Machine Learning Systems for Human Emotional States

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von

David Bethge
M.Sc. Wirtschaftsingenieurwesen

München, den 1. August 2023

Erstgutachter:	Prof. Dr. Albrecht Schmidt
Zweitgutachter:	Prof. Dr. Christian Holz
Drittgutachter:	Prof. Dr. Bastian Pfleging

Tag der mündlichen Prüfung: 26.5.2023

ABSTRACT

Human emotions play a vital role in our everyday lives and influence our communication, perception and experience in the real world. However, detecting, modeling, and predicting human emotional states remains challenging. Emotions are subjective, ambiguous, volatile, and current sensing modalities lack robustness in realistic environments and are potentially intrusive. The central question of this thesis is: How can we design emotion-aware systems under these constraints that utilize recent advances in machine learning?

We present novel input for human emotions based on multimodal and contextual data. We furthermore propose machine learning frameworks that provide more explanatory, generalizable, privacy-preserving, and multi-sensory emotion state predictions.

We introduce two ubiquitous mobile sensing applications for human emotional states [P1, P5]. The applications are designed to be minimally intrusive and do not require body contact by acquiring contextual data only to infer emotions. We show that the re-formalization of the classification problem to an indirect estimation context-emotion pattern recognition works better than existing (facial expressions) baseline approaches. Our findings have major implications for future emotional state recognition systems because ubiquitous devices can easily acquire contextual data and transform them into smart emotion-estimation sensors. In the second pillar of this thesis, we focus on machine learning architectures for human emotional states. We present three different deep learning architectures for learning emotion representations from multivariate time-series data [P4, P3, P2]. We show how the representations can be designed to be emotion predictive but also participant-invariant [P2]. We furthermore propose an architecture that can learn a shared emotional representation from multiple noisy datasets with different inherent sensing characteristics [P3, P4]. In summary, the proposed approaches are tailored to capture detailed semantics in noisy input data and also propose measures for making the prediction more interpretable, domain-invariant and scalable. Furthermore, we present a novel transformer-based deep learning architecture for emotion prediction that is able to derive explanations alongside its predictions [P6]. Thereby, a better understanding of the link between emotion sensor input and the model decision can be derived that ultimately helps designing appropriate user interfaces powered by emotion prediction methods. Ultimately, we show a machine learning application of contextual emotion prediction that enables an emotion-aware experience in a dynamic driving environment [P7]. We present the first navigation application that is able to route after emotions. We thereby leverage the concept of context-emotion linking to predict emotional-states on future road-segments and optimize after travel-time and predicted emotional states. We show in a blind-user study, that emotional navigation has a positive effect on valence after riding the predicted happy route.

Although our work is concentrated on human emotional states, we argue that this enables other human-states (e.g., fatigue, cognitive load) prediction systems to be designed to be more robust, scalable, privacy-aware. These systems also allow the user to understand and investigate the decision process and offer explanations. Furthermore, our work provides a foundation for emotion-aware systems in unconstrained, dynamic environments.

ZUSAMMENFASSUNG

Menschliche Emotionen spielen in unserem täglichen Leben eine wichtige Rolle und beeinflussen unsere Kommunikation, Wahrnehmung und Erfahrung in der realen Welt. Die Erkennung, Modellierung und Vorhersage menschlicher Gefühlszustände bleibt jedoch eine Herausforderung. Emotionen sind subjektiv, mehrdeutig und flüchtig. Den derzeitigen Erfassungsmodalitäten mangelt es an Robustheit in realistischen Umgebungen und sie sind potenziell aufdringlich. Die zentrale Frage dieser Arbeit ist: Wie können wir unter diesen eingeschränkten Bedingungen emotionsbewusste Systeme entwerfen, die die jüngsten Fortschritte im Bereich Machine Learning nutzen?

Das Ziel dieser Arbeit ist es, emotionsbewusste Systeme in solchen eingeschränkten Szenarien zu entwickeln, die fortschrittliche Machine Learning Techniken zur Erkennung und Adaption des emotionalen Zustands des Benutzers einsetzen. Wir stellen intelligente Eingabe- und Machine Learning Frameworks für menschliche Emotionen vor, die multimodal und kontextabhängig sind und die erklärende, verallgemeinerbare, datenschutzfreundliche und multisensorische Zustandsvorhersagen ermöglichen. Schließlich zeigen wir eine Machine Learning Anwendung der kontextuellen Emotionsvorhersage, die ein emotionsbewusstes Erlebnis in einer dynamischen Fahrumgebung ermöglicht. Obwohl sich unsere Arbeit auf menschliche Gefühlszustände konzentriert, argumentieren wir, dass dies die Entwicklung von Systemen zur Vorhersage anderer menschlicher Zustände (z. B. Müdigkeit, kognitive Belastung) ermöglicht, die robuster, skalierbarer und datenschutzfreundlicher sind. Diese Systeme ermöglichen es dem Benutzer auch, den Entscheidungsprozess zu verstehen und zu untersuchen und Erklärungen anzubieten. Darüber hinaus bietet unsere Arbeit eine Grundlage für das Design von emotionsbewussten Systemen in eingeschränkten, dynamischen Umgebungen.

ACKNOWLEDGMENTS

My dissertation would not have been possible without the amazing people I had the pleasure to meet and work with every day. Some of you have influenced me through encounters or transient projects, others have been there for years on end. I want to say thank you to everybody who accompanied me along the way and impacted me and the work I do. Thank you to all the great people who collaborated with me, who I had coffee with at conferences, or met elsewhere. Thank you to everyone I forgot to name. I am incredibly grateful for the professional and personal guidance I have received during this time.

Thank you, **Albrecht Schmidt**, for the special PhD and life advice and outstanding support throughout my PhD, your enthusiasm towards new ideas and sharing your endless expertise in AI and HCI. Meeting you and your group was a fortunate event I will never forget.

Thank you, **Christian Holz** and **Bastian Pfleging**, who took their time to serve as external supervisors for this thesis.

Thank you, **Tobias Grosse-Puppendahl**, for the opportunity to do my PhD at Porsche, for introducing me to the Emerging Technologies team and the exciting research in the field of Machine Learning and HCI, and for being a role model in determination and creativity.

Thank you, **Thomas Kosch** for your great advice, teaching me how to draft a good HCI storyline, relentless efforts to organize a thriving HCI community, building cool interactive systems and your research calendar.

Thank you, **Lewis Chuang** for your feedback from a neuroscience's point of view, for asking bigger questions and giving great feedback.

Thank you, **Mohamed Kari** for being a great friend, a supporter on all the crazy ideas we had during coffee and beers and for supporting to find my way for the future.

In my daily work, I received plenty of help from my colleagues at Porsche: Thank you, **Philipp Hallgarten** for your amazing PyTorch support, novel paper ideas, and implementation efforts in representation learning. Thank you, **Constantin Patsch** for your in-depth transformer expertise. Thank you, **Luis Coelho** for your great dev support night and day. Thank you, **Marco Wiedner** for the endless and amazing discussions how to perform good research in a company and what needs to be changed. Thank you, **Satiyabooshan Murugaboopathy** for the great startup discussions. Thank you, **Jochen Gross** for being a great manager and friend. You made the PhD experience in a company worthwile. Thank you, **Aikaterina Kourti** for the Mischwoch coffee get-togethers to talk about research and the life beyond.

I further want to take a bow in front of my friends and fellow researchers at LMU, TU Graz, and other institutions, for the tremendous collaboration throughout the years. Thank you, **Ozan Özdenizci** for your great machine learning wisdom and asking the right questions for disentangling representation learning problems. Thank you, **Daniel Bulanda** and **Adam Kozlowski** for your amazing coding knowledge, research- and bring-to-market-mentality. Thank you, **Francesco Chiossi** for letting me borrow your EEG, expertise in neuroscience, and introducing me to the italian way of life.

Thank you, **Artur Dubrawski** for hosting me at CMU and sparking my interest of doing a PhD in ML. Thank you, **Ting Zhang**, **Naveen Sendhilnathan**, **Tanya Jonker**, **Michael Nebeling**, **Hrvoje**

Benko, and **Tovi Grossmann** for providing an exciting visiting research stay at Meta Reality Labs in Redmond at the final stages of my PhD. Thank you, **Marco Wetter** for teaching me the beauty of math at the age of 6 using your binary watch. Thank you, **Nicklas Lauinger**, **Karin Lauinger**, **Matthias Kuhlemann**, **Robert Steiner**, **Moritz Fasan** and **Nina Fasan** for supporting me as caring friends to keep my sanity throughout my PhD. Thank you, **Jonas Rothfuss**, **Fabio Ferreira**, **Tim Haubeil** and **Marius Causemann** for the endless machine learning discussions and climbing adventures at $3000m$. Thank you, **Truc Ho** and **Tu Trinh** for the great working environment on the weekends with all-you-can-eat rice and spaghetti and for letting me exploit your coffee machine.

Finally, I want to thank my family. Thank you for supporting me since before I can remember, for the backup I could always rely on. Thank you, **Esther**, for your endless openness, love, encouragement and for your unlimited generosity and affection. Thank you, **Daniel**, for putting such an importance on education and lifelong learning and for teaching me to be kind and think in-depth about the world around us. Thank you, **Raphael**, for being my great big brother and for the joy you bring to people and your "construction" attitude. Thank you for showing me and taking the express tunnels during my PhD. Thank you, **Thalia**, for being the best sister one could imagine and for bringing your open attitude and thinking to my life. Thank you, **Han**, for being by my side through thick and thin. For supporting me, challenge me and making me a better version David v2.0. Thank you for grabbing my hand, showing me the world, teaching me surfing, vietnamese food vocabulary and striving and failing daily for going to bed early. Without you, this would have not been possible.

COLLABORATION STATEMENT

None of the work comprised in this thesis would have been possible without the support of many great people: my supervisors, colleagues, students and collaborating partners. I gratefully acknowledge their efforts towards my research by using the scientific *"we"* throughout the text of this dissertation. In this statement I illustrate my personal contribution and the help I received along the way.

All publications are results of a close collaboration with my supervisor Albrecht Schmidt, who was involved at the very start to define the vision guiding my research. He was involved in every project from idea to presentation. My thesis supervisor at Porsche, Tobias Grosse-Puppendahl, helped me to identify open research questions and to define the vision guiding my research. Tobias Grosse-Puppendahl further facilitated my research work through the creative and strategic guidance. Lewis Chuang contributed feedback to study designs and manuscripts. Thomas Kosch contributed feedback to experimental designs and provide guidance for paper manuscripts. These attributions apply to the publications on which co-authorship is stated.

The publications featured in this thesis are based on projects with diverse organizational backgrounds. The students involved either wrote their bachelor's or master's theses or interned with me at Porsche. I determined their research topics, supervised the progress through regular feedback, enabled their work with ample assistance, and implemented and coded the systems side-to-side. Other contributors were colleagues from several institutions whom I collaborated with in shared responsibility. If not stated otherwise, included publications are based on original concepts and prototypes of myself, with feedback from the respective co-authors. I verified all analyses, created the visualizations and wrote and revised the manuscripts.

Machine learning systems: I supervised several student theses on methodological machine learning system questions. Philipp Hallgarten programmed a machine learning system for his master thesis, to which he contributed the implementation and prediction evaluation.

Methodical explorations The intern Alexander Jagaciak implemented a prior version of a smartphone application. My bachelor student Luis Coelho built up on the prototyping work and implemented the experimental design application for iOS smartphones, which was later used as an emotion sampling tool for experimental studies.

Table 1 clarifies the contributions of others to individual projects and publications.

Project & Publications	Contributions of Others
Emotion Sampling [P1, P5]	My bachelor student Luis Coelho and intern Alexander Jagaciak implemented the Swift smartphone application and visual processing pipeline. Luis Coelho conducted part of the user study. Thomas Kosch, Tobias Grosse-Puppendahl, Ulrich von Zadow, Satiyabooshan Murugaboopathy, and Albrecht Schmidt reviewed the paper.
Emotion-Aware Application [P7]	My collaborators at GrapeUp, Daniel Bulanda, and Adam Kozlowski, implemented the visualization, java optimization, and routing application. Tobias Grosse-Puppendahl, Albrecht Schmidt and Thomas Kosch reviewed the paper.
Representation Learning Models [P2, P3, P4]	My master student Philipp Hallgarten programed the neural network design for [P3, P4] and conducted the evaluation analysis. Ozan Özdenizci, Lewis Chuang, Ralf Mikut and Tobias Grosse-Puppendahl revised the paper.
Emotion Explainability Models [P6]	My intern Constantin Patsch implemented and evaluated the neural network architecture. My master student Philipp Hallgarten and collaborator Thomas Kosch reviewed the paper.

Table 1: Clarification of the author's personal contribution in each incorporated publication. If not mentioned otherwise, I conceived and executed the work, provided the visual materials, composed the written manuscript, and edited the final version. The co-authors which are not named individually contributed with supervision, feedback on concepts and manuscripts, and/or organizational help. A more in-depth author contribution description is provided in the publication section 3 of this thesis.

Table of Contents

1 Introduction 1

1.1 Thesis Statement 1

1.2 Problem Statement 1

1.3 Contributing Publications 2

1.4 Guiding Research Questions 3

1.5 Research Approach 3

1.6 Research Context 5

2 Background 7

2.1 Concepts of Human Emotions 7

2.2 Emotion Detection and Emotion Prediction 7

2.2.1 General Approach 8

2.2.2 Difficulties 9

2.2.3 Milestones 10

3 Publications 13

3.1 Intelligent Input Modalities 13

3.2 Explanatory Emotion Predictions Models 17

3.3 Privacy-Preserving Human-Emotion Models 18

3.4 Human-Emotion Application in Dynamic Environments 21

4 Results 23

5 Conclusion 27

5.1 Reflection 27

5.2 Future Work 28

5.2.1 Immediate Future Work 29

5.2.2 Going Beyond The Lab 29

5.2.3 Long Term Horizon 30

5.3 Concluding Remarks 31

List of Tables and Figures 33

References 35

Original Contributing Publications 39

1 Introduction

"An IQ test?"
"No. Empathy."
"I'll have to put on my glasses."

Philip K. Dick, Do Androids Dream of Electric Sheep?

1.1 Thesis Statement

Digital user interfaces are more and more accessible in everyday situations. They are designed to perform tasks explicitly given by the user. At the same time, user interfaces are increasingly conceived to promote the user's well-being. Recent research in the field of affective computing investigates the possibilities for the technical interface to interact with humans by recognizing and considering their emotions. This offers computers to not only provide better performance in assisting humans but also might enhance computers' abilities to make decisions. Recognizing human emotions helps the computer to be more effective in human-computer interaction by, e.g., promoting a workout when the user feels angry. Furthermore, the computer can also employ a happy navigation route for a user when he is detecting angry emotion, thereby increasing mental well-being and reducing traffic accident probability. However, the long-term implications of deploying such emotion detection algorithms and emotion-aware user interfaces are manifold and often hard to predict.

This thesis investigates the field of affective computing from a machine learning (ML) standpoint. The technology of interfaces offers the possibility to access a vast stream of multimodal sensor data. Therefore, we investigate sensing input modalities [P1] and their technical design space for in-the-wild use [P5]. Next, we discuss how contextual affect recognition can be designed to be more explainable and help to understand better the link between context and emotion [P6]. Thereafter, we provide machine learning models to incorporate multi-domain affective data [P2, P3, P4]. At last, we provide an in-the-wild, emotion-aware navigation application using scalable contextual emotion prediction [P7].

1.2 Problem Statement

Human emotions are complex: they are volatile, highly person-dependent, and hard to express, among other factors. Due to this, computers with the ability to sense, process, detect and react to human emotional states are hard to build. How can we design interfaces that sense emotions, process them, and provide insights and interface adaptations?

In computer science, the problem of detecting emotion is to process changes in physiological, behavioral, or contextual ongoings to deduct the current emotion of a human. Many attempts to deduct human emotions in specific settings have been employed, e.g., detecting emotions while playing video games. However, due to the nature of human emotions, a general-purpose method to detect human emotions has yet to be developed.

This thesis takes another angle at detecting and building emotion-aware computing systems by 1) integrating intelligent ubiquitous input, 2) proposing general-purpose encoder models for domain-privacy-aware machine learning methods, 3) providing machine learning architectures for emotion model decisions, and 4) proposing human-emotion applications in uncontrolled, dynamic environments.

1.3 Contributing Publications

This dissertation accumulates the findings of our research on affective automotive user interfaces based on five peer-reviewed publications (four conference papers, one journal paper) and two publications currently under review. We illustrate the bigger picture in which our work is located and discuss its implications from a further distance than possible within the single papers. Contributing publications are marked with a prefixed [P] throughout the document and are available by following the DOIs provided below.

[P1] David Bethge, Thomas Kosch, Tobias Grosse-Puppendahl, Lewis L. Chuang, Mohamed Kari, Alexander Jagaciak, and Albrecht Schmidt. *VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time*. In: *The 34th Annual ACM Symposium on User Interface Software and Technology*. UIST '21. Association for Computing Machinery, 2021, pp. 638–651. DOI: 10.1145/3472749.3474775. URL: https://doi.org/10.1145/3472749.3474775.

[P2] David Bethge, Philipp Hallgarten, Tobias Grosse-Puppendahl, Mohamed Kari, Lewis L. Chuang, Ozan Özdenizci, and Albrecht Schmidt. *EEG2Vec: Learning Affective EEG Representations via Variational Autoencoders*. In: *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2022, pp. 3150–3157. DOI: 10.1109/SMC53654.2022.9945517.

[P3] David Bethge, Philipp Hallgarten, Tobias Grosse-Puppendahl, Mohamed Kari, Ralf Mikut, Albrecht Schmidt, and Ozan Özdenizci. *Domain-Invariant Representation Learning from EEG with Private Encoders*. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 1236–1240. DOI: 10.1109/ICASSP43922.2022.9747398.

[P4] David Bethge, Philipp Hallgarten, Ozan Özdenizci, Ralf Mikut, Albrecht Schmidt, and Tobias Grosse-Puppendahl. *Exploiting Multiple EEG Data Domains with Adversarial Learning*. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022, pp. 3154–3158. DOI: 10.1109/EMBC48229.2022.9871743.

[P5] David Bethge, Luis Falconeri Coelho, Thomas Kosch, Satiyabooshan Murugaboopathy, Ulrich von Zadow, Albrecht Schmidt, and Tobias Grosse-Puppendahl. *Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context*. In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.4 (2023). DOI: 10.1145/3569466. URL: https://doi.org/10.1145/3569466.

[P6] David Bethge, Constantin Patsch, Philipp Hallgarten, and Thomas Kosch. *Interpretable Time-Dependent Convolutional Emotion Recognition with Contextual Data Streams*. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. Association for Computing Machinery, 2023. DOI: `10.1145/3544549.3585672`. URL: `https://doi.org/10.1145/3544549.3585672`.

[P7] David Bethge, Daniel Bulanda, Adam Kozlowski, Thomas Kosch, Albrecht Schmidt, and Tobias Grosse-Puppendahl. *HappyRouting: Learning Emotion-Aware Route Trajectories for Scalable In-The-Wild Navigation*. In: *submit to ACM Conference on Human Factors in Computing Systems* (2024). submitted.

1.4 Guiding Research Questions

This thesis consists of contributions of empirical, conceptual and technical nature, structured into research questions:

RQ1: How can we design machine learning systems for human state detection based on multimodal context sensing?

RQ2: How can machine learning systems be designed to provide meaningful explanations of their classification of human states?

RQ3: How can the required information for machine learning systems for human state classification be enabled to increase privacy?

RQ4: How can machine learning systems be applied to robustly predict human emotions in dynamic and constrained environments?

RQ5: How can emotion-aware and emotion-sensitive systems be designed to improve a user's experience?

1.5 Research Approach

This section presents the operating approach for designing, implementing, and evaluating machine learning systems for human emotional states.

Research Settings We perform multiple research settings depending on the objective of the emotion study. At the start of each research setting, we reviewed relevant literature to obtain an understanding of open research questions.

In two projects [P1, P5], we used an in-the-wild setting to acquire in-car emotions. We target obtaining real-world environment data to deduct in-situ context and emotion understanding. These experiments required us to review our experimental design as the drivers were required to drive on public roads. Initially, we reviewed our experiment design for safety and privacy concerns and obtained approval from a university institutional review board. We designed the data-gathering device with a minimum graphical user interface so that drivers could concentrate on first-level driving tasks, and safety concerns were reduced. The participants were recruited by an institutional mailing list or co-workers who were frequently willing to participate in a study.

For the emotion-aware application research project [P7], we performed an in-the-wild driving assessment of our developed application and performed a driver survey after the ride. Furthermore, we evaluated the general user needs for this emotion-aware application using an online survey. This procedure enabled us to recruit participants from diverse backgrounds and communities. We used the results of this online survey to refine our emotion-aware application user interface and model. Furthermore, it provided us with critical input for the discussion.

Three of our studies [P4, P3, P2] investigate machine learning models for scalable and domain-aware emotion recognition systems. In these studies, we used already available open-source datasets commonly used in the communities. The evaluation of these studies was done by quantitatively comparing our model results to other models' prediction capabilities.

Machine Learning Modeling Many of our research settings included some data analysis and modeling phases. The data analysis included techniques from the field of exploratory data science. We used the python programming language and deep learning frameworks such as Tensorflow and PyTorch to develop novel machine learning systems. We trained the models mostly in a private cloud setting where massive data amounts can be stored and training time can be reduced. The majority of our machine learning models are made available open-source.

Prototyping We developed several prototypes in the sphere of data acquisition [P1, P5], real-time emotion prediction [P1], and emotion-aware interfaces [P7]. All prototypes were implemented as mobile apps to be easily deployed in an in-the-wild environment. Since the prototypes contained a complex stack of software, we made most parts available open-source to facilitate research in this research area.

Two prototypes for collecting in-the-wild driver emotions while acquiring context data were developed for iPhones using the Swift framework. We gathered multiple contextual properties in-the-wild using a mix of already-available geocoding APIs (such as Microsoft Maps and OpenStreetMaps) and internally developed reverse-geocoding software. For some features, such as real-time prediction, we used software packages with additional mobile connectivity features to connect to a cloud backend to retrieve predictions.

Data Collection We used questionnaires to gather information about the personality, user experience, and general demographics. We used these questionnaires to evaluate the effectiveness of our system performance [P7]. We also developed a simulation study to research the effect of hyperparameters on our developed machine learning systems [P7]. Therein, we simulated different parameter settings and analyzed the dependence of the parameter setting on the system's behavior and model output. Additionally, we relied on information gathered through subjectively felt emotional data on the user's behavior when driving in-the-wild. We included the analysis of facial expressions [P1, P5].

Participants We recruited participants mainly from company sources for the in-the-wild research projects [P1, P5, P7]. We used an internal company email to invite participants from Porsche. Working with employees made it easier to show new prototypes and provided us with the easy accessibility of a car and related insurance.

1.6 Research Context

I had the immense pleasure of conducting my doctoral research in residence at the Porsche IT Department for Emerging Technologies in close cooperation with the LMU Munich Chair of Human Ubiquitous Media.

All experimental studies were conducted at Porsche. Porsche funded all project expenses and my salary. Many projects at this time were realized in cooperation with researchers from other institutions. I collaborated with my peers at TU Darmstadt, Utrecht University, CODE University Berlin, Humboldt University, Karlsruhe Institute of Technology, Technical University of Munich, and TU Graz. I received incredible support from these excellent people, for which I am very grateful. I also collaborated with the excellent GrapeUp automotive startup team to implement the emotion-aware application project.

2 Background

> *"A computer that can express itself emotionally will some day act emotionally, and the consequences will be tragic."*
>
> *Rosalind W. Picard, Affective Computing, 1997*

This chapter conveys background information on the fundamental concepts of machine learning systems for human emotional states. First, we introduce the concept of emotion to describe emotions adequately. Second, we present a short overview of emotion state informative signals and detection models. Third, we describe proposed machine learning models for emotions and emotion-aware interfaces. We keep this background section brief to maintain clarity.

2.1 Concepts of Human Emotions

Emotions are complex psycho-physiological ongoings that affect humans' behavior, perception, and interaction. Human emotions are hard to describe, and there exist multiple representation models[1].

Several attempts to describe emotional states have been proposed, while we propose the two most popular ones: (1) emotions can be described as discrete categories e.g., by the basic emotions 'anger', 'disgust', 'happiness', 'sadness', 'surprise', and 'fear' proposed by Ekman [9] or (2) emotions can be represented by dimensional ratings using Russel's [29] circumplex model of affect including the two dimensions valence (positiveness) and arousal (intensity). Figure 2.1 shows the dimensional emotion representation model being including the discrete emotion states. As with any theory, extensions [4] and limitations of the emotion description models have been noted [26].

2.2 Emotion Detection and Emotion Prediction

A critical pillar of emotional intelligence is the ability to detect, react, and manage emotions. Some researchers argue that detecting emotions is even more important than mathematical and verbal intelligence as this is part of daily human interactions [24].

[1] The terms emotion and affect has been used interchangeably in literature and is explained in detail in the paper 'The human Affectome' [30]. In general, affect is the superordinate category - an umbrella term that covers emotions and moods. Emotions and moods are primarily differentiated by duration and whether they are directed at a specific cause. Emotions are short-lived and intense experiences elicited in response to specific external stimuli (i.e., objects or events) and may arise relatively automatically or following a concrete stimulus [25]. For the sake of brevity, we will mainly use the term emotions.

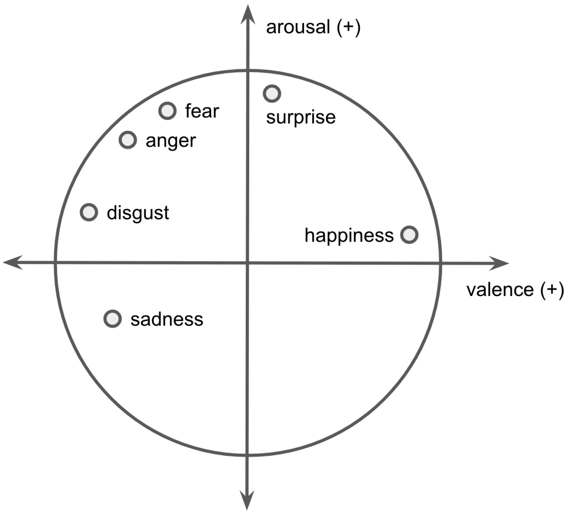


Figure 2.1: Dimensional emotion representation model by [9] showing a 2D space of emotional states characterized by valence and arousal. In the circumplex model, emotional states can be represented at any level of valence and arousal. The discrete emotion categories of [9] are mapped into the continuous space (own elaboration based on [28]).

2.2.1 General Approach

In general, emotions are assessed by finding patterns in human behavior that correlate with emotional expression. In effect, this reduces the task of studying emotion detection models to a supervised learning problem, by which a set of input signal values from, e.g., human physiological signals are transformed to predict an observed emotional state, i.e., the label. This mapping operation ($f : f(X) = y$) from the input signal (X) to the emotional state[2] (y) can be described with arbitrary complexity:

$$X \xrightarrow{f(\cdot)} y \tag{2.1}$$

$$\text{sensor observation} \xrightarrow{\text{machine learning}} \text{emotion label} \tag{2.2}$$

Input Modalities A wide variety of input modalities has been used to assess the emotional states of humans, including physiological signals, facial expressions, self-reports, voice analysis, or contextual signals [40]. We will present the most prominent input modalities. Analyzing the person's facial expression using a camera pointed at the human's face has been researched intensely [16]. Furthermore, analyzing the human's speech signals (e.g., the semantics of spoken text and pitch of voice) from a microphone has been used to classify emotions [2]. Other signals aim at accessing body vital signs such as heart rate variability or galvanic skin responses to deduct emotion-relevant information [31]. In recent times, researchers have focused on getting more introspective signals from the emotional ongoings of the brain via fNIRS or EEG technologies [12, 38]. Recent research also derives context, e.g., via accessing the smartphone to infer emotions [P1, 37]. In many settings, multi-modal sensor input (e.g., face and voice analysis) to detect emotion is used [11].

[2] This thesis allows the label space y to be continuous or discrete and therefore makes no assumption whether $f(\cdot)$ is defined as a regression or classification problem.

Emotion Labels Methods for obtaining the emotional label can be classified into two main groups: 1) self-report techniques based on emotions self-assessment via, e.g., a questionnaire or 2) machine assessment techniques based on measurements of various parameters of the human body, e.g., heart rate variability. While the self-reporting techniques give deep insights into the subjective, introspective emotional states, machine-assessment techniques offer a more objective view of the person [7, 36].

Machine Learning Models A multitude of heuristics and machine learning models have been proposed to learn a mapping function $f(\cdot)$ of input signals X to infer emotions y. Depending on the input signals used for emotion classification, different preprocessing steps and architectures of machine learning models are used to classify the emotional label (see [41] for an overview). As the mapping $f(\cdot)$ can be arbitrarily complex, machine learning methods of diverse flexibility have been applied to classify emotions. Most commonly, a variant of deep neural networks extracts meaningful representations from the input signals, which are then used for classification. Many researchers use convolutional neural networks for analyzing image inputs (such as facial camera streams) or fast-Fourier-transformed heatmaps of time-series signals. Other common machine learning architectures for emotion classification include recurrent neural networks such as LSTMs, graph-neural networks, SVMs, or "weak learners" such as Random Forest or Gradient-Boosting classifiers.

2.2.2 Difficulties

Defining machine learning models for emotion detection is inherently difficult. Multiple challenges occur due to input sensor qualities, robustness of emotional labels, and general learning settings. This section will describe the most frequently encountered difficulties in learning machine learning systems for human emotions.

Availability of Datasets One major obstacle in developing machine learning systems for human emotional states is the lack of high-quality datasets. Currently, several limitations exist, such as the lack of open-source sharing willingness and low-quality and modality-limited input data in real-life environments. In effect, the need for datasets hinders the creation of real-world complexity human emotional state detection systems and emotion-aware applications.

Missing Ground-Truth Labels One major limitation is the lack of ground-truth labels. The emotional label is often assigned by third-person labelers rating the emotional expression, facial expression classifying systems, or by instructing the person to act specific emotions. In these specific cases, the true subjectively felt emotion, i.e., ground truth, cannot be inspected. However, the ground truth can be assessed when asking the participants about their subjectively felt emotions.

Real-Life Data Another drawback of many datasets corresponds to their real-world applicability. Many datasets, especially those that track physiological signals to estimate emotional states, are acquired in a lab environment. Though lab environment offers a more-controlled setting, the data and model's transferability to in-the-wild settings is often substantial. Therefore, the gathering of real-life in-the-wild data is often necessary. Measuring and adapting to emotions during natural settings remains difficult due to many challenges (e.g., uncomfortable sensors, unreliable ground truth data, uncontrolled environment). This thesis will present a machine learning architecture, and input signal sets to predict human emotions robustly in dynamic and unconstrained environments (**RQ4**).

Extensive Samples For machine learning systems to successfully extract emotion-relevant input signal patterns, the dataset should consist of enough samples which cover a variety of input and label space. More extensive training data usually results in a better modeling approximation and hence possesses lower estimation variance and, consequently, better predictive performance. Furthermore, sufficient subject samples are required to adapt to a specific person's characteristics and learn person-dependent models. Many times, obtaining a large number of emotional samples for an individual participant is time-consuming and requires high effort. This hinders the learning of personalized and thus possibly more privacy-preserving emotion models (**RQ3**).

Context One of the biggest obstacles has been the need for context: the fact that emotions cannot be understood in isolation. To improve understanding of why a specific emotional state is observed and to further improve detection performance, machine learning systems could heavily improve in incorporating the environmental context. Recent breakthroughs in ubiquitous sensing allow in-the-wild data, including real-world context, to inform emotion detection models. This thesis presents a mobile and personal computing software that predicts emotions based on contextual data only (**RQ1**).

Interpretability Although there are approaches to link input data to emotional states, understanding the relationship between emotion-relevant data to detected emotion is difficult. Traditional machine learning methods aim to provide an accurate prediction but offer little interpretation of its output. Therefore providing meaningful explanations of an emotion prediction output remains a difficulty. This thesis tackles the problem of proposing a system for providing explanations for emotion prediction using deep neural network architectures (**RQ2**).

Privacy User privacy is a major concern when developing emotion detection systems and applications. First, emotions are very personal such that users may prefer to share their data, and person-dependent modeling may be required. Second, machine learning models for human emotional states in emotion-aware applications should be treated cautiously because emotions can be considered private. Third, this thesis proposes research on machine learning systems that enable the learning of more privacy-preserving machine learning models (**RQ3**).

2.2.3 Milestones

Significant milestones in the area of human emotional state detection and application will be presented in the following section.

Measuring the user's emotions is a compelling topic that has been addressed by previous research. Picard coined the term 'Affective Computing', envisioning computers to express or sense emotions to provide a computerized interface that mimics human-like capabilities [21, 23].

One prominent breakthrough was the automatic extraction of facial features to predict emotional states [8]. Although the robustness of facial expression emotion detection is highly controversial, this method enabled many emotion-aware applications to use a cheap camera to infer emotional states.

Another milestone in the emotion detection domain was the successful extraction of emotion patterns from neurological brain-computer interface data [44]. Analyzing signals such as EEG offers an "inner" recording via a non-intrusive electrical recording of variations, locations, and functional interactions of brain activity. Here, one substantial challenge in the robustness of the label and input space variability across persons persists.

3 Publications

"Any sufficiently advanced technology is indistinguishable from magic."

Arthur C. Clarke, A Space Odyssey.

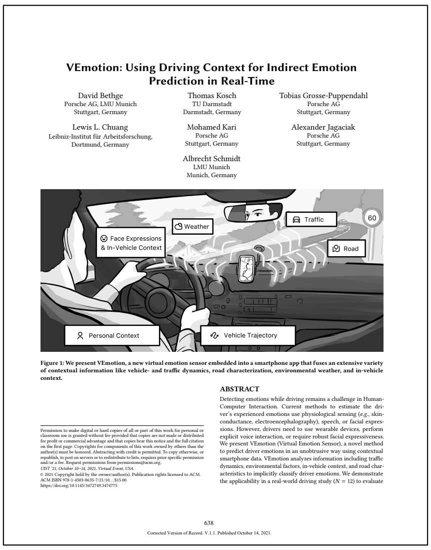
This chapter presents our contributions to the field of machine learning for human emotional states. We summarize an overview of the publications by research questions, method, and contribution type in Table 3.1. The contribution types follow Laudan's taxonomy [15] adapted for HCI by Ouslavirta et al. [19] listed in order of presentation in this dissertation. The publications at the top of the table focus on understanding user emotions and their link to in-the-wild context. The publications towards the bottom of the table shift their focus increasingly towards building machine learning systems, building and testing applications for human emotions. The publications are grouped into four sub-domains: 'Intelligent Input Modalities', 'Explanatory Emotion Prediction Models', 'Privacy-Preserving Human-Emotion Models', and 'Human-Emotion Application in Dynamic Environments'. We present the publications in the order of the research questions.

3.1 Intelligent Input Modalities

Recent developments in ubiquitous sensing and the emergence of AI enable the use of machine learning for human state detection systems. The objective of [P1] is to research how a machine learning system for human emotional states can be built based on contextual features only. The subsequent paper [P5] aim is to deduct a technical design space for unobtrusive human emotional state detection systems.

RQ1: How can we design machine systems for human state detection based on multimodal context sensing?

VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time [P1] This paper presents a novel input modality to estimate driver emotions in an in-the-wild driving setting. We define a novel virtual emotion sensor (VEmotion) and define processing steps to predict driver emotions in an unobtrusive way using contextual smartphone data. We construct intelligent human state input by analyzing traffic dynamics, environmental factors, in-vehicle context, and road characteristics to classify driver emotions implicitly. We demonstrate the applicability in a real-world driving study ($N = 12$) to evaluate emotion prediction performance. Facial expression classification only often yields a 'neutral' emotion class due to the inexpressiveness of the driver, ignoring and misclassifying heavily other felt emotions of the driver. Our machine learning model



VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time

David Bethge
Porsche AG, LMU Munich
Stuttgart, Germany

Thomas Kosch
TU Darmstadt
Darmstadt, Germany

Tobias Grosse-Puppendahl
Porsche AG
Stuttgart, Germany

Lewis L. Chuang
Leibniz-Institut für Arbeitsforschung,
Dortmund, Germany

Mohamed Kari
Porsche AG
Stuttgart, Germany

Alexander Jagaciak
Porsche AG
Stuttgart, Germany

Albrecht Schmidt
LMU Munich
Munich, Germany

Figure 1: We present VEmotion, a new virtual emotion sensor embedded into a smartphone app that fuses an extensive variety of contextual information like vehicle- and traffic dynamics, road characterization, environmental weather, and in-vehicle context.

ABSTRACT

Detecting emotions while driving remains a challenge in Human-Computer Interaction. Current methods to estimate the driver's experienced emotions use physiological sensing (*e.g.*, skin-conductance, electroencephalography), speech, or facial expressions. However, drivers need to use wearable devices, perform explicit voice interaction, or require robust facial expressiveness. We present VEmotion (Virtual Emotion Sensor), a novel method to predict driver emotions in an unobtrusive way using contextual smartphone data. VEmotion analyzes information including traffic dynamics, environmental factors, in-vehicle context, and road characteristics to implicitly classify driver emotions. We demonstrate the applicability in a real-world driving study ($N = 12$) to evaluate

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
UIST '21, October 10–14, 2021, Virtual Event, USA
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8635-7/21/10…$15.00
https://doi.org/10.1145/3472749.3474775

638

Corrected Version of Record. V.1.1. Published October 14, 2021.

Table 3.1: Overview of publications organized by research question, methods, and contribution type.

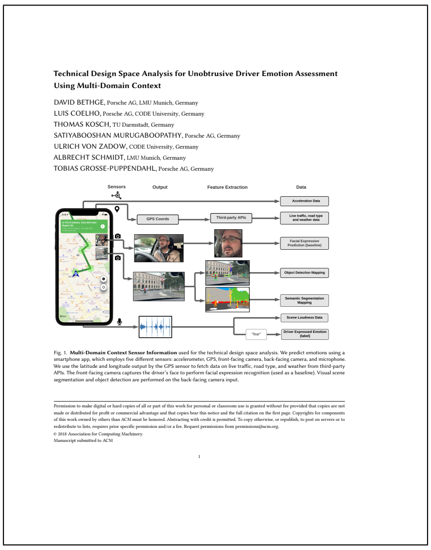
				Contribution			
Paper	Type	Research Question	Method(s)	Empirical	Conceptual	Constructive	Key Outcome
RQ1: *How to design machine learning systems for human state detection based on physiological sensing and context?*							
[P1]	UIST'21 full conference paper	How can we predict emotions using context?	prototyping, implementation ($N = 13$)	evaluation of system	learning architecture, evaluation systems, participant fine-tuning	implementation of unobstrusive mobile machine learning system	novel virtual emotion sensor
[P5]	IMWUT'23 journal paper	What is the technical design space for human-state recognition systems?	implementation, in-the-wild study ($N = 26$)	analysis of context emotion relationship	synthesis of important and privacy-relevant features for emotion classification	technical design space for development of human state machine learning systems	dataset for emotions and context in-the-wild
RQ2: *How can machine learning systems be designed to provide meaningful explanations of their classification of human-states?*							
[P6]	CHI'23 workshop paper	How can machine learning systems provide time-granular emotion state explanations?	implementation	comparative analysis on multi-context emotion database	interpretable time-dependent feature explanations	CNN Transformer architecture with attention mechanisms	novel interpretable neural network architecture for explainable emotion prediction
RQ3: *How can the required information for machine learning systems for human state classification be enabled to increase privacy?*							
[P3]	ICASSP'22 full conference paper	How can we learn common emotion representations being data-source independent?	implementation, quantitative evaluation	analysis of shared latent space across emotional input	private encoding mechanisms for shared emotion latent space generation	novel encoder architecture for high signal-to-noise emotion data input	relevant information extraction method from multiple EEG data domains
[P4]	EMBC'22 full conference paper	How can we exploit multiple emotion sources to learn a common embedding?	implementation, quantitative evaluation	analysis of shared latent space across emotional input	adversarial multi-source learning for emotions	weight updating mechanisms for emotional machine learning systems with contrarian objectives	relevant information extraction method from multiple EEG data domains
[P2]	SMC'22 full conference paper	How can we learn a general purpose emotion representation?	implementation, quantitative evaluation	analysis of shared latent space for data generation and state prediction tasks	trade-off of participant variability and prediction performance of emotion machine learning systems	understanding of emotional latent space for generation and prediction	relevant information extraction method from emotional EEG data
RQ4: *How can machine learning systems be applied to robustly predict emotions in dynamic and unconstrained environments?*							
RQ5: *How can emotion-aware and emotion-sensitive systems be designed to improve a user's experience*							
[P7]	to be published	How can use emotion prediction to navigate users in real-world environments?	implementation, quantitative evaluation ($N = 12$)	effect of emotional navigation on perception	trade-off between happiness machine learning prediction and traveltime	scalable, mobile emotion-navigation system	novel emotion-aware application

using only contextual data could predict emotions confidently with an accuracy of 72% and F_1 score of 0.71. Our results show that VEmotion outperforms facial expressions by 29% in a person-dependent classification and 8.5% in a person-independent classification. Our results show that contextual information can significantly improve the classification of emotional states, especially in detecting 'surprise' situations and discriminating between 'neutral' and 'happiness' states. We show that we can learn a global system for recognizing emotions 'on the go' with contextual and facial expressions. However, this comes at higher computational costs of accessing all participants' data and learning a participant-independent classifier. We also learn an uncalibrated model that can predict emotions for a new driver's emotions. The emotions predicted by our machine learning classifier improve if more person-dependent information is available. Analyzing the feature importances, we found that 'vehicle dynamics', 'weather', and 'traffic flow' were highly predictive of emotions. This implies that the designer of empathic car interfaces should focus on the reliable measurement of these features when assessing emotions is a critical task. We discuss how VEmotion enables empathic car interfaces to sense the driver's emotions and will provide in-situ interface adaptations on the go.

Authors contribution: I came up with the original research idea of using context for predicting driver emotions and was the lead-author. I developed the machine learning pipeline and supervised Alexander Jagaciak, who implemented the experimental iOS app for gathering in-the-wild context and emotion data. Thomas Kosch came up with the experimental study design and provided a paper review. Lewis Chuang reviewed the paper and outlined ethical limitations of the emotion detection approach. Tobias Grosse-Puppendahl provided research guidance throughout the paper writing and reviewed the paper. Mohamed Kari implemented the beta-testing framework for our study-application and provided a paper review. Albrecht Schmidt provided feedback on the paper and the publication.

Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context [P5]

This paper describes a technical-design space analysis for remote sensing human states in a dynamic driving environment. In a user study, we investigate how emotions can be unobtrusively predicted by analyzing a rich set of contextual features captured by a smartphone, including road and traffic conditions, visual scene analysis (i.e., the outside and inside view), audio, weather information, and car speed. We derive a technical design space to inform practitioners and researchers about the most indicative sensing modalities, the corresponding impact on users' privacy, and the computational cost associated with processing this data. Our analysis shows that contextual emotion recognition is significantly more robust than facial recognition, leading to an overall improvement of 41% using a participant-dependent cross-validation.Finally, we present a technical framework showing how contextual and audio-visual sensing modalities influence the accuracy of emotion classification.



Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context

DAVID BETHGE, Porsche AG, LMU Munich, Germany
LUIS COELHO, Porsche AG, CODE University, Germany
THOMAS KOSCH, TU Darmstadt, Germany
SATIYABOOSHAN MURUGABOOPATHY, Porsche AG, Germany
ULRICH VON ZADOW, CODE University, Germany
ALBRECHT SCHMIDT, LMU Munich, Germany
TOBIAS GROSSE-PUPPENDAHL, Porsche AG, Germany

Sensors
Output
Feature Extraction
Data
Acceleration Data
GPS Coords
Third-party APIs
Live traffic, road type and weather data
Facial Expression Prediction (baseline)
Object Detection Mapping
Semantic Segmentation Mapping
Scene Loudness Data
"fear"
Driver Expressed Emotion (label)

Fig. 1. **Multi-Domain Context Sensor Information** used for the technical design space analysis. We predict emotions using a smartphone app, which employs five different sensors: accelerometer, GPS, front-facing camera, back-facing camera, and microphone. We use the latitude and longitude output by the GPS sensor to fetch data on live traffic, road type, and weather from third-party APIs. The front-facing camera captures the driver's face to perform facial expression recognition (used as a baseline). Visual scene segmentation and object detection are performed on the back-facing camera input.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
Manuscript submitted to ACM

1

Our work discusses how designers can select sensing strategies to prototype empathic interfaces considering trade-offs related to computational cost and privacy concerns. For example, users can opt-in for contextual data only and leave out the environmental data in case of privacy concerns. In the setting of a production car, the manufacturer should focus on an easy-to-compute and privacy-preserving set of features. Hence we recommend a feature set without sensitive data and where all features are preferred to be locally computable and third-party API independent. For research purposes and apart from GPS-features, we propose to use computer vision extracted features, i.e., object detection and visual scene segmentation features. They show high feature importance while also offering possible local computability through on-device computer vision inference. Third, research on driver-view context affecting driver's well-being is underexplored [3]. Overall, we do not recommend empathic application designers to acquire emotions through facial expression analysis due to their non-robust detection and privacy-related concerns [34]. Still, many car companies employ facial monitoring software as driver-facing cameras are already equipped in-car, and facial expression software is easy to integrate [18]. Although our work intelligent input work for emotion detection is demonstrated in a car setting, the findings can be translated to other user interface systems where these inputs are available. Since the results of our study are obtained using a smartphone only, we envision that developers and designers can inexpensively prototype novel empathic interfaces using the evaluated data streams.

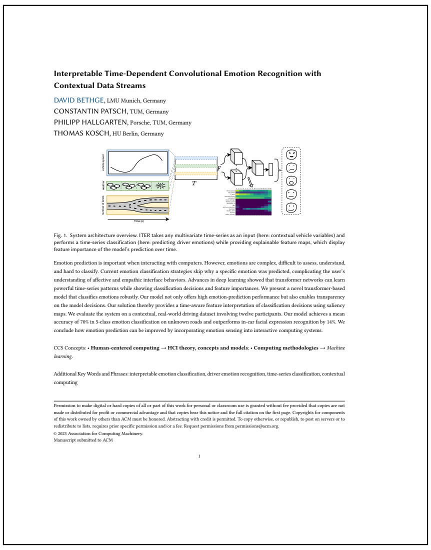
Authors contribution: I came up with the original research idea. Together with Luis Coelho, we extended the input modalities of the VEmotion system [P1] to include audio-visual features. Luis Coelho implemented the data acquisition smartphone app. Luis Coelho and I evaluated the framework. Satiyabooshan Murugaboopathy designed the privacy-awareness table and performed acquisition sessions. Thomas Kosch and Tobias Grosse-Puppendahl provided feedback throughout the paper writing and evaluation stages. Ulrich von Zadow guided the wording of the paper by ensuring conciseness. Albrecht Schmidt reviewed the paper.

3.2 Explanatory Emotion Predictions Models

While better machine learning modeling accuracy is of great benefit, the interpretability of a model is crucial for HCI researchers to gain new knowledge and to advance the field. The objective of **RQ2** is to analyze and define ways for human-emotional state detection systems to provide meaningful explanations. In particular, [P6] demonstrates a novel interpretability mechanism to analyze the model's reasoning "when, where and why" a decision has been made. The proposed methodology in [P6] for generating interpretable feature maps applies to a wide range of HCI scenarios. Using the proposed mechanism, we could analyze which contextual feature changes induced an emotional change and thus infer specific (contextual) emotional triggers.

RQ2: How can machine learning systems be designed to provide meaningful explanations of their classification of human states?

Interpretable Time-Dependent Convolutional Emotion Recognition with Contextual Data Streams [P6] This paper proposes a method to extract the temporal explainability of human-emotional state detection models. Especially when dealing with the challenge of emotion recognition in noisy environments, state-of-the-art approaches utilize mainly physiological and facial data. However, emotions are complex, difficult to assess, understand, and hard to classify. Current emotion classification strategies do not reveal why a specific emotion was predicted, making it difficult for the user to understand the interface behavior. Advances in deep learning showed that transformer networks can learn powerful time-series patterns while showing classification decisions and feature importances. We present a transformer-based deep learning model that classifies emotions robustly. Our model not only offers high emotion-prediction performance but also enables transparency on the model decisions. Our model thereby provides a time-aware feature interpretation of classification decisions using saliency maps. We focus on visual interpretability in the form of saliency maps as they display feature time dependencies and propose a new method to derive activation aggregation. 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. 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%. Finally, we conclude how the proposed post-hoc visualizations help to improve opening the “black-box” emotion prediction model and propose applications of our model in interactive computing environments.



Interpretable Time-Dependent Convolutional Emotion Recognition with Contextual Data Streams

DAVID BETHGE, LMU Munich, Germany
CONSTANTIN PATSCH, TUM, Germany
PHILIPP HALLGARTEN, Porsche, TUM, Germany
THOMAS KOSCH, HU Berlin, Germany

Fig. 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.

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 transformer networks can learn powerful time-series patterns while showing classification decisions and feature importances. We present a novel transformer-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*.

Additional Key Words and Phrases: interpretable emotion classification, driver emotion recognition, time-series classification, contextual computing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Association for Computing Machinery.
Manuscript submitted to ACM

1

Author contribution: I came up with the concept idea to make temporal influences of input on the road for the emotion prediction explanatory. I developed use-cases for the application of interpretable time-dependent emotion recognition and was the lead author for the paper. Together with Constantin Patsch we developed the neural network architecture, which Constantin Patsch evaluated. Philipp Hallgarten and Thomas Kosch contributed to the overall conciseness by reviewing the paper.

3.3 Privacy-Preserving Human-Emotion Models

While it is promising to use deep learning methods in human behavior modeling, there are several challenges. Foremost, deep learning methods are often data hungry, while interaction data is scarce compared to classic machine learning problems such as computer vision or natural language processing. **RQ3** aims to research the question of how to build privacy-constrained human-state systems that are still able to predict human states confidently. We concentrate on a particular aspect of privacy by researching machine learning methods for making emotion prediction domain-independent. By learning domain-invariant representations for emotion classification, we preserve emotion-relevant information but constrain the privacy-related information specific to a particular data domain.

In [P3] and [P4], the aim is to provide novel machine learning architectures to learn from multiple HCI data sources to learn common latent emotion representations which can predict emotions and be data source invariant. The objective of [P2] is to learn an emotion embedding that does not include participant-dependent information but is predictive for emotion and can generate synthetic data. The presented papers outline a novel way to create deep learning for human-state detection by exploiting multiple databases with increased privacy requirements.

RQ3: How can the required information for machine learning systems for human state classification be enabled to increase privacy?

Domain-Invariant Representation Learning from EEG with Private Encoders [P3] This paper presents a multi-source learning architecture where we extract domain-invariant representations from dataset-specific private encoders for human-emotion modeling. Our model utilizes a maximum-mean-discrepancy (MMD) based domain alignment approach to impose domain-invariance for encoded representations. Our framework consists of a private feature encoder per domain and a cross-domain shared classifier. The novelty of our approach is that we utilize a maximum-mean discrepancy (MMD) [10] based domain alignment loss across private feature encoders to minimize domain-specific leakage within the learned representations. Our method outperforms state-of-the-art approaches in EEG-based emotion classification, such as domain-specific EEG learning, global modeling, or adversarial techniques. Although our method is evaluated on emotion-related EEG data, the machine learning framework can be applied to any multivariate time series input. Thus, our model could also be used for estimating other human states, e.g., driver fatigue estimation using heart-rate time series data.

DOMAIN-INVARIANT REPRESENTATION LEARNING FROM EEG WITH PRIVATE ENCODERS

David Bethge[1,2], *Philipp Hallgarten*[1,3], *Tobias Grosse-Puppendahl*[1], *Mohamed Kari*[1], *Ralf Mikut*[3], *Albrecht Schmidt*[2], *Ozan Özdenizci*[4,5]

[1] Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany
[2] Ludwig-Maximilians University Munich, Germany
[3] Karlsruhe Institute of Technology, Germany
[4] Institute of Theoretical Computer Science, Graz University of Technology, Austria
[5] TU Graz - SAL Dependable Embedded Systems Lab, Silicon Austria Labs, Austria

ABSTRACT

Deep learning based electroencephalography (EEG) signal processing methods are known to suffer from poor test-time generalization due to the changes in data distribution. This becomes a more challenging problem when privacy-preserving representation learning is of interest such as in clinical settings. To that end, we propose a multi-source learning architecture where we extract domain-invariant representations from dataset-specific private encoders. Our model utilizes a maximum-mean-discrepancy (MMD) based domain alignment approach to impose domain-invariance for encoded representations, which outperforms state-of-the-art approaches in EEG-based emotion classification. Furthermore, representations learned in our pipeline preserve domain privacy as dataset-specific private encoding alleviates the need for conventional, centralized EEG-based deep neural network training approaches with shared parameters.

1. INTRODUCTION

Over the past decades, electroencephalography (EEG) based emotion recognition gained significant interest towards developing affective neural-machine interfaces [1]. However, due to the costly data collection processes, large EEG datasets recorded under emotion eliciting experimental paradigms are not ubiquitous. Several small scale affective EEG datasets were previously introduced, which however highly differ in their experimental setups (*e.g.*, stimuli, emotion labels). This introduces the well-known *domain adaptation* problem, i.e., models trained to recognize emotions on one of these datasets commonly fails to solve the same task for another dataset. In order to make EEG-based emotion recognition algorithms suitable for a variety of experiments and real-life scenarios, it is crucial to obtain a model that can tackle this task across multiple datasets. One approach to achieve this from a deep learning perspective is to extract and exploit domain-invariant representations from multi-channel EEG data.

Recent work have shown significant promise in using invariant representation learning from time-series EEG data for cross-subject generalization [2–4], while learning across multiple dataset sources remains an open question. Notable methods successfully introduced adversarial censoring for domain invariance, widely known as DANN [5], in various EEG decoding tasks [2, 3], including emotion recognition [4]. Adversarial domain regularization with DANN [5] considers a cross-domain shared encoder that extracts features from data of all subjects, and two classifiers: a task and a domain classifier. During training the encoder is adversarially penalized on the domain classification loss, which enforces the model towards learning domain-invariant representations. In this work we take a different approach to invariant EEG representation learning by further considering to preserve domain privacy that is of critical importance in clinical settings [6, 7].

We propose a multi-source learning framework for domain invariant representation learning from time-series signals such as multi-channel EEG recordings. From a different perspective than adversarial learning methods, our framework consists of a private feature encoder per domain and a cross-domain shared classifier, where we utilize a maximum-mean-discrepancy (MMD) [8] based domain alignment loss across private feature encoders to minimize domain-specific leakage within the learned representations. Our contributions in this work are three-fold: (1) We introduce a deep neural signal processing architecture where EEG time-series are privately encoded at the source end and only the learned representations are shared to a global classifier that enables decentralized cross-dataset learning by preserving domain privacy. (2) We reveal large dataset domain specific variances in conventionally trained centralized pipelines, and demonstrate that regularizing latent representations via an MMD-based domain alignment loss enables dataset source independent representation learning. (3) We show that the use of adaptive batch normalization layers in such multi-source settings prevent performance decrease at inference time.

Authors contribution: I came up with the original research idea and reserarch contribution. Together with Philipp Hallgarten, we developed the machine learning architecture for domain alignment of the data source invariance for emotional representation learning. Ozan Özdenizci contributed ideas to improve the convergence of the learning architecture and supervised the paper writing. Mohamed Kari contributed feedback throughout the paper writing process. Tobias Grosse-Puppendahl, Ralf Mikut and Albrecht Schmidt reviewed the paper.

Exploiting Multiple EEG Data Domains with Adversarial Learning [P4] This paper proposes the idea of privacy-aware multi-source learning via learning domain-invariant representations. We argue that learning emotion representations from multiple data sources is a viable alternative, as the available data from different EEG data-source domains (e.g., subjects, sessions, experimental setups) will grow massively. We propose an adversarial inference approach to learn data-source invariant representations in this context, enabling multi-source learning for EEG-based brain-computer interfaces. We unify EEG recordings from different source domains (i.e., emotion recognition datasets SEED [43], SEED-IV [42], DEAP [14], DREAMER [13]). We demonstrate the feasibility of our invariant representation learning approach in suppressing data-source-relevant information leakage by 35% while still achieving stable EEG-based emotion classification performance.

