I see it in your eyes: Training the shallowest-possible CNN to recognise emotions and pain from muted web-assisted in-the-wild video-chats in real-time

Vedhas Pandit[*,a], Maximilian Schmitt[a], Nicholas Cummins[a], Björn Schuller[a,b]

[a] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*
[b] *GLAM – Group on Language, Audio & Music, Imperial College London, UK*

1. Introduction

The simultaneous emergence of social media environments as a new source of data and unprecedented advancement of artificial intelligence (AI) has opened doors to many next-generation applications in healthcare. We discuss first precisely this ever-increasing synergy between the three domains (cf. Figure 1), the related challenges, followed by our contribution to the emotion-aware and pain-aware AI for remote patient monitoring and counselling.

[*] Corresponding author.
E-mail address: panditvedhas@gmail.com (V. Pandit).

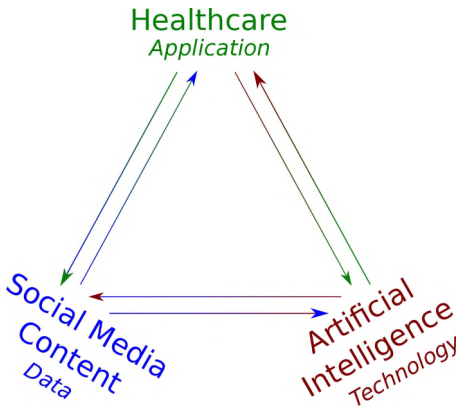


Fig. 1. Synergy between social media content (the data), artificial intelligence (the technology), and healthcare – an application end of the utmost importance.

1.1. The avenues for the social media data analytics / AI in healthcare

Because the social media provides an unparalleled insight into the lives, emotions of the real people, it can be leveraged to recognise patterns relating to their health needs. For example, a patient's feedback on his/her treatment, in the context of their ethnicity, age, gender, smoking, drinking habits gives health professionals a better insight into the patient-education and demographic-tailored remedies (Antheunis, Tates, & Nieboer, 2013). Social media data helps create strategies for improving customer engagement for different healthcare services, e. g. targeted stimulus to motivate a fitness regime (Korda & Itani, 2013). Social media enables everyone to exchange their healthcare experience, and to learn from others (Spink et al., 2004). Information sharing also helps healthcare professionals to identify the gaps towards a better health outcome, e. g. reduction of waiting times, improved customer satisfaction and doctor-patient relationship (Smailhodzic, Hooijsma, Boonstra, & Langley, 2016). In a pandemic like COVID-19, or even otherwise, a web-assisted video platforms can be used to address customer grievances, and for remote patient consultations, counselling and health monitoring (Armfield, Gray, & Smith, 2012) – the related research being the prime focus of this paper. While not the main focus of this paper, but for the sake of completeness, the novel challenges posed by AI to the 'AI-Healthcare-Social data' synergy are identified next, when considering widespread adaption of any application including the research presented herein.

1.2. The anomalous challenge of AI to the healthcare-social data-AI synergy

1.2.1. AI-Driven data augmentation

The rapid AI advancements have made high volume, high velocity social media analytics a feasible reality (Amiriparian, Schmitt, Hantke, Pandit, & Schuller, 2019b; Pandit, Amiriparian, Schmitt, & Schuller, 2019a). Such is the pace of this progress that the finesse with which AI performs complicated tasks remains hard to believe, yet is a common knowledge ironically – be it coherent synthetic text generation, voice mimicking, or a high resolution 'deep-fake' video production. For example, it might be hard to believe that a few applications of social data in healthcare in this very manuscript were suggested by the GPT-2 English language model (Radford et al., 2019)[1], with a frighteningly impressive human-like coherency. Because a human-like content generation is now this easy, the same AI advancement simultaneously manifests itself both as a novel challenge and as an opportunity.

Likewise, issues related to data privacy, a central consideration in healthcare, have become complex. Highly personal data (e. g. health issues, vital signs) is what drives the healthcare. AI helps with synthetic yet realistic data augmentation, coupled with more advanced anonymisation techniques addressing privacy-related issues. Simultaneously and ironically, research in AI-based deanonymisation has made anonymisation of data lot more challenging (Malin & Sweeney, 2004).

1.2.2. White-box AI: A prerequisite in healthcare

The earliest and the main criticism of deep learning technology has long been its inexplicability. A white-box AI is a *necessity* in healthcare, where an incorrect treatment can be fatal, where each diagnosis should be based on a reliable *understanding* of the data available. In recent years, there are monumental developments with several approaches proposed for model explainability (Alber et al., 2019). In this paper as well, in addition to reporting performance metrics, we compute *feature attributions* towards better understanding of the models proposed.

[1] https://talktotransformer.com/

1.3. Prerequisites of an emotion-aware/unaware AI in healthcare

Data mining has helped improve diagnosis and treatment of various diseases (Rodríguez-González et al., 2012), e. g. diagnosis of cancer (Ruiz et al., 2018), sleep apnea (Janott et al., 2018; Qian et al., 2017), diabetes (Pratt, Coenen, Broadbent, Harding, & Zheng, 2016), cardiovascular diseases (Goldberger et al., 2000), and assessment of psychological stress (Thelwall, 2017; Yoo, Lee, & Ha, 2019). It can also be used as a preventive and diagnostic method to identify 'red flag' situations in real-time, e. g. to help identify people prone to suicide, or those with mental conditions such as bipolar disorder (Amiriparian et al., 2019a), autism (Roche et al., 2018), depression (Schuller, 2016), undergoing pain (Lucey, Cohn, Prkachin, Solomon, & Matthews, 2011; Walter et al., 2013) or stress (Zhou, Hansen, & Kaiser, 2001). As discussed previously, an assistive technology recording dyadic conversations and estimating the psychological and physiological state of the recorded individuals can be envisioned.

However, in order for its widespread use in the health sector, it needs meet very high standards of requirements, aside from its explainability of its predictions. First, it needs to be robust for use in a non-laboratory, unconstrained, noisy, i. e. 'in-the-wild' conditions, with data featuring spontaneous behaviours. It should ideally capture nuances of emotions/affects of people of different backgrounds, rather than a crude classification into three (positive, negative, and neutral), or six or such tiny number of basic emotion classes. It should ideally track one's emotions continuously in time, in real-time. In summary, an explainable value-continuous, time-continuous, multidimensional, subject-independent, light-weight, real-time, physiological and affect prediction model trained on the in-the-wild data is desired; the scope of this paper.

1.3.1. Scope of this paper

In this paper, we unearth learnings of a shallowest possible CNN one can ever realise, as it learns to predict human emotions from in-the-wild videos. The training and predictions are of value- and time-continuous affect dimensions of individuals coming from very different cultures, contexts, gender and age-groups featured in the *SEWA* database, without making use of any audio or textual data. Inspired by these findings, we also explore use of this network topology for remote video-based pain monitoring on the *UNBC-McMaster Shoulder Pain Expression* database, and the *BioVid* database.

1.4. Organisation of the paper

In Section 2, we discuss each previous research directly relevant to our current study. We discuss in detail the performance metric we chose for evaluation in Section 3 and the datasets we used in Section 4, complete with statistical analysis of the features and the labels – which is crucial before beginning with the experiments. We detail next the experimental design pipeline, different types of models we tried, and how the powerful, yet shallowest-realisable CNN evolved through these experiments in Section 5. We present the insights gained by analysing the trained weights mapping the features to the output labels in Section 6. We re-evaluate our proposed approach for the time-continuous pain prediction problem in Section 7. We summarise our findings, mentioning briefly the limitations of this study in Section 8. We also list various avenues for future work as the logical next step – including the research paths we have already begun venturing into.

2. Related research

The target research problem here is the explainable, robust, value- and time-continuous recognition of affect dimensions (e. g. arousal and valence) on in-the-wild audiovisual recordings, featuring spontaneous behaviours in the conversational context. The publicly available databases (e. g. SMARTKOM (Schiel, Steininger, & Türk, 2002), IEMOCAP (Busso et al., 2008), RECOLA (Ringeval, Sonderegger, Sauer, & Lalanne, 2013), MAHNOB Mimicry (Bilakhia, Petridis, Nijholt, & Pantic, 2015), 4D CCDb (Vandeventer, Aubrey, Rosin, & Marshall, 2015)) do not typically meet this criteria, featuring either the non-spontaneous behaviours, and/or the laboratory recordings, and/or only the categorical and/or only the sample-level labels, and/or are too small Kossaifi et al.. The only exceptions are SEWA Kossaifi et al. and GRAS^2 (Eyben, Weninger, Paletta, & Schuller, 2013). The GRAS^2 database has arguably more Gabor effect-free recordings; likely to generate a more robust model (Pandit et al., 2018a; 2018b). However, its time-bound consent stands expired.

The 'Automatic Sentiment Analysis in-the-wild' (SEWA) corpus Kossaifi et al. has consistently featured in the 'Affect Recognition' sub-challenge of the Audio/Visual Emotion Challenge since a few years (AVEC 2017, 2018, 2019) (Ringeval et al., 2018a; 2018b; 2017). It is arguably the most popular in-the-wild public database available to date, featuring time-continuous, high resolution labels for multiple dimensions of affect. The participants of the AVEC challenges (Chen et al., 2019; Kaya et al., 2019; Zhao, Li, Liang, Chen, & Jin, 2019) compete to correctly predict these affect labels for different cultures, based on the audio, textual, and video features provided. The bag-of-words representation computed using openXBOW (Schmitt & Schuller, 2017) has been shown to perform well across all the three modalities.

In this paper, we combine and extend two of our previous research works. With Pandit, Schmitt, Cummins, and Schuller (2019b), we ventured into the territory of 'explainable AI' for time- and value-continuous in-the-wild affect predictions for healthcare. We investigated how a trained model uses only 521 text features to predict the three affect dimensions by computing their feature attributions. These 521 features represent the frequency of those specific 521 words in a 6 second window. Consistent to human perception, the model relies heavily on action markers such as <laughter> and <slight_laughter> to predict arousal and valence. Even more interestingly, the model reassigned high relevance to words implying contexts when predicting liking, e. g. 'dazu'/'außerdem' (therefore), 'endlich' (at last), 'über' (over (something)), 'Zusammenhang' (context), 'weil' (because). We also established

that the text-features themselves were informative enough that even a simple feedforward network utilised these features effectively to predict the affect dimensions.

