been most certain with their judgement for the happy and sad face. They have been most uncertain for the disgusted facial expression, followed by neutral. Anger placed in the middle regarding their certainty. Also, no one of the participants did classify any of the presented emotions wrong.

8 Discussion

The aim of our study was to gain insights into the interaction between non-experts and machine learning models using NOVA. In the following, the results obtained will be discussed.

8.1 NOVA Is Helpful for Non-Experts

The results of our study show that users who have little or no experience with Machine Learning are able to use NOVA for labeling data in the revision step of the CML workflow (see Section 3 step 4). Even though all of the participants have never worked with NOVA before, they found

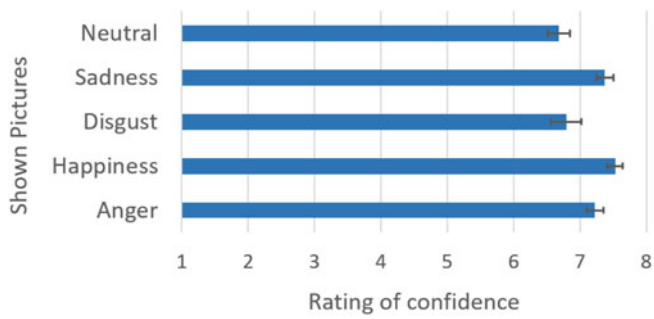

Fig. 8. Rating of the participants to what extent they are confident in their classification of the emotion pictures. The rating was scaled between 1=unsure to 8=fully sure. Bars represent the standard error.

it easy to use and had the impression that NOVA helps them to understand the machine learning model.

Similar results were found for the given XAI information. Confidence values, as well as XAI visualisations generated by LIME were rated as easy to understand and helpful by the participants.

Also, the CSE values show that the participants have a high computer self-efficacy towards NOVA. They believed that they would be able to do similar tasks with NOVA in the future.

8.2 XAI does not Automatically Influence Users' Perceptions

We found no difference in the CSE values between the four experimental conditions. Instead, the participants in all conditions achieved high CSE values. Similar results were found for subjective trust and cognitive workload. A cause for this could be the easy handling and usage of NOVA as well as the domain of the classification task of the neural network model. Emotion expression recognition is a task where (most) humans perform quite well. This could have led to more self-confidence and increased trust in the system, compared to the work of [50], where the domain explained with XAI visualisations (spoken words in form of spectrograms) was not familiar to humans.

But the use of different XAI methods seems to influence the subjectively perceived simplicity of the specific method. The fact that users found the use of confidence values harder to understand when they also saw LIME visualisations may be a first indication that XAI visualisations give users the impression of being easier to be interpreted. De Graaf and Malle [51] assumed that people apply human traits to AI systems. This leads to the expectation that the AI system should explain its behaviour in a human-like manner. Miller [52] points out that explanations including probabilities are not necessarily the best explanations for a user.

8.3 Users Create Assumptions About AI

An interesting result we observed is that even without XAI information, participants in the baseline condition formulated extensive explanations about the behaviour of the neural network model and were also very confident in their reasoning. This is an indication that with high computer self-efficacy and a very well-known application domain (e.g., emotion expression recognition), users tend to equate their own assumptions with those of the Machine Learning model. This assumption can have devastating consequences if it does not hold because people do not question whether the model has learned what it should have learned (see Fig. 6).

We found a difference regarding the users' mental models of the AI system and their assessment of how helpful and easy to understand the XAI methods were. Although the users had the impression that the XAI methods were helpful and easy to understand, only the two conditions with LIME visualisations helped the users to create more correct mental models.

Even if the explanations of the participants about the behaviour of the neural network model in the conditions of the LIME visualisations were more accurate and correct than in the other conditions, it must be pointed out that visualisations alone are not sufficient to generate exhaustive explanations. For example, many participants in the two LIME visualisation conditions still assumed additional information that is not part of the visualisations themselves (e.g.image sharpness, image exposure). XAI visualisations alone do not explain anything, they only provide information that has to be interpreted by the user. But the interpretation itself may again be flawed. Therefore, it is necessary to go beyond visualisations and provide additional information, for example, in form of combining LIME with linguistic explanations about relational concepts [53] (e.g., "The classification was happiness because the raised corners of the mouth were relevant") in order not to leave the interpretation completely to the imagination of the user.