Exploiting Multiple EEG Data Domains with Adversarial Learning

David Bethge[1,2], Philipp Hallgarten[1,3], Ozan Özdenizci[4,5],
Ralf Mikut[3], Albrecht Schmidt[2], Tobias Grosse-Puppendahl[1]

***Abstract*— Electroencephalography (EEG) is shown to be a valuable data source for evaluating subjects' mental states. However, the interpretation of multi-modal EEG signals is challenging, as they suffer from poor signal-to-noise-ratio, are highly subject-dependent, and are bound to the equipment and experimental setup used, (*i.e. domain*). This leads to machine learning models often suffer from poor generalization ability, where they perform significantly worse on real-world data than on the exploited training data. Recent research heavily focuses on cross-subject and cross-session transfer learning frameworks to reduce domain calibration efforts for EEG signals. We argue that multi-source learning via learning domain-invariant representations from multiple data-sources is a viable alternative, as the available data from different EEG data-source domains (e.g., subjects, sessions, experimental setups) grow massively. We propose an adversarial inference approach to learn data-source invariant representations in this context, enabling multi-source learning for EEG-based brain-computer interfaces. We unify EEG recordings from different source domains (i.e., emotion recognition datasets SEED, SEED-IV, DEAP, DREAMER), and demonstrate the feasibility of our invariant representation learning approach in suppressing data-source-relevant information leakage by 35% while still achieving stable EEG-based emotion classification performance.**

***Index Terms*— adversarial learning, domain invariance, EEG.**

I. Introduction

Electroencephalogram (EEG) based brain-computer interface (BCI) systems aim to identify users' intentions from brain recordings with potential uses in neurorehabilitation systems [1]. However, moderate decoding accuracies have limited the practical use of BCIs [2], [3]. Due to the high data collection efforts and costs, EEG datasets highly diverge in their recording environment (*e.g.*, stimulus), the equipment and devices, and the ground truths derived. Shortage of large and homogeneous BCI datasets limits the choice of applicable models and causes a high effort if individual models are to be used for each domain. Imbalance of EEG data source domains for classification is therefore prevalent and posing important challenges for EEG-based BCIs.

Transfer learning across different data domains as such has been extensively studied over the past decades in computer vision [4], [5], proposing convolutional neural networks (CNNs) to extract domain-invariant features for image search and classification across domains. Subsequently, transfer learning on neurophysiological recording datasets (*e.g.*, EEG) is becoming an active research field [6]. Generalized neural decoder learning for across recording modalities (multi-corpus) on electrocorticography data has been recently proposed by [7]. Their approach was shown to generalize to new participants and recording modalities, robustly handle variations in electrode placement, and allow participant-specific fine-tuning with minimal data. Also recently, [2] discussed an online pre-alignment strategy for aligning the motor imagery EEG recording distributions of different subjects before training and inference processes, and showed to significantly improve generalization across datasets. Towards a similar goal, [8], [9] proposed an invariant representation learning scheme using adversarial inference to enable cross-nuisance transfer learning in EEG signal classification with deep neural networks. Empirical assessments on EEG decoding tasks revealed that cross-subject [8] or cross-session [9] representations can be learned with such models. Cross-subject EEG transfer learning have been also explored for emotion recognition to generalize existing models to new subjects, and thereby reducing the demand for the calibration data amount effectively for new subjects [10].

In light of recent work on enabling multi-corpus learning from neurophysiological data [2], [11], [12], [13], we propose an adversarial machine learning approach to unify different raw EEG time-series and pre-process them accordingly. Unlike previous work that has focused on learning scenarios across subjects or sessions, we explore dataset-invariant representations via an adversarial learning framework that can be used in EEG multi-label settings. Our approach aims at expressing robust task-relevant EEG features in a latent representation for emotion recognition across several datasets, by limiting the representation to not learn nuisances specific to these datasets, hence being dataset invariant. We evaluate our framework against the competing baseline of a state-of-the-art deep learning encoder-classifier network trained on the unified set of all data sources.

Our contributions in this work are as follows: (1) We present a unifying EEG pre-processing framework for fusing different raw EEG time-series datasets and associated emotion state labels for transfer learning. (2) We propose an adversarial machine learning framework on multivariate EEG time-series to learn dataset-invariant representations to predict EEG class labels. (3) We present an experimental study on assembling four publicly-available EEG datasets in the field of emotion recognition, and show that our approach can learn dataset-invariant representations *i.e.*, transfer emotion-relevant EEG signals across datasets containing data from different subjects and measurement conditions.

[1] Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany.
[2] Ludwig-Maximilians University, Munich, Germany.
[3] Karlsruhe Institute of Technology, Karlsruhe, Germany.
[4] Institute of Theoretical Computer Science, TU Graz, Austria.
[5] TU Graz-SAL DES Lab, Silicon Austria Labs, Graz, Austria.
Corresponding author: P. Hallgarten (philipp.hallgarten1@porsche.de).

Our work can be extended by adapting the encoder framework to be able to use different EEG input shapes according to the specified data source, and as a result, a different number of channels and sampling frequencies can be learned. We envision an adversarial shared-private model, where some channels are shared among data sources (as in our approach) but private (data-source-specific) input can be incorporated. Our approach can also easily be adapted to learn representations invariant to other EEG variation factors, e.g., participant ID, by adding an additional adversarial classifier.

Authors contribution: I came up with the original research idea. I was the lead author of the paper and implemented the preprocessing of the domain data sources. Philipp Hallgarten implemented the adversarial learning framework. Ozan Özdenizci contributed to the resulting publication by co-supervising the development of the architecture. He revised the method and discussion section of the paper. Tobias Grosse-Puppendahl, Mohamed Kari, Ralf Mikut, and Albrecht Schmidt reviewed the paper and provided feedback to the concicesness of the paper's contribution.

EEG2Vec: Learning Affective EEG Representations via Variational Autoencoders [P2] In this paper, we explore whether representing neural data, in response to emotional stimuli, in a latent vector space can predict emotional states and generate synthetic EEG data that are participant- and emotion-specific. We propose a conditional variational-autoencoder-based framework called EEG2Vec to learn generative and emotion-discriminative representations. Experimental results on emotional EEG recording datasets demonstrate that our model is suitable for emotional time series modeling using a general-purpose latent representation. Our model achieves a robust performance of 68.49% in the emotion classification task with three distinct emotion categories (positive, neutral, negative). Furthermore, our model generates synthetic EEG sequences resembling actual EEG data inputs. Especially low-

EEG2Vec: Learning Affective EEG Representations via
Variational Autoencoders

David Bethge[1,2,*], Philipp Hallgarten[1,3], Tobias Grosse-Puppendahl[1], Mohamed Kari[1],
Lewis L. Chuang[4], Ozan Özdenizci[5,6], Albrecht Schmidt[2]

***Abstract*— There is a growing need for sparse representational formats of human affective states that can be utilized in scenarios with limited computational memory resources. We explore whether representing neural data, in response to emotional stimuli, in a latent vector space can serve to both predict emotional states as well as generate synthetic EEG data that are participant- and/or emotion-specific. We propose a conditional variational autoencoder based framework, EEG2Vec, to learn generative-discriminative representations from EEG data. Experimental results on affective EEG recording datasets demonstrate that our model is suitable for unsupervised EEG modeling, classification of three distinct emotion categories (positive, neutral, negative) based on the latent representation achieves a robust performance of 68.49%, and generated synthetic EEG sequences resemble real EEG data inputs to particularly reconstruct low-frequency signal components. Our work advances areas where affective EEG representations can be useful in e.g., generating artificial (labeled) training data or alleviating manual feature extraction, and provide efficiency for memory constrained edge computing applications.**

I. Introduction

The emphasis on human-centric computing in recent years has accelerated efforts in affective computing to develop effective computer-aided approaches for recognizing, interpreting, processing, and simulating a person's emotions. Recent notable successes include affect-adaptive robot-child feedback in education [1], mobile real-time facial emotion annotation systems [2], as well as detecting and influencing drivers' emotions [3]. In particular, the growing availability of high-resolution wearable physiological measurement systems has, in combination with powerful machine learning methods, increased recognition performance of affective user states for various applications in-the-wild [4].

In this paper, we focus on electroencephalographic (EEG) measurements that were elicited in response to affective emotional stimuli. Primarily we seek to determine whether a user-specific affective representation from raw EEG can be learned in an end-to-end *representation learning* framework. In such a setting, representations of affective states are learned from input data—typically by transforming it or extracting informative features from it (the useful vantage point of the data's key qualities)—towards the objective of performing particular tasks like prediction of affective states from noisy EEG data. Traditional discriminative machine learning approaches have the sole objective of classifying distinct affective categories. In opposition, our focus on representation learning aims at estimating a powerful abstraction of affect-relevant multi-channel sensor input. This approach encodes the signal generative components from the training data distribution in a learned latent space. Once a representation is learned, semi-supervised learning schemes can be performed by applying the representation learning framework on unlabeled data which can be combined with labeled data representations to obtain a larger training base in order to, for instance, predict emotions.

Our work is inspired by recent advances in word representations, also denoted as *embeddings* [5], [6]. A prominent success story is *word2vec* [7] in natural language processing, which uses a neural network model to learn word representations from a large text corpus. Once trained, such representation models can detect synonymous words or accurately suggest additional words for a partial sentence. This has given rise to numerous natural language applications that were previously unimaginable (e.g., predicting the right next words in chats, sentiment analysis of messages, machine translation, click session advertisement-recommendation, automatic topic clustering). State-of-the-art natural language processing algorithms can even learn cross-lingual concepts [8], generate complete texts [9], [10] or infer emotion-related text sentiments [11]. When trained with enough data, word *embeddings* tend to capture word concepts and meanings, and are even able to perform analogies, e.g., the vector for "Paris" minus the vector for "France" plus the vector for "Italy" is very close to the vector for "Rome". Thereby, word representations can bridge the human understanding of language to that of a machine.

We seek to contribute towards ubiquitous wearable physiological systems and allow for computing systems that are responsive to user affective states in real-world scenarios. Here, we focus on high-dimensional EEG data and share how a highly informative and compressed representation could be derived from it to support this vision. Once learned, such representations can be useful for downstream tasks, such as predicting emotions to better understand the affective states in the brain through representations in a lower dimensional space, or simulating synthetic EEG data.

To date, affective brain-computer interface (BCI) methodologies are often hindered by the lack of large labeled datasets, low signal-to-noise (SNR) ratio in real-time EEG data acquisition, or non-reproducible handcrafted feature ex-

[1] Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany
[2] Ludwig Maximilian University, Munich, Germany
[3] University of Tübingen, Tübingen, Germany
[4] Chemnitz University of Technology, Chemnitz, Germany
[5] Institute of Theoretical Computer Science, TU Graz, Graz, Austria
[6] TU Graz-SAL DES Lab, Silicon Austria Labs, Graz, Austria
[*] Corresponding author: david.bethge@ifi.lmu.de

frequency signal components can be reconstructed with high robustness.

Our work advances areas where affective EEG representations can be helpful to generate artificial (labeled) training data or alleviate manual feature extraction and provide efficiency for memory-constrained edge computing applications.

Authors contribution: I developed the concept idea and came up with the original research questions. I implemented the conditional variational autoencoder architecture and evaluated the framework. Ozan Özdenizci reviewed the paper, supervised the original contribution conciseness, and provided ideas for the convergence of the machine learning framework. Philipp Hallgarten visualized latent space results and reviewed the paper for conciseness. Lewis L. Chuang provided a neuroscience interpretation on the latent space findings. Tobias Grosse-Puppendahl, Mohamed Kari, and Albrecht Schmidt reviewed the paper.

3.4 Human-Emotion Application in Dynamic Environments

While designing robust human-state detection systems with high accuracy, privacy awareness, and explanations capabilities is favorable, their use in applications provides further HCI challenges. In [P7], we aim to deploy robust state detection systems in dynamic and unconstrained environments by developing the first scalable emotional-routing application. We begin by discovering the degrees of freedom to design a scalable affective navigation system applicable to unknown users, environments, and roads. Next, we demonstrate that theoretical psychological assumptions hold for the experienceable system, showing for the first time a navigation system that regulates emotions positively.

RQ4: How can machine learning systems be applied to robustly predict human emotions in dynamic and constrained environments?

RQ5: How can emotion-aware and emotion-sensitive systems be designed to improve a user's experience?

AffectRoute: Learning Emotion-Aware Route Trajectories for Scalable In-The-Wild Navigation

David Bethge
Porsche AG, LMU Munich
Stuttgart, Germany

Daniel Bulanda
GrapeUp
Warsaw, Poland

Adam Kozlowski
GrapeUp
Warsaw, Poland

Thomas Kosch
TU Darmstadt
Darmstadt, Germany

Albrecht Schmidt
LMU Munich
Munich, Germany

Tobias Grosse-Puppendahl
Porsche AG
Stuttgart, Germany

Figure 1: We present AffectRoute, a new navigation system able to route after positive emotions. We predict emotional weights for every road coordinate based on environmental, personal, and dynamic road context and find the optimal driving trajectory.

ABSTRACT

Routes represent an integral part of triggering emotions in drivers. Many navigation systems allow users to choose a navigation strategy, such as the fastest or shortest route, but they do not take the emotional well-being of the driver into account. We present AffectRoute, a novel navigation-based empathic car interface that guides drivers through real-world traffic while evoking positive emotions. We propose a set of design considerations, derive a technical architecture, and implement an optimization framework. The central part of our contribution is a custom emotion map layer generated by a machine learning classifier that predicts emotions along a route based on static and dynamic contextual data. Based on this, we developed a real-time mobile navigation app to interactively predict routes that evoke feelings of happiness. We evaluated this system in a real-world driving study ($N = 13$) and found that happy routes increase subjectively perceived valence by 11% ($p = .007$). Finally, we show how emotion-based routing can be integrated into common navigation apps to promote drivers' emotional well-being.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
UIST '22, October 10–14, 2022, Bend, Oregon, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

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

KEYWORDS

empathic interfaces, affective computing, navigation, machine learning, contextual computing

HappyRouting: Learning Emotion-Aware Route Trajectories for Scalable In-The-Wild Navigation [P7] This paper proposes a novel emotion-aware application that uses previously discussed remote sensing inputs [P1, P5] to propose route recommendations. Many navigation systems allow users to choose a navigation strategy, such as the fastest or shortest route, but they do not take the emotional well-being of the driver into account. We present HappyRouting, a novel navigation-based empathic car interface that guides drivers through real-world traffic while evoking positive emotions.

While the vision, preferences, and design of empathic navigation have been presented in prior work [20], its technical concept, implementation, and concrete evaluation have yet to be the subject of research. We propose a set of design considerations, derive a technical architecture, and implement an optimization framework. The central part of our contribution is a custom emotion map layer generated by a machine learning classifier that predicts emotions along a route based on static and dynamic contextual data. Based on this, we developed a real-time mobile navigation app to interactively predict routes that evoke feelings of happiness. We evaluated this system in a real-world driving study ($N = 13$) and found that happy routes. increase subjectively perceived valence by 11% ($p = .007$). Our results show a significant effect in perceived valence between the fast and happy routes, showing that the happy route selected leads to an improvement in valence. Furthermore, our participants were willing to use our system, although positive routes consumed more time. Moreover, we conducted a simulation study in a whole region to compare the differences between the optimization objectives. Finally, we show how emotion-based routing can be integrated into common navigation apps to promote drivers' emotional well-being.

Author contribution: I was the leading author of this publication. Together with Tobias Grosse-Puppendahl, I came up with the idea of affective navigation through context. I developed the navigation prediction and did the quantitative and qualitative study. Daniel Bulanda designed the figure on the first page and together with Adam Kozlowsky set up the navigation and mobile app infrastructure. Thomas Kosch revised the paper. Tobias Grosse-Puppendahl, and Albrecht Schmidt revised the paper for conciseness and readability.

4 Results

*"I visualise a time when we will be to robots what dogs are to humans,
and I'm rooting for the machines."*

Claude Shannon

In this section, we will briefly summarize the findings of the individual papers contributing to answer the guiding research questions.

RQ1: How to design machine learning systems for human state detection based on multimodal context sensing?

In our papers [P1, P5] we show that a consumer smartphone paired with machine learning modeling can predict emotions in the wild. Our work shows that contextual data is a reliable classification input for emotions [P1], where adding environmental data streams (i.e., the outside and inside view) can improve the overall emotion classification performance. Our results show that emotions for unknown drivers can be classified with up to 59% when using contextual and audio-visual features, an improvement of 7% over emotion detection using facial expressions. We find that emotion prediction is better when fine-tuned on participants' labeled data. However, acquiring labeled data in the wild is costly, hence we propose a few-shot learning approach to fine-tune the model by using the first minutes of input data of the participants' data.

Furthermore, we show a technical design space for designing the appropriate input feature set for contextual emotion modeling [P5]. We analyzed an extensive set adding auditory context and driver-view features analyzed by computer vision approaches (visual object detection and visual scene segmentation). By analyzing the audio-visual complexity of the outer-car ongoings, driver emotions can be predicted with 59% accuracy in a leave-one-participant-out cross-validation. In contrast, only-outside view information using the smartphone's camera stream on the road offers a recognition accuracy of 54% while providing a less driver-privacy intrusive sensing system. Our smartphone-based sensor fusion implementation is uncomplicated to integrate into other ubiquitous sensor streams with GPS or camera functionality. The robust performance provides the designer of affective in-car systems with new possibilities that do not involve cameras directed at the driver, which might raise a feeling of surveillance. Instead, our approach may only require an image representing what the driver sees. Furthermore, current driver assistance systems already obtain fine-grained outside-view information from sensors attached to the vehicle. Therefore, outside-view features, e.g., potentially already provided by an autonomous driving sensor, could directly serve as input for a potential in-cabin emotion classifying system based on visual features.

Our work advances future machine learning systems for human emotional states in different axes. The capabilities and variety of sensors in our smartphones will increase in the future, and head-worn devices such as augmented reality glasses are already in development for large-scale use. This poses a challenge for future remote sensing systems, as small ubiquitous devices can infer context from little sensory information to predict emotional states.

RQ2: How can machine learning systems be designed to provide meaningful explanations of their classification of human states?

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 particular classification result. Hereby, explainability is visualized by saliency maps that are created with a gradient-based and a forward-score-based method. Explaining the model's classification decision by inferring the importance of certain feature aspects is crucial to help humans understand the model's reasoning process. Our method allows us to understand better the relationship between environmental, emotional triggers, and emotional states. The linking of context and emotional triggers was first stressed by [6] and [39], which outlined the importance of knowing "when, where, and why" emotional triggers affect the emotional state to improve emotion recognition accuracies. The time-feature-dependent understanding is favorable for the emotion recognition developer in knowing why a specific decision has been made and offers the driver a transparent way of knowing why a machine learning decision based on his emotional state was made. In effect, our method allows to understand better the relationship between environmental, emotional triggers, and emotional states. 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. The interaction between humans and machine learning systems is crucial, especially when developing empathic car interfaces for in-the-wild use [32].

RQ3: How can the required information for machine learning systems for human state classification be enabled to increase privacy?

Emotion-related information is highly personal therefore, this sensitive data must be handled appropriately. We present three deep learning architecture methods (EEG2Vec [P2], aDAPE [P3], and ACSE [P4]) to learn domain-invariant representations in order to increase domain privacy. We evaluate all approaches on EEG datasets with human emotional state labels. Unlike previous work that has focused on learning scenarios across subjects or sessions, we explore data source invariant representations via an adversarial learning framework that can be used in EEG multi-label settings. Our approach aims at expressing robust emotion-relevant EEG features in a latent representation for emotion recognition across several datasets by limiting the representation not to learn nuisances specific to these datasets, hence being dataset invariant.

Furthermore, we find that aDAPE can learn from multiple EEG data sources and extract meaningful latent representations. We reveal large dataset domain-specific variances in conventionally trained centralized pipelines. We demonstrate that regularizing latent representations via an MMD-based domain alignment loss enables data-source-independent representation learning.

We propose EEG2Vec as an algorithm to learn latent representations of affective EEG data that allow for general use in various generative and discriminative machine learning paradigms. Our model learns vectorized representations (i.e., embeddings) of EEG responses to emotional stimuli that are discriminative of the affective states, as well as sufficiently representative to generate synthetic EEG data. In doing so, learned embeddings can also be used to generate synthetic EEG data that is both participant- and emotion-specific, simply by sampling from the latent state probability function. One important limitation of our approach lies in the accessible training dataset infrastructure. It is naturally likely that the amount of participant-specific data can impact optimization if not accounted for. So far, we have only considered learning from a balanced training data set in terms of participant IDs and class labels by stratifying our available training set size.

In general, the proposed latent representation learning pipelines with sufficient data allow future research to exploit low-rank emotion-input representations with less memory demand for general-purpose edge applications (e.g., wearable computing or human-robot interaction).

RQ4: How can machine learning systems be applied to robustly predict human emotions in dynamic and constrained environments?

Defining a robust emotion detection sensor in dynamic and unconstrained environments is presented in [P1, P5]. These unobtrusive context-to-emotion machine learning systems can be understood as a building block to predict human emotions in the wild. In [P7], we applied a context-input-only emotion prediction in the wild and evaluated its usefulness in a navigation application.

We use personal, environmental, and road-specific information to define a custom emotion routing graph that optimizes routes for predicted happy emotions. Our system predicts emotions for unknown roads before optimizing the route choice. Overall, our model is able to achieve a mean emotion recognition accuracy of 63% with a balanced F_1 score of 53.4%. These results are slightly inferior to current subject-independent contextual emotion classifiers but are also based on a remotely acquirable and thus much-reduced feature set. As a baseline in our dataset, we recorded a driver-facing camera stream and applied a facial-recognition classifier, showing that the collected contextual features still outperform facial expressions. We then use the robust emotion machine learning system to predict emotions on unknown roads to find the happy route (HappyRouting). A real-world user study shows that our happy routing system elicits positive emotions by navigating after the predicted emotional values. Consequently, HappyRouting requires more driving time which was accepted by our participants as long as the circumstances allowed it (e.g., no time pressure). Our work is not only relevant to driving but can also be applied to other areas of mobility and autonomous driving. We are confident that the presented process of simulating emotions and evaluating different paths through many potential user journeys can be generalized to an even wider variety of use cases (e.g., bicycle riding).

RQ5: How can emotion-aware and emotion-sensitive systems be designed to improve a user's experience?

Using our machine learning models based on contextual input, we designed an emotion-aware application to improve a user's experience [P7]. We design HappyRouting, a new type of empathic interface capable of navigating by positive emotions. We use personal, environmental, and road-specific information to define a custom emotion routing graph that optimizes routes for happy emotions. A real-world user study shows that HappyRouting elicits positive emotions through navigation. As a consequence, HappyRouting requires more driving time which was accepted by our participants as long as the circumstances allowed it (e.g., no time pressure). Our work is not only relevant to driving but can also be applied to other areas of mobility, such as biking navigation and autonomous driving. We find that the capability of simulating emotions and evaluating different paths through many potential user journeys can be generalized to an even wider variety of HCI areas. Our results suggest a tradeoff between the duration of the fastest route and the perceived valence of driving the happy route. Although the happy route takes more time than the fastest route, our participants would use the HappyRouting for their navigation to improve their emotional well-being.

Undeniably, the regulation of emotions by technological systems is highly controversial, as psychological effects are largely unknown. Avoidance of negative situations, for example, is an essential strategy of human emotion (self-)regulation [17], but also an implicit result of our system's promotion of positive emotions. Studies with individuals have shown that situation avoidance results in decreased learning and adaptation abilities, as well as social and anxiety disorders [1]. Therefore, we emphasize that such short- and long-term effects must be investigated in future work.

5 Conclusion

> *"We know the past but cannot control it. We control the future but cannot know it."*
>
> *Claude Shannon*

This chapter reflects on the developed research works and provides an outlook for future work.

5.1 Reflection

To conclude this thesis, we reflect upon machine learning models for human emotional states and their corresponding application areas. The presented thesis provides a variety of frameworks to tackle problems of e.g., robustness, privacy, data and domain variability, which require in-depth discussion and reflection in themselves. This critical discussion will span the included publications and provide directions towards the advancement machine learning systems for human emotional states. Limitations and reflections on particular items are discussed in the individual publications.

Everyday Available Emotional AI. Our work shows that we are able to learn powerful emotion representations from diverse input data. A regular smartphone for example can already serve as a powerful data input device and prediction medium for emotional machine learning models. In the future, novel sensor inputs e.g., from wearables, AR glasses and other disruptive, interactive technologies will find their way into our live, where it will be easy to provide the user with emotion-aware interfaces. This way the development of machine learning models, input modalities and adaptation interventions become easier and less expensive. On top, users may become used to machine learning algorithms and interfaces detecting and reacting to their human states in everyday scenarios. In the long term, we expect to see companion-like interfaces, similar to speech assistants like Amazon Alexa or Apple Siri, that interact emotionally with the user and potentially replace some human-to-human interactions.

Long-Term Effects of Machine Learning Systems for Emotions. Emotion-aware systems with the ability to influence the people might pose a risk for societies in the long run. By providing the user with only positive suggestions e.g., proposing only the happy route reinforces the echo chamber phenomenon. Humans routinely detect other human emotions and manipulate them, e.g., by playing happy music to cheer up a sad friend. Computers might be able to do the same emotion detection and manipulation. That said, we agree that there can be unethical uses of emotion detection and manipulation — both by people and by people employing affective computers [22]. Consequently, society should restrain emotion-aware interfaces that are mal-intended in their goal, e.g., by misleading humans, profiting from bad emotional states, or violating privacy norms. Deployed emotional interfaces running over long time horizons might also reinforce an echo chamber phenomenon that amplifies specific emotional states. For example, our HappyRouting application provides the user with only positive suggestions by proposing only the happy route [5].

Intrusive Data Usage. We emphasize an ethical as well as transparent use of emotion prediction models and applications. We stress that emotions are intimate, personal, and vulnerable, where potential emotional insights can be manipulated to impact behavior in the long run. Until now, many resources went into in-vehicle sensing which has resulted in much debate about the need for limiting facial recognition technology due to privacy and ethical considerations [3, 51]. This has implications on several fronts: (1) We have been collecting data for the last 15 years, yet a potential exploit of this data might enable to backwards-infer human's perceived feelings given the features presented are available in the data. (2) Environmental contextual data offers a potentially more privacy-preserving and discomfort-reducing alternative to measure emotions in the wild. However, many other data variables can infer emotions without the need for recording affective or physiological variables. Our current work broadens the debate as to what type of data should be accessible by whom and for what purposes.

Participant-Dependent Learning Systems. Furthermore, a powerful method could be the use of learning global emotion prediction models that are exported to the users device and then fine-tuned to the specific user. This would not require the user to send emotion-relevant data to a centralized database for learning a global model and reduce the exchange of personal data.

Autonomous Cars. We see application areas of the proposed machine learning systems in automated vehicles. Given a more mature autonomous driving unit (SAE Level 3 or higher [35]), the driver will likely spend less time on first-level driving tasks. This way emotion-aware applications such as emotion-aware routing may become more prevalent as this contributes to the improvement of driving safety and well-being. Furthermore, the amount of available sensor information in the car itself will likely increase due to mandatory driver monitoring. This development will help machine learning systems to better predict emotional states used for various driver monitoring and empathic recommendation functions.

5.2 Future Work

This thesis applies principles of machine learning systems to the domain of affective computing. We show novel input and machine learning frameworks for estimating and modeling human emotional states. Furthermore, we present a novel scalable machine learning method emotion-ware and emotion-sensitive system that is able to improve the user's experience in a driving domain. In this conclusive part, we propose research directions to advance the field in continuation of our work.

The contribution of this dissertation represents a step towards better understanding on how to model and design of machine learning systems for human emotional states. However, there remain unanswered questions and unresolved challenges. Arguably, the rapid development of the larger machine learning space has opened up more questions than this dissertation managed to address during the same time frame. The studies presented in this dissertation naturally face limitations, which are laid out in detail in the individual publications. Overall, the presented insights resulted from studies conducted mainly in Europe. Studies in other geographical and political contexts may bring forward differences with regard to users behavior, motivation, or perceived utility. While we are confident that the presented results for the robust emotion detection modeling [P1, P5] and emotional navigation [P7] are robustly evaluated and employed in a to in-the-wild use, further research is needed to confirm the systems. Rooted in the presented findings, we therefore discuss how future Human-Computer Interaction research may overcome these limitations and address new questions that have emerged from the evolving machine learning systems for emotions landscape.

5.2.1 Immediate Future Work

Research projects that can be **immediately** picked up concern the application of the tools provided by our work for more general validation purposes, but also for applying them to new application scenarios.

Self-Supervised Emotion Representation Learning. Practically speaking, it is impossible to label everything in the world. Producing a dataset with clean emotion labels is expensive, but computing devices generate unlabeled data all the time. For example, think about the number of cars driving with a camera or, in the future, augmented reality applications sensing our everyday surroundings. One way to use this large amount of unlabeled data is to use self-supervised learning paradigms in which the model obtains supervisory signals from the data itself. The model leverages the underlying structure in the data and can have a more nuanced understanding of reality beyond the specified training data.

Hence, interesting future work is to learn a general encoder network using a proxy task by masking input and predicting the masked input using the other sensor data. For example, the network learns to predict the current weather from the other context input such as acceleration and camera views. In the second step, the learned encoder weights are fixed, and only a few emotional labels are used to train a prediction layer from the encoder output via supervised learning. This way, the emotional classification becomes more robust as the intermediate representation has good semantic and structural meanings. Furthermore, the representations can be used for various practical downstream tasks, e.g., activity or fatigue detection.

5.2.2 Going Beyond The Lab

Contextual Physiological Input Systems. Future work could explore the gap between EEG and contextual representations. By acquiring EEG and contextual sensor data simultaneously, the individual sensor inputs can be encoded. In the next step, the mismatch in EEG and contextual input representation can be assessed and minimized using the presented domain alignment learning framework [P3]. This alignment would result in a general affective state representation format and enable future emotion-aware applications that can operate on the same representation space. By visualizing and analyzing the shared latent space, the research may move closer toward understanding the psycho-physiological ongoings of human (emotional) states.

Emotional User Prediction in the Loop. While we propose machine learning systems that are able to predict emotions in the wild, further evaluation of the machine learning system output is needed. A promising outlook for the affective navigation framework [P7] would be to show the user the real-time emotion prediction for possible driving turns and assess the user's emotional response when taking a specific turn. This continuous stream of emotional interactions could employ feedback loops of detected user reactions and system adaptations over time to optimize any system for maximum joy or other desired experiences. Thereby machine learning frameworks in the field of online learning could be employed.

5.2.3 Long Term Horizon

As for the long-term time horizon (10 years), we see the main challenge in the wide-scale adoption of the technologies and vision proposed in this thesis.

Ethical Machine Learning Systems. Machine learning systems for emotions help us to understand users. Therefore, this technology can not only be used to improve the product but also to assert influence on users. As machine learning models often learn better predictions from more data, we hope to see regulatory decisions to protect users from greedy technologies. Failing to limit the exploitation of this new technology could be a significant error able to change the future of societies. In the long term, we expect society and technology to define patterns of ethical guidelines and data usage for large-scale adaptation of machine learning systems for emotions. One potential challenge is to know when precisely a human-state prediction is made by a machine learning system. Most machine learning systems solve a proxy optimization problem, e.g., predicting the time spent on a user interface by which the machine learning system also captures and likely deducts elements of human states. The disentanglement of human-state machine learning systems and those that do not contain any human-state information becomes more blurry as non-user data, such as context, can be used for human-state prediction. In our papers, we show in our papers [P1, P5] that we can predict human states from contextual data only with reasonable accuracy, demonstrating a challenge from the regulatory and ethical perspective.

Emotion Conservation. In today's world, we can immerse ourselves in videos to the time back when the video is recorded. If this video contains vivid childhood memories, the conservation of feelings can be re-evoked and "relived" again. However, videos only portray a subset of past experiences by capturing visual and auditory stimuli. As emotions are naturally multi-modal, capturing a broader dataset of past events is necessary to recreate the immersion and re-evocation of past feelings. In the future, one interesting direction is to research immediate emotion re-evocation - which we call "emotion conservation". The idea is that past emotional events are recorded, e.g., via a more contextual approach of a surround-view input of context, video avatars, smell, and/or more introspective signals coming from the brain (e.g., EEG) observing the scenario. The sensor input is then encoded and saved as a latent representation. The system loads the specific latent representation whenever the user wants to relive the emotional feeling of a past experience. Then, the system recreates the stimuli as close to the past experience / latent representation as possible. We envision a control-based system that can strengthen individual stimuli components of the re-stimuli input to the user based on the current physiology and context of the human. For example, the system would amplify the voice of the past experience to the user if the effect of emotional reliving is insufficient. One way to assess the mismatch in the reliving experience is to capture the human input when reliving and encoding it in the same latent representation format. Thereby, the two past and present latent representations share a common hyper-dimensional space, and the mismatch of experiences resolves to the distance of the two representations. Similar to this idea is the stream of research called life-logging. Life-logging preliminary proposed by Sellen et al. [33] is a technology in the form of a wearable camera, which aims to capture data about everyday life to enable people's memory for past, personal events. Reliving past experiences is being experimented with Alzheimer's patients to facilitate people's ability to connect to their past.

Our paper [P2, P3, P4] offers preliminary work on latent representations for emotional states. In the case of the EEG2Vec architecture [P2], our latent representation can re-generate input data from the latent representation. This way, the work of this thesis can be seen as a first step towards understanding the emotional stimuli, representing and providing emotion-aware (re-) evocation strategies. However, many technological advances to create such a system are currently missing. One key aspect is capturing a sufficient set o multi-modal components of an experience. One way closer to this reality is the current emergence of 3D human body avatars [27] instead of 2D videos, which allow capturing a much finer-grained context of human experiences. Furthermore, more research in testing computer-controlled emotional stimuli to humans has to be done to re-create a complete immersion of a past experience.

Everyday Emotional AI. While we provide the first investigations in applying machine learning systems for human-state prediction to car navigation, a fully developed system that comprehensively supports users with their information tasks throughout the day could boost an entire society's productivity and safety level.

Going beyond the ability to detect human emotional states, future research could address how technologies can stage interventions and thus induce favorable human (emotional) states. Similarly, people drink a cup of coffee when they begin to feel tired, future machine learning systems for emotional states could detect other human states, such as fatigue, and proactively help users change their current state. This intervention is similar to recommended meditation exercises to increase general focus or wind down after a long workday. In addition, some systems have been proposed that support users' creativity process. However, systems could go beyond single use cases and schedule a variety of interventions throughout the day to help users get to and remain in highly productive states. We envision an everyday emotional companion powered by machine learning that detects and proposes recommendations. Such potentially emotion-invasive applications are controversial, and we recommend empowering a research pillar on users' long-term effects, including behavior changes and potential health impact. Looking at other human states, research has shown that chronically increased levels of mental could cause various health problems, such as stress and depression. On the other side, such a system can be beneficial in helping people cope better with everyday life by balancing, e.g., extreme emotional burdens.

5.3 Concluding Remarks

In the last few decades, we have seen a vast increase in computing power and the capability of machines to capture real-world complexity. Especially the field of machine learning systems has evolved rapidly over the last five years. New machine learning models are published daily that can handle complex problem spaces. Recently, Chat-GPT [1], a machine learning model trained by OpenAI, was released, offering a human-like conversation with a machine conversationally with realistic-sounding answers. Chat-GPT provides human-like, detailed answers to inquiries, e.g., the chatbot can write an

[1] `https://openai.com/blog/chatgpt/`, accessed 25.01.2023

article on any topic efficiently within seconds, potentially eliminating the need for a human writer or generating a program code on its own. Thus the model can capture the complexity of human language like never before, pushing the field closer to ‘Artificial General Intelligence’ (AGI) and advancing the field of affective computing. The advances in machine learning show that machine learning systems are moving fast, and previously proposed human state recognition architectures may become obsolete.

I hope that the work presented in this dissertation contributes its humble part to the field of research on machine learning systems for human states. Furthermore, the work can serve as a foundation for future research and practices. The current results show valuable tools for machine learning systems for emotional states, thereby providing a better understanding of human behavior. In the future, iteratively testing and improving machine learning models and their user-interface recommendations to accurately detect and react to human states will be highly important.

LIST OF TABLES AND FIGURES

1 Author’s personal contribution towards the publications. x

2.1 Dimensional emotion representation model. 8

3.1 Publications organized by research question, methods, and contribution type. 14

References

[1] Amelia Aldao, Gal Sheppes, and James J Gross. *Emotion regulation flexibility*. In: *Cognitive Therapy and Research* 39.3 (2015), pp. 263–278. DOI: 10.1007/s10608-014-9662-4.

[2] Tanja Bänziger, Didier Grandjean, and Klaus R Scherer. *Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT).* In: *Emotion* 9.5 (2009), p. 691.

[3] Cristina Bustos, Neska Elhaouij, Albert Sole-Ribalta, Javier Borge-Holthoefer, Àgata Lapedriza, and Rosalind Picard. *Predicting Driver Self-Reported Stress by Analyzing the Road Scene*. In: 2021, pp. 1–8. DOI: 10.1109/ACII52823.2021.9597438.

[4] Alan S Cowen and Dacher Keltner. *Self-report captures 27 distinct categories of emotion bridged by continuous gradients*. In: *Proceedings of the national academy of sciences* 114.38 (2017), E7900–E7909.

[5] Roddy Cowie. *Ethical issues in affective computing*. In: *The Oxford handbook of affective computing* (2015), pp. 334–348.

[6] Monique Dittrich. *Why Drivers Feel the Way they Do: An On-the-Road Study Using Self-Reports and Geo-Tagging*. In: *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 2021, pp. 116–125.

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

[8] Paul Ekman. *Facial expression and emotion.* In: *American psychologist* 48.4 (1993), p. 384. DOI: 10.1037/0003-066X.48.4.384.

[9] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. *Universals and cultural differences in the judgments of facial expressions of emotion.* In: *Journal of personality and social psychology* 53.4 (1987), p. 712. DOI: 10.1037/0022-3514.53.4.712.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. *A kernel two-sample test*. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[11] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. *Emotion representation, analysis and synthesis in continuous space: A survey*. In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE. 2011, pp. 827–834.

[12] Robert Jenke, Angelika Peer, and Martin Buss. *Feature extraction and selection for emotion recognition from EEG*. In: *IEEE Transactions on Affective computing* 5.3 (2014), pp. 327–339.

[13] Stamos Katsigiannis and Naeem Ramzan. *DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices*. In: *IEEE journal of biomedical and health informatics* 22.1 (2017), pp. 98–107.

[14] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. *Deap: A database for emotion analysis; using physiological signals*. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[15] Raymond Shih Ray Ku. *The creative destruction of copyright: Napster and the new economics of digital technology*. In: *The University of Chicago Law Review* (2002), pp. 263–324.

[16] Shan Li and Weihong Deng. *Deep facial expression recognition: A survey*. In: *IEEE Transactions on Affective Computing* (2020).

[17] Kateri McRae and James J Gross. *Emotion regulation.* In: *Emotion* 20.1 (2020), p. 1. DOI: 10.1037/emo0000703. URL: http://dx.doi.org/10.1037/emo0000703.

[18] Fariz Redzuan bin Monir, Rusyaizila Ramli, and Nabilah Rozzani. *Driving Alert System Based on Facial Expression Recognition*. In: *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS)*. IEEE. 2021, pp. 104–109. DOI: 10.1109/I2CACIS52118.2021.9495910.

[19] Antti Oulasvirta and Kasper Hornbæk. *HCI research as problem-solving*. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 4956–4967.

[20] Bastian Pfleging, Stefan Schneegass, Alexander Meschtscherjakov, and Manfred Tscheligi. *Experience Maps: Experience-Enhanced Routes for Car Navigation*. In: AutomotiveUI '14. Association for Computing Machinery, 2014, pp. 1–6. DOI: 10.1145/2667239.2667275. URL: https://doi.org/10.1145/2667239.2667275.

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

[22] Rosalind W Picard. *Affective computing: challenges*. In: *International Journal of Human-Computer Studies* 59.1-2 (2003), pp. 55–64.

[23] Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.

[24] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. *Toward machine emotional intelligence: Analysis of affective physiological state*. In: *IEEE transactions on pattern analysis and machine intelligence* 23.10 (2001), pp. 1175–1191.

[25] David J Rachlin. *Encyclopedia of Behavioral Medicine*. In: *Reference Reviews* (2013).

[26] Nancy A Remington, Leandre R Fabrigar, and Penny S Visser. *Reexamining the circumplex model of affect.* In: *Journal of personality and social psychology* 79.2 (2000), p. 286.

[27] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. *Audio-and gaze-driven facial animation of codec avatars*. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 41–50.

[28] James A Russell, Maria Lewicka, and Toomas Niit. *A cross-cultural study of a circumplex model of affect.* In: *Journal of personality and social psychology* 57.5 (1989), p. 848.

[29] James A. Russell. *A Circumplex Model of Affect.* In: *Journal of Personality and Social Psychology*. Vol. 39. 6. American Psychological Association (APA), 1980, pp. 1161–1178. DOI: 10.1037/h0077714.

[30] Daniela Schiller, NC Alessandra, Nelly Alia-Klein, Susanne Becker, Howard C Cromwell, Florin Dolcos, Paul J Eslinger, Paul Frewen, Andrew H Kemp, Edward F Pace-Schott, et al. *The human affectome*. In: (2022).

[31] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. *Introducing wesad, a multimodal dataset for wearable stress and affect detection*. In: *Proceedings of the 20th ACM international conference on multimodal interaction*. 2018, pp. 400–408.

[32] Bridie Scott-Parker. *Emotions, behaviour, and the adolescent driver: A literature review*. In: *Transportation research part F: traffic psychology and behaviour* 50 (2017), pp. 1–37.

[33] Abigail J Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. *Do life-logging technologies support memory for the past? An experimental study using SenseCam*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, pp. 81–90.

[34] Luke Stark. *Facial recognition is the plutonium of AI*. In: *XRDS: Crossroads, The ACM Magazine for Students* 25.3 (2019), pp. 50–55. DOI: 10.1145/3313129.

[35] Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. Standard SAE J3016_201806. SAE International, 2018. URL: https://www.sae.org/standards/content/j3016_201806.

[36] Harald G Wallbott and Klaus R Scherer. *Assessing emotion by questionnaire*. In: *The measurement of emotions*. Elsevier, 1989, pp. 55–82.

[37] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R Schinazi, Markus Gross, and Christian Holz. *Affective State Prediction from Smartphone Touch and Sensor Data in the Wild*. In: *CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–14.

[38] Feichi Wang, Mengchai Mao, Lian Duan, Yuxia Huang, Zheng Li, and Chaozhe Zhu. *Intersession instability in fNIRS-based emotion recognition*. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.7 (2018), pp. 1324–1333.

[39] Sebastian Zepf, Monique Dittrich, Javier Hernandez, and Alexander Schmitt. *Towards empathetic Car interfaces: Emotional triggers while driving*. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–6. DOI: 10.1145/3290607.3312883.

[40] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W. Picard. *Driver Emotion Recognition for Intelligent Vehicles: A Survey*. In: *ACM Comput. Surv.* 53.3 (2020). DOI: 10.1145/3388790. URL: https://doi.org/10.1145/3388790.

[41] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. *Emotion recognition using multimodal data and machine learning techniques: A tutorial and review*. In: *Information Fusion* 59 (2020), pp. 103–126.

[42] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki. *EmotionMeter: A Multimodal Framework for Recognizing Human Emotions*. In: *IEEE Transactions on Cybernetics* (2018), pp. 1–13. DOI: 10.1109/TCYB.2018.2797176.

[43] Wei-Long Zheng and Bao-Liang Lu. *Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks*. In: *IEEE Transactions on Autonomous Mental Development* 7.3 (2015), pp. 162–175. DOI: `10.1109/TAMD.2015.2431497`.

[44] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. *Identifying stable patterns over time for emotion recognition from EEG*. In: *IEEE Transactions on Affective Computing* 10.3 (2017), pp. 417–429.

Original Contributing Publications

This appendix contains the publications contributing to this thesis in their original format.

[P1] VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time 41

[P5] Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context 55

[P7] HappyRouting: Learning Emotion-Aware Route Trajectories for Scalable In-The-Wild Navigation 85

[P2] EEG2Vec: Learning Affective EEG Representations via Variational Autoencoders . . . 107

[P3] Domain-Invariant Representation Learning from EEG with Private Encoders 115

[P4] Exploiting Multiple EEG Data Domains with Adversarial Learning 121

[P6] Interpretable Time-Dependent Convolutional Emotion Recognition with Contextual Data Streams 127

VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time

David Bethge
Porsche AG, LMU Munich
Stuttgart, Germany

Thomas Kosch
TU Darmstadt
Darmstadt, Germany

Tobias Grosse-Puppendahl
Porsche AG
Stuttgart, Germany

Lewis L. Chuang
TU Dortmund
Dortmund, Germany

Mohamed Kari
Porsche AG
Stuttgart, Germany

Alexander Jagaciak
Porsche AG
Stuttgart, Germany

Albrecht Schmidt
LMU Munich
Munich, Germany

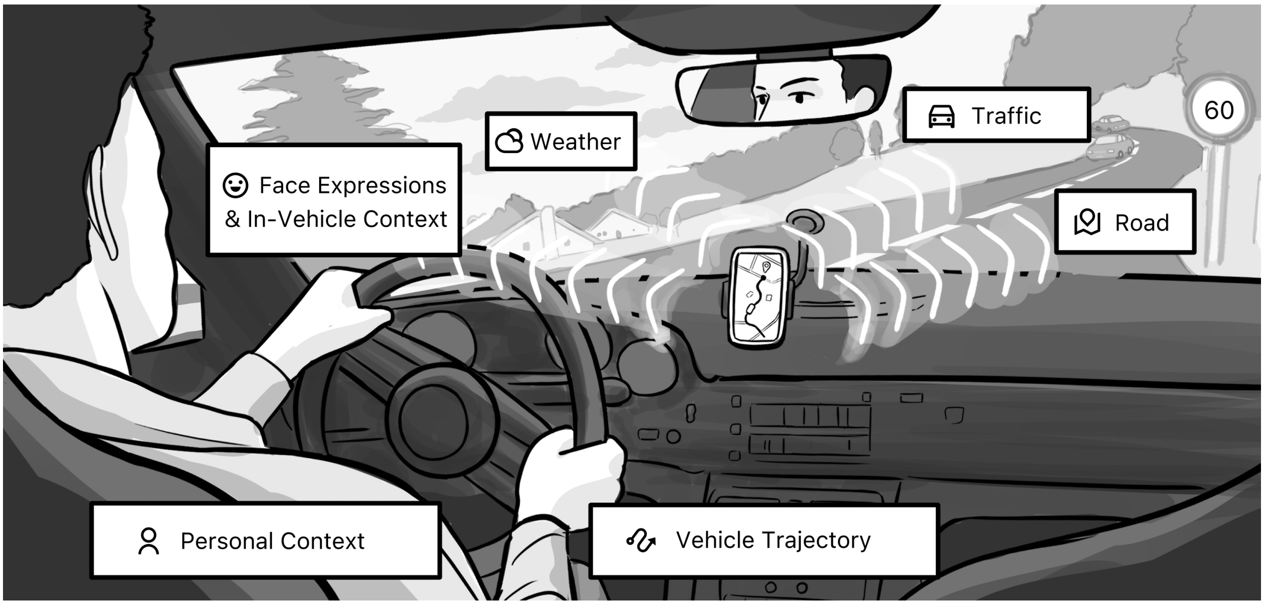


Figure 1: We present VEmotion, a new virtual emotion sensor embedded into a smartphone app that fuses an extensive variety of contextual information like vehicle- and traffic dynamics, road characterization, environmental weather, and in-vehicle context.

ABSTRACT

Detecting emotions while driving remains a challenge in Human-Computer Interaction. Current methods to estimate the driver's experienced emotions use physiological sensing (*e.g.,* skin-conductance, electroencephalography), speech, or facial expressions. However, drivers need to use wearable devices, perform explicit voice interaction, or require robust facial expressiveness. We present VEmotion (Virtual Emotion Sensor), a novel method to predict driver emotions in an unobtrusive way using contextual smartphone data. VEmotion analyzes information including traffic dynamics, environmental factors, in-vehicle context, and road characteristics to implicitly classify driver emotions. We demonstrate the applicability in a real-world driving study ($N = 12$) to evaluate

UIST '21, October 10–14, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8635-7/21/10…$15.00
https://doi.org/10.1145/3472749.3474775

the emotion prediction performance. Our results show that VEmotion outperforms facial expressions by 29% in a person-dependent classification and by 8.5% in a person-independent classification. We discuss how VEmotion enables empathic car interfaces to sense the driver's emotions and will provide *in-situ* interface adaptations on-the-go.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

driver emotion detection, mobile sensory system, contextual affective state prediction, machine learning

ACM Reference Format:
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 (UIST '21), October 10–14, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3472749.3474775

1 INTRODUCTION

Driving can elicit many emotional and cognitive states. The experience of driving — a combination of how one feels before entering the vehicle, the context of neighboring traffic, the behavior of other road users, the car aesthetics, and one's own driving style, among other factors — induces a wide range of emotions in drivers [48]. There is a growing interest in developing automotive user interfaces that allow for implicit and explicit interactions that are aware of how the driver is feeling [4]. This rests on the viability of the system in accurately estimating the driver's emotions, a field referred to as affective or empathic computing [7, 39].

Recent breakthroughs in ambient ubiquitous sensing [33] allow in-the-wild driver data, including real-world driving context, to inform emotion classification models. In principle, this could allow for empathic car interfaces [4] that could plan routes to invoke specific emotions, raise the user's engagement when detecting boredom by playing the user's preferred music, or mitigate undesirable driving styles that result from negative emotions (e.g., anger, sadness). The viability of such interfaces rests on the accurate, robust, and real-time classification of a driver's emotions. This remains an ongoing research challenge.

What are emotions and how do we measure them? Ekman has proposed six basic and pancultural emotions that can be inferred from one's facial features [15, 16]. This has motivated the development of computer vision for recognizing emotions from camera-captured facial expressions [2]. Besides this, implicit physiological activity could also be relied on for estimating the user's emotion. Some modalities include electroencephalography [1], electrodermal activity [8], or heart rate variability [40]. Nonetheless, physiological sensing often requires user contact with the measurement sensor, which impacts the user acceptance [25] and the overall driving experience. In comparison, remote cameras are less intrusive [31, 38]. For this reason, state-of-the-art algorithms for facial expression recognition are now commercially and widely available — such as the Affectiva SDK [34], or the Microsoft Azure face detection API[1]. These systems have been deployed on a large scale and are utilized to measure drivers' emotions and stress [9, 23]. The correlation between facial expressions and their underlying emotion can vary across individuals [45], where the emotion detection quality depends on the driver's facial expressiveness, brightness levels, and the driver's willingness to be video recorded. Here, previous research suggests that the individual driving style and driving performance are indicative of the driver's experienced emotions [22, 35]. With this in mind, we investigate whether the analysis of driving styles and driving-related information can be used to predict driver emotions? This is a counter-intuitive proposition, given that we are sensing driving information instead of sensing the driver themselves.

This paper presents VEmotion, a smartphone system that uses internal sensors only to measure driving information and estimate the perceived emotions in real-time. VEmotion analyzes the user's driving behavior through the car's surroundings variables including speed, weather, road types, and traffic flow. In contrast to previous emotion assessment modalities, VEmotion relies only on the contextual data from the vehicle that does not require modifying the car itself. To elaborate, we recorded high-dimensional contextual driving data on different routes and derived common environmental influences on emotional states. We collected data with VEmotion in a user study with twelve participants and evaluated its classification accuracy. Our results show that vehicle speed, traffic flow, and weather terms are assigned the highest feature importance from all recorded context variables. We conclude that VEmotion is an appropriate and generalizable approach for predicting the driver's emotions, achieving up to 72.4% accuracy in real-world driving scenarios.

CONTRIBUTION STATEMENT

Our work makes four contributions: (1) We present VEmotion, a mobile and personal computing software that predicts driver emotions based on contextual driving data. (2) We report an *in-the-wild* study and demonstrate that emotion recognition from camera-captured facial expressions can be improved by 28.5% using VEmotion. (3) We provide a machine learning-based processing pipeline that analyzes the relative importance of the various contextual features and, hence, their respective contribution to emotion prediction accuracy. (4) Finally, we discuss how VEmotion enables seamless emotion prediction for future empathic car interfaces. Altogether, this paper demonstrates that contextual measurements can support emotion state classification, not only of the user themselves but also of contextual variables that invoke the state (e.g., weather, traffic flow) or result from the vehicle state (e.g., car speed).

2 RELATED WORK

This section presents previous work about emotion assessment, detection of emotions in driving scenarios, and the use of emotions in interactive systems.

[1] https://azure.microsoft.com/services/cognitive-services/face, last access 2021-04-07

2.1 Emotion Assessment

There is a tendency in computer science to treat affect and emotion as the same phenomenon inferring and understanding human emotion primarily through the expression of physiological signals such as facial expression, gait, or blood conductivity [52]. Although they are different, necessary distinctions occur. Affect has been described by Deborah Gould [20] as "non-conscious and unnamed, but nonetheless registered, experiences of bodily energy and intensity that arise in response to stimuli" and thereby describes a "compound phenomenon variously consisting of evaluative, physiological, phenomenological, expressive, behavioral, and mental components" [52]. Emotion is regarded as "what from the potential of [affective] bodily intensities gets actualized or concretized in the flow of living" [20]. Treating Stark and Hoye [52] as a starting point, our current work is physiological and adopts a motivational model of emotion. We address criticism against this conflation of our chosen approach in the discussion section.

Measuring the user's emotions is a compelling topic that has been addressed by previous research. Picard coined the term *Affective Computing*, envisioning computers to express or sense emotions to provide a computerized interface that mimics human-like capabilities [39]. Modern user interfaces, such as voice or speech interfaces, benefit from understanding the user's currently perceived emotions or cognitive states and can adjust their interface according to the user's mood [54]. However, investigating robust modalities that *sense* emotions in real-time is still an ongoing research field.

Early work looked at facial expressions as a marker for perceived emotions. Ekman [14] and Ekman and Rosenberg [18] concluded that a connection between emotions and facial expressions exists. Numerous frameworks exist which can recognize emotional states using facial expressions.

However, facial expressions are considered an individual property that is different across the user's culture [43] or their gender [17]. Hence, facial expressions for interactive applications require users to calibrate towards their individual facial expressions. Kosch et al. [27] investigated if the detection of facial expressions via computer vision is feasible for mobile in-the-wild studies. They find that a re-calibration of the individual facial expressions on a per-user basis increases the correctness of emotions detected through facial expressions by 33%. However, detecting facial expressions using computer vision requires installing cameras and can compromise the user's privacy. External factors, such as illumination, can influence the quality of facial expression detection. Wearable sensors that provide a direct assessment of the user's physiological states can be used to infer the perceived emotions. Other wearable sensors exploited alternative physiological sensing modalities, such as electrodermal activity, heart rate, muscle tension, breathing rate, and electroencephalography [28]. However, wearable devices must provide a sufficient utility to the user to justify the user's effort of using the wearable sensor [59]. Also, the obtained physiological signals require a certain quality level and the suitable measurement modality for the right job to provide a meaningful assessment over the emotions [12].