In another research (Schmitt, Cummins, & Schuller, 2019), we questioned necessity of a recurrent neural network (RNN) for a value-continuous time-series prediction problem. While the RNNs learn the long-term dependencies in theory, they fail to do so in practice for some fundamental reasons (Hochreiter & Schmidhuber, 1997). RNNs with nodes with memory units (e. g. Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM)) alleviating this problem continue to suffer from the vanishing and exploding gradient problems (Pascanu, Mikolov, & Bengio, 2013). The CNNs utilise the adjacencies and contexts, while also allowing data parallelisation. CNNs are, thus, much faster to train and test (Pandit et al., 2019a). In Schmitt et al. (2019), we established that CNNs are as effective as RNNs, if not more, in predicting arousal and valence dimensions even in the cross-cultural context. In this paper, we use the video features to predict affect dimensions using CNNs similar to those used in Schmitt et al. (2019), and explain the model workings through feature attribution computation similar to Pandit et al. (2019b). We later test the approach on the time-continuous pain intensity prediction problem.

3. Choice of evaluation metric

The primary focus of this paper is the explainable, time-continuous prediction of the affective and physiological state of a subject under observation, e. g. a patient. Choosing a metric for evaluation of the predictive capability of the system is a crucial step. For a time-series prediction that is *value-continuous* (i. e. a regression problem), arguably the most popular performance metrics are: the Mean Squared Error (MSE), the Mean Absolute Error (MAE), the Pearson Correlation Coefficient (CC or ρ) and the Concordance Correlation Coefficient (*CCC* or ρ_c). Thus, given a bivariate population, i. e. the two time-series, $X{:}=(x_i)_1^N$ and $Y{:}=(y_i)_1^N$, following are the definitions for each of the metrics:

$$\textit{Error statistics:}\quad MSE = \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}, \qquad MAE = \frac{\sum_{i=1}^{N}|x_i - y_i|}{N}, \tag{1}$$

$$\textit{Covariability statistics:}\quad \text{Covariance, } \sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{N}, \tag{2}$$

$$\text{Correlation Coefficient, } \rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{3}$$

$$\text{Concordance Correlation Coefficient, } \rho_c = \frac{2\rho\sigma_X\sigma_Y \quad (=2\sigma_{XY})}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}, \tag{4}$$

$$\text{where,}\quad \mu_X = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu_X)^2}{N}}, \quad \mu_Y = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \mu_Y)^2}{N}}. \tag{5}$$

$$\text{Note that, we always have } \; -1 \le -|\rho| \le \rho_c \le |\rho| \le 1. \tag{6}$$

The first two (MAE and MSE) quantify the extent to which the two time-series differ. Neither translates to any meaningful inference without the prior knowledge of the magnitude of the variables. The two also fail to capture correlated variation of the quantities being measured (i. e. whether a greater value of one corresponds to a greater value of the other). While a covariance metric quantifies such relationships, this metric too is impossible to interpret without knowing the variable magnitudes. The normalised covariance metric, called the Pearson correlation coefficient, quantifies the strength of the linear relationship between two variables, ignoring the bias and the scale. The CCC metric goes a step further, and penalises any deviation from the identity relationship, i. e. the non-unity scaling and the non-zero bias. For a standardised, subject-independent, 'in-the-wild' health monitoring system, knowing and proving existence of a strong linear relationship is not enough. Rather, an accurate prediction (and not the scaled prediction) is the central necessity. For this reason, we use the most stringent 'CCC' metric as the evaluation metric.

4. The datasets

4.1. Overview in terms of the subjects, and the data-splits

The *SEWA* dataset features spontaneous dyadic internet-based conversations between the participants, discussing an advertisement they watched. The recordings were not standardised intentionally, and are truly in-the-wild. The participants were allowed to converse from wherever they wished (e. g. home, office, cafeteria), with no set requirement on the devices (notebooks, microphones, cameras) and internet connection in use. Many of the audiovisuals suffer from bad lighting conditions, noise, echoes and even frame freezing. The participants came from several age-groups, with various degrees of mutual acquaintance, and different cultural backgrounds. With 201 male and 197 female participants, and with nearly 60 pairs from each of the 6 linguistic populations (Chinese, English, German, Greek, Hungarian, Serbian), the dataset features behavioural data from diversified and balanced demographics.

Table 1
Participant count and duration for the data splits of the SEWA dataset for the AVEC 2019 challenge (Ringeval et al., 2018b), and for the experiments featured in this paper.

Culture	Partition	#Subjects	Duration (mmm : ss)
German	Training	34	093 : 12
German	Development	14	037 : 46
German	Test	16	046 : 38
Hungarian	Test	66	133 : 12
Chinese	Test	70	200 : 46
Total Test	**Test**	**152**	**380 : 36**
Total		**200**	**511 : 34**

Only the Chinese, German, and Hungarian recordings have been annotated, and are available for research (cf. Table 1). Note that in our experiments, the test split is huge compared to the training and validation splits.

As for the experiments relating to pain monitoring, we use the *UNBC-McMaster Shoulder Pain Archive* (Lucey et al., 2011) containing 200 video sequences with more than 48,000 labelled frames from 25 subjects suffering from shoulder pain problems. The subjects undergo eight standard motion tests; abduction, flexion, and internal and external rotation of each arm separately. The number sequences annotated for a subject vary from 2 to 16. We use 9 subjects (71 sequences) for training, and 7 subjects (68 sequences) as a validation set. The models are tested on the remaining 9 subjects (the remaining 61 sequences).

We also used the *BioVid database* (Lucey et al., 2011) featuring 87 subjects, subjected to 5 levels of pain using a different heat stimulus 20 times in a random order, each for 4 s minimum. The 5 pain stimulation levels were calibrated separately for each participant, ranging from no stimulation, to the 'beginning of feeling pain' (pain threshold), with a gradual increase to the unacceptable pain-level (tolerance threshold). We report the results for training to validation to test split ratio of 1:1:1. In the pain-related experiments, the main focus is to understand the learnings of the model and whether it is consistent to the provided definition of the pain label, thus the human understanding of the pain.

4.2. Overview of the labels and the features contained in the dataset

The SEWA dataset features value- and time-continuous labels for arousal, valence, and liking (i. e. extent liking for the advertisement is expressed), annotated by five to six annotators speaking the language of the participants of the audiovisuals they annotate. The annotators, therefore, can use each data modality available to them, namely, the linguistic content (what is expressed), the audio content (how it is expressed, e. g. prosody, tone, pitch), and the visual content (non-verbal cues, facial expressions). In the current research we make use of the FAU features extracted using OPENFACE[2] (Baltrusaitis, Zadeh, Lim, & Morency, 2018) (cf. Table 2).

For experiments related to the pain intensity prediction problem, we use the manually coded frame-level FAU features as an input, and the featured Prkachin and Solomon Pain Intensity (PSPI) (Lucey et al., 2011) labels as the prediction target. Likewise, the BioVid database contains manually coded FAU features and PSPI labels (PSPI_PA4, PSPI_SD) for 435 (87 subjects $\times$ 5 pain levels) from the total of 8 700 videos (87 subjects $\times$ 5 pain levels $\times$ 20 stimuli). Additionally, We extract FAU features for all of the 8 700 videos using the OPENFACE toolkit (Baltrusaitis et al., 2018), and generate the PSPI labels from the extracted features. The PSPI labels generated using the extracted FAUs are highly correlated with the PSPI_PA4, PSPI_SD featured in the dataset, and in turn, the subjective human-annotated pain intensity labels of the BioVid dataset.

4.3. FAU label statistics and insights for affect prediction

In the affect prediction study, we restrict our attention to employing the video-based features only, i. e. FAUs and the speaker-turn activation feature indicating that at least one participant is speaking. FAU is defined by the Facial Action Coding System (FACS) – a predefined set of different simultaneous facial muscle movements. We employ the 17 FAUs available to us, and the associated confidence scores. As discussed previously, the aim of the research is primarily to investigate how the neural networks utilise FAUs effectively for affect prediction, what makes their learnings generalise across different cultures, and if these learnings are consistent to human perception of emotion expressions.

To explore the utility and challenges presented by the extracted FAU feature set, we first prepare a statistical summary of the activation values of the individual features (cf. Fig. 2). These input features are identical to those provided in the AVEC 2018 and AVEC 2019 challenges, i. e. the moving average and the moving standard deviations of 17 FAU activations computed for consecutive 50 frames (1 s) with a hop of 5 frames (100 ms). Fig. 2 presents grouped box-plots of the corresponding moving statistics, for better comparison of the features across cultures. As can be seen from the plot, the ranges, quartiles, and outlier distributions are similar across the different cultures.

Considering that the three cultures are vastly different from each another, the consistent statistics of FAU activations across cultures – in some sense – hint at the universality of the FAU features (Russell, 1994). One could expect, therefore, a superior pan-

[2] https://github.com/TadasBaltrusaitis/OpenFace .

Table 2
Illustrations of what the extracted FAU activations look like. (Image credits: with permission from imotions.com/blog/facial-action-coding-system/).

Action Unit (AU)	Facial Muscles	Description	Example Movement	
			From	To
1	Frontalis, pars medialis	Inner Brow Raiser		
2	Frontalis, pars lateralis	Outer Brow Raiser		
4	Depressor Glabellae, Depressor Supercilli Currugator	Brow Lowerer		
5	Levator palpebrae superioris	Upper Lid Raiser		
6	Orbicularis oculi, pars orbitalis	Cheek Raiser		
7	Orbicularis oculi, pars palpebralis	Lid Tightener		
9	Levator labii superioris alaquae nasi	Nose Wrinkler		
10	Levator Labii Superioris, Caput infraorbitalis	Upper Lip Raiser		
12	pars palpebralis	Lip Corner Puller		
14	Buccinator	Dimpler		
15	Depressor anguli oris (Triangularis)	Lip Corner Depressor		
17	Mentalis	Chin Raiser		
20	Risorius	Lip stretcher		
23	Orbicularis oris	Lip Tightener		
25	Depressor Labii, Relaxation of Mentalis, Orbicularis Oris	Lips part		
26	Masetter; Temporal and Internal Pterygoid relaxed	Jaw Drop		
45	Levator Palpebrae Relaxation, Orbicularis Oculi Contraction, Pars Palpebralis.	Blink		

cultural performance of the models using the FAU features to predict affect. Because the median utility function is less affected than the mean by the outliers in the data, we consider the median heatmap column to be the better representatives of the features in the discussions next.

