Interpretable Time-Dependent Convolutional Emotion Recognition with Contextual Data Streams

David Bethge
david.bethge@um.ifi.lmu.de
Dr. Ing. h.c. F. Porsche AG, LMU Munich
Munich, Germany

Constantin Patsch
constantin.patsch@porsche.de
Dr. Ing. h.c. F. Porsche AG
Stuttgart, Germany

Philipp Hallgarten
philipp.hallgarten1@porsche.de
Dr. Ing. h.c. F. Porsche AG, TUM
Stuttgart, Germany

Thomas Kosch
thomas.kosch@hu-berlin.de
HU Berlin
Berlin, Germany

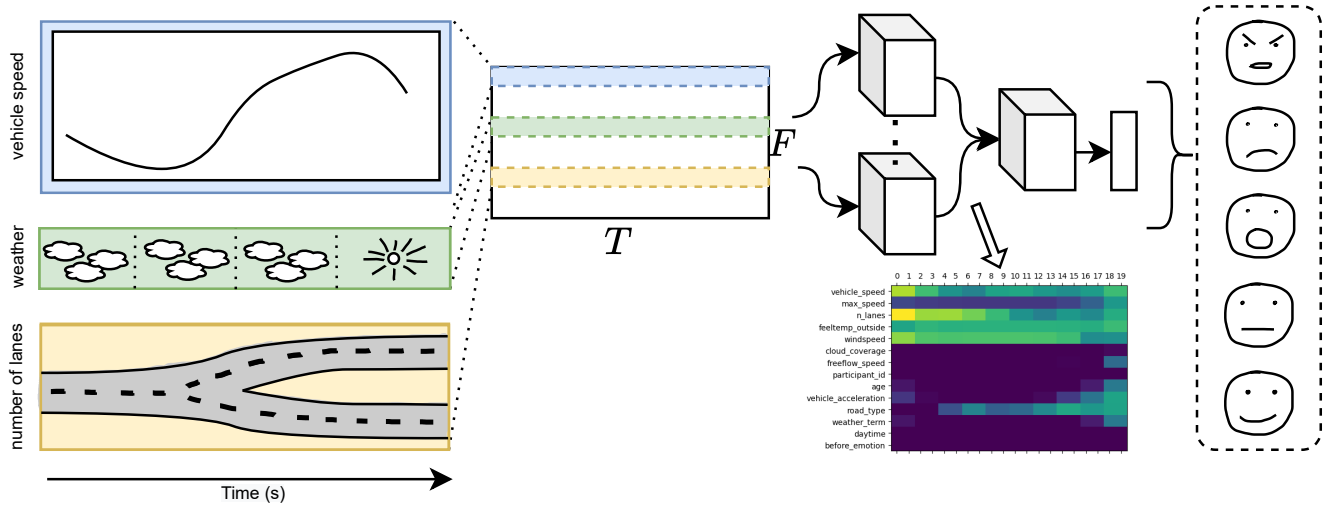

Figure 1: System architecture overview. ITER takes any multivariate time-series as an input (here: contextual vehicle variables) and performs a time-series classification (here: predicting driver emotions) while providing explainable feature maps, which display feature importance of the model's prediction over time.

ABSTRACT

Emotion prediction is important when interacting with computers. However, emotions are complex, difficult to assess, understand, and hard to classify. Current emotion classification strategies skip why a specific emotion was predicted, complicating the user's understanding of affective and empathic interface behaviors. Advances in deep learning showed that convolutional networks can learn powerful time-series patterns while showing classification decisions and feature importances. We present a novel convolution-based model that classifies emotions robustly. Our model not only offers high emotion-prediction performance but also enables transparency on the model decisions. Our solution thereby provides a time-aware feature interpretation of classification decisions using saliency maps. We evaluate the system on a contextual, real-world driving dataset involving twelve participants. Our model achieves a mean accuracy of 70% in 5-class emotion classification on unknown roads and outperforms in-car facial expression recognition by 14%. We conclude how emotion prediction can be improved by incorporating emotion sensing into interactive computing systems.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Explainable AI, Emotion Classification, Time-Series Classification, Affective Computing

ACM Reference Format:
David Bethge, Constantin Patsch, Philipp Hallgarten, and Thomas Kosch. 2023. Interpretable Time-Dependent Convolutional Emotion Recognition

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585672

with Contextual Data Streams. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3544549.3585672