2.2 Detecting Emotions while Driving

Facial expressions have a long tradition as an indicator for the expressed emotions [14]. Typical facial expressions include smiling or frowning as well as head gestures, such as nods and tilts. The detection of facial expressions requires an additional camera in the driver's cabin, including RGB cameras [9, 31, 38], infrared cameras [19] or thermal cameras [26]. Physiological sensing utilizes the driver's direct bodily responses to draw conclusions about the emotional states. Several physiological sensing modalities, such as heart rate, electrodermal activity, and electroencephalography [13, 62], are indicative of the driver's perceived emotions. However, to measure such physiological signals, sensors require direct contact with the user while driving (e.g., electrodermal activity sensor attached to the driver's hand). This can impact the driving experience and usability negatively [63]. In-car speech interfaces have been investigated as a modality to measure the driver's emotions. The way the driver talks to the voice assistant or co-drivers can indicate the user's perceived emotions. A variety of studies focus on paralinguistic features and how drivers are verbally interacting with the environment [21, 44, 46] by analyzing the sound's loudness, pitch, and spectral features [62]. However, the driver needs to communicate with an entity in the car while driving to enable robust detection of emotions which is not feasible during stressful or cognitively demanding driving scenarios.

Previous research hypothesizes that the driving behavior, style, and the driver's context are indicative of the currently perceived emotions [35]. Here, behavioral characteristics are viewed as emotional markers. For example, the grip strength applied on the steering wheel varies with the driver's emotional states [30, 36, 49]. Other factors include the interaction with the gas and brake pedals [32] as well as changes in body posture using pressure sensors [53]. Similarly, the driver's context and driving behavior are reliable factors to predict emotions. Navon et al. investigated how a driver's driving style is influenced under different emotions, finding that maladaptive driving styles are closely related with participants who have difficulties in emotion regulation and forgivingness [35]. Hancock et al. [22] show that negative emotions impact driver performance and driving styles, impacting the number of lane excursions and lateral control of the car.

Based on previous work, we expect correlations between the driving style and driver emotions. However, developers and researchers must access the car's sensor layer, which is often kept confidential, to infer the user's driving style. Standards for obtaining these data streams exist (e.g., OBD II) but are limited to specific measures, such as acceleration, braking, or steering behavior [50]. Furthermore, these standards have to be implemented by the individual car manufacturers and often miss environmental factors, including road context variables. So far, previous research has informed how emotions can be sensed in-car interfaces. Sensing the driver's emotions by utilizing the driver's driving context and behavior without modifying the user's car on the go has not been studied so far. We close this gap by presenting a study that classifies the driver's emotions by solely analyzing the context and driving behavior.

2.3 Considering Emotion Expressiveness in Real Driving Environments

Detecting emotions in the wild is a challenging task. From a machine learning perspective, most recognition models are trained with data from a constrained environment (e.g., driving simulators) and perform poorly in unconstrained scenarios. To evaluate our contribution to existing work in driver emotion recognition (e.g., with other modalities), the most recent systematic literature survey by Zepf et al. [62] provides a detailed understanding. The survey systematically reviews literature back to 2002 and identifies 63 papers on this topic. Out of 63 identified articles in the survey, only 19 papers measure emotions in natural, non-simulated settings (i.e., not induced or acted). Looking at the expressed emotion categories of the 19 papers, 16 papers were measuring stress while three papers were measuring emotions. One of these papers was predicting aggressive driving behavior without taking emotional states into account [24]. Another one used electroencephalography and electrodermal activity to predict concentration, tension, tiredness, and relaxation [41]. Finally, Riener et al. [42] inferred arousal states using electrocardiography and GPS data. Contrary to related work, our approach does not require modifying the user's car and utilizes only smartphone sensors to determine the user's driving context and behavior, hence implying the user's perceived emotions. We present the system and classification pipeline in the following section.

3 VEMOTION

In this section, we present VEmotion, a system that captures the driver's contextual driving data from the smartphone alone. We present the software architecture and the measures of our implementation in the following.

3.1 System Architecture

We implemented a smartphone app that captures contextual smartphone data to train a classifier that predicts the driver's emotions. We perform a layered approach of extracting relevant context information to learn as much as possible from the driver's driving context using a minimum set of input streams. The selected features are based on Braun et al. work [5] where driving behavior, traffic, vehicle performance, and environmental factors are relevant. We filtered the variables based on the following requirements: (1) on-device computation without accessing the vehicle itself, (2) no direct user interaction, and (3) non-critical consumption of device resources. We capture the smartphones' fused sensory data and use it as an input for a machine learning predictor. Figure 2 provides an overview of the VEmotion system architecture. VEmotion utilizes the speed of the vehicle, current weather, traffic context, road context using GPS data, and the driver's facial expressions along a perceived emotion baseline to train a predictive classifier. A prototype is developed as an iOS app, in which location-based data is sensed in a $1Hz$ (Hertz) interval, whereas the video produces

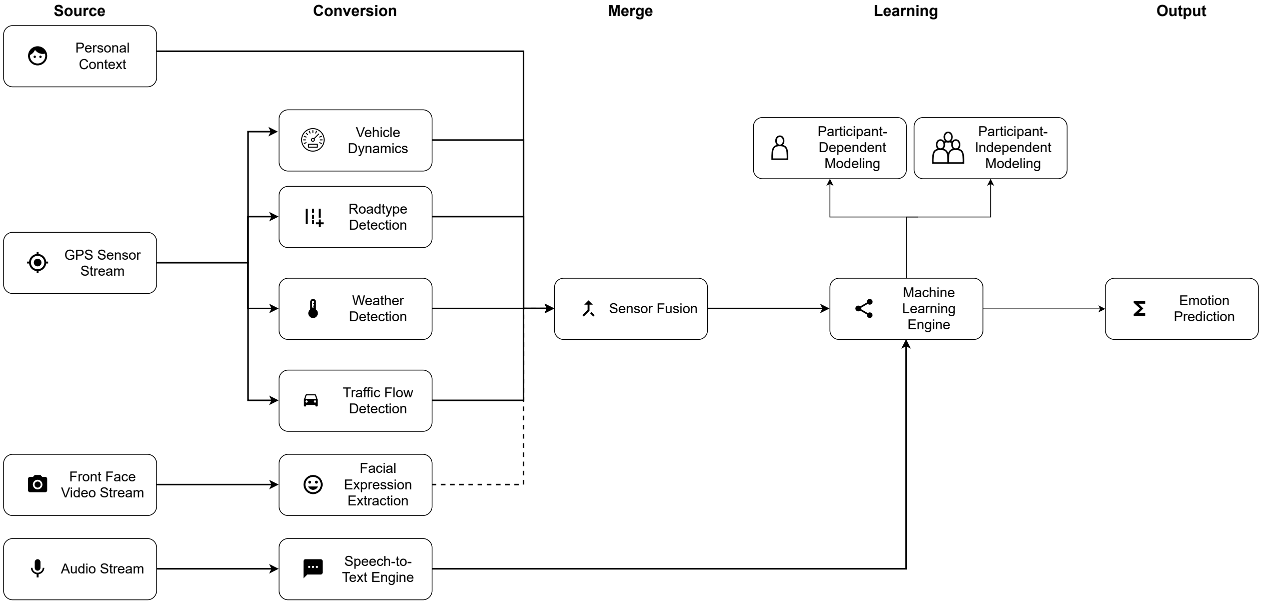


Figure 2: Overview of the VEmotion system architecture. We record contextual data (e.g., weather, road type, traffic flow) and the driver's facial expressions while driving. We fuse the collected data and use it as an input for a machine learning predictor that predicts the driver's emotions. The audio stream is used to detect the baseline emotion in our study experiment. Facial expressions can be included as a feature in VEmotion based on individual privacy policies and is therefore depicted as a dashed line. The audio stream input analyzed via a speech-text-engine is used to extract the label for our system and is not included as an input feature to the machine learning engine.

approximately 30 frames per second. In the following, we present the features and data that are recorded by VEmotion.

3.1.1 GPS Sensor: ***Vehicle Dynamics*** . We interpolate the speed of the vehicle (v) between two consecutive GPS waypoints ($(lat_1, long_1), (lat_2, long_2)$) and the time between t via the Haversine formula [58]. We also calculated the vehicle's acceleration by computing the change in velocity divided by the time between using two consecutive vehicle speed measurements.

3.1.2 GPS Sensor: ***Weather***. We request weather information of each incoming GPS coordinate from the Microsoft Azure Maps API[2] to reflect the weather context conditions in real-time. Thereby, we include the following weather conditions: weather description called '*weather_term*' (e.g., '*sunny*'), the approximated outside-temperature '*feeltemp_outside*' (in $^\circ C$), cloud coverage '*cloud_coverage*' (in %), and wind speed '*windspeed*' (in km/h).

3.1.3 GPS Sensor: ***Trafficflow***. We also include the traffic flow in VEmotion by providing information about the speeds and travel times of the road fragment closest to the given coordinates using the Microsoft Maps Traffic Flow API. Thereby, for each GPS point we include the variable '*freeflow_speed*', which is the speed of traffic expected under ideal conditions. The freeflow speed can be different from the maximum speed limit of the road, for example, in case narrow roads force driver to slow down. To account for slow-moving traffic and jams, we define a feature called '*trafficflow_reducedspeed*'. The reduced speed of the traffic flow is calculated by the freeflow speed on the road *freeflow_speed*$(lat, long)$ minus the actual traffic flow speed on this segment *current_speed*$(lat, long)$: *trafficflow_reducedspeed*$(lat, long)$ = *freeflow_speed*$(lat, long)$ − *current_speed*$(lat, long)$ measured in km/h.

3.1.4 GPS Sensor: ***Road Type***. We extract the nearest roads from OpenStreetMap[3] via reverse geocoding to detect the surrounding infrastructure for every GPS coordinate. We download a $200m \times 200m$ high-definition map of the current GPS coordinate and perform a map matching by calculating the euclidean distance of each node in the map to the current GPS coordinate and select the road node object that is the closest. We thereby extract the following features: road-type (e.g., '*highway*'), maximum speed on the current road (in km/h), and the number of available lanes on the current road.

3.1.5 Front-Facing Smartphone Camera: ***Facial Expressions***. We decided to include and evaluate the basic emotions captured through facial expressions [14] into our classification pipeline. The facial expression does not represent our label for predicting the emotions of the driver but is rather a way to have more inside-view information. The smartphone app obtains an image stream with 30 frames per second from the driver-facing camera and cuts it into frames to assess the driver's facial expressions. Up to 10 frames per second are sent via a cloud platform to be analyzed for facial expression features. Here, the Microsoft Face Recognition API is used to detect facial expressions that indicate specific emotions. The API returns confidences for eight basis emotions ('anger', 'contempt', 'disgust', 'fear', 'happiness', 'neutral', 'sadness', 'surprise'). No emotion value is recorded if no faces are detected (e.g., due to occlusion or shaky video stream). To have distinct emotions corresponding to a GPS coordinate rather than confidences of the eight basic emotions, we take the emotion with the maximum confidence and call this variable *facial_expression*. The validity of different cloud-based, commercial facial expression SDKs has been researched by Yang et al. [60] using a multitude of data sets such as ADFES [56], RaFD [29], WSEFEP [37]. The overall emotion recognition accuracy of Microsoft Azure is higher 84.7% compared to the 67% accuracy from the Affectiva SDK, especially 'angry', 'sad', and 'happy' facial expressions can be predicted more confidently with Microsoft Azure [60].

3.1.6 Per-Ride User-Input: ***Personal Context***. To include more subject-variant features in our analysis, we selected '*daytime*' of the ride, '*age*' of the driver, and felt emotions before the ride ('*before_emotion*') as variables to our system. Their values remain constant over the driving time.

3.2 Synchronizing Data Streams: Sensor Fusion

In the system's sensor fusion module, we make sure that all incoming sensor streams from GPS and camera are aligned along the time- and spatial dimensions. The GPS module exports its latitude and longitude signals together with the current timestamp of the sensor system in a GPX-XML format. The frontal face video stream is divided into individual frames and attached metadata about their time-occurrence based on the camera's frames per second. The output emotion categories are merged with the GPS sensor stream by the timestamp values after analyzing the individual frames and GPS-derived information. Table 1 shows the used features with example values.

Table 1: List of available features to predict emotions on the ride.

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'

We performed several steps to clean the data before training a suitable context-emotion classifier. The labeling process is defined in the user study section 4. We excluded all observations, where

[2] https://azure.microsoft.com/services/azure-maps, last access 2021-04-07

[3] https://nominatim.openstreetmap.org/ui/search.html, last access 2021-04-07

the 'expressed_emotion' label is outside our specified emotion categories (e.g., one participant P11 once labeled he is 'stressed'). The string-based features are encoded into integer categories ('before_emotion', 'daytime', 'weather_term', 'road_type' and 'facial_expression') for appropriate use in the classification algorithm. Next, we cleaned the data of missing values by setting the default number of lanes 'n_lanes' to 1 and set missing entry values for 'max_speed' to 0. We selected a Random Forest Ensemble Learning as a default classifier based on a 10-fold grid-search cross-validation (using Support Vector Machines, KNeighbors, Decision Tree, Adaboost, and Random Forest classifier from scikit-learn with default parameters), in which the Random Forest achieved the highest average F_1 score. The type of modeling procedure (person-dependent and person-independent) is explained in detail in the sections 5.3 and 5.5.

We also developed a real-time prediction app of VEmotion to classify emotions on unknown roads based on the learned classifier in which the mean emotion inference took 1.36s (SD: 0.246, min: 0.962, max: 1.996) in a 30-minute test ride.

4 USER STUDY

We conduct a user study to understand the impact of the VEmotion's contextual data on emotion prediction.

4.1 Apparatus and Method

We built a vehicle-usable iOS app that records the individual GPS and video stream and computes the variables described in Section 3 continuously during the ride[4]. We asked the participants to use this app the next time they used their personal car to ride and attach their phone to the windscreen. We recorded the daytime and asked the participant about their currently perceived emotion at the beginning of the ride. To collect a baseline of the participant's own interpretation of emotional states during the ride, we trigger a beep tone every 60 seconds for the participants to verbally provide their currently perceived emotions. We designed this emotion probing in correspondence to the *in-situ* categorical emotion response (CER) rating for collecting data on emotional experiences in vehicles [11]. Participants were instructed about the set of available emotions before starting the experiment (i.e., the basic emotions after Ekman [14]). The verbally expressed emotion was recorded while driving and is analyzed after the driving scenarios with a speech-to-text algorithm. As this procedure requires the passenger to talk during the ride and can be a distraction from first-order driving tasks, in a pre-study ($N = 5$) we optimized the time interval not to be annoying, ensure safety, and simultaneously cover the felt emotions appropriately. A post-hoc driving questionnaire showed that 9/12 participants were not bothered by the beep. The mean time-to-beep-response was 1.8 seconds. For an in-the-wild system that uses our architecture, the ground truth emotion assessment will not be required, and therefore, the system will not interact with the driver. A printout of the basic emotions was given to the participants before the start of the experiment. After the ride, the participant answers several subjective questions, including remarkable incidents.

4.2 Procedure

Twelve participants were invited through a mailing list from a pool of colleagues willing to participate in research studies. They were asked to download our iOS app beforehand and were equipped with a windshield smartphone retainer. The participants were asked before their next ride to call the study instructor via a remote call. In this call, the participants were asked about their demographics, frequency of driving, and feelings before the ride. Then we gave an introduction to our app. The participants were then asked to hang up, start the app recording, and drive freely to their chosen destination and after the ride to save the recordings and call the instructor. The instructor asked the participants about notable incidents while driving, emotions while and after driving.

4.3 Participants

We recruited 12 participants (eight self-identified as male, two self-identified as female) with an average age of 27 years (SD = 4.73). Six 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 16 minutes (SD = 11, min=7, max=52). The road type changed on average 7.9 times per ride. Participants expressed on average 4.41 distinct emotions during their ride (the duration between different expressed emotions across all users was 2 minutes 43 (SD=3 minutes 59).

5 RESULTS

We analyze the prediction performance of driver's emotions using the data captured by VEmotion. First, we evaluate the relative importance of single features of the data set collected by VEmotion. Then, we investigate the classification accuracy for emotion recognition based on facial expressions alone. Finally, we performed the following model evaluations: (1) a Leave-One-of-10-Road-Segments-Out cross-validation, (2) a participant-dependent Leave-One-of-10-Road-Segments-Out cross-validation, and (3) a Leave-One-Participant-Out cross-validation for evaluating VEmotion on unseen participants (i.e., participant-independent evaluation).

5.1 Relevant Features for Predicting Emotions

We collected 8986 instances of labeled data, namely a GPS location with a ground-truth label of the user's self-reported emotion. This corresponds to 1.1 seconds of driving depending on data validity, such as GPS fixes. Overall, 5780 were labeled as 'neutral' (64%), 2839 as 'happy' (32%), 177 as 'surprise' (2%), 130 as 'angry'(1%), and 60 as 'disgust'(< 1%). We start by investigating how decisive each feature was for creating a classification model. For this, we extracted the feature importance of the context variables, provided by VEmotion, in a leave-one-participant-out situation in Figure 3. As we employed a Random Forest classifier for emotion prediction, feature importance is measured as the popular mean decrease in impurity — this is defined as the total decrease in node Gini-impurity (weighted by the probability of reaching that node), averaged over all trees of the ensemble [6].

Of the context variables, 'vehicle_speed' was ranked highest in terms of feature importance. This might be because 'happy'

[4] Ethical approval was granted by the institutional review board of the university department

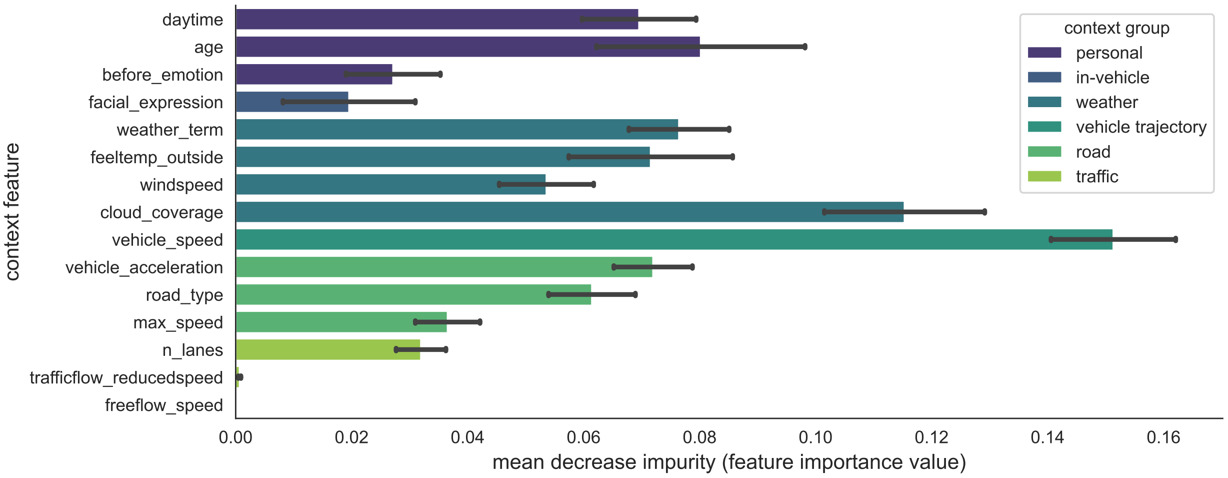


Figure 3: Feature importances measured by the mean decrease of Gini-impurity for the Leave-One-Participant-Out cross-validation.

emotions are often reported in unhindered speed scenarios. In contrast, related research [61] report higher negative emotions (i.e., 'anger' and 'fear') during unforeseen traffic incidents (e.g., high traffic densities or red light series) that require high cognitive demands. The information extracted by the traffic variables ('trafficflow_reducedspeed' and 'freeflow_speed') is assigned the lowest feature importances overall, which might be due to the vehicle trajectory features (acceleration and speed) working as proxy variables for unhindered traffic rides. Interestingly 'weather' and 'daytime' were assigned a medium level of feature importance. These environmental variables have been observed to impact emotional states in related psychological research [10, 61]. Related research has weakly associated negative emotional states to 'temperature' and positive emotions to 'sunlight'. However, weather influences tend to be highly dependent on person and age, which are additional context variables in VEmotion. Contrary to our expectations, the emotion reported before the ride was not assigned a very high feature importance, which may be due to mood changes when driving and unforeseen traffic incidents. The facial expressions captured by the frontal face camera have low feature importance. In contrast, all other recorded context inputs have medium-level importance. This underlines the usefulness of personal- and environmental input based on GPS location.

In a subsequent analysis, we evaluated the learned feature importances assigned conditional to the emotional class labels. We observe that 'cloud_coverage' and 'max_speed' information contribute highly to 'happy' emotions. Interestingly, 'freeflow_speed' has high feature importance conditioned on 'disgust' emotional states.

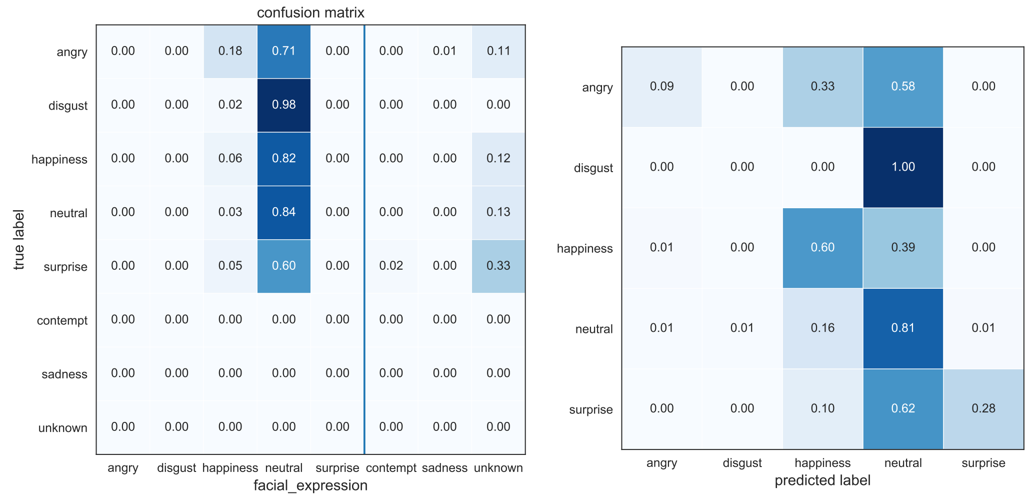
5.2 Validity of Facial Expressions

We analyzed the validity of 'facial expressions' as the sole indicator for the driver's emotions. We observe that 'facial expressions' predict five distinct emotional categories on the complete data set ('neutral', 'happiness', 'surprise', 'contempt', 'sadness', and 'unknown' if no face is detected). The emotions reported by participants on the ride are 'angry', 'disgust', 'happiness', 'neutral', and 'surprise', and thereby a subset of the facial expressions detected. Figure 4a shows the output of the facial expression engine and the self-reported emotion in our data set. The confusion matrix indicates that the facial expression engine detects many 'neutral' emotions (which are, in most cases, the true self-reported emotions). However, the self-reported emotion is often not correctly detected: 60% of 'surprised' emotions are predicted as 'neutral' emotions. At the same time, 82% of 'happy' and 98% of 'disgust' emotions are predicted as being 'neutral'. To conclude, the facial expression engine often yields a 'neutral' emotion class, ignoring and misclassifying heavily other felt emotions of the driver.

5.3 Leave-One-of-10-Road-Segments-Out Cross-Validation

Emotion recognition from 'facial expressions' alone is limited. To overcome this, we trained a Random Forest classifier (random state = 0, n_estimators = 50, max_features= $log2^5$) on the whole data set in the study, which included 'context variables'. We performed unshuffled cross-validation with 10-folds from all participants by segmenting the participant data into ten distinct consecutive folds (time-dependent road segments). Thereby, we construct a training set from nine training folds and one test set consisting of the remaining folds. This avoids a common constraint posted by a traditional 10-fold shuffled cross-validation evaluation since neighboring samples can be present in both training and test sets, resulting in trivial classification models. We term our evaluation approach 'Leave-One-of-10-Road-Segments-Out cross-validation', as this provides a better picture of the potential performance and robustness for

[5] The hyperparameters are found using a 10-fold hyperparameter tuning grid search.



	angry	disgust	happiness	neutral	surprise	contempt	sadness	unknown
angry	0.00	0.00	0.18	0.71	0.00	0.00	0.01	0.11
disgust	0.00	0.00	0.02	0.98	0.00	0.00	0.00	0.00
happiness	0.00	0.00	0.06	0.82	0.00	0.00	0.00	0.12
neutral	0.00	0.00	0.03	0.84	0.00	0.00	0.00	0.13
surprise	0.00	0.00	0.05	0.60	0.00	0.02	0.00	0.33
contempt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sadness	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
unknown	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(a)

	angry	disgust	happiness	neutral	surprise
angry	0.09	0.00	0.33	0.58	0.00
disgust	0.00	0.00	0.00	1.00	0.00
happiness	0.01	0.00	0.60	0.39	0.00
neutral	0.01	0.01	0.16	0.81	0.01
surprise	0.00	0.00	0.10	0.62	0.28

(b)

Figure 4: Comparison between facial expression output vs. VEmotion in predicting self-reported emotions on the road. The values of the confusion matrices are normalized on the true emotion class occurrences. The confusion matrix have different sizes as the facial expression engine falsely outputs a larger set of emotions (indicated by the blue vertical line). (a): Direct output of the facial expressions from Microsoft Azure. The detection accuracy of the self-reported emotions is 55.57%. (b): 10-fold cross-validation on the participants' unshuffled data using VEmotion accessing contextual data (no facial expression features included) trained with a Random Forest classifier trained with an accuracy of 71.70%. VEmotion achieves an unweighted average of the class specific recalls of 0.41 in (a) vs. a worse close-to-chance 0.18 when using facial expressions alone (b).

evaluating the classification performance. Furthermore, we believe that realistic data sets will contain relatively few, hence insufficient, 'angry' and 'disgust' emotion categories for model learning. Due to the imbalance of emotional classes that may be apparent in specific rides, we set the class weight of observations to 'balance'. This means that the Random Forest in VEmotion uses the values of the emotion class to automatically adjust weights inversely proportional to class frequencies in the training folds. Hence, we calculate the weighted average F_1 score over all emotion classes, which is defined as the harmonic mean of precision and recall, as an evaluation measure of classification performance.

VEmotion prediction performance of self-reported emotions in a 10-fold cross-validation on unseen ride segments is shown in Figure 4b. The overall accuracy of emotions is 71.70%. In other words, it is 29% better than relying on the 'facial expression' engine alone. We validated the facial expression engine by using other common facial expression classifier systems. We explored and applied a locally computable EmoPy trained on FER 2013 dataset [55] and AWS Emotion Recognition [47] to our data showing similar, sub-par results (predicting neutral/calm states is prevalent, accuracy: 0.55 and 0.07). VEmotion achieves an weighted average F_1 score of 71.30 (SD: 0.0713) across all emotional classes and outperforms the facial-expression-only system by 20 percentage points. We also observe that VEmotion only predicts classes that are actually expressed during the ride. In contrast, the 'facial expression' engine predicted contempt' or 'sad' emotions. Furthermore, VEmotion predicts 60% of 'happy' emotions vs. 6% using facial expressions only by only losing 3%. of correct 'neutral' emotion predictions. 'surprise' emotions can be accurately predicted with 28%. In contrast, 'angry' and 'disgust' emotions cannot be properly detected by VEmotion. The results indicate that contextual information can significantly improve the classification of emotional states, especially in detecting 'surprise' situations. VEmotion additionally discriminates better between 'neutral' and 'happiness' states of the driver. This evaluation is based on a 10-fold cross-validation and has access to training data of individual participants. We show that we can learn a global system for recognizing emotions 'on-the-go' with contextual and facial expressions. However, this comes at higher computational costs of having access to all participants' data and learning a participant-independent classifier. If the system should be used for uncalibrated modeling of a new driver's emotions, we perform an extensive evaluation in the next paragraph.

5.4 Participant-Dependent Leave-One-of-10-Road-Segments-Out Cross-Validation

Furthermore, we analyzed participant-dependent modeling using a participant-dependent Leave-One-of-10-Road-Segments-Out cross-validation. This means that we are training a participant-dependent model and validating on a holdout set of the participant using a 10-fold cross-validation scheme. The results are presented in Table 2.

Table 2: Accuracy, precision, recall, and weighted F_1 scores of the global 10-fold cross validation on unseen consecutive driving segments, aggregate of participant-dependent Leave-One-of-10-Road-Segments-Out cross validation, as well as leave-one-participant-out cross-validation.

	Leave-One-of-10-Road-Segments-Out Cross-Validation				**Participant-Dependent Leave-One-of-10-Road-Segments-Out Cross-Validation**				**Leave-One-Participant-Out Cross-Validation**			
Input	Accuracy	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1
Facial Expressions	.56	.57	.56	.51	.57	.66	.57	.56	.59	**.63**	.59	.54
VEmotion + Facial Expressions	.72	1.0	.72	**.72**	**.70**	.86	.70	.73	.64	.58	.64	**.57**
VEmotion	**.72**	**1.0**	**.72**	.71	**.71**	**.89**	**.71**	**.73**	**.64**	.56	**.64**	.56

5.5 Leave-One-Participant-Out Cross-Validation

We evaluate the possibility of a general classification model using all participant data except for one for training and using the last participant for evaluation. Semantically, this approach learns a model without knowing anything about the driver in advance and predicts the drivers' emotions independent from individual context emotion preferences. As we have a more complex prediction problem by not having learned from the held-out participant, we expect the overall prediction to decrease. The results of the experiment are shown in Table 2.

5.6 Comparison of Model Performances

Table 2 provides an overview of the prediction performance scores for the different evaluation procedures based on VEmotion. The Leave-One-of-10-Road-Segments-Out cross-validation approach, which uses all data samples from all participants, yields 71.70% accuracy. This is considerably higher than relying on 'facial expressions' only, which achieves a mean accuracy of 55.58%. Looking at the F_1 score, weighted for the class labels, VEmotion achieves a score of 0.7130 and 0.7164 with inclusion of facial expressions. In the next step, we performed a 10-fold cross validation only based on individual participants' data and aggregate the results over all participants (i.e., a participant-dependent Leave-One-of-10-Road-Segments-Out cross-validation). Here, we observe a similar prediction performance compared to the global model. VEmotion achieves here an average accuracy of 70.67% which is approximate 1.03%. smaller than in the global 10-fold cross-validation step. Also, the weighted F_1 score increased slightly to 0.7282.

In the participant-dependent cross-validation, we also observe that the VEmotion without the variables from the 'facial expression' (VEmotion) engine has a marginally higher precision of 88.59 % than VEmotion including facial expressions. This indicates that a high fraction of emotions are predicted with a low false-positive fraction. Looking at a much more challenging problem of predicting emotion categories of unseen participants in the Leave-One-Participant-Out scheme, the virtual sensor also outperforms the other models with an F_1 score of 0.56. The average accuracy of VEmotion is 63.71%. This is less than the achieved accuracy in the 10-fold cross-validation but remains a high-quality predictor if no information about the user is known in this multiple class output prediction. Since global and participant-dependent modeling of contextual emotions yield similar prediction qualities, we conclude that it is computationally favorable to learn participant-wise models over various rides, instead of learning global models that require data exchange of all participants. This also ensures that the inter-person and trip variety is sufficiently accounted for in the training sample. We stress the fact of imbalanced emotion classes that can only be acquired mainly through global data acquisition. Thereby learning a solely participant-dependent model puts the detection of emotions at a disadvantage that are not occurring frequently (e.g., 'surprise', 'angry', 'fear', and 'disgust'). Hence, facial expressions that do not occur frequently can still contribute to a robust model when collecting them from multiple participants. Finally, our 'in-the-wild study' does not show a significant benefit in including 'facial expressions' as features in our classification model. Thus, we propose omitting 'facial expressions' in practice, which would further reduce computational costs. Furthermore, facial expressions inhibit largely privacy concerns of the end-users and might raise a feeling of video surveillance while driving.

To answer the question of how many minutes of driving data is needed for VEmotion to be accordingly calibrated to predict emotions on the road with high accuracy,

5.7 Participant Fine-Tuning

We added a learning scheme below that illustrates how many minutes of driving data is needed for VEmotion to be calibrated for a high accuracy emotion prediction on the road. We used a leave-one-participant-out classifier to assess the emotion classification performance by incorporating the first x minutes of additional participants' driving data and evaluated the performance of VEmotion on the remaining driving data. Figure 5 presents the results of the analysis. We found that the first five to ten minutes have to be captured to achieve a mean precision of over 75% across participants due to the better discrimination of the classifier between neutral and happy states during the first 10 minutes (F_1 = .61). The drop in accuracy and F_1 after 10 minutes of training data is due to little held out test data which increases the variability of the prediction intervals. High precision of 80% and recall of 63.5% can be achieved when fine-tuning the classifier on the first 14 minutes. However, this requires the driver to label his perceived emotions 14 times, which may be annoying if done on every ride. We suspect our performances to increase heavily if multiple rides with fine-tuning in the first minutes are performed. Furthermore, VEmotion's application in practice would benefit from perceived emotion labeling in any special scenario within the ride and not just during the first minutes of driving.

6 DISCUSSION

Can we predict driver emotions based on driving context? To the best of our knowledge, VEmotion provides the first in-car sensor that combines implicit and non-intrusive measures to detect the driver's emotional states. In a user study with twelve participants, we find the highest classification accuracy when training a global model. We discuss the implications of our results in the following.

6.1 Driving Context Implies Emotions

Previous work hypothesized that the observation of driving behavior can be indicative of driver emotions [22, 35]. Indeed, our results show that certain features are predictive for driver emotions. In analyzing the feature importances, we found that 'vehicle dynamics', 'weather', and 'traffic flow' were highly predictive for emotions. This implies that the designer of empathic car interfaces should focus on the reliable measurement of these features when assessing emotions is a critical task. These can be integrated into existing emotion recognition engines or car navigation systems that are already integrated into vehicles or smartphones. Our results demonstrate that different users share common emotional categories influenced by the same contextual and environmental factors. In our real-world study, we notice a high imbalance of self-reported emotions, as most people either respond to feeling 'happy' or 'neutral' along their ride. This provides a challenging task for an appropriate data basis and proper classification of 'sad', 'fear', or 'disgust' states which are often observed with higher safety concerns [62]. These imbalanced emotion class distributions in the wild should therefore be extended in future data acquisition.

6.2 Comparing the Classification Performance between Facial Expressions and Driving Behavior

We find a difference between the classification performance for VEmotion, facial expressions, and a combination of VEmotion and facial expressions. Our study shows that the use of facial expressions alone results in the lowest classification accuracy compared to either VEmotion or VEmotion in combination with facial expressions in a real-world driving setting. Furthermore, our results show that not all emotions can be reliably detected using facial expressions. This includes the emotions 'angry', 'surprise', or 'disgust'.

Our results show that the emotion class 'neutral' is predicted most often by the facial expression engine. We suspect that the 'neutral' emotion class occurs frequently due to the low facial expressiveness in driving scenarios. Also, facial expressions are affected by user-to-user variability, resulting in individual differences in facial expressiveness and self-reported emotions. Further limitations include a moving driving environment, occlusion, and changing visibility conditions (e.g., sudden darkness in a tunnel). In contrast, VEmotion captures the driving behavior of the user in addition to facial expressions, which introduced performance increases of 38% in person-dependent (Leave-One-of-10-Road-Segments-Out cross-validation) and 10% in person-independent cross-validation schemes.

While our results show an improved classification performance for VEmotion, we find that the driving behavior and the perceived emotions are individual factors. Here, the resulting general model results in poor classification performances. However, training the model for each user individually yields a higher classification accuracy. VEmotion has to learn person-dependent discriminatory features from the contextual data to achieve acceptable accuracies. Therefore, the emotions predicted by VEmotion improves if more person-dependent information is available.

6.3 Enabling Empathic Vehicle-Applications with VEmotion

VEmotion allows the implementation of several use cases, however, our work intends to make a sensory system contribution of unobtrusively measuring emotions in the wild. VEmotion is beneficial in providing direction into what enjoyable drives are, and VEmotion's predictions[6] can inform infrastructure and road planning policies. For instance, it might be meaningful to enforce speed limits or narrow roads on some road segments to increase the overall road safety based on VEmotion. For example, VEmotion enables navigation functionality to invoke positive emotions. This idea has been proposed but yet has to be implemented [4]. Unknown route segments can be labeled with the respectively measured emotions. Car navigation can then be extended by routing after emotions. Other applications include the reflection of emotions after a ride. For example, a post-driving tool can visualize the perceived emotions for single road segments. Furthermore, future empathic car interfaces can utilize VEmotion to modulate driver emotions in real-time, for example, by playing pleasurable music [57].

6.4 Ethical Considerations

We emphasize an ethical as well as transparent use of VEmotion for application purposes and stress that emotions are intimate, personal, and vulnerable, where potential emotional insights can be manipulated to impact behavior in the long term [3]. Until now, many resources went into in-vehicle sensing which has resulted in much debate about the need for limiting facial recognition technology due to privacy and ethical considerations [3, 51]. The current work objectively looks at the significance of facial recognition and other data regarding what they might be telling us about the human perceived emotion. To the best of our knowledge, this is the first study where volunteers have allowed the recording of facial expressions together with contextual vehicular data in the wild. Our analysis reveals that contextual data obtained from a vehicle-CAN or smartphone is more efficient than actual facial recognition technologies. This has implications on several fronts: (1) We have been collecting vehicle data for the last 15 years, yet a potential exploit of this data might enable to backwards-infer human's perceived

[6] given a more broaden data acquisition

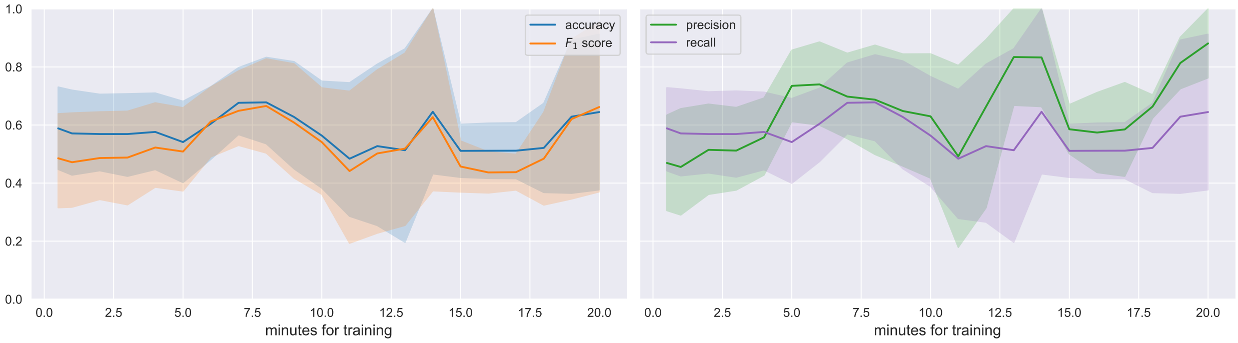


Figure 5: Classification performance for fine-tuning the Random Forest classifier with the first x minutes of participants riding data in a Leave-One-Participant-Out training scheme, in which the performance is computed on the available rest-duration of the ride data. The mean performance across all participants, as well as the 95% confidence intervals, are shown. Overall the performance of the classifier increases in all metrics (accuracy, F_1 score, precision and recall) knowing the first 5 minutes of personal driving data. Subsequently, the metrics converge, but precision is steadily increasing. We stopped computing the performance after 20 minutes due to little held out remaining driving test data.

feeling on this road given the features presented are available in the data. (2) Environmental contextual data offers a potentially more privacy-preserving and discomfort-reducing alternative to measure emotions in the wild. The connection between affect and emotions has always been emphasized. However, many other data variables can infer emotions without the need for recording affective or physiological variables. Our current work broadens the debate as to what type of data should be accessible by whom and for what purposes.

6.5 Limitations and Future Work

The robustness of our approach relies heavily on the quality of contextual input sensors. Thus, reliable in-vehicle emotion classification becomes less reliable as more features drop out. For example, facial expressions require a particular "expressiveness" of the driver to detect the emotion. Another example includes the dropout of contextual driving data, such as GPS connectivity, when driving through a tunnel. We also do not gain introspective insights on the on-goings of the driver's mind and instead describe the driver's perceived emotions via eight primary states; this abstracts a significant part in the much broader assessments of the multitude of felt psychological on-goings of the driver. To further reflect the relationship between emotional contextual triggers and emotional states, we will expand our work to include outside-view-camera input. Expressions via ecstatic hand gestures indicating angry affective states could not be found in the video stream but may also provide a direction for future camera-based affect features. Future work might also extend the outside-view of VEmotion by looking at other car's behavior through the use of more privacy concerning frontal video stream input. A more extensive database of rides with a wider variety and distinction of emotions and more extended personal driving history enables longitudinal studies. Here, we strongly stress acknowledging the context of the driver and surroundings besides the emotion prediction, which should be represented in the decision space of empathic car interfaces and data basis for emotion recognition engines. Finally, our results show that an 8-minute calibration procedure on unseen drivers is sufficient to achieve a satisfying accuracy of over 68%, while the beep sound was not perceived as annoying by the participant. However, different unobtrusive strategies for a suitable calibration of VEmotion, such as incident-based sampling, will be evaluated in future work, having the caveat of not accessing a high-resolution emotion assessment on all road types.

7 CONCLUSION

This paper presents VEmotion, a system that derives user emotions by assessing driving information. We found that context variables can be captured in real-time using GPS at low cost, optionally accompanied by a camera monitoring the driver. This finding is unique as comparatively few studies are performed 'in-the-wild' and with the use of personal computing devices as opposed to the bespoke in-vehicle sensors. We gain many insights by having the ability to record a much more fine-grained picture of the driver and its surroundings and potential influences on emotion with VEmotion in a noisy real-world environment. This provides automotive user interface designers with an additional tool to design unobtrusive empathic car interfaces deployed in real-world scenarios. Here, we are confident that VEmotion advances the field of emotion-aware car interfaces. To encourage research in this area, we publish the source code of VEmotion and the data set for further analysis by the research community[7].

ACKNOWLEDGMENTS

This research was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in TRR161 (Quantitative methods for visual computing, Project ID 251654672) in Project C06.

[7] https://github.com/davebeght/VEmotion

REFERENCES

[1] S. M. Alarcão and M. J. Fonseca. 2019. Emotions Recognition Using EEG Signals: A Survey. *IEEE Transactions on Affective Computing* 10, 3 (2019), 374–393. https://doi.org/10.1109/TAFFC.2017.2714671

[2] Victor M Álvarez, Claudia N Sánchez, Sebastián Gutiérrez, Julieta Domínguez-Soberanes, and Ramiro Velázquez. 2018. Facial emotion recognition: a comparison of different landmark-based classifiers. In *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*. IEEE, 1–4.

[3] Nazanin Andalibi and Justin Buss. 2020. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] Michael Braun, Jingyi Li, Florian Weber, Bastian Pfleging, 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

[5] Michael Braun, Florian Weber, and Florian Alt. [n.d.]. Affective Automotive User Interfaces - Reviewing the State of Emotion Regulation in the Car. In *To appear in ACM Copmuting Surveys*.

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[7] Yang Cai. 2006. Empathic computing. In *Ambient Intelligence in Everyday Life*. Springer, 67–85.

[8] Delphine Caruelle, Anders Gustafsson, Poja Shams, and Line Lervik-Olsen. 2019. The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action. *Journal of Business Research* 104 (2019), 146–160. https://doi.org/10.1016/j.jbusres.2019.06.041

[9] 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

[10] Marie Connolly. 2013. Some like it mild and not too wet: The influence of weather on subjective well-being. *Journal of Happiness Studies* 14, 2 (2013), 457–473. https://doi.org/10.1007/s10902-012-9338-2

[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] 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 The proceedings of AmI, the 2018 European Conference on Ambient Intelligence.

[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] Paul Ekman. 1992. Are there basic emotions? (1992). https://doi.org/10.1037/0033-295X.99.3.550

[16] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384. https://doi.org/10.1037/0003-066X.48.4.384

[17] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology* 53, 4 (1987), 712. https://doi.org/10.1037/0022-3514.53.4.712

[18] Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[19] 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

[20] Deborah Gould. 2010. On affect and protest. In *Political emotions*. Routledge, 32–58.

[21] Michael Grimm, Kristian Kroschel, Helen Harris, Clifford Nass, Björn Schuller, Gerhard Rigoll, and Tobias Moosmayr. 2007. On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving. In *Affective Computing and Intelligent Interaction*, Ana C. R. Paiva, Rui Prada, and Rosalind W. Picard (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 126–138. https://doi.org/10.1007/978-3-540-74889-2_12

[22] GM Hancock, PA Hancock, and CM Janelle. 2012. The impact of emotions and predominant emotion regulation technique on driving performance. *Work* 41, Supplement 1 (2012), 3608–3611. https://doi.org/10.3233/WOR-2012-0666-3608

[23] Javier Hernandez, Daniel McDuff, Xavier Benavides, Judith Amores, Pattie Maes, and Rosalind Picard. 2014. AutoEmotive: Bringing Empathy to the Driving Experience to Manage Stress. In *Proceedings of the 2014 Companion Publication on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS Companion '14)*. Association for Computing Machinery, New York, NY, USA, 53–56. https://doi.org/10.1145/2598784.2602780

[24] Ozgur Karaduman, Haluk Eren, Hasan Kurum, and Mehmet Celenk. 2013. An effective variable selection algorithm for Aggressive/Calm Driving detection via CAN bus. In *2013 International Conference on Connected Vehicles and Expo (ICCVE)*. IEEE, 586–591. https://doi.org/10.1109/ICCVE.2013.6799859

[25] Armağan Karahanoğlu and Çiğdem Erbuğ. 2011. Perceived Qualities of Smart Wearables: Determinants of User Acceptance. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces* (Milano, Italy) *(DPPI '11)*. Association for Computing Machinery, New York, NY, USA, Article 26, 8 pages. https://doi.org/10.1145/2347504.2347533

[26] A. Kolli, A. Fasih, F. A. Machot, and K. Kyamakya. 2011. Non-intrusive car driver's emotion recognition using thermal camera. In *Proceedings of the Joint INDS'11 ISTET'11*. 1–5. https://doi.org/10.1109/INDS.2011.6024802

[27] 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

[28] Janina Künecke, Andrea Hildebrandt, Guillermo Recio, Werner Sommer, and Oliver Wilhelm. 2014. Facial EMG responses to emotional expressions are related to emotion perception ability. *PloS one* 9, 1 (2014), e84053. https://doi.org/10.1371/journal.pone.0084053

[29] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. 2010. Presentation and validation of the Radboud Faces Database. *Cognition and emotion* 24, 8 (2010), 1377–1388.

[30] Y. Lin, H. Leng, G. Yang, and H. Cai. 2007. An Intelligent Noninvasive Sensor for Driver Pulse Wave Measurement. *IEEE Sensors Journal* 7, 5 (2007), 790–799. https://doi.org/10.1109/JSEN.2007.894923

[31] 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

[32] L. Malta, C. Miyajima, N. Kitaoka, and K. Takeda. 2011. Analysis of Real-World Driver's Frustration. *IEEE Transactions on Intelligent Transportation Systems* 12, 1 (2011), 109–118. https://doi.org/10.1109/TITS.2010.2070839

[33] E. Massaro, C. Ahn, C. Ratti, P. Santi, R. Stahlmann, A. Lamprecht, M. Roehder, and M. Huber. 2017. The Car as an Ambient Sensing Platform [Point of View]. *Proc. IEEE* 105, 1 (2017), 3–7. https://doi.org/10.1109/JPROC.2016.2634938

[34] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3723–3726. https://doi.org/10.1145/2851581.2890247

[35] 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

[36] Michael Oehl, Felix W. Siebert, Tessa-Karina Tews, Rainer Höger, and Hans-Rüdiger Pfister. 2011. Improving Human-Machine Interaction – A Non Invasive Approach to Detect Emotions in Car Drivers. In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, Julie A. Jacko (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 577–585.

[37] Michal Olszanowski, Grzegorz Pochwatko, Krzysztof Kuklinski, Michal Scibor-Rylski, Peter Lewinski, and Rafal K Ohme. 2015. Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. *Frontiers in psychology* 5 (2015), 1516. https://doi.org/10.3389/fpsyg.2014.01516

[38] 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

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

[40] Daniel S. Quintana, Adam J. Guastella, Tim Outhred, Ian B. Hickie, and Andrew H. Kemp. 2012. Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition. *International Journal of Psychophysiology* 86, 2 (2012), 168–172. https://doi.org/10.1016/j.ijpsycho.2012.08.012

[41] Genaro Rebolledo-Mendez, Angelica Reyes, Sebastian Paszkowicz, Mari Carmen Domingo, and Lee Skrypchuk. 2014. Developing a body sensor network to detect emotions during driving. *IEEE transactions on intelligent transportation systems* 15, 4 (2014), 1850–1854. https://doi.org/10.1109/TITS.2014.2335151

[42] Andreas Riener, Alois Ferscha, and Mohamed Aly. 2009. Heart on the road: HRV analysis for monitoring a driver's affective state. In *Proceedings of the 1st*

international conference on automotive user interfaces and interactive vehicular applications. 99–106. https://doi.org/10.1145/1620509.1620529

[43] James A Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological bulletin* 115, 1 (1994), 102.

[44] Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining* 8, 1 (2018), 1–26. https://doi.org/10.1007/s13278-018-0505-2

[45] Karen L Schmidt and Jeffrey F Cohn. 2001. Dynamics of facial expression: Normative characteristics and individual differences. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.* Citeseer, 547–550. https://doi.org/10.1109/ICME.2001.1237778

[46] B. W. Schuller. 2008. Speaker, Noise, and Acoustic Space Adaptation for Emotion Recognition in the Automotive Environment. In *ITG Conference on Voice Communication [8. ITG-Fachtagung]*. 1–4.

[47] Amazon Web Services. [n.d.]. AWS Recognition API. https://docs.aws.amazon.com/rekognition/

[48] Mimi Sheller. 2004. Automotive Emotions: Feeling the Car. *Theory, Culture & Society* 21, 4-5 (2004), 221–242. https://doi.org/10.1177/0263276404046068

[49] Felix W Siebert, Michael Oehl, and HR Pfister. 2010. The measurement of grip-strength in automobiles: A new approach to detect driver's emotions. *Advances in Human Factors, Ergonomics, and Safety in Manufacturing and Service Industries* (2010), 775–783.

[50] A. X. A. Sim and B. Sitohang. 2014. OBD-II standard car engine diagnostic software development. In *2014 International Conference on Data and Software Engineering (ICODSE)*. 1–5. https://doi.org/10.1109/ICODSE.2014.7062704

[51] Luke Stark. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 50–55.

[52] Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 782–793.

[53] Ronnie Taib, Jeremy Tederry, and Benjamin Itzstein. 2014. Quantifying Driver Frustration to Improve Road Safety. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 1777–1782. https://doi.org/10.1145/2559206.2581258

[54] Jianhua Tao and Tieniu Tan. 2005. Affective Computing: A Review. In *Affective Computing and Intelligent Interaction*, Jianhua Tao, Tieniu Tan, and Rosalind W. Picard (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 981–995. https://doi.org/10.1007/11573548_125

[55] ThoughtWorksArts. [n.d.]. EmoPy - Python Emotion Recognition Toolkit. https://github.com/thoughtworksarts/EmoPy

[56] Job Van Der Schalk, Skyler T Hawk, Agneta H Fischer, and Bertjan Doosje. 2011. Moving faces, looking places: validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion* 11, 4 (2011), 907. https://doi.org/10.1037/a0023853

[57] Marjolein D van der Zwaag, Joris H Janssen, Clifford Nass, Joyce HDM Westerink, Shrestha Chowdhury, and Dick de Waard. 2013. Using music to change mood while driving. *Ergonomics* 56, 10 (2013), 1504–1514. https://doi.org/10.1080/00140139.2013.825013

[58] Kiel von Lindenberg. 2014. Comparative analysis of gps data. *Undergraduate Journal of Mathematical Modeling: One+ Two* 5, 2 (2014), 1. https://doi.org/10.5038/2326-3652.5.2.1

[59] Heetae Yang, Jieun Yu, Hangjung Zo, and Munkee Choi. 2016. User acceptance of wearable devices: An extended perspective of perceived value. *Telematics and Informatics* 33, 2 (2016), 256–269. https://doi.org/10.1016/j.tele.2015.08.007

[60] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2020. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer* (2020), 1–20. https://doi.org/10.1007/s00371-020-01881-x

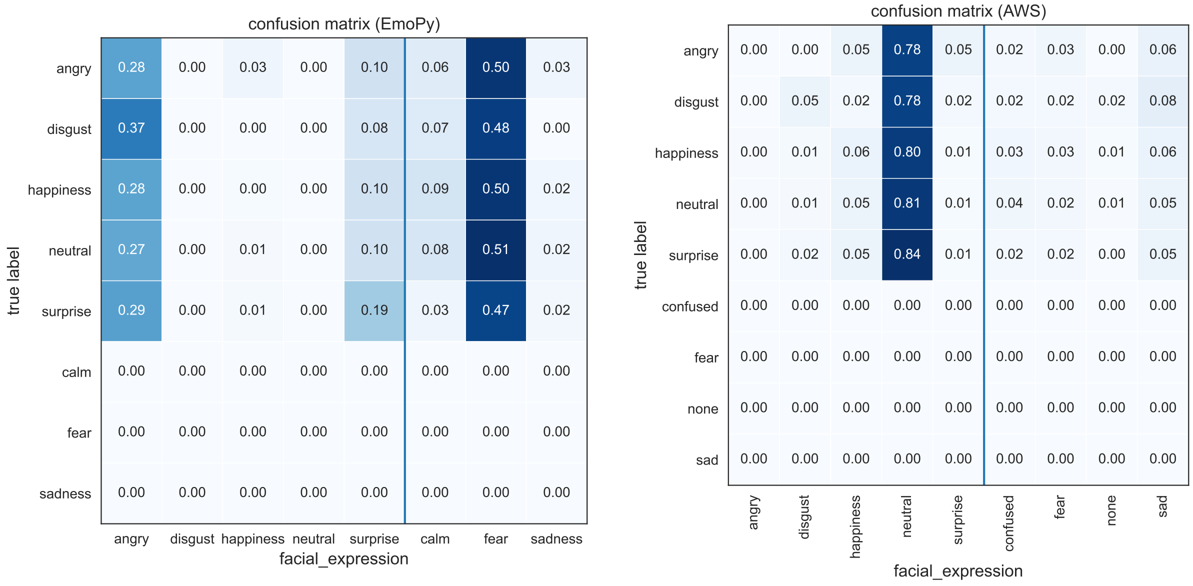
[61] Sebastian Zepf, Monique Dittrich, Javier Hernandez, and Alexander Schmitt. 2019. Towards empathetic Car interfaces: Emotional triggers while driving. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6. https://doi.org/10.1145/3290607.3312883

[62] 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

[63] 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.

APPENDIX

A BASELINE FACIAL EXPRESSION ANALYSIS



confusion matrix (EmoPy)

true label \ facial_expression	angry	disgust	happiness	neutral	surprise	calm	fear	sadness
angry	0.28	0.00	0.03	0.00	0.10	0.06	0.50	0.03
disgust	0.37	0.00	0.00	0.00	0.08	0.07	0.48	0.00
happiness	0.28	0.00	0.00	0.00	0.10	0.09	0.50	0.02
neutral	0.27	0.00	0.01	0.00	0.10	0.08	0.51	0.02
surprise	0.29	0.00	0.01	0.00	0.19	0.03	0.47	0.02
calm	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fear	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sadness	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(a)

confusion matrix (AWS)

true label \ facial_expression	angry	disgust	happiness	neutral	surprise	confused	fear	none	sad
angry	0.00	0.00	0.05	0.78	0.05	0.02	0.03	0.00	0.06
disgust	0.00	0.05	0.02	0.78	0.02	0.02	0.02	0.02	0.08
happiness	0.00	0.01	0.06	0.80	0.01	0.03	0.03	0.01	0.06
neutral	0.00	0.01	0.05	0.81	0.01	0.04	0.02	0.01	0.05
surprise	0.00	0.02	0.05	0.84	0.01	0.02	0.02	0.00	0.05
confused	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fear	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
none	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sad	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(b)

Figure 6: Facial Expression analysis using a) publicly available facial expression analysis tool EmoPy b) cloud-service AWS Facial Recognition service. For AWS, we assigned 'calm' recognition labels to 'neutral'. Both classification system offer little predictive power in explaining perceived emotions on the ride. The accuracy overall of a) is 0.0076 **and b)** 0.5445**.**

Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context

DAVID BETHGE*, Porsche AG, LMU Munich, Germany
LUIS FALCONERI COELHO*, Porsche AG, CODE University, Germany
THOMAS KOSCH, HU Berlin, Germany
SATIYABOOSHAN MURUGABOOPATHY, Porsche AG, Germany
ULRICH VON ZADOW, CODE University, Germany
ALBRECHT SCHMIDT, LMU Munich, Germany
TOBIAS GROSSE-PUPPENDAHL, Porsche AG, Germany

Driver emotions play a vital role in driving safety and performance. Consequently, regulating driver emotions through empathic interfaces have been investigated thoroughly. However, the prerequisite - driver emotion sensing - is a challenging endeavor: Body-worn physiological sensors are intrusive, while facial and speech recognition only capture overt emotions. In a user study (N=27), we investigate how emotions can be unobtrusively predicted by analyzing a rich set of contextual features captured by a smartphone, including road and traffic conditions, visual scene analysis, audio, weather information, and car speed. We derive a technical design space to inform practitioners and researchers about the most indicative sensing modalities, the corresponding impact on users' privacy, and the computational cost associated with processing this data. Our analysis shows that contextual emotion recognition is significantly more robust than facial recognition, leading to an overall improvement of 7% using a leave-one-participant-out cross-validation.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Computing methodologies →** *Machine learning*.