8.4 XAI Perception Differs From Human Perception

In Section 7.5 we presented the results for the task of manually highlighting facial regions that are supposedly relevant for a specific emotion and compared those with the marked regions generated by LIME, in which the output of LIME describes areas that have been crucial for the classification. We found that the participants identified the eye and mouth area to be most important. However, depending on the presented emotion they weighted those areas differently, e.g., for the angry facial expression the eye region was considered more important whereas for the happy face the focus was on the mouth. Moreover, they tended to value specific facial features more than a holistic approach to recognize emotions in facial expressions. Those findings are interesting when put into context with the research of Bombari *et al.* [54]. They investigated the role of featural (e.g., shape of the mouth) and configural face information (relational information, e.g., the distance between the nose and the mouth) when it comes to recognizing emotions. For their experiments, they used faces representing four different emotions (happiness, sadness, anger, and fear). They reported that happiness has been recognized more easily and rapidly when compared to other emotions. Also, they stated that the mouth region has been particularly important for recognizing happiness. This is in line with our finding that the participants have been most confident in their classification for the happy facial expression (see Fig. 8) and they highlighted the mouth as most relevant for their classification. It is important to note that in our study we explicitly asked the participants what they think the important regions for recognizing a specific emotion are, whereas Bombari *et al.* gathered that information by using eye tracker systems. When we compare the results of Bombari *et al.* with the generated heatmaps of the facial expressions in Fig. 7, it seems that when asked what the relevant information for recognizing a specific emotion is, humans tend to focus more on the featural aspects of faces rather than the configural information. We mentioned earlier in Section 7.5 the impression that our trained neural network model follows a rather holistic approach to recognize emotional expressions. When we now inspect the visualisation for the relevant areas generated by LIME, it is visible that a large area of the face is marked as especially important for classification. The participants identified specific features to be

important, whereas the neural network model focuses on larger facial areas. It is important to understand that depending on the emotion, either configural or featural information is more relevant for humans to visually classify facial expressions [54], but when asked, people tend to state that mainly featural information is considered important. This should be kept in mind when providing additional information to humans about the inner workings of machine learning models. It could be similar to our case that the model actually imitates a human-like holistic perception behaviour, but the users may not appreciate the explanation as they feel like irrelevant information is considered important. Further, generating explanations should be in line with human expectation while mapping the actual behaviour of machine learning models.

8.5 Implications for Other Emotion Recognition Domains

In our proposed study we investigated how XAI techniques can assist non-experts in terms of trust, perceived self-efficacy, cognitive workload and creating a correct mental model about a system. However, we solely considered a non-verbal aspect of affective computing namely the recognition of emotional facial expression. In fact, recent studies in the field of affective computing also focus on sentiment analysis and natural language processing [55]. As a result, innovative approaches emerged like using stacked ensemble to predict the intensity of sentiments and emotions [56] or applying novel semi-supervised learning techniques to extract knowledge from unstructured social data [57]. For future work, it would be interesting to examine how XAI methods perform on black-box models that predict emotion from text.

9 Conclusion

In our study, we showed that interactive machine learning applications like NOVA are helpful for tasks that involve non-experts in the process. Even non-experts found NOVA easy to use and to understand. Moreover, the participants were confident about their ability to employ NOVA for similar affective computing annotation tasks. We have further shown that XAI information is considered comprehensible and helpful to our participants that had no or only little expertise in data annotation and machine learning. We, therefore, conclude that incorporating such techniques in end-user applications offers value to users in the interactive machine learning loop and machine learning enthusiasts alike.