1 INTRODUCTION AND BACKGROUND

Interacting with computing systems can induce a variety of emotions due to a combination of how one feels before, during, and after an interaction. Knowing the user's emotional state offers numerous possibilities for empathic and affective interfaces (e.g., emotion-adaptive lighting and design of emotion-dependent interaction patterns). However, due to the person-specific and privacy-concerning characteristics of emotions, there is a pressing need for emotion recognition engines to provide explanations on how the emotion prediction was made by opening the "black-box" prediction model. Providing explainable post-hoc visualizations can help the users to understand better employed empathic controls (e.g., changing of lighting because of detected emotions [17]) and reduce privacy concerns. Already in 2000, Picard [32] coined the term "Affective Computing", envisioning computers to express, sense, and predict emotions. Such interfaces have gained increased attention in numerous areas, such as the automotive sector or within the domain of recommender systems, to sense and regulate user emotions. Different sensors were investigated to detect emotional states, such as facial expression [20, 25], voice analysis [16], self-reports [7], or physiological sensing [5, 12].

Facial expressions have a long tradition as an indicator for the expressed emotions [14] and are used in a variety of software frameworks[1]. Typical facial expressions include smiling or frowning as well as head gestures (e.g., nods and tilts). The detection of facial expressions requires a remote camera within the user's environment, such as RGB cameras [10, 27, 30] or infrared cameras [15]. However, facial expressions can be misinterpreted without involving the user's context [20] and subjective interpretation [23]. In contrast, physiological sensing utilizes the user's direct bodily responses to draw conclusions about the emotional states. Several physiological sensing modalities, such as heart rate, electrodermal activity, and electroencephalography [4, 13, 37], are indicative of the user's perceived emotions. However, such sensors require direct contact with the user (e.g., an electrodermal activity sensor attached to the user's hand). Body-worn sensors can thus impact the user experience and usability negatively [40].

Various emotions can be elicited depending on the user's context. For example, driving is a common use case when studying user emotions [6, 9, 37]. Thus, various datasets exist that allow to compare the performance of different classification techniques [3, 6]. Previous research hypothesizes that the driving behavior, style, and context are indicative of the currently perceived emotions [29]. Here, behavioral characteristics are viewed as emotional markers.

However, improving time, impact, and the temporal context of emotion classification has not been studied so far. We close this gap by presenting an explainable model called ITER - **I**nterpretable, **T**ime-**B**ased **E**motion **R**ecognition, where time-dependent contextual features can be analyzed for their influence on emotions. Since driving datasets contain a large variety of perceived emotions, they are interesting to evaluate emotion classification techniques. Emotion estimation from contextual data is favorable as it becomes less privacy intrusive and no body-worn sensors are required. Thus, we are focusing on emotion classification using driving datasets.

Numerous emotion classification methods exist. In general, there are several methods for explaining model decisions, where essential ones can be grouped into local approximation [26, 33], backpropagation [34, 39] and input-masking based [31, 36] approaches. Using convolutional networks, we focus on visual interpretability in the form of saliency maps as they display feature time dependencies. They are defined as the weighted combination of the model's feature maps which provide insights into the network's attention toward feature-time instances within a specific sample. The feature maps are weighted by individual scores based on their contribution to the classification process. We propose to use a gradient-based method [34] combined with a forward-scoring method [36] to build interpretable feature maps. Our feature map thereby determines and visualizes the importance of the neurons for the classification decision. We omit the disadvantages of gradient-based methods for importances suffering from vanishing gradient problems by considering forward- and backward-importance derivatives when calculating the feature map.

Assaf et al. [1] introduce the MTEX-CNN, an architecture that performs time-series classification and explains its predictions by generating saliency maps from a convolutional layer. The creation of these saliency maps relies on the aforementioned Grad-CAM approach. By applying a convolution along the time dimension for each feature, they can retain the importance of an individual feature for time for a classification decision. Furthermore, in order to account for inter-feature dependencies, they apply a 1D convolution. When extracting saliency maps from this layer, they can infer the network's attention over all features towards specific time steps.

Tang et al. [35] propose an omni-scale 1D-CNN architecture for time-series classification that aims to cover a wide range of different receptive fields while relying on only a few layers. In contrast to related work, which defines a kernel configuration for parallel 1D-CNNs, we define a configuration for parallel 2D convolutions to retain the individual feature importance over time and thereby ensure feature-wise interpretability. Furthermore, to the best of our knowledge, no emotion predictor model exists that models the time-feature correspondence.

CONTRIBUTION STATEMENT

This paper makes the following contributions: (C1) We propose a learning architecture to include time as a variable in emotion recognition systems. (C2) We perform emotion classification with respect to time and contextual dependencies. These dependencies are interpreted with saliency maps that are extracted with a gradient-based and a forward-score-based approach. (C3) We present a novel parameter-efficient modeling structure for interpretable time-feature machine learning classification, making it useful for small-scale HCI datasets.