Additional Key Words and Phrases: Emotion Sensing, Affective Computing, Remote Sensors, Automotive, Empathic Interfaces, In-the-Wild Analysis

ACM Reference Format:
David Bethge, Luis Falconeri Coelho, Thomas Kosch, Satiyabooshan Murugaboopathy, Ulrich von Zadow, Albrecht Schmidt, and Tobias Grosse-Puppendahl. 2022. Technical Design Space Analysis for Unobtrusive Driver Emotion Assessment Using Multi-Domain Context. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 159 (December 2022), 30 pages. https://doi.org/10.1145/3569466

*Both authors contributed equally to the paper

Authors' addresses: David Bethge, Porsche AG, LMU Munich, Stuttgart, Germany, david.bethge@ifi.lmu.de; Luis Falconeri Coelho, Porsche AG, CODE University, Berlin, Germany, luis.coelho@code.berlin; Thomas Kosch, HU Berlin, Berlin, Germany; Satiyabooshan Murugaboopathy, Porsche AG, Stuttgart, Germany; Ulrich von Zadow, CODE University, Berlin, Germany; Albrecht Schmidt, LMU Munich, Munich, Germany; Tobias Grosse-Puppendahl, Porsche AG, Stuttgart, Germany.

2474-9567/2022/12-ART159 $15.00
https://doi.org/10.1145/3569466

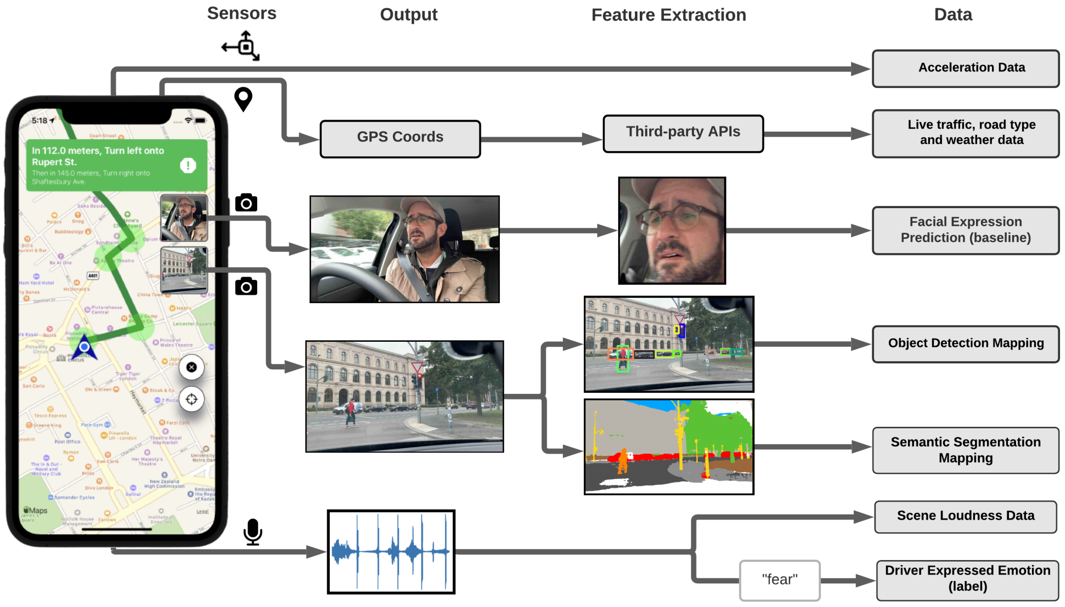

Sensors
Output
Feature Extraction
Data
Acceleration Data
GPS Coords
Third-party APIs
Live traffic, road type and weather data
Facial Expression Prediction (baseline)
Object Detection Mapping
Semantic Segmentation Mapping
Scene Loudness Data
"fear"
Driver Expressed Emotion (label)
In 112.0 meters, Turn left onto Rupert St.
Then in 145.0 meters, Turn right onto Shaftesbury Ave

Fig. 1. Multi-Domain Context Sensor Information used for the technical design space analysis. We predict emotions using a smartphone app, which employs five different sensors: accelerometer, GPS, front-facing camera, back-facing camera, and microphone. We use the latitude and longitude output by the GPS sensor to fetch data on live traffic, road type, and weather from third-party APIs. The front-facing camera captures the driver's face to perform facial expression recognition (used as a baseline). Visual scene segmentation and object detection are performed on the back-facing camera input.

1 INTRODUCTION

Emotions considerably impact drivers' performance, safety, and health [51]. Aggressive driving, for instance, plays a significant role in most fatal highway collisions each year [19, 45], leading to more severe injuries and fatalities [14]. Statistics show that more than 90% of traffic accidents can be attributed to human errors [57]. Negative emotional states while driving are associated with "poorer physical and mental health and quality of life", leading to an overall deteriorating driving performance [24].

Even sadness can seriously increase driving errors and decrease driving efficiency [25]. Consequently, the design, implementation, and evaluation of so-called *empathic car interfaces* has been the subject of previous research [8, 22, 62]. Empathic car interfaces aim to regulate driver emotions, thus improving the driving experience and reducing the risk of accidents. However, the unobtrusive assessment of driver emotions remains an open challenge.

Facial expression recognition, a frequently used method to detect emotions in driving contexts, often performs poorly [23, 34]. It requires subjects to overtly express their emotion through their facial muscles, failing to detect covert affective states. [6, 30]. As an alternative, past research used physiological sensing as a real-time measure to estimate driver emotions [22, 62]. Although this provides accurate assessments, it requires body-worn physiological sensors that reduce user acceptance. To address these shortcomings, contextual and behavioral driving data analysis emerged as an unobtrusive alternative to detect driver emotions [6, 26, 36, 44, 46]. Previous work in this area focused on a limited set of features frequently requiring internal car data access. Furthermore, it did not approach the topic of how different sensing modalities contribute to the classification performance of emotions.

Contrary to previous approaches that require body-mounted sensors [42], access to internal car information [36], or rely on a limited set of contextual features [6]; this paper investigates how emotions can be predicted by analyzing a rich set of contextual features unobtrusively captured by a smartphone, including audio-visual data. Furthermore, we derive a technical design space analysis to inform practitioners and researchers about the most indicative sensing modalities, their advantages and drawbacks.

We collect contextual and audio-visual data in an in-the-wild study with 27participants. The collected data comprises different context domains: weather, traffic, road type, and motion, including speed and acceleration. We record in-cabin audio and video, with the front-facing camera recording the driver and the back-facing camera recording the road view. We annotate the data using the participants' self-assessed emotional states. We compare our classification approach against emotion classification through facial expression as a baseline. A Random Forest classification using all features yields a classification accuracy of 59% (F_1: 0.45), outperforming facial expression classification by 7% and contextual classification by 13%. Finally, we present a technical framework showing how contextual and audio-visual sensing modalities influence the accuracy of emotion classification. Our work discusses how designers can select sensing strategies to prototype empathic car interfaces considering trade-offs related to computational cost and privacy concerns.

CONTRIBUTION STATEMENT

The contribution of this paper is threefold:

C1: An analysis of the technical design space for empathic car interfaces using a rich set of sensor streams from smartphones.

C2: A smartphone system and extensive data collection from in-the-wild driving evaluating contextual and audio-visual driving data for ubiquitous driver emotion assessments.

C3: Guidelines and considerations for application developers taking specific features for computational costs and privacy into account.

2 RELATED WORK

This section outlines the current understanding of emotions, how they are affected in a driving context, current emotion assessment practices, and driver emotion regulation methods.

2.1 Understanding Emotions while Driving

Emotional states can be schematized in different ways, the two most common categories being discrete and continuous emotion representations. Discrete representations of emotions derive from the works of Paul Ekman [15] who identified six basic emotions (i.e., anger, disgust, happiness, sadness, surprise, and fear) which are universally recognizable and encodable in facial muscles. In contrast, continuous emotion representation models encode the emotional state into a continuous value spectrum. Russell's circumplex model of affect [50] is one of the most commonly adopted continuous approaches. In our work, we employ a discrete emotion categorization.

Research on driving contexts has found interesting relations between specific scenarios and their potential for eliciting emotional states. Dittrich considered the "spatial-temporal distribution of drivers' emotions and their determinants" [12]. The study found that road intersections cause considerable amounts of emotional activity in drivers. Positive emotions are more likely at the beginning and end of a ride, adding strength to the claim made in this work that drivers' emotions can be inferred from contextual information.

Hancock et al. [20] concluded that as drivers' affective states change, so do the "measures of both longitudinal and lateral control of the vehicle", indicating that different emotions correlate with different mean vehicle speeds and the number of lane excursions. Another study [43] further examined the link between driver affect and driving styles, verifying that maladaptive driving styles, including reckless, careless, angry, hostile, and anxious, were associated with a lower capacity for emotional self-regulation. This finding is confirmed by Mesken et al. [41], which investigates the impact of driving context in eliciting certain emotions, listing anxiety, anger, and happiness as the most likely emotions to fluctuate in traffic situations.

A dynamically changing external environment can manipulate drivers' perceived workload and emotions. Faure et al. [18] showed that visually changing driving environments influence the driver's perceived cognitive workload. Frequent and unexpected changes in visual processing can change driver stress levels [52], resulting in differently perceived emotions by the driver [51].

2.2 Emotion Assessment

Detecting the driver's emotional state is essential for developing in-vehicle empathic systems to improve the driving experience. Related psychological research has shown that negative affective states can negatively impact driving performance [25], potentially causing traffic violations, driver distraction, and accidents. Therefore, previous research on in-car interventions has been designed to alleviate extreme emotional states. Braun et al. [8] show that these extreme states correspond to danger, while states with medium arousal levels and positive valence are recognized as optimal for safe driving.

Different approaches have been used to assess drivers' affective states, including physiology, facial expression, self-reports, or biosignals [4, 62]. To an extent, driving behavior and context have also been researched as an alternative assessment of emotions. Liu et al. [36] presented an emotion sensor based on CAN-BUS data and external environmental factors. In a long-term user study, they collected facial expressions, heart rate variability, CAN-BUS data, and environmental data to predict driver emotions using three classification models: CAN-BUS data only, video-only, and a fusion of both models. Their results show a participant-dependent classification accuracy of 71% and a leave-one-participant-out accuracy of 59.2% using a fusion-based model considering both video and CAN-BUS data. Our system is inspired by this approach, extending the collected video-only data by contextual driving semantics such as live traffic, weather, and audio data.

Furthermore, we evaluate the impact of different variables, including facial expression analysis as a common emotion assessment [15, 16], on the classification accuracy. Universal emotion assessments through facial expressions are disputed in previous work [23, 34], potentially requiring individual training for each person [30]. Here, we aim for a universally applicable approach using the driver's smartphone only to collect contextual and environmental data. We label the collected data using the participant's verbally self-assessed emotions.

In addition to driver context, previous research pointed out that environmental events influence driver emotions and stress levels—however, exclusively using driving context for predicting driver emotions is relatively new. Recently, Bustos et al. [10] proposed a system that recognizes driver stress levels by analyzing outside-view camera input during real-world driving conditions. The authors propose three models to predict a three-class stress level (i.e., low, medium, and high) from the image stream: (1) image classification with object presence features, (2) end-to-end image classification via a CNN, and (3) end-to-end video classification by temporal segment networks. Their results showed that the best average test accuracy of 72% was obtained using a video CNN. While their work focused on a second-person annotated stress label which should reflect the driving scene complexity [21], our work uses self-reported subjective emotion ratings. Bethge et. al [6] proposed a smartphone application that detects subjective discrete driver emotions. Their app uses GPS and third-party APIs to obtain road and traffic data representing environmental characteristics. While their approach offers a rich set of features to classify emotional states, their sensor set is constrained, containing no visual or auditory features. We added their work as a baseline for our study.

2.3 Driver Emotion Regulation

Empathic car interfaces can counteract emotion-related hazards by sensing the driver's state and intervening when potentially dangerous behavior is detected. Different mechanisms, including interventions, adaptive music [28], and lighting [9] were proposed in the literature. Such empathic car interfaces require continuous monitoring of the driver's emotional state, preferably via remote sensing, making our contribution relevant for application designers.

Summary

Previous research informs how emotions are interpreted, how they change while driving, and how they can be assessed in real-time to implement empathic car interfaces. However, they present drawbacks that may hinder the adoption of empathic car interfaces in the real world. Currently, most reliable assessments rely on body-worn sensors or are not universally applicable, e.g., by relying on internal car interfaces. Though there is a large availability of research evaluating different data streams, it remains unclear which features indicate emotions. We address this gap by collecting a rich contextual and audio-visual data set in an in-the-wild study using consumer smartphones. We analyze the indicativeness of the data streams to present a technical framework, depicting the contribution of contextual and audio-visual feature sets for the accuracy of driver emotion classifications.

3 DATA COLLECTION SYSTEM

In this section, we present a system that captures the contextual driving data from a smartphone using a combination of virtual and on-device sensor streams. We built an end-to-end data pipeline composed of a data-gathering mobile application with remote-sensing features and a data-processing pipeline. Informed by related work, we considered the following requirements: (1) in-the-wild data acquisition with a smartphone, (2) seamless integration of usable in-the-wild emotion sensing, (3) acquisition of features related to driving tasks or emotional state, (4) unobtrusive remote sensing and (5) the effect of the environment's physical characteristics (e.g., weather, road type and motion metrics) and visual complexity-related features [6, 36]. The end-user application seamlessly integrates in-the-wild contextual gathering to everyday driving tasks; hence, it features standard navigation functionality.

3.1 Mobile Application

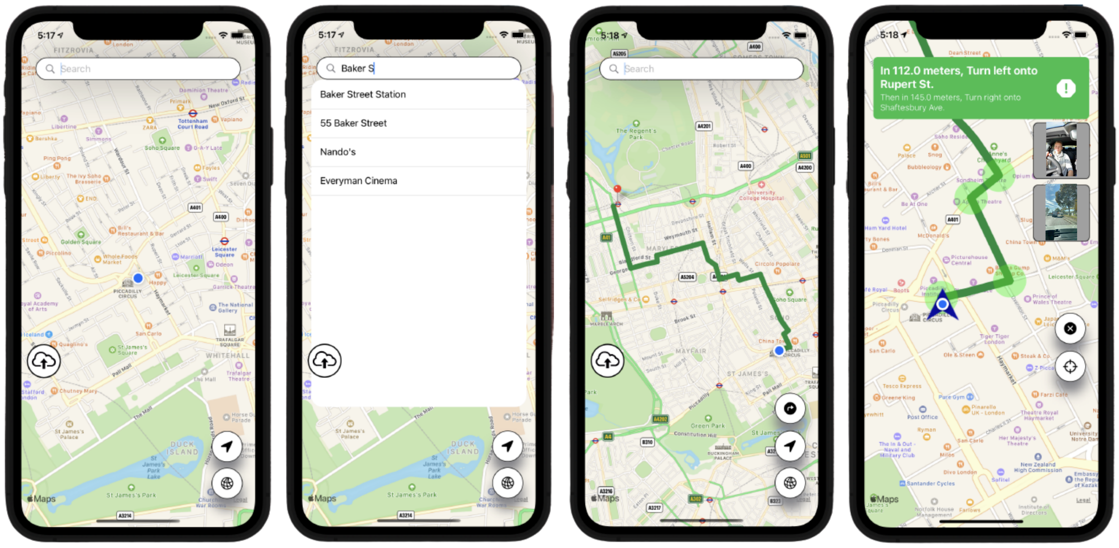
The mobile application, written in Swift[1], allows users to enter text-based descriptions of locations to obtain turn-by-turn spoken directions. We show the application user interface in Figure 2. The app requires an internet connection and runs on iPhones [2] that have iOS 13 or higher installed.

Figure 3 presents the mobile application's application architecture. The input modalities utilized by the application as data sources are the smartphone's front-facing camera, back-facing camera, microphone, GPS sensor, and accelerometer.

3.1.1 Application Main Loop. The audio/video controller monitors the application's main loop. Its output frequency is configured to 10 frames per second (FPS) for optimal performance and constitutes the central processing trigger. When the user activates the navigation mode, session recording begins. The application starts writing a sequence of RGB images from the front-facing camera, facing the driver, and back-facing camera, directed at the road, to local storage. A journey snapshot summary JSON object is generated for each frame pair. Each summary includes the frame number reference for posterior retrieval of images and a summary of the most up-to-date sensor-merged data.

[1] https://developer.apple.com/swift

[2] 11 Pro Max, 11 Pro, 11, Xs Max, Xs, Xr, SE 2. These iPhones enable acquiring front-facing and back-facing camera input at the same time.


Search
Baker S
Baker Street Station
55 Baker Street
Nando's
Everyman Cinema
In 112.0 meters, Turn left onto Rupert St.
Then in 145.0 meters, Turn right onto Shaftesbury Ave.
Maps

(a) (b) (c) (d)

Fig. 2. Navigation User Flow. Screenshots of the mobile application demonstrate user flow from the home screen (a), followed by route search (b), route overview (c) and after navigation is activated and session recording beings (d). In (d), we also show a small preview of the front-facing and back-facing camera stream to the user, e.g., to allow for position adjustments and to be transparent about all recorded information.

3.1.2 Data Synchronization. At the end of each session, the application concatenates the 'journey snapshot' objects accumulated and adds non-sensitive user-specific information (self-identified gender, age, car model - provided on the first app launch) and session metadata, e.g., start time, end, day of the week, to produce a JSON file summarizing the whole ride (see section 5 for a more detailed description of the dataset). At this point, the app outputs the audio recording file and uploads only the JSON summary to cloud storage (we used AWS S3[3] bucket) to register the session's occurrence. This upload does not include the audio and the images. Due to the size of the file bundles (i.e., approximately 2.5 gigabytes per session), we designed the upload process to be initiated by users at their leisure, making usage more convenient and avoiding unnecessary mobile network charges. This process does not conflict with future real-time capabilities and local predictions, but rather is necessary to explore potential design decisions on a fully functional dataset of source data. For real-time predictions, we imagine that remote context data (e.g., current weather) is collected from remote services, and ML models are run locally using frameworks such as CoreML[4]. Developers of future applications will additionally have to consider the trade-off between prediction frequency and energy consumption.

3.2 Data Post-Processing Pipeline

First, the data processing pipeline ingests data from the cloud storage and checks that the session data has been uploaded and there are no corrupted files. After, we perform multiple extraction mechanism steps: (1) visual features extraction, (2) audio analysis, (3) facial expression classification, (4) road data acquisition, (5) weather

[3] https://aws.amazon.com/s3

[4] https://developer.apple.com/documentation/coreml

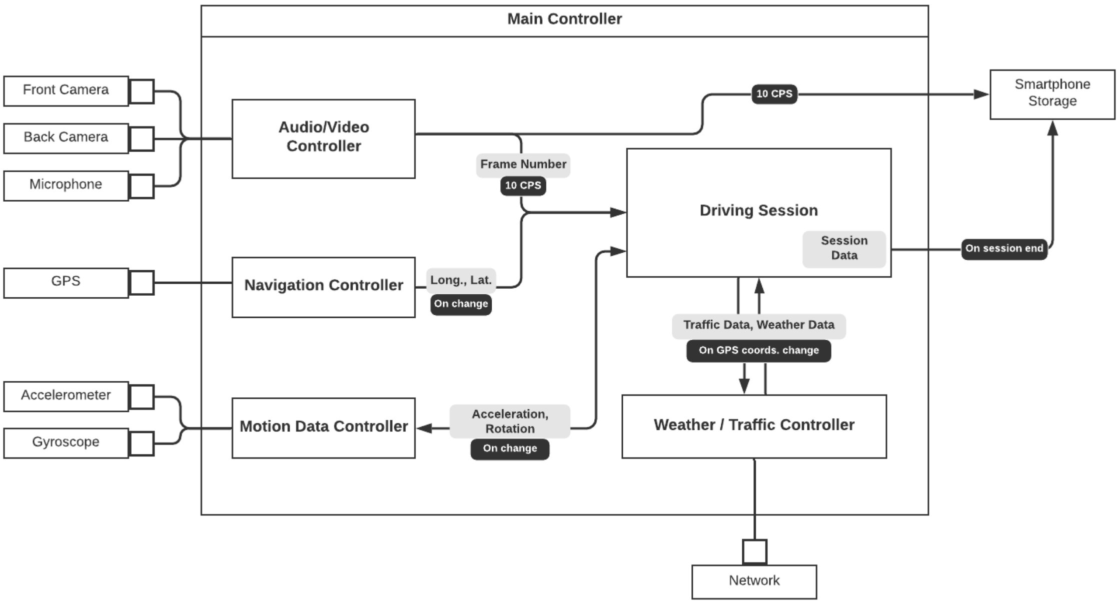


Fig. 3. Application's Reference Model. To the left of the *Main Controller*, sensors employed are displayed. *Sensing Controllers* intermediate the *sensor fusion* process by providing the *Driving Session* with up-to-date data from sensors. The Main Controller houses all the other controllers and manages data exchange. The Navigation Controller employs Apple's Mapkit framework to implement navigation and gather location data. The Weather/Traffic Controller calls both a weather API and a traffic API to fetch live data. The Session Data model holds at all times the most up-to-date value for each of the variables.

and traffic flow estimation, and (6) modeling. We describe the data pipeline in detail in the following paragraphs. The complete list of available features is presented in Table 1.

3.2.1 Visual Feature Extraction. We employ two parallel approaches to extract visual-related features from the road-facing frames: (a) object detection and (b) semantic segmentation.

Object Detection. For the object detection module, we used a PyTorch implementation[5] of a Yolo5 object detection model pre-trained on the COCO dataset[6]. The machine learning model outputs a list of objects detected, and their respective 2D bounding boxes (BB) expressed as normalized pixel coordinates (x, y, width, and height). The object classes are filtered to include only those of interest: cars, people, bicycles, motorcycles, buses, trucks, traffic lights, and traffic signs. We use the BBs to calculate the relative area of the object to the complete frame, representing the object's perceived size. We classify the relative area values using predefined thresholds into five different distance/perceived size classes (very far, far, medium, close, very close). Thresholds are devised based on observations about the frequency of occurrence of relative area values. The final output is a dictionary providing a scene summary with the number of objects for each detected class and the number of objects in each distance/perceived size class.

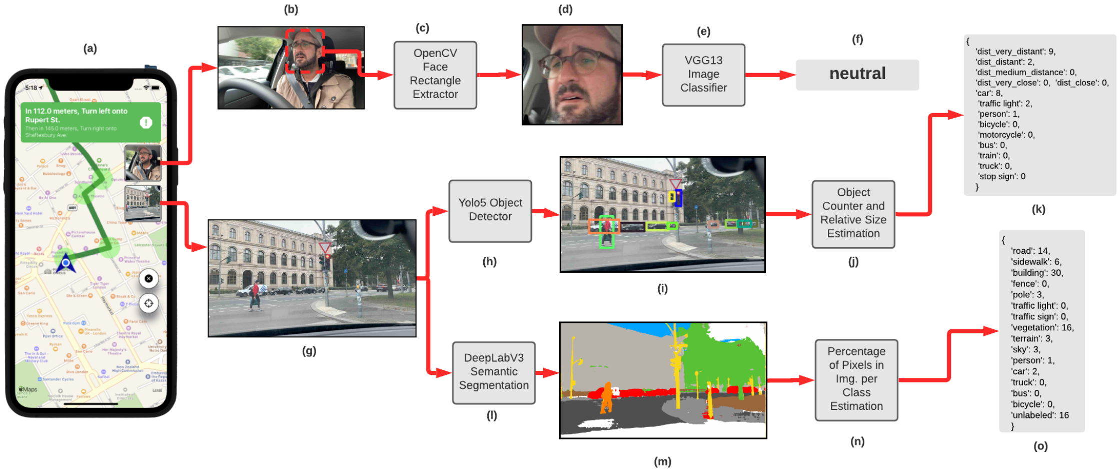
[5] https://pytorch.org/hub/ultralytics_yolov51
[6] https://cocodataset.org

Table 1. List of all available features in the dataset. We group the features by context and provide exemplary values or a description in the details column. The columns ‘frame_number’, ‘timestamp’, ‘audio_file_path’, ‘front_file_path’, ‘latitude’ and ‘longitude’ are not used as input for the machine learning models.

Context	Feature	Details
Reference Data	frame_number timestamp audio_file_path front_frame_path back_frame_path	The number reference for the session snapshot frame pair. e.g. 21/10/15, 18: 55:39:0025 p_01/session_id/audio.mp4 p_01/session_id/imgs/front_frame_501.jpg p_01/session_id/imgs/back_frame_501.jpg
Personal	sex car_model age participant_id emotion_before	male, female, other e.g. VW Polo, Porsche Taycan Participant’s age. e.g. p_01, p_02 Emotion before ride.
Session	session_id session_start session_end	e.g. 0751B8E9-3357-47E3-A862-CBFC60B88555 e.g. 21/10/15, 18: 54:49:0015 eg. 21/10/15, 19: 14:69:0485
Session Time	weekday daytime	Mon. Tue. Wed., Thurs. Fri., Sat., Sun. Morning, Afternoon, Evening, Night.
Motion	acceleration_x acceleration_y acceleration_z vemotion_acceleration	Acceleration on the x axis. Acceleration on the y axis. Acceleration on the z axis. (or acceleration_v1) Acceleration as in VEmotion [6].
GPS	speed latitude longitude	Vehicle speed in km/h. Latitude value of current location. Longitude value of current location.
Traffic Data	current_travel_time free_flow_speed current_speed free_flow_travel_time reduced_speed	Current travel time in seconds. The free flow speed expected under ideal conditions. The current average speed at the selected point. The travel time in seconds under ideal free flow conditions. Calculated with *free_flow_speed* minus *current_speed.*
Weather Data	wind_speed precipitation_24h_mm feel_temp_outside cloud_cover weather_term	Outside wind speed in km/h. Rain fall measurement in millimetres. "Feels like" temperature in Celsius. Percent representing cloud cover. e.g. cloudy, mostly cloudy, mostly sunny, sunny
Road Data	road_type max_speed num_lanes	e.g. cycleway, footway, living_street, motorway, residential. Maximum allowed speed for the current road. Count of available lanes on the road.
Facial Expression Pred.	facial_expression_label	Front-facing camera’s classified emotion|.
Perceived Emotion	label	Emotion expressed by the participant during the experiment.
Audio	audio_amplitude audio_loudness audio_zero_crossings	Audio amplitude averaged for duration of correspondent chunk. Audio recording average loudness for duration of correspondent chunk. Audio zero crossing rate of correspondent chunk.
Visual Complexity (Object Detection)	num_cars, num_people, bycicles, pedestrians, motorcycles, buses, trucks, traffic_lights, traffic_signs	Num. of objects detected in the back-facing camera frame per class.
	num_med_close_objs, num_very_close_objs, num_close_objs, num_very_far_objs, num_far_objs	Num. of objects at an estimated distance from camera.
Visual Complexity (Segmentation)	road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle	Percentage pixels in back-facing frame representing class.

Semantic Segmentation. We trained a Deeplabv3-ResNet model[7] on the Cityscapes Dataset[8] to perform semantic segmentation on the back-facing frames. We limited the training data to the classes relevant to our study: road,

[7] https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101


(a)
(b)
(c)
OpenCV Face Rectangle Extractor
(d)
(e)
VGG13 Image Classifier
(f)
neutral
(g)
Yolo5 Object Detector
(h)
(i)
Object Counter and Relative Size Estimation
(j)
(k)
DeepLabV3 Semantic Segmentation
(l)
(m)
Percentage of Pixels in Img. per Class Estimation
(n)
(o)

```
{
   'dist_very_distant': 9,
   'dist_distant': 2,
   'dist_medium_distance': 0,
   'dist_very_close': 0,  'dist_close': 0,
   'car': 8,
    'traffic light': 2,
    'person': 1,
    'bicycle': 0,
    'motorcycle': 0,
    'bus': 0,
    'train': 0,
    'truck': 0,
    'stop sign': 0
   }
```

```
{
   'road': 14,
   'sidewalk': 6,
   'building': 30,
   'fence': 0,
   'pole': 3,
   'traffic light': 0,
   'traffic sign': 0,
   'vegetation': 16,
   'terrain': 3,
   'sky': 3,
   'person': 1,
   'car': 2,
   'truck': 0,
   'bus': 0,
   'bicycle': 0,
   'unlabeled': 16
   }
```

Fig. 4. Contextual Unobtrusive Sensor Feed Pipeline showing image processing from acquisition (a) to output features used for emotion prediction (f, k, o). The frame containing the driver's face (b) is cropped into a face rectangle (d) with Python OpenCV (c) and run through an emotion classifier (e). The back-facing camera frames (g) are processed with: 1) a Yolo5 object detection model (h), whose output (i) is further processed by (j), 2) a DeepLabV3 semantic segmentation model (m) which outputs a dictionary (o) of pixel per class.

sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, bicycle. This model's processed output consists of an array of shapes in which each pixel is assigned a class (m). Following, we calculate the percentage of pixels occupied by each class into a dictionary (o). The percentage of pixels associated with a specific road class attribute helps the system to understand how complex the visual field may be to the driver. For instance, a high number of pixels associated with cars and pedestrians may be due to a traffic jam and challenging driving scenarios. The visual segmentation engine's output is shown in Figure 5.

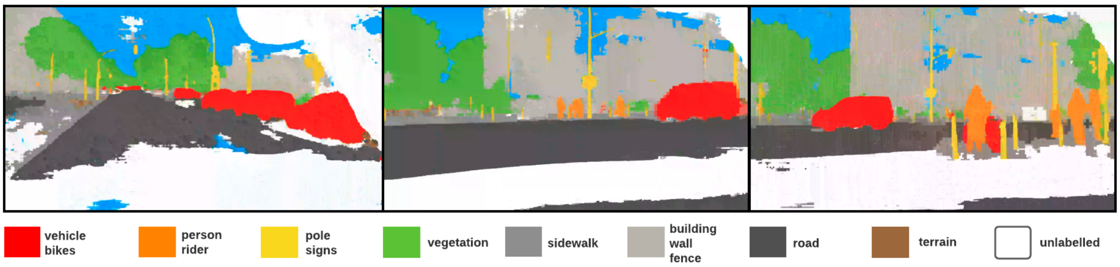


Fig. 5. Semantic Image Segmentation. Images showing segmentation results with different colors representing the predicted semantic classes. We color vehicles (cars, trucks, buses, trains) and bikes (motorcycles and bicycles) red. Yellow shows poles, traffic lights, and traffic signs.

[8] https://www.cityscapes-dataset.com

3.2.2 Audio Analysis. The smartphone's microphone is used in two ways: (1) to extract in-car loudness, (2) to compute the zero-crossing rate of the audio signal, and (3) to extract the annotated emotional labels expressed by the participants.

Loudness Extraction. We use audio amplitude and loudness in decibels (dB) to represent in-cabin driver auditory stimuli. We segment the audio stream into chunks of 0.5 seconds to calculate its mean amplitude. The formula for deriving loudness in dB is as follows:

$$loudness = 20 * log_{10}(\sqrt{\overline{chunk^2}}) \quad (1)$$

$$amplitude = \overline{chunk^2} \quad (2)$$

Zero-Crossing Rate. The Zero-Crossing Rate (ZCR) of an audio frame is a measurement of an audio signal's human-perceived *noisiness*. We calculate ZCR by counting the number of times a given audio signal crosses the zero axis and dividing it by the length of the frame. Unlike *loudness*, it incorporates spectral aspects of the signal and is widely used by applications in speech analysis [38, 48, 56] and musical genre classification [59]. Therefore, it is a good representation of driver auditory stimuli. Again, we derive a ZCR value for each chunk of 0.5 seconds.

Emotion Labels. Annotated labels are extracted manually onto a text file from the session's audio recordings. Label definitions and procedure of labeling are discussed later in the experiments section 4.

3.2.3 Facial Expression Classification. We use facial expression recognition to obtain baseline metrics for model performance comparison. Thus facial expressions are not included as features in the modeling phase. To extract facial recognition predictions, we use a face rectangle extractor and run its resulting image through a VGG13-based image classifier trained on the FERPlus Dataset[9]. Furthermore, as an additional baseline, we apply the Microsoft Face Recognition API classifier[10]. This step outputs a facial expression prediction for each of the session's front-facing frames. We stress that the emotion label provided by the facial expression classifier is not used as a feature for our model but represents a baseline metric.

3.2.4 Road Type Data. In order to detect the road infrastructure components of in-the-wild driving thoroughly, we acquire road-type-related features via reverse geocoding from OpenStreetMap[11] with the Python package OSMnx[12]. We download a high-definition map for each unique combination of GPS coordinates and search for the closest road node object to extract the relevant data. From the closest road node object, we extract the following attributes: 'road_type' (e.g., residential), the number of available lanes on the current road ('n_lanes'), and the maximum allowed speed on the current road ('max_speed').

3.2.5 Weather and Traffic Flow. We request weather information for each distinct GPS coordinate pair from the Microsoft Azure Maps API[13]. We include the following weather conditions: weather description, approximate outside temperature, cloud coverage, and wind speed. We also infer the traffic flow by requesting the speeds and travel times of the road fragment closest to the given coordinates using the Microsoft Maps Traffic Flow API[14].

[9] https://github.com/microsoft/FERPlus

[10] https://azure.microsoft.com/en-gb/services/cognitive-services/face

[11] https://www.openstreetmap.org

[12] https://github.com/gboeing/osmnx

[13] https://atlas.microsoft.com

[14] https://atlas.microsoft.com/traffic

4 EXPERIMENT

In the following section, we describe the details of the in-the-wild driving experiment.

4.1 Participants

In total, 27 participants (five female, ages 21 to 63, $\mu_{age} = 30.9$, $\sigma_{age} = 9.8$) took part in the experiment. The participants drove in total 48 sessions, with a total duration of 663.93minutes and a mean duration of 13.83 minutes (min = 9.0, max = 28.62, $\sigma_{duration} = 3.54$). The average number of unique emotions reported per session was 2.79.

4.2 Procedure

We asked participants to download the mobile app from Apple's beta-testing platform "TestFlight" and instructed them to use it as their navigation tool during two to three journeys with a duration of between 10 and 15 minutes. We also recommended that journeys be at different times of the day and, preferably, on different days to diversify the data collected as much as possible.

We needed the participant's front-facing camera to be pointed towards their face and the back-facing video stream to capture the road from the driver's perspective. Participants had to attach their smartphones to the windscreen.

The first time the subjects launched the app, they were asked for driver-specific context information, precisely their age, self-perceived gender, and car model. After that, we instructed the participants to choose their destination freely. Upon entering navigation mode after accepting the route proposed by the app, subjects were asked to select their pre-ride felt emotion and provide their emotion after the ride. Ethical approval for the experimental procedure was granted by the institutional review board of the university department.

4.2.1 Annotation of Emotions. To link the acquired contextual data with emotions on the road, we present the experimental design of emotion annotation in the following section. First, we explain how subjectively felt emotions can be acquired in the vehicle context. Second, we explain the mapping procedure and trade-offs between the expressed emotion and contextual ongoings. In our case, the driver expresses their emotions during the ride via voice without the need to let go of the steering wheel. In preparation for the experiments, subjects were asked whether they felt confident expressing their emotional states while driving. The verbally expressed emotion was recorded while driving and analyzed offline using a speech-to-text algorithm. We triggered a beep tone every 60 seconds for the drivers to express their discrete emotional state (a list of valid responses was given to the participants beforehand). Based on Ekman's basic emotion theory [16], we selected eight basic categorical emotion categories as possible response values: 'happiness', 'anger', 'fear','surprise', 'neutral', 'contempt', 'disgust' and 'sadness'.

Similarly to Bethge et al. [6], we adopt the *in-situ* categorical emotion response (CER) rating [13] for labeling in-the-wild emotions. We opted against continuous emotion labeling (DER) in the form of valence-arousal ratings, as users would need to select their continuous emotional rating via touch on an in-cabin device. Touch interactions are shown to distract from first-level driving tasks and pose a risk factor in this study [35]. The free categorical emotion response method is found to have practical limitations as it can generate a large number of labels. Consequently, Dittrich et al. [13] recommend adopting in-situ categorical emotion ratings (CER) with an "appropriate number and naming of categories that cover a significant range of emotions".

Furthermore, it is challenging to find the optimal time interval between prompting the driver for their emotion. On the one hand, we do not want to distract, bias, and annoy the driver when asking too often for an emotion. On the other hand, we want to ask as frequently as possible to have a granular resolution of the emotional ground truth that helps to learn a better link between our features and emotions. Using this annotation procedure, we link the expressed emotions to contextual data within the window of the previous 60 seconds. However, this approach is deliberately oversimplified, as emotion transitions might not be correctly reflected in the annotated data. We accept this trade-off in favor of a more realistic driving experience with the highest acceptable amount of interruptions. We further address the limitations of the experimental choice of defining emotional labeling in section 7.5.

5 DATASET CHARACTERISTICS

The following section presents an overview and necessary preprocessing steps of our data. The dataset consists of 48 sessions of different participants driving in the wild with our system explained in Section 3 and the data-gathering procedure described in Section 4.

5.1 Data Preprocessing

We experienced performance oscillations with the data-gathering application due to, e.g., battery state differences. These performance inconsistencies caused some session fragments to have a higher frame output than others, resulting in inconsistent data distribution data over time. Therefore, the dataset was down-sampled to 3 Hz, i.e., using three frames per second as data entries. We decided to drop the first minute of each session due to the time difference between starting our app and actual driving. Our preprocessing removes, on average, 7% of data per participant due to, e.g., invalid sensor information.

Some participants had difficulty adhering to the experiment's predefined emotions and used non-complying labels. We substituted some of these labels with synonyms. 'Scared' and 'concerned' were renamed to 'fear'. 'Annoyed' and 'frustrated' were replaced with 'anger'. Occurrences of 'stressed' were attributed to the label 'unknown' due to its ambiguity (could be interpreted as 'anger' or 'fear'). Other non-complying labels such as 'curious' and 'confused' were also changed to'unknown'. Emotion labels with 'unknown', duplicate values, and other rows with missing data were removed from the dataset.

5.2 Data Summary

We briefly give an overview of the preprocessed dataset in the following section and will recap the feature streams from our system thereafter.

After preprocessing, the dataset comprises 97020 samples (663.93 min) of labeled driving data from 48 sessions with 27 participants. We present an overview of our dataset in Figure 6. Due to the wide variety and depth of the acquired data, we only plot a subset of available features in Figure 6. We plot the Pearson-correlation matrix of all extracted features in the appendix Figure 11.

5.2.1 Perceived Emotion Labels. An overview of the perceived emotion per participant is shown in the upper left plot of Figure 6. Overall, we observe many real-world traffic conditions where drivers felt 'neutral' (57%), which is unsurprising given normal traffic conditions during many rides. The participants also perceived happy emotions 17% of the time. Negative emotional states did not occur frequently and were expressed primarily by a few participants. Drivers expressed numerous times to feel 'fear' which can be explained by some participants driving non-frequently and feeling nervous in complex traffic situations. Apart from sadness, all other pre-selected emotions (neutral, fear, happiness, anger, surprise, disgust, contempt) occurred throughout the sessions. In average, 2.79 distinct emotion categories were expressed per ride. There was only one participant who expressed a sad emotional state, while all other states were felt by multiple drivers.

5.2.2 Speed. The histogram in Figure 6 (right plot in the second row) shows the distribution of 'speed' (in km/h) across all sessions, revealing that in most rides, the range goes from 0 to 50 km/h. Two sessions have speeds surpassing the 50 km/h mark and going up to 200 km/h. Most sessions took place in cities, whereas two collected data on a motorway with no maximum speed limit.

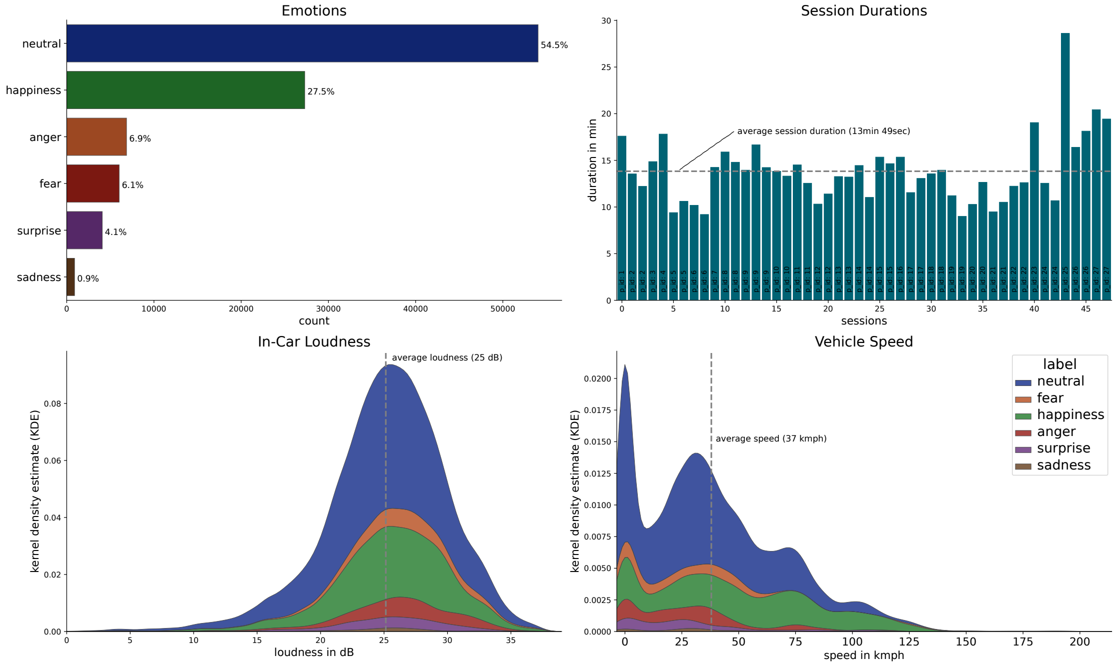


Fig. 6. Overview of the acquired data. While different personal meta-data statistics are shown in the figures in the top row, the bottom row shows the kernel density estimation (KDE) [47] of some contextual data including their label distribution.

5.2.3 *Visual Object Detection.*

The computer vision module of our system tracks granular changes in environmental visual ongoings. Figure 7 shows an exemplary output of the sensor feed of the visual features 'num_cars', 'num_people' and 'num_traffic_lights' being computed. We observe that the computer vision module can detect ongoing traffic situations, e.g., pedestrians or traffic lights while driving.

6 RESULTS

The following section addresses how context, captured driving data, and environmental factors predict driver emotions. First, we analyze the features' importance for predicting driver emotions. After that, we evaluate the prediction performance of the different features in an extensive cross-validation setup and compare them against several baselines. At last, we compare the features' characteristics in terms of privacy and computational costs.

6.1 Emotion Classification Module

We set a Random Forest Ensemble Learning as a default classifier based on a 10-fold grid-search cross-validation[15]. The 'class_weight' parameter of the random forest is set to 'balanced' to ensure that the algorithm can handle an unbalanced label distribution. Further hyperparameters are found using a 10-fold hyperparameter tuning grid search (random state = 0, n_estimators = 50, max_features= log_2). This machine learning model is kept

[15] Using Support Vector Machines, KNeighbors, Decision Tree, Adaboost, and Random Forest classifier from scikit-learn with default parameters. The Random Forest achieved the highest average F_1 score.

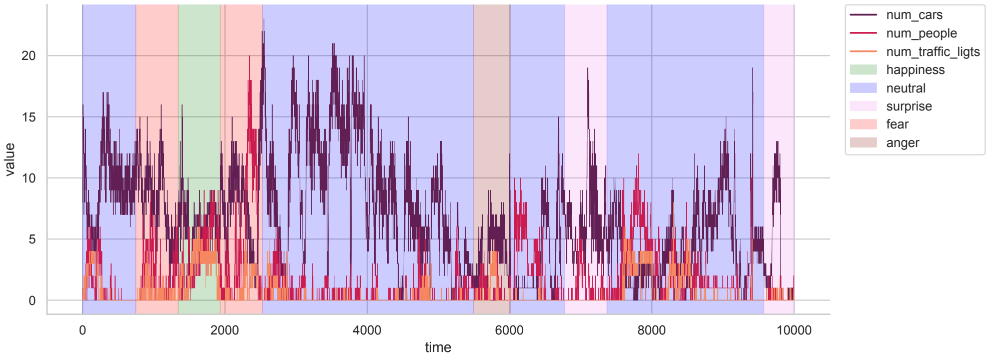


Fig. 7. Stream of outside-view visual complexity features for a sample session (participant 1). The y-axis shows the individual feature values, whereas the x-axis denotes the number of distinct, consequential data entries (sampled at 100 Hz). The object detection module recognizes the number of cars on the road, the number of people in the visual field, and the number of traffic lights at every moment of the drive. The prediction of the number of elements in the visual field varies as occlusion, shaky frames, and lighting conditions can occur.

identical to related work [6] to enable the comparison of results. We further evaluate the performance of a deep-learning-based feedforward neural network using all features. The neural network parameter settings are explained in detail in Appendix A.3. We explain the evaluation procedure in detail in the Section 6.4.

6.2 Importance of Features for In-The-Wild Emotion Recognition

We start by investigating how decisive each feature is for creating a classification model. Thus, we extracted the feature importance of the contextual variables: In a leave-one-participant-out situation, we assess the permutation importance for each variable, which is defined as the decrease in the balanced F_1 score of the classification algorithm. The permutation importance can be seen as a metric of how much performance we lose (here measured in F_1 score) if we do not have access to a specific system feature. This is done by randomly sampling the specific variable and thereby making the variable not-containing any meaningful information. The higher the permutation importance of a feature, the higher its predictive power, i.e., the more performance the classification decreases when it is unavailable. The feature importance does not provide information on which feature value contributes to a specific label prediction. We refer the reader to the concept of local feature importance computation, e.g., SHAP values, which could explain feature importances of a value range given an individual data object [40]. These importances are specific to a subject; therefore, this paper does not provide a local feature assessment. Figure 8 shows the calculated permutation importance for all features. In general, we observe that some features show very high importance, and most features do not. Overall, we detect a high importance of vehicle speed for the emotion classification decision. This finding overlaps with related work, which shows that free-flow highway driving and emotional happy states are tied. In contrast, low-speed values, combined with unforeseen traffic incidences such as traffic jams, have been associated with negative emotional feelings such as 'anger' and 'contempt' [61]. The available number of lanes on the road is a significant predictor in classifying the driver's emotional states, which is unsurprising as a high number of lanes is weakly correlated with the traffic conditions, i.e., speed and acceleration behavior [33].

Furthermore, the felt temperature outside and the number of pixels associated with the sky ('segment_sky') also show high-importance measures. The high feature importance in both sky and environment temperature is

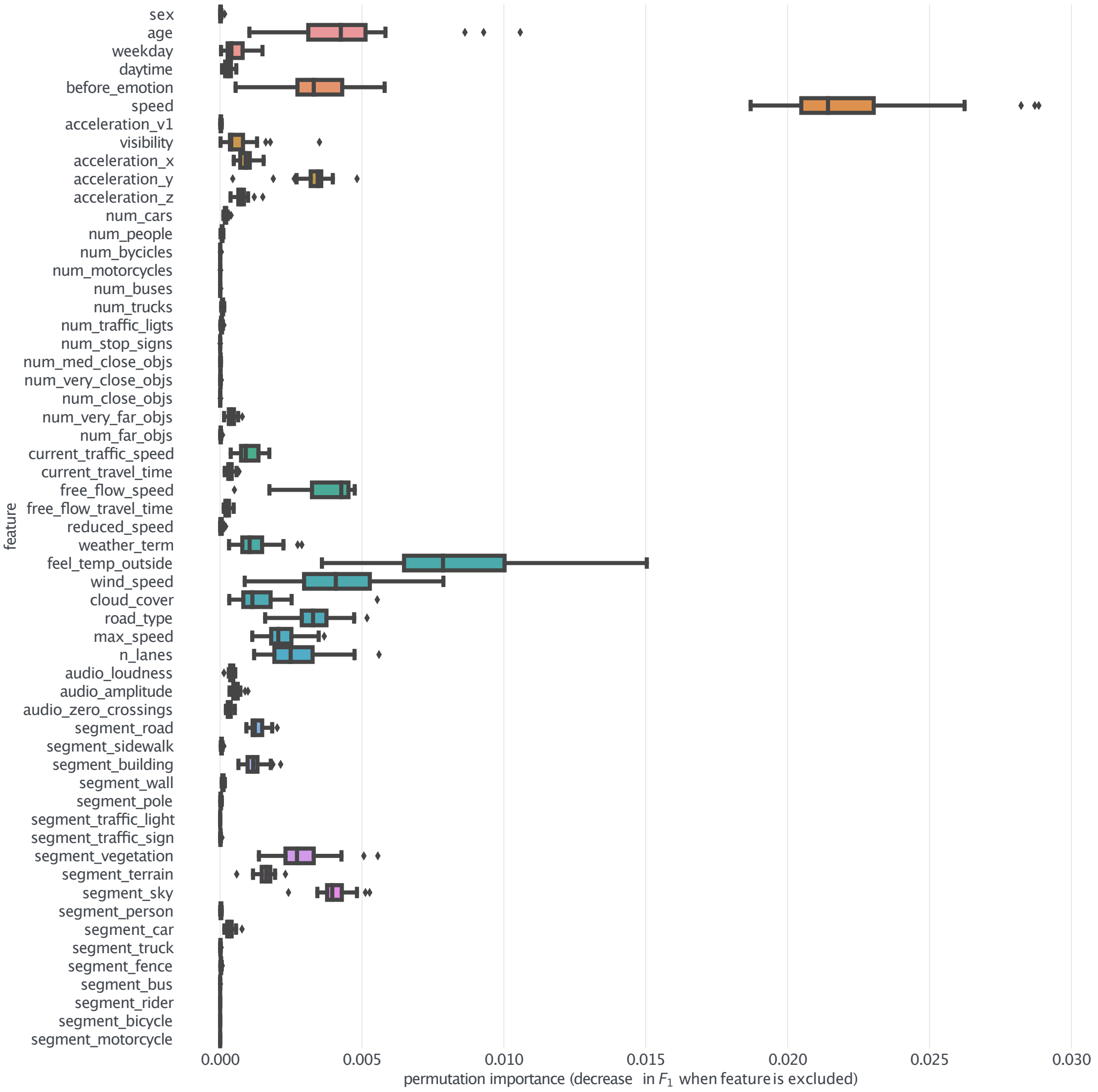


Fig. 8. Permutation importance of all available features. The permutation importance is characterized by a decrease in the model's F_1 score if the individual feature values are randomly permuted, i.e., made uninformative. Therefore, a high F_1 score indicates high feature importance to the emotion classification decision. Low permutation importance suggests that omitting the specific feature would not result in large prediction performance loss.

interesting, as related work has shown that greater sky exposure and air temperature tended to make drivers report lower stress levels and lower negative emotional states [3, 29].

Interestingly, the features representing drivers auditory complexity (‘audio_loudness’, ‘audio_amplitude’ and ‘audio_zero_crossings’) have low permutation importance scores making them not highly useful for classifying subjective emotional states. Driver speech analysis can predict human emotions successfully, and loud in-cabin sounds are potentially associated with driver distractions and the prevalence of annoyed and angry emotions [27, 49]. However, our results do not detect high audio feature importance, indicating that the other assessed contextual variables are more indicative of driver emotions. Although our in-the-wild results recommend omitting in-cabin audio recordings, further research is necessary to evaluate the impact of different audio features on emotions. For example, more advanced speech semantics, including tonality, pitch, or frequencies, can be more indicative of emotion recognition. Here, we expect individual differences in the audio data to mitigate the general classification performance at the cost of disclosing more privacy-sensitive data. In Section 6.5.2, we will further discuss the privacy-related concerns of in-cabin microphones.

6.3 Visual Driving Scene Features for Emotion Recognition In-The-Wild

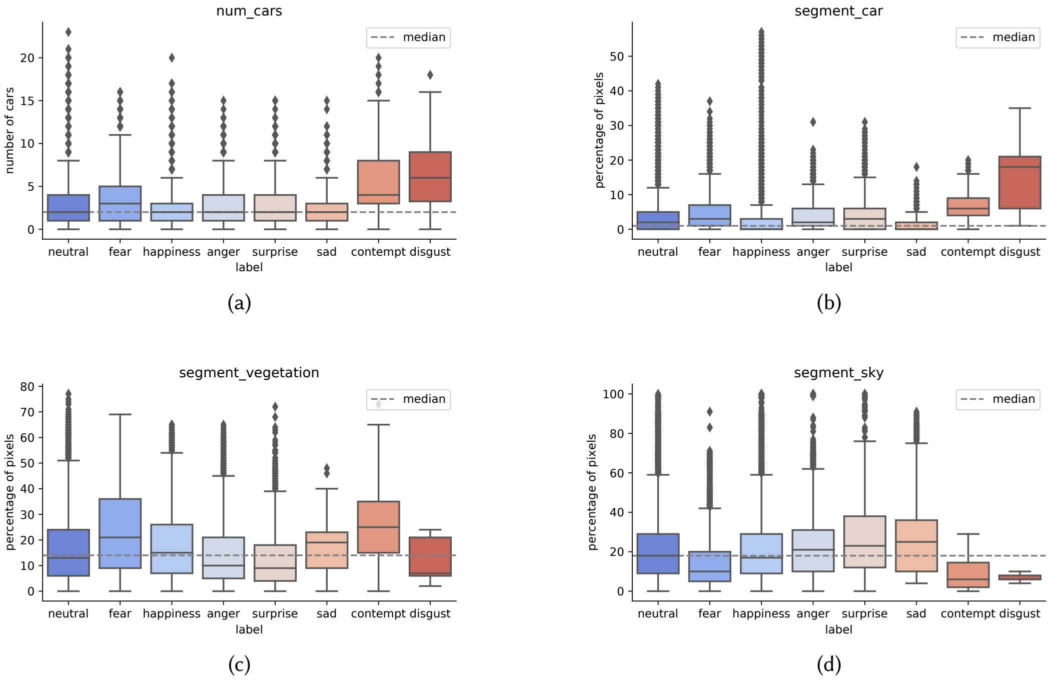
The feature importance analysis provides insights into how indicative a feature is for predicting subjective emotions. Analyzing the driver’s visual scene is beneficial for understanding the driving task’s complexity and the environment’s aesthetics [10]. This visual information can help to deduct the driver’s well-being. The following section will further analyze the link between the extracted visual features and emotions in the wild.