It can be inferred from the heatmap in Figure 2 that not all FAU features convey equally useful information. For example, we notice that:

- The close to zero standard deviations ($\in [0.00, 0.07)$) of all of the AU12 to AU45 activation statistics indicate that AU12 to AU45-related inputs remain almost constant across the entire dataset. Medians of these features are close to zero likewise ($\in [0.00, 0.05]$), implying that the constant activation, too, is very close to zero (< 0.05) likewise.
- The range for the moving-mean and moving-standard deviations are drastically different for different FAUs. For example, the maximum moving-mean activation for the AU1 to AU10 FAUs is close to 5.0, while the AU12 to AU45 moving-mean activations rarely go past 1.0 (cf. the boxplots).
- This dissimilarity between the different FAU features is so stark that even maxima (i. e. the outliers in one direction) of certain features (e. g. AU12_mean) are less than even medians of most other features. Such extremely high or low values, in the absence of a batch-normalisation layer, can drive prediction of a neural network astray – unless the features with stronger activations are more informative already. It is only then that a model can possibly assign quasi-zero weights to the erratic features.
- Interestingly, in spite of the featured high maximas and near-zero minimas (i. e. the high range) for AU 4, 6, 7, 10 -related statistics, the medians of the moving-standard deviations of activations are zero. As for the moving-standard deviations of activations, we note that more than 50% of the dataset is dominated by the zero-valued samples. This tells us that these four activations hardly ever fluctuate too much from their respective median values (.560, .835, .230, .100) in time.
- As for the remaining AU1, AU2, AU5, and AU9 activations, the medians of the moving-mean and the moving-standard deviations are (.000, .175), (.000, .440), (.603, .835), (.000, .235) respectively. As implied earlier, the first quantity represents roughly the degree of the FAU activation in time. The second quantity, i. e. the median of the standard deviation is a representative of the perturbation of that FAU activation. The latter is affected by both the magnitude and the frequency of perturbation.

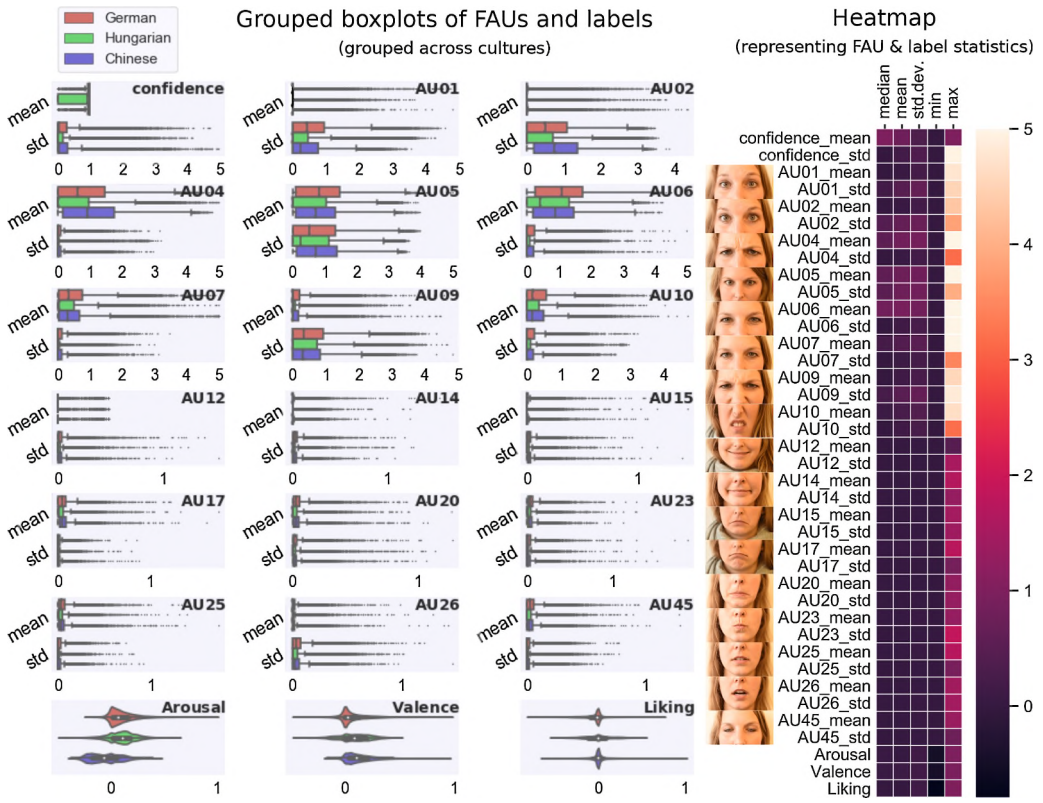


Fig. 2. Boxplot of FAU activation levels and the three affect labels. Note that the range (cf. x-axes) for different FAU features are drastically different, and only a few FAUs dominate over the rest (cf. max values). Some of the FAUs (e. g. AU12_mean) rarely ever manifest, even their maximum values are close to minimum of almost all other FAU activations.

5. Affect prediction: Experiment and model design pipeline

We train various minimalist CNN topologies using the Keras (v2.2.2) library with the Tensorflow (v1.11.0) backend. Training and evaluations are run on regular notebook with a Nvidia GeForce GTX 1050 Ti GPU-card. The FAU features obtained from 34 video chat sequences of German participants (cf. Figure 3) and the RMSprop optimiser (learning rate = 0.001) are used for training, which runs for 2 000 epochs. We choose the model weights based on their performance on the development set. Because an attempt at minimisation of *MSE* does not necessarily translate into maximisation of *CCC* Pandit and Schuller, we use the deviation from the maximally achievable *CCC* i. e. '1 − *CCC*' as the loss function of choice. This strategy of employing *CCC* as part of the loss function has proven to be successful in practice as well (Pandit et al., 2019b; Schmitt et al., 2019; Trigeorgis et al., 2016). We choose the optimal delay compensation of 4.0 seconds and 2.8 seconds for arousal and valence respectively, based on our previous study (Schmitt et al., 2019).

5.1. Minimalist model A

In Schmitt et al. (2019), we demonstrated the suitability of CNNs over RNNs for the high noise, high resolution time-series prediction problems. We begin with an identical CNN architecture, except that we use the 37 visual features instead of the 47 acoustic ones, consequently resulting in a huge reduction in the number of trainable parameters. An output neuron in this architecture has a receptive field of about 10 s ($\because$ 50 + 30 + 20 + 5-3 = 102 time-steps = 10.2 s, cf. Figs. 4 and 5).

5.2. Minimalist model B

We gradually remove the intermediate layers from the model A, while keeping about the same receptive field of the output neurons (cf. Fig. 5). While we experimented with number of different combinations of the number of intermediate layers (L_N), the receptive field of the intermediate layers (T_F) and the number of filters for each of the intermediate layers (N_F), the most important revelation came from the observation that changing the ReLu activation to a linear one does not deteriorate the model performance. Because the input to output mapping is reasonably linear, there is no inherent requirement of too many filters for each layer. Unless stated otherwise, we discuss the extreme case, where number of intermediate Conv1D layers = 1, and where $N_F = 1$, $T_F = 100$.

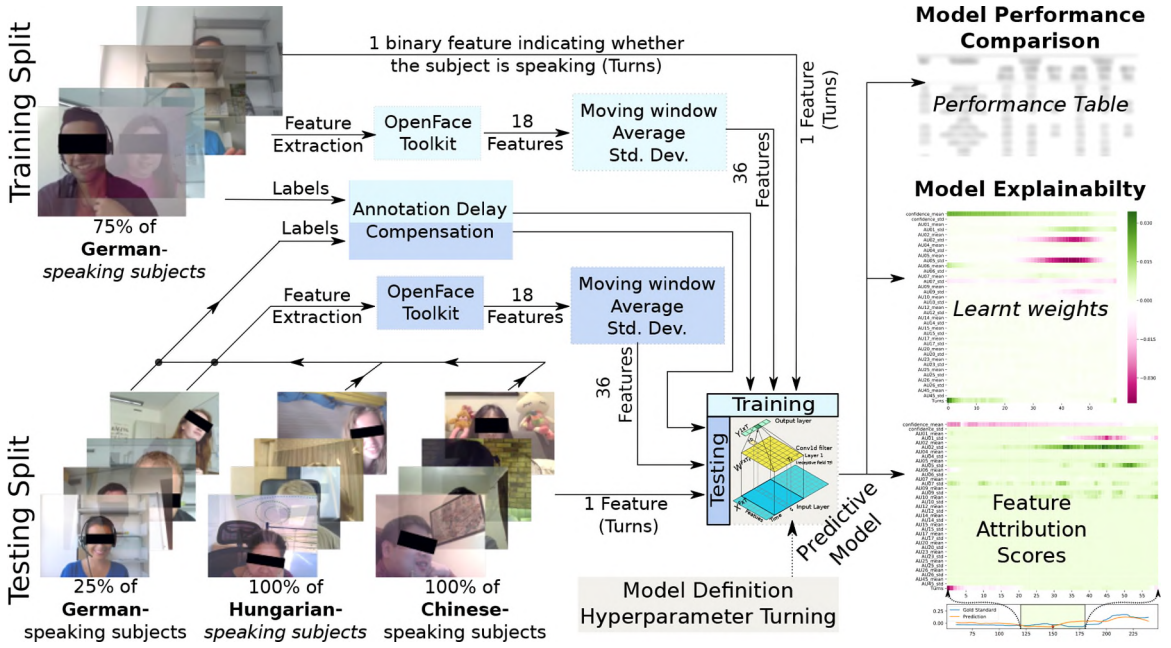


Fig. 3. The entire experimental design pipeline.

5.3. Minimalist model C

The last layer of the Model B (cf. Fig. 5) is a Time-Distributed Dense layer. Because the output of the previous layer is a $N \times 1768 \times 1$ matrix (as we use only 1 filter), this dense layer only scales every output by a constant factor adding a constant offset. Because the relationship continues to be linear, and because a neural network, by definition, learns weights (scalings) and biases (offsets), this last dense layer is redundant. However, removing the dense layer (thus, resulting in Model C) deteriorates the performance. This is not exactly surprising, since an explicit post-processing step, instead of a linear dense layer at the output, is often witnessed in the literature (Schmitt et al., 2019; Trigeorgis et al., 2016). The newly reduced topology leads us to the model D where, while maintaining minimalism of model C, we also ensure the superior predictive performance of the trained model B.