2 SYSTEM

In the following section, we describe our technique in detail. Our system needs to make sure to entail the following requirements:

[1]For example Affectiva: www.affectiva.com

Figure 2: Network architecture of our interpretable time-series classi cation system. The architecture consists of two stages, where the rst consists of parallel 2D convolution layers that preserve the feature dimension. The second stage consists of a 1D convolution layer, where the resulting feature maps are attened and forwarded to a dense and the nal classi cation layer.

(1) learning time dependencies in the input space, (2) applicable to small-scale HCI datasets, and (3) preserving feature explainability over time. The architecture proposed is outlined in Figure 2. It is composed of two subsequent parts, where the former deals with determining individual time-dependent feature importance while the latter focuses on determining time importance over the complete feature set. We consider a multivariate time series input for our multi-class classi cation problem. In order to make our model adaptive to small-scale experimental HCI datasets, we aim to minimize the number of trainable parameters. Inspired by [35], we de ne an architecture that captures a maximal variation of receptive elds while using a minimal amount of layers by applying di erent kernel sizes in parallel at several stages in the network. While [35] apply this approach to 1D convolutions, we apply it on our parallel 2D convolution layers where the kernel size is kept constant along the feature dimension. Thus, when applying the gradient and forward score-based approaches, we can distinguish between individual feature contributions toward the decision. This is essential for the user to infer the in uence of the time-context instances on the emotion classi cation. Due to the success of Grad- and Score-CAM in explaining image classi cation decisions we choose a CNN-based architecture due to their compatibility with these explainability methods.

Building Interpretable Feature Maps. The feature maps that we generate are saliency maps that help the user understand the model's decisions. The activation feature maps that are extracted from the last 2D convolution layers represent a visualization of the network's attention towards speci c features over time to a particular classi cation decision.

We determine activation feature maps based on the Grad-CAM method introduced by [34] and the Score-CAM method from [36]. Both Grad-CAM and Score-CAM are needed, as Grad Cam uses backward gradient calculation of feature importances, whereas Score-CAM is able to escape the vanishing gradient problem and uses forward-pass scores concerning the target class. In Grad-CAM, we calculate the gradients of the class with respect to the activations and average over the number of time instances of all features. A high value indicates a strong contribution of the individual instances in the feature maps towards the classi cation of the specific class. On the other hand, we use the Score-CAM method from [36], which deals with possible shortcomings of gradient-based methods like the vanishing gradient problem. The approach does not rely on the gradient-based weights by determining the activation map weighting through the forward pass scores concerning the target class. We achieve an interpretable feature map by summing up and normalizing the resulting weighted feature maps from the two convolution layers of the second stage for each of those methods. We describe the detailed calculation method in the Appendix. A comparison between feature maps of those two methods will be presented in Section 4.

3 DATA

The data used for ITER consists of acquired contextual driving data from an in-the-wild study [6] and is published open-source.[2] In total, 12 participants (• 12 self-identi ed as female) with an average age of 27 years (SD = 4.73). Six of the participants occasionally drive (i.e., less than 10,000 kilometers per year), where three participants drive moderate distances (i.e., between 10,000 and 20,000 kilometers per year), and three participants drive more frequently (i.e., more than 20,000 kilometers per year). The mean duration of the rides is 10 minutes (min = 6, max = 44). The data from all participants consists of 160 driving minutes sampled at 1 Hz, which corresponds to 9600 samples. The ground-truth emotion label capturing is designed in correspondence to the in-situ categorical emotion response (CER) rating for collecting data on emotional experiences in cars [11]. We consider the emotions `angry', `disgust', `happiness', `neutral', and `surprise'. A speech-to-text engine from the smartphone audio recording is used to extract the emotion labels, as the participant had to verbally provide their discrete emotion label every 60 seconds

[2] https://github.com/david-bethge/VEmotion

after a beep tone. A windshield-mounted smartphone recorded the driver's facial expression and contextual data. However, during our evaluation, we do not rely on these facial expressions. The list of available contextual features with exemplary values is shown in Table 2. We refer to the original paper for more details on the dataset and acquired emotional markers.

During preprocessing, we replace missing categorical and discontinuous values with the last recorded valid value and further replace the rest of the missing values by backpropagating a subsequent value to past time steps. Missing continuous numerical values like vehicle speed are replaced by applying kNN imputation [38]. This method ensures that we can prevent discontinuous changes between valid recorded and imputed values. As our architecture expects a xed input size, we use a sliding window approach similar to [24] with a stride of 1 on the multivariate time series to generate samples of size) . The corresponding label for each window is de ned as the label with the most occurrences within the window. We address the challenge of learning long-term emotion dependencies in the discussion section. We choose a window size of 20 as this has shown the best experimental recognition performance validated in an extensive window-size grid-search[3].