One of the key revelations of our work has been that humans create assumptions about AI. In our study, we found that especially when users get presented little to no additional information about the inner workings of the system, they start to apply their own mental model upon the machine learning model. This became evident when investigating the reported feedback of the participants about the predictions the system made. We argue that this is connected to the high levels of self-efficacy and a domain (emotion expression recognition) the participants are familiar with. Further, we want to stress that such behaviour is to be seen critical, especially when the computational model does not align with the mental model of the user. In those cases, the users might stop questioning what the computational model actually has learned.

Moreover, it became evident that explanations in the form of visualisations are helpful to create a correct mental model but alone are not sufficient to provide enough transparency and insight about a given system. This claim is grounded on the fact that the participants in the two LIME conditions - even though they referred in their feedback to the given visualisations - still made up additional reasons that were not accessible from the information they were provided. Further, we argue that such visualisations themselves are not explanations but offer additional information that has to be interpreted by the user. Therefore we recommend to use this kind of visual feedback and combine it with additional information or interpretation to provide the user with more holistic explanations. A possible implementation for our use case could be to add some kind of textual or verbal explanation in the form of "The person seems to be happy because the raised corners of the mouth were of high relevance and indicate a smile". In such a case the user would have access to the actual image with the marked areas that have been considered important by the machine learning model, as well as an interpretation of what the model actually focused on.

At last, we want to stress the fact that the context and domain of a classification task might influence how XAI visualisations are perceived and interpreted. Interpreting the results of the task where participants were asked to identify important information in given images of facial expressions, we found that there is a discrepancy between what people consider important as to how they actually process certain emotions. This could potentially lead to less acceptance of a machine learning model even though the behaviour might be in line with the human approach of processing information. Therefore we suggest to generate explanations that align with human perception of a given problem domain. This is highly connected to our earlier recommendation to provide holistic explanations that are easier for the user to comprehend and assist him or her when it comes to interpreting the presented behaviour.

In this article, we applied the cooperative machine learning workflow that incorporates explanations in an affective computing problem domain. We strongly believe that other disciplines such as health care, psychotherapy, and others may also benefit from such technologies. Especially in high-risk environments that apply artificial intelligence, it is crucial to not only rely on high prediction accuracies but also to fully understand the underlying processes that led to a classification result. Tools such as NOVA prove to be valuable as they can potentially help domain experts (e.g., physicians, psychotherapists) with little to no expertise in machine learning to better assess the behavior of their ML models.

Acknowledgments

This work was supported by DFG under project number 392401413, DEEP. Further this work presents and discusses results in the context of the research project ForDigitHealth. The project is a part of the Bavarian Research Association on Healthy Use of Digital Technologies and Media (ForDigitHealth), funded by the Bavarian Ministry of Science and

Arts. Moreover, the presented study has been approved by the data protection officer of the University of Augsburg. Alexander Heimerl, Katharina Weitz, Tobias Baur, and Elisabeth André contributed equally to this work.

References

[1] A. Heimerl, T. Baur, F. Lingenfelser, J. Wagner, and E. André, "Nova - a tool for explainable cooperative machine learning," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 109–115.

[2] Y. Zhang, A. Michi, J. Wagner, E. André, B. Schuller, and F. Weninger, "A generic human-machine annotation framework based on dynamic cooperative learning," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1230–1239, Mar. 2020.

[3] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (SSI) framework: Multimodal signal processing and recognition in real-time," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 831–834.

[4] T. Baur *et al.*, "Explainable cooperative machine learning with nova," *KI - Künstliche Intelligenz*, vol. 34, pp. 143–164, Jan. 2020.

[5] B. Settles, *Active Learning: Synthesis Lectures on Artificial Intelligence and Machine Learning*. San Rafael, CA, USA: Morgan & Claypool, 2012.

[6] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2012, pp. 467–474.

[7] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Mar. 2002.

[8] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 10:1–10:21, Feb. 2011.

[9] M. Stikic, K. V. Laerhoven, and B. Schiele, "Exploring semi-supervised and active learning for activity recognition," in *Proc. 12th IEEE Int. Symp. Wearable Comput.*, 2008, pp. 81–88.