5.4. Minimalist models D

If W_L^M and B_L^M are the weights and biases of the layer L of model M, and I and O are the inputs and outputs of the model, we have:

$$O = (W_{Conv1D}^{Model_B} * I + B_{Conv1D}^{Model_B}) * W_{Dense}^{Model_B} + B_{Dense}^{Model_B} \tag{7}$$

$$O = (W_{Conv1D}^{Model_D} * I + B_{Conv1D}^{Model_B}) \tag{8}$$

Comparing Equation (7) and Equation (8), we have:

$$W_{Conv1D}^{Model_D} = W_{Conv1D}^{Model_B} * W_{Dense}^{Model_B}, \tag{9}$$

$$\text{and} \quad B_{Conv1D}^{Model_B} = B_{Conv1D}^{Model_B} * W_{Dense}^{Model_B} + B_{Dense}^{Model_B} \tag{10}$$

5.5. Minimalist model E

In an attempt to make the model computationally even more inexpensive, we experiment with a few input feature combinations. This feature selection, too, is data-driven, gaining insights from the visualisations of the filter weights and the feature attribution heatmaps, as we discuss in Section 6.

5.6. Minimalist model F

Likewise, we note that the most of the high magnitude contributions come from the top right quarter of the feature attribution matrix, indicating a likely correction in the annotator delay compensation. We have verified these findings – i. e. regarding the consistent correction in the annotator delay compensation – by advancing the training labels even more (by 40 samples for the model represented in Figure 6) and by rolling over the filter coefficients likewise (by 40 samples) for retraining of the models. The exact

tabulation of CCC scores across different receptive field values and different annotation delay compensation values is a work currently in progress.

6. Affect prediction: Results and insights

6.1. Feature attribution calculation

Consider model D consisting of only one intermediate Conv1D layer between the input and the output layer. Because the output is just a single channel (we predict only one emotion dimension at a time), the number of filters associated with the last and the only intermediate Conv1D layer is one. Each output is the element-wise product of trained filter weights and a section of the input matrix, added to the trained bias value. Thus, the Conv1D layer represents the degree of similarity of the input features (in an interval equal to its receptive field) against the pattern described by the weights of the trained filter.

Let the inputs and outputs of the model, and the filter weights and the bias of Conv1D be the matrices $X^{F \times T}$ $Y^{1 \times T}$, $W^{F \times T_F}$, and the scalar b respectively.

$$
\begin{aligned}
&\therefore \; X^{F \times T} = \{x_{f,t}\}, && \text{with } f \in [0,\, F-1],\, t \in [0,\, T-1],\\
&\quad\; Y^{1 \times T} = \{y_t\}, && \text{with } t \in [0,\, T-1].\\
&\quad\; W^{F \times T_F} = \{w_{f,t}\}, && \text{with } f \in [0,\, F-1],\, t \in [0,\, T_F-1].\\
&\therefore \; y_{t_0} = g\left(\left[\sum_{f=0}^{F-1}\sum_{t=0}^{T_F-1} x_{f,t_0+t} \times w_{f,t}\right] + b\right) && \text{with } g(\cdot) \text{ as the activation.}\\
&\therefore \; y_{t_0} = \left[\sum_{f=0}^{F-1}\sum_{t=0}^{T_F-1} x_{f,t_0+t} \times w_{f,t}\right] + b && \text{for the linear activation.}
\end{aligned}
\tag{11}
$$

Distributing the constant bias b across the associated input features evenly, we get the contribution to the output y_{t_0} by the feature x_{f,t_0+t}, i. e. the feature attribution score S_{f,t_0+t} as:

$$
S_{f,t_0+t} = \left[x_{f,t_0+t} \times w_{f,t}\right] + \frac{b}{F \times T_F}. \tag{12}
$$

If the multiplication of the weight matrix and the input sub-sequence generates the output at a time-step that is central to the input sub-sequence, e. g. as in Keras implementation (cf. Fig. 6(a)), the contribution by x_{f,t_0+t} is:

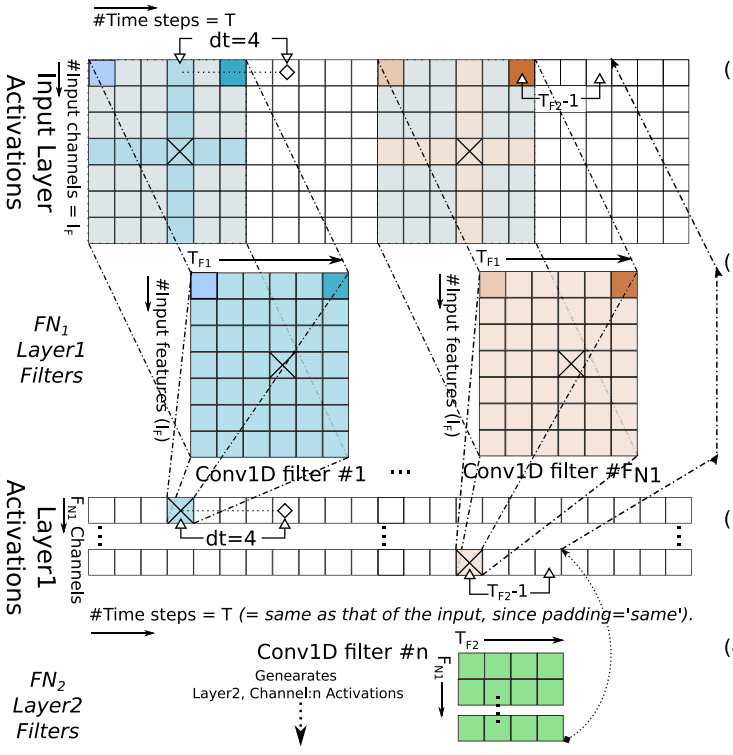
$$
S_{f,t_0-int\left(\frac{T_F}{2}\right)+t} = \left[x_{f,t_0-int\left(\frac{T_F}{2}\right)+t} \times w_{f,t}\right] + \frac{b}{F \times T_F}. \tag{13}
$$

We present insights gained by inspecting the feature attribution heatmaps at various time-steps (t_0) for the arousal and valence labels, for several subjects. Fig. 6 presents an example.

6.2. Interpreting the feature attribution matrix and the filter weights

As discussed, one interpretation of the filter weights is that the model seeks an input pattern that is most similar to the filter weights. The more similar the input pattern, the higher is the output activation. Because the output is ultimately a weighted linear combination of the filter coefficients (weighted by the input activations), the output can be regrouped into several column-wise or several row-wise summations. We make the following observations.

- For the strictly non-negative feature activations (e. g. inputs in this study, the FAU activations), the sign of a weight indicates whether a certain feature activation is positively or negatively correlated with the output.
- A dominant *row* (featuring high magnitude weights) amplifies the corresponding activations heavily; suggesting a high importance of the feature.
- This interpretation needs a careful reconsideration in the case of non-normalised inputs. Through high magnitude weights, the model also ensures that the low activations are represented enough to drive the output.
- A row also represents a temporal trend of the specific feature activation in the input matrix that the filter expects to see. More similar an input row (a trend of a specific feature activation) is to the corresponding row in the weight matrix, more is the cosine similarity between the two vectors, driving the output to be more positive. Essentially, an output represents the extent to which the input encapsulates the various expected temporal trends of the individual feature activations (cf. Fig. 7).
- Likewise, a *column* represents a specific feature activation combination. As for the FAUs, an example combination could be 0.71 activation of lip corner puller (a smile), 0.51 cheek raiser, and so on, mapping to arousal label of 0.83, through weighted summations (cf. Fig. 7).
- Thus, the output represents the extent to which an input encapsulates the various expected feature combinations at specific time-



(1) Displacement of the filter by 'dt' time-steps results in the output displaced in time by 'dt' time-steps.

(2) On the other hand,
receptive field of every filter of Layer1
= T_{F1}
= Receptive field of the Layer1 neurons.
= Number of timesteps of the previous layer, i.e., number of timesteps of the input layer that a Layer1 neuron captures the trend of.

(3) Thus, effective receptive field of the next layer (i.e., of the Layer2) neurons
= $T_{F1} + (T_{F2} - 1)$. (From (1) and (2)).

(4) When L number of Conv1D layers are stacked, effective receptive field of the output neuron
= $T_{F1} + (T_{F2} - 1) + (T_{F3} - 1) \ldots + (T_{FL} - 1)$.

Fig. 4. Receptive field of the output neuron when L number of Conv1D layers are stacked $= \left(\sum_{i=1}^{L} T_{Fi} \right) + 1 - L$, where T_{Fi} is the receptive field of the filter in the i^{th} Conv1D layer.

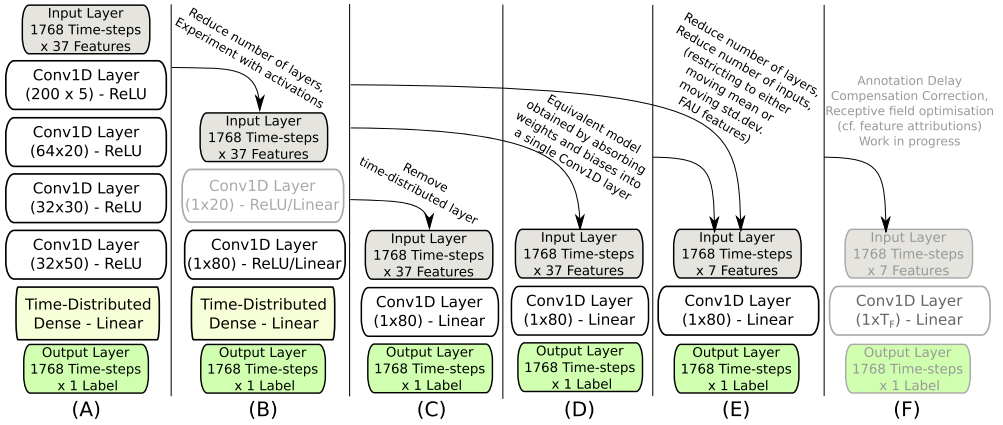


Fig. 5. Evolution of the models. Model 'A' has a topology similar to Schmitt et al. (2019). We derive the rest by reducing the number of layers, the number of filters and the number of input features. Feature selection is achieved by visualising feature attribution matrices.

steps, relative to the centre of the filter – moving across in the temporal dimension.

- The filter-weights alone do not quantify the relevance of individual features, since the scale of the input features is a crucial missing context, but the feature attribution scores do.