4 RESULTS

In this section, we analyze the emotion recognition performance of our system and compare it with related work. Furthermore, we explain and interpret the feature maps that our model outputs and compare the feature maps resulting from the gradient and forward score-based approaches.

For a baseline comparison, we evaluate our model using a 10-fold cross-validation similar to [6]. For each participant within the dataset, we leave one of the ten road segments out for evaluation and use the remaining road segments for training. A road segment is obtained by splitting the participant's driving session into ten parts. This evaluation teaches a global participant-independent model that can predict emotions on unknown road segments. Our model's results are depicted in the confusion matrix in Figure 3a. Overall our model achieves an accuracy of 70% and a $_1$ score of 69%. Besides that, ITER reaches a recall value of 51% on the 'happy' emotion, 82% on the 'neutral' emotion and 40% on the 'surprise' emotion. Nonetheless, we detect a poor classi cation performance for `angry' and `disgust' states. In particular, this is likely due to the skewed distribution of subjectively felt emotions. The emotions `angry' and `disgust' are underrepresented in the data as they only account for 1.3% in the former and 0.6% in the latter case.

We compare our model to the Random Forest classi er of VEmotion [6] and the Microsoft Face Recognition API [28], which both do not consider time during classi cation. Furthermore, for comparison, we choose models that also consider the temporal dimension. In particular, these are a two-layer LSTM model, a two-layer 1D-CNN, as well as the MTEX-CNN [1], which utilizes 2D and 1D convolution layers. From Table 1, we can observe that the accuracy and the $_1$ score of our approach are 2% lower than the ones of the VEmotion model. The di erence in performance is likely caused by our system's windowing preprocessing of the data, leading to an even smaller training dataset during cross-validation. In order to be able to exploit time dependencies more e ciently, the average time of a driving session and the number of participants will have to be extended. In the case of a larger dataset, time-series-based methods like our approach are likely to improve their performance results.

Emotion classi cation with the Microsoft Face Recognition API based on facial video data is outperformed by our system by 14% in terms of accuracy and 18% in terms of the $_1$ score. This indicates that facial expressions in a driving context are less expressive than time-dependent context features. When comparing our architecture to the two-layer 1D-CNN architecture, we can see that ITER achieves 5% better accuracy and 6% better $_1$ score. This increase implies that capturing a large range of receptive elds improves classi cation performance. Similarly, the two-layer LSTM model struggles to classify infrequent classes, which is indicated by the 14% lower $_1$ score compared to our model. In the case of the MTEX-CNN, the model seems to be less adapted towards imbalanced datasets, which is indicated by the 8% lower $_1$ score compared to our model. Furthermore, our model consists of about 20% of the trainable parameters of the MTEX-CNN. The models' relatively higher accuracies result from the dataset's imbalanced nature, where the neutral class is the most frequent.

Overall our model performs better in terms of accuracy and F1-score than the other models except for the Random Forest classi er introduced by [6]. However, their approach and the Microsoft Face Recognition API do not consider time dependencies in the data and cannot provide per-sample feature-wise explanations for emotion classi cation. While being able to consider time dependencies, up to our knowledge, there is no method to recover individual feature-time contributions from cell states in the LSTM model able to provide visual explanations. The 1D-CNN cannot provide feature-wise explanations as it applies a kernel over the whole feature dimension. The MTEX-CNN and our ITER model can consider time dependencies and provide feature maps that display the feature-wise importance over time.

Table 1: Emotion recognition performances of di erent classi cation models. The table further includes the models' properties time dependency and interpretability. Hereby, interpretability refers to the feature-wise explanation for classi cation decisions based on saliency maps. We compare our system to VEmotion [6], a facial expression classi cation system Face [28], a LSTM deep learning model [18], a 1d-CNN [19], MTEX-CNN [1].

	VEmotion	Face	LSTM	1D-CNN	MTEX-CNN	ITER (ours)
accuracy	.72	56	.64	.65	.68	.70
$_1$ score	.71	51	.55	.63	.66	.69
time dependency	7	7	3	3	3	3
interpretability	7	7	7	7	3	3

In this section, exemplary interpretable feature maps that result from the normalized weighted summation over the feature maps from the last 2D convolution layers are examined. Figure 3b displays an example of a multivariate time series within a 20-second window labeled with a happy emotion. Furthermore, we visualize the feature

[3] the window-size search space was set to {30,25,20,15,10}.

(a) Normalized confusion matrix. (b) Visualization of the input features.