We analyzed the driver’s visual field in two ways: (1) via a computer-vision-based object detector for every outside view image frame of our system, and (2) via a visual segmentation engine. Figure 9 shows a boxplot of emotions over visual scene extracted features. Interestingly, compared to happy states, we see that a high number of detected cars in the visual field (high value of ‘num_cars’) is prevalent in conjunction with ‘fear’ driver states with a high degree of certainty ($p < .01$). Many cars in the visual driver scene are observed in dense traffic scenarios on the highway or in the city, often with traffic jams. Rural areas often have a lower incidence of cars. The median number of cars in happy states (2.0 - SD: 2.03) implies that fewer cars are prevalently observed in ‘happy’ states. We observe that the participants ‘disgust’ and ‘contempt’ states are in high traffic conditions, i.e., in conjunction with many cars. We note that the object detector also counts parked cars on the side of the road. However, the object detector only recognizes 2-3 parked vehicles in one frame due to occlusion, limiting visual scene object counting.

Looking at the visual segmentation features ‘segment_car’, we detect a higher degree of car scene pixels in the frame in negative emotional states. Similar to the ‘num_car’ feature, a lower percentage of car presence in the visual scene links to ‘happy’ emotional states. This observation validates previous research in driver stress recognition, which showed that high stress levels often happen in highway and city driving conditions [10]. Interestingly, the ‘segment_vegetation’ feature is not increased for happy emotional states compared to neutral states. The degree of sky presence (‘segment_sky’) for the driver visual field is highly relevant for increased for ‘sad’ emotional states, e.g., the presence of blue skies has been shown to affect personal well-being positively [58]. However, only the number of pixels associated with the sky is non-complete for defining a specific emotional state, as, e.g., rainy weather conditions in combination with the presence of the sky can induce negative emotional states [2]. As a result, the segmentation features of the outside-view can be regarded as a non-complete feature set for predicting subjective driver emotions in the wild. Further studies, including a broader range of study participants, should validate this visual scene object detection and segmentation findings.

6.4 Comparison of Recognition Performances

We perform an extensive evaluation setup to assess the feasibility of using diverse contextual, audio, and visual data streams for recognizing emotions in the wild. We compare the performance of the machine learning classifier system using different sensor stream inputs and evaluate their prediction performance. Table 2 provides the



(a)

(b)

(c)

(d)

Fig. 9. Visual segmentation and object detection features relation to expressed emotional state. We analyze the mean differences of the features between the emotional states via a one-sided t-test controlled for sample size differences. **(a)** Boxplot of 'num_cars' feature. We observe a significant higher mean number of cars of the visual field in 'fear' ($p < 0.01$) emotional states than when people felt 'happiness'. The difference in observed number of cars between 'neutral' and 'happiness' is significant. **(b)** Boxplot of 'segment_car' showing the differences in percentage of pixels in visual fields showing cars per emotion. **(c)** Boxplot of 'segment_vegetation'. **(d)** Boxplot of 'segment_sky'.

prediction performance scores for a leave-one-participant-out evaluation procedure based on our system and baseline approaches.

6.4.1 Baseline Approaches. We compare our classification system against several state-of-the-art baselines. As a common remote sensing technology for emotion recognition in-vehicle, we set facial expression recognition as a baseline. FERPlus is a facial recognition classifier (VGG16) learned on the popular FERPlus dataset, and Facial Expressions (Azure) defines the classifying system via the Microsoft Face Recognition API. Furthermore, we compare our system to VEmotion proposed by Bethge et al. [6], a machine learning classifier system based on GPS-based context and driver-related metadata.

6.4.2 Emotion Extraction from Facial Expressions. Appendix Figure 10a shows a confusion matrix giving a comparison between the output from the facial expression recognition (FER) model against the true labels (i.e. emotions expressed by the participants during the experiment). FER outputs contained all of the possible labels[16], while the true labels had all but 'sadness'. FER achieved an overall mean accuracy of 43%, most of it owning to

[16] 'anger', 'contempt', 'disgust', 'fear', 'happiness', 'neutral', 'sadness', and 'surprise'.

Table 2. Random Forest Evaluation Results. Table 2 shows averaged evaluation results for each of the feature groups in a global classifier learning leave-one-participant-out evaluation. We compute accuracy (Acc.), class-weighted precision (Prec.), unweighted average recall (UAR) and F_1 scores. Best values are indicated in bold.

	Global (Leave-One-Participant-Out)			
Feature Group	**Accuracy**	**Precision**	**UAR**	F_1
Facial Expressions (FERPlus)	.43 ± .12	.46 ± .13	.13 ± .03	.41 ± .13
Facial Expressions (Azure)	.55 ± .13	.48 ± .15	.17 ± .03	.47 ± .13
VEmotion (VE)	.52 ± .16	.38 ± .17	.26 ± .09	.42 ± .17
Visual Complexity Segmentation (VC-Seg.)	.54 ± .14	.48 ± 0.12	.18 ± .03	.44 ± 0.16
Visual Complexity - Object Detection (VC-ObjD.)	.50 ± .11	.32 ± .13	.09 ± .02	.37 ± .13
VC-Seg. + Audio	.55 ± .14	.49 ± .09	.17 ± .03	.44 ± .17
VC-ObjD. + Audio	.48 ± .12	**.50 ± .11**	.12 ± .01	.45 ± 0.13
Audio only	.37 ± .06	.49 ± .1	.10 ± .01	.40 ± .09
Audiovisual (OjbD. + Seg. + Audio)	.56 ± .14	.48 ± .15	.19 ± .03	.44 ± .17
GPS-infered features only	.57 ± .16	.44 ± .19	.29 ± .11	**.46 ± .18**
All features	**.59 ± .15**	.43 ± .17	**.29 ± .08**	.45 ± .17
Neural Network (All features)	.43 ± .14	.52 ± .15	.18 ± .06	.43 ± .14

the model predicting ‘neutral’ correctly while failing to do so in a significant way on any of the other classes. ‘Anger’, for instance, is often miss-predicted as ‘sadness’ (26%) or ‘neutral’ (64%). ‘Fear’ emotional states are never correctly predicted as such, and even predicted as ‘happiness’ in 4% of the cases.

Although our results show that facial expressions poorly predict facial expressions, we acknowledge that our used API (i.e., Microsoft Azure) and training dataset (i.e., FERPlus) are not optimized for in-vehicle use. Training a participant-dependent model about the contextual vehicle data can improve classification performance. However, this would require prior individual data collection. Other platforms, such as Affectiva[17], offer car-specific classifications but are costly to deploy. Comparing different facial expression classification platforms for in-vehicle use is a research topic for future work. In summary, off-the-shelf facial expression classification substantially over-predicts ‘neutral’ and under-predicts all other emotional states, showing a worse prediction performance driver emotions in-the-wild.

6.4.3 Global Modeling: Leave-One-Participant-Out. We evaluate the feasibility of a general classification model using all participant data except for one for training and using the last participant for evaluation. Semantically, this approach learns a model without knowing anything about the driver in advance and predicts the drivers’ emotions independent from individual context emotion preferences. In production, such a model could be trained once on a set of participants and then shipped to the customer’s vehicle without retraining. By using this

[17] www.affectiva.com

cross-validation setting, the chance of overfitting to participants or specific road properties is very low. Besides accuracy, class-weighted recall, precision, and F_1 score, we address the issue of unbalanced emotional class labels by reporting the unweighted average recall (UAR). UAR calculates the recall for each label and finds its unweighted mean.

Looking at the accuracy of our sensory system, all features combined achieve the highest average prediction accuracy of 59%. This is significantly better than using facial recognition engines alone which achieve an accuracy of 43% (FERPlus) and 55% (Azure). The difference of the all features model to VEmotion is 7 percentage points showing that incorporating additional features for global modeling is favorable. Furthermore, visual complexity segmentation features alone can achieve a high emotion recognition performance of 54%. The high performance of using only visual segmentation shows that outside-view information only by camera systems can already predict emotions for unknown drivers. This result offers the chance of using the already in-the-car integrated segmentation results of some autonomous driving control units for scene understanding to infer visual complexity and possible subjective emotions.

Interestingly, using the audio-visual complexity measures (visual object detection, visual segmentation, and in-cabin audio features) seems to increase the performance of the classifying system by only 2%, so that acquiring inside-cabin audio information does not improve results significantly. We also explored the possibility of learning participant-dependent models i.e., models that are trained only on an individual participant's data. We report the evaluation results in the Appendix Table 4.

6.4.4 Conclusion of Model Performances. In general, the hierarchy of prediction performances for specific feature sets remains constant across the evaluation settings, i.e., visual complexity only features show high recognition performances. A promising alternative is using only GPS-inferred features. We observe the highest performance of 59% to predict subjective emotions on unknown participants. Overall, a participant-independent classifier is able to predict emotions on unknown participants confidently and enables a promising alternative to e.g., facial expression detection. Furthermore, the global model can be employed as is and enables the possibility to be retrained on individual participants context preferences.

6.5 Technical Design Considerations

In-car real-time applications such as driver emotion monitoring are developed under strong computing power constraints [55]. The number of extracted features can strongly affect the computing time of the algorithm. Furthermore, features that require intensive computation and may only perform equally well as other features are often not used. An additional constraint to be respected is the degree of privacy erosion caused by each feature. With that in mind, in the next section, we discuss the importance and relevance of all features regarding their computational cost, privacy impact, and influence on model performance. We present the results of this trade-off in Table 3.

6.5.1 Computational Cost Factors. We choose two factors as the base for specifying computational complexity: (1) local computability and (2) third-party-API dependence [39]. These factors are crucial for the time delay caused by a single feature. Differences in computing time are insignificant as long as the feature can be computed locally. Equal differences in computing time are negligible as soon as the feature needs to be inferred externally. Further notable increases occur when a third-party API is required, as described in the second parameter. Also, the dependence of a third-party API negatively impacts privacy [17].

6.5.2 Privacy-Eroding Factors. We treat Stark et al.'s [53] work as a starting point defining the emotional context of information privacy. One factor to characterize the degree of privacy erosion is the type of sensors required to collect a particular feature. For example, in-cabin audio or video recording may contribute to a feeling of surveillance more than just an accelerometer or GPS tracking. Next, a sensitive topic is whether user data had to be transferred over the internet, which, e.g., can be a potential data leak and, therefore, potentially privacy erosive. Finally, another erosive privacy aspect is the need for private and/or sensitive data as defined by Zainab et al. [60]. Overall, emotion prediction is a highly personal prediction decision that should be treated with caution.

Table 3. Trading off ubiquitous feature stream importance. We show different cost computational and privacy eroding factors of features while trading them off against their influence on performance in the form of F_1 decrease.

		Feature Acquisition Factors				Privacy Factors				Prediction Importance
Context	**Feature**	**Required Sensors**	**Complex Preprocessing**	**Locally Computable**	**Third-Party API Dependent**	**Transfer of User Data over the Internet**	**Personal Data**	**Sensitive Data**	**In-Cabin Recording**	F_1 **Decrease**
Personal	*sex* *car_model* *age* *before_emotion*	–	✗	✓	✗	✗	✓	✗	✗	0.0 0.0 0.001 0.003
Session Time	*weekday* *daytime*	–	✗	✓	✗	✗	✓	✗	✗	0.0 0.001
Motion	*acceleration_x* *acceleration_y* *acceleration_z* *vemotion_acceleration*	*Accelerometer*	✗	✓	✗	✗	✓	✗	✗	0.002 0.005 0.001 0.0
GPS	*speed* *latitude* *longitude*	*GPS*	✗	✓	✗	✗	✓	✗	✗	0.025 – –
Traffic Data	*current_travel_time* *free_flow_travel_time* *current_traffic_speed* *free_flow_speed* *reduced_speed*	*GPS*	✗	✗	✓	✓	✗	✗	✗	0.0 0.0 0.001 0.008 0.0
Weather Data	*wind_speed* *precipitation_24_hours_mm* *feel_temp_outside* *cloud_cover* *weather_term*	*GPS*	✗	✗	✓	✓	✗	✗	✗	0.001 0.0 0.009 0.001 0.0
Road Data	*road_type* *max_speed* *num_lanes*	*GPS*	✗	✗	✓	✓	✗	✗	✗	0.006 0.002 0.011
Facial Expression Pred.	*facial_expression_label*	*In-Cabin Camera*	✓	✓	✗	✗	✓	✓	✓	–
Audio	*audio_amplitude* *audio_loudness*	*In-Cabin Microphone*	✓	✓	✗	✗	✓	✓	✓	0.001 0.001
Visual Complexity (Object Detection)	*buses, bicycles, cars, close_objs, far_objs, med_close_objs, motorcycles, people, trucks, stop_signs, traffic_ligts, very_far_objs, very_close_objs*	*Outside-View Camera*	✓	✓	✗	✗	✗	✗	✗	0.0, 0.0, 0.0, 0.0 0.0, 0.0 0.0, 0.0, 0.0, 0.0 0.0, 0.0 0.001
Visual Complexity (Segmentation)	*bicycle, building, bus, car, fence, motorcycle, person, pole, terrain, rider, road, sidewalk, vegetation, sky, traffic_light, truck, traffic_sign, wall*	*Outside-View Camera*	✓	✓	✗	✗	✗	✗	✗	0.0, 0.005, 0.0, 0.001, 0.0, 0.0, 0.0, 0.0, 0.0, 0.001, 0.0, 0.009, 0.001, 0.0, 0.0, 0.0, 0.002, 0.0

6.5.3 Prediction Importance. The final factor is the influence on the model performance, which is evaluated by the decrease in F_1 score if the respective feature is removed.

Based on Table 3 and the above-described evaluation parameters, use-case-oriented feature sets can be built. In the setting of a production car, the manufacturer should focus on an easy-to-compute and privacy-preserving set of features. Hence we recommend a feature set without sensitive data and where all features are preferred to be locally computable and third-party API independent. We propose a performance-oriented feature set designed to allow more privacy erosion to have higher accuracy while maintaining a low computational cost that the user can manually select. This feature set would neglect the privacy protection to improve performance and use all of our proposed features. For research purposes, we propose to use computer vision extracted features, i.e., object detection and visual scene segmentation features. First, these features are not directed at the drivers themselves and offer unobtrusive sensing of emotions benefiting driver privacy factors. Second, they show high feature importance while also offering possible local computability. Third, research on driver-view context affecting driver's well-being is underexplored [10]. Overall, we do not recommend empathic application designers to acquire emotions through facial expression analysis due to their non-robust detection and privacy-related concerns [54]. Still, many car companies employ driver facial monitoring software as driver-facing cameras are already equipped in-car, and facial expression software is easy to integrate [7, 11].

7 DISCUSSION

We propose a technical design space that extracts a large bandwidth of streams, giving information about the contextual in- and outside cabins ongoings using a consumer smartphone only. Our results show that contextual features are highly informative for recognizing the driver's emotional states in the wild. The approach inhibits several limitations and ethical considerations. The following section will discuss our findings and propose endeavors for future work.

7.1 Context and Audio-Visual Features Predict Emotions

We show that a consumer smartphone paired with machine learning modeling and computer vision can predict emotions in the wild. The capabilities and variety of sensors in our smartphones will increase in the future, and head-worn devices such as augmented reality glasses are already in development for large-scale use. This poses a challenge for future remote sensing systems, as small ubiquitous devices can infer context from little sensory information to predict emotional states. Our results show that driver emotions can be classified with up to 59% when using contextual and audio-visual features, an improvement of 7% over emotion detection using facial expressions. Our work confirms previous results using contextual data as a reliable classification input for emotions [6, 36], where adding environmental data streams (i.e., the outside and inside view) can improve the overall emotion classification performance. This conforms with previous work that showed how fast-paced changing driving situations influence the state of drivers, such as stress [10, 52]. Our results show that this concept can also be translated to emotions: environmental conditions are indicative of emotions and improve the overall classification accuracy when analyzed together with contextual data.

7.2 A Technical Framework to Prototype Empathic Car Interfaces

We separated and investigated different data streams for their influence on the overall classification performance. Including all features (i.e., contextual and audio-visual) provided the most efficient classification performance. We derived a technical design framework (see Table 2), separating the influence of the different data streams on the overall accuracy. On one side, designers and developers of empathic car interfaces can choose which data streams are available on the hardware or which data streams are necessary to achieve a particular classification performance. On the other side, users can enable or disable specific data streams to their preferences and desired classification accuracy. For example, users can opt-in for contextual data only and leave out the environmental data in case of privacy concerns. Developers and users can suit their sensing preferences according to the use case. Since the results of our study are obtained using a smartphone only, we envision that developers and designers can inexpensively prototype novel empathic car interfaces using the evaluated data streams. We are confident that our work will encourage researchers to investigate additional data streams for emotion-sensing while fostering rapid prototyping of novel empathic car interfaces.

7.3 Ethics and Privacy

We emphasize an ethical and cautious use of the context features explored in this study. Emotion-related information is highly personal and, thus, this sensitive data must be handled appropriately. The proposed approach uses the external-view camera stream, which may capture other people's information. Nevertheless, it offers a potentially more driver-privacy-preserving and discomfort-reducing alternative to the driver for measuring emotions in the wild than using facial expression or voice analysis, which requires in-cabin audio and video recordings. Filming outside, however, affects other people's privacy so that obfuscating faces are necessary, which is discussed thoroughly in related work [1, 37]. Moreover, our system can further alleviate privacy concerns by running object detection locally. This on-device run approach would make the system completely independent from an internet connection or GPS coordinates and third-party APIs, enabling broader coverage, e.g., in tunnels or remote country roads.

The analysis shows that visual features alone can predict emotions with reasonably high accuracy of 54%, outperforming facial expression analysis significantly. The robust performance provides the designer of affective in-car systems with new possibilities that do not involve cameras directed at the driver, which might raise a feeling of surveillance. Instead, our approach may only require an image representing what the driver sees. Furthermore, current driver assistance systems already obtain fine-grained outside-view information from sensors attached to the vehicle, which could directly serve as input for a potential in-cabin emotion classifying system based on visual features.

7.4 Reproducibility

We gain many insights by recording a fine-grained picture of the driver, its surroundings, and possible influences on emotion in a noisy real-world environment. Our work equips automotive user interface designers with an additional tool to design unobtrusive empathic car interfaces deployed in real-world scenarios. Furthermore, to encourage research in this area, we enable other researchers to access the data, reproduce our results and use the smartphone sensing architecture on their own by making our source code publicly available at https://github.com/msatiya/unobtrusive_driver_emotion_ds/.

7.5 Limitations

7.5.1 Emotion Annotation. Emotions are complex psycho-physiological phenomena and, as such, are difficult to study, especially in experiments in the wild. Several participants raised concerns regarding the emotion representation model, expressing that the predefined set was hard to memorize, had a priming effect, and did not allow them to truly express their emotions. Besides, they mentioned not being able to differentiate between 'contempt' and 'disgust' or, in some cases, did not even know what 'contempt' meant. This raises transparency concerns about the functionality and accuracy of AI-related classifications [32]. Furthermore, the difficulties of tracking emotions in driving contexts and, in particular, the pitfalls of using discrete emotions have been discussed thoroughly in related work [13, 62]. We recognize that this methodology is prone to noise and renders diminished nuance, but it is practically viable for in-the-wild driving contexts.

Our emotional annotation process is designed to collect ground-truth emotional labels of drivers in the wild. Due to the individual subjective nature of the expressed emotion, the label's robustness is based on trustworthiness of the participants to provide their true subjective feelings and cannot be verified by outsiders. Furthermore, the acquired emotional labels exhibit that the driver's emotions are heavily class-imbalanced. The emotion distribution has high support for neutral and happy classes, whereas there is little data support for, e.g., surprise emotions. This heavy class-tailed emotion distribution reflects the true underlying distribution of driver emotions in the wild and is not an effect of the annotation process.

7.5.2 Feature Importance Interpretations. Regarding the feature importance interpretation, we do not have sufficient data to make explicit statements about whether a specific visual feature can increase the probability of a particular emotion. Related work has shown that emotions can be assessed in various ways and a variety of factors can induce subjectively felt emotions. In our study, we used various environmental, visual, auditory, and contextual data, however, this subset is non-complete, and further efforts on, e.g., cultural aspects influencing subjective emotions can be evaluated. In our study, we could show that a high number of traffic participants in the driver's visual field are indicative of negative emotional states, however, this visual complexity assessment could be different in, e.g., India, where dense traffic is the norm.

7.5.3 Generalizability and Real-World Applicability. Our dataset contains a preliminary study of in-the-wild driver emotions. Our dataset contains multiple caveats that affect the model's generalizability: imbalanced emotion class labels, not all registered participants drove multiple sessions, and heterogeneous in-the-wild data acquisition setting. To not overfit specific participants, we decided to report the results of a leave-one-participant-out cross-validation and were able to show that the model outperforms baselines. Furthermore, the average session duration is 13.83 minutes, while the general daily usage of vehicles in the US is 27.6 minutes. Therefore the gathered dataset is acquired under realistic circumstances, but longer commute times are unavailable.

7.5.4 Sensor Data Quality. Our work presents results based on the current state-of-the-art gathering and analysis of smartphone data captured in the wild. The model's performances and sensor data quality thus should be regarded as a pillar of what is currently possible, however, some sensor data quality limitations still exist. The dataset contains features with high variability (e.g., speed), many of which cannot be controlled in an in-the-wild setting. Therefore, a more extensive dataset could also lead to better global models by covering more situations than currently represented in our dataset. This could also have a positive impact on the performance of a global model.

The back-facing camera frames present considerable variance across sessions due to different camera positioning and dashboard settings in different cars, resulting in inconsistent angles relative to the road. The visual field differences may result in uneven representations of the driving context. A camera-calibration step could be introduced in the app to add consistency to the collected data. The classification of relative distance to camera fails when a portion of an object is occluded, resulting in it being classified as further than it is. Too many distance classes create irrelevance for the less frequent ones. Some vehicles with lower incidence, like motorcycles, should be included with similar ones (e.g., buses and trucks as large vehicles). Our visual segmentation engine shows good recognition performance for subjective emotions, however, using a mounted camera inhibits several limitations. The camera frame quality (shaky video streams, low-resolution frames, and low frame rates) and occlusion due to, e.g., a truck occluding the other ongoing traffic participants affects the segmentation performance. Ongoing advancements in the smartphone camera quality render some of the issues mentioned above unimportant. Furthermore, reverse-geocoding might fail to recognize the exact road type for every geolocation due to imprecise GPS or nearer pedestrian road elements.

7.6 Future Work

Our study can be extended using a more extensive database of rides with a wider variety and distinction of emotions and more extended personal driving history. We propose a longitudinal study including more participants and longer sessions. Including a broader range of traffic scenarios while addressing the previously mentioned labeling issue by grouping emotions and adding other possible categories likely to arise in traffic contexts (e.g., stressed, confused).

To address the approach's limitations, we recommend revisiting the choice of features. Emulating relevant non-visual features would certainly help increase performance. For example, image segmentation could easily extract the number of lanes from the camera stream, with the advantage of more precision and independence than using GPS and third-party APIs. The modeling approach does not learn time-dependent information across multiple GPS traces and image frames. Future work can address this time-relationship learning of emotional context by fitting a time-aware model, e.g., recurrent neural networks.

Furthermore, several features used in the study can provide a more detailed emotion assessment. For example, advanced audio features, including pitch, frequencies, or lexical density, can provide more insights into driver emotions. However, this requires recording the voice and environmental sounds, impacting the driver's privacy. In future work, we will investigate how advanced audio features relate to emotions and how privacy-preserving emotion prediction through audio can be implemented.

7.6.1 Additional Sensor Stream Integration. Additional sensor streams can be easily integrated to infer a broader range of environmental ongoings [5]. For example, newer smartphone generations offer a light intensity sensor which could be an informative feature for explaining emotional feelings in the wild. Furthermore, we envision a non-remote sensing scenario in the future, where additional physiological information from a wearable smartwatch is connected to the smartphone. For example, physiological signals such as galvanic skin response and heart-rate-variability have shown to be a good predictor of arousal levels.

7.6.2 Additional Application Scenarios. The proposed sensor system stream can infer contextual, environmental, and visual-auditory scene understanding for various in-the-wild application scenarios. Furthermore, our easy-to-integrate smartphone app can be used in bike studies to infer emotions for bike riders, e.g., urban areas, to provide infrastructure planers feedback of bike riders' emotions. This approach can be combined with advanced immersive technologies [31] to obtain more accurate emotion assessment results.

8 CONCLUSION

This paper presents a novel technical design space using contextual and audio-visual data for unobtrusive driver emotion detection. We show that by analyzing the audio-visual complexity of the outer-car ongoings, driver emotions can be predicted with 59% accuracy in a leave-one-participant-out cross-validation using a smartphone only. In contrast, only-outside view information using the smartphone's camera stream on the road offers a recognition accuracy of 54% while providing a less driver-privacy intrusive sensing system. Our smartphone-based sensor fusion implementation is uncomplicated to integrate into other ubiquitous sensor streams with GPS or camera functionality. We make our implementation and data publicly available to foster research in this area. We encourage the research community to participate in improving on-the-road emotion classifications and discuss the ethical implications of using empathic car interfaces.

ACKNOWLEDGMENTS

This work has been partly funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A (MCML).

REFERENCES

[1] Prachi Agrawal and PJ Narayanan. 2011. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 3 (2011), 299–310. https://doi.org/10.1109/TCSVT.2011.2105551

[2] Liisi Ausmees, Anu Realo, and Jüri Allik. 2011. The Influence of the Weather on Affective Experience: An Experience Sampling Study. *Journal of Individual Differences* 32 (01 2011), 74–84. https://doi.org/10.1027/1614-0001/a000037

[3] Francisco Benita and Bige Tunçer. 2019. Exploring the effect of urban features and immediate environment on body responses. *Urban Forestry & Urban Greening* 43 (2019), 126365. https://doi.org/10.1016/j.ufug.2019.126365

[4] David Bethge, Lewis Chuang, and Tobias Grosse-Puppendahl. 2020. Analyzing Transferability of Happiness Detection via Gaze Tracking in Multimedia Applications. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 34, 3 pages. https://doi.org/10.1145/3379157.3391655

[5] David Bethge, Philipp Hallgarten, Tobias Grosse-Puppendahl, Mohamed Kari, Ralf Mikut, Albrecht Schmidt, and Ozan Özdenizci. 2022. Domain-Invariant Representation Learning from EEG with Private Encoders. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1236–1240.

[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.* Association for Computing Machinery, New York, NY, USA, 638–651. https://doi.org/10.1145/3472749.3474775

[7] Fariz Redzuan bin Monir, Rusyaizila Ramli, and Nabilah Rozzani. 2021. Driving Alert System Based on Facial Expression Recognition. In *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS)*. IEEE, 104–109. https://doi.org/10.1109/I2CACIS52118.2021.9495910

[8] Michael Braun, Jonas Schubert, Bastian Pfleging, and Florian Alt. 2019. Improving Driver Emotions with Affective Strategies. *Multimodal Technologies and Interaction* 3, 1 (March 2019), 21. https://doi.org/10.3390/mti3010021

[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 (sep 2021), 26 pages. https://doi.org/10.1145/3460938

[10] Cristina Bustos, Neska Elhaouij, Albert Sole-Ribalta, Javier Borge-Holthoefer, Àgata Lapedriza, and Rosalind Picard. 2021. Predicting Driver Self-Reported Stress by Analyzing the Road Scene. 1–8. https://doi.org/10.1109/ACII52823.2021.9597438

[11] Silvia Ceccacci, Maura Mengoni, Andrea Generosi, 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. 504–517. https://doi.org/10.1007/978-3-030-49108-6_36

[12] Monique Dittrich. 2021. Why Drivers Feel the Way they Do:An On-the-Road Study Using Self-Reports and Geo-Tagging. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '21)*. Association for Computing Machinery, New York, NY, USA, 116–125. https://doi.org/10.1145/3409118.3475130

[13] Monique Dittrich and Sebastian Zepf. 2019. Exploring the Validity of Methods to Track Emotions Behind the Wheel. 115–127. https://doi.org/10.1007/978-3-030-17287-9_10

[14] Xinyu Du, Yue Shen, Ruosong Chang, and Jinfei Ma. 2018. The exceptionists of Chinese roads: The effect of road situations and ethical positions on driver aggression. *Transportation Research Part F: Traffic Psychology and Behaviour* 58 (Oct. 2018), 719–729. https://doi.org/10.1016/j.trf.2018.07.008

[15] Paul Ekman. 1992. Are there basic emotions? *Psychological Review* 99, 3 (1992), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

[16] Paul Ekman. 1999. Basic Emotions. *Handbook of Cognition and Emotion* (1999).

[17] Benjamin Eriksson, Jonas Groth, and Andrei Sabelfeld. 2019. On the Road with Third-party Apps: Security Analysis of an In-vehicle App Platform. 64–75. https://doi.org/10.5220/0007678200640075

[18] Vérane Faure, Régis Lobjois, and Nicolas Benguigui. 2016. The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation Research Part F: Traffic Psychology and Behaviour* 40 (2016), 78–90. https://doi.org/10.1016/j.trf.2016.04.007

[19] Tara E Galovski and Edward B Blanchard. 2004. Road rage: a domain for psychological intervention? *Aggression and Violent Behavior* 9, 2 (2004), 105–127. https://doi.org/10.1016/S1359-1789(02)00118-0

[20] G. M. Hancock, P. A. Hancock, and C. M. Janelle. 2012. The impact of emotions and predominant emotion regulation technique on driving performance. *Work* 41, Supplement 1 (Jan. 2012), 3608–3611. https://doi.org/10.3233/WOR-2012-0666-3608 Publisher: IOS Press.

[21] Neska El Haouij, Jean-Michel Poggi, Sylvie Sevestre-Ghalila, Raja Ghozi, and Mériem Jaïdane. 2018. AffectiveROAD system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. 800–803. https://doi.org/10.1145/3167132.3167395

[22] 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.). Vol. 11746. Springer International Publishing, Cham, 721–742. https://doi.org/10.1007/978-3-030-29381-9_43 Series Title: Lecture Notes in Computer Science.

[23] Douglas Heaven. 2020. Why faces don't always tell the truth about feelings. *Nature* 578, 7796 (Feb. 2020), 502–504. https://doi.org/10.1038/d41586-020-00507-5 Bandiera_abtest: a Cg_type: News Feature Number: 7796 Publisher: Nature Publishing Group Subject_term: Psychology, Society, Computer science.

[24] Megan E. Hempel, Joanne E. Taylor, Martin J. Connolly, Fiona M. Alpass, and Christine V. Stephens. 2017. Scared behind the wheel: what impact does driving anxiety have on the health and well-being of young older adults? *International Psychogeriatrics* 29, 6 (June 2017), 1027–1034. https://doi.org/10.1017/S1041610216002271

[25] Myounghoon Jeon. 2016. Don't Cry While You're Driving: Sad Driving Is as Bad as Angry Driving. *International Journal of Human–Computer Interaction* 32, 10 (Oct. 2016), 777–790. https://doi.org/10.1080/10447318.2016.1198524

[26] O. Karaduman, H. Eren, H. Kurum, and M. Celenk. 2013. An effective variable selection algorithm for Aggressive/Calm Driving detection via CAN bus. In *2013 International Conference on Connected Vehicles and Expo (ICCVE)*. 586–591. https://doi.org/10.1109/ICCVE.2013.6799859

[27] Costas I. Karageorghis, Garry Kuan, William Payre, Elias Mouchlianitis, Luke W. Howard, Nick Reed, and Andrew M. Parkes. 2021. Psychological and psychophysiological effects of music intensity and lyrics on simulated urban driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 81 (2021), 329–341. https://doi.org/10.1016/j.trf.2021.05.022

[28] Mohamed Kari, Tobias Grosse-Puppendahl, Alexander Jagaciak, David Bethge, Reinhard Schütte, and Christian Holz. 2021. *SoundsRide: Affordance-Synchronized Music Mixing for In-Car Audio Augmented Reality.* Association for Computing Machinery, New York, NY, USA, 118–133. https://doi.org/10.1145/3472749.3474739

[29] Won Hee Ko, Stefano Schiavon, Hui Zhang, Lindsay T. Graham, Gail Brager, Iris Mauss, and Yu-Wen Lin. 2020. The impact of a view from a window on thermal comfort, emotion, and cognitive performance. *Building and Environment* 175 (2020), 106779. https://doi.org/10.1016/j.buildenv.2020.106779

[30] Thomas Kosch, Mariam Hassib, Robin Reutter, and Florian Alt. 2020. *Emotions on the Go: Mobile Emotion Assessment in Real-Time Using Facial Expressions.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3399715.3399928

[31] Thomas Kosch, Andrii Matviienko, Florian Müller, Jessica Bersch, Christopher Katins, Dominik Schön, and Max Mühlhäuser. 2022. NotiBike: Assessing Target Selection Techniques for Cyclist Notifications in Augmented Reality. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 197 (sep 2022), 24 pages. https://doi.org/10.1145/3546732

[32] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The Placebo Effect of Artificial Intelligence in Human-Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* (mar 2022). https://doi.org/10.1145/3529225 Just Accepted.

[33] Michael Kyte, Zaher Khatib, Patrick Shannon, and Fred Kitchener. 2000. Effect of environmental factors on free-flow speed. In *Fourth International Symposium on Highway Capacity.* Citeseer, 108–119.

[34] 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

[35] John D Lee and David L Strayer. 2004. Preface to the special section on driver distraction. *Human factors* 46, 4 (2004), 583–586. https://doi.org/10.1518/hfes.46.4.583.56811

[36] Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, and Felix Wortmann. 2021. The empathetic car: Exploring emotion inference via driver behaviour and traffic context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–34. https://doi.org/10.1145/3478078

[37] Sascha Löbner, Frédéric Tronnier, Sebastian Pape, and Kai Rannenberg. 2021. Comparison of De-Identification Techniques for Privacy Preserving Data Analysis in Vehicular Data Sharing. In *Computer Science in Cars Symposium.* 1–11. https://doi.org/10.1145/3488904.3493380

[38] Marshall Long. 2014. 3 - Human Perception and Reaction to Sound. In *Architectural Acoustics (Second Edition)* (second edition ed.), Marshall Long (Ed.). Academic Press, Boston, 81–127. https://doi.org/10.1016/B978-0-12-398258-2.00003-9

[39] Andre Luckow, Ken Kennedy, Fabian Manhardt, Emil Djerekarov, Bennie Vorster, and Amy Apon. 2015. Automotive big data: Applications, workloads and infrastructures. In *2015 IEEE International Conference on Big Data (Big Data).* IEEE, 1201–1210. https://doi.org/10.1109/BigData.2015.7363874

[40] 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.

[41] Jolieke Mesken, Marjan P Hagenzieker, Talib Rothengatter, and Dick De Waard. 2007. Frequency, determinants, and consequences of different drivers' emotions: An on-the-road study using self-reports,(observed) behaviour, and physiology. *Transportation research part F: traffic psychology and behaviour* 10, 6 (2007), 458–475. https://doi.org/10.1016/j.trf.2007.05.001

[42] Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata, and Enzo Pasquale Scilingo. 2015. Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. *IEEE Transactions on Affective Computing* 6, 4 (2015), 385–394. https://doi.org/10.1109/TAFFC.2015.2432810

[43] 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 (Aug. 2019), 1–9. https://doi.org/10.1016/j.trf.2019.07.005

[44] Michael Oehl, Felix W. Siebert, Tessa-Karina Tews, Rainer Höger, and Hans-Rüdiger Pfister. 2011. Improving Human-Machine Interaction – A Non Invasive Approach to Detect Emotions in Car Drivers. In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments (Lecture Notes in Computer Science)*, Julie A. Jacko (Ed.). Springer, Berlin, Heidelberg, 577–585. https://doi.org/10.1007/978-3-642-21616-9_65

[45] Rajesh Paleti, Naveen Eluru, and Chandra R. Bhat. 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident; Analysis and Prevention* 42, 6 (Nov. 2010), 1839–1854. https://doi.org/10.1016/j.aap.2010.05.005

[46] Pablo E. Paredes, Francisco Ordonez, Wendy Ju, and James A. Landay. 2018. Fast & Furious: Detecting Stress with a Car Steering Wheel. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3174239

[47] Emanuel Parzen. 1962. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* 33, 3 (1962), 1065 – 1076. https://doi.org/10.1214/aoms/1177704472

[48] Esther Ramdinmawii, Abhijit Mohanta, and Vinay Kumar Mittal. 2017. Emotion recognition from speech signal. In *TENCON 2017-2017 IEEE Region 10 Conference*. IEEE, 1562–1567.

[49] Alicia F Requardt, Klas Ihme, Marc Wilbrink, and Andreas Wendemuth. 2020. Towards affect-aware vehicles for increasing safety and comfort: recognising driver emotions from audio recordings in a realistic driving study. *IET Intelligent Transport Systems* 14, 10 (2020), 1265–1277. https://doi.org/10.1049/iet-its.2019.0732

[50] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714 Place: US Publisher: American Psychological Association.

[51] B. Scott-Parker. 2017. Emotions, behaviour, and the adolescent driver: A literature review. *Transportation Research Part F: Traffic Psychology and Behaviour* 50 (2017), 1–37. https://doi.org/10.1016/j.trf.2017.06.019

[52] Maria Seitz, Thomas J. Daun, Andreas Zimmermann, and Markus Lienkamp. 2013. Measurement of Electrodermal Activity to Evaluate the Impact of Environmental Complexity on Driver Workload. In *Proceedings of the FISITA 2012 World Automotive Congress*. Springer Berlin Heidelberg, Berlin, Heidelberg, 245–256. https://doi.org/10.1007/978-3-642-33838-0_22

[53] Luke Stark. 2016. The emotional context of information privacy. *The Information Society* 32 (01 2016), 14–27. https://doi.org/10.1080/01972243.2015.1107167

[54] Luke Stark. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 50–55. https://doi.org/10.1145/3313129

[55] Rainer Steffen, Richard Bogenberger, Joachim Hillebrand, Wolfgang Hintermaier, Andreas Winckler, and Mehrnoush Rahmani. 2008. Design and realization of an ip-based in-car network architecture. In *Proceedings of the First Annual International Symposium on Vehicular Computing Systems (ISVCS 2008)*. https://doi.org/10.4108/ICST.ISVCS2008.3543

[56] Alejandro A. Torres-García, Omar Mendoza-Montoya, Marta Molinas, Javier M. Antelis, Luis A. Moctezuma, and Tonatiuh Hernández-Del-Toro. 2022. Chapter 4 - Pre-processing and feature extraction. In *Biosignal Processing and Classification Using Computational Learning and Intelligence*, Alejandro A. Torres-García, Carlos A. Reyes-García, Luis Villaseñor-Pineda, and Omar Mendoza-Montoya (Eds.). Academic Press, 59–91. https://doi.org/10.1016/B978-0-12-820125-1.00014-2

[57] Xiaoyuan Wang, Yongqing Guo, Jeff Ban, Qing Xu, Cheng-Lin Bai, and Shanliang Liu. 2020. Driver Emotion Recognition of Multiple-ECG Feature Fusion based on BP Network and D-S Evidence. *IET Intelligent Transport Systems* 14 (03 2020). https://doi.org/10.1049/iet-its.2019.0499

[58] Stephen Westland, Yuan Li, Dabo Guan, Yanni Yu, P Wang, Xuejun Wang, Kebin He, Shu Tao, and Jing Meng. 2019. A psychophysical measurement on subjective well-being and air pollution. *Nature Communications* 10 (11 2019). https://doi.org/10.1038/s41467-019-13459-w

[59] Changsheng Xu, Namunu C Maddage, Xi Shao, Fang Cao, and Qi Tian. 2003. Musical genre classification using support vector machines. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 5. IEEE, V–429.

[60] Syeda Sana e Zainab and Tahar Kechadi. 2019. Sensitive and Private Data Analysis: A Systematic Review. In *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems* (Paris, France) *(ICFNDS '19)*. Association for Computing Machinery, New York, NY, USA, Article 12, 11 pages. https://doi.org/10.1145/3341325.3342002

[61] Sebastian Zepf, Monique Dittrich, Javier Hernandez, and Alexander Schmitt. 2019. Towards Empathetic Car Interfaces: Emotional Triggers while Driving. *CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312883

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

A APPENDIX

A.1 Confusion Matrices

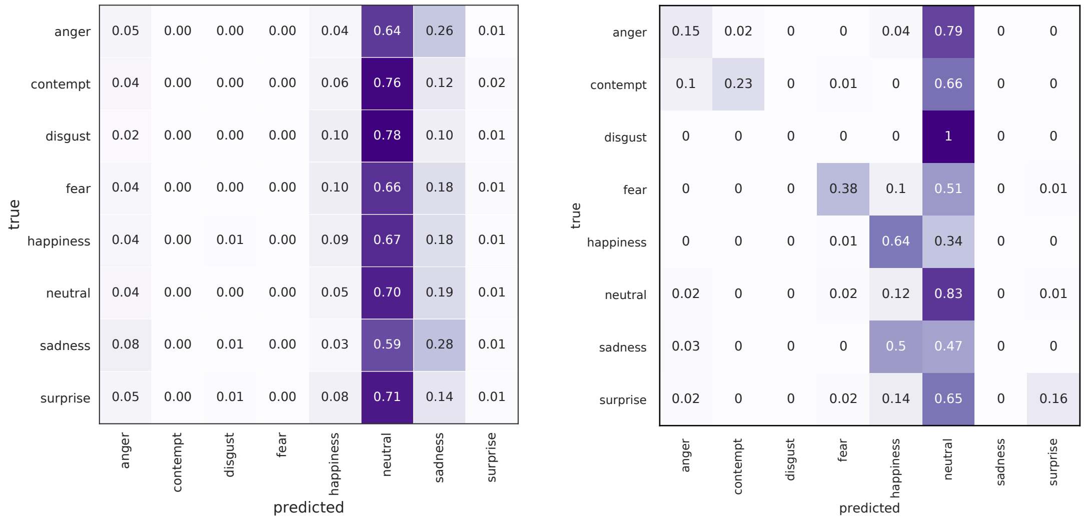
The confusion matrices of the facial expression baseline and the Random Forest predictor is shown in Fig 10.

A.2 Input Feature Correlation

The Pearson correlation of the input features of our system is shown in Figure 11.

A.3 Neural Network Architecture

The specification of the used neural network architecture is explained in the following section. We employ a feedforward fully-connected neural network, with two hidden layers and one output layer. The first layer contains 100 neurons, the second layer contains 50 neurons, and both use a relu activation function. The output layer



true \ predicted	anger	contempt	disgust	fear	happiness	neutral	sadness	surprise
anger	0.05	0.00	0.00	0.00	0.04	0.64	0.26	0.01
contempt	0.04	0.00	0.00	0.00	0.06	0.76	0.12	0.02
disgust	0.02	0.00	0.00	0.00	0.10	0.78	0.10	0.01
fear	0.04	0.00	0.00	0.00	0.10	0.66	0.18	0.01
happiness	0.04	0.00	0.01	0.00	0.09	0.67	0.18	0.01
neutral	0.04	0.00	0.00	0.00	0.05	0.70	0.19	0.01
sadness	0.08	0.00	0.01	0.00	0.03	0.59	0.28	0.01
surprise	0.05	0.00	0.01	0.00	0.08	0.71	0.14	0.01

(a) Confusion matrix comparing facial expression recognition output (FERPlus model) against true labels.

true \ predicted	anger	contempt	disgust	fear	happiness	neutral	sadness	surprise
anger	0.15	0.02	0	0	0.04	0.79	0	0
contempt	0.1	0.23	0	0.01	0	0.66	0	0
disgust	0	0	0	0	0	1	0	0
fear	0	0	0	0.38	0.1	0.51	0	0.01
happiness	0	0	0	0.01	0.64	0.34	0	0
neutral	0.02	0	0	0.02	0.12	0.83	0	0.01
sadness	0.03	0	0	0	0.5	0.47	0	0
surprise	0.02	0	0	0.02	0.14	0.65	0	0.16

(b) Confusion matrix comparing the results of the Random Forest prediction model against true labels.

Fig. 10. Confusion matrices of the facial expression baseline (a) and (b) the Random Forest trained on the participants multi-domain contextual data. The values in the matrix are normalized on the true emotion class occurrences. (a) Only 5% of 'anger' emotional states are recognized by the facial expression classifier and 64% of all true 'anger' emotions are falsely predicted as being 'neutral'. (b) The model is participant-dependent trained model. Although, the model overpredict 'neutral' states, the random model performs significantly better in predicting 'fear' (38%), 'happiness' (64%), and 'neutral' (83%) states.

outputs one-hot-encoded emotion labels using a sigmoid function. The network is trained with a batch size of 64 for 3000 epochs using the adam optimizer for backpropagation. We employ early stopping criteria from avoiding overfitting after a waiting period of 50 epochs. To counterbalance the imbalance of classes, we assign a loss weight according to the inverse frequency of class observation to the categorical cross-entropy loss optimization function. We report the neural network performance in Table 2.

A.4 Person-Dependent Modeling

We analyzed participant-dependent modeling using a participant-dependent Leave-One-of-10-Road-Segments-Out cross-validation. This setting denotes that we are training a participant-dependent model and validating the participant's holdout set using a 10-fold cross-validation scheme. In general, participant-dependent models can adapt to specific persons and provide a possibly more privacy-aware and personal emotion predictor as data is not shared globally. However, each person-dependent model has only limited training data available, so longer drive durations are needed to reach a satisfactory recognition performance. We acquired multiple sessions for some participants to circumvent the issue of having too little data on individual participants. We did not employ a leave-one-session cross-validation as not every participant acquired driving data in multiple sessions. From the last four rows in the emotion recognition performance table, we see that the overall recognition performance of the models using our features is best. The combined classification model with all input features reaches an accuracy of 66% and an F_1 score of 67%. Therefore this model is significantly better than the baselines and the

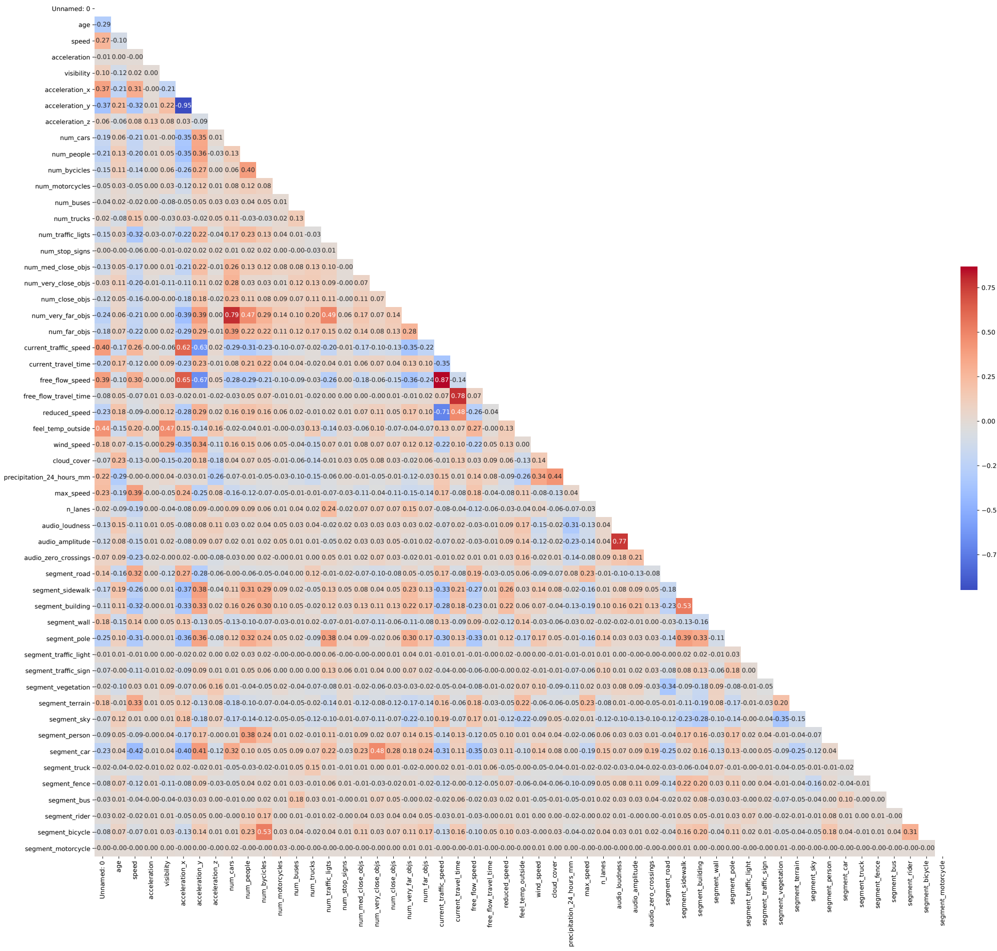


Fig. 11. Correlation matrix of the available features of our system. High positive correlations are depicted in red and high negative correlations are shown in deep blue. The feature 'segment_train' has no variance and is therefore left blank, since the segmentation module has not detected any trains in the in-the-wild driving.

performances of global modeling procedures. The audio-visual feature set predicts emotions confidently with 62% accuracy (F_1: 62%), whereas GPS-sensor-only extracted features are able to predict subjective emotions with 65% accuracy (F_1: 57%).

Table 4. Random Forest Evaluation Results. We report the averaged evaluation results for each of the feature groups across all evaluation steps: global classifier learning leave-one-participant-out evaluation and learning participant-dependent models. Accuracy, class-weighted precision, unweighted average recall (UAR) and F_1 scores. Values are averages from the 10-fold cross-validation (best values are indicated in bold).

	Participant-Dependent (Leave-One-Road-Segment-Out)			
	Accuracy	**Precision**	**UAR**	F_1
Facial Expressions (FERPlus)	.33 ± .0	.41 ± .09	.23 ± .04	.36 ± .09
Facial Expressions (Azure)	.39 ± .04	.42 ± .04	.27 ± .0	.38 ± .03
VEmotion (VE)	.65 ± .04	.63 ± .04	.59 ± .01	.64 ± .03
Visual Complexity Segmentation (VC-Seg.)	.6 ± .04	.50 ± .04	.38 ± .01	.62 ± .03
Visual Complexity - Object Detection (VC-ObjD.)	.51 ± .06	.51 ± .06	.27 ± .01	.56 ± .03
VC-Seg. + Audio	.61 ± .04	.51 ± .04	.39 ± .01	.61 ± .03
VC-ObjD. + Audio	.58 ± .04	.42 ± .04	.32 ± .01	.61 ± .02
Audio only	.52 ± .02	.39 ± .02	.28 ± .01	.57 ± .02
Audiovisual (OjbD. + Seg. + Audio)	.62 ± .03	.52 ± .03	.43 ± .01	.62 ± .02
GPS-infered features only	.65 ± .04	.63 ± .04	.57 ± .11	.57 ± .01
All features	**.66 ± .03**	**.65 ± .03**	**.6 ± .01**	**.67 ± .03**

HappyRouting: Learning Emotion-Aware Route Trajectories for Scalable In-The-Wild Navigation

DAVID BETHGE*, Porsche AG, LMU Munich, Germany
DANIEL BULANDA, GrapeUp, Poland
ADAM KOZLOWSKI, GrapeUp, Poland
THOMAS KOSCH, HU Berlin, Germany
ALBRECHT SCHMIDT, LMU Munich, Germany
TOBIAS GROSSE-PUPPENDAHL, Porsche AG, Germany

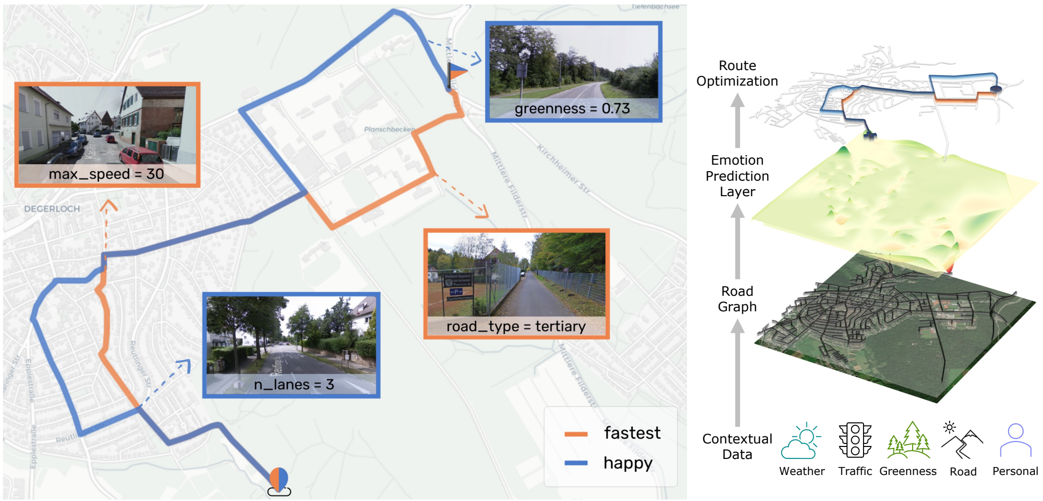


Fig. 1. We present HappyRouting, a new navigation system able to route after positive emotions. We predict emotional weights for every road coordinate based on environmental, personal, and dynamic road context and find the optimal driving trajectory.

Routes represent an integral part of triggering emotions in drivers. Many navigation systems allow users to choose a navigation strategy, such as the fastest or shortest route. However, they do not consider the driver's emotional well-being. We present HappyRouting, a novel navigation-based empathic car interface guiding drivers through real-world traffic while evoking positive emotions. We propose a set of design considerations, derive a technical architecture, and implement an optimization framework. Our contribution is a

Authors' addresses: David Bethge, Porsche AG, LMU Munich, Stuttgart, Germany, david.bethge@ifi.lmu.de; Daniel Bulanda, GrapeUp, Warsaw, Poland; Adam Kozlowski, GrapeUp, Warsaw, Poland; Thomas Kosch, HU Berlin, Berlin, Germany; Albrecht Schmidt, LMU Munich, Munich, Germany; Tobias Grosse-Puppendahl, Porsche AG, Stuttgart, Germany.

Manuscript submitted to ACM

machine learning-based generated emotion map layer, predicting emotions along a route based on static and dynamic contextual data. We developed HappyRouting, a real-time mobile navigation app to predict routes evoking positive emotions interactively. We evaluated HappyRouting in a real-world driving study ($N = 13$), finding that happy routes increase subjectively perceived valence by 11% ($p = .007$). Finally, we show how emotion-based routing can be integrated into common navigation apps, promoting emotional well-being for general mobility use.

CCS Concepts: • **Human-centered computing** → *Interactive systems and tools*; *HCI theory, concepts and models*; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Empathic Interfaces, Affective Computing, Navigation, Machine Learning, Contextual-Aware Computing

ACM Reference Format:
David Bethge, Daniel Bulanda, Adam Kozlowski, Thomas Kosch, Albrecht Schmidt, and Tobias Grosse-Puppendahl. 2023. HappyRouting: Learning Emotion-Aware Route Trajectories for Scalable In-The-Wild Navigation. 1, 1 (January 2023), 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Today's car navigation systems allow users to navigate according to various objectives, such as the fastest route, the shortest distance, or routes that require the lowest energy consumption [24, 59]. In contrast to these routing modalities, we investigate a new objective by optimizing routes for positive emotions. Emotions play an important role in driving [25], as certain states of arousal and valence can lead to more thoughtful decisions. In contrast, exaggerated states can significantly increase the driver's willingness to take risks and thus endanger the safety of all road users [16, 44]. Subsequently, we propose HappyRouting, a system navigating drivers through routes that elicit positive (i.e., happy) emotions.

While the vision, preferences, and design of empathic navigation have been presented in prior work [45], its technical concept, implementation, and concrete evaluation have rarely been the subject of research. In particular, the field of in-vehicle emotion assessment [6, 36] has evolved strongly over the past decade, while empathic real-world applications remain the exception [8, 61]. Based on an increasing number of available datasets that classify driver emotions based on driving context [3, 6, 36], we conceptualize and implement the missing building blocks for an end-to-end empathic navigation interface. Consequently, HappyRouting predicts possible emotions for thousands of unseen roads throughout a road graph and optimizes for the best tradeoff between positive emotions and travel time.