[10] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 275–278.

[11] J. Poignant *et al.*, "The CAMOMILE collaborative annotation platform for multi-modal, multi-lingual and multi-media documents," in *Proc. 10th Int. Conf. Lang. Resour. Eval.*, 2016, pp. 1421–1425.

[12] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "ihearu-play: Introducing a game for crowdsourced data collection for affective computing," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 891–897.

[13] J. Cheng and M. S. Bernstein, "Flock: Hybrid crowd-machine learning classifiers," in *Proc. 18th ACM Conf. Comput. Supported Cooperative Work Soc. Comput.*, 2015, pp. 600–611.

[14] B. Kim and B. Pardo, "I-sed: An interactive sound event detector," in *Proc. 22nd Int. Conf. Intell. User Interfaces*, 2017, pp. 553–557.

[15] C. Kate *et al.*, "AI now 2019 report," AI Now Institute, 2019. [Online]. Available: https://ainowinstitute.org/AI_Now_2019_Report.html

[16] A. Rai, "Explainable ai: from black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020.

[17] C. Molnar, "Interpretable machine learning a guide for making black box models explainable," 2019. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[18] P. Hall and N. Gill, *Introduction to Machine Learning Interpretability*. Sebastopol, CA, USA: O'Reilly Media, 2018.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[20] M. Alber *et al.*, "Innvestigate neural networks!," *J. Mach. Learn. Res.*, vol. 20, no. 93, pp. 1–8, Feb. 2019. [Online]. Available: https://jmlr.org/papers/v20/18-540.html

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[22] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[23] B. Petrak, K. Weitz, I. Aslan, and E. André, "Let me show you your new home: Studying the effect of proxemic-awareness of robots on users' first impressions," in *Proc. 28th IEEE Int. Conf. Robot Human Interactive Commun.*, 2019, pp. 1–7.

[24] T. Huber, K. Weitz, E. André, and O. Amir, "Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps," 2020, *arXiv: 2005.08874*.

[25] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ""Let me explain!": Exploring the potential of virtual agents in explainable ai interaction design," *J. Multimodal User Interfaces*, pp. 1–12, 2020.

[26] S. M. Merritt and D. R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *J. Human Factors Ergonomics Soc.*, vol. 50, no. 2, pp. 194–210, 2008.

[27] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[28] S. Marsh and M. R. Dibben, "Trust, untrust, distrust and mistrust – an exploration of the dark(er) side," in *Trust Management*, P. Herrmann, V. Issarny, and S. Shiu, Eds. Berlin, Germany: Springer, 2005, pp. 17–33.

[29] A. Bandura, "Self-efficacy," *Corsini Encyclopedia Psychol.*, pp. 1–3, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470479216.corpsy0836

[30] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS Quart.*, vol. 19, pp. 189–211, 1995.

[31] T. Hill, N. D. Smith, and M. F. Mann, "Role of efficacy expectations in predicting the decision to use advanced technologies: The case of computers," *J. Appl. Psychol.*, vol. 72, no. 2, 1987, Art. no. 307.

[32] J. B. Wiggins, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Do you think you can? the influence of student self-efficacy on the effectiveness of tutorial dialogue for computer science," *Int. J. Artif. Intell. Educ.*, vol. 27, no. 1, pp. 130–153, 2017.

[33] EuropeanCommission, "Special eurobarometer 460," *Attitudes Towards the Impact of Digitisation Autom. Daily Life. Eurobarometer Report*, 2017.

[34] P. M. Nugent, "Performance," 2013. Accessed: Sept. 2, 2020. [Online]. Available: https://psychologydictionary.org/performance/

[35] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Advances Psychol.*, vol. 52, pp. 139–183, 1988.

[36] A. Ramkumar *et al.*, "Using GOMS and NASA-TLX to evaluate human–computer interaction process in interactive segmentation," *Int. J. Human–Comput. Interaction*, vol. 33, no. 2, pp. 123–134, 2017.