Figure 3: (a): Normalized confusion matrix of the results of ITER with a mean accuracy of 70% based on a 10-fold cross-validation. (b): Visualization of the normalized input features from a multivariate time series sample corresponding to a happy emotion.

(a) Feature map based on Grad-CAM (b) Feature map based on Score-CAM

Figure 4: Feature maps based on the Grad-CAM and Score-CAM approaches resulting from the input of Figure 3b. The ~-axis corresponds to the contextual feature streams, whereas the G-axis shows the ascending time towards the most recent timestamp (the timestep 20 contains the most recent data).

vehicle speed exemplarily. The ascending time scale corresponds to the progress towards the most recent timestep. We normalized the input over the features, where yellow indicates the highest value and dark blue indicates the lowest value. Additionally, the vehicle speed feature column is visualized in a graph for the 20 seconds time window.

The interpretable feature maps displayed in Figure 4 represent the network's attention towards speci c time instances of features which are, on the one hand, determined by the gradient-based approach and, on the other hand, based on the forward pass scores of the masked inputs. As the whole feature map is normalized, yellow spots represent high attention, green spots medium attention, and dark blue spots low attention.

When looking at the feature map resulting from the Grad-CAM approach, which is shown in Figure 4a, we can observe that especiallyE4 82;4 B?44B<14A >5 ;0=4B<?4A0CD7A4=3 B?443 andA>03 C~?4em to be essential for the classi cation decision of this happy sample. Moreover, when comparing A>03 C~?4 instances of the feature map with the input, especially changes in A>03 C~?4em relevant for the classi cation decision. Furthermore, the model puts a higher focus on low acceleration values as the speci c time instances of the input have a higher weighting in the feature map. From the feature map in Figure 4b created based on the Score-CAM approach, we can observe that the attention intensity di ers from the Grad-CAM feature map. For example, the most recent time instances of the essential features are weighted

relatively higher in Figure 4b compared to the Grad-CAM feature map. However, the general importance of a feature's relevance for the classi cation decision is comparable to the Grad-CAM feature map. We showed that we could extract time-dependent feature interpretations for an emotion classi cation in the form of saliency maps. Furthermore, we provided sample-speci c explanations for a classi cation decision based on contextual features. The two proposed feature map generation methods have shown valid outputs and can both be used for emotion classi cation interpretation from contextual data streams.

5 DISCUSSION

Human-in-the-Loop for Emotion Recognition Models. Our method allows us to understand better the relationship between environmental, emotional triggers, and emotional states. The time-feature-dependent understanding is favorable for the emotion recognition developer in knowing why a speci c decision has been made and offers the user a transparent way of knowing why a machine learning decision based on his emotional state was made. This interactivity between humans and machine learning systems is crucial, especially when developing empathic interfaces for in-the-wild use. Furthermore, by providing a more direct assessment of emotion detection, our model can be seen as another step toward transparency in empathic interfaces, which are a major limiting factor in the development of large-scale employment [8].

The proposed methodology for generating interpretable feature maps can be applied to a wide range of HCI scenarios. We could analyze which contextual feature changes induced an emotion change in an automotive context and thus infer speci c emotional triggers. These could then be consumed by a routing algorithm that adapts correspondingly, e.g., by avoiding speci c road attributes. In the case of developing empathic car interfaces, being able to detect emotions and interpret the classi cation process is essential. The system could display its reasoning process with the help of feature maps to the driver and thus improve the transparency of model decisions. This could further improve the driver's trust in the system. Recent research by Atakishiyeva et al. [2] stresses the signi cance of explainability for autonomous driving decisions and highlights the approach of using post-hoc explanation visualizations.

Limitations and Future Work. In general, the features corresponding to an emotion that the model explicitly nds important might only partly match with the features that the driver perceives as most in uential in a particular situation. For example, features or modalities not captured in the dataset, like in-car volume or voice intensity, might be more expressive in certain situations. As the driver is exposed to a vast range of modalities in the environmental context (e.g. cognitive workload [21] or biased expectations towards a reactive AI-based emotion feedback system [22]), the interpretation of emotion for a limited number of features might only re ect the emotional reasoning to a certain extent. For the model to learn long-term dependencies (e.g., 5 minutes), the input window must be at least this speci c size. As a result, the number of samples in the training and test set decreases. This poses a problem in small-scale experimental datasets as, in our case, the mean duration of a participant's driving session is only 10 minutes. Thus, large input windows cannot be chosen due to the relatively short driving sessions, which is why we set a time window of 20 seconds. Furthermore, the interpretable feature maps we extract only o er a local per-sample explanation concerning an emotion. Thus, these representations allow no implications about global feature importance over the whole data set. We focused on emotion classi cation based on contextual driving data in this work. However, for future work, one might also consider physiological data of the participants or even further in-car modalities, like in-car volume levels.