The early feature relevance score matrices were hard to interpret, as these consisted of too many positive and negative scores. We added L_1 regularisation loss, compelling the model to use a lot sparser filter weight matrix. Without much of a loss in predictive performance, this results in a sparser feature attribution matrix, which is a lot easier to analyse for a human and to gain insights. We use heatmaps of seaborn library with a diverging PiYG colormap centred at 0.0 to visualise the matrices, irrespective of the range of the scores.

We experimented with the receptive field for the model D, and yet:

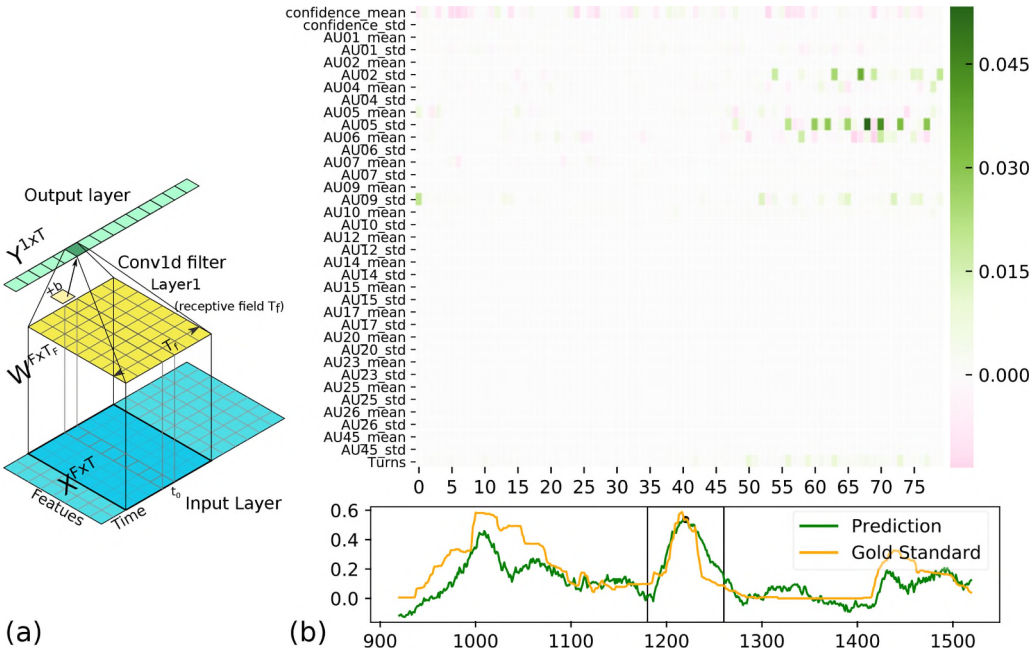


Fig. 6. (a) Product of the filter weights and the input sub-sequence generates the output at a time-step that is central to the input sub-sequence. (b) An example feature attribution matrix S (upper subplot), the summation of the elements of matrix S (the black circle in the lower subplot), the gold standard and the prediction sequences (the lower subplot). The plot generation helps us verify that $y_{t_0} = \sum_{f=0}^{F-1} \sum_{t=0}^{TF-1} S_{f,t_0-int\left(\frac{T_F}{2}\right)+t}$. For SEWA, the bottom half of S is typically quasi zero-valued, and most of the contributions to the output come from the top right quarter of S, suggestive of a correction in the annotation delay compensation.

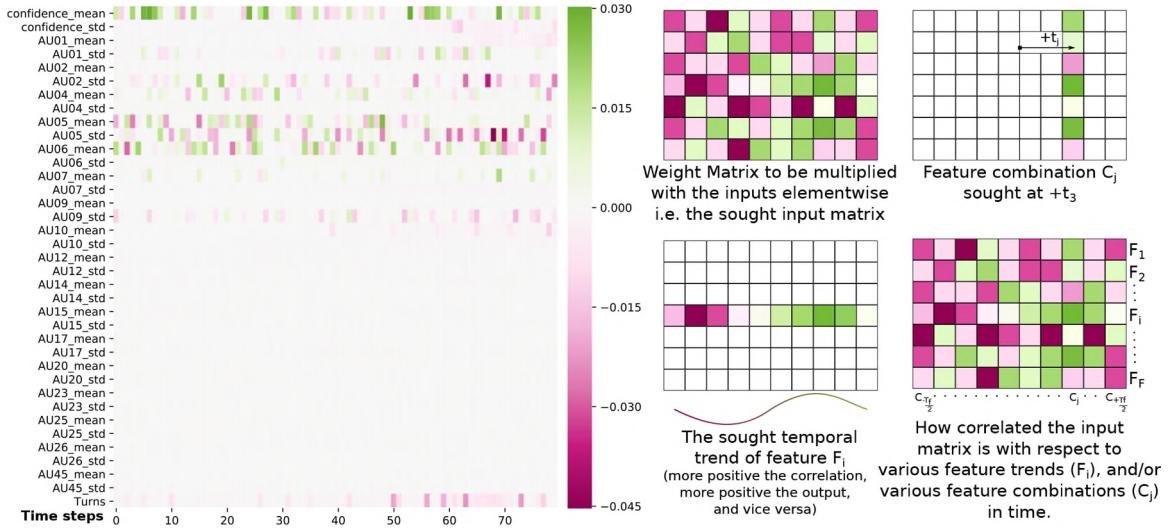


Fig. 7. The filter weight obtained for the 37-input arousal model (the model D in Fig. 5), and several ways of interpreting the filter weights.

- The model consistently relied on the following features, 1. AU02_std, 2. AU05_std, 3. Turns, and 4. Confidence_mean.
- A few other FAUs, namely AU01, AU04, AU09, and AU10 were also found to be seldom making non-zero contributions to the prediction.
- The contributions coming from the rest of the features were consistently close to zero, inspiring us to investigate data-driven feature selection.
- The majority of the contributions come from the right half of the matrix, suggestive of a correction in the annotation delay compensation.

Table 3
Performance comparison of the trained models in terms of CCC. Each of the models was trained using an identical Training and Development set featuring only the German-speaking subjects. The trained model was tested on 16 German(DE)-, 66 Hungarian(HU)- and 70 Chinese(ZH)-speaking participants. The results presented here are entirely reproducible using the scripts made available at https://github.com/vedhasua/tiny_emotionet. For the sake of completeness, we also note the state of the art performance on the AVEC 2019 Chinese test-split as reported by the challenge winners. As these models use recurrent neural networks, data parallelisation is not possible. These models also necessitate much complex multimodal feature extraction steps, followed by explicit early- and/or late- fusion steps, and were trained using a much larger training and validation data from subjects from two cultures (Ringeval et al., 2018b), as opposed to ours which used the data from only 48 German-speaking subjects for training and validation (cf. Table 1).

Model Topology	Affect Dimension	Test Splits		
		DE	HU	ZH
Similar to Schmitt et al. (2019): 37 Features, [200x5] Relu, [64x20] Relu [32x30] Relu, [32x50] Relu [1] Linear	Arousal	0.578	0.393	0.438
	Valence	0.573	**0.477**	0.315
Model D (cf. Fig. 5): 37 Features, [80x1] Linear, [1] Linear	Arousal	**0.606**	0.525	0.367
	Valence	**0.625**	0.420	**0.451**
Model E (cf. Fig. 5): 19 Features, [80x1] Linear, [1] Linear	Arousal	0.566	0.532	**0.481**
	Valence	0.611	0.422	0.431
Model E (cf. Fig. 5): 7 Features, [80x1] Linear, [1] Linear	Arousal	0.569	**0.534**	0.337
	Valence	0.601	0.390	0.430
AVEC 2019 Winning Submissions				
Chen et al. (2019): Data Augmentation, ResNet-based feature extraction, Early- and Late-Fusion, DBLSTM	Arousal	Not Applicable, Used in training		0.513
	Valence			0.515
Kaya et al. (2019): PCA of eGeMAPS, Early fusion with FAUs, [200,100,200] GRU-RNN, Weighted late-fusion	Arousal	Not Applicable, Used in training		0.466
	Valence			0.499
Zhao et al. (2019): Multimodal feature extraction, Adversarial Domain Adaptation, LSTM, fully connected	Arousal	Not Applicable, Used in training		0.400
	Valence			0.471

6.3. Feature selection experiments and insights

As discussed in the concluding remarks of the previous section, because the moving mean statistics almost always never contribute to the output (barring a lone exception of confidence_mean), we reconfigured the model to consume only the moving standard deviations and the turns information, reducing the number of input features from 37 to 19, thus halving simultaneously also the number of learnable parameters (cf. Model E of Fig. 5). For the receptive field equalling 80 time-steps, the number of learnable parameters get reduced down to 1 523 ($80 \times 19 + 3$) from 2 963 ($80 \times 37 + 3$), i. e. 51.4%.

Based on the feature attribution scores of this derived model, we further reduce these down to only 7 features, namely: Confidence_mean, AU01_std, AU02_std, AU04_mean, AU04_std, AU05_std, and Turns. Consequently, the learnable parameters reduce down to merely 563 ($80 \times 7 + 3$), i. e. 19% of the initial 2 963 learnable parameters. By optimising the delay compensation and the receptive field of the Conv1D filter, the number of learnable parameters can be reduced even further as discussed in Section 5.6 (cf. Model F in Fig. 5). We note that the majority, i. e. five of the seven features are FAUs, all of which relate to the eye movement and not the mouth, nose or any other facial features. The Table 3 lists the selected of the several model topologies we built and their corresponding predictive performance on the test splits.

7. Pain intensity prediction

Inspired by predictive performance of the minimalist CNN we proposed to model emotional state of the human subjects cross-culturally, that uses only a handful FAU activations recorded in the noisy conditions, we now turn our attention to the pain intensity prediction problem – equally relevant for many healthcare applications, including the automated remote patient monitoring system. An interpretable and explainable AI continues to be the primary focus of the experiments. One of the popularly used pain intensity metric is the Prkachin and Solomon Pain Intensity (PSPI) Scale (Lucey et al., 2011) given by the following equation:

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + binary(AU43). \tag{14}$$

While we know that there exists a direct mapping from a single feature vector to the corresponding PSPI label, we do not use the available feature vectors in isolation. Instead, consistent to the original problem statement, we treat this as the *time-series* prediction problem. We add, therefore, a temporal dimension to our feature context, making the non-linear mapping slightly more challenging. That is to say, ideally, the trained model should learn to completely ignore the neighbouring feature vectors, and use specifically only the lone feature vector at a relevant time-step, and only the set of relevant activations. But because the neighbouring feature vectors correspond to consecutive frames of a video sequence, they are largely correlated with one another. In the case of UNBC-McMaster Shoulder Pain Archive, even this assumption is not necessarily true, as we do observe the sudden spikes in the manual annotations, consistently representing the FAU onsets. The *max* function makes the FAU to PSPI mapping non-linear. We explore whether the proposed minimalist model is able to learn the challenging target mapping using linear activations alone.