[37] F. G. Halasz and T. P. Moran, "Mental models and problem solving in using a calculator," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 1983, pp. 212–216.

[38] D. A. Norman, "Some observations on mental models," in *Mental Models*. London, U.K: Psychology Press, 2014, pp. 15–22.

[39] H. Rutjes, M. Willemsen, and W. IJsselsteijn, "Considerations on explainable ai and users' mental models," in *Where is the Human? Bridging the Gap Between AI and HCI*, Glasgow, U.K.: Association for Computing Machinery, 2019.

[40] A. Richardson and A. Rosenfeld, "A survey of interpretability and explainability in human-agent systems," in *Proc. 2nd Workshop Explainable Artif. Intell.*, 2018, pp. 137–143.

[41] K. Balaji and K. Lavanya, "Chapter 5 - medical image analysis with deep neural networks," in *Deep Learning and Parallel Computing Environment for Bioengineering Systems*, A. K. Sangaiah, Ed. Cambridge, MA, USA: Academic Press, 2019, pp. 75–97. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128167182000129

[42] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, 2017. [Online]. Available: https://doi.org/10.1162/neco_a_00990

[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.

[45] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 18–31, 2017.

[46] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, 1971, Art. no. 124.

[47] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: challenges and prospects," 2018, *arXiv:1812.04608*.

[48] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognitive Ergonom.*, vol. 4, no. 1, pp. 53–71, 2000.

[49] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," in *Advances in Psychology*. Amsterdam, The Netherlands: Elsevier, 1988, pp. 139–183.
[50] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ""Do you trust me?": Increasing user-trust by integrating virtual agents in explainable ai interaction design," in *Proc. 19th ACM Int. Conf. Intell. Virt. Agents*, 2019, pp. 7–9.
[51] M. M. A. De Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *AAAI Fall Symp. AI-HRI*, 2017, pp. 19–26.
[52] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2018.
[53] J. Rabold, H. Deininger, M. Siebers, and U. Schmid, "Enriching visual with verbal explanations for relational concepts–combining lime with aleph," 2019, *arXiv: 1910.01837*.
[54] P. Schmid, M. Mast, S. Birri, F. Mast, and J. Lobmaier, "Emotion recognition: The role of featural and configural face information," *Quart. J. Exp. Psychol.*, vol. 66, pp. 2426–2442, 2013.
[55] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
[56] M. S. Akhtar, A. Ekbal, and E. Cambria, "How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]," *IEEE Comput. Intell. Magazine*, vol. 15, no. 1, pp. 64–75, Feb. 2020.
[57] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231217316363

Alexander Heimerl received the master's degree in computer science and information technology from the University of Augsburg, Germany. He is currently working as a research associate with the Lab for Human-Centered AI, University of Augsburg. His research interests include machine learning in the context of affective computing and explainable AI.

Katharina Weitz received the master of science degree in psychology, and the master of science degree in computing from the Humanities (applied computer science), University of Bamberg, Germany. She is currently working with the Lab for Human-Centered AI, University of Augsburg. She is interested in machine learning topics in the field of social Human-Computer Interaction. The influence of explainability and transparency of AI systems on people's trust and mental models is a central point of her research activities.

Tobias Baur received the PhD (Doktor rer. nat.) degree, in 2018. He is a postdoctoral researcher with the Human-Centered AI Lab, Augsburg University, Augsburg, Germany. His main research focuses on human-centered tools that help (non-) experts to create and understand Machine Learning models. His research interests include artificial emotional intelligence; social signal processing and ethical & explainable AI

Elisabeth André is currently a full professor in computer science with Augsburg University, Augsburg, Germany, and the chair of the Laboratory for Human-Centered AI. She holds a long track record in embodied conversational agents, multimodal interfaces, and social signal processing. She was elected as a member of the German Academy of Sciences Leopoldina, the Academy of Europe and AcademiaNet. She is also a fellow of the European Coordinating Committee for Artificial Intelligence.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.