6 CONCLUSION

We introduced ITER, a model that classi es drivers' emotions based on contextual driving data represented as multivariate time-series. We showed that by considering time as a variable in the emotion recognition system, we are able to interpret the importance of individual feature instances with respect to a speci c classi cation result. Hereby, explainability is visualized by saliency maps that are created with a gradient-based and a forward-score-based method. Being able to explain the model's classi cation decision by inferring the importance of certain feature aspects might be crucial to help humans understand the model's reasoning process. In driving scenarios, empathic car interfaces and emotional routing might be suitable applications for such a system. In general, being able to interpret model decisions might help to better understand the input data by analyzing conspicuities within a sample.

ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) CRC 1404 FONDA.

REFERENCES

[1] Roy Assaf, Ioana Giurgiu, Frank Bagehorn, and Anika Schumann. 2019. Mtex-cnn: Multivariate time series explanations for predictions with convolutional neural networks. In2019 IEEE International Conference on Data Mining (ICDM). IEEE, 952 957.

[2] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2021. Explainable arti cial intelligence for autonomous driving: a comprehensive overview and eld guide for future research directions. arXiv preprint arXiv:2112.11561 (2021).

[3] David Bethge, Luis Falconeri Coelho, Thomas Kosch, Satiyabooshan Murugaboopathy, Ulrich von Zadow, Albrecht Schmidt, and Tobias Grosse-Puppendahl. 2023. Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6, 4 (2023), 1 30. https://doi.org/10.1145/3569466

[4] David Bethge, Philipp Hallgarten, Tobias Grosse-Puppendahl, Mohamed Kari, Lewis L. Chuang, Ozan Özdenizci, and Albrecht Schmidt. 2022. EEG2Vec: Learning A ective EEG Representations via Variational Autoencoders. In2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 3150 3157. https://doi.org/10.1109/SMC53654.2022.9945517

[5] David Bethge, Philipp Hallgarten, Ozan Özdenizci, Ralf Mikut, Albrecht Schmidt, and Tobias Grosse-Puppendahl. 2022. Exploiting multiple eeg data domains with adversarial learning. In2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 3154 3158. https://doi.org/10.1109/EMBC48229.2022.9871743

[6] David Bethge, Thomas Kosch, Tobias Grosse-Puppendahl, Lewis L Chuang, Mohamed Kari, Alexander Jagaciak, and Albrecht Schmidt. 2021. VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time. In The 34th Annual ACM Symposium on User Interface Software and Technology. 638 651. https://doi.org/10.1145/3472749.3474775

[7] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic di erential. Journal of Behavior Therapy and Experimental Psychiatry 25, 1 (1994), 49 59. https://doi.org/10.1016/0005-7916(94)90063-9

[8] Michael Braun, Jingyi Li, Florian Weber, Bastian P eging, Andreas Butz, and Florian Alt. 2020. What If Your Car Would Care? Exploring Use Cases For

Affective Automotive User Interfaces. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 37, 12 pages. https://doi.org/10.1145/3379503.3403530

[9] Michael Braun, Florian Weber, and Florian Alt. 2021. Affective Automotive User Interfaces–Reviewing the State of Driver Affect Research and Emotion Regulation in the Car. *ACM Comput. Surv.* 54, 7, Article 137, 26 pages. https://doi.org/10.1145/3460938

[10] Silvia Ceccacci, Maura Mengoni, Generosi Andrea, Luca Giraldi, Giuseppe Carbonara, Andrea Castellano, and Roberto Montanari. 2020. A Preliminary Investigation Towards the Application of Facial Expression Analysis to Enable an Emotion-Aware Car Interface. In *Universal Access in Human-Computer Interaction. Applications and Practice*, Margherita Antona and Constantine Stephanidis (Eds.). Springer International Publishing, Cham, 504–517. https://doi.org/10.1007/978-3-030-49108-6_36

[11] Monique Dittrich and Sebastian Zepf. 2019. Exploring the validity of methods to track emotions behind the wheel. In *International Conference on Persuasive Technology*. Springer, 115–127. https://doi.org/10.1007/978-3-030-17287-9_10

[12] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. 2020. Human Emotion Recognition: Review of Sensors and Methods. *Sensors* 20, 3 (2020). https://doi.org/10.3390/s20030592

[13] Maria Egger, Matthias Ley, and Sten Hanke. 2019. Emotion recognition from physiological signal analysis: a review. *Electronic Notes in Theoretical Computer Science* 343 (2019), 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

[14] Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to emotion* 3, 19 (1984), 344.