Table 4
Statistics relating to UNBC McMaster Shoulder Pain Archive data splits used.

Split	Num. Subjects	Num. Sequences		Num. Frames = Num. Feature vectors (Sequence length)					
		Total	Valid	Total	Mean	Median	Std.Dev.	Min	Max
Training	9	71	47	17,353	244.41	217.0	114.73	75	683
Validation	7	68	45	14,809	217.78	202.5	90.51	68	518
Testing	9	61	29	16,236	266.16	240.0	105.94	48	495
Total	25	200	121	48,398	241.99	225.0	106.12	48	683

7.1. UNBC McMaster shoulder pain archive: Experiments and insights

The database contains 200 FAU and PSPI sequences for 25 subjects, amounting to 48 398 feature vector-label pairs in total. The statistics of the training, validation and test splits are reported in Table 4. We note the large variance in sequence lengths. We also note that the three splits we chose through iterations are fairly evenly distributed in terms of sequence lengths, numbers of sequences and subjects. The database features not only the FAU activations 4, 6, 7, 9, 10, 12, 20, 25, 26 and 43 (Lucey et al., 2011), but also the FAU 0, 15, 27 and 50 activations. The feature matrices are sparse, with many zero-valued rows. The number of zero-valued feature vectors in the training, validation and testing split are 13 423, 11 739 and 11 094 respectively (i. e. nearly 75% of the frames).

The only other annotations available in the database are the sequence-level self-reported and observer-reported pain ratings, and the frame-level Active Appearance Model (AAM) facial landmark points. In the interest of explainable AI, we use the FAU activation sequences as the inputs, assisting training of a lot sparser filter-weight matrix. We train the models to predict PSPI levels, with varying receptive fields for the output neurons. Just as one would expect, typically:

- The predictions improve with the reduction in receptive field if every other hyperparameter, including number of epochs is kept constant. This is somewhat expected, since a single feature vector governs the output at any time-step, albeit non-linearly.
- As a corollary, the model takes more epochs to arrive at a better generalising mapping i. e. higher performance on the unseen test partition, when the receptive field is larger.
- Even with large number of epochs of training, a model with receptive field of 1 time-step performs consistently and significantly better than the one with a larger receptive field of 60 time-steps (cf. Fig. 9, p-value $<$ 0.01 and $<$ 0.001 for 'UNBC McMaster' and 'BioVid' databases respectively).
- The model relies mostly on only the 6 relevant FAU activations out of 14, occurring at precisely the time-step at its centre (cf. Fig. 8(a)).
- The model has been observed to assign comparable weights to some 'irrelevant' activations at non-central time-steps, likely in an attempt to correct the errors due to non-linearity of the target mapping, to help fine-tune the predictions, e. g. AU20 in Fig. 8(a).
- While the weights assigned to most other activations are largely close to zero; these tiny non-zero weights act as low-pass filters for the output time-series; an effect especially observed at the PSPI onsets (cf. Fig. 8(b)). For multiple models with receptive fields

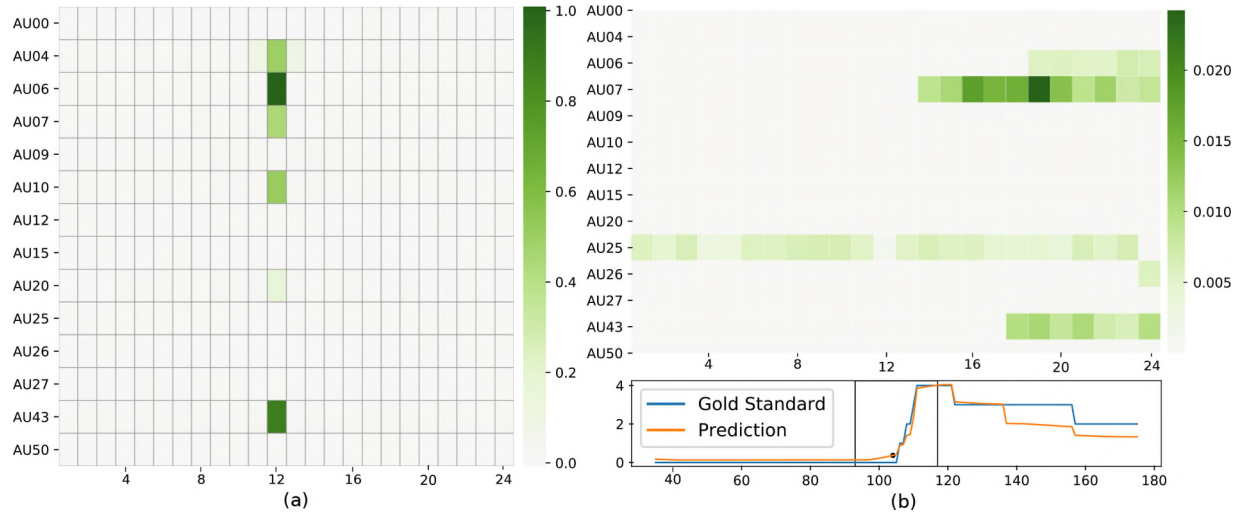


Fig. 8. (a) W_f trained using UNBC-McMaster Shoulder Pain Archive. We observe the high weights assigned to mostly the relevant activations (AU 4, 6, 7, 10, 43) at the relevant time-step even though the receptive field ($= 24$) is far larger than 1. (b) The non-zero tiny contributions from other FAU activations (e. g. AU 25, 26) at a non-central time-step becoming the most relevant in predicting low PSPSI value. Notice the resulting low pass filter effect at the PSPI onset.

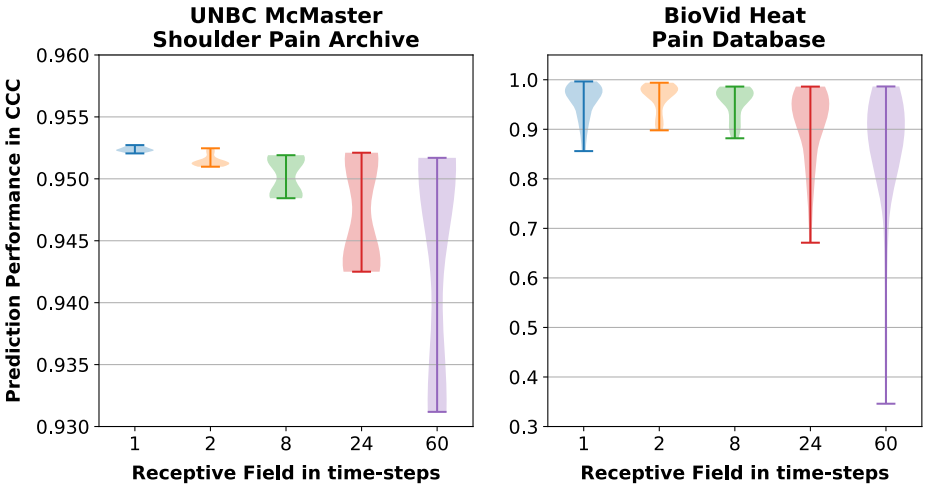


Fig. 9. Violinplot of the predictive performances of the models with different T_F.

other than 24 time-steps, we witness, in principle, very similar assignment of filter weights.

We are able to achieve *CCC* greater than 0.8, irrespective of the receptive field we use for the intermediate Conv1D layer. However, such a high performance means a little, and needs to be looked at carefully – especially when many of the entire output sequences are constantly zero-valued. We find that for the database contained 121 sequences in total. For the data-splits we made, these were divided into training, validation and testing split as 47 (out of 71), 45 (out of 68), and 29 (out of 61) zero-valued sequences respectively. To more stringently and more meaningfully evaluate the trained models, we get rid of all of the zero-valued input matrices, and test using the rest of the test split. Still, the CCC consistently turned out to be greater than to 0.8 with enough epochs of training.

7.2. BioVid database: Experiments and insights

Likewise, we train and test our models on part A of the BioVid Database featuring 8 700 video recordings (87 subjects $\times$ 5 pain levels $\times$ 20 stimuli) as discussed in Section 4.1 (Walter et al., 2013; Werner, Al-Hamadi, & Walter, 2017). Only 435 out of 8700 have been manually annotated with FAU activations, and two novel sequence-level pain annotations as a measure of 'expressiveness' of the subject (Werner et al., 2017). They compute first the frame-level PSPI values for the 435 videos (87 subjects $\times$ 5 pain levels $\times$ 1 chosen stimulus out of 20). They then extract maximum PSPI in each of the five video sequences for every subject, as the representative PSPI of that sequence. The normalised standard deviation of the five maximum PSPI values (PSPI_SD) per subject, and the maximum PSPI for the sequence with the maximum pain stimulus (PSPI_PA4) are considered to be measure of 'expressiveness' of the subject in response to the pain stimulus (Werner et al., 2017) The PSPI_SD and PSPI_PA4 are reported to be highly correlated (Pearson correlation coefficient $>$ .63) with the subjective annotations coming from human observers. This expressiveness measure has been proven to be useful to meaningfully split the data, improving performance of the pain stimulus prediction (Werner et al., 2017). In other words, late fusion of the two models – one for more expressive, and the other for less expressive subjects, based on individual expressiveness measure – has been shown to perform better in predicting the pain stimulus (e. g. pain threshold: BLN, tolerance threshold: PA4, intermediate stimuli: PA2, PA3) they were subjected to.

In our experiments, we extract FAU activations for all of the 8700 videos using OpenFACE toolkit, and consequently the frame-level PSPI values. The OpenFACE toolkit does not extract AU43 (eyes closed), but AU45 (blink) instead. In total we have 41 feautures, namely the 17 FAU intensities, 18 binary values depicting FAU presence, and 6 pose related features. We generate the PSPI labels with the following equation.

$$PSPI = AU04_r + max(AU06_r, AU07_r) + max(AU09_r, AU10_r) + AU45_c. \tag{15}$$

Each video sequence is 5.5 seconds in duration, featuring 138 frames. The standard deviation of the maximum PSPIs (PSPI_SD) represents extent of variations in expressions in response to the pain. This measure takes into account subject's responses to all of the pain stimuli, and is not based on just one stimuli of PA4. We compute PSPI_SD_New using the newly extracted 435 $\times$ 138 frame-level FAU activations and PSPI labels. We note that this newly obtained subject-level PSPI_SD_New is highly correlated with PSPI_SD metric (ρ = .6542), with PSPI_PA4 metric (ρ = .6463), and the human annotator rating (ρ = .7056)[3]. Thus, the newly computed PSPI sequences can be considered to be consistent with the labels in the original dataset. While the database originally featured manually annotated PSPI labels for only 435 videos, we now extend the database to 8700 videos, i. e. 8 700 $\times$ 5.5 $\times$ 25 = 1 196 250 feature-

[3] Reproducible using scripts available at https://github.com/vedhasua/tiny_emotionet .

label pairs.