In this paper, we present design considerations, the resulting architecture, and an experienceable implementation for driving with positive emotions in real-world environments. We begin by discovering the degrees of freedom to design a scalable affective navigation system applicable to unknown users, environments, and roads. We demonstrate that theoretical psychological assumptions hold for the experienceable system, showing for the first time a navigation system that regulates emotions positively. Based on this, we derive the technical architecture for HappyRouting. An in-the-wild driving study with 13 participants investigates the effect on arousal and valence between choosing the *fastest* route, and the predicted *happier* route. Our results show a significant effect in perceived valence between the fast and happy route, showing that the happy route selected by HappyRouting leads to an improvement in valence. Furthermore, our participants were willing to use HappyRouting although positive routes consumed more time. Moreover, we conducted a simulation study in a whole region to compare the differences between the optimization objectives. Finally, we conclude our work by discussing ethics, the applicability of HappyRouting for other transport modalities, generalization for unseen roads, limitations, and future work.

CONTRIBUTION STATEMENT

The contribution of this work is threefold:

(1) We present the degrees of freedom to design a scalable affective navigation system that is applicable to unknown users and unseen environments.
(2) With HappyRouting, we demonstrate that guiding design decisions hold for an experienceable end-to-end system and show for the first time that a navigation system can regulate emotions positively.
(3) We characterize the qualitative and quantitative properties of our proposed affective navigation system in an in-the-wild user study ($N = 13$), as well as with detailed simulations.

2 RELATED WORK

HappyRouting's idea of routing after positive emotions builds on concepts found in driver emotion assessment, contextual computing, and empathic car interfaces.

2.1 Inferring Driver Emotions

Empathic car interfaces benefit from understanding the driver's emotions to adapt their interface, contributing positively to the user's emotional state [55, 61]. Emotion assessment can be achieved through *direct* and *indirect* user observation.

Direct observation methods, such as recognizing facial expressions [17, 19], are a convenient method to infer emotions while driving. Although facial expressions are a commonly used modality [10], it remains controversial in research [28, 40]. Alternatively, emotions can be derived from psychophysiological signals such as electrodermal activity, heart rate, muscle tension, respiratory rate, and electroencephalography [3, 54]. The setup of in-car physiological sensing is often problematic due to insufficient signal quality levels [15] and missing user acceptance [58].

Indirect user observation through analyzing contextual driving data has gained increasing attention for emotion recognition. Zepf et al. [60] surveyed affective automotive user interfaces and identified several factors causing emotional triggers and changes, including driving behavior, music, and road conditions. This fact was exploited by Liu et al. by analyzing vehicle CAN-bus data [36], reaching subject-independent F1-scores of 59%. Bethge et al. [6] showed that contextual driving data captured with a smartphone result in subject-independent F1-scores of 56%, an improvement over using facial expressions as a baseline.

An empathic navigation system poses additional constraints on the observation method since it is required to predict emotions on thousands of possible road segments to find the optimal emotion-aware route. Since, in most cases, there is no direct observation input (e.g., crowd-sourced facial expressions) accessible for every unseen road segment, algorithms trained on remotely accessible observations (e.g., traffic, road properties or weather) are needed.

2.2 Affective Routing

Routing is considered to be a factor that strongly influences the driver's emotions. In their detailed study, Braun et al. [8] explored 20 concepts for empathic car interfaces, finding that empathic navigation represents the highest demand among German and Chinese users. Pfleging et al. [45] evaluated the general idea of experience-based navigation in a web survey and identified the fastest route and the route with the least stress as the most important factors for route selection. At the same time, users often bypass the fastest route, for example, to avoid stressful situations and negative emotions [12]. Zepf et al. [60] showed that the majority of positive emotional triggers are associated with the environment. Accordingly, positive and negative experiences with a route play a crucial role for the acceptance of future route recommendations [52]. Previous work has focused on various routing concepts that may indirectly influence emotion. This is in contrast to HappyRouting, which directly optimizes for positive emotions by applying a diverse set of features.

Quercia et al. [48] investigate a scenic routing concept using crowd-sourced images associated with POIs. Similarly, Runge et al. [50] identify scenic rides by applying a pre-trained neural network to street view imagery. Using physiological data, Tavakoli et al. [56] introduce a framework for routing recommendations based on the driver's heart rate collected in a three-month in-the-wild study. The authors also note that the proposed framework is capable of finding infrastructural elements in a route that can potentially affect a driver's well-being. Hernandez et al. [29] proposed the long-term vision of crowd-sourced driver stress detection [42] using "Empathetic GPS" - a vision of a navigation system that geographically identifies routes minimizing stress whilst taking the driver to a given destination.

2.3 Summary

Previous work shows that empathic navigation is a highly desired feature among drivers and co-drivers [8, 45]. Our work leverages such initial concepts and contributes with the technical building blocks to ultimately present HappyRouting, a real-world, end-to-end affective navigation system. To the best of our knowledge, HappyRouting is the first experienceable system that predicts emotions for thousands of possible routes on a map and optimizes the route for the best tradeoff between positive emotions and travel time.

3 DESIGN CONSIDERATIONS

The following section describes our design consideration for HappyRouting. We start by describing how HappyRouting is going to affect the driver's emotions, mood, and well-being. Then, we look into different routing concepts and conclude with relevant objectives as well as the modeling of driver routes. A particular focus on ethics and limitations can be found in our discussion in Section 7.

3.1 User Emotions, Mood & Wellbeing

Our goal is to create a joyful driving experience that is implicitly composed of contextual data such as traffic, road characteristics, and weather. In general, our approach can be considered as a method for regulating emotions [38] during navigation, in particular aiming for an up-regulation. Emotions can be regarded as situationally bound, limited in time, with either a positive or a negative state [38]. This applies, for example, to traffic flow or route characteristics, which are among the primary sources of information for HappyRouting.

In contrast to emotions, the user's mood is less intense and specific and often not caused by a particular event or situation [21], such as the current weather [33]. HappyRouting primarily aims to elicit positive emotions, eventually leading to a positive influence on the user's mood. However, this approach is deliberately oversimplified as it is necessary to consider the overall process of mood adjustment and counter-hedonistic effects [34]. This means that positive mood is not only established by a simple aggregation of positive emotions but rather a complex interplay of positive *and* negative emotions (e.g., people like to listen to sad music to adjust their mood positively) [41].

Another constraint of our approach is that we focus on primary emotions and, in particular, on positive affect. However, positive affect and the absence of negative affect represent only a subset of possible dimensions to improve subjective well-being [22]. Other important factors, particularly in the dimensions of social well-being and eudaimonic well-being (e.g., self-acceptance), are currently well outside our scope of work. In summary, HappyRouting can be seen as a first important step towards a more detailed understanding of how technical systems can have a positive impact on emotions. On the other hand, the aforementioned limitations raise many important questions for future work.

3.2 Routing Concepts

All routing concepts have in common that they operate on a graph consisting of nodes and edges with associated weights. Edges represent a road segment in the overall road network, while nodes connect the various road segments. The weights associated with a road segment can represent different optimization objectives, for example, routes with the fine particulate (PM2.5) intake [37] or those requiring least energy [59]. For most use-cases, the primary optimization objective is tightly coupled with travel time and distance between the two nodes. Eventually, this fact leads to the need for multi-objective optimizations, achievable through single-stage optimization and multi-stage optimization.

Single-stage optimization combines multiple optimization objectives into one optimization method. For this purpose, multiple weights corresponding to the different objectives are associated with each edge, sometimes referred to as layers. In the most simple case, the final weight can then be determined by a weighted addition of the individual weights [57] or the introduction of a penalty factor [31]. If multiple objectives have statistical dependencies, more complex models like Bayesian Belief Networks can be used to determine the combined weight [53].

Multi-Stage Optimization conducts multiple optimizations in succeeding steps, with the first steps representing the most important optimization objectives. This optimization procedure can be used if the optimization problem is expressed through multiple models, e.g., road graphs and lists of POIs. For example, Quercia et al. [48] apply Eppstein's algorithm [20] to find the N shortest paths, and then, in a second stage, rank those paths by user scores for POIs.

For HappyRouting, we apply single-stage optimization, as we can associate emotions to each road segment, enabling us to express the problem in a uniform way. HappyRouting's primary objective is *travel time*, while emotions are added to the graph's weights as a penalty term [31]. The penalty term is computed through a machine-learning model, which takes various emotion-related features into account. Routes can be computed with efficient graph-based algorithms like Dijkstra or A*, or in our case, the contraction hierarchies algorithm specifically designed for vehicle navigation optimization [23].

3.3 Optimization Objectives

Considering human wayfinding, Golledge [24] ranked various route selection criteria. *Shortest distance* ranked first and *least time* second, followed by *fewest turns* and *most scenic*. Less generic approaches consider criteria like *least energy* [59], *least fine particulate (PM2.5) intake* [37], *optimal physical exercise* [53], or *personalized accessibility metrics* [32, 57]. We can categorize these criteria into the following optimization objectives:

- *Environment-dependent* objectives, e.g., *shortest distance* do not change over the duration of the trip. In HappyRouting, we utilize properties of the environment like the number of lanes and speed limits to derive emotions.
- *Time-dependent* objectives change over the duration of the trip, such as *least time* would be affected by time-dependent traffic. Our primary optimization objective in HappyRouting is the travel time, which in turn is penalized by negative or neutral emotions.
- *User-dependent* objectives depend on personal criteria, such as accessibility needs. HappyRouting attempts to scale across a variety of users, including unknown ones. Therefore, we include some user-dependent features, i.e., personal context as input in our architecture but identify a need for further exploration in future work.

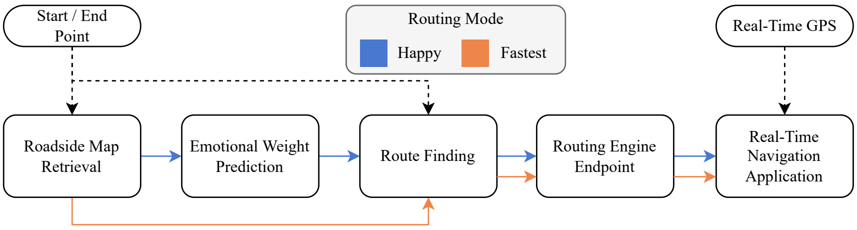


Fig. 2. Architecture of happy navigation computation.

- *State-dependent* objectives consider the state of a object, such as an electric vehicle's charging state [59]. For practical reasons and generalization purposes, we do not consider this in our objectives.

Different optimization objectives raise the question of their societal impact, particularly when applied at large-scale. Johnson et al. discuss such potentially negative influence of scenic routing algorithms and their optimization objectives on neighborhoods and parks [30]. We refer to Section 7 for a more detailed discussion.

3.4 Modeling and Simulation

Most sophisticated optimization objectives require an approach to express their influence on the weights of a graph.

The most common form is modeling based on historical observations, especially of travel times [46] or least fine particulate (PM2.5) intake [37]. However, the two examples differ greatly in how they can be applied to a graph network. In the case of travel times, observations can directly be linked to edges in the graph. For fine particulate (PM2.5), an intermediate interpolation and edge association step is needed, as observations are linked to measurement stations [37]. HappyRouting applies both methods for different features: On the one hand, the characteristics of the road segments are used as direct parameters, and on the other hand, metrics of the surrounding landscape such as the green index are interpolated.

Models that take travel time into account require time-dependent modeling, as traffic and therefore weights in the graph change over the duration of the trip. Such look-ahead models are often based on historical observations. Except for *travel time*, HappyRouting currently does not consider additional fast-changing environmental parameters. In particular, the weather will be considered static throughout the trip. This design choice reflects in lowered computational effort at the cost of potentially less accurate predictions in the future.

In many use-cases like travel time prediction [46], the penalty term for each road segment can be represented as a regression model. In contrast, HappyRouting is based on a multi-class model for predicting emotions (e.g., happy, neutral, sad), where the inputs consist of road parameters and the outputs represent the pseudo-likelihood of each class. To synthesize the penalty term for the graph edges, we decided to use only the pseudo-likelihood of the class *happy*. Alternative methods comprise representing the model as a binary classifier (e.g., happy against all other classes) at the cost of decreased performance due to increased imbalance of classes.

4 ARCHITECTURE

In the following section, we describe the architecture of our system and the necessary steps to provide the user with an emotion-optimized route. We derive the technical considerations from the concept design considerations presented in Section 3.

Table 1. List of available features to predict drivers emotions.

Context	Feature	Example	Description	Source
weather	feeltemp_outside	13.0	temperature outside of car	Azure Weather
	windspeed	5.6	windspeed in km/h	
	cloud_coverage	76	relative cloud coverage in %	
	weather_term	‘clear’	description of weather condition	
traffic	reducedspeed	7.295495	current reduced speed to freeflow speed	Azure Traffic
	freeflow_speed	115.0	freeflow speed expected under ideal conditions	
road	road_type	‘residential’	road type of current position	OpenStreetMap
	max_speed	120.0	maximum allowed speed on the road	
	n_lanes	2	number of available lanes	
greeness	satellite_greeness	0.2	percentage of green pixels in environment	Mapbox Satellite
time	daytime	‘afternoon’	current daytime	system input
personal	age	21	age of the driver	user input
	before_emotion	‘happiness’	subjective expressed emotion before driving	

4.1 Problem Statement

Finding an emotional-relevant route is complex due to several reasons. The optimization of the route must be executed in near real-time, and all information required for the routing algorithm must be available (see Section 3).

Given a user’s starting point a and selected destination b as GPS coordinates, we search for a route that likely makes the user happy. The following requirements must be fulfilled by HappyRouting:

Req 1: The emotional component of the route is subject to context, person and traffic characteristics
Req 2: The happiness weight of road segments has to be assigned before starting the navigation
Req 3: The system should be usable like a common smartphone navigation system
(a) The system should enable destination search functionality (e.g., finding a train station)
(b) The system should re-locate given the smartphones geolocation and show the trajectory of the happy route
(c) The system should output turn-by-turn navigation instructions to the driver in real-time
Req 4: The navigation engine should be designed as a scalable system
(a) Provide happy routes in every geolocation (no pre-annotated or historic routes)
(b) Optimize the route trajectories without delay so that the user receives the route recommendation $< 2s$ after entering the destination

4.2 General Framework

Figure 2 provides an overview of the system architecture. Depending on the start to end point, a roadside map is created via OpenStreetMap (OSM)[1]. We perform a custom map layer computation in the subsequent step in which we predict emotional weights for every edge in the road graph. The happy route is then found via an optimization procedure with the newly created map. We expose the endpoint of the navigation engine and build a real-time navigation smartphone app on the basis of the routing engine.

[1] https://www.openstreetmap.org

4.3 Input Features

We used a reduced number of contextual road features of the original dataset [6] for our custom context-emotion classifier model. The selected features are based on Braun et al. work [11] where driving behavior, traffic, vehicle performance, and environmental factors are found to be discriminative of emotions. We filtered the variables based on the following requirements: (1) real-time, on-device computation without accessing the vehicle itself, (2) no direct user interaction, (3) non-critical consumption of device resources, and (4) time-critical computability. We restrict the model to only those input features that can be pre-computed before driving (**Req 2**)[2]. Furthermore, personal factors such as age are used to adapt to user-dependent emotion-route preferences. The selected features are shown in Table 1.

We compute the weather and traffic features for every road segment using Microsoft Azure's Weather and Maps API. The road type features are gathered from OSM by selecting the nearest OSM node with its corresponding parameters. Based on satellite imagery, we quantify the vegetation and determine the green index [13] at any given geolocation. We compute the curviness using a weighted measure of the length of curves, which depends on a radius of a circumscribed circle that passes through all three consecutive geocoordinates in a route. Given a, b, c as the length of the three sides of a triangle, the radius of the circumcircle is given by the formula:

$$r = \frac{abc}{\sqrt{(a+b+c)(b+c-a)(c+a-b)(a+b-c)}} \tag{1}$$

4.4 Emotion Prediction

The foundation for our emotional routing is a computational behavior model for predicting emotion using road context. We thereby achieve to learn subject-independent emotion labels for previously unseen road segments (**Req 4**). Recently, Bethge et al. [6] proposed an in-car remote-sensing system able to predict emotions on unknown roads for unknown users with very high confidence. The model is able to predict discrete emotion categories ('happy', 'sad', 'neutral', 'angry', 'contempt', 'disgust', 'fear', 'surprise') using contextual road information (**Req 1**). Although many affective representation models exist (e.g., Plutchik's wheel of emotions describing 56 emotions [47] or Russel's circumplex model [51]), we have selected the seven emotion categories, as well as the category neutral [3]. The choice of our set of discrete emotions is practically grounded in Ekman's theory which is often associated with emotion detection by the analysis of facial features. We exploit this well-known model for our optimization and build a bridge to previous work [14, 61].

In their in-the-wild driving study, the authors collected contextual driving data and subjective emotional states expressed by drivers while driving [6]. To not distract the driver and bias the ground-truth labeling, a beep tone every 60 seconds was triggered for the driver to verbally express their felt emotion according to a predefined set. We acquired the dataset and extended it by another 14 participants to 26 participants in total, reflecting in 31 sessions consisting of 438 min of emotion-labeled driving and eight classes of emotions in total.

After defining the input features, we selected a Random Forest Ensemble Learning as classifier based on a 10-fold grid-search cross-validation (using Support Vector Machines, Feedforward Neural Network, Decision Tree, Adaboost, and Random Forest classifier from scikit-learn with hyperparameter optimized parameters) in which the Random Forest achieved the highest average F1 score. The prediction model [4] is tested via a leave-one-subject-out cross-validation on

[2] Contextual variables such as the current acceleration cannot be pre-computed.

[3] Our model is designed to predict multiple emotions to ensure adaptability for navigation use-cases where other emotions predictions are needed, rather than simplifying it to a binary classification setting for just predicting 'happiness'.

[4] model parameter: class_weight = 'balanced', max_features= log_2, n_estimators= 50.

unseen participants. Compared to deep neural network architectures, our Random Forest model has the advantage of being relatively easily deployable on-device, and its inference time is short. The results are outlined in Table 2. Overall, our model is able to achieve a mean emotion recognition accuracy of 63% with a balanced F_1 score of 53%[5]. These results are slightly inferior to current subject-independent contextual emotion classifiers [6, 36], but are also based on a remotely acquirable, and thus much reduced, feature set. As a baseline in our dataset, we recorded a driver-facing camera stream and applied a FERPlus-trained classifier [5], showing that the collected contextual features still outperform facial expressions [6].

Table 2. Mean (standard deviation) accuracy, class-weighted precision, recall, and F_1 scores of the cross-validation on unseen participants i.e., leave-one-participant-out cross-validation. The model predicts 8 emotion classes in total.

	Leave-One-Participant-Out Cross-Validation			
Input	Accuracy	Precision	Recall	F_1
Facial Expr. [5]	.55 ± .18	**.53** ± .19	.55 ± .18	.49 ± .19
Our model	**.63** ± .16	.49 ± .21	**.63** ± .16	**.53** ± .20

4.5 Routing Map and Navigation

Having defined the predictive model required to simulate emotions based on contextual information collected remotely, we now present the system required to provide the user with a route optimized for emotions. In Figure 3 we display how a happy path may differ from the fastest one based on a custom emotion map layer.

Routing Map Generation. We define a custom emotion map layer that contains predicted emotions and optimize the route thereafter. Given a road graph G with vertices V and edges E, we predict emotion weights for every driveable segment E. We then apply the contraction hierarchies algorithm [23] to the road graph by optimizing for the following equation with the user's start point a and end point b:

$$route(a, b) = min \sum_{i,j \in [a,b]; i \neq j \in E} \frac{d(i, j)}{\lambda * e(i, j) * c(e(i, j))} \tag{2}$$

In contrast to the fastest route, our optimizer minimizes the sum of the travel-time of each edge $d(i, j)$ and penalizes its decision by the happiness weighing factor λ and its corresponding predicted happiness value $e(i, j)$, multiplied by the confidence of the individual happiness prediction $c(e(i, j))$. Here, the last part ensures that it is favorable for the optimizer to choose edges with high predicted happiness values[6]. In our simulation study (see Section 6), we found that a happiness weighing factor of $\lambda = 20$ yields a good tradeoff between travel time and positive emotions.

Optimization Backend. To implement the optimization procedure, we use the open-source, Java-based framework GraphHopper[7]. GrapHopper offers a fast and memory-efficient routing engine including a web-frontend and a standalone web-server to calculate the distance, time, turn-by-turn instructions, and trajectory properties for a route. We adopt

[5] Neutral emotions represent the majority class of our dataset, while happy emotions are at 23%, being predicted second best (after neutral) in terms of precision/recall

[6] We opt the routing decision formula to be influenced by the predicted emotion value in the denominator as the travel times have no equal lengths and a regularizing longer route segments (high $d(i, j)$) with the emotion scaling is more beneficial than e.g., substracting the emotion values.

[7] https://github.com/graphhopper/graphhopper

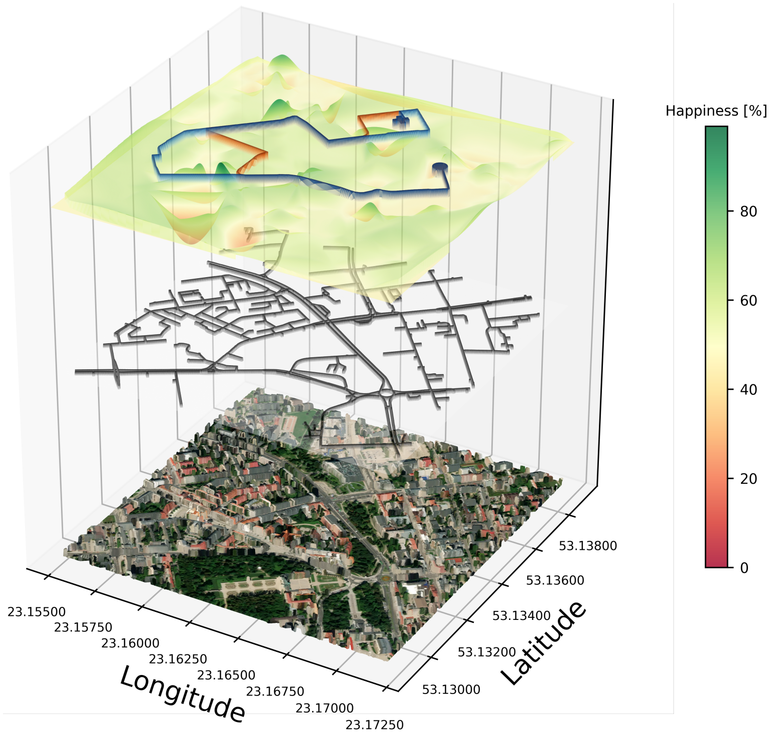
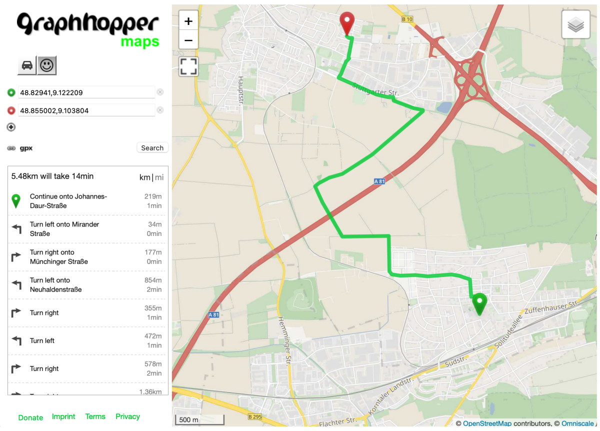


Fig. 3. Graph Building for Happy Route Optimization. The navigation finds the optimal emotional path according to the emotion-road-weight regularization (equation 2). The bottom layer is a satellite image. The layer above represents the routable roads. Above is emotion heatmap based on interpolation of the computed happiness points. The red path is the fastest path offered by navigation, while the blue path is the happy path.

the routing optimization according to Equation 2. We do not employ a standard A* algorithm [26] for optimal route finding due to performance reasons. Instead, we disable all initial edge weight calculations for happy routing and build a prominently-used CH (Contraction Hierarchy) graph [23] with precalculated happiness weights to speed-up optimization (**Req 4**). Thereafter, our system exposes a happy and fastest routing computation endpoint. We show the interactive GraphHopper routing endpoint for a happy route computation in Figure 4.

Smartphone Navigation App. To provide users with the ability to navigate, we implemented a scalable mobile application. Therefore, we customized the Android application PocketMaps[8] to use our optimization engine (**Req 3**). Our mobile application tracks the current smartphone geolocation using GPS and is able to search for destinations on the map via Google Maps search. The application then performs map matching of the current geocoordinate to the road segment and outputs turn-by-turn navigation instructions (via text and voice). Users can choose between fastest and happy routing in our app. Figure 5 shows the navigation screen of our customized PocketMaps application in the wild.

[8] https://github.com/junjunguo/PocketMaps



48.82941,9.122209

48.855002,9.103804

5.48km will take 14min

Continue onto Johannes-Daur-Straße	219m	1min
Turn left onto Mirander Straße	34m	0min
Turn right onto Münchinger Straße	177m	0min
Turn left onto Neuhaldenstraße	854m	2min
Turn right	355m	1min
Turn left	472m	1min
Turn right	578m	2min
	1.36km	

Fig. 4. GraphHopper web-server for Happy Route Optimization in a 2D-layout.

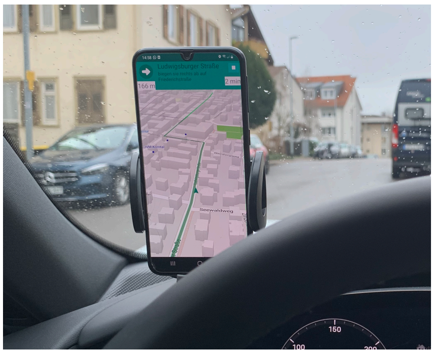


Fig. 5. Implemented navigation app that supports normal and happy routing. The app is placed on the windshield and has the same functionality as normal navigation apps (turn-by-turn navigation, voice output for hinting next directions).

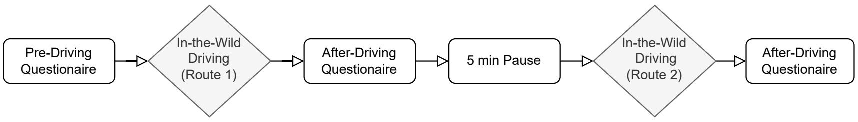

Pre-Driving Questionnaire
In-the-Wild Driving (Route 1)
After-Driving Questionnaire
5 min Pause
In-the-Wild Driving (Route 2)
After-Driving Questionnaire

Fig. 6. Experimental design of the emotional navigation driving study. The endpoint of the second drive was set to be the start point of the first drive.

5 DRIVING STUDY

The goal of our driving study is to gain an understanding of HappyRouting's user experience and its influence on a driver's emotions.

5.1 Within-Subject Study Design

We conducted a within-subject driving study to evaluate the experience with HappyRouting. We recruited 13 participants (11 self-identified as male, two self-identified as female) with an average age of 27 ± (8.51) years. Six participants drive occasionally (i.e., less than 10,000 km/year), six participants drive moderate distances (i.e., between 10,000 and 20,000 km/year), and one participant is a frequent driver (>20,000 km/year). The participants accessed a vehicle with a standard Android smartphone attached to the windscreen (see Figure 5). We gave the participants time to get familiar with the car and explained that they could drive like they normally do (e.g., listening to music). We asked the participant to use our HappyRouting application just like a common mobile navigation app. The start and end location was chosen to be a 15-minute drive away with segments including rural and urban segments. The calculated routes were kept consistent for all participants to ensure comparability. The routing choice (fastest or happy routing) was hidden in the mobile application to avoid confirmation bias (i.e., blind route choice). The routing choice was randomized so that 7 participants drove the happy route first, while 6 drivers were assigned to the fastest route first. Overall the one-way driving had a duration of approximately twelve minutes for the fastest route and 14 minutes for the happy route, depending on individual traffic conditions.

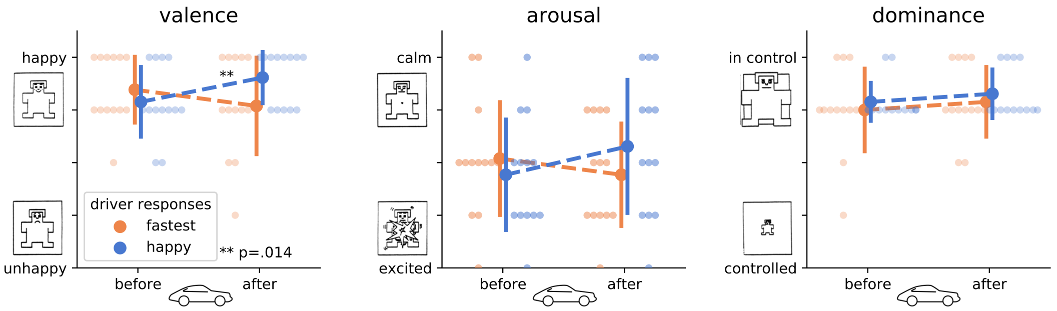


Fig. 7. Before and after driving analysis of valence (left), arousal (middle) and dominance (right) questionnaire answers of the driving study. The lines indicate the standard deviation (vertical) of the responses where the means are connected via the dashed line. The asterisk indicates significance. Fast and happy routes were assigned blindly and by random succession.

For each assessment of the driver's emotional state (valence, arousal), we apply the self-assessment manikin (SAM) framework [7] with a five-point Likert scale. The outline of the study is presented in Figure 6.

Valence-Arousal-Dominance Analysis. We present the before and after analysis of valence, arousal and dominance scores assessed with the self-assessment manikin questionnaire in Figure 7. We find that people gave higher valence ratings, i.e., positive attitudes, after taking the happy route. The mean valence score for happy routing before driving is 4.15 and increases to 4.62 after driving (11% valence score increase). Applying a Shapiro-Wilk test revealed a non-normal distribution, $p < .001$. A Wilcoxon signed rank test found a significant difference in valence before and after navigating through a happy route, $p = .014$. In addition, we did not find significant before-after differences in valence or arousal when driving the fastest route. Overall, we found a positive trend in arousal when driving the happy route, though all expressed arousal levels have high variance. The high variance likely comes from the fact that the driving task is perceived as relaxing or exciting on an individual driver's basis. A Shapiro-Wilk test showed a non-normal distribution for arousal, $p = .025$. There was no significant difference in arousal before and after driving the happy route according to a Wilcoxon signed rank test, $p = .1$. This finding stands in contrast to many empathic car applications that seek to optimize arousal levels for safety reasons [8, 9]. Furthermore, we did not detect any significant changes in dominance scores i.e. we detect no effect in how controlled or submissive one feels after driving the happy route.

Happy Navigation Driving Behavior. In our driving questionnaire, we saw high variability when and how drivers wanted to use happy navigation functionality. After the driving experiment, we asked the participants how much time they were willing to sacrifice for a happy route, assuming 20 minutes for the fastest route. 9 of 13 participants answered with 3 – 5 minutes, while 3 of 13 drivers would only spend 1 – 3 minutes additional drive time. One participant stated that he would even spend more than 10 minutes of additional drive time to drive the happy route. These results are consistent with the web survey by Pfleging et al. [45] which states that participants would take 20.9% more time for an experience-optimized route compared to the fastest route. While the fastest route took on average 2 minutes less time, 8 of 13 participants perceived the happy route as shorter. Combined with the finding that subjects had a more positive emotional state after driving the happiness route, we conclude that a happiness route may positively influence the perceived travel time. Furthermore, in our study, 11 out of 13 participants stated that they would use the app in their leisure time when they do not have time pressure. Interestingly, many participants responded to use our navigation only on the weekend (P9, P10, P12), preferably in the summer (P1, P2, P3, P4, P9, P10, P12), and not at night when the driving scenery is not visible (P8, P13). P2 mentioned that he would use happy routes “if a traffic jam occurs and he could take lesser crowded, more relaxed and unknown routes”.

System Acceptance. In response to the question "How likely would you be to use this system?" on a scale of 1 (not at all likely) to 5 (very likely), 11 of 13 participants gave scores of 4 and 5. The study participants further introduced ideas for pairing happy navigation with other further in-car technology. The most prominent responses were that many participants associate happiness with music while driving. Therefore, many suggestions were made to automatically select music while driving to match the route, or vice versa, to select the route to better match the music.

We also asked the participants in a free-response question, "Do you think there are any societal and ethical implications of this navigation functionality? And if yes, which one?". Many participants said they do not see any ethical or societal implications (P6, P7, P9, P11, P13). Participants also responded with higher energy consumption costs and a more environmentally harmful behavior when driving a happy route (P1, P3, P10). P10 stated that he sees a problem with happy routing only recommending pleasant driving routes so that other less happy predicted locations do not get visibility, creating a self-reinforcing effect of what people see.

6 SIMULATION STUDY

To offer a broad assessment of the recommended happy routes by our system, we perform an offline numerical simulation analysis.

6.1 Experiment Design

First, we download and compute the emotion prediction layer for a map of a medium-sized city (12 × 12 km). We sample a large number of equally-distributed, random start and end points ($N = 1000$) and search for the happy and fastest routes. We then analyze the route trajectories segments by computing several characteristics such as road types, greenness, traffic conditions, and curviness. Furthermore, we compute the travel time, distance, and the overlap of fastest and happy routes.

6.2 Route Time Analysis

We anticipated that taking the happy route would increase the travel time. Figure 8 shows the relationship of the navigation mode on travel times. Using linear regression, we find that a one-minute increase in fastest routing requires on average 1.25 minutes (75 sec.) more time to drive using happy routing. Only 9% of the start-end coordinates result in a situation where the happy route is identical to the fastest route ($overlap = 100\%$). The time difference can be substantial in individual cases, therefore, we stress a transparent time forecast when recommending happy routes to drivers. We conclude that the factor λ should rather be regarded as an internal technical parameter (see the influence of λ in Figure 9) instead of a user-adjustable parameter. Higher λ results in increased average travel time and therefore cause longer travel times. Therefore, λ can be adjusted dynamically to suit the societal driving context.

6.3 Road Characteristics

We analyze the recommended happy and fastest route for their road types in Figure 10 and Figure 11. As the drive-time is normalized per individual route, the values of the bars do not add up to 100%. We tested whether the distribution of the different road characteristics are significantly different ($p < .01$) using a non-parametric Mann-Whitney U

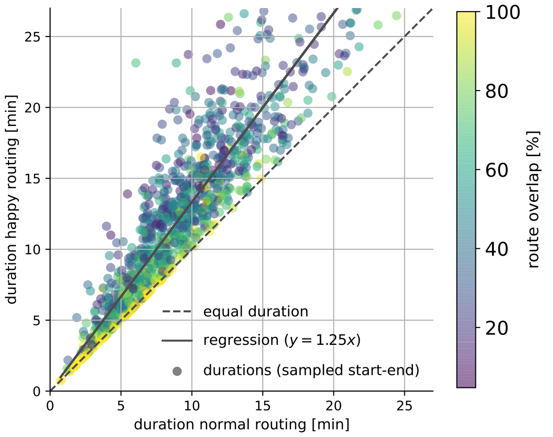


Fig. 8. Scatterplot of drive duration of normal vs. happy routing. The points are mostly on the top-left of the equal-traveltime line, meaning that happy routing takes generally longer to drive. The fitted regression ($R^2 = 0.82$, $BIC = 4715$) with a slope of $\beta_1 = 1.25$ ($p = .00$) means that a 1-minute increase in normal routing will take 1.25 minutes (75 *sec*.) more time to drive using happy routing.

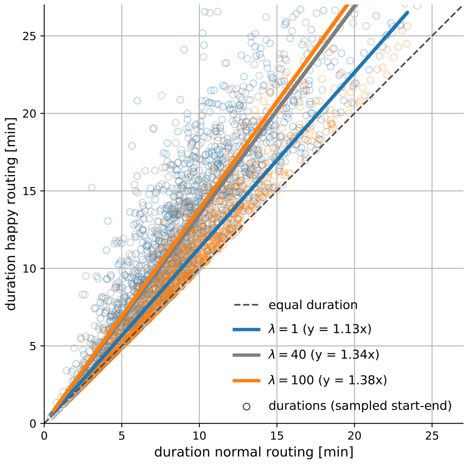


Fig. 9. Influence of emotional weighing factor λ on happy routes. The additional travel time for happy routing does not scale linearly with the emotional weighing factor λ. On average, the setting $\lambda = 40$ achieves a similar time divergences as $\lambda = 100$.

test. Compared to the fastest route, we find that happy routes consist of more road segments with a higher predicted happiness score, higher curviness, higher freeflow speed, and maximum speed.

Curvy roads tend to increase driving enjoyment but also inhibit driving accident risks [27]. Unhindered traffic scenarios can be captured by our proxy variable freeflow speed, which is higher for happy routes, and increase driver well being [49]. Contrary to our initial hypothesis, we detect no significant effect of the satellite-image-derived greenness (known part of the HSV spectrum) in happy routes compared to the fastest route ($p = 0.29$). Finally, we find that on-average happy routing includes significantly more residential roads. Residential roads often have reduced traffic and may reduce drivers' stress, leading to a more happy emotional state. In contrast, the recommended fastest routes contain significantly more living street and primary road segments, which often require more driver attention. As stated before, these findings are made based on a large sample size and do not represent an individual recommended route.

6.4 Computational Characteristics

Navigation systems deployed in the wild require high scalability. To assess the computational complexity of our system, we computed the execution time of the routing endpoint (GraphHopper). On the 12 × 12 km map, our system needs to perform emotion prediction on 21, 673 unique edges and caches the corresponding data in the optimization graph. The cache is needed because the input data is collected from various APIs, which makes on-demand prediction attainable when optimizing the route. In a subsequent step, the execution time for recommending happy routes is 0.08 ± 0.075 seconds and takes longer to compute than fastest routing 0.01 ± 0.004. With recommendation times smaller than 1 second, our system is highly time-efficient and user-friendly.

7 DISCUSSION

With HappyRouting, drivers perceived a higher valence when using the happy route compared to the fastest route, showing that choosing emotionally positive routes contribute to a driver's well-being. In the following, we discuss the implications of our results.

7.1 Tradeoff Between Valence and Route Duration

Our results suggest a tradeoff between the duration of the fastest route and the perceived valence of driving the happy route. Although the happy route takes more time, our participants would use the HappyRouting for their navigation to improve their emotional well-being. However, due to the longer travel times, the majority of our participants indicated that they would prefer the HappyRouting if they were not pressed for time. In addition, our study results suggest other modalities for controlling driver emotions by combining the in-vehicle environment with the suggested happy route. For example, participants proposed to explore music in combination with happy routes to enhance feelings of happiness. Using individual preferences for the in-vehicle environment as an additional variable can lead to emotion prediction models that ultimately reduce driving time. However, the combination of in-vehicle adaptions with happy routes proposed by HappyRouting requires further research.

7.2 Using HappyRouting for Other Transport Modalities

HappyRouting generates routing decisions that can be used in various other transportation modalities once the foundation for a context-aware machine learning classifier is established. With a few modifications, HappyRouting can apply emotion-based navigation for cyclists by predicting the emotionally pleasant cycling route. We propose to incorporate advanced contextual sensors when optimizing happy routes for other road users (e.g., pedestrians or cyclists) by extending the feature set to include elevation information and information about road intersections. For the application of HappyRouting in pedestrian routing, we recommend extending our feature set to include traffic banning features, as these have been shown to influence valence [43] positively.

7.3 Ethics & Societal Impact

We emphasize an ethical as well as transparent use of HappyRouting for application purposes and stress that emotions are intimate, personal, and vulnerable [2]. The Emotional Artificial Intelligence ethics guidelines by McStay et al. [39] provided us with a meaningful reference to cover personal, relationship, and societal aspects.

Our approach is privacy-aware because it uses a machine learning model based on an aggregate, anonymized dataset provided in advance by a set of volunteers, rather than subconsciously assessing the emotions of individual HappyRouting users. On the other hand, we also see clear limitations of our dataset in the area of cultural and regional diversity, as well as the explainability of resulting algorithm choices. Future empathic car interfaces must clearly communicate how and what data is assessed to clarify how this subsequently affects the users' privacy.

Undeniably, the regulation of emotions by technological systems is highly controversial, as psychological effects are largely unknown. Avoidance of negative situations, for example, is an essential strategy of human emotion (self-)regulation [38], but also an implicit result of our system's promotion of positive emotions. Studies with individuals have shown that situation avoidance results in decreased learning and adaptation abilities, as well as social and anxiety disorders [1]. Therefore, we emphasize that such short- and long-term effects must be investigated in future work.

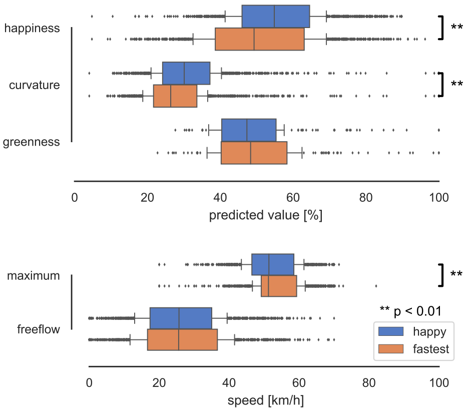


Fig. 10. Characteristics of happy route vs. fastest route. Distribution of happiness, curviness, greeness, max_speed and freeflow_speed for the two routing modes.

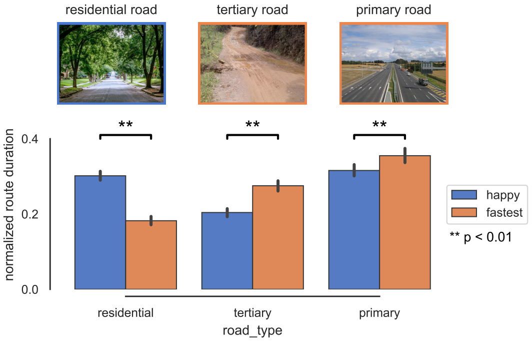


Fig. 11. Analysis of road types of happy routing vs. fastest routing. We assess the road type of every road segment (x-axis) and compute the drive-time normalized route duration (y-axis). All the presented road types have been tested to be significantly different ($p < .01$). Residential roads are are found in living areas, primary road types are major highways linking large towns and tertiary roads connect minor streets to more major roads.

Our study of route characteristics shows that heavily traveled routes are often avoided in favor of quieter routes. To us, this is a clear indication for future work, as these externalities at large-scale can potentially affect residential areas, parks, or nature, as Johnson et al. note [30].

We showed that routes proposed by HappyRouting result in increased travel times which are ultimately bound to higher energy consumption. Certain route choices might affect the safety of traffic participants, for example, due to a model's preference for specific road types. These and many other route characteristics must be communicated transparently to the users to promote their autonomy and enable highly informed choices [39]. Alternative strategies could comprise correction terms applied to our optimization, for example, when the routing choice is not desirable on a societal basis (e.g., routing through densely populated areas) [30].

7.4 Limitations & Future Work

This work takes the first steps towards a novel type of empathic car interface based on emotional predictions and optimizations through routing. We accepted several limitations in the domains of psychology, algorithms, and the user experience to achieve this goal. First and foremost, the psychological model of fostering well-being through aggregation of positive emotions is deliberately oversimplified, as discussed in Section 3. Future models could operate on a diverse emotional flow [34], which requires significant changes to the optimization method and its proven graph algorithms. The utilization of emotion-related signals during driving would enable the dynamic updating of the predicted emotional weights and real-time adaptation of the happy route. This feature can be easily integrated into the current system architecture, but it should be approached with caution as it has the potential to be perceived as privacy-intrusive. However, the benefit of our non-interactive emotion navigation system is that it allows for an empathic interface without compromising privacy during operation, and the option to switch to a different routing modality can be easily selected at any point during the journey.

HappyRouting requires the ability to simulate the driver's emotions for any road segment at any time while considering contextual information like traffic, road types, and speed limits. A key design decision for simulation lies in the choice between subjective and objective metrics for characterizing user emotions. HappyRouting relies on a dataset containing self-expressed and thus subjectively perceived emotions for prediction. Consequently, we base the simulated emotions on discrete representations of emotions, as identified by Ekman [18]. However, a detailed comparison with objective metrics should be subject to future work, such as using facial expressions [36] or muscle stiffness [3]. At the same time, the car offers only a limited set of remotely accessible contextual features for predicting driver emotions, making the modeling complex.

Our navigation framework is based on an emotion prediction layer which can be adapted easily towards additional modalities. Weighting in objective parameters such as the greenness score [4] could promote user-specific preferences without the need for a personalized emotion model. On the other hand, user-dependent models can further increase the accuracy, as shown in related work [3, 6]. Finally, we see limitations in explaining the overall recommendation process to the end-user which is ultimately very important for the ethics and transparency of our system. The intransparency can lead to placebo effects, where the description of using an allegedly adaptive AI-driven system biases the perceived system utility for drivers [35]. In future work, we plan to summarize how route recommendations were computed on an individual user's basis and research how to communicate key emotional route segments [8]. Finally, further long-term experiments with a larger variety of roads and routes under vastly different conditions are needed to produce necessary evidence of the proposed models ability to find happy routes. These long-term studies in the wild may help to better understand the effects and societal impact of affective routing.

8 CONCLUSION

This paper presents HappyRouting, a new type of empathic interface capable of navigating by positive emotions. We use personal, environmental, and road-specific information to define a custom emotion routing graph that optimizes routes for happy emotions. A real-world user study shows that HappyRouting elicits positive emotions through navigation. As a consequence, HappyRouting requires more driving time which was accepted by our participants as long as the circumstances allowed it (e.g., no time pressure). Our work is not only relevant to driving but can also be applied to other areas of mobility and autonomous driving. We are confident that the presented process of simulating emotions and evaluating different paths through many potential user journeys can be generalized to an even wider variety of use cases. To encourage research in this area, we publish the source code of our system and the data set for further analysis by the research community [9].

REFERENCES

[1] Amelia Aldao, Gal Sheppes, and James J Gross. 2015. Emotion regulation flexibility. *Cognitive Therapy and Research* 39, 3 (2015), 263–278. https://doi.org/10.1007/s10608-014-9662-4

[2] Nazanin Andalibi and Justin Buss. 2020. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–16. https://doi.org/10.1145/3313831.3376680

[3] Stephanie Balters, Nikhil Gowda, Francisco Ordoñez, and Pablo Paredes. 2021. Individualized stress detection using an unmodified car steering wheel. *Scientific Reports* 11 (10 2021). https://doi.org/10.1038/s41598-021-00062-7

[4] Fouad Baouche, Romain Billot, Rochdi Trigui, and Nour Eddin El Faouzi. 2014. Electric vehicle green routing with possible en-route recharging. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2787–2792. https://doi.org/10.1109/ITSC.2014.6958136

[9] https://anonymous.4open.science/r/affectroute-CD20/

[5] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 279–283. https://doi.org/10.1145/2993148.2993165

[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.* Association for Computing Machinery, New York, NY, USA, 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 differential. *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 Pfleging, 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, Jonas Schubert, Bastian Pfleging, and Florian Alt. 2019. Improving Driver Emotions with Affective Strategies. *Multimodal Technologies and Interaction* 3, 1 (March 2019), 21. https://doi.org/10.3390/mti3010021

[10] 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 Computing Surveys (CSUR)* 54, 7 (2021), 1–26. https://doi.org/10.1145/3460938

[11] 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 Computing Surveys (CSUR)* 54, 7 (2021), 1–26.

[12] Vaida Ceikute and Christian S. Jensen. 2013. Routing Service Quality – Local Driver Behavior Versus Routing Services. In *2013 IEEE 14th International Conference on Mobile Data Management*, Vol. 1. 97–106. https://doi.org/10.1109/MDM.2013.20

[13] Agatha Czekajlo, Nicholas C Coops, Michael A Wulder, Txomin Hermosilla, Yuhao Lu, Joanne C White, and Matilda van den Bosch. 2020. The urban greenness score: A satellite-based metric for multi-decadal characterization of urban land dynamics. *International Journal of Applied Earth Observation and Geoinformation* 93 (2020), 102210. https://doi.org/10.1016/j.jag.2020.102210

[14] Monique Dittrich and Sebastian Zepf. 2019. Exploring the Validity of Methods to Track Emotions Behind the Wheel. In *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems*, Harri Oinas-Kukkonen, Khin Than Win, Evangelos Karapanos, Pasi Karppinen, and Eleni Kyza (Eds.). Springer International Publishing, Cham, 115–127. https://doi.org/10.1007/978-3-030-17287-9_10

[15] 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 The proceedings of AmI, the 2018 European Conference on Ambient Intelligence..

[16] Ahinoam Eherenfreund-Hager, Orit Taubman – Ben-Ari, Tomer Toledo, and Haneen Farah. 2017. The effect of positive and negative emotions on young drivers: A simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour* 49 (2017), 236–243. https://doi.org/10.1016/j.trf.2017.07.002

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

[18] Paul Ekman. 1992. Are there basic emotions? *Psychological Review* 99, 3 (1992), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

[19] Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[20] D. Eppstein. 1994. Finding the k shortest paths. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*. 154–165. https://doi.org/10.1109/SFCS.1994.365697

[21] Nico H Frijda. 1993. Moods, emotion episodes, and emotions. (1993).

[22] Matthew Gallagher, Shane Lopez, and Kristopher Preacher. 2009. The Hierarchical Structure of Well-Being. *Journal of personality* 77 (06 2009), 1025–50. https://doi.org/10.1111/j.1467-6494.2009.00573.x

[23] Robert Geisberger, Peter Sanders, Dominik Schultes, and Daniel Delling. 2008. Contraction hierarchies: Faster and simpler hierarchical routing in road networks. In *International Workshop on Experimental and Efficient Algorithms*. Springer, 319–333. https://doi.org/10.1007/978-3-540-68552-4_24

[24] Reginald G. Golledge. 1995. Path selection and route preference in human navigation: A progress report. In *Spatial Information Theory A Theoretical Basis for GIS*, Andrew U. Frank and Werner Kuhn (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 207–222.

[25] GM Hancock, PA Hancock, and CM Janelle. 2012. The impact of emotions and predominant emotion regulation technique on driving performance. *Work* 41, Supplement 1 (2012), 3608–3611. https://doi.org/10.3233/WOR-2012-0666-3608

[26] Peter E Hart, Nils J Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107.

[27] Robin Haynes, Andrew Jones, Victoria Kennedy, Ian Harvey, and Tony Jewell. 2007. District variations in road curvature in England and Wales and their association with road-traffic crashes. *Environment and planning A* 39, 5 (2007), 1222–1237. https://doi.org/10.1068/a38106

[28] Douglas Heaven. 2020. Why faces don't always tell the truth about feelings. *Nature* 578, 7796 (Feb. 2020), 502–504. https://doi.org/10.1038/d41586-020-00507-5

[29] Javier Hernandez, Daniel McDuff, Xavier Benavides, Judith Amores, Pattie Maes, and Rosalind Picard. 2014. AutoEmotive: Bringing Empathy to the Driving Experience to Manage Stress *(DIS Companion '14)*. Association for Computing Machinery, New York, NY, USA, 53–56. https://doi.org/10.1145/2598784.2602780

[30] I. Johnson, J. Henderson, C. Perry, J. Schöning, and B. Hecht. 2017. Beautiful…but at What Cost? An Examination of Externalities in Geographic Vehicle Routing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 15 (jun 2017), 21 pages. https://doi.org/10.1145/3090080

[31] Ioannis Kaparias, Michael G. H. Bell, Klaus Bogenberger, and Yanyan Chen. 2007. Approach to Time Dependence and Reliability in Dynamic Route Guidance. *Transportation Research Record* 2039, 1 (2007), 32–41. https://doi.org/10.3141/2039-04 arXiv:https://doi.org/10.3141/2039-04

[32] Piyawan Kasemsuppakorn and Hassan A. Karimi. 2009. Personalised Routing for Wheelchair Navigation. *J. Locat. Based Serv.* 3, 1 (mar 2009), 24–54. https://doi.org/10.1080/17489720902837936

[33] Matthew C. Keller, Barbara L. Fredrickson, Oscar Ybarra, Stéphane Côté, Kareem Johnson, Joe Mikels, Anne Conway, and Tor Wager. 2005. A Warm Heart and a Clear Head: The Contingent Effects of Weather on Mood and Cognition. *Psychological Science* 16, 9 (2005), 724–731. https://doi.org/10.1111/j.1467-9280.2005.01602.x PMID: 16137259.

[34] Silvia Knobloch. 2003. Mood Adjustment via Mass Communication. *Journal of Communication* 53, 2 (2003), 233–250. https://doi.org/10.1111/j.1460-2466.2003.tb02588.x

[35] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The Placebo Effect of Artificial Intelligence in Human-Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* (mar 2022). https://doi.org/10.1145/3529225 Just Accepted.

[36] Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, and Felix Wortmann. 2021. The empathetic car: Exploring emotion inference via driver behaviour and traffic context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–34. https://doi.org/10.1145/3478078

[37] Sachit Mahajan, Yu-Siou Tang, Dong-Yi Wu, Tzu-Chieh Tsai, and Ling-Jyh Chen. 2019. CAR: The Clean Air Routing Algorithm for Path Navigation With Minimal PM2.5 Exposure on the Move. *IEEE Access* 7 (2019), 147373–147382. https://doi.org/10.1109/ACCESS.2019.2946419

[38] Kateri McRae and James J Gross. 2020. Emotion regulation. *Emotion* 20, 1 (2020), 1. https://doi.org/10.1037/emo0000703

[39] A. McStay and P. Pavliscak. 2019. Emotional Artificial Intelligence: Guidelines For Ethical Use. Website. Retrieved March 10, 2022 from https://drive.google.com/file/d/1frAGcvCY_v25V8ylqgPF2brTK9UVj_5Z/view.

[40] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[41] Robin L. Nabi and Melanie C. Green. 2015. The Role of a Narrative's Emotional Flow in Promoting Persuasive Outcomes. *Media Psychology* 18, 2 (2015), 137–162. https://doi.org/10.1080/15213269.2014.912585

[42] Clifford Nass, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. 2005. Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1973–1976. https://doi.org/10.1145/1056808.1057070

[43] Felix Ortag and Haosheng Huang. 2011. Location-based emotions relevant for pedestrian navigation. In *Proceedings of the 25th International Cartographic Conference*.

[44] Christelle Pêcher, Céline Lemercier, and Jean-Marie Cellier. 2009. The influence of emotions on driving behaviour. *Traffic Psychology and driving behaviour, New-York: Hindawi Publishers* (2009).

[45] Bastian Pfleging, Stefan Schneegass, Alexander Meschtscherjakov, and Manfred Tscheligi. 2014. Experience Maps: Experience-Enhanced Routes for Car Navigation *(AutomotiveUI '14)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/2667239.2667275

[46] D. Pfoser, N. Tryfona, and A. Voisard. 2006. Dynamic Travel Time Maps - Enabling Efficient Navigation. In *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*. 369–378. https://doi.org/10.1109/SSDBM.2006.19

[47] Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*. Vol. 1. Academic Press.

[48] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. 2014. The Shortest Path to Happiness: Recommending Beautiful, Quiet, and Happy Routes in the City. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. Association for Computing Machinery, New York, NY, USA, 116–125. https://doi.org/10.1145/2631775.2631799

[49] Ernst Roidl, Berit Frehse, and Rainer Höger. 2014. Emotional states of drivers and the impact on speed, acceleration and traffic violations—A simulator study. *Accident Analysis & Prevention* 70 (2014), 282–292. https://doi.org/10.1016/j.aap.2014.04.010

[50] Nina Runge, Pavel Samsonov, Donald Degraen, and Johannes Schöning. 2016. No More Autobahn! Scenic Route Generation Using Googles Street View. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 147–151. https://doi.org/10.1145/2856767.2856804

[51] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714 Place: US Publisher: American Psychological Association.