[15] H. Gao, A. Yüce, and J. Thiran. 2014. Detecting emotional stress from facial expressions for driving safety. In *2014 IEEE International Conference on Image Processing (ICIP)*. 5961–5965. https://doi.org/10.1109/ICIP.2014.7026203

[16] Teddy Surya Gunawan, Muhammad Fahreza Alghifari, Malik Arman Morshidi, and Mira Kartiwi. 2018. A review on emotion recognition algorithms using speech analysis. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)* 6, 1 (2018), 12–20.

[17] Mariam Hassib, Michael Braun, Bastian Pfleging, and Florian Alt. 2019. Detecting and Influencing Driver Emotions Using Psycho-Physiological Sensors and Ambient Light. In *Human-Computer Interaction – INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, Cham, 721–742.

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[19] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 2021. 1D convolutional neural networks and applications: A survey. *Mechanical systems and signal processing* 151 (2021), 107398.

[20] Thomas Kosch, Mariam Hassib, Robin Reutter, and Florian Alt. 2020. Emotions on the Go: Mobile Emotion Assessment in Real-Time Using Facial Expressions. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI '20)*. Association for Computing Machinery, New York, NY, USA, Article 18, 9 pages. https://doi.org/10.1145/3399715.3399928

[21] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* (jan 2023). https://doi.org/10.1145/3582272

[22] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* 29, 6, Article 56 (jan 2023), 32 pages. https://doi.org/10.1145/3529225

[23] Tuan Le Mau, Katie Hoemann, Sam H Lyons, Jennifer Fugate, Emery N Brown, Maria Gendron, and Lisa Feldman Barrett. 2021. Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nature communications* 12, 1 (2021), 1–13. https://doi.org/10.1038/s41467-021-25352-6

[24] Chien-Liang Liu, Wen-Hoar Hsaio, and Yao-Chung Tu. 2018. Time series classification with multivariate convolutional neural network. *IEEE Transactions on Industrial Electronics* 66, 6 (2018), 4788–4797.

[25] André Teixeira Lopes, Edilson De Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. 2017. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern recognition* 61 (2017), 610–628.

[26] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[27] Zhiyi Ma, Marwa Mahmoud, Peter Robinson, Eduardo Dias, and Lee Skrypchuk. 2017. Automatic Detection of a Driver's Complex Mental States. In *Computational Science and Its Applications*, Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Giuseppe Borruso, Carmelo M. Torre, Ana Maria A.C. Rocha, David Taniar, Bernady O. Apduhan, Elena Stankova, and Alfredo Cuzzocrea (Eds.). Springer International Publishing, Cham, 678–691. https://doi.org/10.1007/978-3-319-62398-6_48

[28] Microsoft. [n. d.]. Azure Facial recognition API. https://azure.microsoft.com/de-de/services/cognitive-services/face/

[29] Meital Navon and Orit Taubman – Ben-Ari. 2019. Driven by emotions: The association between emotion regulation, forgivingness, and driving styles. *Transportation Research Part F: Traffic Psychology and Behaviour* 65 (2019), 1–9. https://doi.org/10.1016/j.trf.2019.07.005

[30] M. Paschero, G. Del Vescovo, L. Benucci, A. Rizzi, M. Santello, G. Fabbri, and F. M. F. Mascioli. 2012. A real time classifier for emotion and stress recognition in a vehicle driver. In *2012 IEEE International Symposium on Industrial Electronics*. 1690–1695. https://doi.org/10.1109/ISIE.2012.6237345

[31] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).

[32] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[35] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Jing Jiang, and Michael Blumenstein. 2020. Rethinking 1d-cnn for time series classification: A stronger baseline. *arXiv preprint arXiv:2002.10061* (2020).

[36] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 24–25.

[37] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W Picard. 2020. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–30. https://doi.org/10.1145/3388790

[38] Shichao Zhang. 2012. Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software* 85, 11 (2012), 2541–2552.

[39] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[40] Feng Zhou, Yangjian Ji, and Roger J. Jiao. 2014. Augmented Affective-Cognition for Usability Study of In-Vehicle System User Interface. *Journal of Computing and Information Science in Engineering* 14, 2 (02 2014). https://doi.org/10.1115/1.4026222 arXiv:https://asmedigitalcollection.asme.org/computingengineering/article-pdf/14/2/021001/6099446/jcise_014_02_021001.pdf 021001.

A APPENDIX

Ethical Impact Statement. Our emotion model is privacy-sensitive as it offers the possibility to recognize subjectively-felt emotions for drivers with good recognition performance. In addition, this work looks at contextual data only, thereby being less privacy intrusive than facial expression or voice analysis systems. If a system would be employed to trigger in-cabin adaptations (e.g., emotion-adaptive lighting, displays, sounds), the user might get the feeling of not being in control. However, due to the lightweight structure of our model, we can integrate it into the car, and thus it would be able to provide significantly more feedback to the driver than current systems. Our current work objectively tries to provide more interpretable feedback for model decisions. Therefore, we stress a transparent and ethical use of our system.