With the workflow mostly remaining identical to the one described in Section 7.1, we evaluate the proposed model on the BioVid dataset as well, with one exception that relates to scaling of the input features. We standardise the input features, using the scaling and offset factors learnt from the training split. This is because the ranges for the individual features (AU activations and head-pose) are drastically different. We also check if the model decisionss continue to remain interpretable, and are consistent to with our human understanding, the known mapping. Just as we saw in the case of UNBC-McMaster Shoulder Pain database, we get high predictive performance with model's dependence on mostly the most relevant features at many of the time-steps (cf. AU45 in Fig. 10(a)). However, contrary to the expectations we otherwise would have from the filter weights (cf. Fig. 10(c)), the model was also seen to make creative use of non-relevant activations at certain time-steps (cf. Fig. 10(b)) for predicting low target values. Analysing occurrences of such 'spurious' or unexpected feature attribution scores (e. g. Figs. 8(b) and 10(c)), we note that these occur usually for predicting relatively small target outputs. Therefore, it is likely that these less influential weights assigned to the less relevant inputs are used to drive the output to a general bare minimum output, that is typically obtained in the absence of the most relevant activations. Because the FAU activations are strictly non-negative, an absence of the more relevant FAUs such as AU 4, 6, 7, 9, 10 (i. e. zero-valued input) translates to corresponding re-scaled inputs becoming negative input upon standardisation of the input matrix. This explains the negative contributions from AU 4, 6, 7, 9, 10 in predicting a small-valued label close to 1.0 in Fig. 10(c). The presented models are largely transparent; i. e. the inner workings are typically well-understood.

8. Conclusions

Towards explainable, robust, value- and time-continuous, real-time affect prediction on social media-based, web-assisted 'in-the-wild' video chat sessions, we presented a shallow CNN-based model consisting of a single one dimensional convolutional layer. Through statistical analysis of the input features, we investigated as to how and why the model assigns the filter weights the way it does. Because we used linear activation, we computed the feature attribution scores of the individual features for every model (i. e. the arousal and valence predicting models, consuming 37, 19 or 7-dimensional input feature vectors) and derived the most relevant FAU features. Thus, we effectively applied the computed feature attribution scores for feature selection and annotator delay compensation, from which new models evolved. Because there are currently only a few hyperparameters, namely the delay compensation, receptive field, and feature indices), in our planned future work, we intend to use genetic algorithms to automate the process of model learning, and arrive at even more computationally inexpensive and better performing affect-predicting models. To account for the ambiguity and simultaneous existence of multiple emotions (Gibaja & Ventura, 2015), the models could be retrained by incorporating ambiguity-aware emotion labels Sethu et al..

In spite of the observation that all of the models end up utilising primarily 7 out of 37 features, one can not definitively claim that the remaining 30 features are not informative enough in general. The reason for under-utilisation of 30 features is the inherent skew in the entropy across different features, that could be specific to the SEWA database. Also, because only 7 features were utilised at this point, we would refrain from making any strong claims in regards to consistency of the model with respect to human utilisation of these features to predict affective behaviours. The 7 emphasised features were all derived from different eye and eyebrow activity-related FAUs, which is consistent with the human emotion expression (Cavé et al., 1996). Because we have modelled the input to an output relationship using only the linear activations, methods utilising non-negative or sparse matrix factorisation are very likely directly relevant (Trigeorgis, Bousmalis, Zafeiriou, & Schuller, 2017).

As we have shown through statistical analysis, the 30 derived features were not informative enough, but this is likely specific to the AVEC 2019 challenge dataset. Different FAUs span over different timescales (Valstar & Pantic, 2006), and 1 second moving window-average and -standard deviation features might not capture the sporadic, sparse, yet certain highly informative facial muscle activations (e. g. eye-roll) that effectively reduce down to the zero-valued utility functions thanks to the considered 1 second moving window. A possible scope for a future study is to train models on the raw FAU features without any moving window-based statistics. The correlation analysis for the individual FAUs, similar to Sargin, Yemez, Erzin, and Tekalp (2007), is likely to help shed light on how decoupled these individual features are. Also, it should be noted that a one-dimensional convolution operation is inherently a statistical summarisation over a moving window, and the weights of the model offer more degrees of freedom for this implicit feature extraction compared to the plain moving average. As a result, the feature attribution scores for the raw FAU features will help establish better understanding of human perception of emotion expression – similar to our previous study on textual features.

We devised and trained the shallow models that achieved good predictive performance on German, Hungarian, and Chinese-speaking subjects, comparable to that of the state of the art models. Further, we demonstrated the versatility of the proposed approach through a set of experiments on two pain prediction corpora. The inherent transparency of the models allowed us to verify the model assigned high weights to FAU's relevant for pain detection. In future work relating to these expirements, we plan to test the approach on across other modalities, and explore it's usefulness as an adaptive, explainable feature fusion methodology.

Finally, our proposed network architecture uses the features that can be computed in real-time, and is computationally very inexpensive due to its size and small number of learnable parameters. The proposed models, therefore, can be easily integrated into devices such as smartglasses, or CCTV cameras. The study marks likely the first attempt at real-time, explainable AI for purely video-based time- and value- continuous affect prediction, pushing the frontiers of the research in advanced surveillance and healthcare monitoring systems.

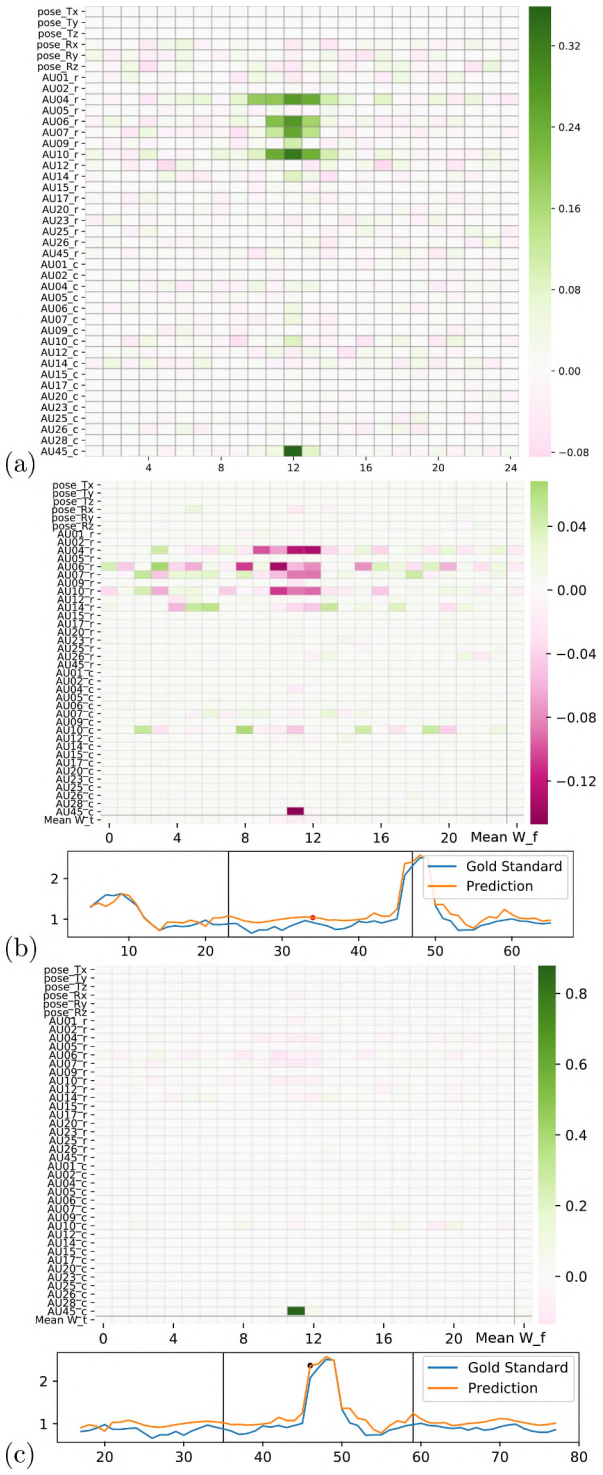


Fig. 10. (a) W_f trained with the BioVid database, (b) S matrix where the re-scaled FAU 4, 5, 6, 7, 9, 10 and 45 activations (consequently, can be negative) contribute negatively to the output. (c) The high AU45_c attribution, consistent to Eq. (15).

CRediT authorship contribution statement

Vedhas Pandit: Conceptualization, Methodology, Software, Validation, Data curation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Maximilian Schmitt:** Methodology, Software, Data curation, Writing -

review & editing. **Nicholas Cummins:** Writing - review & editing, Resources. **Björn Schuller:** Writing - review & editing, Resources, Project administration, Funding acquisition, Supervision.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2020.102347

References

Amiriparian, S., Awad, A., Gerczuk, M., Stappen, L., Baird, A., Ottl, S., & Schuller, B. (2019a). Audio-based recognition of bipolar disorder utilising capsule networks. Proc. 32nd International Joint Conference on Neural Networks, IJCNN, 1–7INNS/IEEEBudapest, Hungary.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., & Kindermans, P. J. (2019). iNNvestigate neural networks. *Journal of Machine Learning Research, 20*(93), 1–8.

Amiriparian, S., Schmitt, M., Hantke, S., Pandit, V., & Schuller, B. (Schmitt, Hantke, Pandit, Schuller, 2019b). *Humans inside: Cooperative big multimedia data mining. Innovations in big data mining and embedded knowledge: Domestic and social context challenges Intelligent Systems Reference Library (ISRL), vol. 159*, Springer235–257.

Antheunis, M. L., Tates, K., & Nieboer, T. E. (2013). Patients' and health professionals' use of social media in health care: Motives, barriers and expectations. *Patient Education and Counseling, 92*(3), 426–431.