[52] Briane Paul V. Samson and Yasuyuki Sumi. 2019. Exploring Factors That Influence Connected Drivers to (Not) Use or Follow Recommended Optimal Routes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300601

[53] Monir H. Sharker, Hassan A. Karimi, and Janice C. Zgibor. 2012. Health-Optimal Routing in Pedestrian Navigation Services. In *Proceedings of the First ACM SIGSPATIAL International Workshop on Use of GIS in Public Health* (Redondo Beach, California) *(HealthGIS '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2452516.2452518

[54] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A Review of Emotion Recognition Using Physiological Signals. *Sensors* 18, 7 (2018). https://doi.org/10.3390/s18072074

[55] Jianhua Tao and Tieniu Tan. 2005. Affective Computing: A Review. In *Affective Computing and Intelligent Interaction*, Jianhua Tao, Tieniu Tan, and Rosalind W. Picard (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 981–995. https://doi.org/10.1007/11573548_125

[56] Arash Tavakoli, Mehdi Boukhechba, and Arsalan Heydarian. 2021. *Leveraging Ubiquitous Computing for Empathetic Routing: A Naturalistic Data-Driven Approach.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451693

[57] Thorsten Völkel and Gerhard Weber. 2008. RouteCheckr: Personalized Multicriteria Routing for Mobility Impaired Pedestrians. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 185–192. https://doi.org/10.1145/1414471.1414506

[58] Heetae Yang, Jieun Yu, Hangjung Zo, and Munkee Choi. 2016. User acceptance of wearable devices: An extended perspective of perceived value. *Telematics and Informatics* 33, 2 (2016), 256–269. https://doi.org/10.1016/j.tele.2015.08.007

[59] Jyun-Yan Yang, Li-Der Chou, and Yao-Jen Chang. 2016. Electric-Vehicle Navigation System Based on Power Consumption. *IEEE Transactions on Vehicular Technology* 65, 8 (2016), 5930–5943. https://doi.org/10.1109/TVT.2015.2477369

[60] Sebastian Zepf, Monique Dittrich, Javier Hernandez, and Alexander Schmitt. 2019. Towards empathetic Car interfaces: Emotional triggers while driving. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6. https://doi.org/10.1145/3290607.3312883

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

A APPENDIX

Curviness Computation

We compute the curviness using a weighted measure of the length of curves, which depends on a radius of a circumscribed circle that passes through all three consecutive geocoordinates in a route. Given a, b, c as the length of the three sides of a triangle, the radius of the circumcircle is given by the formula:

$$r = \frac{abc}{\sqrt{(a+b+c)(b+c-a)(c+a-b)(a+b-c)}} \tag{3}$$

Classifier Feature Importances

We analyze how decisive each contextual input feature is for our human emotional state classification model . We extract the feature importance (Gini impurity) of the input variables in a leave-one-participant-out situation in Figure 12. The variable 'greeness' shows the highest importance for the classifier to predict the likely emotional state on the road. However, these feature importances are aggregate metrics and do not convey any participant-dependent importance measures for a specific routing choice (local feature importance measures such as SHAP values are needed for this). Here, we only analze the feature importance of the emotion classification model, a route-specific analysis of the road properties can be found in section 6.3.

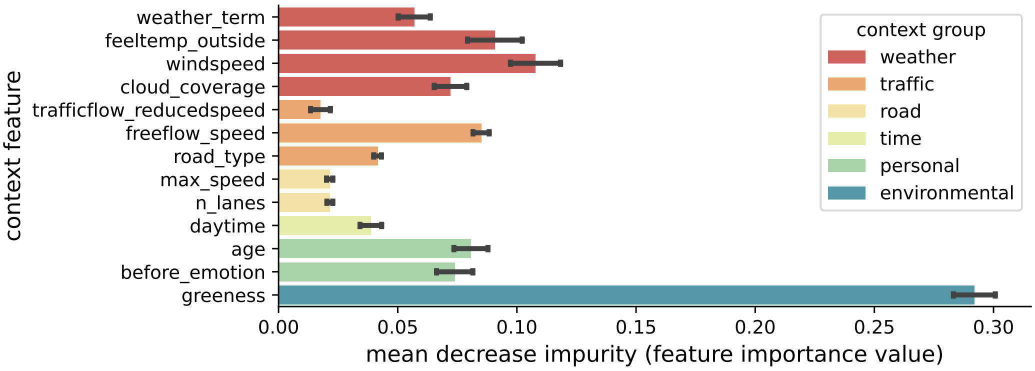


Fig. 12. Feature importances measured by the mean decrease of Gini-impurity for the Leave-One-Participant-Out cross validation.

EEG2Vec: Learning Affective EEG Representations via Variational Autoencoders

David Bethge[1,2,*], Philipp Hallgarten[1,3], Tobias Grosse-Puppendahl[1], Mohamed Kari[1], Lewis L. Chuang[4], Ozan Özdenizci[5,6], Albrecht Schmidt[2]

Abstract—**There is a growing need for sparse representational formats of human affective states that can be utilized in scenarios with limited computational memory resources. We explore whether representing neural data, in response to emotional stimuli, in a latent vector space can serve to both predict emotional states as well as generate synthetic EEG data that are participant- and/or emotion-specific. We propose a conditional variational autoencoder based framework, EEG2Vec, to learn generative-discriminative representations from EEG data. Experimental results on affective EEG recording datasets demonstrate that our model is suitable for unsupervised EEG modeling, classification of three distinct emotion categories (positive, neutral, negative) based on the latent representation achieves a robust performance of 68.49%, and generated synthetic EEG sequences resemble real EEG data inputs to particularly reconstruct low-frequency signal components. Our work advances areas where affective EEG representations can be useful in e.g., generating artificial (labeled) training data or alleviating manual feature extraction, and provide efficiency for memory constrained edge computing applications.**

I. Introduction

The emphasis on human-centric computing in recent years has accelerated efforts in affective computing to develop effective computer-aided approaches for recognizing, interpreting, processing, and simulating a person's emotions. Recent notable successes include affect-adaptive robot-child feedback in education [1], mobile real-time facial emotion annotation systems [2], as well as detecting and influencing drivers' emotions [3]. In particular, the growing availability of high-resolution wearable physiological measurement systems has, in combination with powerful machine learning methods, increased recognition performance of affective user states for various applications in-the-wild [4].

In this paper, we focus on electroencephalographic (EEG) measurements that are elicited in response to affective emotional stimuli. Primarily we seek to determine whether a user-specific affective representation from raw EEG can be learned in an end-to-end *representation learning* framework. In such a setting, representations of affective states are learned from input data—typically by transforming it or extracting informative features from it (the useful vantage point of the data's key qualities)—towards the objective of performing particular tasks like prediction of affective states from noisy EEG data. Traditional discriminative machine learning approaches have the sole objective of classifying distinct affective categories. Differently, our focus on representation learning aims at estimating a powerful abstraction of affect-relevant multi-channel sensor input. This approach encodes the signal generative components from the training data distribution in a learned latent space. Once such models are learned, semi-supervised learning schemes can be adopted by e.g., applying the representation learning encoder on unlabeled data for class-conditional data synthesis, which can be combined with labeled data representations for a larger training base to, for instance, predict emotions.

Our work is inspired by recent advances in word representations, also denoted as *embeddings* [5], [6]. A prominent success story is *word2vec* [7] in natural language processing, which uses a neural network model to learn word representations from a large text corpus. Once trained, such representation models can detect synonymous words or accurately suggest additional words for a partial sentence. This has given rise to numerous natural language applications that were previously unimaginable (e.g., predicting the right next words in chats, sentiment analysis of messages, machine translation, click session advertisement-recommendation, automatic topic clustering). State-of-the-art natural language processing algorithms can even learn cross-lingual concepts [8], generate complete texts [9], [10] or infer emotion-related text sentiments [11]. When trained with enough data, word *embeddings* tend to capture word concepts and meanings, and are even able to perform analogies, e.g., the vector for "Paris" minus the vector for "France" plus the vector for "Italy" is very close to the vector for "Rome". Thereby, word representations can bridge the human understanding of language to that of a machine.

We seek to contribute towards ubiquitous wearable physiological systems and allow for computing systems that are responsive to user affective states in real-world scenarios. Here, we focus on high-dimensional EEG data and share how a highly informative and compressed representation could be derived from it to support this vision. Once learned, such representations can be useful for downstream tasks, such as predicting emotions to better understand the affective states in the brain through representations in a lower dimensional space, or simulating synthetic EEG data.

To date, affective brain-computer interface (BCI) methodologies are often hindered by the lack of large labeled datasets, low signal-to-noise (SNR) ratio in real-time EEG data acquisition, or non-reproducible handcrafted feature extraction [12]. Our work investigates variational representation learning for affective EEG data [13], that is able to mitigate these issues by learning a suitable and easy-to-use emotion representation trained for data augmentation and emotion recognition. Furthermore, many in-the-wild affective applications share the common obstacle of efficiently processing raw EEG data [12], [14], which can be reduced by processing an information condensed latent space vector, which ultimately

[1]Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany
[2]Ludwig Maximilian University, Munich, Germany
[3]University of Tübingen, Tübingen, Germany
[4]Chemnitz University of Technology, Chemnitz, Germany
[5]Institute of Theoretical Computer Science, TU Graz, Austria
[6]TU Graz-SAL DES Lab, Silicon Austria Labs, Graz, Austria
*Corresponding author: david.bethge@ifi.lmu.de

can be also streamed to the cloud with bandwidth restrictions.

We term our proposed model *EEG2Vec*, which exploits a conditional variational autoencoder (cVAE) [15] structure by encoding raw multi-channel EEG signals into a shared latent vector space while a simple feed-forward emotion classification neural network is simultaneously harnessing these latent representations as input. Subsequently, the resulting latent representations can be used to (1) predict the affective state of user from their current EEG recordings within a discriminative framework, as well as (2) to generate emotion- and subject-specific synthetic multi-channel EEG signals.

II. Related Work

A. EEG-based Affective State Estimation

Affective state estimation from EEG recordings have gained significant interest over the past decades [16]. Majority of existing methods rely on extracting single-channel features such as statistically derived features [17]–[19], fractal dimension [20], power-spectral-density (PSD) based features [21], differential entropy [22] or wavelet features [23]. Multiple features across several channels are then fused to exploit the inter-channel asymmetry or connectivity relationships. Beyond using traditional classifiers to discriminate such hand-crafted features, deep learning based end-to-end feature extraction and classification methods were also explored in EEG-based emotion recognition. Notably, [24] introduced a deep EEG classification neural network EEGNet, as they designed a generic and compact convolutional neural network (CNN) to accurately classify EEG signals from different tasks. Similarly specialized networks with hierarchical spatial and temporal EEG feature extraction for emotion recognition have also been proposed [25]–[28] (see [29] for a review).

B. Synthetic EEG Data Generation

Various approaches to generate synthetic EEG data have been recently explored to learn shared EEG components of across dataset samples. (cf. [30] for a recent review). One proposed framework for the generation of artificial data is generative adversarial networks (GANs) [31] which showed significant results for the generation of artificial images. This framework was applied to EEG data [32]–[34] revealing generated EEG signals by GANs resemble the temporal, spectral and spatial characteristics of real EEG. Another line of BCI studies have used variants of variational autoencoders (VAEs) [15], for unsupervised feature learning [30], [35], [36]. In contrast to GANs, VAEs optimize a parameterization of a low-dimensional representation space of the training data, and hence more suitable for learning compressed data representations which is of our main interest in the affective computing applications. Recently [35] proposed to use a standard VAE to learn latent codes containing emotion-related information and use in the downstream emotion classification task via an RNN-LSTM.

C. Deep EEG Latent Representation Learning

Several approaches in EEG representation learning use deep learning models for deterministic feature learning. One of the earlier works by [37] aims at finding robust representations from EEG data, that would be invariant to inter- and intra-subject differences and to inherent noise associated with EEG data collection. Their approach used “EEG movies” (topology-preserving multi-spectral images) and a CNN that is applied to a cognitive load classification task. [38] proposes a temporal and spatial feature concatenated vector representation learned with a compact deep multi-scale neural network, which is applied to diverse EEG tasks such as motor imagery, seizure or drowsiness detection.

Recently [39] proposed generative VAE and GAN models for data augmentation in emotion classification. They show that either full or partial selection of VAE or GAN results can be used to augment EEG training datasets and demonstrate an increase in affective state classification performances. In contrast to our approach that utilizes raw, multi-channel EEG signals in an end-to-end manner, their method does not consider temporal dependencies in input EEG and uses hand-crafted power spectral density features as network inputs. Another work employs a conditional VAE model based feature encoder on EEG data, and a CNN for downstream task classification [40]. Proposed approach aims to learn subject-invariant representations by simultaneously training a cVAE and an adversarial censoring network (similar to the idea from discriminative-adversarial settings [41]–[43]), for transfer generalization of the feature encoder that can efficiently process unseen users' EEG data for decoding. While prior work of EEG-based emotion recognition have mostly focused on defining explicit feature extraction and/or model architectures for detecting human emotions, our work is explicitly designed for generative-discriminative representation learning allowing for multiple classification tasks.

III. Methods

A. Notation

We denote the labeled dataset by $\{(\boldsymbol{X}_i, y_i, p_i)\}_{i=1}^n$. Here $\boldsymbol{X}_i \in \mathbb{R}^{C \times T}$ defines the EEG data matrix at trial i recorded from C channels (i.e., EEG sensors) for T discretized time samples. Accordingly, $p_i \in \{1, \ldots, P\}$ is the participant ID label and $y_i \in \{1, \ldots, L\}$ is the emotion category/class label of the corresponding trial. A deep latent representation learning model encodes an input $\boldsymbol{X}$ to learn a latent vector which we will denote by z.

B. Conditional Variational Autoencoder (cVAE)

In vanilla autoencoders, a deterministic encoder and decoder network pair learns a latent representation vector z that is sufficient to encode the underlying shared structure across the data samples $\boldsymbol{X}_i$, such that the decoder counterpart can fully reconstruct the input samples from these learned representations. In generative modeling with VAEs [15], however, the encoder network parameterized by ϕ is stochastic and estimates a true prior distribution $p(z)$ of latent z via a variational posterior distribution denoted by $q_\phi(z|\boldsymbol{X}) \sim N(\mu_z, \sigma_z)$. In practice the encoder network estimates the parameters μ_z and σ_z, and latent vectors are obtained by sampling from the estimated variational distribution $z \sim q_\phi(z|\boldsymbol{X})$ at the bottleneck. The subsequent decoder network parameterized by θ is then a generative model denoted by $p_\theta(\boldsymbol{X}|z)$, provided with these drawn samples z. VAEs are trained to jointly learn better approximations of

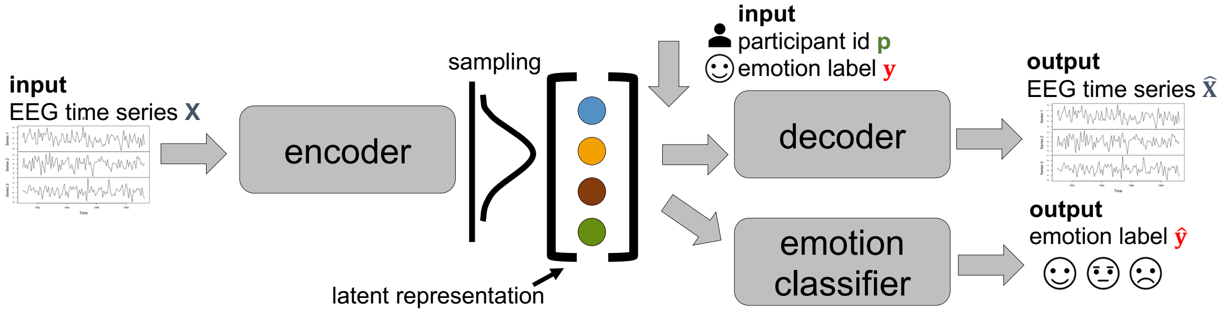


Fig. 1: Overview of the EEG2Vec model architecture. The variational autoencoder learns a subject- and emotion-dependent representation of EEG data that incorporates domain specific information regarding the classification task (i.e., emotion recognition). The parameters of the models are learned by minimizing the decoder reconstruction error, the loss regarding the variational posterior estimation, and the emotion classification loss.

the latent prior $p(z)$ via the variational posterior $q_\phi(z|\boldsymbol{X})$, and successful reconstructions of $\boldsymbol{X}$ by the decoder.

In the conditional VAE (cVAE) framework [44], the decoder is further conditioned on at least one additional variable provided as input besides z. We will define a cVAE to have a decoder posterior distribution that is conditioned on both p and y, thus modeling $p_\theta(\boldsymbol{X}|z, y, p)$. In a cVAE, the encoder is expected to learn representations z that are invariant of p and y, since p and y are already provided as input to the decoder to reconstruct the input $\boldsymbol{X}$. Training objective for variational autoencoder based models consists of maximizing a derived lower bound for the log likelihood of the data, which is usually referred as the evidence lower bound (ELBO) [15], [44]. Accordingly, the cVAE loss function to be minimized (i.e., negated ELBO) is given by:

$$\begin{aligned}\mathcal{L}_{\mathbf{cVAE}} = &\underbrace{\mathbf{E}_{q_\phi(z|\boldsymbol{X})}[-\log p_\theta(\boldsymbol{X}|z, p, y)]}_{\mathcal{L}_{recon}} \\ &+ \underbrace{D_{\mathbf{KL}}(q_\phi(z|\boldsymbol{X})||p(z))}_{\mathcal{L}_{KL}}.\end{aligned} \tag{1}$$

Here the first loss term $\mathcal{L}_{recon}$ corresponds to minimizing the reconstruction loss of the decoder, which is usually defined as the mean-squared-error between $\boldsymbol{X}$ and $\widehat{\boldsymbol{X}}$, and the second loss term $\mathcal{L}_{KL}$ corresponds to minimizing the Kullback–Leibler (KL) divergence between the encoder-estimated variational posterior $q_\phi(z|\boldsymbol{X})$ and the true distribution of z, which is usually defined as $p(z) \sim \mathcal{N}(0, I)$.

C. Beta Conditional Variational Autoencoder (β-cVAE)

We extend the cVAE models with what we will refer to as a beta conditional variational autoencoder (β-cVAE). In contrast to the original β-VAEs [45], the model simply utilizes a conditional decoder architecture as in a cVAE. Overall, this model proposes a modification to the objective in Eq. (1) by introducing a hyperparameter β for the KL-divergence term in the training loss function as follows:

$$\mathcal{L}_{\beta\text{-}\mathbf{cVAE}} = \mathcal{L}_{recon} + \beta\mathcal{L}_{KL}. \tag{2}$$

Intuitively, due to the traditional choice of $p(z) \sim \mathcal{N}(0, I)$, a higher β value will converge latent representation units to more strictly follow a standard normal distribution, leading to a unit diagonal covariance matrix and (ideally) statistically independent latent representation units. Hence, during representation learning optimization, imposing a higher weight for the KL term $\beta > 1$ is expected to successfully disentangle the latent representation units [46]. Therefore we utilize this constrained variational β-VAE [45] approach in our conditional representation learning setting in order to impose tunable regularization of the latent space.

D. EEG2Vec: A Generative-Discriminative EEG Representation Learning Framework

In the proposed EEG2Vec framework, a conditional β-VAE and a classifier to predict y (i.e., emotion category) from latent representations z are simultaneously trained. We extend the objective in Eq. (2) to obtain the EEG2Vec training objective function as given in Eq. (3).

$$\mathcal{L}_{\mathbf{EEG2Vec}} = \mathcal{L}_{\beta\text{-}\mathbf{cVAE}} + \lambda\underbrace{\mathbf{E}[-\log r_\varphi(y|z)]}_{\mathcal{L}_{cla}}. \tag{3}$$

Here we aim to minimize the cross-entropy loss of the emotion category classification network with the additional $\mathcal{L}_{cla}$ term. The emotion predictor function is described as r_φ with parameters φ, using the latent representation z to predict y. We introduce a tunable parameter $\lambda > 0$ in order to control the objective weighting for the model between a generative or discriminative behavior. For the deterministic decoder, the reconstruction loss is determined by the mean squared error of the estimated time-series EEG data.

Given the high dimensional input EEG data $\boldsymbol{X}$, its corresponding emotion label y and the participant ID p, the goal of the EEG2Vec model is to learn (1) a variational feature encoder $q_\phi(z|\boldsymbol{X})$ with parameters ϕ which can be generalized across subjects and emotion categories, (2) an adjacent decoder where novel data samples can be synthesized by exploiting this variational distribution, and (3) latent features from input EEG data $\boldsymbol{X}$ which are simultaneously representative in discriminating tasks or brain states associated with their corresponding emotion label y.

E. Generative-Discriminative Inference with EEG2Vec

After the EEG2Vec model is learned, the inference process with our model proceeds as follows. Given some participant p's EEG data sample $\boldsymbol{X}$ for inference, we encode $\boldsymbol{X}$ into a latent representation distribution and obtain a sampled vector z. From the learned representation and the classifier network we can predict the corresponding emotion category label, and thereby performing a conventional, discriminative emotion recognition task. Furthermore we can generate synthetic EEG data samples using this sample z, by only providing an emotional state label y and a participant ID p to the decoder network due to its conditional nature. Using this manipulation scheme, we also provide a generative model that can synthesize EEG data samples specific to a particular affective state y and participant ID p.

We regard the conditioning parameters at training and inference time to be known. However, it is possible to use the estimated values $\hat{y}$ of the emotion classifier network and the nearest neighbor p of X in the bottleneck as conditioning parameter estimates similarly proposed in [47] as conditional values in our architecture.

IV. Experimental Study

A. Experimental Dataset

There are a few publicly available EEG datasets with affective labels [13], [48]–[51]. For our study, we decided to use the STJU Emotion EEG Dataset (SEED) dataset [13] as it uses a rather simple labeling system with three distinct classes: negative, neutral and positive. This facilitates the learning of embeddings and reduces complexity for the emotion classifier. The dataset contains 62-channel EEG recordings sampled at 1000Hz from 15 participants recorded from 3 sessions. During each session, the participants were shown 15 film clips that should elicit either negative, neutral, or positive emotions representing the emotional label for that trial.

Since the duration of each experiment was different, to unify, we determined a 185 seconds duration (being the shortest duration of all experiments) as the standard experiment duration. For those experiments which duration is longer than 185s, the last 185s segment were selected. In order to avoid the possible interference or the possible emotions has not been elicited at the beginning of the experiment, we removed the first 30 seconds of EEG data, (i.e., only 155 seconds of segments were used [52]). Data were further preprocessed as in [13], [52] and accordingly first downsampled to 200Hz, and then a 2-40 Hz Butterworth bandpass filter was applied for low and high frequency band artifacts. EEG data were segmented into 2 second non-overlapping time intervals in accordance with previous work [53]. Finally, data is normalized to the range of $[0, 1]$. No offline channel selection was performed.

B. EEG2Vec Model Specifications

We developed our feature encoder backbone based on the well-known convolutional EEGNet architecture [24] due to its multi-purpose EEG representation learning capabilities. We modified the architecture based on the input representations of our dataset (e.g., sampling rate or number of EEG sensors). The decoder is implemented using inverse versions of the encoder layers. The encoded latent representation z is used as an input to the classifier that aims to accurately predict the corresponding emotional state. Herein, z is propagated through three fully-connected layers where the last layer is equipped with a softmax activation function. We initially set $\beta = \lambda = 1$, such that the reconstruction and classification performances are equally weighted in the loss function. Depending on the practical applications of our embedding, we can set these parameters accordingly.

C. Model Training and Evaluation

1) Implementation: Networks were trained with 50 training trials per batch for at most 2000 epochs with early stopping based on the model loss on the validation set. Parameter updates were performed once per batch with Adam. The input EEG data matrices are of dimensions $C = 62$ times 400 discretized time samples. Dimensionality of latent z was determined as 1000. We formulate p as a one-hot encoded vector (i.e., a P-dimensional vector with a value of 1 at the s'th index and zero in other indices) and y to be the one-hot encoded emotion class vector. We use both p and y as conditioning parameters for the decoder to enforce the learning of subject- and emotion-dependent generated EEG. We used the TensorFlow libraries with the Keras API.

2) Neural Network Architectures: To measure the effectiveness of our results we measure both signal reconstruction ability and emotion classification performance against a baseline discriminative model. We report the emotion classification results of a discriminative EEGNet [24] model, which we used as the backbone in an accuracy trade-off to also realize generative EEG data modeling.

3) Evaluation: Our experiments evaluate the performance of EEG2Vec in comparison to a state-of-the-art discriminative baseline model in terms of emotion state classification, and also demonstrates its EEG signal reconstruction ability. We initially separate a holdout fixed testing set consisting of 10% of the complete dataset. Then we split the remaining 90% of the data into training and validation sets by 5-fold cross-validation. We ensure that in both training, validation and testing sets, all affective states (i.e., classes) and subjects are represented equally in terms of number of data samples.

V. Experimental Results

A. Learning Deep Latent Representations

We first investigate the structuring of the learned embedding. Figure 2 visualizes the 1000-dimensional learned embedding z into a two-dimensional scatterplot via t-distributed Stochastic Neighbor Embedding (t-SNE) [54]. All tSNE visualizations are generated with default parameters: perplexity $= 30$, number of iterations $= 1000$. We observe a discriminative pattern of different emotion types in z as the EEG data of positive emotions is found prevalently in the left half of the scatterplot. This indicates that the encoder can embed affective state information in the latent representation, and the auxiliary classifier can predict emotions easier from this representation (see Section V-B). Thus z incorporates specific information about the emotion category. However, we also observe that data from negative and neutral emotion class are overlapping in Figure 2, which is further investigated in Section V-B and mainly due to harder discriminability [55] and changing spatio-temporal patterns for these emotions affecting the input EEG.

Figure 3 demonstrates similar t-SNE embeddings of z separated by the participants. We observe no particular global

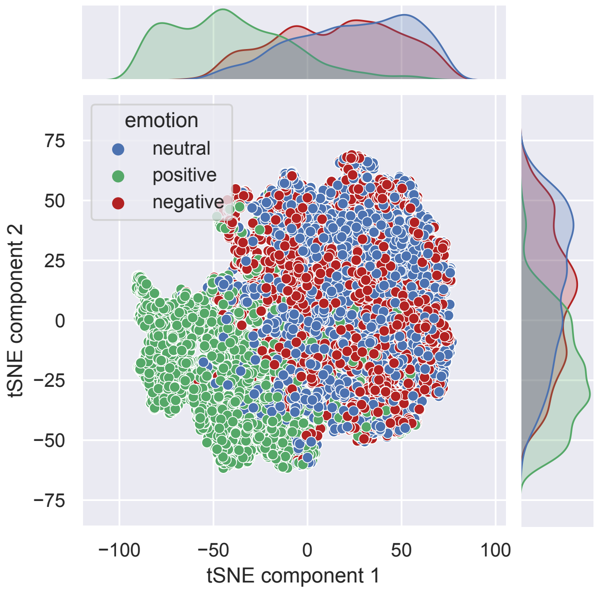


Fig. 2: Latent space visualization of learned representation z using a t-distributed Stochastic Neighbor Embedding (t-SNE) with two components. We used the encoder network of the EEG2Vec model ($\beta = 1$) to transform all observations from the validation set to z.

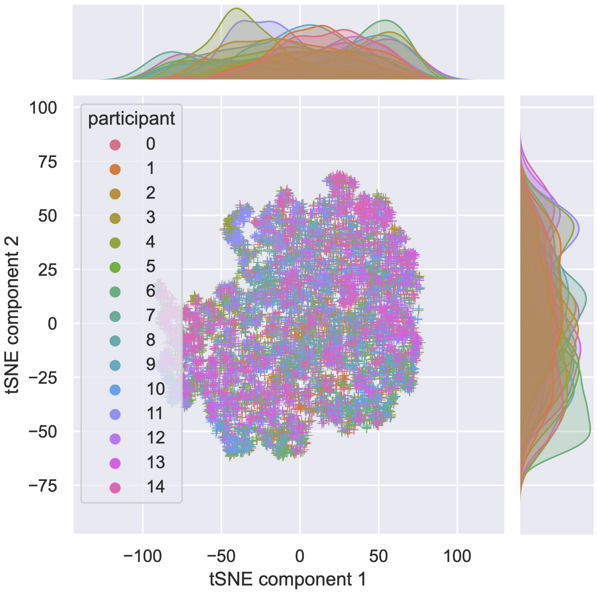


Fig. 3: Latent space visualization of learned representation z using a t-distributed Stochastic Neighbor Embedding (t-SNE) with 2 components. Colors represent different participant IDs.

clustering in z based on the participant. This is an expected behavior as the participant ID is provided to the decoder network as an additional input besides z and hence z is expected to be invariant to participant IDs.

Participant-dependent affective state information is encoded into a low-dimensional space with 1000 dimensional latent variables, which is only 4.03% of the original EEG data size (62×400). This compressed representation can thereby be efficiently used for EEG signal processing with low computational cost and memory requirements as demonstrated practically with similar techniques in [56], [57].

B. Emotion Classification from EEG

Within our multi-outcome EEG2Vec framework (both reconstructing EEG and predicting emotional labels), we weigh the importance of signal reconstruction and emotion classification equally via setting λ appropriately. We train our model as described in Section IV-C. We demonstrate the differences in emotion classification across individual participants in Figure 4.

Our model is able to achieve a 68.49% testing classification accuracy (significantly above three-class decoding chance-level accuracy with $p = 10^{-4}$ using a Wilcoxon signed rank test), and reliably predict positive emotions with a very low false-prediction error rate: very high precision of 82% and recall of 89%. Figure 4 further supports the existence of a significantly larger prediction performance for positive affective state classification (i.e., green curve) on an individual participant-level as well. These results are computed on the basis of a latent representation invariant of participant information and thereby represent a participant-independent emotion recognition performance.

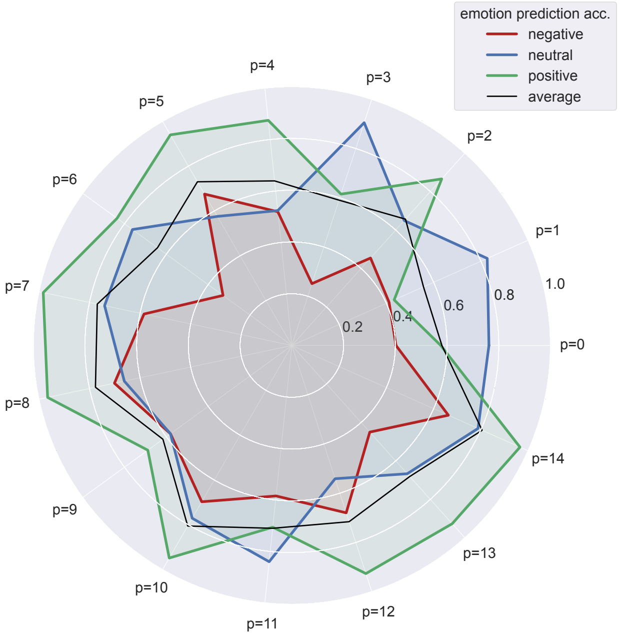


Fig. 4: Emotion prediction accuracy of EEG2Vec for different participants in the testing set. Different colored polar-plots represent the achieved accuracy of the prediction of the specific emotion class. Black curve indicates the mean accuracy per participant over all emotional classes.

On the other hand, the fully-discriminative EEGNet baseline model is able to achieve an affective state prediction accuracy of 77.27% on the same testing set. The increase in classification accuracy with EEGNet can be simply explained

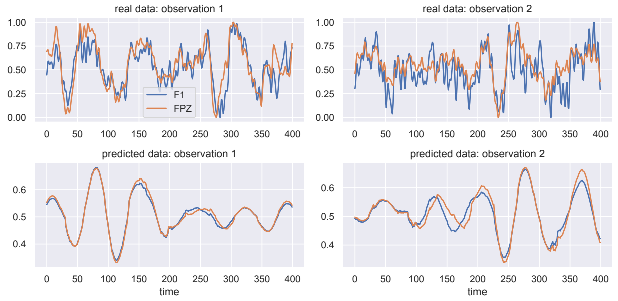


Fig. 5: Real sample versus generated time-series EEG data for two EEG channels F1 and FPZ. Observation 1 has positive affective state, whereas observation 2 is of negative affective state, both from the same participant.

by using a sole model objective of deterministic affective state classification, versus learning robust variational embeddings for emotion prediction and EEG data augmentation simultaneously as in our EEG2Vec model. Compared to the feature-optimized emotion detectors such as [13] or more general deep learning approaches such as EEGNet, EEG2Vec compensates a slightly lower classification accuracy while being able to provide a low-dimensional, affect-distinctive representation. However our model classification performance converges to the overall EEGNet accuracy as λ increases, i.e., weighing a stronger classification performance more than generative capabilities of the model.

C. *Condition-Specific Artificial EEG Synthesis*

Figure 5 visualizes generated EEG examples, where sampled latent representations z are exploited with the decoder network for reconstruction. To further validate the usefulness of the synthetic EEG data, we employ an evaluation of an emotion classification on original and synthetic data in Table I. In the analysis of the prediction performance, we observe that original data merged with 20% synthetic data from our model can improve classification accuracy by several percentage points (from 66% to 69%) vs. only using the original data. We detect an overall modest increase in many participants' classification accuracy with the synthetically boosted model (20% synthetic data), however we see the largest increase in emotion recognition increased by 42.85% by participant 12's and 31.57% when looking at participant 11, while the largest decrease in accuracy of -17.02% is observed when decoding the data of participant 4.

High increases in performances for specific subjects using EEG2Vec's synthetic data augmentation technique are likely due to the subject- and session-variant nature of EEG data which is learned by our model. This phenomenon is similar to the work by [36], which showed that generating supplementary synthetic EEG improved steady state visually evoked potential classification across subjects. We opted against using a higher percentage of synthetic data since the models can then easily overfit to the distribution of the synthetic data. Overall, our results show, that EEG2Vec's synthetic EEG data is beneficial to overall increase the detection ability of newly trained models and can boost the emotion recognition performance of particular subjects.

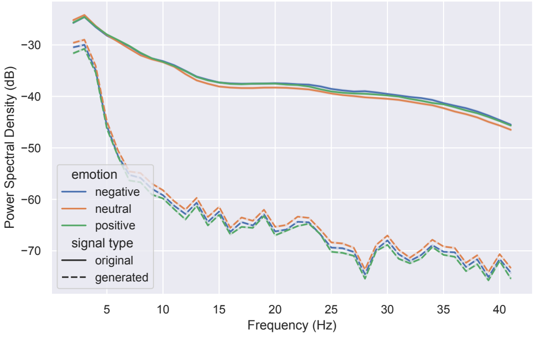


Fig. 6: Mean PSD of original and generated EEG data (channel: F3, validation set) with EEG2Vec. Affective states are depicted in different colors. Low frequency can be generated accurately, while higher frequency signals cannot be appropriately synthesized as can be seen from the diverging lines of reconstructed and original PSD-Frequency lines.

TABLE I: Evaluation of synthetic EEG2Vec data that is additionally used for classification of emotional states.

model	**accuracy**	F_1	**precision**	**recall**
			negative / neutral / positive	
original	.66	.66	.55/.60/.83	.52/.63/.82
original + 5% synthetic data	.68	.68	.53/.56/.82	.65/.56/.82
original + 20% synthetic data	.69	.65	.51/.57/.71	.68/.57/.71

Our architecture is able to generate visually-representative EEG data with relevant low-frequency components. Looking at the power-spectral density (PSD) plot in Figure 6, the reconstruction ability for lower frequency ($\leq 10Hz$) of the original EEG signal is good, because at a low-frequency of $1Hz$ both reconstructed and original signal starting with a PSD of -25 to -30 dB followed by the same downward PSD curve trend. Here, the mean PSD is calculated by averaging over all observation EEG sequences the Welch's power-spectral density results with frequency min/max of 2/41 Hz, FFT length of 200 and 50 observations overlap. However, higher frequency ($> 10Hz$) cannot be learned appropriately as the PSD values differ heavily (generated PSD values range from -60 to -75 dB whereas the original signal are from -35 to -40 dB). This can be partly explained by the up-sampling components of the generator network which introduce aliasing frequency artifacts [32] as well as noise artifacts in the inputs are not captured.

D. *Effects of Disentangling in Latent Space*

Lastly, we examine reconstruction performance of EEG2Vec by varying the KL-divergence factor β while regularizing the latent space. With increasing β, the factors of the learned embedding are directed towards a more disentangled form, i.e., statistical independence than on reconstruction, which results in more uncorrelated latent features with less reconstruction ability. If each variable in the inferred latent representation z is only sensitive to one single generative factor and relatively invariant to other factors, we will say this representation is disentangled. Having disentangled representations can help to reduce information overlay from

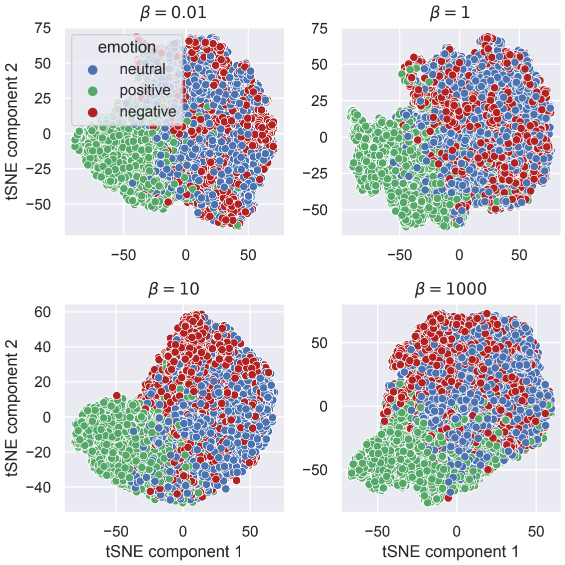


Fig. 7: t-SNE visualizations of latent embeddings z for varying latent space regularization weights β.

different factors, and promote better interpretability and easier generalization to a variety of tasks [45].

For all β values depicted in Figure 7, z is showing positive emotions in a discriminative region. We also observe that with higher β (lower plots) negative and neutral affective state z are showing more discriminative clusters which are depicted more clearly in different subspaces. Results indicate that higher disentanglement can favor higher distinction between affective states of the clusters.

VI. Discussion

We propose EEG2Vec as a mechanism to learn latent representations of affective EEG data that allow for general use in various generative and discriminative machine learning paradigms. Our model learns vectorized representations (i.e., *embeddings*) of EEG responses to emotional videos that are discriminative of the affective states, as well as sufficiently representative to generate synthetic EEG data. In doing so, learned *embeddings* can also be used to generate synthetic EEG data that is both participant- and emotion-specific, simply by sampling from the latent state probability function. Our results altogether show that the proposed architecture is able to learn both efficient (lower dimensionality with z) and expressive (able to maintain useful properties with z) representations. One important limitation of our approach lies on the accessible training dataset infrastructure. It is naturally likely that the amount of participant-specific data can impact optimization if not accounted for. So far we only considered learning from a balanced training data set in terms of participant IDs and class labels by stratifying our available training set size. Nevertheless the proposed EEG2Vec pipeline with sufficient amount of data allows future research to exploit low-rank EEG representations with less memory demand for general purpose edge applications (e.g., wearable computing [58], [59] or human-robot interaction [60]).

Acknowledgments

This research was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in TRR161 (Quantitative methods for visual computing, Project ID 251654672) in Project C06. LLC is supported by the research initiative “Instant Teaming between Humans and Production Systems” co-financed by tax funds of the Saxony State Ministry of Science and Art (SMWK3-7304/35/3-2021/4819).

References

[1] M. Tielman *et al.*, “Adaptive emotional expression in robot-child interaction,” in *9th ACM/IEEE International Conference on Human-Robot Interaction*, 2014, pp. 407–414.

[2] T. Zhang *et al.*, “RCEA: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–15.

[3] M. Hassib *et al.*, “Detecting and influencing driver emotions using psycho-physiological sensors and ambient light,” in *IFIP Conference on Human-Computer Interaction*, 2019, pp. 721–742.

[4] X. Hu *et al.*, “Ten challenges for EEG-based affective computing,” *Brain Science Advances*, vol. 5, no. 1, pp. 1–20, 2019.

[5] H. Fei *et al.*, “Latent emotion memory for multi-label emotion classification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7692–7699.

[6] Y.-P. Ruan and Z. Ling, “Emotion-regularized conditional variational autoencoder for emotional response generation,” *IEEE Transactions on Affective Computing*, 2021.

[7] T. Mikolov *et al.*, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[8] J. Williams *et al.*, “Exploring disentanglement with multilingual and monolingual vq-vae,” *arXiv preprint arXiv:2105.01573*, 2021.

[9] J. Devlin *et al.*, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[10] T. B. Brown *et al.*, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.

[11] Z. Yang *et al.*, “Improved variational autoencoders for text modeling using dilated convolutions,” in *International Conference on Machine Learning*, 2017, pp. 3881–3890.

[12] S. Saha *et al.*, “Progress in brain computer interfaces: challenges and trends,” *arXiv preprint arXiv:1901.03442*, 2019.

[13] W.-L. Zheng and B.-L. Lu, “Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks,” *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[14] R. G. Lupu *et al.*, “Brain-computer interface: Challenges and research perspectives,” in *22nd International Conference on Control Systems and Computer Science*, 2019, pp. 387–394.

[15] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[16] A. Al-Nafjan *et al.*, “Review and classification of emotion recognition based on EEG brain-computer interface system research: a systematic review,” *Applied Sciences*, vol. 7, no. 12, p. 1239, 2017.

[17] K. Takahashi and A. Tsukaguchi, “Remarks on emotion recognition from multi-modal bio-potential signals,” in *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, vol. 2, 2003, pp. 1654–1659.

[18] C. Tang *et al.*, “EEG-based emotion recognition via fast and robust feature smoothing,” in *International Conference on Brain Informatics*, 2017, pp. 83–92.

[19] O. Özdenizci and D. Erdoğmuş, “Information theoretic feature transformation learning for brain interfaces,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 69–78, 2019.

[20] Y. Liu and O. Sourina, “Real-time fractal-based valence level recognition from EEG,” in *Transactions on Computational Science XVIII*, 2013, pp. 101–120.

[21] Y.-P. Lin *et al.*, “EEG-based emotion recognition in music listening,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[22] L.-C. Shi *et al.*, “Differential entropy feature for EEG-based vigilance estimation,” in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 6627–6630.

[23] M. Akin, "Comparison of wavelet transform and FFT methods in the analysis of EEG signals," *Journal of Medical Systems*, vol. 26, no. 3, pp. 241–247, 6 2002.

[24] V. J. Lawhern *et al.*, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[25] Y. Li *et al.*, "From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition," *IEEE Transactions on Affective Computing*, 2019.

[26] P. Zhong *et al.*, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, 2020.

[27] D. Bethge *et al.*, "Domain-invariant representation learning from EEG with private encoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1236–1240.

[28] ——, "Exploiting multiple EEG data domains with adversarial learning," in *44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2022.

[29] A. Craik *et al.*, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, 2019.

[30] O. Özdenizci and D. Erdoğmuş, "On the use of generative deep neural networks to synthesize artificial multichannel EEG signals," in *10th International IEEE/EMBS Conference on Neural Engineering*, 2021, pp. 427–430.

[31] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[32] K. G. Hartmann *et al.*, "EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals," *arXiv preprint arXiv:1806.01875*, 2018.

[33] F. Fahimi *et al.*, "Towards EEG generation using GANs for BCI applications," in *IEEE EMBS International Conference on Biomedical & Health Informatics*, 2019, pp. 1–4.

[34] Y. Luo and B.-L. Lu, "EEG data augmentation for emotion recognition using a conditional wasserstein GAN," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2535–2538.

[35] X. Li *et al.*, "Variational autoencoder based latent factor decoding of multichannel EEG for emotion recognition," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2019, pp. 684–687.

[36] N. K. N. Aznan *et al.*, "Simulating brain signals: Creating synthetic EEG data via neural-based generative models for improved SSVEP classification," in *International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[37] P. Bashivan *et al.*, "Learning representations from EEG with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.

[38] W. Ko *et al.*, "Multi-scale neural network for EEG representation learning in BCI," *IEEE Computational Intelligence Magazine*, vol. 16, no. 2, pp. 31–45, 2021.

[39] Y. Luo *et al.*, "Data augmentation for enhancing EEG-based emotion recognition with deep generative models." *Journal of Neural Engineering*, vol. 17, no. 5, pp. 056 021–056 021, 2020.

[40] O. Özdenizci *et al.*, "Transfer learning in brain-computer interfaces with adversarial variational autoencoders," in *9th International IEEE/EMBS Conference on Neural Engineering*, 2019, pp. 207–210.

[41] Y. Li *et al.*, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 494–504, 2021.

[42] O. Özdenizci *et al.*, "Learning invariant representations from EEG via adversarial inference," *IEEE Access*, vol. 8, pp. 27 074–27 085, 2020.

[43] ——, "Adversarial deep learning in EEG biometrics," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 710–714, 2019.

[44] K. Sohn *et al.*, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[45] I. Higgins *et al.*, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017.

[46] C. P. Burgess *et al.*, "Understanding disentangling in beta-VAE," *arXiv preprint arXiv:1804.03599*, 2018.

[47] M. Lopez-Martin *et al.*, "Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot," *Sensors*, vol. 17, no. 9, p. 1967, 2017.

[48] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[49] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, pp. 1–13, 2018.

[50] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[51] J. A. Miranda-Correa *et al.*, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2018.

[52] C. Qing *et al.*, "Interpretable emotion recognition using EEG signals," *IEEE Access*, vol. 7, pp. 94 160–94 170, 2019.

[53] H. Candra *et al.*, "Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 7250–7253.

[54] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[55] W. Liu *et al.*, "Multimodal emotion recognition using multimodal deep learning," *arXiv preprint arXiv:1602.08225*, 2016.

[56] S. Li *et al.*, "A continuous biomedical signal acquisition system based on compressed sensing in body sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 3, pp. 1764–1771, 2013.

[57] Y. Shin *et al.*, "Simple adaptive sparse representation based classification schemes for EEG based brain–computer interface applications," *Computers in Biology and Medicine*, vol. 66, pp. 29–38, 2015.

[58] M. Han *et al.*, "Disentangled adversarial autoencoder for subject-invariant physiological feature extraction," *IEEE Signal Processing Letters*, vol. 27, pp. 1565–1569, 2020.

[59] ——, "Universal physiological representation learning with soft-disentangled rateless autoencoders," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 2928–2937, 2021.

[60] O. Özdenizci *et al.*, "Hierarchical graphical models for context-aware hybrid brain-machine interfaces," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1964–1967.

ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747398

DOMAIN-INVARIANT REPRESENTATION LEARNING FROM EEG WITH PRIVATE ENCODERS

David Bethge[1,2], *Philipp Hallgarten*[1,3], *Tobias Grosse-Puppendahl*[1], *Mohamed Kari*[1],
Ralf Mikut[3], *Albrecht Schmidt*[2], *Ozan Özdenizci*[4,5]

[1] Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany
[2] Ludwig-Maximilians University Munich, Germany
[3] Karlsruhe Institute of Technology, Germany
[4] Institute of Theoretical Computer Science, Graz University of Technology, Austria
[5] TU Graz - SAL Dependable Embedded Systems Lab, Silicon Austria Labs, Austria

ABSTRACT

Deep learning based electroencephalography (EEG) signal processing methods are known to suffer from poor test-time generalization due to the changes in data distribution. This becomes a more challenging problem when privacy-preserving representation learning is of interest such as in clinical settings. To that end, we propose a multi-source learning architecture where we extract domain-invariant representations from dataset-specific private encoders. Our model utilizes a maximum-mean-discrepancy (MMD) based domain alignment approach to impose domain-invariance for encoded representations, which outperforms state-of-the-art approaches in EEG-based emotion classification. Furthermore, representations learned in our pipeline preserve domain privacy as dataset-specific private encoding alleviates the need for conventional, centralized EEG-based deep neural network training approaches with shared parameters.

1. INTRODUCTION

Over the past decades, electroencephalography (EEG) based emotion recognition gained significant interest towards developing affective neural-machine interfaces [1]. However, due to the costly data collection processes, large EEG datasets recorded under emotion eliciting experimental paradigms are not ubiquitous. Several small scale affective EEG datasets were previously introduced, which however highly differ in their experimental setups (*e.g.*, stimuli, emotion labels). This introduces the well-known *domain adaptation* problem, i.e., models trained to recognize emotions on one of these datasets commonly fails to solve the same task for another dataset. In order to make EEG-based emotion recognition algorithms suitable for a variety of experiments and real-life scenarios, it is crucial to obtain a model that can tackle this task across multiple datasets. One approach to achieve this from a deep learning perspective is to extract and exploit domain-invariant representations from multi-channel EEG data.

Recent work have shown significant promise in using invariant representation learning from time-series EEG data for cross-subject generalization [2–4], while learning across multiple dataset sources remains an open question. Notable methods successfully introduced adversarial censoring for domain invariance, widely known as DANN [5], in various EEG decoding tasks [2, 3], including emotion recognition [4]. Adversarial domain regularization with DANN [5] considers a cross-domain shared encoder that extracts features from data of all subjects, and two classifiers: a task and a domain classifier. During training the encoder is adversarially penalized on the domain classification loss, which enforces the model towards learning domain-invariant representations. In this work we take a different approach to invariant EEG representation learning by further considering to preserve domain privacy that is of critical importance in clinical settings [6, 7].

We propose a multi-source learning framework for domain invariant representation learning from time-series signals such as multi-channel EEG recordings. From a different perspective than adversarial learning methods, our framework consists of a private feature encoder per domain and a cross-domain shared classifier, where we utilize a maximum-mean-discrepancy (MMD) [8] based domain alignment loss across private feature encoders to minimize domain-specific leakage within the learned representations. Our contributions in this work are three-fold: (1) We introduce a deep neural signal processing architecture where EEG time-series are privately encoded at the source end and only the learned representations are shared to a global classifier that enables decentralized cross-dataset learning by preserving domain privacy. (2) We reveal large dataset domain specific variances in conventionally trained centralized pipelines, and demonstrate that regularizing latent representations via an MMD-based domain alignment loss enables dataset source independent representation learning. (3) We show that the use of adaptive batch normalization layers in such multi-source settings prevent performance decrease at inference time.

2. MATERIALS AND METHODS

2.1. Experimental Data

Benchmark Affective EEG Datasets: We used four publicly available EEG-based emotion recognition datasets: *DEAP* [9], *DREAMER* [10], *SEED* [11,12], *SEED-IV* [13]. During EEG recordings subjects were presented audio-visual stimuli that are expected to elicit distinct emotions. Differences in the experimental setups lead to a variability in the structure of data samples and their labels. *SEED* and *SEED-IV* contain 62-channel EEG recordings sampled at 200 Hz, whereas *DEAP* contains recordings from 32 electrodes and *DREAMER* from 14, both sampled at 128 Hz. Emotion labels of *SEED* and *SEED-IV* represents one out of three and four discrete emotions respectively, labeled by the stimuli that the subjects were presented. For *DEAP* and *DREAMER* post-recording subject self-assessment ratings on the *valence* and *arousal* continuous scales were included as the labels.

Label Transformation: To realize our cross-dataset experimental analyses, we transformed the label spaces across all four datasets into a common set. We determined the three discrete emotions *Negative*, *Neutral*, and *Positive* (as also used in the *SEED*-dataset) as the common label-space. Using *k-means* clustering in the two-dimensional label space of *Valence* and *Arousal* for *DEAP*, and *DREAMER* datasets, we determined four clusters corresponding to the four discrete emotions used as labels in the *SEED-IV* dataset: *Fear*, *Sad*, *Neutral*, *Happy*. We merged the former two into being the *Negative* class, and considered *Happy* to be the *Positive* class. Hence we obtained three class labels in a unified manner.

2.2. Multi-Source EEG Processing for Private Encoding

We will denote multiple datasets $\mathcal{D}_k$ consisting of time-series EEG signal epochs and corresponding emotion label of the experiment paradigm as pairs (s_j, l_j). We pre-process the signals to serve as input to our architecture as follows. First, we perform baseline correction for signals s_j using a three-second time-window and apply a $4-40$ Hz Butterworth band-pass filter. Since the length of the available s_j differ within each dataset, we used only the last T seconds of each available signal with T being the length of the shortest time-series in the dataset, i.e., 60, 64, 185, 50 for *DEAP*, *DREAMER*, *SEED* and *SEED-IV* respectively. Then, we segment the signals into non-overlapping 2 s windows that will be used as inputs x_i to our model. We assign the label l_{j_i} to each window x_i. All pairs of x_i and corresponding labels y_i obtained from a dataset are considered as a processed data-source $\mathcal{X}_k$.

To balance the number of samples across data-sources and with respect to the label for each data-source individually, we applied stochastic undersampling while constructing our training set. We split the data into the training (60%), validation (20%) and test (20%) sets, while we assure that the stratification constraints hold for each subset individually.

2.3. Multi-Source Learning Framework

Domain Aligned Private Encoders (DAPE): Our model consist of one private encoder per data-source and a shared classifier. While the encoder can be designed arbitrarily for each data-source, for proof of concept we used the *DeepConvNet* architecture [14] for our all data-sources. We replaced the final pooling layer in *DeepConvNet* by an adaptive average pooling layer, that introduces the flexibility to control the number of features n_z in the latent representation z. In our models we chose $n_z = 50$ as the best results in terms of the balance between emotion classification and domain invariance were achieved. One extension of our model towards achieving better domain adaptation, which we denote by adaptive DAPE (aDAPE), considers the use of adaptive batch normalization layers [15]. Such layers were also recently used in deep EEG analysis by [16, 17], instead of conventional batch normalization. By doing so, we ensure each layer of any given encoder to receive data from a similar distribution at test time as well, as samples are normalized by the statistics estimated from the utilized mini-batch.

During each forward pass, a batch of B samples from $x_i \sim \mathcal{X}_i$ are drawn and used as input to each of the M encoders, represented through the parameters $\theta_i, i \in \{1, \ldots, M\}$, with M also being the number of data-sources. We consider the output representations $z_i = f(\theta_i, x_i) \in \mathbb{R}^{B \times n_z}$ of each encoder as samples drawn from independent data-source distributions. To overcome the problem of domain shift across encoders and regularize the encoders towards learning domain-invariant representations, we use an MMD [8] based domain-alignment loss $\mathcal{L}_{\text{DA}}$ using Gaussian kernels with $\sigma_j \in \{10, 15, 20, 50\}$. Rather than computing the MMD for all combinations of data-sources, we randomly sample M pairs of distributions $\{(z_p^{(i)}, z_q^{(i)})\}_{i=1}^{M}, p, q \in \{1, \ldots, M\}$ such that $p \neq q$, calculate the MMD loss [8] between them and accumulate in the domain alignment regularizer:

$$\mathcal{L}_{DA} = \sum_{j=1}^{|\sigma|} \sum_{i=1}^{M} \text{MMD}^2(z_p^{(i)}, z_q^{(i)}, \sigma_j), \tag{1}$$

where $|\sigma|$ denotes the cardinality for the pre-defined set σ_j is chosen from and $\text{MMD}^2(z_p^{(i)}, z_q^{(i)}, \sigma_j)$ is defined as follows:

$$\begin{aligned} \text{MMD}^2 = & \frac{1}{B(B-1)} \sum_{m=1}^{B} \sum_{\substack{n=1 \\ n \neq m}}^{B} \varphi(z_{p_m}^{(i)}, z_{p_n}^{(i)}, \sigma_j) \\ & + \frac{1}{B(B-1)} \sum_{m=1}^{B} \sum_{\substack{n=1 \\ n \neq m}}^{B} \varphi(z_{q_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j) \\ & - \frac{2}{B^2} \sum_{m=1}^{B} \sum_{n=1}^{B} \varphi(z_{p_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j), \end{aligned} \tag{2}$$

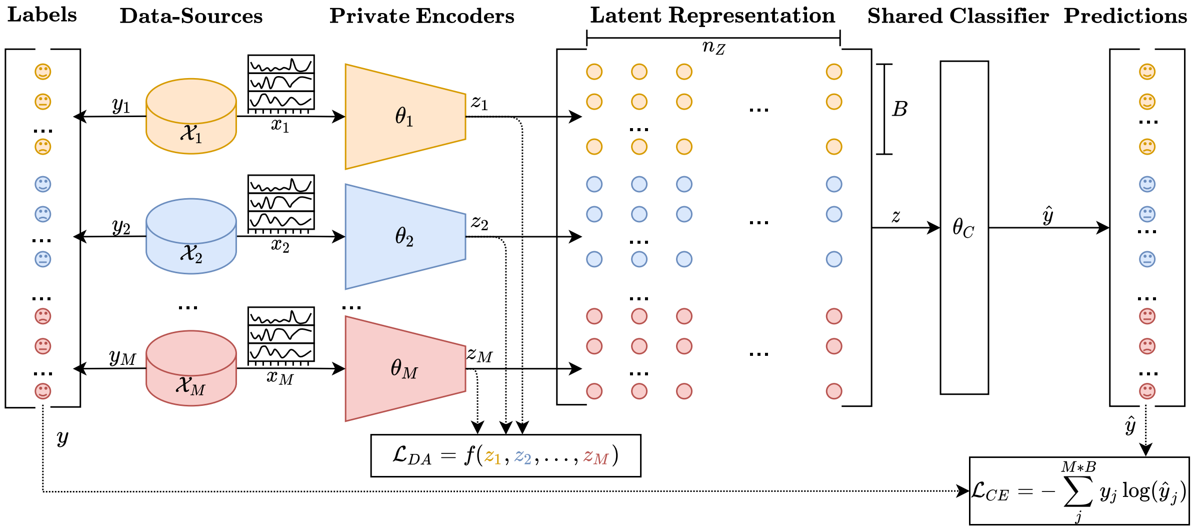


Fig. 1. **DAPE Architecture.** Each color represents one data-source. Signals x_1 from each data-source $\mathcal{X}_i$ are processed by a private encoder. Output representations z_i are then used to calculate the domain-alignment loss $\mathcal{L}_{DA}$ via Eq. (1). In addition, output representations are concatenated to a representation $z \in \mathbb{R}^{(B*M)\times n_z}$ and used as input for a shared classifier. Classifier outputs a batch of predictions $\hat{y}_i$ for each batch of signals x_i. Corresponding batches of labels y_i are concatenated in the same way as the latent representations z_i and used together with the predictions for the calculation of the cross-entropy loss $\mathcal{L}_{CE}$.

with $\varphi(z_{p_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j)$ being the Gaussian kernel function:

$$\varphi(z_{p_m}^{(i)}, z_{q_n}^{(i)}, \sigma_j) = \exp\left(\frac{-\|z_{p_m}^{(i)} - z_{q_n}^{(i)}\|^2}{2\sigma_j^2}\right). \qquad (3)$$

All batches of representations z_i are then stacked to one vector of representations $z \in \mathbb{R}^{(B*M)\times n_z}$, and used as input for the classifier with parameters θ_C to compute predicted emotions $\hat{y}$. We train the network with a categorical cross-entropy loss function $\mathcal{L}_{CE} = -\sum_j^{M*B} y_j \log(\hat{y}_j)$. Note that only the parameters θ_C and θ_i (parameters of the encoder that processed a sample x_j) contribute to the prediction of an individual sample $\hat{y}_j$, which ensures that only the classifier and the encoder i are optimized based on the gradient of the loss.