Table 2: List of available features to predict drivers emotions.

Context	Feature	Example Values
vehicle trajectory	vehicle_speed	2.255133
	vehicle_acceleration	-0.15.
weather	feeltemp_outside	13.0
	windspeed	5.6
	cloud_coverage	76
	weather_term	'clear'
traffic	trafficflow_reducedspeed	7.295495
	freeflow_speed	115.0
road	road_type	'residential'
	max_speed	30.0
	n_lanes	2
in-vehicle	facial expression	'surprise'
personal	daytime	'afternoon'
	age	21
	before_emotion	'happiness'

Neural Network Specification. The time-series input to the network has the dimension $_0 \times)_0$, where $_0$ is the feature dimension, and $)_0$ is the time dimension. In our case, $)_0$ is set to a temporal window size of 20, and $_0$ is equivalent to 14 features. The choice of $_0$ depends on the context features that are recorded in the dataset, while the choice of $)_0$ has been determined experimentally (this is further justified in section 5). The first stage of the architecture consists of # parallel 2D convolution layers with different kernel sizes $1 \times :_n$ with $= \in \{1, \ldots, \#\}$. $:_n$ represents the respective kernel size along the time dimension, while the first dimension of the kernel is set to 1 to retain the individual feature importance for a classification decision. Same padding and a stride size of 1 are used to preserve the original input dimension of $_0 \times)_0$ and allow the concatenation of feature maps resulting from different kernel sizes. After concatenating the number of 3_f feature maps resulting from the convolution layers along the third dimension, a batch normalization, ReLU, and dropout layer are applied onto the feature maps. We again repeat the aforementioned process of parallel 2D convolution layers with the same kernel sizes $1 \times :_n$ with respect to the 3_f feature maps. By using same padding and a stride size of 2, the feature map sizes result in $_0 \times)_1$. In the next stage, we apply a 2D convolution with the kernel sizes 1×1 and 1×2, while using same padding and a stride of 1. The resulting feature maps are again concatenated and reshaped to $_1 \times)_1$.

The second part of the architecture is defined by a 1D convolution layer, a dense layer as well as the final dense classification layer with a softmax activation function. More specifically, we define a 1D convolution with the kernel size $:_{1D}$ that is used to account for dependencies of features between different time steps. The resulting $_2 \times)_1$ feature map is flattened in the last stage to be a suitable input to the following dense layer of size $1 \times \ _2$. As the last step, we define a dense classification layer for the number of classes $=_{cl}$.

Saliency Map Calculation. The feature maps that we generate are saliency maps that help the user understand the model's decisions. The activation feature maps that are extracted from the last 2D convolution layers represent a visualization of the network's attention towards specific features over time to a particular classification decision.

On the one hand, we create activation feature maps based on the Grad-CAM method introduced by [34]. The weight U_k^c of each feature map ^{k} is determined by

$$U_k^c = \frac{1}{k)_k} \sum_{i=1}^{F_k} \sum_{j=1}^{T_k} \frac{\partial ~_c}{\partial \ ^k_{ij}}, \tag{1}$$

where we calculate the gradients of the class $~_c$ with respect to the activations $^k_{ij}$ and average over the number of time instances of all features. A high value of U_k^c would indicate a strong contribution of the individual instances in the feature map ^{k} towards the classification of $~_c$. We sum over the weighted activation maps and apply a ReLU function in order to capture only positive influence with respect to class $~_c$.

On the other hand, we use the Score-CAM method from [36] which deals with possible shortcomings of gradient-based methods like the vanishing gradient problem. The approach does not rely on the gradient-based weights by determining the activation map weighting through the forward pass scores concerning the target class. Therefore, we first have to calculate the masked inputs ^k_M defined by

$$^k_M = \ \circ \ " ^k, \tag{2}$$

where $$ represents the multivariate time window input and $" ^k$ defines activation maps ^{k} that are upsampled to the input and normalized. The masked inputs are then fed into the model to determine their classification score V_k^c for class $~_c$. The higher the classification score of a masked input ^k_M, the stronger ^{k} gets weighted. Like Grad-Cam, a ReLU function is applied to the sum over the weighted activation maps $V_k^c \ ^k$.

Evaluation. For comparison, we provide the confusion matrix of VEmotion and qualitative results that display Grad-CAM and Score-CAM visualizations.

Figure 5: 10-fold cross-validation results of VEmotion [6].