Armfield, N., Gray, L. C., & Smith, A. (2012). Clinical use of skype: A review of the evidence base. *Journal of Telemedicine and Telecare, 18*(3), 125–127.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). Openface 2.0: Facial behavior analysis toolkit. 13th Intl. Conf. Automatic Face & Gesture Recognition, FG'18, 59–66, IEEEXi'an, P. R. China.

Bilakhia, S., Petridis, S., Nijholt, A., & Pantic, M. (2015). The MAHNOB mimicry database: A database of naturalistic human interactions. *Pattern Recognition Letters, 66*, 52–61.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., & Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*(4), 335.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. Proc. 4th Intl. Conf. Spoken Language Processing. ICSLP'96, 4, 2175–2178 IEEE Philadelphia, PA.

Chen, H., Deng, Y., Cheng, S., Wang, Y., Jiang, D., & Sahli, H. (2019). Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. Proc. 9th Intl. Workshop on Audio/Visual Emotion Challenge, AVEC'19, 27th ACM MM, 19–26, ACM Nice, France.

Eyben, F., Weninger, F., Paletta, L., & Schuller, B. (2013). *The acoustics of eye contact – Detecting visual attention from conversational audio cues. Proc. 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction, GAZEIN, 15th ICMI*. Sydney, Australia: ACM7–12 .

Gibaja, E., & Ventura, S. (2015). A tutorial on multilabel learning. *ACM Computing Surveys (CSUR), 47*(3), 52 .

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220 .

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Janott, C., Schmitt, M., Zhang, Y., Qian, K., Pandit, V., Zhang, Z., Heiser, C., Hohenhorst, W., Herzog, M., Hemmert, W., & Schuller, B. (2018). Snoring classified: The Munich Passau snore sound corpus. *Computers in Biology and Medicine, 94*(1), 106–118.

Kaya, H., Fedotov, D., Dresvyanskiy, D., Doyran, M., Mamontov, D., Markitantov, M., Salah, A. A. A., Kavcar, E., Karpov, A., & Salah, A. A. (2019). Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. Proc. 9th Intl. Workshop on Audio/Visual Emotion Challenge, AVEC'19, 27th ACM MM, 27–35, ACM Nice, France.

Korda, H., & Itani, Z. (2013). Harnessing social media for health promotion and behavior change. *Health Promotion Practice, 14*(1), 15–23.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B., Star, K., Hajiyev, E., & Pantic, M. (2019). SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*.

Lucey, P., Cohn, J., Prkachin, K., Solomon, P., & Matthews, I. (2011). Painful data: The UNBC-mcmaster shoulder pain expression archive database. 8th Intl. Conf. and Workshops on Automatic Face and Gesture Recognition, FG'11, 57–64, IEEE Santa Barbara, CA.

Malin, B., & Sweeney, L. (2004). How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics, 37*(3), 179–192.

Pandit, V., Amiriparian, S., Schmitt, M., & Schuller, B. (Amiriparian, Schmitt, Schuller, 2019a). *Big data multimedia mining: Feature extraction facing volume, velocity, and variety. Big Data Analytics for Large-Scale Multimedia Search*. Wiley61–83.

Pandit, V., Cummins, N., Schmitt, M., Hantke, S., Graf, F., Paletta, L., & Schuller, B. (2018a). Tracking authentic and in-the-wild emotions using speech. Proc. 1st ACII Asia, IEEE, AAAC Beijing, P. R. China. doi:10.1109/ACIIAsia.2018.8470340.

Pandit, V., Schmitt, M., Cummins, N., Graf, F., Paletta, L., & Schuller, B. (2018b). How good is your model 'really'? on 'wildness' of the in-the-wild speech-based affect recognisers. Proc. 20th Intl. Conf. Speech and Computer, SPECOM, Springer ISCA Leipzig, Germany. doi:10.1007/978-3-319-99579-3_51.

Pandit, V., Schmitt, M., Cummins, N., & Schuller, B. (Schmitt, Cummins, Schuller, 2019b). *I know how you feel now, and here's why!: Demystifying time-continuous high resolution text-based affect predictions in the wild. Proc. 32nd international symposium on computer-based medical systems, CBMS*. Córdoba, Spain: IEEE465–470.

Pandit, V., & Schuller, B. The many-to-many mapping between the concordance correlation coefficient and the mean square error. arXiv:1902.05180.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). *On the difficulty of training recurrent neural networks. Intl. Conf. Machine Learning*1310–1318.

Pratt, H., Coenen, F., Broadbent, D., Harding, S., & Zheng, Y. (2016). Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science, 90*, 200–205.

Qian, K., Janott, C., Pandit, V., Zhang, Z., Heiser, C., Hohenhorst, W., Herzog, M., Hemmert, W., & Schuller, B. (2017). Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis. *IEEE Transactions on Biomedical Engineering, 64*(8), 1731–1741.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 1–24.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., Ciftci, E., Gülec, H., Salah, A. A., & Pantic, M. (2018a). AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. Proc. 8th Intl. Workshop on Audio/Visual Emotion Challenge, AVEC'18, 26th ACM MMACMSeoul, South Korea.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Soleymani, M., Schmitt, M., Amiriparian, S., Messner, E.-M., Tavabi, L., Song, S., Alisamir, S., Lui, S., Zhao, Z., & Pantic, M. (2018b). AVEC 2019 workshop and challenge: State-of-mind, depression with AI, and cross-cultural affect recognition. Proc. 9th Intl. Workshop on Audio/Visual Emotion Challenge, AVEC'19, 27th ACM MM ACM Nice, France.

Ringeval, F., Schuller, B., Valstar, M., Mozgai, S., Cummins, N., Schmitt, M., & Pantic, M. (2017). *AVEC 2017 – Real-life depression, and affect recognition workshop and challenge. Proc. 7th Intl. Workshop on Audio/Visual Emotion Challenge, AVEC'17, 25th ACM MM*. Mountain View, CA: ACM3–9.

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 10th Intl. Conf. and Workshops on Automatic Face and Gesture Recognition FG'13, 1–8IEEE Shanghai, P. R. China.

Roche, L., Zhang, D., Pokorny, F. B., Schuller, B. W., Esposito, G., Bölte, S., Roeyers, H., Poustka, L., Bartl-Pokorny, K. D., Gugatschka, M., Waddington, H., Vollmann, R., Einspieler, C., & Marschik, P. B. (2018). Early vocal development in autism spectrum disorders, rett syndrome, and fragile x syndrome: Insights from studies using retrospective video analysis. *Advances in Neurodevelopmental Disorders, 2*(1), 49–61.

Rodríguez-González, A., Labra-Gayo, J. E., Colomo-Palacios, R., Mayer, M. A., Gómez-Berbís, J. M., & García-Crespo, A. (2012). Sedelo: Using semantics and description logics to support aided clinical diagnosis. *Journal of Medical Systems, 36*(4), 2471–2481.

Ruiz, E. M., Tuñas, J. M., Bermejo, G., Martín, C. G., Rodríguez-González, A., Zanin, M., de Pedro, C. G., Méndez, M., Zaretskaia, O., Rey, J., et al. (2018). Profiling lung

cancer patients using electronic health records. *Journal of Medical Systems, 42*(7), 126.
Russell, J. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115*(1), 102.
Sargin, M. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2007). Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia, 9*(7), 1396–1403.
Schiel, F., Steininger, S., & Türk, U. (2002). *The smartkom multimodal corpus at BAS. Proc. 3rd Intl. Conf. Language Resources and Evaluation LREC'02* Las Palmas, Canary Islands - Spain
Schmitt, M., Cummins, N., & Schuller, B. W. (2019). Continuous emotion recognition in speech – do we need recurrence? Proc. 20th Annual Conference of The International Speech Communication Association, INTERSPEECH, ISCA Graz, Austria. doi:10.21437/Interspeech.2019-2710.
Schmitt, M., & Schuller, B. (2017). openXBOW – Introducing the passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research, 18*(96), 1–5.
Schuller, B. (2016). Can virtual human interviewers "hear" real humans' depression? *IEEE Computer Magazine, 49*(7), 8.
Sethu, V., Provost, E. M., Epps, J., Busso, C., Cummins, N., & Narayanan, S. (c). The ambiguous world of emotion representation. arXiv:1909.00360,.
Smailhodzic, E., Hooijsma, W., Boonstra, A., & Langley, D. (2016). Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Services Research, 16*(1), 442.
Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D. P., Ozmutlu, S., & Ozmutlu, H. C. (2004). A study of medical and health queries to web search engines. *Health Information & Libraries Journal, 21*(1), 44–51.
Thelwall, M. (2017). Tensistrength: Stress and relaxation magnitude detection for social media texts. *Information Processing & Management, 53*(1), 106–121.
Trigeorgis, G., Bousmalis, K., Zafeiriou, S., & Schuller, B. (2017). A deep matrix factorization method for learning attribute representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(3), 417–429.
Trigeorgis, G., Ringeval, F., Brückner, R., Marchi, E., Nicolaou, M., Schuller, B., & Zafeiriou, S. (2016). *Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. Proc. 41st ICASSP.* Shanghai, P. R. China: IEEE5200–5204.
Valstar, M., & Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. Conference on computer vision and pattern recognition workshop, CVPRW'06, 149–149, IEEENew York, NY.
Vandeventer, J., Aubrey, A. J., Rosin, P. L., & Marshall, D. (2015). *4d cardiff conversation database (4d CCDb): A 4d database of natural, dyadic conversations. Proc. 1st joint conference on facial analysis, animation and auditory-visual speech processing, FAAVSP* Vienna, Austria
Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H., Werner, P., Al-Hamadi, A., Crawcour, S., Andrade, A., & da Silva, G. M. (2013). *The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. Intl. Conf. Cyber., CYBCO*128–131 Lausanne, Switzerland
Werner, P., Al-Hamadi, A., & Walter, S. (2017). Analysis of facial expressiveness during experimentally induced heat pain. 7th Intl. Conf. Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW, 176–180 IEEE San Antonio, TX.
Yoo, M., Lee, S., & Ha, T. (2019). Semantic network analysis for understanding user experiences of bipolar and depressive disorders on reddit. *Information Processing & Management, 56*(4), 1565–1575.
Zhao, J., Li, R., Liang, J., Chen, S., & Jin, Q. (2019). Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions. Proc. 9th Intl. Workshop on Audio/Visual Emotion Challenge, AVEC'19, 27th ACM MM, 37–45, ACM Nice, France.
Zhou, G., Hansen, J., & Kaiser, J. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Process. 9*(3), 201–216.