Multi-Objective Optimization: Training of *DAPE* can be considered as a multi-objective optimization. On one hand, we want to minimize the cross-entropy loss for the classification task and one the other hand we want to minimize the domain-alignment loss $\mathcal{L}_{\mathrm{DA}}$ for learning domain-invariant representations. We assume that these goals are not contradictory and that the classification task gets more robust using domain-invariant representations of the samples. To control the domain-invariance of representations we use a hyperparameter κ as the tunable weight in the optimization problem:

$$\theta_1^*, \ldots, \theta_M^*, \theta_C^* = \underset{\theta_1,\ldots,\theta_M,\theta_C}{\arg\min}\ \mathcal{L}_{CE} + \kappa\mathcal{L}_{DA}. \qquad (4)$$

We use an annealing update scheme for κ as proposed by [18, 19]. To ensure that the domain alignment does not lead the encoders to learn task-irrelevant representations, we start the training with $\kappa = 0$ and increase it with the beginning of the 5th epoch by a rate of 0.25 per epoch. To prevent the domain-alignment loss dominating the cross-entropy loss in the later epochs, we stop updating κ by the 70th epoch and retain the value of 16.25 for all subsequent epochs.

2.4. Quantifying Domain-Invariance of Representations

One of the motivating goals in our architecture is to preserve domain privacy at the data-source end, as well as ensuring that an attacker not being able to deduce the information concerning the data-source from latent representations. To assess invariance of learned representations via DAPE, we firstly compute the latent representations z_i of all samples in the test set. We use 80% of the samples to train a *domain classifier* using the data-source ID as label. We subsequently evaluate the domain-invariance of the representations by calculating the achieved accuracies of these domain classifiers on the remaining 20% of the test set representations. Note that a lower 4-class domain-classifier accuracy corresponds to a higher domain-invariance of the learned representations, which is favorable for multi-source learning. We considered a multitude of learning machines as domain classifiers (e.g., support vector machines (SVM), quadratic discriminant analysis). Since all domain classifiers showed similar results we report our results with linear SVMs in Section 3.2.

Table 1. Accuracies of 3-class emotion classification evaluated with *DAPE*, *aDAPE* and baseline methods: local and global baseline models, and *DANN* architecture. Bottom row shows the results of the linear SVM used to assess the 4-class domain-invariance. Local baseline corresponds to using four independent networks (i.e., dataset-specific encoder and classifiers) without any regularization or domain-alignment.

		Local Models	Global Model	*DANN* [5]	*DAPE* (**Ours**)	*aDAPE* (**Ours**)
Emotion Classification (↗)	*DEAP*	50.01% (±0.58)	40.02% (±1.09)	39.81% (±0.87)	47.81%(±0.24)	47.99%(±1.30)
	DREAMER	59.03% (±2.41)	43.78% (±1.28)	42.30% (±1.60)	49.13%(±1.54)	48.70%(±1.99)
	SEED	56.19% (±1.97)	42.93% (±0.95)	42.63% (±0.26)	51.72%(±0.67)	53.45%(±2.05)
	SEED-IV	42.42% (±3.64)	34.55% (±0.78)	34.57% (±0.56)	41.94%(±0.64)	43.25%(±1.07)
	Mean	51.91% (±2.15)	40.32% (±0.42)	39.82% (±0.33)	47.65%(±0.13)	**49.09**%(±1.49)
Domain Classification (↘)	**Mean**	99.87% (±0.15)	60.11% (±3.47)	66.17%(±4.45)	82.83%(±3.18)	**52.76**%(±2.92)

3. EXPERIMENTAL RESULTS

3.1. Emotion Classification

Table 1 summarizes our results on the emotion classification task. Here *local models* indicate four independent networks processed on each data-source individually (i.e., *DAPE* with $M = 1$), which would be a baseline approach if one is not interested in domain-privacy or representation invariance. To the contrary, the *global model* indicates a baseline where one trains a single, unified model by pooling all data-source inputs into one unified training set, however constraining both the amount of training data and the encoder specifications to be uniform and shared across all data-sources. DANN [5] depicts an invariant representation learning approach that was previously considered for EEG decoding in [2, 4], where one can train a model adversarially by censoring the data-source ID relevant information from latent representations to impose invariance, without considering data-source privacy.

While the global model and DANN perform only slightly above chance level (33%) in emotion classification, we observe that *DAPE* and *aDAPE* clearly outperform the baseline approaches that learns a unified model for emotion classification, while simultaneously ensuring data-source privacy. Furthermore, we observe that *aDAPE* is also only slightly below the average of the extensive local models (49.09% vs. 51.91%), although solving this task across multiple domains poses a more challenging problem than learning locally.

3.2. Domain-Invariance of the Learned Representations

Domain-invariance of the learned representations estimated through the linear SVM are shown at the bottom row of Table 1. We observe that *aDAPE* shows the highest domain-invariance in the learned representations, where in all other models the linear SVM was able to deduce the data-source ID from the representations with $> 60\%$ accuracy. Achieved domain-invariance is $\sim$ 14% higher for *aDAPE* than for *DANN*, and $\sim$ 47% higher than for the local models. Furthermore, it is important to note that the performance of the domain classifier is much higher ($\sim$ 30%) for *aDAPE* than for *DAPE*, from an ablation study perspective. This significant gap results in the use of adaptive batch normalization layers. Overall emotion classification accuracy of *aDAPE* is higher than of *DAPE*, and achieved domain-invariance of representations is also stronger (while still being slightly above the chance-level), confirming our hypothesis that using domain-invariant representations in multi-source settings helps by making the classification task more robust.

4. CONCLUSION

We present *Domain Aligned Private Encoders* as a multi-source learning framework for domain-invariant representation learning, and demonstrate its feasibility on an EEG-based emotion classification task using data from four publicly available datasets. Different than state-of-the-art adversarial training approaches to learn invariant EEG representations [2–4], we utilize an MMD based domain alignment loss [16, 20] across dataset-specific private feature encoders. Proposed deep neural signal processing architecture encodes multi-channel EEG time-series signals privately at the source end and only the learned representations are shared to a global classifier. Our decentralized invariant representation learning approach and use of adaptive batch normalization layers improved performance in our experimental analyses.

Accessible EEG data is generally present in small datasets, which are also distributed across various countries and/or laboratories. Due to naturally occurring data privacy concerns, we believe that such architectures as proposed in this work make common use of private data representations more convenient and real-life applicable (e.g., clinical time-series monitoring [21], personalized BCI-based stroke rehabilitation protocols [6, 22]), hence our work advances the area of cross-domain transfer learning for EEG signals in that sense.

5. REFERENCES

[1] Christian Mühl et al., "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[2] Ozan Özdenizci et al., "Learning invariant representations from EEG via adversarial inference," *IEEE Access*, vol. 8, pp. 27074–27085, 2020.

[3] Eunjin Jeon et al., "Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[4] Soheil Rayatdoost et al., "Subject-invariant EEG representation learning for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 3955–3959.

[5] Yaroslav Ganin et al., "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[6] Anisha Agarwal et al., "Protecting privacy of users in brain-computer interface applications," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 8, pp. 1546–1555, 2019.

[7] Ce Ju et al., "Federated transfer learning for EEG signal classification," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, 2020, pp. 3040–3045.

[8] Arthur Gretton et al., "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[9] Sander Koelstra et al., "DEAP: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[10] Stamos Katsigiannis and Naeem Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[11] Ruo-Nan Duan et al., "Differential entropy feature for EEG-based emotion classification," in *6th International IEEE/EMBS Conference on Neural Engineering*, 2013, pp. 81–84.

[12] Wei-Long Zheng and Bao-Liang Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[13] Wei-Long Zheng et al., "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[14] Robin Tibor Schirrmeister et al., "Deep learning with Convolutional Neural Networks for EEG Decoding and Visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[15] Yanghao Li et al., "Revisiting batch normalization for practical domain adaptation," *ICLR Workshops*, 2017.

[16] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil, "Custom domain adaptation: A new method for cross-subject, EEG-based cognitive load recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 750–754, 2020.

[17] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil, "Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition," *Pattern Recognition Letters*, vol. 141, pp. 54–60, 2021.

[18] Edgar Schonfeld et al., "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8247–8255.

[19] Samuel R Bowman et al., "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[20] Hao Chen et al., "MS-MDA: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition," *arXiv preprint arXiv:2107.07740*, 2021.

[21] Ozan Özdenizci et al., "Time-series prediction of proximal aggression onset in minimally-verbal youth with autism spectrum disorder using physiological biosignals," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2018, pp. 5745–5748.

[22] Anastasia-Atalanti Mastakouri et al., "Personalized brain-computer interface models for motor rehabilitation," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2017, pp. 3024–3029.

Exploiting Multiple EEG Data Domains with Adversarial Learning

David Bethge[1,2], Philipp Hallgarten[1,3], Ozan Özdenizci[4,5],
Ralf Mikut[3], Albrecht Schmidt[2], Tobias Grosse-Puppendahl[1]

Abstract— **Electroencephalography (EEG) is shown to be a valuable data source for evaluating subjects' mental states. However, the interpretation of multi-modal EEG signals is challenging, as they suffer from poor signal-to-noise-ratio, are highly subject-dependent, and are bound to the equipment and experimental setup used, (*i.e. domain*). This leads to machine learning models often suffer from poor generalization ability, where they perform significantly worse on real-world data than on the exploited training data. Recent research heavily focuses on cross-subject and cross-session transfer learning frameworks to reduce domain calibration efforts for EEG signals. We argue that multi-source learning via learning domain-invariant representations from multiple data-sources is a viable alternative, as the available data from different EEG data-source domains (e.g., subjects, sessions, experimental setups) grow massively. We propose an adversarial inference approach to learn data-source invariant representations in this context, enabling multi-source learning for EEG-based brain-computer interfaces. We unify EEG recordings from different source domains (i.e., emotion recognition datasets SEED, SEED-IV, DEAP, DREAMER), and demonstrate the feasibility of our invariant representation learning approach in suppressing data-source-relevant information leakage by 35% while still achieving stable EEG-based emotion classification performance.**

Index Terms— **adversarial learning, domain invariance, EEG.**

I. Introduction

Electroencephalogram (EEG) based brain-computer interface (BCI) systems aim to identify users' intentions from brain recordings with potential uses in neurorehabilitation systems [1]. However, moderate decoding accuracies have limited the practical use of BCIs [2], [3]. Due to the high data collection efforts and costs, EEG datasets highly diverge in their recording environment (*e.g.*, stimulus), the equipment and devices, and the ground truths derived. Shortage of large and homogeneous BCI datasets limits the choice of applicable models and causes a high effort if individual models are to be used for each domain. Imbalance of EEG data source domains for classification is therefore prevalent and posing important challenges for EEG-based BCIs.

Transfer learning across different data domains as such has been extensively studied over the past decades in computer vision [4], [5], proposing convolutional neural networks (CNNs) to extract domain-invariant features for image search and classification across domains. Subsequently, transfer learning on neurophysiological recording datasets (*e.g.,* EEG) is becoming an active research field [6]. Generalized neural decoder learning for across recording modalities (multi-corpus) on electrocorticography data has been recently proposed by [7]. Their approach was shown to generalize to new participants and recording modalities, robustly handle variations in electrode placement, and allow participant-specific fine-tuning with minimal data. Also recently, [2] discussed an online pre-alignment strategy for aligning the motor imagery EEG recording distributions of different subjects before training and inference processes, and showed to significantly improve generalization across datasets. Towards a similar goal, [8], [9] proposed an invariant representation learning scheme using adversarial inference to enable cross-nuisance transfer learning in EEG signal classification with deep neural networks. Empirical assessments on EEG decoding tasks revealed that cross-subject [8] or cross-session [9] representations can be learned with such models. Cross-subject EEG transfer learning have been also explored for emotion recognition to generalize existing models to new subjects, and thereby reducing the demand for the calibration data amount effectively for new subjects [10].

In light of recent work on enabling multi-corpus learning from neurophysiological data [2], [11], [12], [13], we propose an adversarial machine learning approach to unify different raw EEG time-series and pre-process them accordingly. Unlike previous work that has focused on learning scenarios across subjects or sessions, we explore dataset-invariant representations via an adversarial learning framework that can be used in EEG multi-label settings. Our approach aims at expressing robust task-relevant EEG features in a latent representation for emotion recognition across several datasets, by limiting the representation to not learn nuisances specific to these datasets, hence being dataset invariant. We evaluate our framework against the competing baseline of a state-of-the-art deep learning encoder-classifier network trained on the unified set of all data sources.

Our contributions in this work are as follows: (1) We present a unifying EEG pre-processing framework for fusing different raw EEG time-series datasets and associated emotion state labels for transfer learning. (2) We propose an adversarial machine learning framework on multivariate EEG time-series to learn dataset-invariant representations to predict EEG class labels. (3) We present an experimental study on assembling four publicly-available EEG datasets in the field of emotion recognition, and show that our approach can learn dataset-invariant representations *i.e.*, transfer emotion-relevant EEG signals across datasets containing data from different subjects and measurement conditions.

[1] Dr. Ing. h.c. F. Porsche AG, Stuttgart, Germany.
[2] Ludwig-Maximilians University, Munich, Germany.
[3] Karlsruhe Institute of Technology, Karlsruhe, Germany.
[4] Institute of Theoretical Computer Science, TU Graz, Austria.
[5] TU Graz-SAL DES Lab, Silicon Austria Labs, Graz, Austria.
Corresponding author: P. Hallgarten (philipp.hallgarten1@porsche.de).

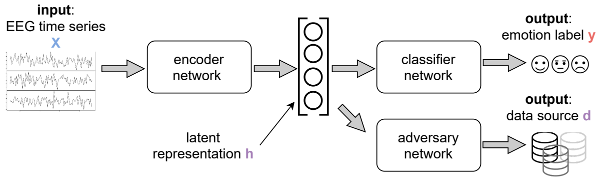


Fig. 1. Overview of the network architecture proposed for adversarial domain adaptation across multiple EEG datasets, consisting of an encoder and two separate dense layer classifiers (*i.e.*, a classifier network for emotions and an adversary network that identifies the EEG data-source ID).

II. Methods

A. Notation and Problem Statement

Let $\{(X_i, y_i, s_i, d_i)\}_{i=1}^n$ denote the data samples consisting observations from a data generation process with $X \sim p(X|y, s, d), y \sim p(y)$, $s \sim p(s)$, and $d \sim p(d)$, where $X_i \in \mathbb{R}^{C_i \times T_i}$ is the raw trial EEG data from data-source d_i during trial i recorded from C_i channels for T_i discretized time samples, y_i is the corresponding emotion label, that can either be a discrete state $y_i \in \{1, .., Y\}$ or a vector $y_i \in \mathbb{R}^Y$ depending on the data-source, and $s_i \in \{1, 2, \dots, S\}$ denotes the subject identification (ID) number for the participant that the trial EEG data is recorded from. Since the subject IDs and emotional labels are defined and used differently within different data-sources, it is necessary to pre-process the data, as described in more detail in Sec. III-C. To describe the data-source origin of a particular EEG epoch, $d_i \in \{1, \dots, D\}$ specifies the data-source ID of X_i. Note that for our problem of interest, the underlying assumption here is s and y as well as d and y being marginally independent, *i.e.* the probability of a certain emotion is the same for all subjects and across all data-sources. We achieve this by balancing the samples with respect to the subject IDs and the data-source IDs, as further described in Sec. III-D.

We can distinguish two approaches to combine multiple data-sources in a learning pipeline: (1) pre-processing the samples and labels of the data-sources, so that they can be processed by the same encoder framework, and (2) training individual encoder frameworks for each data-source, while ensuring a consistent latent representation among all frameworks. For the scope of this paper, we will investigate the first approach, whereas our adversarial training pipeline is applicable to both. Given training data the aim is to learn a discriminative model that predicts y from observations X. For such a model to be generalizable across datasets, ideally, the predictions should be invariant to d, which will be unknown at test time. Herein, we regard d as nuisance parameters involved in the EEG data generation process and aim to learn a parametric model that can be generalized across different data sources and learns robust representations.

B. Adversarially Learned Invariant EEG Representations

We train a deterministic encoder with parameters θ_{enc} to learn representations $h = f(X; \theta_{enc})$ given the training data. We discuss the encoder network specifications in detail in Sec. II-C. Obtained representations h are used as input separately to both a classifier network with parameters θ_{clf} to estimate y, as well as an adversary network with parameters θ_{adv}, which aims to recover the data-source variable d. Respectively, the classifier and adversary networks estimate the likelihoods $q_{clf}(y|h)$ and $q_{adv}(d|h)$.

Algorithm 1: Adversarial multi-source EEG neural network training scheme.

```
input : $\eta$ learning rate, $\lambda$ adversarial reg. weight
for epoch $\leftarrow 0$ to epochs do
    for batch $\leftarrow 0$ to batches do
        # Forward pass the input through the encoder
          to compute representation
        $h \leftarrow f(X; \theta_{enc})$
        # Update parameters of the adversary
        $\theta_{adv} \leftarrow \theta_{adv} - \eta \nabla_{\theta_{adv}} \mathbb{E}[-\lambda \log q_{\theta_{adv}}(d|h)]$
        # Update parameters of the encoder and
          emotion classifier
        $\theta_{enc} \leftarrow \theta_{enc} - \eta \nabla_{\theta_{enc}} \mathbb{E}[-\log q_{\theta_{clf}}(y|h) +$
          $\lambda \log q_{\theta_{adv}}(d|h)]$
        $\theta_{clf} \leftarrow \theta_{clf} - \eta \nabla_{\theta_{clf}} \mathbb{E}[-\log q_{\theta_{clf}}(y|h) +$
          $\lambda \log q_{\theta_{adv}}(d|h)]$
    end
end
```

We aim to filter factors of variation caused by d within h. To achieve this, we propose an adversarial learning scheme. The adversary network is trained to predict d by maximizing the likelihood $q_{adv}(d|h)$. At the same time, the encoder is trying to conceal information regarding d that is embedded in h by minimizing that likelihood, as well as retaining sufficient discriminative information for the classifier to estimate y by maximizing $q_{clf}(y|h)$. Overall, we simultaneously train these networks towards the following objective:

$$\min_{\theta_{enc}, \theta_{clf}} \max_{\theta_{adv}} \mathbb{E}[-\log q_{\theta_{clf}}(y|h) + \lambda \log q_{\theta_{adv}}(d|h)] \quad (1)$$

with θ_{enc} represented through $h = f(X; \theta_{enc})$. A higher adversarial regularization weight $\lambda > 0$ enforces stronger invariance from d trading-off with discriminative performance. We use stochastic gradient descent (or ascent) alternatingly for the adversary and the encoder-classifier networks to optimize Eq. (1). Note that setting the regularization parameter $\lambda = 0$ indicates training a regular neural network.

C. Neural Network Architecture and Training

Proposed model is illustrated in Figure 1. The encoder network maps each input sample X_i to a latent representation vector h_i, which is used as input to two separate single dense layer classifiers. The first classifier, *i.e. emotion classifier*, predicts an EEG class label y_i, *i.e.,* an emotional label. The second classifier, *i.e. adversary network*, serves as an EEG domain classifier and predicts the data-source ID d_i of the current training data sample. To solve the objective in Eq. 1, we update the parameters of the adversary network (i.e., domain classifier) and the encoder-emotion classifier network pair alternatingly on each batch. Our model training pipeline is outlined in Algorithm 1. While the proposed architecture is not restrictive to any neural network specification, during our evaluations, for the encoder we used the state-of-the-art convolutional DeepConvNet EEG encoder backbone [14].

III. Experimental Study Design

A. Datasets

We regard four commonly-used and open-source EEG-based emotion recognition experiment datasets as our data-sources, namely SEED [15], SEED-IV [16], DEAP [17], and DREAMER [18]. All datasets contain EEG signals recorded from multiple subjects that were exposed to audio-visual stimuli such as music videos. The EEG signals are labeled with the emotion, that the subject is assumed to have felt during recordings. The datasets mainly differ from each other in three points. First, the experimental setup used for recording, including the number of sessions performed per subject or the used emotional stimuli. Second, the characteristics of the EEG signals, including the number of electrodes (channels) used or the sampling rate. And third, the emotion representation model used to determine the ground truth for the signals. An overview over the specifications of the used datasets can be found in Table I.

B. EEG Pre-Processing

We use only channels C_i which are within all datasets, *i.e.,* $C = \{$'AF3', 'AF4', 'F3', 'F4', 'F7', 'F8', 'FC5', 'FC6', 'O1', 'O2', 'P7', 'P8', 'T7', 'T8'$\}$. Furthermore, we downsample all recordings to the minimum sampling rate of the datasets, *i.e.,* $128\,\mathrm{Hz}$. This downsampling procedure ensures that the model can analyze the EEG time-frequency patterns coherently with the same encoder architecture. The non-zero averages of some of the EEG Signals would lead to increased activation within the neural network. Therefore we calculate the mean value of each channel during the first three seconds of each experiment and subtract it from the whole time series. As some of the EEG signals provided to us are already bandpass filtered using different cut-off frequencies, we bandpass-filter the signals again, using a Butterworth bandpass filter, preserving the smallest common frequency-band all examples contain, *i.e.,* $4\,\mathrm{Hz}$ to $45\,\mathrm{Hz}$.

Finally, all the time series are cut into 2 seconds non-overlapping windows, resulting in a data sample space of dimension $\mathbb{R}^{n\times 14\times 256}$ where n is the number of window segments [19]. Note that in doing so, we make a rather weak assumption that the emotion representation in EEG is stable throughout the experiment, which makes the problem harder for us with the presence of noisy labels.

Overall, we note that through the downsampling and channel selection (least common divisor approach), we discard valuable (high-frequency) EEG information, which poses a limitation of our model's classification performance.

C. Emotion Category Label Conversion

To date, no unified emotion model across datasets exists, and the various established models can often only be partially compared or mapped into one another. Among the datasets we used, two (SEED and SEED IV) employ a discrete state emotion model. In contrast, DEAP and DREAMER used a dimensional model by assessing each emotion by a quantitative expression in several dimensions. The three established dimensions *Valence, Arousal*, and *Dominance* are used in both DEAP and DREAMER.

For our experimental studies, we transformed the different emotion representations into a common representation. Since the discrete states are not differentiated enough to be reasonably mapped into a dimensional model, we converted all representations into a discrete emotion model with three states (*negative, neutral, positive*). As the SEED dataset already uses these states, no transformation was necessary. By assuming that in the dimensional emotion models of DEAP and DREAMER there are four clusters associated with the states *sad*, *fear*, *happy*, and *neutral* we first transferred these representations into a discrete emotion representation model using k-means clustering. To map the states of SEED-IV to our emotion label representation, we made the rather reasonable assumption that *fear* and *sad* are negative emotions and *happy* is a positive emotion. Merging the two negative states, we were then able to transform the label representation into the required emotion state model.

D. Balancing the Samples Across Data Sources

As described in Sec. II-A, we assume y_i and s_i as well as y_i and d_i to be marginally independent. To obtain the same distributions $p(y|s_i)$ for all subjects s_i and $p(y|d_i)$ for all data-source IDs d_i, we balanced the samples X_i with respect to the emotion label first for all subjects individually and later for all data-sources individually. We also took the same number of subjects with the same data-source IDs, giving us a fully-balanced dataset. Using a fully balanced and stratified dataset as such allows us to eliminate biased predictions due to imbalanced samples and ensures that our approach is enforced to not biased on certain participants, data-sources or emotion class labels.

E. Experimental Configurations

We evaluated our model using (1) pre-processed EEG time series in conjunction with the deep neural network (DNN) architecture, and (2) manually extracted power spectral density (PSD) features from the preprocessed time series as input [16]. [1] In order to also test binary classification performance, in a different set of experiments we omitted the observations with a neutral emotion label and evaluated binary classification using the same time-series DNN architecture.

We performed five repetitions of each experiment by using 60% of the preprocessed dataset as the training set, 20% as the validation set, and 20% as the test set. We ensured that the specified requirements from Sec. III-D was held for each of these sets. Maximum number of epochs was always set to 500 with validation loss based early stopping (which generally resulted in completion around 50 epochs).

[1] PSD features were calculated within the delta ($1\,\mathrm{Hz}$ to $4\,\mathrm{Hz}$), theta ($4\,\mathrm{Hz}$ to $7\,\mathrm{Hz}$), alpha ($8\,\mathrm{Hz}$ to $13\,\mathrm{Hz}$), beta ($13\,\mathrm{Hz}$ to $30\,\mathrm{Hz}$), and gamma ($> 30\,\mathrm{Hz}$) band for each sample and channel individually.

TABLE I

Details on the used four dataset specifications.

		SEED [15]	**SEED-IV** [16]	**DEAP** [17]	**DREAMER** [18]
Exp. Setup	subjects (male / female)	15 (7/8)	15 (7/8)	32 (16/16)	23 (14/9)
	sessions per subject	3	3	1	1
	trials per session	15	24	40	18
	stimuli	film clips	film clips	music videos	film clips or music videos
	provided stimuli duration	∼4 min	∼2 min	1 min	∼1 min to 7 min
EEG	number of channels	62	62	32	14
	sampling rate	200 Hz	200 Hz	128 Hz	128 Hz
	freq. filtering	0 Hz to 75 Hz	1 Hz to 75 Hz	4 Hz to 45 Hz	-
	baseline removal	-	-	3s	-
Labels	emotion representation model	discrete	discrete	dimensional	dimensional
	self-analysis	No	No	Yes	Yes
	discrete or continuous	discrete	discrete	continuous (1–9)	discrete (1–5)
	states / dimensions	*Negative, Neutral, Positive*	*Sad, Fear, Neutral, Happy*	*Valence, Arousal, Dominance, Liking, Familiarity*	*Valence, Arousal, Dominance*

IV. Experimental Results

A. Investigating Domain-Specific Leakage during Training

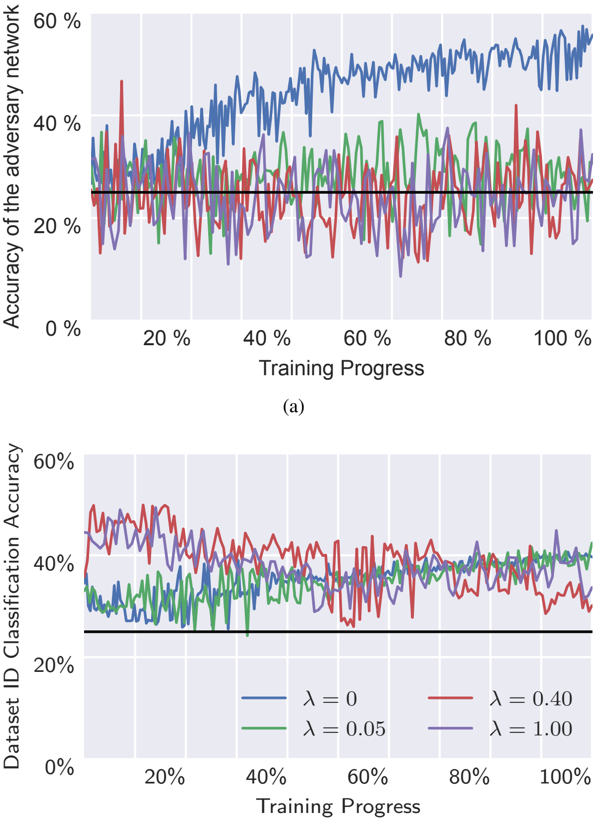
For preliminary verification purposes, we monitored the dataset domain specific information leakage throughout training. We assess this by observing (1) the predictions made by the adversary network throughout epochs, as well as (2) an independent naïve Bayes classifier that is fitted per epoch on the current latent representation to predict the dataset ID.

Figure 2(a) illustrates the prediction accuracies of the adversary network during training. Note that for the baseline model with $\lambda = 0$, an adversary was still trained alongside the classifier to simply monitor d_i-relevant information leakage, without impacting the total loss or gradient-based parameters updates of parameters. We observe that adversarially censored models yield chance-level dataset prediction accuracies, whereas the baseline models show undesired dataset-relevant information leakage throughout training.

We present the results of the independently epoch-wise fitted naïve Bayes classifier in Figure 2(b). We observe that for higher λ values (hence imposing stronger domain-invariance) estimated leakage starts to decrease with trained epochs, which again implies that our approach leads the encoder to reduce the d-relevant leakage in the latent space.

B. Impact of Adversarial Learning on Classification

The classification performance of the emotion classifier depends on the choice of the hyperparameter λ, due to the revealed influence of the d-invariance imposing optimization scheme. Figure 3 shows final test set accuracies of the two classifier ends (emotion and dataset ID classification) of the overall architecture for different λ choices. We consistently observe that the accuracy of the emotion classifier is not significantly impacted with increasing λ, however then starting to decrease due to high adversarial censoring leading to loss of emotion-relevant discriminative information in the latent representations. Regarding the accuracy of the domain (data-source ID) classifier, censoring accordingly with $\lambda > 0$ leads to the data-source no longer be meaningfully decoded by the adversary network, while there was an observed >50% data-source ID classification accuracy by the domain classifier for $\lambda = 0$ baseline models, *i.e.*, regular CNNs.



(a)

(b)

Fig. 2. Domain-relevant leakage throughout training by (a) observing accuracy of the adversary network, (b) fitting Naïve Bayes classifiers to predict d from h, for different adversarial censoring hyperparameter choices λ. The training progress is normalized to percentage by the early stopping end epoch. The black line indicates the chance level.

C. EEG Classification Results

Our method works on a very restricted dataset as described in Sec. III to test representation transfer capabilities across four emotion recognition datasets. Since in our experiments we utilize intersecting subsets of channels in each data-source and filter accordingly, as well as discard observations

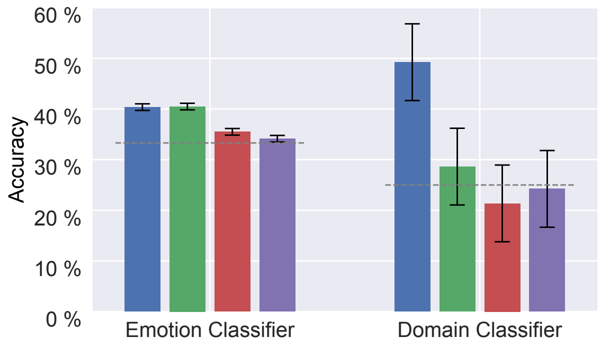


Fig. 3. Comparison of the mean emotion classification and data-source identification accuracies for different hyperparameters λ, averaged over 5 runs. Horizontal dashed lines represent the chance-level accuracy, and black solid lines show the empirical standard deviation.

from specific classes for stratified sampling, the actual emotion classification task becomes highly challenging.

Table II shows averaged accuracies for the adversarially learned model, as well as the baseline global model. Our models achieve an above-chance classification performance for emotion recognition across all four datasets. We further showed that invariant models can be learned by reducing the leakage and maintaining a similar emotional classification quality to $\lambda = 0$ cases (cf. Figure 3).

V. Discussion & Conclusion

In this paper we explore robustly transferable patterns across multiple EEG emotion recognition data-sources. We present an adversarial learning framework to unify different EEG data-sources and labels for multi-source transfer learning by finding data-source-invariant shareable information for multiple EEG-related tasks. Our approach makes significant pre-processing steps to unify the data basis for multi-source transfer learning. Thereby, the results indicate that the pre-processing comes at the cost of classifier performance overall. However our adversarial censoring approach achieves the same classification performance as simply pooling the data domains together (*i.e.*, training regular CNNs with pooled datasets with $\lambda = 0$) while giving us the opportunity to restrict the representation to be highly data-invariant (35% leakage). Our implementations are publicly available at: `https://github.com/philipph77/ACSE-Framework`.

Our work can be extended by adapting the encoder framework to be able to use different EEG input shapes according to the specified data-source, and as a result, different number of channels and sampling frequencies can be learned. We envision an adversarial shared-private model similar to [20] where some channels are shared among data-sources (as in our approach) but private (data-source-specific) input can be incorporated. Our approach can also easily be adapted to learn representations that are invariant corresponding to other EEG variation factors *e.g.*, participant ID, by adding an additional adversarial classifier [21], [22].

TABLE II

Mean emotion classification accuracy by the adversarially learned model and a baseline global model trained without adversarial censoring, averaged over 5 runs.

	Time-Series DNN	PSD Features MLP	Time-Series DNN (binary)
Global	40.37%(±0.65%)	40.26%(±0.36%)	57.63%(±0.77%)
Adversarial	40.48%(±0.70%)	38.74%(±0.65%)	58.17%(±1.63%)

References

[1] S. Machado *et al.*, “EEG-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation,” *Reviews in the Neurosciences*, vol. 21, no. 6, pp. 451–468, 2010.

[2] L. Xu *et al.*, “Cross-dataset variability problem in EEG decoding with deep learning,” *Frontiers in Human Neuroscience*, vol. 14, 2020.

[3] O. Özdenizci and D. Erdoğmuş, “Information theoretic feature transformation learning for brain interfaces,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 69–78, 2019.

[4] E. Tzeng *et al.*, “Adversarial discriminative domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[5] F. Zhuang *et al.*, “A comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[6] Z. Wan *et al.*, “A review on transfer learning in EEG signal analysis,” *Neurocomputing*, vol. 421, pp. 1–14, 2021.

[7] S. M. Peterson *et al.*, “Generalized neural decoders for transfer learning across participants and recording modalities,” *Journal of Neural Engineering*, vol. 18, no. 2, p. 026014, 2021.

[8] O. Özdenizci *et al.*, “Learning invariant representations from EEG via adversarial inference,” *IEEE Access*, vol. 8, pp. 27 074–27 085, 2020.

[9] ——, “Adversarial deep learning in EEG biometrics,” *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 710–714, 2019.

[10] J. Li *et al.*, “Multisource transfer learning for cross-subject EEG emotion recognition,” *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3281–3293, 2019.

[11] K. Ross *et al.*, “Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data,” *arXiv preprint arXiv:2008.10726*, 2020.

[12] P. Rodrigues *et al.*, “Dimensionality transcending: a method for merging BCI datasets with different dimensionalities,” *IEEE Transactions on Biomedical Engineering*, 2020.

[13] D. Bethge *et al.*, “Domain-invariant representation learning from EEG with private encoders,” *arXiv preprint arXiv:2201.11613*, 2022.

[14] R. T. Schirrmeister *et al.*, “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[15] R.-N. Duan *et al.*, “Differential entropy feature for EEG-based emotion classification,” in *6th International IEEE/EMBS Conference on Neural Engineering*, 2013, pp. 81–84.

[16] W.-L. Zheng *et al.*, “Emotionmeter: A multimodal framework for recognizing human emotions,” *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[17] S. Koelstra *et al.*, “DEAP: A database for emotion analysis using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[18] S. Katsigiannis and N. Ramzan, “DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[19] H. Candra *et al.*, “Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine,” in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015, pp. 7250–7253.

[20] P. Liu *et al.*, “Adversarial multi-task learning for text classification,” *arXiv preprint arXiv:1704.05742*, 2017.

[21] M. Han *et al.*, “Disentangled adversarial autoencoder for subject-invariant physiological feature extraction,” *IEEE Signal Processing Letters*, vol. 27, pp. 1565–1569, 2020.

[22] ——, “Universal physiological representation learning with soft-disentangled rateless autoencoders,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 2928–2937, 2021.

Interpretable Time-Dependent Convolutional Emotion Recognition with Contextual Data Streams

DAVID BETHGE, LMU Munich, Germany
CONSTANTIN PATSCH, TUM, Germany
PHILIPP HALLGARTEN, Porsche, TUM, Germany
THOMAS KOSCH, HU Berlin, Germany

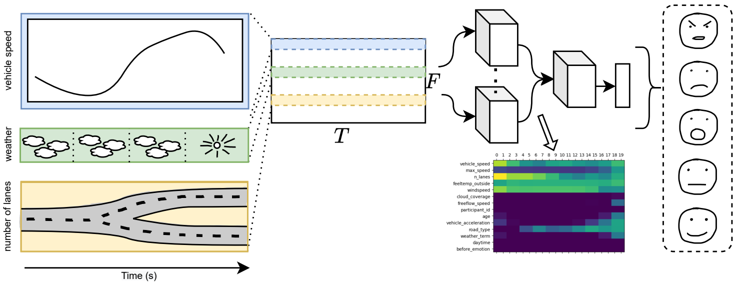


Fig. 1. System architecture overview. ITER takes any multivariate time-series as an input (here: contextual vehicle variables) and performs a time-series classification (here: predicting driver emotions) while providing explainable feature maps, which display feature importance of the model's prediction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

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 transformer networks can learn powerful time-series patterns while showing classification decisions and feature importances. We present a novel transformer-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*.

Additional Key Words and Phrases: interpretable emotion classification, driver emotion recognition, time-series classification, contextual computing

ACM Reference Format:
David Bethge, Constantin Patsch, Philipp Hallgarten, and Thomas Kosch. 2023. Interpretable Time-Dependent Convolutional Emotion Recognition with Contextual Data Streams. In *CHI'23 Extended Abstracts: Late Breaking Works, April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/XXXXXXX.XXXXXXX

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 [15]) and reduce privacy concerns. Already in 2000, Picard [28] 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 [18, 21], voice analysis [14], self-reports [5], or physiological sensing [10].

Facial expressions have a long tradition as an indicator for the expressed emotions [12] 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 [8, 23, 26] or infrared cameras [13]. However, facial expressions can be misinterpreted without involving the user's context [18] and subjective interpretation [19]. 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 [3, 11, 33], 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 [36].

Various emotions can be elicited depending on the user's context. For example, driving is a common use case when studying user emotions [4, 7, 33]. Thus, various datasets exist that allow to compare the performance of different classification techniques [2, 4]. In this context, previous research hypothesizes that the driving behavior, style, and context are indicative of the currently perceived emotions [25]. 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. 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 [22, 29], backpropagation [30, 35] and input-masking based [27, 32] approaches. Using transformer 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

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

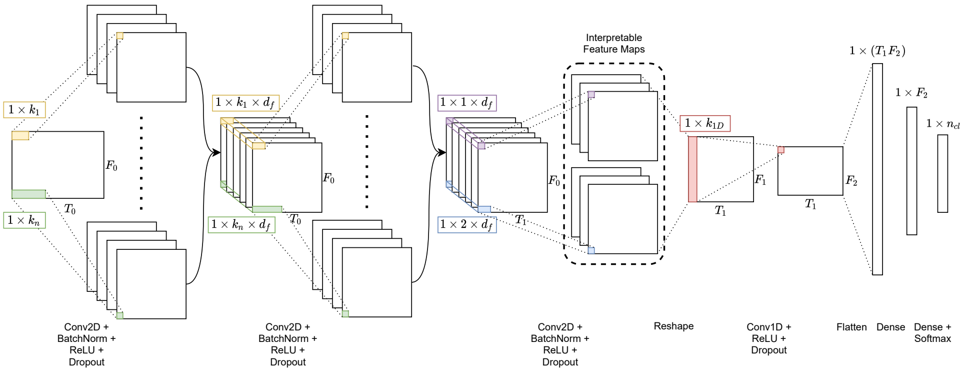


Fig. 2. Network architecture of our interpretable time-series classification system. The architecture consists of two stages, where the first consists of parallel 2D convolution layers that preserve the feature dimension. The second stage consists of a 1D convolution layer, where the resulting feature maps are flattened and forwarded to a dense and the final classification layer.

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 [30] combined with a forward-scoring method [30] 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. [31] 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) 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 classification 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 [31], we define an architecture that captures a maximal variation of receptive fields while using a minimal amount of layers by applying different kernel sizes in parallel at several stages in the network. While [31] 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 classification decision. This is essential for the user to infer the influence of the time-context instances on the emotion classification.

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 specific features over time to a particular classification decision.

We determine activation feature maps based on the Grad-CAM method introduced by [30] and the Score-CAM method from [32]. 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 classification of the specific y class. On the other hand, we use the Score-CAM method from [32], 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 [4] and is published open-source[2]. In total, 12 participants (2/12 self-identified 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 vehicles [9]. We consider the emotions 'angry', 'disgust',

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

'happiness', 'neutral', and 'surprise'. A speech-to-text engine from the smartphone audio recording is used to extract the emotion label y, as the participant had to verbally provide their discrete emotion label every 60 seconds 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 presented in Table 2. We refer the reader to the original paper for more details on the dataset.

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 [34]. This method ensures that we can prevent discontinuous changes between valid recorded and imputed values. As our architecture expects a fixed input size, we use a sliding window approach similar to [20] with a stride of 1 on the multivariate time series to generate samples of size $F \times T$. The corresponding label for each window is defined 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].

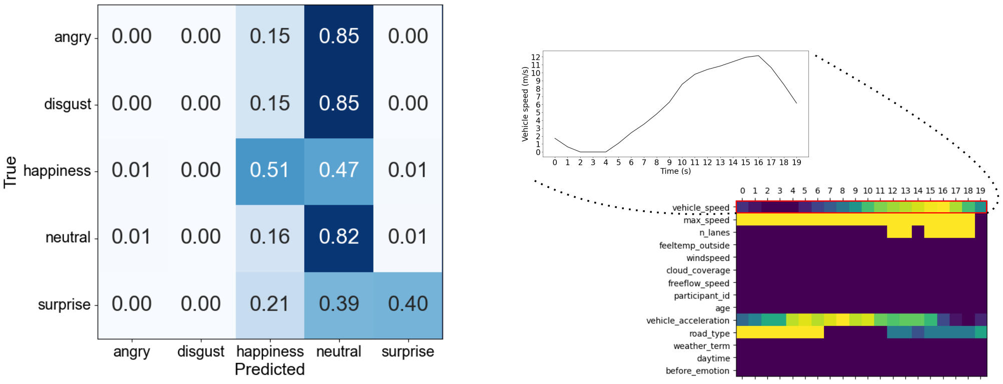
4 RESULTS

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 [4]. 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. The results of our architecture are depicted in the confusion matrix in Figure 3a. Overall our model achieves an accuracy of 70% and a F_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 classification performance for 'angry' and 'disgust' states. In particular, this is likely due to the skewed distribution of subjectively felt emotions in-the-wild. The emotions 'angry' and 'disgust' are underrepresented in the dataset 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 classifier of VEmotion [4] and the Microsoft Face Recognition API [24], which both do not consider time during classification. 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 F_1 score of our approach are 2% lower than the ones of the VEmotion model. The difference 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 efficiently, 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.

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



(a) Normalized confusion matrix.

(b) Visualization of the input features.

Fig. 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.

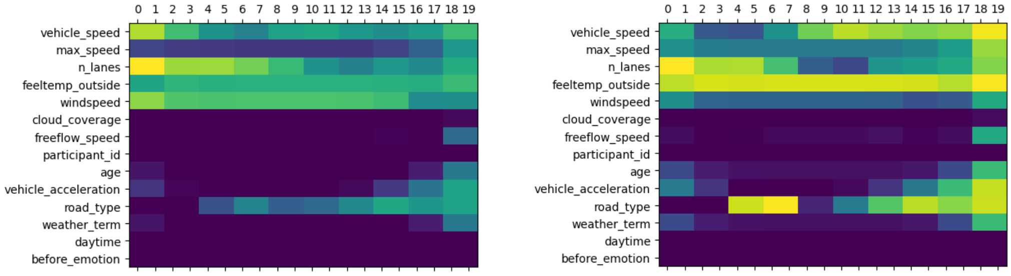
Emotion classification with the Microsoft Face Recognition API based on facial video data is outperformed by our system by 14% in terms of accuracy and 18% in terms of the F_1 score. This indicates that facial expressions in a driving context are less expressive than time-dependent context features. When comparing our architecture to the two-layer 1D-CNN architecture, we can see that ITER achieves a 5% better accuracy and a 6% better F_1 score. This increase implies that capturing a large range of receptive fields improves classification performance. Similarly, the two-layer LSTM model struggles to classify infrequent classes, which is indicated by the 14% lower F_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 3% lower F_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 classifier introduced by [4]. 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 classification. 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 for the classification decision.

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 vehicle speed exemplarily. The ascending time scale corresponds to the progress towards the most recent timestep. We normalized the

Table 1. Emotion recognition performances of different classification models. The table further includes the models' properties time dependency and interpretability. Hereby, interpretability refers to the feature-wise explanation for classification decisions based on saliency maps. We compare our system to VEmotion [4], a facial expression classification system Face [24], a LSTM deep learning model [16], a 1d-CNN [17], MTEX-CNN [1].

	VEmotion	Face	LSTM	1D-CNN	MTEX-CNN	ITER **(ours)**
accuracy	.72	56	.64	.65	.68	.70
F_1 score	.71	51	.55	.63	.66	.69
time dependency	✗	✗	✓	✓	✓	✓
interpretability	✗	✗	✗	✗	✓	✓



(a) Feature map based on Grad-CAM

(b) Feature map based on Score-CAM

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

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 specific 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 especially *vehicle speed*, *number of lanes*, *temperature*, *wind speed* and *road type* seem to be essential for the classification decision of this happy sample. Moreover, when comparing *road type* time instances of the feature map with the input, especially changes in *road type* seem relevant for the classification decision. Furthermore, the model puts a higher focus on low acceleration values as the specific 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 differs 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 classification decision is comparable to the Grad-CAM feature map. We showed that we could extract time-dependent feature interpretations for an emotion classification in the form of saliency maps. Furthermore, we provided sample-specific explanations for a classification decision based on contextual features. The two proposed feature map generation methods have shown valid outputs and can both be used for emotion classification 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 specific 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 [6].

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 specific emotional triggers. These could then be consumed by a routing algorithm that adapts correspondingly, e.g., by avoiding specific road attributes. In the case of developing empathic car interfaces, being able to detect emotions and interpret the classification 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.

Limitations and Future Work. In general, the features corresponding to an emotion that the model explicitly finds important might only partly match with the features that the driver perceives as most influential 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, the interpretation of emotion for a limited number of features might only reflect 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 specific 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 offer 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 classification 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 classifies 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 specific classification 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 classification 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.

REFERENCES

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

[2] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 159 (jan 2023), 30 pages. https://doi.org/10.1145/3569466

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

[4] 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.

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

[6] Michael Braun, Jingyi Li, Florian Weber, Bastian Pfleging, 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

[7] Michael Braun, Florian Weber, and Florian Alt. [n. d.]. Affective Automotive User Interfaces - Reviewing the State of Emotion Regulation in the Car. In *To appear in ACM Copmuting Surveys*.

[8] 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

[9] 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

[10] 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

[11] 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

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

[13] 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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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

[19] 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

[20] 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.

[21] 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.

[22] 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.

[23] 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

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

[25] 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

[26] 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

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

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

[29] 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.

[30] 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.

[31] 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).

[32] 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.

[33] 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

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

[35] 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.

[36] 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 $F_0 \times T_0$, where F_0 is the feature dimension, and T_0 is the time dimension. In our case, T_0 is set to a temporal window size of 20, and F_0 is equivalent to 14 features. The choice of F_0 depends on the context features that are recorded in the dataset, while the choice of T_0 has been determined experimentally (this is further justified in section 5). The first stage of the architecture consists of N parallel 2D convolution layers with different kernel sizes $1 \times k_n$ with $n \in \{1, ..., N\}$. k_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 $F_0 \times T_0$ and allow the concatenation of feature maps resulting from different kernel sizes. After concatenating the number of d_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 $1 \times k_n$ with respect to the d_f feature maps. By using same padding and a stride size of 2, the feature map sizes result in $F_0 \times T_1$. 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 $F_1 \times T_1$.

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 k_{1D} that is used to account for dependencies of features between different time steps. The resulting $F_2 \times T_1$ feature map is flattened in the last stage to be a suitable input to the following dense layer of size $1 \times F_2$. As the last step, we define a dense classification layer for the number of classes n_{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 [30]. The weight α_k^c of each feature map A^k is determined by

$$\alpha_k^c = \frac{1}{F_k T_k} \sum_{i=1}^{F_k} \sum_{j=1}^{T_k} \frac{\partial y_c}{\partial A_{ij}^k}, \tag{1}$$

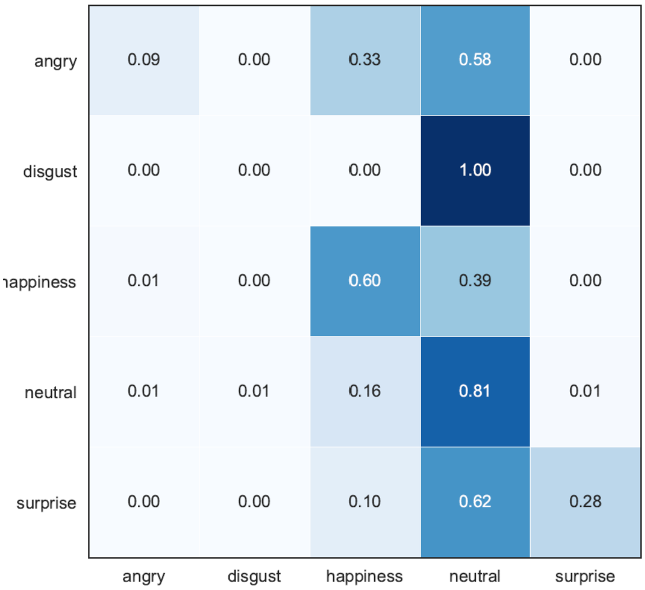
where we calculate the gradients of the class y_c with respect to the activations A_{ij}^k and average over the number of time instances of all features. A high value of α_k^c would indicate a strong contribution of the individual instances in the feature map A^k towards the classification of y_c. We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

On the other hand, we use the Score-CAM method from [32] 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 I_M^k defined by

$$I_M^k = I \circ M^k, \tag{2}$$

where I represents the multivariate time window input and M^k defines activation maps A^k that are upsampled to the input and normalized. The masked inputs are then fed into the model to determine their classification score β_k^c for class y_c. The higher the classification score of a masked input I_M^k, the stronger A^k gets weighted. Like Grad-Cam, a ReLU function is applied to the sum over the weighted activation maps $\beta_k^c A^k$.

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



	angry	disgust	happiness	neutral	surprise
angry	0.09	0.00	0.33	0.58	0.00
disgust	0.00	0.00	0.00	1.00	0.00
ıappiness	0.01	0.00	0.60	0.39	0.00
neutral	0.01	0.01	0.16	0.81	0.01
surprise	0.00	0.00	0.10	0.62	0.28

Fig. 5. 10-fold cross-validation results of VEmotion [4].

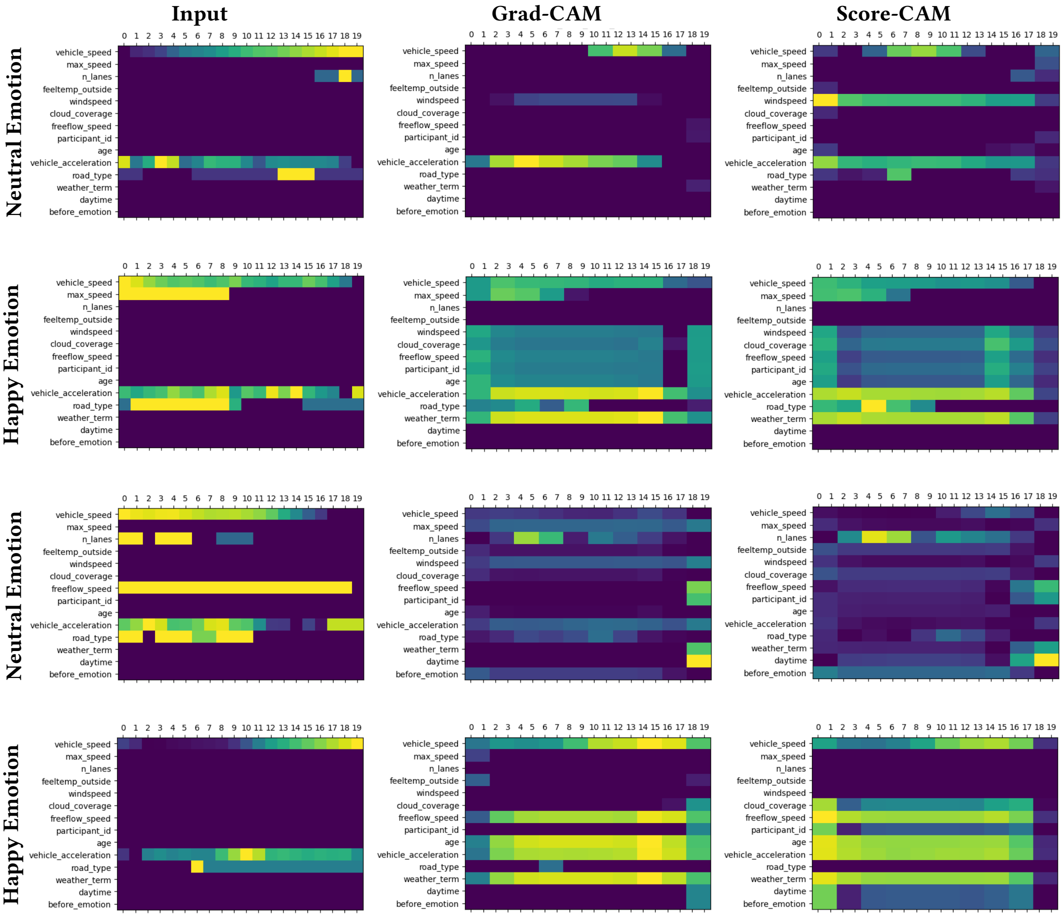


Fig. 6. The rows contain the input and feature maps corresponding to an emotion sample. The first column corresponds to the normalized input sample, the second column to the feature maps resulting from the Grad-CAM approach and the third column to the feature map based on the Score-CAM approach.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt wurde.

München, den 1. August 2023

David Bethge